Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

DISSERTATION

Hybrid Discontinuous Galerkin Methods for the Wave Equation

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Univ.-Prof. Dipl.-Ing. Dr.techn. Joachim Schöberl
E101
Institut für Analysis und Scientific Computing

eingereicht an der Technischen Universität Wien
bei der Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Martin Huber
Matrikelnummer: 0056265
Wilhelmstraße 5/3
1120 Wien

Wien, am 27. Februar 2013

Kurzfassung

Beim Lösen der Wellengleichung mit klassischen Finiten Elementen wächst für hohe Frequenzen ω die Anzahl der für eine vorgegebene Genauigkeit benötigten Unbekannten aufgrund des sogenannten „pollution effects“ stärker als $\mathcal{O}(\omega^d)$ mit d als Raumdimension an. Als eine Möglichkeit dieses Problem in den Griff zu bekommen, werden in der vorliegenden Dissertation hybride Discontinuous Galerkin Methoden für die skalare und die vektorwertige Wellengleichung vorgestellt.

Üblicherweise beruhen klassische Finite Element Methoden (FEM) auf stetigen Ansatzfunktionen, während Discontinuous Galerkin Verfahren Basisfunktionen verwenden, die über Elementgrenzen hinweg unstetig sind. Erst zusätzliche Strafterme führen hier zu einer stetigen Lösung. Hybride FEMen basieren auf den selben Ansatzräumen, nur wird jetzt die Stetigkeit über zusätzliche Lagrangevariablen, die nur auf den Elementgrenzen leben, erzwungen. Im Falle der Wellengleichung ermöglicht die Einführung eines zweiten Satzes von Lagrangevariablen eine vollständige Elimination der ursprünglichen Freiheitsgrade. Dadurch kann das Problem auf ein viel kleineres, nur für die Lagrangevariablen reduziert werden.

Werden Finite Elemente zum Lösen eines auf einem unendlichen Gebiet gestellten Problems verwendet, wird in der Regel das Rechengebiet eingeschränkt, was transparente Randbedingungen am neu entstandenen künstlichen Rand nötig macht. Als Realisierungsmöglichkeiten für solche Randbedingungen werden sowohl die Hardy-Raum Methode als auch eine Zerlegung des Fernfeldes in ebene Wellen diskutiert. Letztere ist gerade für die Simulation von Beugungsgittern besonders geeignet.

Für die zweidimensionale Helmholtzgleichung lässt sich die hybride FEM mittels einer diskreten Eigenfunktionenbasis besonders effizient umsetzen. Auf einem Rechtecksgitter ermöglichen eindimensionale Eigenwertprobleme, die nur von der Seitenlänge und der Polynomordnung abhängen, die Konstruktion einer solchen Basis. Durch die Eigenfunktionenbasis vereinfacht sich die Assemblierung, und die Elimination der Volumenfreiheitsgrade kann billig durchgeführt werden. Zusammen mit der Tatsache, dass sich das Eigenwertproblem auch für Polynomordnungen größer tausend noch schnell lösen lässt, wird die Verwendung von Elementen sehr hoher Ordnung möglich. Durch das Zulassen von Netzen mit hängenden Knoten kann die exponentielle Konvergenz von *hp*-Methoden augenutzt werden.

Eine besondere Herausforderung ist meist die Lösung des resultierenden linearen Gleichungssystems. Hier bietet die hybride FEM auf natürliche Weise die Möglichkeit, dieses über Krylovraummethoden zusammen mit Gebietszerlegungsmethoden effizient zu lösen. Neben additiven und multiplikativen Schwarz Vorkonditionierern mit lokalen Glättern, sowie einem elementweisen BDDC-Vorkonditionierer wird ein neuer Gebietszerlegungsvorkonditionierer vorgestellt, der in jedem Iterationsschritt Teigebietsprobleme mit Robin Randbedingungen direkt löst und sich folglich bestens zur Parallelisierung eignet. Numerische Experimente verdeutlichen schließlich die guten Konvergenzeigenschaften dieser Löser.

Abstract

When the wave equation is solved for high frequencies ω with a classical Finite Element Method (FEM), the number of unknowns required to resolve the strongly oscillating solution grows due to the so called "pollution effect" faster than $\mathcal{O}(\omega^d)$, where d is the space dimension. At the same time the iterative properties of the linear system of equations worsen. One way to cope with this difficulty are hybrid FEMs. In this thesis, we investigate hybrid FEMs for the scalar and vector valued wave equation, which are equivalent to a discontinuous Galerkin method, based on the ultra weak variational formulation.

In Discontinuous Galerkin and hybrid Discontinuous Galerkin methods the continuity of basis functions is broken across element facets, i.e., the interfaces between them. This is contrary to classical FEMs, which are based on continuous function spaces. While a continuous solution is obtained in Discontinuous Galerkin schemes by additional "penalty" terms, hybrid FEMs reinforce continuity via Lagrange multipliers supported only on element facets. For the wave equation a second set of multipliers is necessary to eliminate the original degrees of freedom cheaply element by element. This approach allows to reduce the system of equations to a much smaller system just for the Lagrange multipliers.

If wave type problems posed on an infinite domain are solved by FEMs, generally, the computational domain is restricted, and appropriate transparent boundary conditions are needed. As a realization of such a boundary condition, we discuss the Hardy space infinite element method, and we adjust it to the introduced hybrid FEMs. Additionally, transparent boundary conditions, based on a plane wave expansion for the far field, are provided. It is shown that they are well suited for the simulation of diffraction gratings.

Apart from this, the work presents an optimized implementation technique of the hybrid FEM for the two dimensional Helmholtz equation. There, a discrete eigenfunction basis, which makes the assembly of the system matrix and the elimination of the interior degrees of freedom computationally inexpensive, is used. For rectangular meshes the construction of such a basis requires the solution of a one dimensional eigenvalue problem for each pair

of edge length and polynomial order. The eigenvalue problem can be solved for polynomial orders up to thousands. Combining this with the cheap assembly and the reduction in problem size, we are able to use very high order basis functions efficiently. By allowing for hanging nodes, we can benefit from exponential convergence of *hp*-methods.

A very challenging point is solving the resulting system of equations. Since the hybrid formulation provides appropriate interface conditions, an efficient iterative solution with Krylov space methods combined with domain decomposition preconditioners is possible. Apart from multiplicative and additive Schwarz block preconditioners with local smoothers or an element wise BDDC preconditioner, a new Robin type domain decomposition preconditioner is constructed. This preconditioner solves in each iteration step local problems on subdomains by directly inverting the system matrix. Thus, it is well suited for parallel computations. Good convergence properties of these iterative solvers are demonstrated by numerical experiments.

Acknowledgements

Firstly, I would like to express my special thanks to J. Schöberl for supervising and inspiring my work and for many time-intensive discussions and helpful remarks given during the last years. At the same time I want to thank S. Rotter for co-refereeing this thesis.

Many thanks go to my colleagues an former colleagues H. Egger, A. Hannukainen, K. Hollaus, G. Kitzler, L. Nannen, A. Pechstein and S. Zaglmayr for countless inspiring discussions, for proofreading the thesis and for many helpful comments, which improved my work significantly. Especially, I want to thank for the friendly atmosphere throughout the last years.

I want to acknowledge the scientific and physical environment provided by the Radon Institute for Computational Mathematics of the Austrian Academy of Science, by the Institute MathCCES at RWTH Aachen and the Institute for Analysis and Scientific Computing at TU Wien. Furthermore, I am grateful for the financial support by the Austrian Science Fund "Fonds zur Förderung der wissenschaftlichen Forschung in Österreich" (FWF) via the Start Project Y-192 at the beginning of my work.

Finally, special thanks go to my parents, my brother and all my friends, who helped me to overcome many difficult situations during the last years. Without their help I would not have been able to finish this work.

Contents

1 Introduction 1

2 From Maxwell's equations to Helmholtz equation 6

2.1 Maxwell's equations . . . 6
  2.1.1 Maxwell's equations in media . . . 7
  2.1.2 Interface and boundary conditions . . . 8
  2.1.3 A time harmonic setting . . . 11
  2.1.4 Transversal electric and transversal magnetic modes . . . 12
  2.1.5 Electromagnetic energy . . . 13
2.2 The Helmholtz equation . . . 14
  2.2.1 The primal form of the Helmholtz equation . . . 14
  2.2.2 The mixed form of the Helmholtz equation . . . 15
2.3 The vector valued wave equation . . . 16
  2.3.1 The primal form of the vector valued wave equation . . . 16
  2.3.2 The mixed form of the vector valued wave equation . . . 17

3 The variational framework 18

3.1 Some function spaces . . . 18
  3.1.1 The space $L^2(\Omega)$ . . . 19
  3.1.2 The space $H^1(\Omega)$ . . . 19
  3.1.3 The space $H(\mathrm{div}, \Omega)$ . . . 21
  3.1.4 The space $H(\mathrm{curl}, \Omega)$ . . . 23
3.2 Variational formulations . . . 25
  3.2.1 Existence and uniqueness for primal and mixed problems . . . 25
  3.2.2 The variational form of the primal Helmholtz problem . . . 27
  3.2.3 The variational form of the mixed Helmholtz problem . . . 29

3.2.4 The variational form of the primal vector valued wave problem . . . 31
3.2.5 The variational form of the mixed vector valued wave problem . . . 34

4 The finite element method 35
4.1 Basic ingredients 35
4.1.1 The Galerkin framework and approximation results 35
4.1.2 The triangulation and the finite element 37
4.1.3 The reference element 38
4.2 Some conforming finite elements 40
4.2.1 An H^1 conforming finite element 40
4.2.2 An $H(\mathrm{curl})$ conforming finite element 42
4.2.3 An $H(\mathrm{div})$ conforming finite element 44
4.2.4 An L^2 conforming finite element 46

5 A hybrid finite element method 47
5.1 An introduction to hybridization 48
5.2 A hybrid finite element method for the Helmholtz equation 50
5.2.1 Hybridizing the Helmholtz equation 50
5.2.2 Consistency, existence and uniqueness of the hybrid formulation . . 54
5.2.3 The discrete finite element spaces 56
5.2.4 Conservation of energy 58
5.3 A hybrid finite element method for the vector valued wave equation 62
5.3.1 Hybridizing the vector valued wave equation 62
5.3.2 Consistency existence and uniqueness of the formulation 65
5.3.3 The discrete finite element spaces 67
5.3.4 Conservation of energy 69

6 Iterative solvers 73
6.1 State of the art 75
6.2 Static condensation 78
6.3 Schwarz preconditioners 80
6.3.1 Additive Schwarz preconditioners 80
6.3.2 Multiplicative Schwarz preconditioners 83
6.3.3 Schwarz methods for the mixed hybrid formulation 85
6.4 The BDDC preconditioner 86

6.4.1 Theoretical framework 86
6.4.2 The BDDC preconditioner for the mixed hybrid formulation 91
6.5 A Robin type domain decomposition preconditioner 93
6.6 Numerical results 95
6.6.1 Comparison of the preconditioners 95
6.6.2 Large scale examples for the Helmholtz equation 102
6.6.3 Large scale examples for the vector valued wave equation 105

7 Hardy space infinite elements 109
7.1 Hardy space infinite element method in one dimension 110
7.2 The mixed hybrid formulation in the exterior domain 114
7.2.1 The formal variational formulation 115
7.2.2 The finite element spaces 116
7.2.3 The integrals involved 120
7.3 The finite elements 122
7.3.1 The finite element basis functions 122
7.3.2 The element matrices 124

8 A tensor product implementation 126
8.1 Preliminaries 127
8.1.1 The underlying mixed hybrid formulation 127
8.1.2 The discrete finite element spaces 129
8.2 The eigenfunction basis 130
8.2.1 Legendre and integrated Legendre polynomials 130
8.2.2 The eigenvalue problem 132
8.2.3 The basis functions 134
8.3 The linear system of equations 135
8.3.1 The coupling of the volume basis functions 136
8.3.2 The coupling of the facet basis functions 137
8.3.3 The coupling between facet and volume degrees of freedom 139
8.3.4 Solving the system of equations 140
8.4 Hanging nodes 142
8.4.1 The mesh 142
8.4.2 The structure of the system matrix 143
8.4.3 The preconditioner 145

8.5 Hardy space infinite elements . . . 146
8.5.1 The mixed hybrid formulation . . . 146
8.5.2 The eigenfunction basis . . . 147
8.5.3 The linear system of equations . . . 149
8.5.4 The preconditioner . . . 153
8.6 Numerical results . . . 153
8.6.1 Absorbing boundary conditions . . . 153
8.6.2 Hardy space infinite elements . . . 160

9 Simulation of diffraction gratings 164
9.1 Preliminaries . . . 166
9.1.1 The geometry . . . 166
9.1.2 The Rayleigh expansion . . . 167
9.2 Problem formulations and discretizations for one dimensional gratings . . . 171
9.2.1 The classical formulation . . . 171
9.2.2 The weak formulation . . . 172
9.2.3 The incorporation of quasi periodic boundary conditions . . . 173
9.2.4 Incorporation of the interface conditions by Nitsche's method . . . 174
9.3 Problem formulations and discretizations for biperiodic gratings . . . 177
9.4 Modeling of layers . . . 179
9.4.1 TE-modes . . . 180
9.4.2 TM-modes . . . 182
9.5 Numerical examples . . . 182
9.5.1 A lamellar grating as benchmark problem . . . 183
9.5.2 A large unit cell . . . 185
9.5.3 A biperiodic grating . . . 185

10 Summary and Outlook 189

Bibliography 192

Chapter 1

Introduction

Our daily life is influenced by a big variety of different wave type phenomena. Talking to each other or listening to music is possible only because of sound waves, and the ability to see is based on light waves. If we go surfing on the ocean during the holidays, we can directly experience water waves. But not all types of wave phenomena are useful for humanity. Just imagine the pictures from cities destroyed by earthquakes or a Tsunami. Such seismic waves, acoustic waves or water waves, also known as mechanical waves need a material to propagate. Another large class of waves, not restricted to media, are electromagnetic waves. Devices like a radio or a television set as well as the communication via mobile phones are unthinkable without such waves. Apart from light, ultraviolet radiation, x-rays and gamma rays are other famous examples of electromagnetic waves.

Due to the fact that wave phenomena are present in almost all branches of biological and physical sciences, problems related to waves have a big impact on the design and optimization of devices. Typical questions could be for example how to construct a mobile phone, such that for a strong output signal the field strength in the human head is minimal, or how to design a laser with an optimal output. For industry it is in general much cheaper to solve such problems on a computer instead of carrying out costly experiments or building expensive prototypes. Thus, there is a big demand for simulations of wave problems.

These problems can be best modeled by the Helmholtz equation or, in the case of electromagnetic waves, by Maxwell's equations, leading to a vector valued wave equation. Because the equations can be solved analytically just for very simple settings, numerical methods are needed. Mostly the numerical step is realized by discretizing the underlying partial differential equation, and a possibly linear system of equations is obtained. Nonlinear problems are generally reduced by iterative methods to a sequence of linear problems.

Among finite differences [HT95], finite volumes [PM94] or the boundary element method [N0́1] the finite element method (FEM), which is based on a discretization of the partial differential equation via a variational formulation, is a very powerful tool for solving the Helmholtz equation. The biggest advantage of the finite element method is that it is applicable to a big class of different problems, which can be even nonlinear. Furthermore, it benefits from its flexibility with respect to complex domains, varying coefficients as well as different types of boundary conditions. A profound analysis of the method exists already for various problem settings with important contributions from mechanics, electromagnetics or fluid mechanics. For a detailed introduction to finite elements we recommend the books of [Bra03, BS08].

When the finite element method is applied to a wave type problem posed on an infinite domain, the domain is generally restricted, and on the new artificial boundary appropriate transparent boundary conditions are necessary. Such conditions should let an outgoing wave cross the boundary without or with only little reflections. Easy to implement, but not very effective is a Robin type boundary condition, also known as first order absorbing boundary condition. Much more efficient are the widely used perfectly matched layers [Sim79, Ber94] or Hardy space infinite elements [Nan08, HN09, NS11]. We are going to discuss the latter ones in detail in our work. Additionally, we introduce a method, which realizes transparent boundary conditions by replacing the field outside the computational domain by a plane wave expansion. This technique is well suited for periodic structures.

However, when discretizing the wave equation with a standard numerical method, one faces several difficulties. One is caused by a multiscale character of problems with high frequency waves. For example, when calculating the scattered field of a radar wave at an airplane, the scatterer is several magnitudes bigger than the radar wavelength, which is typically in the centimeter region. The fact that at least a fixed number of unknowns per wavelength has to be used to resolve the highly oscillating solution results in a large system of equations. Thus, the total number of degrees of freedom is at least $N = \mathcal{O}(\omega^d)$ for the angular frequency ω and the space dimension d. The situation gets even worse because of the so called pollution effect [Ihl98]. Especially for low polynomial orders the number of unknowns per wavelength needed to achieve a given accuracy increases with increasing frequency. Although this effect is not so strong for higher polynomial orders, the total number of unknowns grows faster than $\mathcal{O}(\omega^d)$.

These difficulties concerning standard methods have led to the development of a variety of different approaches for solving the wave equation within the finite element framework.

Well known are *hp* methods [IB97] or discontinuous Galerkin methods [FW09], which still use the polynomial approximation for a modified variational formulation. Other methods are based on a set of problem-adapted functions. Most popular among them are the partition of unity method [Mel95, MB96], least square methods [MW99], discontinuous enrichment methods [FHH03, TF06] or the ultra weak variational formulation [CD98, Mon03]. Our work is based on a hybrid discontinuous Galerkin method from [MSS10] which is motivated by hybrid finite elements for the Laplace equation[BF91, CG04].

In an hybrid discontinuous Galerkin method the normal and tangential continuity, respectively, of the unknown field is broken across element interfaces and afterwards enforced again by introducing a set of Lagrangian multipliers, supported only on the element facets. When hybridizing the wave equation, a second set of Lagrangian multipliers representing the normal and tangential component, respectively, of the flux field on the element facets is required additionally. This set allows us, due to local Robin boundary conditions, to eliminate the volume degrees of freedom element by element and to reduce the system of equations to the considerably smaller system just for the Lagrangian multipliers. We will present an optimized version of this method for rectangular meshes, which uses a discrete eigenfunction basis. For such a basis an inexpensive assembly procedure without numerical integration is applicable, and the elimination of the volume unknowns can be done very cheaply. When constructing the basis, one has to solve a one dimensional eigenvalue problem for each pair of edge length and polynomial order. The cheap assembly, the efficient elimination of the volume unknowns and the possibility to solve this eigenvalue problem up to polynomial orders larger than thousand allows the usage of very high order elements. By introducing hanging nodes, one can additionally benefit from the exponential convergence of *hp* methods.

Almost independent of the discretization method for the wave equation, finding a solution for the linear system of equations is challenging, not just because of the big problem dimension. The difficulty with the wave equation is that it leads to an indefinite and in the presence of lossy media or absorbing boundary conditions to a complex system of equations. Consequently many preconditioners which are very effective for elliptic equations, like multigrid, turn out to cause large iteration counts, which even grow when the frequency is increased. Although some advances have been made [EVO04], good preconditioners are not available at the moment.

One big advantage of our mixed hybrid variational formulation is that the resulting linear system of equations for the facet degrees of freedom can be solved efficiently with a Krylov space method preconditioned by domain decomposition preconditioners. These preconditioners [TW05] are based on a division of the computational domain into considerably smaller subdomains. When applying them, the problem is reduced by solving it locally on the subdomain to a much smaller interface problem with better iterative properties. The success of domain decomposition preconditioners for the wave equation depends strongly on the choice of interface conditions. Taking for example nodal values, as it is done for elliptic problems, would cause large iteration counts. The mixed hybrid formulation provides in a natural way, due to the two different types of facet unknowns, interface conditions for the impedance traces, which leads to convergent schemes. In our work, we are going to introduce and discuss several different domain decomposition preconditioners.

The thesis is organized as follows:

In Chapter 2 we introduce Maxwell's equations and discuss possible boundary conditions. We will show that under some simplifying assumptions this set of equations leads to the vector valued wave equation or the Helmholtz equation. Additionally, the problem settings considered throughout the thesis are posed.

Chapter 3 recalls the basic results on Sobolev spaces and variational formulations. In this chapter we present standard variational formulations for the Helmholtz equation and the vector valued wave equation together with existence and uniqueness results.

A short overview on the concept of the finite element method is given in Chapter 4 and conforming finite elements for the spaces $L^2(\Omega)$, $H^1(\Omega)$, $H(\mathrm{div}, \Omega)$, $H(\mathrm{curl}, \Omega)$ are introduced.

Chapter 5 deals with a mixed hybrid discontinuous Galerkin method for the Helmholtz equation which is extended to the vector valued wave equation. After providing the finite element spaces needed for the discrete formulation, we state some conservation properties.

Different solution strategies for the resulting system of equations are explained in Chapter 6. We consider multiplicative and additive Schwarz preconditioners with local smoothers as well as a BDDC preconditioner and a new Robin type domain decomposition preconditioner with exact subdomain solvers. Numerical examples demonstrate good convergence properties of these solvers.

In Chapter 7 we focus on Hardy space infinite elements. We briefly overview basic properties of this realization of transparent boundary conditions and adjust them to the mixed hybrid formulation.

Chapter 8 is devoted to a tensor product implementation of the mixed formulation for the Helmholtz equation which uses a high order eigenfunction basis. We discuss the underlying eigenvalue problem, and we illustrate the consequences onto the sparsity pattern of the system matrix.

The simulation of diffraction gratings is the main topic of Chapter 9. There, a plane wave expansion for the far field is used as a transparent boundary condition in order to restrict the computational domain to the near field region. This type of boundary condition is well suited for periodic structures like gratings.

Chapter 2

From Maxwell's equations to Helmholtz equation

A wide range of examples for the wave equation comes from problems in electromagnetics and optics. Therefore, it is worth to take a closer look onto the basic equations, Maxwell's equations, describing such problems. In the first section of this chapter we introduce Maxwell's equations together with an appropriate set of interface conditions and boundary conditions. The vector valued wave equation and, under additional assumptions, the Helmholtz equation are derived from the time harmonic Maxwell's equation in the two subsequent sections. These two equations are the governing equations for this thesis.

2.1 Maxwell's equations

Maxwell's equations consist of two pairs of differential equations involving six space and time dependent fields, where two of them represent known source fields. We should already note that we are going to introduce Maxwell's equations in SI units with the basic units meter (m), kilogram(kg), second(s), Ampère(A) and Volt (V). The involved fields are

the electric field intensity $\boldsymbol{E}$ with the unit V/m,
the electric displacement $\boldsymbol{D}$ with the unit As/m^2,
the magnetic field intensity $\boldsymbol{H}$ with the unit A/m,
the magnetic flux density $\boldsymbol{B}$ with the unit Vs/m^2,

and the source fields are

the charge density ρ with the unit As/m^3 and
the current density $\boldsymbol{j}$ with the unit A/m^2.

For a detailed discussion on electrodynamics we recommend the books [Jac99, LL84], and for readers interested in optics a nice overview on the "Principles of Optics" is given in [BW99]. Throughout the work vectors are denoted by bold letters, i is the complex unit and complex conjugation of a quantity a is written as $\overline{a}$.

2.1.1 Maxwell's equations in media

The first equation we consider is *Gauss law for the electric field.* A volume containing charges gives rise to an electric field, or, in other words, the charges are the source of the electric field. Translating this into an equation, Gauss law says that the integral of the electric displacement $\boldsymbol{D}$ over a closed surface equals the enclosed charge,

$$\int_{\partial V} \boldsymbol{D} \cdot \boldsymbol{n}_{\partial V} ds = \int_V \rho d\boldsymbol{x}, \tag{2.1}$$

where V is a volume with the outer normal vector $\boldsymbol{n}_{\partial V}$ and ρ is the charge density.

For the magnetic field an equivalent equation can be derived, *Gauss law for the magnetic field.* Since magnetic monopoles do not exist, the corresponding quantity in magnetics, the magnetic induction $\boldsymbol{B}$, has no sources and

$$\int_{\partial V} \boldsymbol{B} \cdot \boldsymbol{n}_{\partial V} ds = 0 \tag{2.2}$$

is valid.

Faraday's law states that a varying magnetic field in time, or more precisely a changing magnetic flux, results in an electric field $\boldsymbol{E}$. The magnetic flux through an area A is defined as the integral of the normal component of the magnetic induction over the area. Thus, Faraday's law reads as

$$\int_{\partial A} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A} dl = -\int_A \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}_A ds, \tag{2.3}$$

where $\boldsymbol{\tau}_{\partial A}$ denotes the tangential vector on the boundary of A.

According to the last Maxwell equation, a magnetic field $\boldsymbol{H}$ is induced by a current through an area A. This current can be either a conduction current with current density $\boldsymbol{j}_C$ or a displacement current caused by a varying electric flux in time. Like in the magnetic case, the electric flux is the integral of the normal component of the electric displacement over A. Summarizing this, we get *Ampère's law*

$$\int_{\partial A} \boldsymbol{H} \cdot \boldsymbol{\tau}_{\partial A} dl = \int_A \left(\boldsymbol{j}_C + \frac{\partial \boldsymbol{D}}{\partial t} \right) \cdot \boldsymbol{n}_A ds. \tag{2.4}$$

Using Gauss' and Stoke's theorem, Maxwell's equations can be derived from (2.1)-(2.4) in differential form

$$\begin{aligned} \operatorname{div} \boldsymbol{D} &= \rho, && (2.5) \\ \operatorname{div} \boldsymbol{B} &= 0, && (2.6) \\ \operatorname{curl} \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, && (2.7) \\ \operatorname{curl} \boldsymbol{H} &= \boldsymbol{j}_C + \frac{\partial \boldsymbol{D}}{\partial t}. && (2.8) \end{aligned}$$

This set of equations for the basic quantities $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{D}$ and $\boldsymbol{B}$ is still under determined. In order to be able to get a unique solution for a given charge and current distribution, we have to consider the influence of materials onto the fields, i.e. we have to complete the system by material relations. They are usually considered to have the form

$$\begin{aligned} \boldsymbol{j}_C &= \boldsymbol{j}_I + \sigma \boldsymbol{E} && (2.9) \\ \boldsymbol{D} &= \epsilon \boldsymbol{E} && (2.10) \\ \boldsymbol{B} &= \mu \boldsymbol{H}, && (2.11) \end{aligned}$$

with material parameters σ as the *conductivity* with the unit A/Vm, μ is called *magnetic permeability* measured in Am/Vs and ϵ is the *electric permittivity* or the *dielectric constant*, respectively, with Vm/As as unit. The current density $\boldsymbol{j}_I$ is called *impressed current.*

2.1.2 Interface and boundary conditions

Till now Maxwell's equations are stated under the assumption that the material parameters are continuous. But in many applications one has to deal with abruptly changing material parameters. Therefore, interface conditions are needed.

Interface conditions

First, we are going to find an interface condition for the electric displacement. Therefore, the following setting is considered. We assume an interface area T where the material parameters can be discontinuous. We consider a cylinder V of a small height δh cut by T, such that the cylinder axis is perpendicular to T. The circular areas on the top and on the bottom of the cylinder which are parallel to T are denoted by δA_1 and by δA_2, respectively. The remaining surface is denoted by F. Considering (2.1) leads for the cylinder V to

$$\int_V \rho\, d\boldsymbol{x} = \int_{\partial A_1} \boldsymbol{D}_1 \cdot \boldsymbol{n}_{\delta A_1} ds + \int_{\partial A_2} \boldsymbol{D}_2 \cdot \boldsymbol{n}_{\delta A_2} ds + \int_{\partial F} \boldsymbol{D}_F \cdot \boldsymbol{n}_F ds,$$

where $\boldsymbol{D}_1, \boldsymbol{D}_2$ and $\boldsymbol{D}_F$ are the electric displacements on $\delta A_1, \delta A_2$ and F, respectively. At the limit $\delta h \to 0$ the surface integral over F vanishes, and assuming that there are charges on T with surface density $\hat{\rho}$,

$$\lim_{\delta h \to 0} \int_V \rho\, d\boldsymbol{x} = \int_{\partial A} \hat{\rho}\, ds$$

is obtained. This leads to

$$\int_{\partial A} \hat{\rho}\, ds = \int_{\partial A} \boldsymbol{D}_1 \cdot \boldsymbol{n}_{\delta A_1} ds + \int_{\partial A} \boldsymbol{D}_2 \cdot \boldsymbol{n}_{\delta A_2} ds.$$

Taking $\boldsymbol{n}_{\delta A_1} = -\boldsymbol{n}_{\delta A_2} =: \boldsymbol{n}_T$ into account, we can conclude

$$\boldsymbol{D}_1 \cdot \boldsymbol{n}_T - \boldsymbol{D}_2 \cdot \boldsymbol{n}_T = \hat{\rho},$$

from the fact that the position and radius were arbitrary. Consequently, in the absence of surface charges, $\boldsymbol{D}$ is normal continuous. With the same technique one can derive normal continuity of $\boldsymbol{B}$ from (2.2), i.e.

$$\boldsymbol{B}_1 \cdot \boldsymbol{n}_T - \boldsymbol{B}_2 \cdot \boldsymbol{n}_T = 0.$$

In order to get an interface condition for the electric field $\boldsymbol{E}$, we assume an arbitrary area $A = A_1 \cup A_2$ where A_1 and A_2 share the interface Γ. $\boldsymbol{E}$ has the values $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ on A_1 and A_2, respectively. When considering (2.3) for this area, we get

$$\begin{aligned} 0 &= \int_{\partial A} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A} dl + \int_{A_1} \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}_{A_1} ds + \int_{A_2} \frac{\partial \boldsymbol{B}}{\partial t} \cdot \boldsymbol{n}_{A_2} ds \\ &= \int_{\partial A} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A} dl - \int_{\partial A_1} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A_1} dl - \int_{\partial A_2} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A_2} dl \\ &= -\int_{\partial A_1 \cap \Gamma} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A_1} dl - \int_{\partial A_2 \cap \Gamma} \boldsymbol{E} \cdot \boldsymbol{\tau}_{\partial A_2} \\ &= -\int_{\Gamma} \big(\boldsymbol{E}_1 - \boldsymbol{E}_2\big) \cdot \boldsymbol{\tau}_{\Gamma} dl. \end{aligned}$$

Here $\boldsymbol{\tau}_{\partial A_1} = -\boldsymbol{\tau}_{\partial A_2} =: \boldsymbol{\tau}_{\Gamma}$ was used on the interface. Since the area was chosen arbitrarily, we can assume tangential continuity of $\boldsymbol{E}$, i.e. $\boldsymbol{E} \times \boldsymbol{n}$ is continuous on an interface with normal $\boldsymbol{n}$.

With similar arguments one can show tangential continuity of the magnetic field

$$\boldsymbol{H}_1 \times \boldsymbol{n} - \boldsymbol{H}_2 \times \boldsymbol{n} = \boldsymbol{j}_S,$$

where $\boldsymbol{j}_S$ is a surface current.

Note that at interfaces with jumps in the material parameters ϵ or μ the quantities $\boldsymbol{E} \cdot \boldsymbol{n}, \boldsymbol{H} \cdot \boldsymbol{n}, \boldsymbol{D} \times \boldsymbol{n}$ and $\boldsymbol{B} \times \boldsymbol{n}$ are not continuous.

Boundary conditions

In order to reduce an infinite computational domain to a finite one, boundary conditions are needed. We give a short list of the most frequently used boundary conditions.

- A perfectly electric conductor
  On a perfectly electric conducting surface the tangential component of the electric field is zero, i.e. $\boldsymbol{E} \times \boldsymbol{n} = \boldsymbol{0}$.

- A perfectly magnetic conductor
  On a perfectly magnetic conducting surface the tangential component of the magnetic field is zero, i.e. $\boldsymbol{H} \times \boldsymbol{n} = \boldsymbol{0}$.

- Surface charges
  A surface charge density $\hat{\rho}$ is prescribed via $\boldsymbol{D} \cdot \boldsymbol{n} = \hat{\rho}$.

- Surface currents
  A surface current $\boldsymbol{j}_S$ is prescribed by $\boldsymbol{H} \times \boldsymbol{n} = \boldsymbol{j}_S$.

- Impedance boundary condition
  The impedance boundary condition for some $\kappa \in \mathbb{C}$ reads as

  $$-\boldsymbol{n} \times \boldsymbol{H} + \kappa \boldsymbol{E}_{\|} = \boldsymbol{g},$$

  where $\boldsymbol{E}_{\|} := \boldsymbol{n} \times (\boldsymbol{E} \times \boldsymbol{n})$ is the tangential component of $\boldsymbol{E}$ on the boundary, while $\boldsymbol{n} \times \boldsymbol{H}$ represents the rotated tangential component of $\boldsymbol{H}$. We will use such a Robin type boundary condition later on as absorbing boundary condition to prescribe transparent boundaries.

2.1.3 A time harmonic setting

Now we assume that the excitations, to be more precise the quantities $\boldsymbol{j}_I$ and ρ, vary periodically in time with the same *angular frequency* ω. Thus,

$$\boldsymbol{j}_I(\boldsymbol{x}, t) = \mathrm{Re}\left(\widetilde{\boldsymbol{j}}_I(\boldsymbol{x}) e^{-i\omega t}\right) \qquad \rho(\boldsymbol{x}, t) = \mathrm{Re}\left(\widetilde{\rho}_S(\boldsymbol{x}) e^{-i\omega t}\right),$$

where $\widetilde{\boldsymbol{j}}_I$ and $\widetilde{\rho}$ are just space dependent. Under this assumption the unknown vector fields will show the same time dependence, and we use the ansatz

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{x}, t) &= \mathrm{Re}\left(\widetilde{\boldsymbol{E}}(\boldsymbol{x}) e^{-i\omega t}\right) & \qquad \boldsymbol{H}(\boldsymbol{x}, t) &= \mathrm{Re}\left(\widetilde{\boldsymbol{H}}(\boldsymbol{x}) e^{-i\omega t}\right) \\ \boldsymbol{D}(\boldsymbol{x}, t) &= \mathrm{Re}\left(\widetilde{\boldsymbol{D}}(\boldsymbol{x}) e^{-i\omega t}\right) & \qquad \boldsymbol{B}(\boldsymbol{x}, t) &= \mathrm{Re}\left(\widetilde{\boldsymbol{B}}(\boldsymbol{x}) e^{-i\omega t}\right). \end{aligned}$$

Inserting this into Maxwell's equations (2.5)-(2.8), using the material laws (2.9)-(2.11) and carrying out the time derivatives leads to the time harmonic Maxwell's equations. For notational reasons we neglect the tilde, so we write $\boldsymbol{E}$ instead of $\widetilde{\boldsymbol{E}}$ and so on. The time harmonic Maxwell's equations are

$$\begin{aligned} \mathrm{div}(\epsilon \boldsymbol{E}) &= \rho, & (2.12) \\ \mathrm{div}(\mu \boldsymbol{H}) &= 0, & (2.13) \\ \mathrm{curl}\, \boldsymbol{E} - i\omega\mu \boldsymbol{H} &= 0, & (2.14) \\ \mathrm{curl}\, \boldsymbol{H} + i\omega\epsilon \boldsymbol{E} - \sigma \boldsymbol{E} &= \boldsymbol{j}_I. & (2.15) \end{aligned}$$

The equations (2.14) and (2.15) lead to the Helmholtz equation which will be discussed in detail later.

2.1.4 Transversal electric and transversal magnetic modes

We assume that no charges are present and that the system is invariant in z-direction. Thus, all z-derivatives vanish and the computational domain can be reduced to a two dimensional one. This leads for the equations (2.14) and (2.15) to

$$\begin{aligned} \partial_y E_z - i\omega\mu H_x &= 0, && (2.16) \\ -\partial_x E_z - i\omega\mu H_y &= 0, && (2.17) \\ \partial_x E_y - \partial_y E_x - i\omega\mu H_z &= 0, && (2.18) \end{aligned}$$

$$\begin{aligned} \partial_y H_z + i\omega\epsilon E_x - \sigma E_x &= j_{Ix}, && (2.19) \\ -\partial_x H_z + i\omega\epsilon E_y - \sigma E_y &= j_{Iy}, && (2.20) \\ \partial_x H_y - \partial_y H_x + i\omega\epsilon E_z - \sigma E_z &= j_{Iz}. && (2.21) \end{aligned}$$

Here ∂_x denotes the x derivative, E_x is the x component of $\boldsymbol{E}$, etc. Taking into account that the two divergence equations just contain E_x, E_y and H_x, H_y, respectively, the total system of Maxwell's equations decouples into two systems of equations.

Transversal electric modes

One system consists (besides (2.13)) of the equations (2.16), (2.17) and (2.21) and just contains the unknowns E_z, H_x and H_y. The fact that the total electric field is perpendicular to the computational domain gives rise to the name *transversal electric* (TE) modes (also known as σ polarization). If we define $\widehat{\boldsymbol{H}} = (-H_y, H_x)^\top$, or in other words rotate the magnetic field, we get from (2.16), (2.17) and (2.21)

$$\begin{aligned} \operatorname{grad} E_z + i\omega\mu\widehat{\boldsymbol{H}} &= 0, && (2.22) \\ \operatorname{div}\widehat{\boldsymbol{H}} + i\omega\epsilon E_z + \sigma E_z &= -j_z. && (2.23) \end{aligned}$$

Continuity of $\boldsymbol{n} \times \boldsymbol{E}$ and $\boldsymbol{n} \times \boldsymbol{H}$ on an interface between two different media is now equivalent to the continuity of E_z and $\boldsymbol{n} \cdot \widehat{\boldsymbol{H}}$ (where $\boldsymbol{n} = (n_x, n_y)^\top$), and the impedance boundary condition reads as

$$-\boldsymbol{n} \cdot \widehat{\boldsymbol{H}} + \kappa E_z = g$$

for some g. Assuming $\boldsymbol{j}_I = \boldsymbol{0}$ and $\sigma = 0$ results in the two dimensional mixed Helmholtz equation discussed below.

Transversal magnetic modes

The remaining equations contain the components E_x, E_y and H_z. Therefore, they are called *transverse magnetic* (TM) (or π-polarized). Defining a rotated electric field $\widehat{\boldsymbol{E}} = (-E_y, E_x)^\top$, equations (2.18)- (2.20) lead to

$$\begin{aligned} \operatorname{div} \widehat{\boldsymbol{E}} + i\omega\mu H_z &= 0, && (2.24) \\ \operatorname{grad} H_z + i\omega\epsilon\widehat{\boldsymbol{E}} - \sigma\widehat{\boldsymbol{E}} &= (-j_y, j_x)^\top. && (2.25) \end{aligned}$$

Again the tangential continuity of the electric and the magnetic field is equivalent to the continuity of H_z and $\boldsymbol{n} \cdot \widehat{\boldsymbol{E}}$, and the impedance boundary condition can be written as

$$-H_z + \kappa \boldsymbol{n} \cdot \widehat{\boldsymbol{E}} = g$$

for some g. In the absence of currents $\boldsymbol{j}$ and with $\sigma = 0$ the two dimensional Helmholtz equation is obtained.

2.1.5 Electromagnetic energy

For an electric field $\boldsymbol{E}(\boldsymbol{x}, t)$ and a magnetic field $\boldsymbol{H}(\boldsymbol{x}, t)$ with the corresponding magnetic induction $\boldsymbol{B}(\boldsymbol{x}, t)$ the Poynting vector is defined as

$$\boldsymbol{S}(\boldsymbol{x}, t) = \frac{1}{\mu}\boldsymbol{E}(\boldsymbol{x}, t) \times \boldsymbol{B}(\boldsymbol{x}, t) = \boldsymbol{E}(\boldsymbol{x}, t) \times \boldsymbol{H}(\boldsymbol{x}, t).$$

The Poynting vector can be seen as an energy flux per time and area. In the time harmonic case one can average this energy flux over time which gives a time averaged Poynting vector

$$\langle \boldsymbol{S} \rangle(\boldsymbol{x}) = \frac{1}{2}\operatorname{Re}\big(\boldsymbol{E}(\boldsymbol{x}) \times \overline{\boldsymbol{H}}(\boldsymbol{x})\big),$$

where $\boldsymbol{E}$ and $\boldsymbol{H}$ represent for notational reasons just the spatial parts of the electric and magnetic field.

If we consider an electromagnetic wave with amplitude $\boldsymbol{E}_0$ and wave vector $\boldsymbol{k}$ (note

$\boldsymbol{E}_0 \perp \boldsymbol{k}$),

$$\boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{E}_0 e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \qquad \text{and} \qquad \boldsymbol{H}(\boldsymbol{x}) = \frac{1}{\omega\mu}\boldsymbol{k} \times \boldsymbol{E}_0 e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$$

the time averaged Poynting vector is

$$\langle \boldsymbol{S} \rangle = \frac{1}{2}\|\boldsymbol{E}_0\|^2 \mathrm{Re}\left(\frac{1}{\omega\mu}\overline{\boldsymbol{k}}\right). \tag{2.26}$$

Thus, the energy density for a plane wave is space independent, and energy is transported into the direction of the wave vector. The energy $\mathcal{E}$ transported through a surface A per time can be calculated simply as

$$\mathcal{E} = \int_A \langle \boldsymbol{S} \rangle \cdot \boldsymbol{n}_A \, ds. \tag{2.27}$$

2.2 The Helmholtz equation

The Helmholtz equation in primal and mixed form is introduced in this section . Although, we motivate the Helmholtz equation via TE and TM problems, which are two dimensional, we will define it for arbitrary dimensions in order to cover also applications for example from mechanics. Thus, let us consider a domain $\Omega \subset \mathbb{R}^d$ with dimension $d = 2, 3$ whose boundary is denoted by $\Gamma = \partial\Omega$.

2.2.1 The primal form of the Helmholtz equation

When solving the Helmholtz equation, one solves for a scalar function $u : \Omega \to \mathbb{C}$ which fulfills

$$\Delta u + \nu^2\omega^2 u = 0 \qquad \text{in } \Omega. \tag{2.28}$$

Additionally, the angular frequency $\omega \in \mathbb{R}$ is supposed to be constant, and ν is positive, real and space dependent. If ν is constant, the fundamental solution of the Helmholtz equation is a plane wave $u = u_0 \exp(i\boldsymbol{k} \cdot \boldsymbol{x})$ with amplitude u_0. The wave vector $\boldsymbol{k}$ describing the direction of propagation has an absolute value $|\boldsymbol{k}| = \omega\nu$, and consequently ν can be interpreted as the inverse of the speed of the wave. In optics, the quantity ν/ν_0, where ν_0 represents the value of ν in vacuum, has the meaning of the refractive index.

For TE problems the Helmholtz equation can be obtained from the governing equations of TE modes (2.22) and (2.23) by eliminating $\widehat{\boldsymbol{H}}$ and assuming that $\sigma = 0$ and $j_z = 0$. The quantity ν^2 is then the product of ϵ and μ where μ was assumed to be constant, and

u represents E_z. Assuming in the TM equations that ϵ is constant, $\sigma = 0$, $j_x = j_y = 0$ leads as well to a Helmholtz equation with u representing H_z and $\nu^2 = \epsilon\mu$. We should mention that for acoustic problems u stands for the pressure, or in geophysics u denotes the displacement.

In order to pose appropriate boundary conditions, we introduce incoming and outgoing impedance traces at the domain boundary Γ as

$$\left(-\frac{1}{i\nu\omega}\frac{\partial u}{\partial \boldsymbol{n}_\Gamma} + u\right)\Big|_\Gamma \qquad \text{and} \qquad \left(\frac{1}{i\nu\omega}\frac{\partial u}{\partial \boldsymbol{n}_\Gamma} + u\right)\Big|_\Gamma$$

where $\frac{\partial}{\partial \boldsymbol{n}_\Gamma}$ denotes the derivative in outer normal direction $\boldsymbol{n}_\Gamma$. Evaluating the incoming impedance trace for an incoming wave of the domain Ω with amplitude A, $u_{in}(\boldsymbol{x}) = Ae^{i\nu\omega(\boldsymbol{n}_\Gamma\cdot\boldsymbol{x})}$, we obtain $\mathrm{In}_\Gamma(u_{in}) = 2Ae^{i\nu\omega(\boldsymbol{n}_\Gamma\cdot\boldsymbol{x})}|_\Gamma$, while an evaluation of the outgoing impedance trace yields zero. For outgoing waves the situation is vice versa. Thus by fixing the incoming impedance trace, which results in an absorbing boundary condition

$$-\frac{1}{i\omega\nu}\frac{\partial u}{\partial \boldsymbol{n}} + u = g \qquad \text{on } \Gamma, \tag{2.29}$$

the energy entering the computational domain is prescribed.

2.2.2 The mixed form of the Helmholtz equation

The mixed form of the Helmholtz equation,

$$i\omega\epsilon u - \operatorname{div}\boldsymbol{\sigma} = 0 \qquad \text{in } \Omega, \tag{2.30}$$

$$\operatorname{grad} u - i\omega\mu\boldsymbol{\sigma} = 0 \qquad \text{in } \Omega, \tag{2.31}$$

is obtained from (2.28) by introducing an additional flux field $\boldsymbol{\sigma} : \Omega \to \mathbb{C}^d$ which fulfills $\boldsymbol{\sigma} = \frac{1}{i\omega\mu}\operatorname{grad} u$. The parameter $\mu \in \mathbb{R}$ is a positive constant, and ϵ is chosen, such that $\nu^2 = \epsilon\mu$. Note, by calling these two material parameters ϵ and μ, we have the TE problem in mind, and $\boldsymbol{\sigma}$ corresponds to $\widehat{\boldsymbol{H}}$ from (2.22). In the TM case $\boldsymbol{\sigma}$ corresponds to $\widehat{\boldsymbol{E}}$ from (2.24), and μ and ϵ interchange roles.

For this system, incoming and outgoing impedance traces can be rewritten as

$$\Big(-\sqrt{\frac{\mu}{\epsilon}}\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma + u\Big)\Big|_\Gamma \qquad \text{and} \qquad \Big(\sqrt{\frac{\mu}{\epsilon}}\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma + u\Big)\Big|_\Gamma.$$

Thus, fixing the incoming impedance trace leads to the boundary condition

$$-\sqrt{\frac{\mu}{\epsilon}}\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma + u = g \qquad \text{on } \Gamma. \tag{2.32}$$

2.3 The vector valued wave equation

In order to get a wave equation for electromagnetic waves on a computational domain $\Omega \subset \mathbb{R}^3$ with boundary Γ, we consider for the time harmonic Maxwell's equations (2.12)-(2.15) a non conducting material without free charges and impressed currents, i.e. $\sigma = 0$, $\rho = 0$ and $\boldsymbol{j}_I = \boldsymbol{0}$. In the following we assume, that $\mu \in \mathbb{R}$ is constant, and ϵ is a real and positive function.

2.3.1 The primal form of the vector valued wave equation

Eliminating the magnetic field $\boldsymbol{H}$ in the equations (2.14) and (2.15) results in the vector valued time harmonic wave equation

$$\operatorname{curl}(\operatorname{curl}\boldsymbol{E}) - \omega^2\epsilon\mu\boldsymbol{E} = 0. \tag{2.33}$$

The fundamental solution of this equation is for constant material parameters a plane wave $\boldsymbol{E} = \boldsymbol{E}_0\exp(i\boldsymbol{k}\cdot\boldsymbol{x})$ with amplitude $\boldsymbol{E}_0$ and a wave vector of absolute value $|\boldsymbol{k}| = \omega\sqrt{\epsilon\mu}$. Note that according to the divergence freeness of the electric field $\boldsymbol{E}_0$ and $\boldsymbol{k}$ are perpendicular. The quantity $1/\sqrt{\epsilon\mu}$ can be interpreted as the speed of the wave, i.e. the speed of light in a medium.

In the same way as for the Helmholtz equation, we define incoming and outgoing impedance traces on Γ

$$\Big(\frac{1}{i\omega\sqrt{\mu\epsilon}}\boldsymbol{n}_\Gamma\times\operatorname{curl}\boldsymbol{E} + \boldsymbol{E}_{\|}\Big)\Big|_\Gamma \qquad \text{and} \qquad \Big(-\frac{1}{i\omega\sqrt{\mu\epsilon}}\boldsymbol{n}_\Gamma\times\operatorname{curl}\boldsymbol{E} + \boldsymbol{E}_{\|}\Big)\Big|_\Gamma$$

where, as already mentioned, $\boldsymbol{E}_{\|} := \boldsymbol{n}_\Gamma \times \big(\boldsymbol{E}\times\boldsymbol{n}_\Gamma\big)$ is the tangential component of $\boldsymbol{E}$ on Γ. Prescribing an incoming plane wave is equivalent to fixing the incoming impedance trace as boundary condition

$$\frac{1}{i\omega\sqrt{\mu\epsilon}}\boldsymbol{n}_\Gamma\times\operatorname{curl}\boldsymbol{E} + \boldsymbol{E}_{\|} = \boldsymbol{g} \qquad \text{on } \Gamma. \tag{2.34}$$

The quantity $\boldsymbol{g}$ corresponds to two times the trace of the incoming wave at the boundary.

2.3.2 The mixed form of the vector valued wave equation

The time harmonic Maxwell's equations (2.14) and (2.15) represent the mixed version of the vector valued time harmonic wave equation,

$$i\omega\epsilon\boldsymbol{E} + \operatorname{curl}\boldsymbol{H} = 0 \qquad \text{in } \Omega, \tag{2.35}$$

$$\operatorname{curl}\boldsymbol{E} - i\omega\mu\boldsymbol{H} = 0 \qquad \text{in } \Omega, \qquad . \tag{2.36}$$

Note that the magnetic field can be interpreted as the flux field corresponding to the electric field. By using (2.36), the incoming and outgoing impedance traces read as

$$\Big(\sqrt{\frac{\mu}{\epsilon}}\,\boldsymbol{n}_{\partial V}\times\boldsymbol{H}+\boldsymbol{E}_{\|}\Big)\Big|_{\Gamma} \qquad \text{and} \qquad \Big(-\sqrt{\frac{\mu}{\epsilon}}\,\boldsymbol{n}_{\partial V}\times\boldsymbol{H}+\boldsymbol{E}_{\|}\Big)\Big|_{\Gamma},$$

and consequently an incoming wave is fixed by the absorbing boundary condition

$$\sqrt{\frac{\mu}{\epsilon}}\boldsymbol{n}_{\Gamma}\times\boldsymbol{H}+\boldsymbol{E}_{\|}=\boldsymbol{g} \qquad \text{on } \Gamma. \tag{2.37}$$

Chapter 3

The variational framework

In this chapter we want to introduce standard variational formulations for both, the primal and the mixed Helmholtz equation, as well as the primal and the mixed form of the vector valued wave equation. Therefore, the Sobolev spaces $L^2(\Omega)$, $H^1(\Omega)$, $H(\mathrm{div}, \Omega)$ and $H(\mathrm{curl}, \Omega)$ are needed. The first section is devoted to these spaces, and a small list of important properties is provided. Based on this basic existence and uniqueness results for variational formulations are given in the second section, and the existence of a unique solution is proven for standard weak forms of the Helmholtz equation and the vector valued wave equation.

3.1 Some function spaces

We are going to introduce the Sobolev spaces $L^2(\Omega)$, $H^1(\Omega)$, $H(\mathrm{div}, \Omega)$ and $H(\mathrm{curl}, \Omega)$ for a Lipschitz domain Ω with boundary Γ which is defined according to

Definition 3.1. *A domain $\Omega \subset \mathbb{R}^d$ is said to have a Lipschitz boundary $\Gamma := \partial\Omega$ if a finite number of domains Ω_i, a local coordinate system $(\xi_{i,1}, \dots, \xi_{i,d})$ and a Lipschitz continuous function $f(\xi_{i,1}, \dots \xi_{i,d-1})$ exist, such that*

(1) $\Gamma \subset \bigcup_i \Omega_i$ *and* $\Gamma \cap \Omega_i = \big\{(\xi_{i,1}, \dots, \xi_{i,d}) \;:\; \xi_{i,d} = f(\xi_{i,1}, \dots, \xi_{i,d-1})\big\}$,

(2) $\Omega \cap \Omega_i = \big\{(\xi_{i,1}, \dots, \xi_{i,d}) \;:\; \xi_{i,d} > f(\xi_{i,1}, \dots, \xi_{i,d-1})\big\}$.

If Ω has a Lipschitz boundary, we call it a Lipschitz domain.

For such a Lipschitz domain Ω the set of all continuous functions is denoted by $C(\Omega)$ and the set of all infinitely continuous differentiable functions by $C^\infty(\Omega)$. The subsets of compact support on Ω we call $C_0(\Omega)$ and $C_0^\infty(\Omega)$, respectively.

3.1.1 The space $L^2(\Omega)$

The Sobolev space of square integrable functions on a Lipschitz domain Ω is defined as

$$L^2(\Omega) := \Big\{ u : \Omega \to \mathbb{R} \mid \int_\Omega |u|^2 \, d\boldsymbol{x} < \infty \Big\}.$$

With the help of the scalar product

$$(u, v)_{L^2(\Omega)} := \int_\Omega uv \; d\boldsymbol{x},$$

a norm is induced which we denote by $\| \cdot \|_{L^2(\Omega)}$. By equipping $L^2(\Omega)$ with this norm, $L^2(\Omega)$ becomes a Hilbert space. In an equivalent way we can define the Hilbert space of all square integrable functions on the boundary $\Gamma = \partial\Omega$, $L^2(\Gamma)$. The definition of the L^2-inner product and the L^2-norm for vector valued functions $\boldsymbol{p}, \boldsymbol{q} \in \big(L^2(\Omega)\big)^d$ are straight forward, i.e.

$$(\boldsymbol{p}, \boldsymbol{q})_{L^2(\Omega)} := \sum_{i=1}^{d} (p_i, q_i)_{L^2(\Omega)} \qquad \text{and} \qquad \|\boldsymbol{p}\|^2_{L^2(\Omega)} := (\boldsymbol{p}, \boldsymbol{p})_{L^2(\Omega)}.$$

3.1.2 The space $H^1(\Omega)$

In order to define the space $H^1(\Omega)$, or more generally $H^k(\Omega)$ for $k \in \mathbb{N}_0$, we have to generalize a partial derivative which can be applied till now just to sufficiently smooth functions, i.e. we have to introduce weak derivatives. We will denote a partial derivative of degree n by ∂^α where α is a multi index with $|\alpha| = n$.

Definition 3.2 (weak partial derivative). *For $u \in L^2(\Omega)$ we call $g = \partial^\alpha u$, $g \in L^2(\Omega)$, the weak partial derivative if*

$$\int_\Omega g \, v \; d\boldsymbol{x} = (-1)^{|\alpha|} \int_\Omega u \, (\partial^\alpha v) \; d\boldsymbol{x} \qquad \forall v \in C_0^\infty(\Omega).$$

By combining the weak partial derivatives of order one, the weak gradient can be defined. In the following we will use the symbols ∂^α and grad when referring to the weak partial derivative and gradient, respectively. This leads to the definition of the space $H^k(\Omega)$.

Definition 3.3. *The space $H^k(\Omega)$ for $k \in \mathbb{N}_0$ is defined as*

$$H^k(\Omega) := \Big\{ u \in L^2(\Omega) \; : \; \partial^\alpha u \in L^2(\Omega) \quad \textit{for all } \alpha \textit{ with } |\alpha| \leq k \Big\}.$$

If the space is equipped with the scalar product

$$(u,v)_{H^k(\Omega)} := \sum_{|\alpha|\le k} (\partial^\alpha u, \partial^\alpha v)_{L^2(\Omega)}$$

and the corresponding norm

$$\|u\|_{H^k(\Omega)} := \sqrt{\sum_{|\alpha|\le k} \|\partial^\alpha u\|^2_{L^2(\Omega)}},$$

it is a Hilbert space. Furthermore, by $|u|^2_{H^k(\Omega)} = \sum_{|\alpha|=k} \|\partial^\alpha u\|^2_{L^2(\Omega)}$ *a seminorm is provided.*

Note that the space $H^0(\Omega)$ is identical to the space $L^2(\Omega)$. For Lipschitz domains Ω $H^k(\Omega)$ can be identified with the closure of $C^\infty(\bar{\Omega})$ with respect to the $H^k(\Omega)$-norm (compare [GR86]), i.e.

$$H^k(\Omega) = \overline{C^\infty(\bar{\Omega})}^{\|\cdot\|_{H^k(\Omega)}}.$$

Thus, $C^\infty(\Omega)$ and $C(\Omega)$ are dense in $H^k(\Omega)$. The dual space we denote as $H^{-k}(\Omega)$ and a norm is given via the duality relation

$$\|u\|_{H^{-k}(\Omega)} = \sup_{v\in H^k(\Omega)} \frac{(u,v)_{H^k(\Omega)}}{\|v\|_{H^k(\Omega)}}.$$

In order to deal with boundary conditions, Sobolev spaces of fractional order are needed on the boundary Γ. For more details on such spaces see [McL00]. Especially, we will make use of the space $H^{1/2}(\Gamma) := \overline{C^\infty(\Gamma)}^{\|\cdot\|_{H^{1/2}(\Gamma)}}$ with the norm

$$\|u\|^2_{H^{1/2}(\Gamma)} := \|u\|^2_{L^2(\Gamma)} + \int_\Gamma\int_\Gamma \frac{|u(x)-u(y)|}{|x-y|}\,dx\,dy.$$

In addition, the trace operator tr_Γ for continuous functions is given as

$$\mathrm{tr}_\Gamma(u) = u|_\Gamma \qquad \text{for } u \in C(\bar{\Omega})$$

which leads us to the trace and inverse trace theorem.

Theorem 3.4 (trace and inverse trace theorem for $H^1(\Omega)$)**.** *Let Ω be a bounded Lipschitz domain.*

(1) Then the operator tr_Γ *can be uniquely extended to a continuous operator mapping*

from $H^1(\Omega)$ *to* $H^{1/2}(\Gamma)$*, and*

$$\|\mathrm{tr}_\Gamma(u)\|^2_{H^{1/2}(\Gamma)} \leq c\|u\|^2_{H^1(\Omega)} \qquad \textit{for } u \in H^1(\Omega).$$

(2) For any $g \in H^{1/2}(\Gamma)$ *there exists a function* $u \in H^1(\Omega)$ *with* $\mathrm{tr}_\Gamma(u) = g$ *and*

$$\|u\|^2_{H^1(\Omega)} \leq c\|g\|^2_{H^{1/2}(\Gamma)}.$$

This theorem allows us to define boundary conditions. Consequently, the space $H^1_0(\Omega) := \big\{v \in H^1(\Omega) \;:\; \mathrm{tr}_\Gamma(v) = 0\big\}$, required to incorporate Dirichlet boundary conditions, is well defined. We will finish the discussion by introducing an essential set of interface conditions for $H^1(\Omega)$ functions which will play an important role in constructing conforming finite elements.

Corollary 3.5. *Let* $\Omega_1, \dots, \Omega_N$ *be Lipschitz domains which form a domain decomposition of* Ω*, i.e.* $\Omega_i \cap \Omega_j = \emptyset$ *and* $\bar{\Omega} = \bigcup_i \bar{\Omega}_i$*. By* $\Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j$ *the interfaces are denoted. For a function* u *with* $u|_{\Omega_i} = u_i$*,* $u_i \in H^1(\Omega_i)$ *and* $\mathrm{tr}_{\Gamma_{ij}}(u_i) = \mathrm{tr}_{\Gamma_{ij}}(u_j)$ *follows that* $u \in H^1(\Omega)$ *and* $(\mathrm{grad}\, u)|_{\Omega_i} = \mathrm{grad}\, u_i$*.*

3.1.3 The space $H(\mathrm{div}, \Omega)$

We start the discussion on the vector valued space $H(\mathrm{div}, \Omega)$ by generalizing the divergence for non continuous functions.

Definition 3.6 (weak divergence)**.** *For* $\boldsymbol{q} \in \big(L^2(\Omega)\big)^d$ *we call* $p = \mathrm{div}\, \boldsymbol{q}$*,* $p \in L^2(\Omega)$ *the weak divergence of* $\boldsymbol{q}$ *if*

$$\int_\Omega pv\, dx = -\int_\Omega \boldsymbol{q} \cdot (\mathrm{grad}\, v)\, dx \qquad \forall v \in C_0^\infty(\Omega).$$

In the following we refer to the weak divergence when using the symbol div. Based on this generalized divergence the space $H(\mathrm{div}, \Omega)$ can be introduced.

Definition 3.7. *The space* $H(\mathrm{div}, \Omega)$ *is defined as*

$$H(\mathrm{div}, \Omega) := \Big\{\boldsymbol{q} \in \big(L^2(\Omega)\big)^d \;:\; \mathrm{div}\, \boldsymbol{q} \in L^2(\Omega)\Big\}.$$

Combined with the inner product and the corresponding norm,

$$(\boldsymbol{p}, \boldsymbol{q})_{H(\operatorname{div},\Omega)} := (\boldsymbol{p}, \boldsymbol{q})_{L^2(\Omega)} + (\operatorname{div} \boldsymbol{p}, \operatorname{div} \boldsymbol{q})_{L^2(\Omega)} \qquad \textit{and} \qquad \|\boldsymbol{p}\|^2_{H(\operatorname{div},\Omega)} = (\boldsymbol{p}, \boldsymbol{p})_{H(\operatorname{div},\Omega)},$$

$H(\operatorname{div}, \Omega)$ *is a Hilbert space.*

As it is shown in [GR86, Mon03], the spaces $\big(C^\infty(\Omega)\big)^d$ and $\big(C(\Omega)\big)^d$ are dense in $H(\operatorname{div}, \Omega)$, and the space can be alternatively written as

$$H(\operatorname{div}, \Omega) = \overline{C^\infty\big(\bar{\Omega}\big)}^{\|\cdot\|_{H(\operatorname{div},\Omega)}}.$$

In order to deal with boundary conditions, a well defined normal trace on Γ is needed. Therefore, we introduce for continuous functions the normal trace operator

$$\operatorname{tr}_{\boldsymbol{n},\Gamma}(\boldsymbol{q}) = \boldsymbol{q} \cdot \boldsymbol{n}_\Gamma \qquad \text{for } \boldsymbol{q} \in \big(C(\bar{\Omega})\big)^d$$

where $\boldsymbol{n}_\Gamma$ represents the outer normal vector. By using density arguments this normal trace operator can be extended to $H(\operatorname{div}, \Omega)$.

Theorem 3.8 (trace and inverse trace theorem for $H(\operatorname{div}, \Omega)$)**.** *Let Ω be a bounded Lipschitz domain.*

(1) Then the operator $\operatorname{tr}_{\boldsymbol{n},\Gamma}$ can be uniquely extended to a continuous operator mapping from $H(\operatorname{div}, \Omega)$ to $H^{-1/2}(\Gamma)$, and

$$\|\operatorname{tr}_{\boldsymbol{n},\Gamma}(\boldsymbol{q})\|^2_{H^{-1/2}(\Gamma)} \leq c\|\boldsymbol{q}\|^2_{H(\operatorname{div},\Omega)} \qquad \textit{for } \boldsymbol{q} \in H(\operatorname{div}, \Omega).$$

(2) For any $q_n \in H^{-1/2}(\Gamma)$ there exists a function $\boldsymbol{q} \in H(\operatorname{div}, \Omega)$ with $\operatorname{tr}_{\boldsymbol{n},\Gamma}(\boldsymbol{q}) = q_n$ and

$$\|\boldsymbol{q}\|^2_{H(\operatorname{div},\Omega)} \leq c\|q_n\|^2_{H^{-1/2}(\Gamma)}.$$

If furthermore $(q_n, 1)_{L^2(\Gamma)} = 0$, then $\operatorname{div} \boldsymbol{q} = 0$ for the extension $\boldsymbol{q}$.

Here, the dual space to $H^{1/2}(\Gamma)$ is denoted as $H^{-1/2}(\Gamma)$ with a norm obtained via a duality identity from the $H^{1/2}$ norm. Based on the trace theorem, the space of functions in $H(\operatorname{div}, \Omega)$ with zero normal trace which is needed to incorporate Dirichlet boundary conditions can be introduced as

$$H_0(\operatorname{div}, \Omega) = \Big\{\boldsymbol{q} \in H(\operatorname{div}, \Omega) \,:\, \operatorname{tr}_{\boldsymbol{n},\Gamma}(\boldsymbol{q}) = 0\Big\}.$$

Additionally, density arguments allow us to generalize the integration by parts formula.

Lemma 3.9. *For a Lipschitz domain Ω and for all $\boldsymbol{q} \in H(\mathrm{div}, \Omega)$, $u \in H^1(\Omega)$ the integration by parts formula holds*

$$\int_\Omega \operatorname{grad} u \, \boldsymbol{q} \, dx = - \int_\Omega u \, \operatorname{div} \boldsymbol{q} \, dx + \int_\Gamma \mathrm{tr}_\Gamma(u) \, \mathrm{tr}_{\boldsymbol{n},\Gamma}(\boldsymbol{q}) \, ds.$$

We conclude again the subsection by presenting an essential set of interface conditions for $H(\mathrm{div}, \Omega)$ functions.

Corollary 3.10. *Let $\Omega_1, \ldots, \Omega_N$ be Lipschitz domains which form a domain decomposition of Ω, i.e. $\Omega_i \cap \Omega_j = \emptyset$ and $\bar{\Omega} = \bigcup_i \bar{\Omega}_i$. By $\Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j$ we denote the interfaces and by $\boldsymbol{n}_i$ their normal vectors. For a function $\boldsymbol{q}$ with $\boldsymbol{q}|_{\Omega_i} = \boldsymbol{q}_i$, $\boldsymbol{q}_i \in H(\mathrm{div}, \Omega_i)$ and $\mathrm{tr}_{\boldsymbol{n}_i,\Gamma_{ij}}(\boldsymbol{q}_i) = \mathrm{tr}_{\boldsymbol{n}_i,\Gamma_{ij}}(\boldsymbol{q}_j)$ follows that $\boldsymbol{q} \in H(\mathrm{div}, \Omega)$ and $(\operatorname{div} \boldsymbol{q})|_{\Omega_i} = \operatorname{div} \boldsymbol{q}_i$.*

3.1.4 The space $H(\mathrm{curl}, \Omega)$

Finally, the space $H(\mathrm{curl}, \Omega)$ for $\Omega \subset \mathbb{R}^3$ is introduced. Therefore, we need a generalized curl operator.

Definition 3.11 (weak curl)**.** *For $\boldsymbol{p} \in \left(L^2(\Omega)\right)^3$ we call $\boldsymbol{c} = \operatorname{curl} \boldsymbol{p}$, $\boldsymbol{c} \in \left(L^2(\Omega)\right)^3$ the weak curl of $\boldsymbol{p}$ if*

$$\int_\Omega \boldsymbol{c} \cdot \boldsymbol{v} \, dx = \int_\Omega \boldsymbol{p} \cdot (\operatorname{curl} \boldsymbol{v}) \, dx \qquad \forall \boldsymbol{v} \in \left(C_0^\infty(\Omega)\right)^3.$$

From now on, we refer to the weak curl when using the "curl" symbol. Based on this, the Sobolev space $H(\mathrm{curl}, \Omega)$ is defined.

Definition 3.12. *The space $H(\mathrm{curl}, \Omega)$ is defined as*

$$H(\mathrm{curl}, \Omega) := \left\{ \boldsymbol{p} \in \left(L^2(\Omega)\right)^3 \ : \ \operatorname{curl} \boldsymbol{p} \in \left(L^2(\Omega)\right)^3 \right\}.$$

Together with the inner product and the corresponding norm,

$$(\boldsymbol{p}, \boldsymbol{q})_{H(\mathrm{curl},\Omega)} := (\boldsymbol{p}, \boldsymbol{q})_{L^2(\Omega)} + (\operatorname{curl} \boldsymbol{p}, \operatorname{curl} \boldsymbol{q})_{L^2(\Omega)} \qquad \text{and} \qquad \|\boldsymbol{p}\|^2_{H(\mathrm{curl},\Omega)} = (\boldsymbol{p}, \boldsymbol{p})_{H(\mathrm{curl},\Omega)},$$

$H(\mathrm{curl}, \Omega)$ is a Hilbert space.

According to theorem 3.26 in [Mon03], $H(\mathrm{curl},\Omega)$ can be defined alternatively as

$$H(\mathrm{curl},\Omega) = \overline{C^\infty(\bar{\Omega})}^{\|\cdot\|_{H(\mathrm{curl},\Omega)}}.$$

Thus, the spaces $\left(C^\infty(\Omega)\right)^3$ and $\left(C(\Omega)\right)^3$ are dense in $H(\mathrm{curl},\Omega)$. This density of continuous functions together with the tangential trace operator

$$\mathrm{tr}_{\boldsymbol{\tau},\Gamma}(\boldsymbol{p}) = \boldsymbol{n}_\Gamma \times \boldsymbol{p} \qquad \text{for } \boldsymbol{p} \in \left(C(\Omega)\right)^3$$

leads to the trace theorem.

Theorem 3.13 (trace theorem for $H(\mathrm{curl},\Omega)$)**.** *Let Ω be a bounded Lipschitz domain. Then the operator $\mathrm{tr}_{\boldsymbol{\tau},\Gamma}$ can be uniquely extended to a continuous operator mapping from $H(\mathrm{curl},\Omega)$ to $\left(H^{-1/2}(\Gamma)\right)^3$, and*

$$\|\mathrm{tr}_{\boldsymbol{\tau},\Gamma}(\boldsymbol{p})\|^2_{H^{-1/2}(\Gamma)} \le c\|\boldsymbol{p}\|^2_{H(\mathrm{curl},\Omega)} \qquad \textit{for } \boldsymbol{p} \in H(\mathrm{curl},\Omega).$$

Note that the stated norm estimate is not sharp, and consequently, the inverse trace theorem does not exist in these norms. Nevertheless, the trace theorem allows us to introduce the space $H_0(\mathrm{curl},\Omega) := \left\{\boldsymbol{p} \in H(\mathrm{curl},\Omega) \;:\; \mathrm{tr}_{\boldsymbol{\tau},\Gamma}(\boldsymbol{p}) = \boldsymbol{0}\right\}$ with homogeneous tangential boundary conditions. Alternatively, $H_0(\mathrm{curl},\Omega)$ can be defined as the closure of $\left(C_0^\infty(\Omega)\right)^3$ with respect to the $H(\mathrm{curl})$-norm (see [Mon03]). In the following we will make use of the partial integration rule for $H(\mathrm{curl},\Omega)$ functions.

Lemma 3.14. *For a Lipschitz domain Ω and for $\boldsymbol{p} \in H(\mathrm{curl},\Omega)$ the integration by parts formula holds*

$$\int_\Omega (\mathrm{curl}\,\boldsymbol{p}) \cdot \boldsymbol{\phi}\, dx = \int_\Omega \boldsymbol{p} \cdot (\mathrm{curl}\,\boldsymbol{\phi})\, dx + \int_\Gamma \mathrm{tr}_{\boldsymbol{\tau},\Gamma}(\boldsymbol{p}) \cdot \boldsymbol{\phi}\, ds \qquad \forall \boldsymbol{\phi} \in \left(C^\infty(\bar{\Omega})\right)^3.$$

Finally, a corollary providing essential interface conditions for functions in $H(\mathrm{curl},\Omega)$ is provided.

Corollary 3.15. *Let $\Omega_1,\dots,\Omega_N$ be Lipschitz domains which form a domain decomposition of Ω, i.e. $\Omega_i \cap \Omega_j = \emptyset$ and $\bar{\Omega} = \bigcup_i \bar{\Omega}_i$. By $\Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j$ we denote the interfaces and by $\boldsymbol{n}_i$ their normal vectors. For a function $\boldsymbol{p}$ with $\boldsymbol{p}|_{\Omega_i} = \boldsymbol{p}_i$, $\boldsymbol{p}_i \in H(\mathrm{curl},\Omega_i)$ and $\mathrm{tr}_{\boldsymbol{\tau},\Gamma_{ij}}(\boldsymbol{p}_i) = \mathrm{tr}_{\boldsymbol{\tau},\Gamma_{ij}}(\boldsymbol{p}_j)$ follows that $\boldsymbol{p} \in H(\mathrm{curl},\Omega)$ and $(\mathrm{curl}\,\boldsymbol{p})|_{\Omega_i} = \mathrm{curl}\,\boldsymbol{p}_i$.*

3.2 Variational formulations

This section is started by recalling existence and uniqueness results for primal and mixed formulations. Then, based on the Sobolev Spaces from above, we introduce variational formulations for the Helmholtz equation and the vector valued wave equation, and existence of a unique solution is shown. For convenience we will use in the following the symbols $U = L^2(\Omega)$, $Q = H^1(\Omega)$, $V = H(\mathrm{div}, \Omega)$, $X = \left(L^2(\Omega)\right)^3$, $Y = H(\mathrm{curl}, \Omega)$ for the Sobolev spaces. Volume integrals over a domain Ω are denoted as $\left(u, v\right)_\Omega := \int_\Omega uv \, d\boldsymbol{x}$ and surface integrals on a surface Γ as $\left\langle u, v\right\rangle_\Gamma := \int_\Gamma uv \, ds$.

3.2.1 Existence and uniqueness for primal and mixed problems

First, we provide basic existence and uniqueness results for variational formulations. We consider for a Hilbert space W with norm $\|\cdot\|_W$ the primal formulation.

Formulation 3.16. *Find $u \in W$, such that*

$$a(u, v) = f(v) \qquad \forall v \in W$$

where $a : W \times W \to \mathbb{C}$ denotes a bilinear form, and $f : W \to \mathbb{C}$ a linear form.

Note that this variational formulation is equivalent to an operator equation $\mathcal{A}u = F$ in the dual space W^* of W. Here $\mathcal{A} : W \to W^*$ is defined via $\langle \mathcal{A}u, v\rangle = a(u, v)$ with the duality product $\langle \cdot, \cdot\rangle$, and $F \in W^*$ is given as $\langle F, v\rangle = f(v)$.

The following theorem (see for example [Bra03]) states existence and uniqueness for this formulation.

Theorem 3.17 (Lax-Milgram)**.** *Let W be a Hilbert space and $a : W \times W \to \mathbb{C}$ a bounded and coercive bilinear form, i.e. there exist constants $\alpha > 0$ and $\beta > 0$ such that*

$$\begin{aligned} |a(u, v)| &\le \alpha \|u\|_W \|v\|_W && \forall u \in W, \forall v \in W, \\ |a(u, u)| &\ge \beta \|u\|_W^2 && \forall u \in W. \end{aligned}$$

Then there exists for each continuous linear form $f : W \to \mathbb{C}$ a unique solution $u \in W$ of the problem in Formulation 3.16 with

$$\|u\|_W \le \frac{1}{\beta} \|f\|_{W^*}.$$

As second problem a saddle point problem for the Hilbert spaces W and M is considered.

Formulation 3.18. *Find $u \in W$ and $\lambda \in M$, such that*

$$\begin{aligned} a(u,v) + b(v,\lambda) &= f(v) && \forall v \in W, \\ b(u,\mu) &= g(\mu) && \forall \mu \in M, \end{aligned}$$

with the bilinear forms $a : W \times W \to \mathbb{C}$ and $b : W \times M \to \mathbb{C}$ and the linear forms $f : W \to \mathbb{C}$ and $g : M \to \mathbb{C}$.

Brezzi's Theorem (Theorem 1.1 in [BF91] or Satz 4.3 in [Bra03]) states existence and uniqueness for the solution of this formulation.

Theorem 3.19 (Brezzi)**.** *Let W and M be Hilbert spaces, and the bilinear forms $a : W \times W \to \mathbb{C}$ and $b : W \times M \to \mathbb{C}$ fulfill the following assumptions.*

(1) The bilinear forms are bounded, i.e.

$$\begin{aligned} |a(u,v)| &\le c_a \|u\|_W \|v\|_M && \forall u \in W, \forall v \in W \\ |b(u,\mu)| &\le c_b \|u\|_W \|\mu\|_M && \forall u \in W, \forall \mu \in M. \end{aligned}$$

(2) b is inf-sup stable. There exists a constant $\beta > 0$, such that

$$\inf_{\mu \in M} \sup_{v \in W} \frac{b(v,\mu)}{\|\mu\|_M \|v\|_W} \ge \beta.$$

(3) a is coercive on the kernel of b. Thus, there exists a constant $\alpha > 0$, such that

$$|a(u,u)| \ge \alpha \|u\|_W^2 \qquad \forall u \in \mathrm{Ker}(b),$$

with $\mathrm{Ker}(b) = \big\{ u \in W \,:\, b(u,\mu) = 0 \;\; \forall \mu \in M \big\}$.

Under these assumptions the problem from Formulation 3.18 has a unique solution $u \in W$, $\lambda \in M$ satisfying

$$\begin{aligned} \|u\|_W &= \frac{1}{\alpha}\|f\|_{W^*} + \frac{1}{\beta}\Big(1 + \frac{c_a}{\alpha}\Big)\|g\|_{M^*} \\ \|\lambda\|_M &\le \frac{1}{\beta}\Big(1 + \frac{c_a}{\alpha}\Big)\|f\|_{W^*} + \frac{c_a}{\beta^2}\Big(1 + \frac{c_a}{\alpha}\Big)\|g\|_{M^*}. \end{aligned}$$

3.2.2 The variational form of the primal Helmholtz problem

The weak formulation of the primal Helmholtz problem is derived by integrating the product of (2.28) with a complex conjugated test function $v \in Q = H^1(\Omega)$ over the domain Ω. Finally, integration by parts and inserting the boundary condition (2.29) leads to

Formulation 3.20 (The standard primal Helmholtz formulation)**.** *Find $u \in Q$, such that for $g \in L^2(\Gamma)$*

$$\big(\operatorname{grad} u, \operatorname{grad} \bar{v}\big)_\Omega - \big(\omega^2\nu^2 u, \bar{v}\big)_\Omega - \big\langle i\omega\nu u, \bar{v}\big\rangle_\Gamma = -\big\langle i\omega\nu g, \bar{v}\big\rangle_\Gamma, \qquad \forall v \in Q. \tag{3.1}$$

Lemma 3.21. *Let Ω be a Lipschitz domain, $\nu \in \mathbb{R}$, $\nu > 0$ and $g \in L^2(\Gamma)$. Then the problem from Formulation 3.20 has a unique solution.*

Note, that the formulation is not coercive, and Lax-Milgram is not applicable. A proof of this Lemma is given in [Mel95]. We will repeat it in a slightly different way. In order to proof this Lemma, we need the Fredholm Alternative (compare theorem 2.33 in [Mon03])

Theorem 3.22 (Fredholm Alternative)**.** *Let W be a Hilbert space and $\mathcal{B} : W \to W$ a bounded linear operator, which can be written as the sum of the identity operator $\mathcal{I}$ and a compact operator $\mathcal{K}$, i.e. $\mathcal{B} = \mathcal{I} + \mathcal{K}$. Then either*

(1) The equation $\mathcal{B}u = 0$ has just the trivial solution $u = 0$ in W, and $\mathcal{B}u = f$ is for $f \in W$ uniquely solvable.

(2) The equation $\mathcal{B}u = 0$ has exactly $n \in \mathbb{N}$ linearly independent solutions.

Proof of Lemma 3.21. During the proof, we will denote (3.1) as

$$B(u, v) = F(v)$$

with a sesquilinear form B and a linear form F. The proof splits into two parts, an existence proof and a uniqueness proof.

The existence proof is realized by transforming the formulation into an operator equation

$$(\mathcal{I} + \mathcal{K})u = \phi$$

with $\phi \in L^2(\Omega)$, the identity operator $\mathcal{I}$ and a compact operator $\mathcal{K} : L^2(\Omega) \to L^2(\Omega)$ and by applying the Fredholm Alternative. In order to define $\mathcal{K}$, we introduce the sesquilinear form

$$B_+(u,v) := B(u,v) + \big((1+\omega^2\nu^2)u, \bar{v}\big)_\Omega,$$

which is bounded, and via

$$\big|\mathrm{Re}\big(B_+(u,u)\big)\big| \geq \|u\|^2_{H^1(\Omega)}$$

coercivity can be shown. Rewriting (3.1), we get $B_+(u,v) - \big((1+\omega^2\nu^2)u, \bar{v}\big)_\Omega = F(v)$, and inserting the operator equation $u = \phi - \mathcal{K}u$ leads to

$$B_+(\phi - \mathcal{K}u, v) - \big((1+\omega^2\nu^2)u, \bar{v}\big)_\Omega = F(v), \qquad \forall v \in Q.$$

This allows us to state for any $u \in L^2(\Omega)$ equations for $\mathcal{K}u \in Q \subset L^2(\Omega)$ and $\phi \in Q \subset L^2(\Omega)$

$$\begin{aligned} B_+(\mathcal{K}u, v) &= -\big((1+\omega^2\nu^2)u, \bar{v}\big)_\Omega & \forall v \in Q, \\ B_+(\phi, v) &= F(v) & \forall v \in Q. \end{aligned}$$

For these two equations Lax-Milgram's theorem is due to the boundedness and coercivity of B_+ applicable. Thus, ϕ and $\mathcal{K}u$ exist and they are unique. From the estimate $\|\mathcal{K}u\|_{H^1(\Omega)} \leq c\|u\|_{L^2(\Omega)}$ and the compact embedding of Q in $L^2(\Omega)$ the compactness of $\mathcal{K}$ follows. The Fredholm Alternative states now, that the operator equation, and consequently Formulation 3.20 has at least one solution.

In order to show uniqueness, we have to show that $B(u,v) = 0$ has just the trivial solution. Testing this equation with u and taking the imaginary part leads to

$$0 = \big|\mathrm{Im}\big(B(u,u)\big)\big| = \langle \omega\nu u, \bar{u}\rangle_\Gamma \geq \omega\nu\|u\|^2_{L^2(\Gamma)}.$$

Thus, u is zero on Γ and u is an eigenfunction to the eigenvalue $\omega^2\nu^2$ of the problem

$$\big(\operatorname{grad} u, \operatorname{grad} \bar{v}\big)_\Omega = \omega^2\nu^2\big(u, \bar{v}\big)_\Omega \qquad \forall v \in Q.$$

Now we choose $R > 0$, such that Ω is contained in a sphere of radius R round the origin,

$\Omega \subset B_R(\mathbf{0})$. Then for $c \geq 1$ the function

$$\tilde{u}(\boldsymbol{x}) = \begin{cases} u(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in \Omega, \\ 0 & \text{else,} \end{cases}$$

is an eigenfunction of

$$\big(\operatorname{grad} \tilde{u}, \operatorname{grad} \bar{v}\big)_{B_{cR}(\mathbf{0})} = \omega^2 \nu^2 \big(\tilde{u}, \bar{v}\big)_{B_{cR}(\mathbf{0})} \qquad \forall v \in H^1\big(B_{cR}(\mathbf{0})\big). \tag{3.2}$$

By defining $u_c(\boldsymbol{x}) = \tilde{u}(c\boldsymbol{x})$, it is obvious that u_c is an eigenfunction of

$$\big(\operatorname{grad} u_c, \operatorname{grad} \bar{v}\big)_{B_R(\mathbf{0})} = c^2 \omega^2 \nu^2 \big(u_c, \bar{v}\big)_{B_R(\mathbf{0})} \qquad \forall v \in H^1\big(B_R(\mathbf{0})\big) \tag{3.3}$$

to the eigenvalue $c^2\omega^2\nu^2$. Taking $c \in \mathbb{R}$ with $c \geq 1$ arbitrarily contradicts to the fact that (3.3) has just countably infinite many eigenvalues. Thus, $u_c = 0$ and consequently $u = 0$.

□

3.2.3 The variational form of the mixed Helmholtz problem

We get the variational formulation of the mixed Helmholtz problem by integrating the product of (2.30) and (2.31), respectively, with a complex conjugated scalar test function $v \in U = L^2(\Omega)$ and a complex conjugated vector valued test function $\tau \in V = H(\operatorname{div}, \Omega)$ over the domain Ω. Integration by parts of the second equation gives

$$\big(i\omega\epsilon u, \bar{v}\big)_\Omega - \big(\operatorname{div} \boldsymbol{\sigma}, \bar{v}\big)_\Omega = 0 \qquad \forall v \in U \tag{3.4}$$

$$-\big(u, \operatorname{div} \bar{\boldsymbol{\tau}}\big)_\Omega - \big(i\omega\mu\boldsymbol{\sigma}, \bar{\boldsymbol{\tau}}\big)_\Omega + \big\langle u, \bar{\boldsymbol{\tau}} \cdot \boldsymbol{n}_\Gamma \big\rangle_\Gamma = 0 \qquad \forall \boldsymbol{\tau} \in V. \tag{3.5}$$

After inserting the boundary condition (2.32) the weak formulation is obtained.

Formulation 3.23 (The standard mixed Helmholtz formulation)**.** *Find $u \in U$ and $\boldsymbol{\sigma} \in V$, such that for $g \in L^2(\Gamma)$*

$$\big(i\omega\epsilon u, \bar{v}\big)_\Omega - \big(\operatorname{div} \boldsymbol{\sigma}, \bar{v}\big)_\Omega = 0 \qquad \forall v \in U \tag{3.6}$$

$$-\big(u, \operatorname{div} \bar{\boldsymbol{\tau}}\big)_\Omega - \big(i\omega\mu\boldsymbol{\sigma}, \bar{\boldsymbol{\tau}}\big)_\Omega + \big\langle \sqrt{\frac{\mu}{\epsilon}}\boldsymbol{\sigma} \cdot \boldsymbol{n}_\Gamma, \bar{\boldsymbol{\tau}} \cdot \boldsymbol{n}_\Gamma \big\rangle_\Gamma = -\big\langle g, \bar{\boldsymbol{\tau}} \cdot \boldsymbol{n}_\Gamma \big\rangle_\Gamma. \qquad \forall \boldsymbol{\tau} \in V. \tag{3.7}$$

Lemma 3.24. *Let Ω be a Lipschitz domain, μ is constant and real, ϵ is a real and piecewise continuous function with $0 < \epsilon_{min} \leq \epsilon \leq \epsilon_{max}$ and $g \in L^2(\Gamma)$. Then the problem from*

Formulation 3.23 has a unique solution.

A proof to this lemma with $\mu = \epsilon = 1$ can be found for three dimensions in [MSS10] which we will repeat for ϵ space dependent.

Proof. We start the proof by testing (3.6) with $v = -\frac{1}{i\omega\epsilon} \operatorname{div} \boldsymbol{\tau}$ and $\boldsymbol{\tau} \in V$ which results in

$$\big(u, \operatorname{div} \bar{\boldsymbol{\tau}}\big)_\Omega = \big(\frac{1}{i\omega\epsilon} \operatorname{div} \boldsymbol{\sigma}, \operatorname{div} \bar{\boldsymbol{\tau}}\big)_\Omega.$$

Inserting this into (3.6) gives for all $\boldsymbol{\tau} \in V$

$$B(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \big(\frac{1}{\epsilon} \operatorname{div} \boldsymbol{\sigma}, \operatorname{div} \bar{\boldsymbol{\tau}}\big)_\Omega - \big(\omega^2 \mu \boldsymbol{\sigma}, \bar{\boldsymbol{\tau}}\big)_\Omega - \langle i\omega \sqrt{\frac{\mu}{\epsilon}} \boldsymbol{\sigma} \cdot \boldsymbol{n}_\Gamma, \bar{\boldsymbol{\tau}} \cdot \boldsymbol{n}_\Gamma \rangle_\Gamma = \langle i\omega g, \bar{\boldsymbol{\tau}} \cdot \boldsymbol{n}_\Gamma \rangle_\Gamma. \tag{3.8}$$

For this equation existence and uniqueness will be proven.

In order to prove existence, we restrict the function space. Therefore, the equation is tested with $\boldsymbol{\tau} = \operatorname{curl} \boldsymbol{q}$, $\boldsymbol{q} \in \big(C_0^\infty(\Omega)\big)^d$, which leads to $-\big(\omega^2 \mu \boldsymbol{\sigma}, \operatorname{curl} \bar{\boldsymbol{q}}\big)_\Omega = 0$. Note that the curl of a two dimensional vector function is defined as $\operatorname{curl} \boldsymbol{q} := \partial_x q_y - \partial_y q_x$. Partial integration and density arguments lead to $\operatorname{curl} \boldsymbol{\sigma} = 0$, and we can exchange V in (3.8) by

$$V^{(0)}(\Omega) := H(\operatorname{div}, \Omega) \cap H(\operatorname{curl}_0, \Omega)$$

with $H(\operatorname{curl}_0, \Omega) := \big\{\boldsymbol{\tau} \in H(\operatorname{curl}, \Omega) \,:\, \operatorname{curl} \boldsymbol{\tau} = 0\big\}$. The rest of the existence proof follows the existence proof for Lemma 3.21. Thus, (3.8) is transformed to an operator equation

$$(\mathcal{I} + \mathcal{K}) \boldsymbol{\sigma} = \boldsymbol{\phi},$$

with the operator $\mathcal{K} \,:\, \big(L^2(\Omega)\big)^d \to \big(L^2(\Omega)\big)^d$ and $\boldsymbol{\phi} \in \big(L^2(\Omega)\big)^d$. The function $\boldsymbol{\phi}$ and the operator $\mathcal{K}$ can be defined uniquely via a sesquilinear form

$$B_+(\boldsymbol{\sigma}, \boldsymbol{\tau}) := B(\boldsymbol{\sigma}, \boldsymbol{\tau}) + \big((1 + \omega^2 \mu) \boldsymbol{\sigma}, \bar{\boldsymbol{\tau}}\big)_\Omega$$

which is bounded and coercive with respect to the norm

$$\|\boldsymbol{v}\|_V^2 := \|\boldsymbol{v}\|_{H(\operatorname{div}, \Omega)}^2 + \|\boldsymbol{v} \cdot \boldsymbol{n}_\Gamma\|_{L^2(\Gamma)}^2.$$

The compact embedding of $V^{(0)}(\Omega)$ in $L^2(\Omega)$ (compare Theorem 3.49 in [Mon03]) leads to compactness of $\mathcal{K}$, and Fredholm's Alternative guarantees at least one solution $\boldsymbol{\sigma}$.

For proving uniqueness, it remains to show that the only solution of $B(\boldsymbol{\sigma},\boldsymbol{\tau}) = 0$ is $\boldsymbol{\sigma} = \mathbf{0}$. The proof starts by testing (3.8) with $\boldsymbol{\sigma}$ and taking the imaginary part which leads to

$$0 = \big|\mathrm{Im}\big(B(\boldsymbol{\sigma},\boldsymbol{\sigma})\big)\big| \geq \omega\sqrt{\frac{\mu}{\epsilon_{max}}}\|\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma\|^2_{L^2(\Gamma)} \geq 0$$

and consequently $\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma = 0$ on the boundary. Next (3.8) is tested with $\boldsymbol{\tau} = \mathrm{curl}\,\boldsymbol{\phi}$, $\boldsymbol{\phi} \in \big(C^\infty(\Omega)\big)^d$ which yields

$$\big(\omega^2\mu\boldsymbol{\sigma}, \mathrm{curl}\,\bar{\boldsymbol{\phi}}\big)_\Omega = \big(\omega^2\mu\,\mathrm{curl}\,\boldsymbol{\sigma}, \bar{\boldsymbol{\phi}}\big)_\Omega + \big\langle\omega^2\mu\boldsymbol{\sigma}\times\boldsymbol{n}_\Gamma, \bar{\boldsymbol{\phi}}\big\rangle_\Gamma = 0.$$

If $\boldsymbol{\phi}$ has compact support, density arguments give $\omega^2\mu\,\mathrm{curl}\,\boldsymbol{\sigma} = 0$. A similar argumentation gives $\boldsymbol{\sigma}\times\boldsymbol{n}_\Gamma = \mathbf{0}$ and consequently $\boldsymbol{\sigma} = 0$ on Γ. Thus, there exists a gradient field w with $w = 0$ on Γ, such that $\omega^2\mu\boldsymbol{\sigma} = \mathrm{grad}\,w$. Inserting this into (3.8), we obtain

$$\begin{aligned} 0 &= \big(\frac{1}{\omega^2\epsilon\mu}\,\mathrm{div}\,\mathrm{grad}\,w, \mathrm{div}\,\bar{\boldsymbol{\tau}}\big)_\Omega - \big(\mathrm{grad}\,w, \bar{\boldsymbol{\tau}}\big)_\Omega \\ &= \big(\Delta w + \omega^2\epsilon\mu w, \mathrm{div}\,\bar{\boldsymbol{\tau}}\big)_\Omega. \end{aligned}$$

Because the operator div maps $H(\mathrm{div},\Omega)$ surjective on $L^2(\Omega)$ [GR86], it follows that $\Delta w + \omega^2\epsilon\mu w = 0$ in Ω with $w = 0$ and $\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma = \frac{1}{i\omega\mu}\frac{\partial w}{\partial \boldsymbol{n}} = 0$ on the boundary. Solving the Helmholtz equation under such boundary conditions leads to $w = 0$, i.e. $\boldsymbol{\sigma} = \mathbf{0}$ and $u = 0$. □

3.2.4 The variational form of the primal vector valued wave problem

The variational formulation corresponding to (2.33) is obtained by testing with $\boldsymbol{e} \in H(\mathrm{curl},\Omega) =: Y$, integrating by parts and inserting the boundary condition (2.34).

Formulation 3.25 (The standard primal vector valued wave formulation)**.** *Find $\boldsymbol{E} \in Y$, such that for $\boldsymbol{g} \in \big(L^2(\Gamma)\big)^3$ with $\boldsymbol{g}\cdot\boldsymbol{n}_\Gamma = 0$ on Γ and for all $\boldsymbol{e} \in Y$*

$$\big(\mathrm{curl}\,\boldsymbol{E}, \mathrm{curl}\,\bar{\boldsymbol{e}}\big)_\Omega - \big(\omega^2\epsilon\mu\boldsymbol{E}, \bar{\boldsymbol{e}}\big)_\Omega - \big\langle i\omega\sqrt{\epsilon\mu}\,\boldsymbol{n}_\Gamma\times\boldsymbol{E}, \boldsymbol{n}_\Gamma\times\bar{\boldsymbol{e}}\big\rangle_\Gamma = -\big\langle i\omega\sqrt{\epsilon\mu}\boldsymbol{g}, \bar{\boldsymbol{e}}\big\rangle_\Gamma. \quad (3.9)$$

Like in the scalar case, we state existence and uniqueness.

Lemma 3.26. *Let μ be a positive constant, $\epsilon \in H^3(\Omega)$ is a real valued function with $0 < \epsilon_{min} \leq \epsilon \leq \epsilon_{max}$, $\boldsymbol{g} \in \big(L^2(\Gamma)\big)^3$ with $\boldsymbol{g}\cdot\boldsymbol{n}_\Gamma = 0$ on Γ and Ω a Lipschitz domain. Then* *a unique solution of the problem from Formulation 3.25 exists.*

Remark 3.27. *In chapter 4 of [Mon03] uniqueness and existence was proven for a more general setting. There Ω is considered to consist of N subdomains, and ϵ is at least in $H^3(\Omega_j)$ for each subdomain Ω_j. Furthermore, ϵ is allowed to be complex with an imaginary part which is either positive and bounded away from zero by a constant or zero on each subdomain.*
We will follow this proof for our simplified setting.

For proving the Lemma, a continuation result stated in theorem 9.3 of [CK98] or theorem 4.13 in [Mon03] is needed.

Theorem 3.28. *Let Ω be an open connected domain, and $\boldsymbol{E}, \boldsymbol{H} \in H(\mathrm{curl}, \Omega)$ fulfill on Ω*

$$\begin{aligned} i\omega\epsilon\boldsymbol{E} + \mathrm{curl}\,\boldsymbol{H} &= \boldsymbol{0}, \\ i\omega\mu\boldsymbol{H} - \mathrm{curl}\,\boldsymbol{E} &= \boldsymbol{0}. \end{aligned}$$

If ϵ is real and continuously differentiable in $\overline{\Omega}$, μ is real and constant, and $\boldsymbol{E}$ is zero in a non empty ball contained in Ω, then $\boldsymbol{E}$ and $\boldsymbol{H}$ are zero on Ω.
For a proof of this theorem see [CK98, Mon03].

Proof of Lemma 3.26. During the proof we will denote (3.9) as

$$B(\boldsymbol{E}, \boldsymbol{e}) = F(\boldsymbol{e}) \qquad \forall \boldsymbol{e} \in Y, \tag{3.10}$$

with the sesquilinear form B and the linear form F. As for the other formulations, the proof is divided in an existence and a uniqueness proof.

For the existence proof we have to restrict the function space. Therefore, we test (3.10) with $\boldsymbol{e} = \mathrm{grad}\, q$, $q \in H_0^1(\Omega)$, and we obtain $\left(\omega^2\mu\epsilon\boldsymbol{E}, \mathrm{grad}\,\bar{q}\right)_\Omega = 0$. Thus we can exchange the space Y by

$$Y^{(0)}(\Omega) = \left\{\boldsymbol{v} \in H(\mathrm{curl}, \Omega) \,:\, \left(\epsilon\boldsymbol{v}, \mathrm{grad}\, q\right)_\Omega = 0 \quad \forall q \in H_0^1(\Omega)\right\}.$$

According to theorem 4.7 in [Mon03] $Y^{(0)}(\Omega)$ can be embedded compactly in $\left(L^2(\Omega)\right)^3$. The rest of the proof follows the existence proof for Lemma 3.21, i.e. (3.10) is transformed to an operator equation

$$(\mathcal{I} + \mathcal{K})\boldsymbol{E} = \boldsymbol{\phi}$$

with $\boldsymbol{\phi} \in \left(L^2(\Omega)\right)^3$ and an operator $\mathcal{K} : \left(L^2(\Omega)\right)^3 \to \left(L^2(\Omega)\right)^3$. These two quantities can be defined uniquely via a sesquilinear form

$$B_+(\boldsymbol{E}, \boldsymbol{e}) := B(\boldsymbol{E}, \boldsymbol{e}) + \left((1 + \omega^2 \epsilon \mu)\boldsymbol{E}, \bar{\boldsymbol{e}}\right)_\Omega$$

which is bounded and coercive with respect to the norm

$$\|\boldsymbol{v}\|_Y^2 := \|\boldsymbol{v}\|^2_{H(\mathrm{curl},\Omega)} + \|\boldsymbol{v} \times \boldsymbol{n}_\Gamma\|^2_{(L^2(\Gamma))^3}.$$

Since there exists a compact embedding of $Y^{(0)}(\Omega)$ in $\left(L^2(\Omega)\right)^3$, the operator $\mathcal{K}$ is compact, and according to the Fredholm Alternative, there exists a solution of the operator equation and consequently of (3.10).

In order to show uniqueness, it is sufficient to show that

$$B(\boldsymbol{E}, \boldsymbol{e}) = 0, \qquad \forall \boldsymbol{e} \in Y^{(0)}(\Omega) \tag{3.11}$$

implies $\boldsymbol{E} = \boldsymbol{0}$. Testing equation (3.11) with $\boldsymbol{E}$ and taking the imaginary part leads to

$$0 = \left\langle \omega \sqrt{\mu\epsilon}\, \boldsymbol{n}_\Gamma \times \boldsymbol{E}, \boldsymbol{n}_\Gamma \times \overline{\boldsymbol{E}} \right\rangle_\Gamma \geq c \|\boldsymbol{n}_\Gamma \times \boldsymbol{E}\|^2_{L^2(\Gamma)}.$$

Thus, $\boldsymbol{n}_\Gamma \times \boldsymbol{E}$ is zero on Γ, i.e. the tangential component of $\boldsymbol{E}$ is zero on the boundary.

Next we extend via the Cálderon extension theorem (see [Ada75], Theorem A.4 in [McL00] or Theorem 3.2 in [Mon03]) the function ϵ, which is according to our assumption in $H^3(\Omega)$, to a continuous differentiable function $\tilde{\epsilon}$ on the whole domain $\mathbb{R}^3$. Now, we take a ball $B_r(\boldsymbol{x}_0)$ with any midpoint $\boldsymbol{x}_0 \in \Gamma$ whose radius $r > 0$ is chosen, such that $\tilde{\epsilon}$ is positive in the ball. This is possible, because $\tilde{\epsilon}$ is continuous and bounded away from zero by ϵ_{min} in Ω. Then a function $\tilde{\boldsymbol{E}}$ on $\tilde{\Omega} := B_r(\boldsymbol{x}_0) \cup \Omega$ is constructed by extending a solution $\boldsymbol{E}$ of (3.11) by zero. Note that $\boldsymbol{n}_{\partial\tilde{\Omega}} \times \tilde{\boldsymbol{E}} = \boldsymbol{0}$ on $\partial\tilde{\Omega}$ and that $\tilde{\boldsymbol{E}}$ is tangential continuous across the interface $\tilde{\Gamma} := \Gamma \cap B_r(\boldsymbol{x}_0)$. Thus, $\tilde{\boldsymbol{E}} \in Y^{(0)}(\Omega)$ (compare Lemma 3.15).
Now we realize that $\tilde{\boldsymbol{E}}$ solves

$$\left(\mathrm{curl}\, \tilde{\boldsymbol{E}}, \mathrm{curl}\, \bar{\boldsymbol{e}}\right)_\Omega - \left(\omega^2 \mu \tilde{\epsilon} \tilde{\boldsymbol{E}}, \bar{\boldsymbol{e}}\right)_\Omega = 0 \qquad \forall \boldsymbol{e} \in Y^{(0)}(\tilde{\Omega}),$$

i.e. it solves the wave equation on $\tilde{\Omega}$. Because $\tilde{\boldsymbol{E}}$ is zero on $B_r(\boldsymbol{x}_0) \setminus \Omega$, Theorem 3.28 implies $\tilde{\boldsymbol{E}} = \boldsymbol{0}$ on $\tilde{\Omega}$ and consequently $\boldsymbol{E} = \boldsymbol{0}$.

□

3.2.5 The variational form of the mixed vector valued wave problem

Testing the equations (2.35) and (2.36) with complex conjugated test functions $\boldsymbol{e} \in X := \big(L^2(\Omega)\big)^3$ and $\boldsymbol{h} \in Y := H(\mathrm{curl}, \Omega)$, integrating the second equation by parts and inserting the boundary condition results in the variational formulation for the mixed vector valued wave problem.

Formulation 3.29 (The standard mixed vector valued wave formulation)**.** *Find* $\boldsymbol{E} \in X$ *and* $\boldsymbol{H} \in Y$*, such that for* $\boldsymbol{g} \in \big(L^2(\Gamma)\big)^3$ *and for all* $\boldsymbol{e} \in X$ *and* $\boldsymbol{h} \in Y$

$$\begin{aligned}
\big(i\omega\epsilon\boldsymbol{E}, \bar{\boldsymbol{e}}\big)_\Omega + \big(\mathrm{curl}\,\boldsymbol{H}, \bar{\boldsymbol{e}}\big)_\Omega &= 0 && (3.12)\\
\big(\boldsymbol{E}, \mathrm{curl}\,\bar{\boldsymbol{h}}\big)_\Omega - \big(i\omega\mu\boldsymbol{H}, \bar{\boldsymbol{h}}\big)_\Omega + \big\langle\sqrt{\frac{\mu}{\epsilon}}\boldsymbol{n}_\Gamma \times \boldsymbol{H}, \boldsymbol{n}_\Gamma \times \bar{\boldsymbol{h}}\big\rangle_\Gamma &= \big\langle\boldsymbol{g}, \boldsymbol{n}_\Gamma \times \bar{\boldsymbol{h}}\big\rangle_\Gamma. && (3.13)
\end{aligned}$$

For this setting we can guarantee existence and uniqueness.

Lemma 3.30. *Let* μ *be a positive constant,* $\epsilon \in H^3(\Omega)$ *is a real valued function with* $0 < \epsilon_{min} \leq \epsilon \leq \epsilon_{max}$*,* $\boldsymbol{g} \in \big(L^2(\Gamma)\big)^3$ *and* Ω *a Lipschitz domain. Then a unique solution of the problem from Formulation 3.29 exists.*

Proof. We rewrite Formulation 3.29 by testing (3.12) with $\boldsymbol{e} = -\frac{1}{i\omega\epsilon}\,\mathrm{curl}\,\boldsymbol{h}$, $\boldsymbol{h} \in H(\mathrm{curl}, \Omega)$, i.e.

$$\big(\boldsymbol{E}, \mathrm{curl}\,\bar{\boldsymbol{h}}\big)_\Omega = -\big(\frac{1}{i\omega\epsilon}\,\mathrm{curl}\,\boldsymbol{H}, \mathrm{curl}\,\bar{\boldsymbol{h}}\big)_\Omega.$$

Inserting this into (3.13) results in the equation

$$\begin{aligned}
B(\boldsymbol{H}, \boldsymbol{h}) &:= \big(\frac{1}{\epsilon}\,\mathrm{curl}\,\boldsymbol{H}, \mathrm{curl}\,\bar{\boldsymbol{h}}\big)_\Omega - \big(\omega^2\mu\boldsymbol{H}, \bar{\boldsymbol{h}}\big)_\Omega - \big\langle i\omega\sqrt{\frac{\mu}{\epsilon}}\boldsymbol{n}_\Gamma \times \boldsymbol{H}, \boldsymbol{n}_\Gamma \times \bar{\boldsymbol{h}}\big\rangle_\Gamma\\
&= \big\langle -i\omega\boldsymbol{g}, \boldsymbol{n}_\Gamma \times \bar{\boldsymbol{h}}\big\rangle_\Gamma && (3.14)
\end{aligned}$$

where $\boldsymbol{H}, \boldsymbol{h} \in H(\mathrm{curl}, \Omega)$. The existence and uniqueness proof for this equation follows the proof of Lemma 3.26. □

Chapter 4

The finite element method

This chapter is devoted to the finite element method. We start by describing the basic concept of finite elements in Section 4.1, and we recall some fundamental discretization results. When solving the problems arising from the Formulations 3.20, 3.23, 3.25 or 3.29 by a finite element method, suitable discrete approximations U_h, V_h, Y_h and Q_h of the infinite dimensional spaces $U = L^2(\Omega)$, $V = H(\mathrm{div}, \Omega)$, $Y = H(\mathrm{curl}, \Omega)$ and $Q = H^1(\Omega)$ are needed. Therefore, we introduce in Section 4.2 conforming finite elements for these spaces.

4.1 Basic ingredients

4.1.1 The Galerkin framework and approximation results

We consider again a standard variational formulation 3.16, i.e. we search for a function $u \in W$ which solves

$$a(u, v) = f(v) \qquad \forall v \in W. \tag{4.1}$$

In a Galerkin approximation the infinite dimensional Hilbert space is replaced by a finite dimensional space W_h, which results in the discrete formulation

Formulation 4.1. *Find $u_h \in W_h$, such that*

$$a(u_h, v_h) = f(v_h) \qquad \forall v_h \in W_h.$$

The subscript h stands for a discretization parameter, for example the mesh size if W_h is constructed via a triangulation. If $W_h \subset W$, we call the method *conforming*. For a basis

$\{\phi_1, \ldots, \phi_n\}$ of W_h the unknown function $u_h \in W_h$ can be expanded into this basis, i.e.

$$u_h = \sum_{i=1}^{n} u_{h,i}\phi_i.$$

Testing the variational equation with these basis functions leads to a linear system of equations $A\boldsymbol{u} = \boldsymbol{f}$ with the matrix $A \in \mathbb{C}^{n\times n}$ where $A_{ij} = a(\phi_i, \phi_j)$, a right hand side vector $\boldsymbol{f} \in \mathbb{C}^n$ with $f_i = f(\phi_i)$ and the coefficient vector $\boldsymbol{u} \in \mathbb{C}^n$ with $u_i = u_{h,i}$.

It is desirable that the approximate solution converges towards the original solution u for finer discretizations, i.e. $h \to 0$. For a coercive bilinear form a, or more precisely, if the problems arising from the Formulations 3.16 and 4.1 fulfill the assumptions of the Lax-Milgram Theorem 3.17, the discretization error $\|u - u_h\|_W$ can be expressed via approximation properties of the space W_h.

Theorem 4.2 (Cea). *Let W_h be a subspace of W, and Formulation 3.16 with the solution $u \in W$ as well as Formulation 4.1 with the solution $u_h \in W_h$ fulfill the assumptions of Theorem 3.17. Then*

$$\|u - u_h\|_W \le \frac{\alpha}{\beta} \inf_{v_h \in W_h} \|u - v_h\|_W$$

where α and β are the continuity and coercivity constants of the formulation for continuous spaces.

In the second part of this chapter, we will give some estimates for the approximation error $\inf_{v_h\in W_h} \|u - v_h\|_W$ of different spaces W_h.

Similar results can be obtained for the mixed formulation 3.18. By discretizing the Hilbert spaces W and M, the discrete formulation is obtained.

Formulation 4.3. *Find $u_h \in W_h$ and $\lambda_h \in M_h$, such that*

$$\begin{aligned} a(u_h, v_h) + b(v_h, \lambda_h) &= f(v_h) && \forall v_h \in W_h,\\ b(u_h, \mu_h) &= g(\mu_h) && \forall \mu_h \in M_h. \end{aligned}$$

The following theorem (compare [BF91]) connects the discretization error with the approximation error for mixed formulations.

Theorem 4.4. *Let W_h and M_h be subspaces of W and M. Furthermore, Formulation 3.18 fulfills the assumptions of Brezzi's theorem (Theorem 3.19) and Formulation 4.3 fulfills these assumptions with constants c_a', c_b', α', β' independent of the discretization parameter* *h. Then*

$$\begin{aligned}\|u-u_h\|_W &\le \Big(1+\frac{c_a'}{\alpha'}\Big)\Big(1+\frac{c_b'}{\beta'}\Big)\inf_{v_h\in W_h}\|u-v_h\|_W+\frac{c_b'}{\alpha'}\inf_{\mu_h\in M_h}\|\lambda-\mu_h\|_M\\ \|\lambda-\lambda_h\|_M &\le \Big(1+\frac{c_b'}{\beta'}\Big)\inf_{\mu_h\in M_h}\|\lambda-\mu_h\|_M+\frac{c_a'}{\beta'}\inf_{v_h\in W_h}\|u-v_h\|_W.\end{aligned}$$

4.1.2 The triangulation and the finite element

For the finite element method these discrete spaces are based on a triangulation of the computational domain Ω and the subscript h refers to the mesh size. The computational domain $\Omega \subset \mathbb{R}^d$ with $d = 2, 3$ is considered to be a polygonal domain with a Lipschitz continuous boundary. This leads to the definition of a (regular) triangulation.

Definition 4.5 (Regular Triangulation)**.** *A finite, non overlapping subdivision* $\mathcal{T} = \{T_1, \ldots, T_M\}$ *of* Ω *into elements* T_i *of simple geometry is called a regular triangulation if*

1. $\overline{\Omega} = \bigcup_{i=1}^{N} T_i$,
2. *the elements are non overlapping,* $interior(T_i) \cap interior(T_j) = \emptyset$ *for* $i \neq j$,
3. *the intersection of two different elements* $T_i \cap T_j$ *is either empty, a vertex, an edge or a face of both.*

Note that the last condition avoids hanging nodes. In the following we will call the set of all vertices $\mathcal{V} := \{V_i\}$, the set of all edges is called $\mathcal{E} = \{E_i\}$, and $\mathcal{F} := F_i$ represents the set of all interface and boundary facets. In two dimensions the facets $\mathcal{F}$ match with the edges $\mathcal{E}$. If the index T is added, i.e. $\mathcal{V}_T, \mathcal{E}_T, \mathcal{F}_T$, we refer to the set of vertices, edges and facets of the element T.

Based on a geometrical element T, a finite element can be constructed according to Ciarlet [Cia78, BS08] as

Definition 4.6 (Finite Element)**.** *A finite element is a triple* (T, W_T, Σ_T)*, where*

1. *the element domain* $T \subset \mathbb{R}^d$ *is a bounded closed set with non empty interior and piecewise smooth boundary,*
2. W_T *is the finite dimensional space of shape functions on* T,

3. *$\Sigma_T = \{S_1, \dots, S_k\}$ is a set of linear functionals on W_T which are also called degrees of freedom.*

Of course, W_T and Σ_T can not be taken arbitrarily. They have to be chosen such that if values of the degrees of freedom Σ_T are specified, a function in W_T can be determined uniquely. If this is the case for the finite element, we call it *unisolvent.*

In general, any linear independent set of functions of W_T can be used to span the space W_T. One option is the nodal basis $\{\phi_1, \dots, \phi_k\}$ which is defined via the linear functionals $S_i \in \Sigma_T$,

$$S_i(\phi_j) = \delta_{ij}, \qquad \text{for } i, j = 1, \dots, k.$$

This requires that the function space W_T is also of dimension k, and any function w in W_T can be written in terms of the nodal basis as

$$w(x) = \sum_{j=1}^{k} S_j(w)\phi_j(x).$$

By identifying the local degrees of freedom Σ_T with global ones, i.e. we take the union of the degrees of freedom $\Sigma = \bigcup_{T\in\mathcal{T}} \Sigma_T$, we are able to construct a global finite element space

$$W_h := \left\{ v \in \prod_{T\in\mathcal{T}} W_T \ : \ S(v|_{T_i}) = S(v|_{T_j}) \text{ for all } S \in \Sigma_{T_i} \cap \Sigma_{T_j} \right\}.$$

Thus, when specifying the values of the degrees of freedom in Σ, the corresponding function in W_h consists of functions in W_T, which are uniquely determined via the functionals Σ_T. It is obvious that different choices of the degrees of freedom on the element result in different continuity properties for the global finite element space W_h. This leads us to the following definition.

Definition 4.7 (conforming finite elements)**.** *For a function space W, the finite element (T, W_T, Σ_T) is called $W-$ conforming if the global finite element space W_h is a subspace of W.*

4.1.3 The reference element

For an actual implementation and for analysis purposes a reference finite element $(\hat{T}, \hat{W}_T, \hat{\Sigma}_T)$ is rather used than the physical element (T, W_T, Σ_T). More precisely, the shape functions are defined on the reference element, and calculations, for example numerical integration, are done there. The results are transfered afterwards by a transformation F_T to the physical element. As a reference element $\hat{T}$ for the element T we usually take an element of simple form like a triangle in 2D or a tetrahedron in 3D with size one. In the following, quantities related to the reference element are indicated with a hat. Thus, $\hat{\boldsymbol{x}}$ represents a point in the reference element, while $\boldsymbol{x}$ represents a point in the physical element.

As transformation

$$F_T \; : \; \hat{T} \to T$$

we consider a continuously differentiable function. The Jacobi matrix with respect to the coordinate system of the reference element we denote as J_T and its functional determinant as $\det J_T$. For a general setting F_T is allowed to be nonlinear. Polynomial transformations are required for example to describe polynomial shapes of elements exactly. If no curved boundaries are involved, F_T can be assumed to be affine linear with a constant Jacobian J_T. In this case polynomials are mapped to polynomials of the same degree which simplifies the analysis considerably.

In our work, the transformation of the normal and the tangential vector is important.

Corollary 4.8. *Let $T \subset \mathbb{R}^d$, such that $T = F_T(\hat{T})$ with Jacobian J_T. By $\boldsymbol{n}_T$,$\boldsymbol{\tau}_T$ and $\boldsymbol{n}_{\hat{T}}$,$\boldsymbol{\tau}_{\hat{T}}$ the outer normal vector and the tangential vector for the elements T and $\hat{T}$, respectively, are denoted. Then the outer normal and the tangential vector transform as*

$$\boldsymbol{n}_T \circ F_T = \frac{J_T^{-\top} \boldsymbol{n}_{\hat{T}}}{\| J_T^{-\top} \boldsymbol{n}_{\hat{T}} \|}, \qquad \text{and} \qquad \boldsymbol{\tau}_T \circ F_T = \frac{J_T \boldsymbol{\tau}_{\hat{T}}}{\| J_T \boldsymbol{\tau}_{\hat{T}} \|}.$$

We define the size h_T of an element T as the diameter of the smallest circle containing T, and the mesh size h is given by

$$h := \sup_{T \in \mathcal{T}} h_T.$$

Based on this, uniform and quasi-uniform triangulations can be introduced.

Definition 4.9. *Let $\{\mathcal{T}_h\}$ denote a family of triangulations.*

(1) A triangulation $\mathcal{T}_h$ is called quasi-uniform if there exists a constant $\kappa > 0$ independent of h, such that each element contains a circle of diameter ρ_T with $\rho_T \geq h_T/\kappa$.

(2) A triangulation $\mathcal{T}_h$ is called uniform if there exists a constant $\kappa > 0$ independent of h, such that each element contains a circle of diameter ρ_T with $\rho_T \geq h/\kappa$.

4.2 Some conforming finite elements

If we want to find H^1, $H(\text{curl})$ and $H(\text{div})$ conforming finite element spaces, the degrees of freedom have to be chosen, such that the continuity constraints on element interfaces stated in the Corollaries 3.5, 3.10 and 3.15 are fulfilled. Thus, functions in $H(\text{curl})$ need to be tangential continuous across element interface while normal continuity is needed for functions in $H(\text{div})$.

4.2.1 An H^1 conforming finite element

Finite elements for $Q = H^1(\Omega)$ can be constructed by functionals Σ_T^Q which are point evaluations with points placed regularly on the element. Because of their bad stability properties for high polynomial orders, they are not used frequently. We will stick to hierarchical finite elements introduced in [AC03, Zag06]. In the last reference the polynomial order can be chosen separately for edge face and cell basis functions. There, the H^1-conforming finite element $(T, W_{T,k}^Q, \Sigma_{T,k}^Q)$ of uniform polynomial order k is defined as

- T is a tetrahedron or triangle,
- $W_{T,k}^Q = P^k(T)$ with the polynomial space $P^k(T)$ of order k,
- $\Sigma_{T,k}^Q = \Sigma_V^Q \cup \Sigma_{E,k}^Q \cup \Sigma_{F,k}^Q \cup \Sigma_{C,k}^Q$ with
  - the vertex degrees of freedom

$$\Sigma_V^Q(p) = \Big\{ p(V) \ : \ \text{for all } V \in \mathcal{V}_T \Big\},$$

  - the edge degrees of freedom

$$\Sigma_{E,k}^Q(p) = \Big\{ \int_E \frac{\partial p}{\partial s} \frac{\partial v_i}{\partial s} ds \ : \ \{v_i\} \text{ a basis of } P_0^k(E),\ \forall E \in \mathcal{E}_T \Big\},$$

  - the face degrees of freedom (only in 3D)

$$\Sigma_{F,k}^Q(p) = \Big\{ \int_F \nabla_F p \cdot \nabla_F v_i dA \ : \ \{v_i\} \text{ a basis of } P_0^k(F),\ \forall F \in \mathcal{F}_T \Big\}$$

with $\nabla_F v := \boldsymbol{n} \times (\nabla v)|_F \times \boldsymbol{n}$,

- and the cell degrees of freedom

$$\Sigma^Q_{C,k}(p) = \Big\{ \int_T \nabla p \cdot \nabla v_i d\boldsymbol{x} \;:\; v_i \text{ a basis of } P^k_0(T) \Big\}.$$

This finite element is unisolvent (see [Mon03]) and in [Zag06] a polynomial basis is constructed explicitly.

As already mentioned in the last section, it is convenient to define the shape functions for a reference element. In order to do this, a conforming transformation from the reference element to the physical element is needed, which preserves the degrees of freedom.

Lemma 4.10 (H^1-conforming transformation)**.** *Let $T \subset \mathbb{R}^d$, such that $T = F_T(\hat{T})$ with Jacobian matrix J_T. By* $\operatorname{grad}_{\boldsymbol{x}}$ *and* $\operatorname{grad}_{\hat{\boldsymbol{x}}}$ *the gradient with respect to the coordinate system of T and $\hat{T}$, respectively, is denoted. Then for $u_{\hat{T}} \in H^1(\hat{T})$ the transformation*

$$u := u_{\hat{T}} \circ F_T^{-1}$$

implies $u \in H^1(T)$ and

$$\operatorname{grad}_{\boldsymbol{x}} u = J_T^{-1} \big(\operatorname{grad}_{\hat{\boldsymbol{x}}} u_{\hat{T}} \big) \circ F_T^{-1}.$$

By identifying the local degrees of freedom $\Sigma^Q_{T,k}$ with global ones, thus, $\Sigma^Q_k = \bigcup_{T \in \mathcal{T}} \Sigma^Q_{T,k}$, we get the global finite element space of order k,

$$Q_{h,k} = \prod_{T \in \mathcal{T}} W^Q_{T,k}$$

where h indicates the largest element diameter of the triangulation $\mathcal{T}$. If we denote the functionals of Σ^Q_k as S^Q_j and the corresponding nodal basis functions as ϕ^Q_j, we are able to define the interpolation operator for any function $u \in H^{\frac{3}{2}+\delta}$ via

$$\Pi^Q_k u = \sum_j S^Q_j(u) \phi^Q_j.$$

For a sufficiently smooth function $u \in H^s(\Omega)$, $\frac{3}{2} + \delta \leq s \leq k + 1$ the following interpolation error estimate holds.

Theorem 4.11. *(see [Mon03], Theorem 5.48) Let $\mathcal{T}$ denote a quasi-uniform triangulation with mesh size h, then there exist constants c_1, c_2 independent of h and u, such that for*

$\frac{3}{2} + \delta \leq s \leq k + 1$

$$\begin{aligned} \left\|u - \Pi_k^Q u\right\|_{H^1(\Omega)} &\leq c_1 h^{s-1} \|u\|_{H^s(\Omega)} \qquad (4.2) \\ \left\|u - \Pi_k^O u\right\|_{L^2(\Omega)} &\leq c_2 h^s \|u\|_{H^s(\Omega)}. \qquad (4.3) \end{aligned}$$

4.2.2 An $H(\mathrm{curl})$ conforming finite element

In order to obtain $H(\mathrm{curl})$ conformity, tangential continuity of the basis functions across element interfaces is required. Finite elements, fulfilling this condition are sometimes called edge elements because their lowest order degrees of freedom are related to the element edges. Two families of these elements were first introduced by Nedelec [N8́0] and in [N8́6] with a uniform polynomial order on the whole mesh. Based on this work, in [SZ05, Zag06] finite elements with variable polynomial order were constructed. In this work we used Nedelec elements of second kind, $(T, W^Y_{T,k}, \Sigma^Y_{T,k})$ with constant polynomial order k, which are given as

- T is a tetrahedron,
- $W^Y_{T,k} = \mathcal{ND}^{II}_k := \left(P^k(T)\right)^3$ with the polynomial space $P^k(T)$ of order k,
- $\Sigma^Y_{T,k} = \Sigma^Y_{E,k} \cup \Sigma^Y_{F,k} \cup \tilde{\Sigma}^Y_{F,k} \cup \Sigma^Y_{C,k} \cup \tilde{\Sigma}^Y_{C,k}$ with
  - the edge degrees of freedom

    $$\Sigma^Y_{E,k}(\boldsymbol{u}) = \left\{ \int_E \boldsymbol{u} \cdot \boldsymbol{\tau} v_i ds \ : \ \{v_i\} \text{ a basis of } P^k(E),\ \forall E \in \mathcal{E}_T \right\}$$

    where $\boldsymbol{\tau}$ is the tangential vector to the edge E,
  - the face degrees of freedom

    $$\begin{aligned} \Sigma^Y_{F,k}(\boldsymbol{u}) &= \left\{ \int_F \mathrm{curl}_F\, \boldsymbol{u}\, \mathrm{curl}_F\, \boldsymbol{v}_i dA \ : \ \{\mathrm{curl}_F\, \boldsymbol{v}_i\} \text{ a basis of } \right. \\ &\qquad \left. \mathrm{curl}_F \left(P^k_{0,\boldsymbol{\tau}}(F)\right),\ \forall F \in \mathcal{F}_T \right\}, \\ \tilde{\Sigma}^Y_{F,k}(\boldsymbol{u}) &= \left\{ \int_F \boldsymbol{u} \cdot \boldsymbol{v}_i dA \ : \ \{\boldsymbol{v}_i\} \text{ a basis of } \nabla_F\left(P_0^{k+1}(F)\right),\ \forall F \in \mathcal{F}_T \right\} \end{aligned}$$

    where $\mathrm{curl}_F\, \boldsymbol{v} = \mathrm{curl}\, \boldsymbol{v} \cdot \boldsymbol{n}$ and $P^k_{0,\boldsymbol{\tau}}(K) = \left(P^k(K)\right)^3 \cap H_0(\mathrm{curl}, K)$,

- the cell degrees of freedom

$$\begin{aligned}\Sigma^Y_{C,k}(\boldsymbol{u}) &= \Big\{\int_T \operatorname{curl}\boldsymbol{u}\cdot\operatorname{curl}\boldsymbol{v}_i d\boldsymbol{x} \;:\; \{\operatorname{curl}\boldsymbol{v}_i\} \text{ a basis of } \operatorname{curl}\big(P^k_{0,\boldsymbol{\tau}}(T)\big)\Big\},\\ \tilde\Sigma^Y_{C,k}(\boldsymbol{u}) &= \Big\{\int_T \boldsymbol{u}\cdot\boldsymbol{v}_i d\boldsymbol{x} \;:\; \{\boldsymbol{v}_i\} \text{ a basis of } \nabla\big(P^{k+1}_0(T)\big)\Big\}.\end{aligned}$$

A set of hierarchical basis functions for this finite element can be found in [SZ05, Zag06].

For defining the shape functions on a reference element, a transformation preserving the degrees of freedom is needed.

Lemma 4.12 ($H(\text{curl})$-conforming transformation)**.** *Let $T \subset \mathbb{R}^3$, such that $T = F_T(\hat T)$ with Jacobian matrix J_T. By $\operatorname{curl}_{\boldsymbol{x}}$ and $\operatorname{curl}_{\hat{\boldsymbol{x}}}$ the curl with respect to the coordinate system of T and $\hat T$, respectively, is denoted. Then, for $\boldsymbol{q}_{\hat T} \in H(\text{curl}, \hat T)$ the transformation*

$$\boldsymbol{q} := J_T^{-\top}\boldsymbol{q}_{\hat T} \circ F_T^{-1}$$

implies $\boldsymbol{q} \in H(\text{curl}, T)$ and

$$\operatorname{curl}_{\boldsymbol{x}}\boldsymbol{q} = \frac{1}{\det J_T} J_T\big(\operatorname{curl}_{\hat{\boldsymbol{x}}}\boldsymbol{q}_{\hat T}\big) \circ F_T^{-1}.$$

The transformation is also known as covariant transformation.
For a proof see section 3.9 in [Mon03]

Based on this finite element, the global functional space Σ^Y_k and the global space Y_{hk} of order k with mesh size h approximating $Y := H(\text{curl}, \Omega)$ read as

$$\Sigma^Y_k = \bigcup_{T\in\mathcal{T}} \Sigma^Y_{T,k}, \qquad Y_{hk} := \prod_{T\in\mathcal{T}} W^Y_{T,k}.$$

By defining for functionals S^Y_j of Σ^Y_k and the nodal basis functions ϕ^Y_j the interpolation operator

$$\Pi^Y_k \boldsymbol{v} := \sum_j S^Y_j(\boldsymbol{v})\phi^Y_j,$$

we get the following interpolation error result.

Theorem 4.13. *(see [Mon03] Theorem 8.15) Let $\mathcal{T}$ denote a quasi-uniform triangulation with mesh size h. Then for $\boldsymbol{u} \in \big(H^{s+1}(\Omega)\big)^3$, $1 \le s \le k$, there exists a constant c*

independent of h and $\boldsymbol{u}$, such that

$$\left\|\boldsymbol{u}-\Pi_k^Y\boldsymbol{u}\right\|_{L^2(\Omega)}+h\left\|\operatorname{curl}(\boldsymbol{u}-\Pi_k^Y\boldsymbol{u})\right\|_{L^2(\Omega)}\leq ch^{s+1}\|\boldsymbol{u}\|_{(H^{s+1}(\Omega))^3}. \tag{4.4}$$

4.2.3 An $H(\mathrm{div})$ conforming finite element

As already discussed, a $H(\mathrm{div})$ conforming finite element space has to contain functions which are normal continuous across element interfaces. There are two widely used finite element families which fulfill this property, the Raviart-Thomas element [RT77] and the Brezzi-Douglas-Marini element [BDM85]. For both elements, the lowest order degrees of freedom are related to facets. Thus, they are sometimes called face elements. We will restrict ourself to the Raviart Thomas element. Such elements of varying polynomial order together with an hierarchical basis can be found in [Zag06].
The Raviart-Thomas element $(T, W_{T,k}^V, \Sigma_{T,k}^V)$ of constant polynomial order k reads as

- T is a tetrahedron or triangle
- $W_{T,k}^V=\mathcal{RT}_k:=\left\{\boldsymbol{v}\in(P^k)^d\ :\ (\boldsymbol{v}\cdot\boldsymbol{n})|_F\in P^{k-1}(F)\ \ \forall F\in\mathcal{F}_T\right\}$,
- $\Sigma_{T,k}^V=\Sigma_{F,k}^V\cup\Sigma_{C,k}^V\cup\tilde{\Sigma}_{C,k}^V$ with
    - the face degrees of freedom

$$\Sigma_{F,k}^V(\boldsymbol{u})=\left\{\int_F\boldsymbol{u}\cdot\boldsymbol{n}v_i dA\ :\ \{v_i\}\text{ a basis of }P^{k-1}(F),\ \forall F\in\mathcal{F}_T\right\},$$

    - the cell degrees of freedom

$$\begin{aligned}\Sigma_{C,k}^V(\boldsymbol{u}) &= \left\{\int_T\operatorname{div}\boldsymbol{u}\operatorname{div}\boldsymbol{v}_i d\boldsymbol{x}\ :\ \{\operatorname{div}\boldsymbol{v}_i\}\text{ a basis of }\operatorname{div}\left(P_{0,\boldsymbol{n}}^k(T)\right)\right\},\\ \tilde{\Sigma}_{C,k}^V(\boldsymbol{u}) &= \left\{\int_T\boldsymbol{u}\cdot\boldsymbol{v}_i\,d\boldsymbol{x}\ :\ \{\boldsymbol{v}_i\}\text{ a basis of }\operatorname{curl}\left(P_{0,\boldsymbol{\tau}}^{k+1}(T)\right)\right\},\end{aligned}$$

where $P_{0,\boldsymbol{n}}^p(K):=\left(P^p(K)\right)^3\cap H_0(\mathrm{div},K)$ and $P_{0,\boldsymbol{\tau}}^p(K)=\left(P^p(K)\right)^3\cap H_0(\mathrm{curl},K)$.

The following transformation preserves these degrees of freedom, and it allows us to define the shape functions on a reference element.

Lemma 4.14 ($H(\mathrm{div})$-conforming transformation)**.** *Let $T \subset \mathbb{R}^3$, such that $T = F_T(\hat{T})$ with Jacobian matrix J_T. By $\mathrm{div}_{\boldsymbol{x}}$ and $\mathrm{div}_{\hat{\boldsymbol{x}}}$ the divergence with respect to the coordinate system of T and $\hat{T}$, respectively, is denoted. Then for $\boldsymbol{p}_{\hat{T}} \in H(\mathrm{div}, \hat{T})$ the transformation*

$$\boldsymbol{p} := \frac{1}{\det J_T} \, J_T \, \boldsymbol{p}_{\hat{T}} \circ F_T^{-1}$$

implies $\boldsymbol{p} \in H(\mathrm{div}, T)$ and

$$\mathrm{div}_{\boldsymbol{x}} \, \boldsymbol{p} = \frac{1}{\det J_T} \left(\mathrm{div}_{\hat{\boldsymbol{x}}} \, \boldsymbol{p}_{\hat{T}} \right) \circ F_T^{-1}.$$

The transformation is also known as contravariant or Piola transformation.
For a proof see section 3.9 in [Mon03]

Like for the other spaces, we are able to define the global functional space Σ_k^V and the global finite element space V_{hk} of order k approximating $V := H(\mathrm{div}, \Omega)$ via

$$\Sigma_k^V = \bigcup_{T \in \mathcal{T}} \Sigma_{T,k}^V, \qquad V_{hk} = \prod_{T \in \mathcal{T}} W_{T,k}^V.$$

Note that the Raviart Thomas space V_{hk} of order k contains the element wise polynomial space $\left(P^{k-1}\right)^d$, but it equals not the element wise polynomial space $\left(P^k\right)^d$. The divergence of a function in V_{hk} is in $P^{k-1}(T)$ on the element, and its normal traces on the facets are in $P^{k-1}(F)$. When defining the interpolation operator for functions $\boldsymbol{u} \in \left(H^{\frac{1}{2}+\delta}(\Omega)\right)^3$,

$$\Pi_k^V \boldsymbol{u} = \sum_j S_j^V(\boldsymbol{u}) \phi_j^V,$$

with the functionals S_j^V from Σ_k^V and the corresponding nodal basis ϕ_j^V, we get the following interpolation error estimates.

Theorem 4.15. *(see [Mon03] Theorem 5.25, Remark 5.26) Let $\mathcal{T}$ denote a quasi-uniform triangulation with mesh size h. Then for $\boldsymbol{u} \in \left(H^s(\Omega)\right)^3$, $\frac{1}{2} + \delta \leq s \leq k$, there exist constants c_1, c_2 independent of h and $\boldsymbol{u}$, such that*

$$\begin{aligned} \left\| \boldsymbol{u} - \Pi_k^V \boldsymbol{u} \right\|_{L^2\Omega} &\leq c_1 h^s \|\boldsymbol{u}\|_{H^s(\Omega)}, & (4.5) \\ \left\| \mathrm{div}(\boldsymbol{u} - \Pi_k^V \boldsymbol{u}) \right\|_{L^2(\Omega)} &\leq c_2 h^s \| \mathrm{div}\, \boldsymbol{u} \|_{H^s(\Omega)}. & (4.6) \end{aligned}$$

4.2.4 An L^2 conforming finite element

We finish the discussion on finite elements by introducing an L^2-conforming finite element. In principle, one could use the H^1-conforming finite element and neglect the continuity constraints, i.e. we do not identify the local functionals Σ_T of different elements with each other on the global level. In the following, we present an L^2 conforming element $(T, W^U_{T,k}, \Sigma^U_{T,k})$ obtained by a different approach.

- T is a tetrahedron or triangle,
- $W^U_{T,k} = P^k(T)$ with the polynomial space $P^k(T)$ of order k,
- the cell degrees of freedom

$$\Sigma^U_{T,k}(u) = \Big\{ \int_T u v_i \, d\boldsymbol{x} \; : \; \{v_i\} \text{ a basis of } P^k(T) \Big\}.$$

Based on the finite element, we obtain the global space of functionals and the global finite element space approximating $U = L^2(\Omega)$ as

$$\Sigma^U_k = \bigcup_{T \in \mathcal{T}} \Sigma^U_{T,k}, \qquad U_{kh} = \prod_{T \in \mathcal{T}} W^U_{T,k}.$$

This suggests the interpolation operator for $u \in L^2(\Omega)$

$$\Pi^U_k u = \sum_j S^U_j(u) \phi^U_j$$

where S^U_j are the functionals in Σ^U_k and the ϕ^U_j are the corresponding nodal basis functions. For u sufficiently smooth, i.e. $u \in H^s(\Omega)$ with $1 \le s \le k{+}1$, the interpolation error estimate

$$\left\| u - \Pi^U_k u \right\|_{L^2(\Omega)} \le c h^s \|u\|_{H^s(\Omega)}, \qquad \text{for } 1 \le s \le k+1, \tag{4.7}$$

with c independent of h and u can be shown.

Chapter 5

A hybrid finite element method for the Helmholtz and the vector valued wave equation

As already mentioned in the introduction, a large number of unknowns is necessary to describe the highly oscillating solution of the wave equation. In fact, one needs at least a fixed number of degrees of freedom per wavelength to get a certain accuracy for the solution. Consequently, the total number of degrees of freedom grows for an increasing frequency ω at least proportional to $\mathcal{O}(\omega^d)$ in d dimensions. Additionally, a classical H^1-conforming finite element method suffers from the pollution effect [Ihl98]. This means that the number of unknowns per wavelength, which is necessary to achieve a given accuracy, increases with increasing frequency. Although this effect is not so strong for higher polynomial orders, the total number of unknowns grows faster than $\mathcal{O}(\omega^d)$.

In order to overcome these difficulties, a variety of different approaches within the finite element framework was developed during the last decades. Widely used are *hp* methods [IB97] or discontinuous Galerkin methods [FW09]. They are still based on a polynomial approximation for a modified variational formulation. Other methods, like the partition of unity method [Mel95, MB96], least square methods [MW99], discontinuous enrichment methods [FHH03, TF06] or the ultra weak variational formulation [CD98, Mon03] make use of problem-adapted functions. In this chapter we focus on a mixed hybrid finite element method.

The discussion on hybrid methods starts with an introduction to this topic in Section 5.1. In Section 5.2 the mixed hybrid finite element method for the Helmholtz equation from

[MSS10] is presented. Apart from consistency of the continuous problem and existence and uniqueness of its solution, energy conservation for both, the discrete and the continuous formulation is shown. In Section 5.3 we extend this hybrid formulation to the time harmonic Maxwell case. This approach we already published in [HPS11]. As for the Helmholtz equation, consistency, existence and uniqueness as well as energy conservation results are proven.

5.1 An introduction to hybridization

Before we concentrate on the hybridization of the Helmholtz equation and the vector valued wave equation, we recall the basic concept of hybridization. For a detailed introduction to hybridization see [BF91]. In [AB85] the hybridization of Raviart-Thomas methods is discussed, and a hybridization technique for the biharmonic problem is introduced. Furthermore, [AB85] provides error estimates for the Lagrangian multipliers obtained via estimates for the primal variables. An error analysis just for the Lagrangian multipliers without using the other variables, especially for finite element spaces of different polynomial order, can be found in [CG05a]. In addition to hybridization of Raviart Thomas elements, [CG04] covers hybridization of Brezzi Douglas Marini finite elements.
The starting point for the hybridization of an elliptic boundary value problem is a mixed method with a linear system of equations typically of the form

$$\begin{pmatrix} A & B^* \\ B & C \end{pmatrix} \begin{pmatrix} \underline{\boldsymbol{\sigma}_h} \\ \underline{u_h} \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}$$

where $*$ denotes the adjoint operator, $\underline{\boldsymbol{\sigma}_h}$ is the coefficient vector of the finite element function $\boldsymbol{\sigma}_h$ representing the flux field, and $\underline{u_h}$ is the coefficient vector of the scalar finite element function u_h. In general this mixed system is contrary to the primal problem indefinite, i.e. it forms a saddle point problem, and solving for the unknown coefficients $\underline{\boldsymbol{\sigma}_h}$ and $\underline{u_h}$ is time consuming. Positive definiteness of the primal problem one could regain by eliminating $\underline{\boldsymbol{\sigma}_h}$ which makes an inversion of the block A necessary. Because the flux field $\boldsymbol{\sigma}_h$ fulfills continuity constraints, i.e. for the Poisson equation it is from the space $H(\mathrm{div}, \Omega)$ and therefore normal continuous, the matrix A has entries which couple degrees of freedom belonging to different elements, and therefore the inverse A^{-1} is a full matrix.

This drawback can be compensated by hybridization which was first used in [dV65]. There, the continuity constraints are neglected at element interfaces, for example for the

Poisson equation $\boldsymbol{\sigma}_h$ is only element wise in $H(\mathrm{div})$. Continuity is regained by adding Lagrangian multipliers λ_h which are supported only on element facets as an additional set of unknowns. The resulting system reads as

$$\begin{pmatrix} A & B^* & D^* \\ B & C & 0 \\ D & 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{\boldsymbol{\sigma}}_h \\ \underline{u}_h \\ \underline{\lambda}_h \end{pmatrix} = \begin{pmatrix} F \\ G \\ 0 \end{pmatrix} \tag{5.1}$$

where $\underline{\lambda}_h$ is the coefficient vector of the multiplier λ_h, and the matrix block D represents the continuity conditions. Due to the loss of continuity in the space of $\boldsymbol{\sigma}_h$ the matrix A is block diagonal, and therefore easy to invert. This is the main advantage of this formulation. We are now in the position to express the volume unknowns $\boldsymbol{\sigma}_h$ and u_h in terms of the Lagrangian multipliers λ_h,

$$\begin{pmatrix} \underline{\boldsymbol{\sigma}}_h \\ \underline{u}_h \end{pmatrix} = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}^{-1} \left(\begin{pmatrix} F \\ G \end{pmatrix} - \begin{pmatrix} D^* \\ 0 \end{pmatrix} \underline{\lambda}_h \right) \tag{5.2}$$

with

$$\begin{pmatrix} A & B^* \\ B & C \end{pmatrix}^{-1} = \begin{pmatrix} I_A & -A^{-1}B^* \\ 0 & I_C \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & (C - BA^{-1}B^*)^{-1} \end{pmatrix} \begin{pmatrix} I_A & 0 \\ -BA^{-1} & I_C \end{pmatrix}$$

where I_A and I_C are identity matrices with the dimensions of A and C, respectively. Note that the scalar field u_h is from a discontinuous function space, for example $L^2(\Omega)$ for the Poisson equation. Therefore, $C - BA^{-1}B^*$ is also block diagonal with blocks corresponding to elements. Now, elimination of the volume unknowns, i.e. inserting (5.2) into the third equation of (5.1), leads to a Schur complement system for the multiplier λ_h

$$S\underline{\lambda}_h = H$$

with

$$\begin{aligned} S &= \begin{pmatrix} D & 0 \end{pmatrix} \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}^{-1} \begin{pmatrix} D^* \\ 0 \end{pmatrix} \\ H &= \begin{pmatrix} D & 0 \end{pmatrix} \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}^{-1} \begin{pmatrix} F \\ G \end{pmatrix}. \end{aligned}$$

Due to the fact that for the Poisson equation this system is symmetric positive definite, iterative solvers, like the preconditioned cg-method, can be used for solving it. The volume solutions $\boldsymbol{\sigma}_h$ and u_h are obtained from the multipliers via (5.2). These multipliers often have a physical interpretation. For example in the Poisson equation they represent the value of u_h at the facet. Note that, because of the block structure of A and C, respectively, the computation of $\boldsymbol{\sigma}_h$ and u_h can be done element wise only using information from the element boundary.

In former times this hybridization technique was often interpreted as an implementation trick to deal with mixed methods. But it was realized [AB85, BDM85], that a more accurate solution for the scalar function u in comparison to (5.2) can be obtained by local post processing. Apart from this, hybridization has another advantage. By static condensation (compare [AB85]), i.e. elimination of the volume unknowns, the dimension of the system can be reduced significantly to the number of interface unknowns λ_h.

Besides of Raviart Thomas and Brezzi Douglas Marini methods, hybridization has been used for the Tangential-Displacement-Normal-Normal-Stress formulation in mechanics [Sin09] and for the Stokes system [CG05b, CG05c, CCS06]. In the last one the method is used to construct an exactly divergence free approximation of the velocity without using stream function variables or globally divergence free finite element basis functions. This is done in two and three dimensions. We also mention recent works on hybrid discontinuous Galerkin methods for Maxwell's equations in 2D [LP11] and for 3D problems [NPC11].

5.2 A hybrid finite element method for the Helmholtz equation

Before continuing, we have to introduce some notations. We use a regular finite element mesh $\mathcal{T}$ with elements T and the set of element facets F is denoted by $\mathcal{F}$. The set of all inner facets is named $\mathcal{F}_I$. The vector $\boldsymbol{n}_T$ is the outer normal vector of an element T, $\boldsymbol{n}_F$ represents the normal vector onto a facet F, and, as already mentioned, $\boldsymbol{n}_\Gamma$ is the outer normal of the domain Ω with boundary Γ.

5.2.1 Hybridizing the Helmholtz equation

When hybridizing the mixed Helmholtz equation, we start with the mixed form of the Helmholtz equation (2.30) and (2.31) and multiply it with test functions $v \in U := L^2(\Omega)$

and $\boldsymbol{\tau} \in V := H(\operatorname{div}, \Omega)$, respectively. Carrying out integration by parts element by element leads to

$$\begin{aligned}
&\sum_{T\in\mathcal{T}} \Big[\quad \big(i\omega\epsilon u, v\big)_T - (\operatorname{div}\boldsymbol{\sigma}, v)_T\Big] && = 0 && \forall v \in V\\
&\sum_{T\in\mathcal{T}} \Big[- \big(u, \operatorname{div}\boldsymbol{\tau}\big)_T - \big(i\omega\mu\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_T + \big\langle u, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T}\Big] && = 0 && \forall u \in U.
\end{aligned}$$

For inner facets the test function $\boldsymbol{\tau}$ is normal continuous, and thus, the boundary integral for one element cancels out with the integral from a neighboring one which would give (3.5).

Next, the normal continuity of the flux function $\boldsymbol{\sigma}$ across element interfaces is broken, i.e. the space $H(\operatorname{div}, \Omega)$ is replaced by a broken $H(\operatorname{div})$ space

$$\widetilde{V} := \Big\{\boldsymbol{\tau} \in \big(L^2(\Omega)\big)^d \ : \ \boldsymbol{\tau}|_T \in H(\operatorname{div}, T)\ \forall T \in \mathcal{T}\Big\}.$$

In order to reinforce continuity again, we introduce, as mentioned in the last section, a Lagrangian multiplier u_F from the space

$$u_F \in U_F := L^2(\mathcal{F}).$$

The normal continuity of the flux $\boldsymbol{\sigma}$ is recovered via an additional equation which forces the jump $\big[\boldsymbol{\sigma}\cdot\boldsymbol{n}\big]_F := \boldsymbol{\sigma}\cdot\boldsymbol{n}_{T_1} + \boldsymbol{\sigma}\cdot\boldsymbol{n}_{T_2}$ for inner facets $F \in \mathcal{F}_I$ with the two adjacent elements T_1 and T_2 to zero. This leads to

$$\sum_{F\in\mathcal{F}_I} \big\langle[\boldsymbol{\sigma}\cdot n, v_F]\big\rangle_F = \sum_{T\in\mathcal{T}} \big(\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, v_F\rangle_{\partial T} - \big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, v_F\big\rangle_{\partial T\cap\Gamma}\big) = 0 \qquad \forall v_F \in V_F.$$

Inserting the boundary condition into this equation results in the system of equations for $(u, \boldsymbol{\sigma}, u_F) \in U \times \widetilde{V} \times U_F$

$$\begin{aligned}
&\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon u, v\big)_T - \big(\operatorname{div}\boldsymbol{\sigma}, v\big)_T\Big] && && = 0 && \forall v \in U,\\
&\sum_{T\in\mathcal{T}} \Big[- \big(u, \operatorname{div}\boldsymbol{\tau}\big)_T - \big(i\omega\mu\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_T + \big\langle u_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T}\Big] && && = 0 && \forall \boldsymbol{\tau} \in \widetilde{V},\\
&\sum_{T\in\mathcal{T}} \Big[\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, v_F\big\rangle_{\partial T}\Big] && - \big\langle\sqrt{\frac{\epsilon}{\mu}}u_F, v_F\big\rangle_\Gamma && = \big\langle\sqrt{\frac{\epsilon}{\mu}}g, v_F\big\rangle_\Gamma && \forall v_F \in U_F.
\end{aligned}$$

Note that in the second equation the Lagrangian multiplier plays the role of the scalar field u evaluated on the element interfaces. This was used in the boundary condition, and u was exchanged by u_F in order to regain symmetry. Due to the fact that there is no coupling between volume basis functions belonging to two different elements, it is possible to eliminate the volume unknowns u and $\boldsymbol{\sigma}$ element by element via static condensation (compare [AB85]). The resulting system of equations has to be solved only for the multiplier u_F. In order to eliminate these inner degrees of freedom, the first two equations of our system need to be solved for any u_F on each element T. This is equivalent to solve a Dirichlet problem consisting of (3.4) and (3.5) for $\Omega = T$ and $u = u_F$ as boundary condition. Hence, uniqueness of the solution is lost if ω matches an eigenvalue of the Dirichlet problem.

Uniqueness can be according to [MSS10] re-established by adding a new facet unknown

$$\sigma_F \in V_F := L^2(\mathcal{F}),$$

representing $\boldsymbol{\sigma} \cdot \boldsymbol{n}_F$ on the facet F via an additional equation

$$\sum_{T\in\mathcal{T}} \beta\big\langle \sigma_F - \boldsymbol{\sigma}\cdot\boldsymbol{n}_F,\ \tau_F\big\rangle_{\partial T} = 0 \qquad \forall \tau_F \in V_F$$

with some stabilization parameter β. For symmetry reasons the consistent term $\sum_{T\in\mathcal{T}} \beta\big\langle \boldsymbol{\sigma}\cdot \boldsymbol{n}_F - \sigma_F,\ \boldsymbol{\tau}\cdot\boldsymbol{n}_F\big\rangle_{\partial T}$ is added to our formulation. Collecting these terms, we get for all $(v, \boldsymbol{\tau}, v_F, \tau_F) \in U \times \widetilde{V} \times U_F \times V_F$

$$\begin{aligned}
&\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon u, v\big)_T - \big(\operatorname{div}\boldsymbol{\sigma}, v\big)_T\Big] && = 0,\\
&\sum_{T\in\mathcal{T}} \Big[-\big(u, \operatorname{div}\boldsymbol{\tau}\big)_T - \big(i\omega\mu\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_T + \beta\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T} \\
&\qquad\qquad + \big\langle u_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T} - \beta\big\langle\sigma_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_F\big\rangle_{\partial T}\Big] && = 0,\\
&\sum_{T\in\mathcal{T}} \Big[\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, v_F\big\rangle_{\partial T}\Big] \qquad - \big\langle\sqrt{\frac{\epsilon}{\mu}}u_F, v_F\big\rangle_\Gamma && = -\big\langle\sqrt{\frac{\epsilon}{\mu}}g, v_F\big\rangle_\Gamma,\\
&\sum_{T\in\mathcal{T}} \Big[\beta\big\langle\sigma_F, \tau_F\big\rangle_{\partial T} - \beta\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_F, \tau_F\big\rangle_{\partial T}\Big] && = 0.
\end{aligned}$$

Now, static condensation reduces the system of equations to the facet degrees of freedom u_F and σ_F. Elimination of the inner degrees of freedom, i.e. solving the first two equations on the element level for any σ_F and u_F is now, for $\beta = \sqrt{\frac{\mu}{\epsilon}}$, equivalent to solve Formulation 3.23 with $\Omega = T$ and $g = -\sqrt{\frac{\mu}{\epsilon}}(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F + u_F$. Taking into account that σ_F plays the role of $\boldsymbol{\sigma} \cdot \boldsymbol{n}_F$ and u_F of u on ∂T, we can redefine in- and outgoing impedance traces in terms of the facet variables.

Definition 5.1. *Let T be an element of the triangulation $\mathcal{T}$, then the in- and outgoing impedance traces in terms of the facet variables on the boundary ∂T are defined via*

$$\begin{aligned} In_{\partial T} &:= -\sqrt{\frac{\mu}{\epsilon}}(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F + u_F \\ Out_{\partial T} &:= \sqrt{\frac{\mu}{\epsilon}}(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F + u_F. \end{aligned}$$

In the same way In_Γ and Out_Γ can be defined.

According to this definition, we solve on the element level the mixed Helmholtz problem for $g = In_{\partial T}$, i.e. with a prescribed incoming impedance trace for the element T. Now, uniqueness and existence of the element solution is guaranteed. By exchanging the Dirichlet and Neumann traces u_F, σ_F with the incoming and outgoing impedance traces, one obtains an equivalent formulation which fits well into the context of the ultra weak variational formulation of [CD98]. From this point of view, static condensation can be interpreted as calculating in the case of $\beta = \sqrt{\frac{\mu}{\epsilon}}$ for a given incoming impedance trace the corresponding outgoing impedance trace. Note that the outgoing impedance trace of an element is the incoming impedance trace of the neighboring element, and the incoming impedance trace on the domain Ω is fixed by the boundary condition.

We will show later on that this hybrid formulation is energy conserving. Because, at least, when using a finite element discretization, a fixed number of degrees of freedom is needed to resolve the solution of the Helmholtz equation (compare [Ain04]), high frequency parts of the solution, which are not resolved by the mesh together with the polynomial order, cause spurious modes. These spurious modes can be damped out by a stabilization term with the parameter α,

$$\sum_{T \in \mathcal{T}} \alpha \langle u - u_F, v - v_F \rangle_{\partial T}$$

which penalizes the jump between the solution u and its corresponding facet value u_F.

Summarizing this, we end up with

Formulation 5.2 (the mixed hybrid formulation for the Helmholtz equation)**.** *Find $\tilde{u} :=$*

$(u, \boldsymbol{\sigma}, u_F, \sigma_F) \in U \times \widetilde{V} \times U_F \times V_F$ such that

$$B_{s\Omega}(\tilde{u}, \tilde{v}) + B_{s\Gamma}(\tilde{u}, \tilde{v}) = F_s(\tilde{v}) \tag{5.3}$$

for all $\tilde{v} := (v, \boldsymbol{\tau}, v_F, \tau_F) \in U \times \widetilde{V} \times U_F \times V_F$, with the bilinear forms

$$\begin{aligned}
B_{s\Omega}(\tilde{u}, \tilde{v}) \;:=\; & \sum_{T \in \mathcal{T}} \Big[\big(i\omega\epsilon u, v\big)_T - \big(\operatorname{div} \boldsymbol{\sigma}, v\big)_T - \big(u, \operatorname{div} \boldsymbol{\tau}\big)_T - \big(i\omega\mu\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_T \\
& \quad + \big\langle u_F, \boldsymbol{\tau} \cdot \boldsymbol{n}_T \big\rangle_{\partial T} + \big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_T, v_F \big\rangle_{\partial T} \\
& \quad + \beta \big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_T - (\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F, \boldsymbol{\tau} \cdot \boldsymbol{n}_T - (\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\tau_F \big\rangle_{\partial T} \\
& \quad + \alpha \big\langle u - u_F, v - v_F \big\rangle_{\partial T} \Big], \\
B_{s\Gamma}(\tilde{u}, \tilde{v}) \;:=\; & -\Big\langle \sqrt{\frac{\epsilon}{\mu}} u_F, v_F \Big\rangle_\Gamma,
\end{aligned}$$

and the linear form

$$F_s(\tilde{v}) := -\Big\langle \sqrt{\frac{\epsilon}{\mu}} g, v_F \Big\rangle_\Gamma.$$

5.2.2 Consistency, existence and uniqueness of the hybrid formulation

Next, we show consistency of our formulation. In [MSS10] and [HHS10] it was already proven that the solution of this formulation is equivalent to the solution of the mixed Helmholtz equations (2.30) and (2.31), i.e.

Lemma 5.3. *Let $(u^e, \boldsymbol{\sigma}^e)$ be the exact solution of (2.30)-(2.32), and let $u_F^e = u^e$ as well as $\sigma_F^e = \boldsymbol{\sigma}^e \cdot \boldsymbol{n}_F$ on the facets F. Then $\tilde{u}^e := (u^e, \boldsymbol{\sigma}^e, u_F^e, \sigma_F^e)$ solves Formulation 5.2.*

Proof. Inserting $\tilde{u}^e$ into Formulation 5.2 and making use of the boundary condition (2.32) results in

$$\begin{aligned}
& \sum_{T \in \mathcal{T}} \Big[\big(i\omega\epsilon u^e, v\big)_T - \big(\operatorname{div} \boldsymbol{\sigma}^e, v\big)_T - \big(u^e, \operatorname{div} \boldsymbol{\tau}\big)_T - \big(i\omega\mu\boldsymbol{\sigma}^e, \boldsymbol{\tau}\big)_T \\
& \qquad + \big\langle u_F^e, \boldsymbol{\tau} \cdot \boldsymbol{n}_T \big\rangle_{\partial T} + \big\langle \boldsymbol{\sigma}^e \cdot \boldsymbol{n}_T, v_F \big\rangle_{\partial T} \Big] - \big\langle \boldsymbol{\sigma}^e \cdot \boldsymbol{n}_\Gamma, v_F \big\rangle_\Gamma = 0.
\end{aligned}$$

Because the solution of the strong problem $\boldsymbol{\sigma}^e$ is normal continuous, the term $\big\langle \boldsymbol{\sigma}^e \cdot \boldsymbol{n}_T, v_F \big\rangle_{\partial T}$ cancels for inner edges when summing up over all elements. On boundary edges $\boldsymbol{n}_T = \boldsymbol{n}_\Gamma$

and the term cancels with the boundary integral. Exchanging u_F^e by u^e yields

$$\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon u^e, v\big)_T - \big(\operatorname{div}\boldsymbol{\sigma}^e, v\big)_T \\ - \big(i\omega\mu\boldsymbol{\sigma}^e, \boldsymbol{\tau}\big)_T - \big(u^e, \operatorname{div}\boldsymbol{\tau}\big)_T + \big\langle u^e, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T} \Big] = 0.$$

Partial integration gives

$$\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon u^e - \operatorname{div}\boldsymbol{\sigma}^e, v\big)_T - \big(i\omega\mu\boldsymbol{\sigma}^e - \operatorname{grad} u^e, \boldsymbol{\tau}\big)_T \Big] = 0$$

which completes the proof. □

The next lemma states existence and uniqueness for a solution of Formulation 5.2.

Lemma 5.4. *Let $\beta \neq 0$, $\alpha = 0$, μ be constant and real, ϵ is a real and piecewise continuous function with $0 < \epsilon_{min} \leq \epsilon \leq \epsilon_{max}$ and $g \in L^2(\Gamma)$. Then there exists a unique solution of Formulation 5.2.*

Proof. Lets consider an arbitrary inner facet $F \in \mathcal{F}_I$. We start the proof by testing (5.3) with any $v_F \in L^2(\mathcal{F})$ such that v_F is just nonzero on F. If T_1 and T_2 are the two adjacent elements of F, we get $\boldsymbol{\sigma}\cdot\boldsymbol{n}_T|_{T_1} + \boldsymbol{\sigma}\cdot\boldsymbol{n}_T|_{T_2} = 0$, and because of the different signs of $\boldsymbol{n}_T$ the function $\boldsymbol{\sigma}$ is normal continuous across element interfaces. Thus the space $\widetilde{V}$ in the variational formulation can be exchanged by $V = H(\operatorname{div}, \Omega)$.

Testing now (5.3) with any $\tau_F \in L^2(\mathcal{F})$ such that τ_F is just nonzero on F leads to $2\sigma_F = \boldsymbol{\sigma}\cdot\boldsymbol{n}_F|_{T_1} + \boldsymbol{\sigma}\cdot\boldsymbol{n}_F|_{T_2}$, and because of normal continuity $\sigma_F = \boldsymbol{\sigma}\cdot\boldsymbol{n}_F$ follows. For boundary facets, this can be obtained directly by testing with τ_F supported just on the boundary.

Consequently, the β term in $B_{s\Omega}$ vanishes, and $\sum_{T\in\mathcal{T}} \big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, v_F\big\rangle_{\partial T}$ as well as $\sum_{T\in\mathcal{T}} \big\langle u_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T}$ simplify to $\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma, v_F\big\rangle_\Gamma$ and $\big\langle u_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_\Gamma\big\rangle_\Gamma$, respectively. Collecting everything gives

$$
\begin{aligned}
\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon u, v\big)_T - \big(\operatorname{div}\boldsymbol{\sigma}, v\big)_T \Big] &= 0 && \forall v \in U,\\
\sum_{T\in\mathcal{T}} \Big[-\big(u, \operatorname{div}\boldsymbol{\tau}\big)_T - \big(i\omega\mu\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_T \Big] + \big\langle u_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_\Gamma \big\rangle_\Gamma &= 0 && \forall \boldsymbol{\tau} \in V,\\
\big\langle \boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma, v_F \big\rangle_\Gamma - \big\langle \sqrt{\frac{\epsilon}{\mu}} u_F, v_F \big\rangle_\Gamma &= -\big\langle \sqrt{\frac{\epsilon}{\mu}} g, v_F \big\rangle_\Gamma && \forall u_F \in U_F.
\end{aligned}
$$

From the last equation it follows that $u_F = \sqrt{\mu/\epsilon}\boldsymbol{\sigma}\cdot\boldsymbol{n}_\Gamma + g$. Inserting this into the second equation yields the standard mixed formulation 3.23, and Lemma 3.24 guarantees existence and uniqueness. □

5.2.3 The discrete finite element spaces

To discretize Formulation 5.2, we have to define appropriate discrete counterparts $U_{hp}, V_{hp}, U_{F,hp}, V_{F,hp}$ to the function spaces $U, \widetilde{V}, U_F, V_F$. Here, by h we denote the maximal mesh size. In this work, we will use a broken Raviart Thomas space of order $p+1$ for the approximation of the discrete flux field, i.e.

$$V_{hp+1} := \Big\{ \boldsymbol{v} \in \big(L^2(\Omega)\big)^d \ : \ \boldsymbol{v}|_T \in \mathcal{RT}_{p+1}(T) \text{ for all } T \in \mathcal{T} \Big\}.$$

Thus, the normal continuity is broken across element interfaces. By testing Formulation 5.2 with a test function $(v, \mathbf{0}, 0, 0)$ it becomes obvious that the quantity $\operatorname{div}\boldsymbol{\sigma}$ with $\boldsymbol{\sigma} \in \widetilde{V}$ should be approximated by $u \in U$. Because the divergence of a function in $\mathcal{RT}_{p+1}$ is a polynomial of order p, we define

$$U_{hp} := \Big\{ u \in L^2(\Omega) \ : \ u|_T \in P^p(\Omega) \text{ for all } T \in \mathcal{T} \Big\}.$$

Taking into account that $u_F \in U_F$ and $\tau_F \in V_F$ play the role of u and $\boldsymbol{\sigma}\cdot\boldsymbol{n}_F$ at the element interfaces together with the fact that the normal trace of a function from $\mathcal{RT}_{p+1}$ is a polynomial of order p motivates the definitions

$$U_{F,hp} = V_{F,hp} := \Big\{ u \in L^2(\mathcal{F}) \ : \ u|_F \in P^p(F) \text{ for all } F \in \mathcal{F} \Big\}.$$

We give an example of the approximation properties of the mixed hybrid finite element

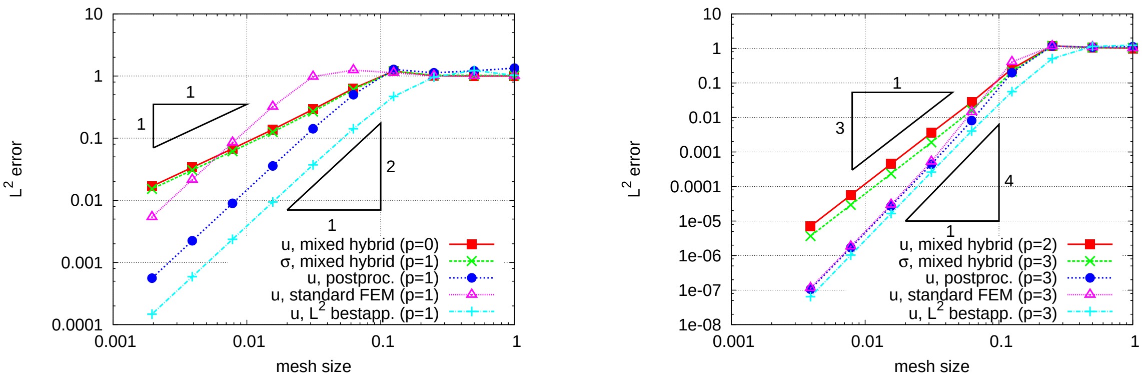

Figure 5.1: The L^2-error of the mixed hybrid solution (red and green) and of a post processed solution (blue) compared to the error of a standard finite elemental solution (magenta) and the L^2 best approximation for the low order case (left) and $p = 2, 3$ (right)

method with a two dimensional problem on the unit square. There, the right hand side of the equation, i.e. the function g, was chosen such that the solution u of the Helmholtz equation is a plane wave with wave length 0.2.

In Figure 5.1 the L^2-error of the approximations u_h for u (red) and $\boldsymbol{\sigma}_h$ for the flux field $\boldsymbol{\sigma}$ (green), obtained by the mixed hybrid finite element method together with the choice of spaces from above, is plotted against the mesh size h. For the left hand plot $p = 0$, i.e. $u_h \in U_{h0}$, $\boldsymbol{\sigma}_n \in V_{h1}$, $u_{Fh} \in U_{F,h0}$, $v_{Fh} \in V_{F,h0}$, was used, while $p = 2$ was taken for the right hand plot. From the slopes of the curves in both plots, we can conclude that u_h and $\boldsymbol{\sigma}_h$ converge with order h^{p+1},

$$\left\|u - u_h\right\|_{L^2(\Omega)} \approx c_1 h^{p+1}, \qquad \left\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\right\|_{L^2(\Omega)} \approx c_2 h^{p+1},$$

which goes along with the interpolation results given in (4.7) and (4.5). Thus the convergence rates are optimal.

In addition, the mixed hybrid formulation offers us the possibility, as already mentioned in the introduction, to obtain a more accurate solution by local post processing [AB85, BDM85]. More precisely, the fact, that $\boldsymbol{\sigma} \in V_{hp+1}$, i.e. it is on the element level a polynomial of order $p + 1$, allows us to compute an approximation $\tilde{u}_h \in U_{hp+1}$ element by element.

Therefore, we solve on each element

$$\begin{aligned}
\big(\operatorname{grad}\tilde{u}_h, \operatorname{grad} v\big)_T &= \big(i\omega\mu\boldsymbol{\sigma}_h, \operatorname{grad} v\big)_T && \forall v \in P^{p+1}(T)\\
\int_T \tilde{u}_h \, d\boldsymbol{x} \int_T v \, d\boldsymbol{x} &= \int_T \frac{1}{i\omega\epsilon} \operatorname{div}\boldsymbol{\sigma}_h \, d\boldsymbol{x} \int_T v \, d\boldsymbol{x} && \forall v \in P^{p+1}(T)
\end{aligned}$$

for $\tilde{u}_h \in P^{p+1}(T)$ with $\boldsymbol{\sigma}_h \in \mathcal{RT}_{p+1}$ as the solution of the mixed hybrid problem. Note that in the first equation, $\operatorname{grad} u = i\omega\mu\boldsymbol{\sigma}$ is used to fix $\tilde{u}_h$ up to a constant. This constant is computed with the help of the equation $i\omega\epsilon u = \operatorname{div}\sigma$. The L^2 error of the post processed solution $\tilde{u}_h$ is plotted in blue in Figure 5.1. It is clearly visible that $\tilde{u}_h$ has the same order of convergence as the standard finite element solution (plotted in magenta) with an H^1-conforming finite element space of the same polynomial order, $p+1$, and as the L^2 best approximation in the space U_{hp+1} (cyan). Thus, we get

$$\big\|u - \tilde{u}_h\big\|_{L^2(\Omega)} \approx ch^{p+2}$$

which is according to (4.7) optimal.

As a conclusion, we remark that from a facet solution u_{Fh} and σ_{Fh} which is approximated with polynomials of order p a solution for the flux field can be reconstructed by local computations. This solution converges optimally with the order h^{p+1}. Additionally an element wise post processing step provides an approximation for u of the order $p+1$, which even converges with order h^{p+2}.

5.2.4 Conservation of energy

We already mentioned that by exchanging the variables u_F, σ_F with the incoming and outgoing impedance traces on the element an equivalent formulation inspired by the ultra weak variational formulation is obtained. [MSS10] showed that, without the absorption term ($\alpha = 0$), the operator representing the Schur complement, i.e. which calculates the outgoing trace for a given incoming trace, is an isometry. This leads us to the following Lemma

Lemma 5.5. *Let ϵ, μ, β be real, and the unknowns g_I^β and g_O^β are defined on each element T as*

$$\begin{aligned}
g_I^\beta &:= -\beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F + u_F\\
g_O^\beta &:= \beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F + u_F,
\end{aligned}$$

where σ_F *and* u_F *solve the problem arising from Formulation 5.2, then*

$$\big\|g_I^\beta\big\|_{L^2(\Gamma)}^2 = \big\|g_O^\beta\big\|_{L^2(\Gamma)}^2.$$

Proof. Consider an arbitrary element $T \in \mathcal{T}$. By testing (5.3) with $(v, \mathbf{0}, 0, 0)$ for any $v \in C_0^\infty(T)$, the density of $C_0^\infty(T)$ in $L^2(T)$ yields $i\omega\epsilon u = \operatorname{div}\boldsymbol{\sigma}$. Consequently, on the element boundary ∂T it holds that that $\alpha(u - u_F) = 0$.

Now, consider an arbitrary inner facet $F \in \mathcal{F}_I$ with adjacent elements T_1 and T_2, and test the underlying problem with any test function v_F which is nonzero just on F. Because of $v_F \in L^2(F)$ and $\alpha(u - u_F) = 0$, we get $\boldsymbol{\sigma} \cdot \boldsymbol{n}_T|_{T_1} + \boldsymbol{\sigma} \cdot \boldsymbol{n}_T|_{T_2} = 0$. Consequently, $\boldsymbol{\sigma}$ is normal continuous across F.

Using τ_F with τ_F just nonzero on F as test function leads to $2\sigma_F = \boldsymbol{\sigma}\cdot\boldsymbol{n}_F|_{T_1} + \boldsymbol{\sigma}\cdot\boldsymbol{n}_F|_{T_2}$, and because of normal continuity $\sigma_F = \boldsymbol{\sigma} \cdot \boldsymbol{n}_F$ follows. This result is obtained directly for boundary facets by testing with τ_F.

Finally, we again consider an arbitrary element T, and test $B_{s\Omega}(\tilde{u}, \tilde{v}) + B_{s\Gamma}(\tilde{u}, \tilde{v}) = F_s(\tilde{v})$ with $(v, \boldsymbol{\tau}, 0, 0)$, where $v = -\overline{u}$ and $\boldsymbol{\tau} = \overline{\boldsymbol{\sigma}}$ in T and zero elsewhere. This leads with $\alpha(u - u_F) = 0$ on ∂T to

$$\begin{aligned}
&- i\omega\big(\epsilon u, \overline{u}\big)_T + \big(\operatorname{div}\sigma, \overline{u}\big)_T - \big(u, \operatorname{div}\overline{\boldsymbol{\sigma}}\big)_T - i\omega\big(\boldsymbol{\sigma}, \overline{\boldsymbol{\sigma}}\big)_T \\
&\quad + \big\langle u_F, \overline{\boldsymbol{\sigma}} \cdot \boldsymbol{n}_T\big\rangle_{\partial T} + \beta\big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_T, \overline{\boldsymbol{\sigma}} \cdot \boldsymbol{n}_T\big\rangle_{\partial T} - \big\langle (\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F, \overline{\boldsymbol{\sigma}} \cdot \boldsymbol{n}_T\big\rangle_{\partial T} = 0.
\end{aligned}$$

Note that $\big(\operatorname{div}\sigma, \overline{u}\big)_T - \big(u, \operatorname{div}\overline{\boldsymbol{\sigma}}\big)_T = 2i\operatorname{Im}\big(\operatorname{div}\boldsymbol{\sigma}, \overline{u}\big)_T$ and that $i\omega\big(\epsilon u, \overline{u}\big)_T$ and $i\omega\big(\mu\boldsymbol{\sigma}, \overline{\boldsymbol{\sigma}}\big)_T$ are purely imaginary. Consequently taking the real part of the equation and inserting g_I^β yields

$$\operatorname{Re}\big(\big\langle g_I^\beta, \overline{\boldsymbol{\sigma}} \cdot \boldsymbol{n}_T\big\rangle_{\partial T}\big) + \beta\|\boldsymbol{\sigma} \cdot \boldsymbol{n}_T\|_{L^2(\partial T)} = 0.$$

Based on this, we get by using $\sigma_F = \boldsymbol{\sigma} \cdot \boldsymbol{n}_F$

$$\begin{aligned}
\big\|g_O^\beta\big\|_{L^2(\partial T)}^2 &= \big\|g_I^\beta + 2\beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_F\big\|_{L^2(\partial T)}^2 = \big\|g_I^\beta + 2\beta\boldsymbol{\sigma} \cdot \boldsymbol{n}_T\big\|_{L^2(\partial T)}^2 \\
&= \big\|g_I^\beta\big\|_{L^2(\partial T)}^2 + 4\beta^2\big\|\boldsymbol{\sigma} \cdot \boldsymbol{n}_T\big\|_{L^2(\partial T)}^2 + 4\beta\operatorname{Re}\big\langle g_I^\beta, \overline{\boldsymbol{\sigma}} \cdot \boldsymbol{n}_T\big\rangle_{\partial T} \\
&= \big\|g_I^\beta\big\|_{L^2(\partial T)}^2.
\end{aligned}$$

By summing up over all elements, we obtain

$$\sum_{T\in\mathcal{T}} \|g_I^\beta\|^2_{L^2(\partial T)} = \sum_{T\in\mathcal{T}} \|g_O^\beta\|^2_{L^2(\partial T)}.$$

The fact that g_I^β on the boundary ∂T of one element equals g_O^β for the neighboring element allows us to ignore the sum over inner facets, which yields

$$\left\|g_I^\beta\right\|^2_{L^2(\Gamma)} = \left\|g_O^\beta\right\|^2_{L^2(\Gamma)}.$$

□

Remark 5.6. *If we choose $\beta = \sqrt{\mu/\epsilon}$ at least on Γ, $g_I^\beta = In_\Gamma$ and $g_O^\beta = Out_\Gamma$ on the boundary. Because g represents the given incoming impedance trace on the domain, we have*

$$\|g\|^2_{L^2(\Gamma)} = \|In_\Gamma\|^2_{L^2(\Gamma)} = \left\|Out_\Gamma\right\|^2_{L^2(\Gamma)}.$$

As already mentioned, the incoming impedance trace can be, up to a phase factor, interpreted as the amplitude of an incoming wave, and the outgoing impedance trace is related to the amplitude of outgoing or scattered waves. Knowing that the square of the amplitudes absolute value is proportional to the energy of the wave, the lemma states for $\beta = \sqrt{\mu/\epsilon}$ energy conservation.

Note that for the continuous problem we have energy conservation for $\alpha \neq 0$. This is not the case for the discrete problem, as we will see in the following. Using the discrete spaces from the last subsection, we obtain the discrete variational formulation:

Formulation 5.7. *Find $\tilde{u}_h := (u_h, \boldsymbol{\sigma}_h, u_{Fh}, \sigma_{Fh}) \in U_{hp} \times V_{hp+1} \times U_{F,hp} \times V_{F,hp}$ such that*

$$B_{s\Omega}(\tilde{u}_h, \tilde{v}_h) + B_{s\Gamma}(\tilde{u}_h, \tilde{v}_h) = F_s(\tilde{v}_h) \tag{5.4}$$

for all $\tilde{v}_h := (v_h, \boldsymbol{\tau}_h, v_{Fh}, \tau_{Fh}) \in U_{hp} \times V_{hp+1} \times U_{F,hp} \times V_{F,hp}$, with the bilinear forms and the linear form from Formulation 5.2.

If the damping parameter α is zero, energy conservation is guaranteed also for the discrete formulation.

Lemma 5.8. *Let μ, ϵ and β be real, $\alpha = 0$, and the unknowns g_I^β and g_O^β are defined on*

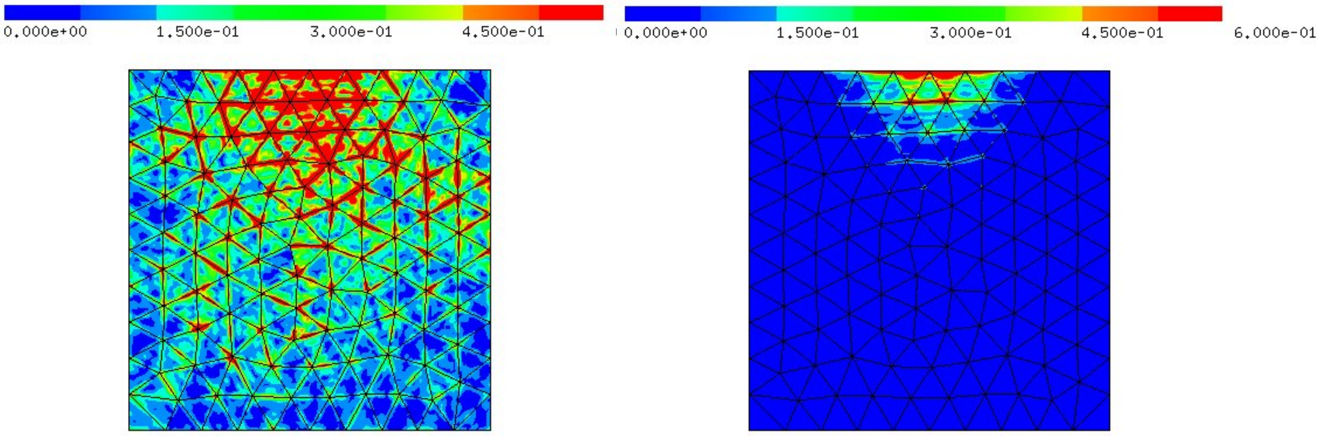

Figure 5.2: $|u_h|$ computed with $\alpha = 0$ (left) and $\alpha \neq 0$ (right) for a wave which can not be resolved by the discrete space U_{hp}

each element T as

$$\begin{aligned} g_I^\beta &:= -\beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_{Fh} + u_{Fh} \\ g_O^\beta &:= \beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sigma_{Fh} + u_{Fh}, \end{aligned}$$

where σ_{Fh} and u_{Fh} solve the problem arising from Formulation 5.7, then

$$\|g_I^\beta\|^2_{L^2(\Gamma)} = \|g_O^\beta\|^2_{L^2(\Gamma)}.$$

Proof. Due to the choice of spaces $\boldsymbol{\sigma}_h \cdot \boldsymbol{n}_F$ is a polynomial of order p on any facet F. Hence it is contained in U_{Fh} and V_{Fh}, respectively. Consequently, as in the continuous case, one can show by testing with v_{Fh} and τ_{Fh} that $\boldsymbol{\sigma}_h$ is normal continuous and $\sigma_{Fh} = \boldsymbol{\sigma}_h \cdot \boldsymbol{n}_F$ on F.

The rest of the proof follows the proof of Lemma 5.8. One just hast to exchange u, $\boldsymbol{\sigma}$, u_F, σ_F and the corresponding test functions by the discrete quantities.

□

Because of this conservation property spurious modes can be generated for $\alpha = 0$ if the solution can not be resolved by the discrete spaces. In Figure 5.2 this is demonstrated for an example with the unit square as computational domain. Choosing $\alpha = 0$, and a polynomial order which is too small to resolve the wave causes spurious modes (left hand plot). Taking a non-zero α leads to a damping out of these spurious oscillations (right hand plot).

5.3 A hybrid finite element method for the vector valued wave equation

In this section, we extend the hybridization technique used for the Helmholtz equation to the vectorial case. We follow our presentation of the method in [HPS11].

5.3.1 Hybridizing the vector valued wave equation

As in the Helmholtz case, we start from the mixed form of the wave equation, (2.35)-(2.37), multiply (2.35) and (2.36) with test functions $\boldsymbol{e} \in X = (L^2(\Omega))^3$ and $\boldsymbol{h} \in Y = H(\mathrm{curl}, \Omega)$, respectively, and integrate over the domain Ω. Element wise integration by parts in the second equation yields

$$\begin{aligned}
\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon\boldsymbol{E}, \boldsymbol{e}\big)_\Omega + \big(\operatorname{curl}\boldsymbol{H}, \boldsymbol{e}\big)_\Omega \Big] &= 0 \qquad && \forall \boldsymbol{e} \in X \\
\sum_{T\in\mathcal{T}} \Big[\big(\boldsymbol{H}, \operatorname{curl}\boldsymbol{h}\big)_\Omega - \big(i\omega\mu\boldsymbol{H}, \boldsymbol{h}\big)_\Omega - \big\langle \boldsymbol{E}, \boldsymbol{n}_\Gamma \times \boldsymbol{h} \big\rangle_\Gamma \Big] &= 0 \qquad && \forall \boldsymbol{h} \in Y.
\end{aligned}$$

Because of the tangential continuity of the test function $\boldsymbol{h}$, the element boundary term cancels for inner edges. Inserting the boundary condition for surface edges leads to Formulation 3.29.

Hybridization is carried out by breaking the tangential continuity of the flux field $\boldsymbol{H}$. Thus, we exchange the space Y by

$$\widetilde{Y} := \Big\{ \boldsymbol{v} \in \big(L^2(\Omega)\big)^3 \;:\; \boldsymbol{v}\big|_T \in H(\mathrm{curl}, T) \quad \forall T \in \mathcal{T} \Big\},$$

a space containing functions which are element wise in $H(\mathrm{curl})$. In order to regain tangential continuity, a vector valued Lagrangian parameter $\boldsymbol{E}_F$ which is supported only on element facets is introduced. This Lagrangian multiplier needs to be vector valued with a direction parallel to the facet, and we have to take

$$\boldsymbol{E}_F \in X_F := \Big\{ \boldsymbol{e} \in \big(L^2(\mathcal{F})\big)^3 \;:\; \boldsymbol{e} \cdot \boldsymbol{n}_F = 0 \text{ for all } F \in \mathcal{F} \Big\}$$

where $\boldsymbol{E}_F$ is component wise in L^2. Continuity is now achieved by setting the jump $[\boldsymbol{n} \times \boldsymbol{H}]_F = \boldsymbol{n}_{\partial T_1} \times \boldsymbol{H} + \boldsymbol{n}_{\partial T_2} \times \boldsymbol{H}$ to zero for inner facets with neighboring elements T_1 and T_2, i.e.

$$\sum_{F \in \mathcal{F}_\mathcal{I}} \big\langle [\boldsymbol{n} \times \boldsymbol{H}]_F, \boldsymbol{e}_F \big\rangle_F = \sum_{T \in \mathcal{T}} \Big(\big\langle \boldsymbol{n}_T \times \boldsymbol{H}, \boldsymbol{e}_F \big\rangle_{\partial T} - \big\langle \boldsymbol{n}_T \times \boldsymbol{H}, \boldsymbol{e}_F \big\rangle_{\partial T \cap \Gamma} \Big) = 0 \quad \forall \boldsymbol{e}_F \in X_F.$$

Inserting the absorbing boundary condition leads to

$$\begin{aligned}
&\sum_{T \in \mathcal{T}} \Big[\big(i\omega\epsilon \boldsymbol{E}, \boldsymbol{e} \big)_T + \big(\operatorname{curl} \boldsymbol{H}, \boldsymbol{e} \big)_T \Big] && = 0 && \forall \boldsymbol{e} \in X && (5.5)\\
&\sum_{T \in \mathcal{T}} \Big[\big(\boldsymbol{E}, \operatorname{curl} \boldsymbol{h} \big)_T - \big(i\omega\mu \boldsymbol{H}, \boldsymbol{h} \big)_T - \big\langle \boldsymbol{E}_F, \boldsymbol{n}_T \times \boldsymbol{h} \big\rangle_{\partial T} \Big] && = 0 && \forall \boldsymbol{h} \in \widetilde{Y} && (5.6)\\
&\sum_{T \in \mathcal{T}} \Big[- \big\langle \boldsymbol{n}_T \times \boldsymbol{H}, \boldsymbol{e}_F \big\rangle_{\partial T} \Big] \qquad - \big\langle \sqrt{\frac{\epsilon}{\mu}} \boldsymbol{E}_F, \boldsymbol{e}_F \big\rangle_\Gamma && = -\big\langle \sqrt{\frac{\epsilon}{\mu}} \boldsymbol{g}, \boldsymbol{e}_F \big\rangle_\Gamma && \forall \boldsymbol{e}_F \in X_F.
\end{aligned}$$

Note that the Lagrangian multiplier plays the role of $\boldsymbol{E}_{||}$ on the facets which was already used when the boundary condition was inserted, and therefore the system of equations stays symmetric.

As in the scalar case, static condensation of the element wise volume unknowns $\boldsymbol{E}$ and $\boldsymbol{H}$ is problematic. Elimination of the volume unknowns is equivalent to solving the equations (5.5) and (5.6) for any $\boldsymbol{E}_F$, and this means solving a Dirichlet problem on the element level. If ω is an eigenvalue, such a solution is not unique.

Uniquely solvable local problems are obtained by adding a facet unknown

$$\boldsymbol{H}_F \in Y_F := \Big\{ \boldsymbol{h} \in \big(L^2(\mathcal{F}) \big)^3 \ : \ \boldsymbol{h} \cdot \boldsymbol{n}_F = 0 \text{ for all } F \in \mathcal{F} \Big\}$$

which is set with the help of an additional equation

$$\sum_{T \in \mathcal{T}} \beta \big\langle \boldsymbol{H}_F - \boldsymbol{n}_F \times \boldsymbol{H}, \boldsymbol{h}_F \big\rangle_{\partial T} = 0 \qquad \forall \boldsymbol{h}_F \in Y_F$$

to $\boldsymbol{n}_F \times \boldsymbol{H}$ on the facets with a stabilization parameter β. Furthermore, symmetry is achieved by subtracting the consistent term $\sum_{T \in \mathcal{T}} \big\langle \boldsymbol{H}_F - \boldsymbol{n}_F \times \boldsymbol{H}, \boldsymbol{n}_F \times \boldsymbol{h} \big\rangle_{\partial T}$. This leads to

$$\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon\boldsymbol{E},\boldsymbol{e}\big)_T + \big(\operatorname{curl}\boldsymbol{H},\boldsymbol{e}\big)_T\Big] = 0 \tag{5.7}$$

$$\begin{aligned}\sum_{T\in\mathcal{T}} \Big[\big(\boldsymbol{E},\operatorname{curl}\boldsymbol{h}\big)_T - \big(i\omega\mu\boldsymbol{H},\boldsymbol{h}\big)_T + \beta\big\langle\boldsymbol{n}_T\times\boldsymbol{H},\boldsymbol{n}_T\times\boldsymbol{h}\big\rangle_{\partial T}&\\ -\beta\big\langle(\boldsymbol{n}_T\cdot\boldsymbol{n}_F)\boldsymbol{H}_F,\boldsymbol{n}_T\times\boldsymbol{h}\big\rangle_{\partial T} - \big\langle\boldsymbol{E}_F,\boldsymbol{n}_T\times\boldsymbol{h}\big\rangle_{\partial T}\Big] &= 0 \end{aligned}\tag{5.8}$$

$$\sum_{T\in\mathcal{T}} \Big[-\big\langle\boldsymbol{n}_T\times\boldsymbol{H},\boldsymbol{e}_F\big\rangle_{\partial T}\Big] - \big\langle\sqrt{\frac{\epsilon}{\mu}}\boldsymbol{E}_F,\boldsymbol{e}_F\big\rangle_\Gamma = -\big\langle\sqrt{\frac{\epsilon}{\mu}}\boldsymbol{g},\boldsymbol{e}_F\big\rangle_\Gamma$$

$$\sum_{T\in\mathcal{T}} \Big[\beta\big\langle\boldsymbol{H}_F,\boldsymbol{h}_F\big\rangle_{\partial T} - \beta\big\langle\boldsymbol{n}_F\times\boldsymbol{H},\boldsymbol{h}_F\big\rangle_{\partial T}\Big] = 0.$$

Now, static condensation, i.e. solving (5.7) and (5.8) for $\boldsymbol{E}$ and $\boldsymbol{H}$, is for $\beta = \sqrt{\frac{\mu}{\epsilon}}$ equivalent to solve the equations (2.35)-(2.37) on $\Omega = T$ with right hand side $\sqrt{\frac{\mu}{\epsilon}}(\boldsymbol{n}_T\cdot\boldsymbol{n}_F)\boldsymbol{H}_F + \boldsymbol{E}_F$, and existence and uniqueness are given by Lemma 3.30.

If we take into account that $\boldsymbol{H}_F$ plays the role of $\boldsymbol{n}_F\times\boldsymbol{H}$ and $\boldsymbol{E}_F$ approximates $\boldsymbol{E}_{\|}$ on the facet, we can again redefine incoming and outgoing impedance traces as functions of the facet variables:

Definition 5.9. *Let T be an element of the triangulation $\mathcal{T}$. Then the incoming and outgoing impedance traces for the facet variables $\boldsymbol{E}_F$ and $\boldsymbol{H}_F$ are defined as*

$$\begin{aligned}\boldsymbol{In}_{\partial T} &:= \sqrt{\frac{\mu}{\epsilon}}(\boldsymbol{n}_T\cdot\boldsymbol{n}_F)\boldsymbol{H}_F\big|_{\partial T} + \boldsymbol{E}_F\big|_{\partial T}\\ \boldsymbol{Out}_{\partial T} &:= -\sqrt{\frac{\mu}{\epsilon}}(\boldsymbol{n}_T\cdot\boldsymbol{n}_F)\boldsymbol{H}_F\big|_{\partial T} + \boldsymbol{E}_F\big|_{\partial T}.\end{aligned}$$

In the same way, incoming and outgoing impedance traces for the domain $\boldsymbol{In}_\Gamma$ and $\boldsymbol{Out}_\Gamma$ can be defined.

With this definition and for $\beta = \sqrt{\mu/\epsilon}$, we can interpret the elimination of the inner degrees of freedom as calculating the outgoing impedance trace $\boldsymbol{Out}_{\partial T}$ for an incoming impedance trace $\boldsymbol{In}_{\partial T}$. Thus, the reaction of an element for an incoming wave is computed. If one replaces the facet unknowns $\boldsymbol{E}_F$ and $\boldsymbol{H}_F$ by the impedance traces, one obtains an equivalent formulation which again fits into the context of the ultra weak variational formulation.

Because, as we will see in the following, this system of equations is energy conserving, and because a finite number of degrees of freedom per wavelength is needed to resolve the wave, high frequency waves, which can not be resolved, cause spurious modes. This fact motivates an additional symmetric stabilization term with parameter α,

$$\sum_{T\in\partial\mathcal{T}} \alpha\big\langle \boldsymbol{E}_F-\boldsymbol{E}_\parallel, \boldsymbol{e}_F-\boldsymbol{e}_\parallel\big\rangle_{\partial T}$$

which damps out unresolved frequencies. Collecting everything, we end up with the hybrid formulation.

Formulation 5.10 (the mixed hybrid formulation for the vector valued wave equation)**.** *Find* $\widetilde{\boldsymbol{u}} := (\boldsymbol{E},\boldsymbol{H},\boldsymbol{E}_F,\boldsymbol{H}_F) \in X\times\widetilde{Y}\times X_F\times Y_F$ *such that*

$$B_v(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}) = F_v(\widetilde{\boldsymbol{v}}) \tag{5.9}$$

holds for all $\widetilde{\boldsymbol{v}} := (\boldsymbol{e},\boldsymbol{h},\boldsymbol{e}_F,\boldsymbol{h}_F) \in X\times\widetilde{Y}\times X_F\times Y_F$ *with the bilinear form*

$$\begin{aligned} B_v(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}) \;:=\; \sum_{T\in\mathcal{T}} \Big[& \big(i\omega\epsilon\boldsymbol{E},\boldsymbol{e}\big)_T + \big(\operatorname{curl}\boldsymbol{H},\boldsymbol{e}\big)_T + \big(\boldsymbol{E},\operatorname{curl}\boldsymbol{h}\big)_T - \big(i\omega\mu\boldsymbol{H},\boldsymbol{h}\big)_T \\ & -\big\langle\boldsymbol{E}_F,\boldsymbol{n}_T\times\boldsymbol{h}\big\rangle_{\partial T} - \big\langle\boldsymbol{n}_T\times\boldsymbol{H},\boldsymbol{e}_F\big\rangle_{\partial T} + \alpha\big\langle\boldsymbol{E}_F-\boldsymbol{E}_\parallel,\boldsymbol{e}_F-\boldsymbol{e}_\parallel\big\rangle_{\partial T} \\ & +\beta\big\langle\boldsymbol{H}_F-\boldsymbol{n}_F\times\boldsymbol{H},\boldsymbol{h}_F-\boldsymbol{n}_F\times\boldsymbol{h}\big\rangle_{\partial T}\Big] - \big\langle\sqrt{\frac{\epsilon}{\mu}}\boldsymbol{E}_F,\boldsymbol{e}_F\big\rangle_\Gamma \end{aligned}$$

and the linear form

$$F_v(\widetilde{\boldsymbol{v}}) := -\big\langle\sqrt{\frac{\epsilon}{\mu}}\boldsymbol{g},\boldsymbol{e}_F\big\rangle_\Gamma.$$

5.3.2 Consistency existence and uniqueness of the formulation

The following Lemma now states the consistency of this formulation.

Lemma 5.11. *Let* $(\boldsymbol{E}^e,\boldsymbol{H}^e)$ *be the exact solution of (2.35)-(2.37), and let* $\boldsymbol{E}_F^e = \boldsymbol{E}_\parallel^e$ *and* $\boldsymbol{H}_F^e = \boldsymbol{n}_F\times\boldsymbol{H}^e$ *on the facets* F*. Then* $\widetilde{\boldsymbol{u}}^e := (\boldsymbol{E}^e,\boldsymbol{H}^e,\boldsymbol{E}_F^e,\boldsymbol{H}_F^e)$ *solves the variational formulation 5.10.*

Proof. First, we insert $\widetilde{\boldsymbol{u}}^e$ together with the boundary condition (2.37) into Formulation

5.10, and we obtain

$$\sum_{T\in\mathcal{T}}\Big[\big(i\omega\epsilon\boldsymbol{E}^e,\boldsymbol{e}\big)_T+\big(\operatorname{curl}\boldsymbol{H}^e,\boldsymbol{e}\big)_T+\big(\boldsymbol{E}^e,\operatorname{curl}\boldsymbol{h}\big)_T-\big(i\omega\mu\boldsymbol{H}^e,\boldsymbol{h}\big)_T$$
$$-\big\langle\boldsymbol{E}^e_F,\boldsymbol{n}_T\times\boldsymbol{h}\big\rangle_{\partial T}-\big\langle\boldsymbol{n}_T\times\boldsymbol{H}^e,\boldsymbol{e}_F\big\rangle_{\partial T}\Big]\quad+\big\langle\boldsymbol{n}_\Gamma\times\boldsymbol{H}^e,\boldsymbol{e}_F\big\rangle_\Gamma=0.$$

Now, we make use of the fact that the exact solution $\boldsymbol{H}^e$ is tangential continuous across element interfaces. Thus, the term $\big\langle\boldsymbol{n}_T\times\boldsymbol{H}^e,\boldsymbol{e}_F\big\rangle_{\partial T}$ cancels out for all inner facets. On boundary facets, where $\boldsymbol{n}_T=\boldsymbol{n}_\Gamma$, this term is eliminated by the boundary term. Via $\boldsymbol{E}^e_\parallel=\boldsymbol{E}^e_F$, this leads to

$$\sum_{T\in\mathcal{T}}\Big[\big(i\omega\epsilon\boldsymbol{E}^e,\boldsymbol{e}\big)_T+\big(\operatorname{curl}\boldsymbol{H}^e,\boldsymbol{e}\big)_T$$
$$-\big(i\omega\mu\boldsymbol{H}^e,\boldsymbol{h}\big)_T+\big(\boldsymbol{E}^e,\operatorname{curl}\boldsymbol{h}\big)_T-\big\langle\boldsymbol{E}^e_\parallel,\boldsymbol{n}_T\times\boldsymbol{h}\big\rangle_{\partial T}\Big]=0\qquad.$$

Note that $\boldsymbol{E}^e_\parallel$ in the facet term can be exchanged by $\boldsymbol{E}^e$, because it is only tested against vector fields tangential to the facet. Integration by parts of the second equation results now in

$$\sum_{T\in\mathcal{T}}\Big[\big(i\omega\epsilon\boldsymbol{E}^e+\operatorname{curl}\boldsymbol{H}^e,\boldsymbol{e}\big)_T+\big(\operatorname{curl}\boldsymbol{E}^e-i\omega\mu\boldsymbol{H}^e,\boldsymbol{h}\big)_T\Big]=0,$$

and the Lemma is proven. □

Additionally, the existence and uniqueness for the solution of Formulation 5.10 can be shown.

Lemma 5.12. *Let μ be a positive constant, $\epsilon\in H^3(\Omega)$ is a real valued function with $0<\epsilon_{min}\le\epsilon\le\epsilon_{max}$, $g\in\big(L^2(\Gamma)\big)^3$, $\beta\neq0$, $\alpha=0$ and Ω a Lipschitz domain. Then a unique solution of Formulation 5.10 exists.*

Proof. Consider an arbitrary inner facet $F\in\mathcal{F}_I$. We start the proof by testing (5.9) with any $\boldsymbol{e}_F\in X_F$ supported only on F. If T_1 and T_2 are the two adjacent elements of F, we get $\boldsymbol{n}_T\times\boldsymbol{H}|_{T_1}+\boldsymbol{n}_T\times\boldsymbol{H}|_{T_2}=0$, and due to $\boldsymbol{n}_{T_1}=-\boldsymbol{n}_{T_2}$ the field $\boldsymbol{H}$ is tangential continuous across element interfaces. Thus, the space $\widetilde{Y}$ in the variational formulation can be exchanged by $Y=H(\operatorname{curl},\Omega)$.

Testing now (5.9) with any $\boldsymbol{h}_F\in Y_F$ such that $\boldsymbol{h}_F$ is just nonzero on F leads to $2\boldsymbol{H}_F=\boldsymbol{n}_F\times\boldsymbol{H}|_{T_1}+\boldsymbol{n}_F\times\boldsymbol{H}|_{T_2}$, and because of tangential continuity $\boldsymbol{H}_F=\boldsymbol{n}_F\times\boldsymbol{H}$

follows. For boundary facets, this can be obtained directly by testing with $\boldsymbol{h}_F$ supported just on the boundary.

Consequently, the β term in B_v vanishes, and $\sum_{T\in\mathcal{T}} \langle \boldsymbol{n}_T \times \boldsymbol{H}, \boldsymbol{e}_F \rangle_{\partial T}$ as well as $\sum_{T\in\mathcal{T}} \langle \boldsymbol{E}_F, \boldsymbol{n}_T \times \boldsymbol{h} \rangle_{\partial T}$ simplify to $\langle \boldsymbol{n}_\Gamma \times \boldsymbol{H}, \boldsymbol{e}_F \rangle_\Gamma$ and $\langle \boldsymbol{E}_F, \boldsymbol{n}_\Gamma \times \boldsymbol{h} \rangle_\Gamma$, respectively. Collecting everything gives

$$\begin{aligned}
&\sum_{T\in\mathcal{T}} \Big[\big(i\omega\epsilon \boldsymbol{E}, \boldsymbol{e}\big)_T + \big(\operatorname{curl} \boldsymbol{H}, \boldsymbol{e}\big)_T \Big] &&= 0 &&\forall \boldsymbol{e} \in X, \\
&\sum_{T\in\mathcal{T}} \Big[\big(\boldsymbol{E}, \operatorname{curl} \boldsymbol{h}\big)_T - \big(i\omega\mu \boldsymbol{H}, \boldsymbol{h}\big)_T \Big] - \big\langle \boldsymbol{E}_F, \boldsymbol{n}_\Gamma \times \boldsymbol{h} \big\rangle_\Gamma &&= 0 &&\forall \boldsymbol{h} \in Y, \\
&-\big\langle \boldsymbol{n}_\Gamma \times \boldsymbol{H}, \boldsymbol{e}_F \big\rangle_\Gamma - \big\langle \sqrt{\frac{\epsilon}{\mu}} \boldsymbol{E}_F, \boldsymbol{e}_F \big\rangle_\Gamma &&= -\big\langle \sqrt{\frac{\epsilon}{\mu}} \boldsymbol{g}, \boldsymbol{e}_F \big\rangle_\Gamma &&\forall \boldsymbol{e}_F \in X_F.
\end{aligned}$$

From the last equation it follows, that $\boldsymbol{E}_F = \boldsymbol{g} - \sqrt{\mu/\epsilon}\, \boldsymbol{n}_\Gamma \times \boldsymbol{H}$. Inserting this into the second equation yields the standard mixed Formulation 3.29, and Lemma 3.30 guarantees existence and uniqueness. ☐

5.3.3 The discrete finite element spaces

Now, we are in the position to search for appropriate discrete spaces $X_{hp}, Y_{hp}, X_{F,hp}, Y_{F,hp}$ for the function spaces $X, \widetilde{Y}, X_F, Y_F$ involved in Formulation 5.10. A natural choice for the space of the flux field $\boldsymbol{H}$ is a broken Nédélec space

$$Y_{hk} := \Big\{ \boldsymbol{h} \in \big(L^2(\Omega)\big)^3 \ : \ \boldsymbol{h}|_T \in \mathcal{ND}_k^{II}(T) \text{ for all } T \in \mathcal{T} \Big\}$$

with polynomial order $k = p+1$, i.e. a Nédélec space without any continuity constraints on the element interfaces. The fact that the field $\boldsymbol{E} \in X$ approximates $\operatorname{curl} \boldsymbol{H}$ on the element level, which can be seen by testing Formulation 5.10 with $(\boldsymbol{e}, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0})$, and that the curl of a function in $\mathcal{ND}_{p+1}^{II}(T)$ is in $\big(P^p(T)\big)^3$ suggests the definition

$$X_{hp} := \Big\{ \boldsymbol{e} \in \big(L^2(\Omega)\big)^3 \ : \ \boldsymbol{e}|_T \in \big(P^p(T)\big)^3 \text{ forall } T \in \mathcal{T} \Big\}.$$

Because for $\boldsymbol{H} \in \mathcal{ND}_{p+1}^{II}(T)$ the tangential trace of $\boldsymbol{H}$ is also a polynomial of order $p+1$, and because the facet variables $\boldsymbol{E}_F, \boldsymbol{H}_F$ are used to ensure the tangential continuity of $\boldsymbol{H}$,

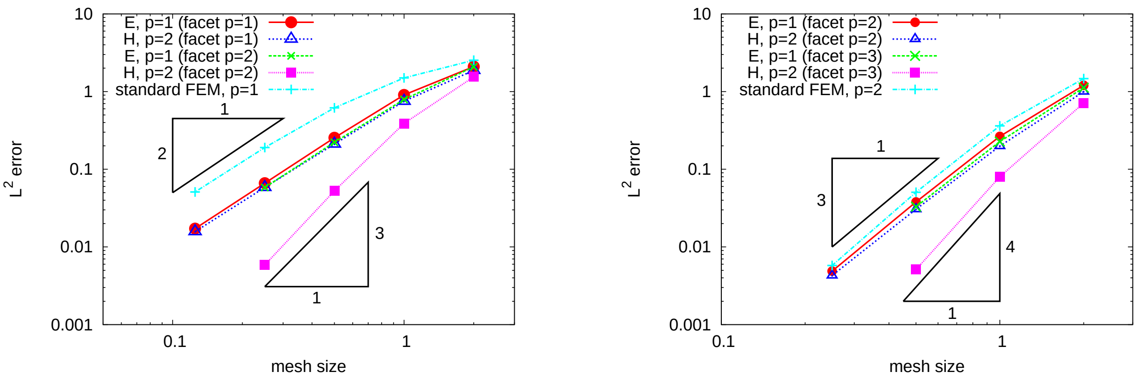

Figure 5.3: The L^2 error of the mixed hybrid solutions $\boldsymbol{E} \in X_{hp}$, $\boldsymbol{H} \in Y_{hp+1}$ calculated with p order polynomials (red and blue) and $p+1$ order polynomials (green and magenta) on the facet. The error of a standard finite element solution is plotted in cyan.

i.e. $\boldsymbol{e}_F \in X_{F,hp}, \boldsymbol{h}_F \in Y_{F,hp}$ are tested against $\boldsymbol{n}_T \times \boldsymbol{H}$, the discrete function spaces

$$X_{F,hk} = Y_{F,hk} := \left\{ \boldsymbol{u} \in \left(L^2(\mathcal{F})\right)^3 \ : \ \boldsymbol{u} \in \left(P^k(F)\right)^3, \ \boldsymbol{n}_F \cdot \boldsymbol{u} = 0 \text{ for all } F \in \mathcal{F} \right\}$$

with polynomial order $k = p + 1$ can be motivated. On the other hand, $\boldsymbol{E}_F$ plays the role of the tangential component of the unknown field $\boldsymbol{E}$ on the facet, which is for $\boldsymbol{E} \in X_{hp}$ a polynomial of order p. This would justify a choice of a polynomial order $k = p$ for the facet spaces.

The approximation properties for these choices of spaces were tested by solving a problem on a cube of side length two with a plane wave solution of wavelength 1.7 fixed by appropriate absorbing boundary conditions. For the calculations the parameters β and α were one and zero, respectively. In Figure 5.3 the L^2 error for $\boldsymbol{E} \in X_{hp}$ and $\boldsymbol{H} \in Y_{hp+1}$, with $p = 1$ (left) and $p = 2$ (right) is plotted via the mesh size h. The red and blue lines were calculated by using polynomial order p for the facet spaces, while the polynomial order was increased by one for the green and the magenta lines. For comparison the error of a standard finite element method using Nédélec elements of order p is plotted in cyan.

The slopes of the lines indicate that when using the same polynomial order for the facet spaces and the unknown field $\boldsymbol{E}$, the approximations $\boldsymbol{E}_h$ and $\boldsymbol{H}_h$ converge as

$$\left\| \boldsymbol{E} - \boldsymbol{E}_h \right\|_{L^2(\Omega)} \approx c_1 h^{p+1}, \qquad \left\| \boldsymbol{H} - \boldsymbol{H}_h \right\|_{L^2(\Omega)} \approx c_2 h^{p+1}.$$

Thus, we obtain for the electric field the same convergence order as for a standard finite element method, and according to the interpolation results from (4.7) and (4.4) this is optimal. Because for the magnetic field finite elements of a larger polynomial order are used, and the tangential trace of its approximation $\boldsymbol{H}_h$ can not be approximated exactly by the facet variables, the convergence rate for $\boldsymbol{H}$ is according to (4.4) one order smaller as the optimal one.

If one would increase the polynomial order of the facet spaces by one such that it matches the polynomial order of the tangential trace of the flux field, the approximation of $\boldsymbol{H}$ converges according to the figure with the optimal rate h^{p+2}. But, considering that in the solution process the volume degrees of freedom are eliminated on the element level and that the resulting system of equations needs to be solved just for the facet unknowns, this increase in polynomial order leads effectively to a much bigger linear system of equations. Thus, the computational cost is comparable to solve the mixed problem with $\boldsymbol{E} \in X_{hp+1}$, $\boldsymbol{H} \in Y_{hp+2}$, $\boldsymbol{E}_F \in X_{F,hp+1}$, $\boldsymbol{H}_{F,hp+1}$, which provides because of the higher polynomial order for the volume functions a better convergence rate of the unknown field $\boldsymbol{E}$. Therefore, it is preferable to choose for the facet spaces the same polynomial order as for the unknown $\boldsymbol{E}$, although this leads to a non optimal convergence rate for the flux field.

5.3.4 Conservation of energy

Finally, we are going to prove the beforehand mentioned energy conservation for the continuous problem. Thus, we need the following Lemma.

Lemma 5.13. *Let ϵ, μ, β be real, and the unknowns $\boldsymbol{g}_I^\beta$ and $\boldsymbol{g}_O^\beta$ are defined on each element T as*

$$\begin{aligned} \boldsymbol{g}_I^\beta &:= \beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\boldsymbol{H}_F + \boldsymbol{E}_F \\ \boldsymbol{g}_O^\beta &:= -\beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\boldsymbol{H}_F + \boldsymbol{E}_F \end{aligned}$$

where $\boldsymbol{H}_F$ and $\boldsymbol{E}_F$ solve the variational formulation 5.10, then

$$\|\boldsymbol{g}_I^\beta\|^2_{L^2(\Gamma)} = \|\boldsymbol{g}_O^\beta\|^2_{L^2(\Gamma)}.$$

Proof. First, we assume T to be any element in $\mathcal{T}$. We start by testing (5.9) with $\boldsymbol{e} \in$

$\big(C_0^\infty(T)\big)^3$. Because of the density of $C_0^\infty(T)$ in $L^2(T)$ it follows that $i\omega\epsilon\boldsymbol{E} + \operatorname{curl}\boldsymbol{H} = \boldsymbol{0}$, and consequently $\alpha(\boldsymbol{E}_F - \boldsymbol{E}_\parallel)$ has to be zero on ∂T, too.

Next, we choose any inner facet $F \in \mathcal{F}_I$ with adjacent elements T_1 and T_2, and we take any $\boldsymbol{e}_F \in X_F$ as test function for (5.9) such that $\boldsymbol{e}_F$ is just on F different from zero. Because $\alpha(\boldsymbol{E}_F - \boldsymbol{E}_\parallel) = 0$, we get $\boldsymbol{n}_T \times \boldsymbol{H}|_{T_1} + \boldsymbol{n}_T \times \boldsymbol{H}|_{T_2} = 0$, and $\boldsymbol{H}$ is tangential continuous across element interfaces. Testing our formulation with $\boldsymbol{h}_F \in Y_F$ with $\boldsymbol{h}_F$ only supported on F yields $2\boldsymbol{H}_F = \boldsymbol{n}_F \times \boldsymbol{H}|_{T_1} + \boldsymbol{n}_F \times \boldsymbol{H}|_{T_2}$ which results due to the tangential continuity of $\boldsymbol{H}$ in $\boldsymbol{H}_F = \boldsymbol{n}_F \times \boldsymbol{H}$. The same result is obtained for boundary facets just by testing with $\boldsymbol{h}_F$.

Finally, we take again an arbitrary element $T \in \mathcal{T}$. Choosing $(\boldsymbol{e}, \boldsymbol{h}, \boldsymbol{0}, \boldsymbol{0})$ with $\boldsymbol{e} = -\overline{\boldsymbol{E}}$ and $\boldsymbol{h} = \overline{\boldsymbol{H}}$ on T as test function in (5.9) yields under the consideration that $\boldsymbol{E}_F = \boldsymbol{E}_\parallel$ on ∂T

$$\begin{aligned}
&-\big(i\omega\epsilon\boldsymbol{E}, \overline{\boldsymbol{E}}\big)_T - \big(\operatorname{curl}\boldsymbol{H}, \overline{\boldsymbol{E}}\big)_T + \big(\boldsymbol{E}, \operatorname{curl}\overline{\boldsymbol{H}}\big)_T - \big(i\omega\mu\boldsymbol{H}, \overline{\boldsymbol{H}}\big)_T \\
&+ \beta\big\langle \boldsymbol{n}_T \times \boldsymbol{H}, \boldsymbol{n}_T \times \overline{\boldsymbol{H}}\big\rangle_{\partial T} - \big\langle \beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\boldsymbol{H}_F + \boldsymbol{E}_F, \boldsymbol{n}_T \times \overline{\boldsymbol{H}}\big\rangle_{\partial T} = 0.
\end{aligned}$$

Because $\big(\boldsymbol{E}, \operatorname{curl}\overline{\boldsymbol{H}}\big)_T - \big(\operatorname{curl}\boldsymbol{H}, \overline{\boldsymbol{E}}\big)_T = 2i\mathrm{Im}\big(\boldsymbol{E}, \operatorname{curl}\overline{\boldsymbol{H}}\big)_T$, and $\big(i\omega\epsilon\boldsymbol{E}, \overline{\boldsymbol{E}}\big)_T$ and $\big(i\omega\mu\boldsymbol{H}, \overline{\boldsymbol{H}}\big)_T$ are purely imaginary, taking the real part of the equation from above gives

$$\beta\big\|\boldsymbol{n}_T \times \boldsymbol{H}\big\|^2_{L^2(\partial T)} = \mathrm{Re}\big\langle \boldsymbol{g}_I^\beta, \boldsymbol{n}_T \times \overline{\boldsymbol{H}}\big\rangle_{\partial T}. \tag{5.10}$$

By using this together with $\boldsymbol{H}_F = \boldsymbol{n}_F \times \boldsymbol{H}$, we obtain

$$\begin{aligned}
\big\|\boldsymbol{g}_O^\beta\big\|^2_{L^2(\partial T)} &= \big\|\boldsymbol{g}_I^\beta - 2\beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\boldsymbol{H}_F\big\|^2_{L^2(\partial T)} = \big\|\boldsymbol{g}_I^\beta - 2\beta\boldsymbol{n}_T \times \boldsymbol{H}\big\|^2_{L^2(\partial T)} \\
&= \big\|\boldsymbol{g}_I^\beta\big\|^2_{L^2(\partial T)} + 4\beta^2\big\|\boldsymbol{n}_T \times \boldsymbol{H}\big\|^2_{L^2(\partial T)} - 4\beta\mathrm{Re}\big\langle \boldsymbol{g}_I^\beta, \boldsymbol{n}_T \times \overline{\boldsymbol{H}}\big\rangle_{\partial T} \\
&= \big\|\boldsymbol{g}_I^\beta\big\|^2_{L^2(\partial T)}.
\end{aligned}$$

By summing up over all elements $\sum_{T\in\mathcal{T}} \big\|\boldsymbol{g}_I^\beta\big\|^2_{L^2(\partial T)} = \sum_{T\in\mathcal{T}} \big\|\boldsymbol{g}_O^\beta\big\|^2_{L^2(\partial T)}$ immediately follows. Taking into account that $\boldsymbol{g}_I^\beta$ of one element is $\boldsymbol{g}_O^\beta$ of the neighboring element, the sum over inner facets can be neglected, and the proof is complete.

□

Note, for $\beta = \sqrt{\mu/\epsilon}$ the facet functions $\boldsymbol{g}_I^\beta$ and $\boldsymbol{g}_O^\beta$ are equal to the incoming and outgoing impedance traces $\boldsymbol{In}_\Gamma$ and $\boldsymbol{Out}_\Gamma$, and we have

$$\|\boldsymbol{g}\|_{L^2(\Gamma)} = \|\boldsymbol{In}_\Gamma\|_{L^2(\Gamma)} = \|\boldsymbol{Out}_\Gamma\|_{L^2(\Gamma)}.$$

Again, this can be interpreted as conservation of the physical energy. Note that the damping term has no influence for the continuous problem. If we choose the discrete versions of the facet spaces X_F and Y_F, such that they contain the tangential component of the magnetic field, we are able to show energy conservation for the discrete problem with $\alpha = 0$, too. Thus, we consider in the following the formulation

Formulation 5.14. *Find* $\widetilde{\boldsymbol{u}}_h := (\boldsymbol{E}_h, \boldsymbol{H}_h, \boldsymbol{E}_{Fh}, \boldsymbol{H}_{Fh}) \in X_{hp} \times \widetilde{Y}_{hp+1} \times X_{F,hp+1} \times Y_{F,hp+1}$ *such that*

$$B_v(\widetilde{\boldsymbol{u}}_h, \widetilde{\boldsymbol{v}}_h) = F_v(\widetilde{\boldsymbol{v}}_h) \tag{5.11}$$

holds for all $\widetilde{\boldsymbol{v}}_h := (\boldsymbol{e}_h, \boldsymbol{h}_h, \boldsymbol{e}_{Fh}, \boldsymbol{h}_{Fh}) \in X_{hp} \times \widetilde{Y}_{hp+1} \times X_{F,hp+1} \times Y_{F,hp+1}$ *with the bilinear form and the linear form from Formulation 5.10.*

Lemma 5.15. *Let* ϵ, μ, β *be real,* $\alpha = 0$*, and the unknowns* $\boldsymbol{g}_I^\beta$ *and* $\boldsymbol{g}_O^\beta$ *are defined on each element* T *as*

$$\begin{aligned} \boldsymbol{g}_I^\beta &:= \beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\boldsymbol{H}_{Fh} + \boldsymbol{E}_{Fh} \\ \boldsymbol{g}_O^\beta &:= -\beta(\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\boldsymbol{H}_{Fh} + \boldsymbol{E}_{Fh} \end{aligned}$$

where $\boldsymbol{H}_{Fh}$ *and* $\boldsymbol{E}_{Fh}$ *solve the variational formulation 5.14, then*

$$\|\boldsymbol{g}_I^\beta\|^2_{L^2(\Gamma)} = \|\boldsymbol{g}_O^\beta\|^2_{L^2(\Gamma)}.$$

Proof. The proof of this Lemma is similar to the proof of Lemma 5.13. Because for any facet F, the tangential component of $\boldsymbol{H}_h$ on F is of polynomial order $p+1$ and therefore contained in $X_{F,hp+1}$ and $Y_{F,hp+1}$, respectively, testing (5.11) with $\boldsymbol{h}_{Fh}$ and $\boldsymbol{e}_{Fh}$ implies tangential continuity of $\boldsymbol{H}_h$ and $\boldsymbol{H}_{Fh} = \boldsymbol{n}_F \times \boldsymbol{H}_h$ on F.

The rest of the proof follows the proof of Lemma 5.13. There, the functions $\boldsymbol{E}$, $\boldsymbol{H}$, $\boldsymbol{H}_F$, $\boldsymbol{E}_F$ and the corresponding test functions have to be exchanged by their discrete representations.

□

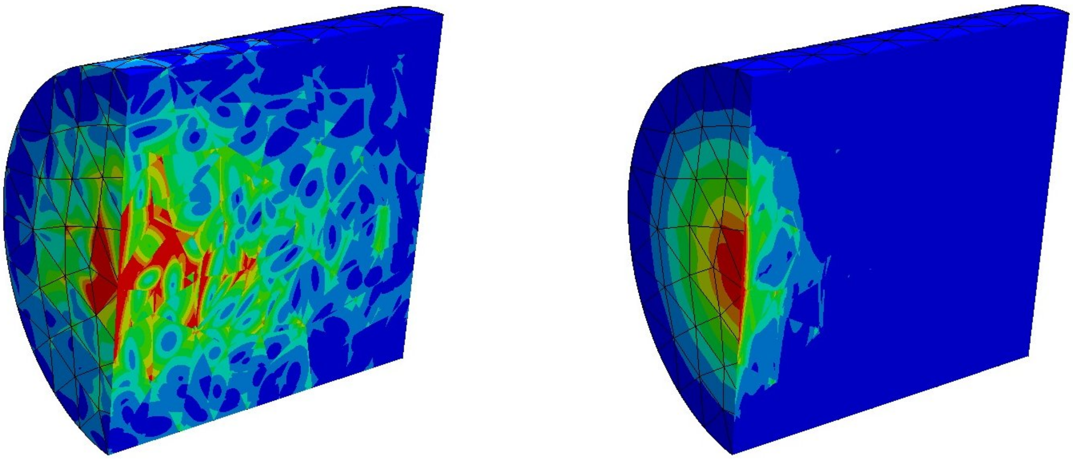

Figure 5.4: $|\boldsymbol{E}_h|$ calculated with $\alpha = 0$ (left) and $\alpha \neq 0$ (right) for a wave which can not be resolved by the discrete space U_{hp}

As for the mixed hybrid Helmholtz equation, we remark that if the solution can not be resolved by the discrete spaces, this conservation property for $\alpha = 0$ can cause spurious modes. We demonstrate this with the example from Figure 5.4. There on a cylindric computational domain the mesh size and the polynomial order were chosen such that a wave propagating through the cylinder can not be resolved. Taking $\alpha = 0$ leads to spurious modes (compare left hand plot), while for $\alpha \neq 0$ the additional term damps these modes (compare left hand plot).

Chapter 6

Iterative solvers for the mixed hybrid formulation

When solving a partial differential equation with the finite element method, a very challenging part is to solve the resulting linear system of equations

$$A\boldsymbol{x} = \boldsymbol{b} \tag{6.1}$$

with $A \in \mathbb{C}^{N\times N}$, and $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{C}^N$. In general, efficient solvers have to find a good compromise between two criteria, a short computational time and small memory consumption. Both criteria depend strongly on the properties of the system matrix A. The difficulty with the Helmholtz equation is that it leads to a system matrix which is indefinite and complex in the presence of absorbing boundary conditions or lossy media. The indefiniteness explains best the lack of good preconditioners. Furthermore, as already mentioned, the solution is oscillatory with a wavelength inversely proportional to the angular frequency ω. Since in numerical applications one is mainly interested in large frequencies, a large number of unknowns N is needed and numerical solvers are expensive.

When applying a direct solver to (6.1), the matrix is most of the time brought to a form which can be solved easier. For example in Gauss elimination the matrix is transformed to an upper triangular one, and LU or Cholesky factorization decompose A into a product of two triangular matrices. The resulting systems can be solved simply by backward substitution. One bottle neck of direct solvers is most of the time their memory consumption. The number of non-zero entries for matrices obtained via a finite element discretization is proportional to the number of unknowns, i.e. the matrix is sparse. Transforming such a matrix to triangular form destroys some of the sparsity pattern, and many zeros are exchanged by non-zeros. Additionally, a direct solver requires at least $\mathcal{O}(N \log N)$ operations in two dimensions and $\mathcal{O}(N^2)$ operations in three dimensions which causes long computational times for a large number of unknowns.

An iterative solver constructs a sequence of approximate solutions which converges against the exact one. One big advantage is that instead of the whole matrix A just its application to a vector $\boldsymbol{y}$ is needed, i.e. we have only to know how $A\boldsymbol{y}$ is calculated for any $\boldsymbol{y}$, and memory can be saved. For iterative methods mostly the computational times and the number of iterations needed to get an accurate approximation, respectively, is the challenging criterion. Especially, wave type problems suffer from large iteration counts which even increase with growing frequency. The iteration number of an iterative solver depends strongly on the condition number of the matrix A. Therefore, it is useful to improve the properties of the matrix A by preconditioning, which can be interpreted as multiplying (6.1) with a preconditioner matrix C^{-1},

$$C^{-1}A\boldsymbol{x} = C^{-1}\boldsymbol{b}.$$

A good preconditioner has to satisfy two conditions. On the one hand, an application of C^{-1} has to be cheap. Thus, the computation of $C^{-1}\boldsymbol{y}$, which corresponds to solve $C\boldsymbol{x} = \boldsymbol{y}$, should be inexpensive. On the other hand, C has to match A as well as possible, i.e. the condition number of $C^{-1}A$ has to be small.

In this chapter, which is organized as follows, new preconditioners for the wave equation are presented. The first section gives an overview on existing iterative solvers, and our preconditioners are put into the context of these methods. Section 6.2 is devoted to the topic of static condensation. In the Sections 6.3 and 6.4 Schwarz and BDDC preconditioners, respectively, are introduced. There, we discuss how these preconditioners need to be adapted to the mixed hybrid formulation such that convergent schemes are obtained. A new Robin type domain decomposition preconditioner is described in Section 6.5. Finally, in Section 6.6 these preconditioners are compared with the help of numerical examples, and it is demonstrated that the new Robin type domain decomposition preconditioner as well as the BDDC preconditioner are well suited to solve problems with high wave numbers which is the main result of the thesis.

The preconditioners and some of the results we already published in [HPS11] and in [HS12].

6.1 State of the art

Many standard preconditioners, like a Gauss-Seidel preconditioner or the incomplete LU factorization, which have good convergence properties for the Poisson equation turn out to fail when they are applied to the indefinite linear system of equations arising from a discretization of the wave equation. Even the multigrid method, which is known for its robustness and efficiency in the elliptic case, is ineffective for the Helmholtz equation. Convergence of multigrid can be achieved by combining it with a Krylov subspace solver at the price of a rapidly growing iteration number with growing frequency.

Krylov subspace solvers are characterized by the fact that they search for an approximation in the so called Krylov space $K_n = \operatorname{span}\{\boldsymbol{b}, A\boldsymbol{b}, A^2\boldsymbol{b}, \ldots, A^{n-1}\boldsymbol{b}\}$. Most popular among them are CG (conjugate gradient), GMRES (generalized minimum residual), QMR (quasi minimal residual) or Bi-CGSTAB (biconjugate gradient stabilized). For further reading on Krylov space solvers we recommend the book [vdV03].

The bad performance of standard preconditioners when applied to the wave equations has led to the development of a variety of methods designed especially for this problem type. For overviews on such methods see [EG12, Erl08].

One popular possibility is the shifted Laplace preconditioner [Erl08] combined with a Krylov space iteration which was first introduced in [EVO04], and which is based on an idea of [BGT83]. The preconditioner is obtained from a discretization of the Helmholtz equation with a rescaled complex wave number, i.e. the corresponding operator reads as

$$\mathcal{C} := -\Delta - (\beta_1 + i\beta_2)\omega^2.$$

A delicate task is to find a good choice for the parameters β_1 and especially for the complex shift β_2. On the one hand, a small value of β_2 (for $\beta_1 \approx 1$) is needed to cluster the spectrum of the preconditioned operator away from zero which is essential for the convergence of Krylov space iterations. On the other hand, it can be shown that for a large value of β_2 standard multigrid methods work, and the cost of the application of the preconditioner can be reduced. In [Erl08] it was suggested to use $(\beta_1, \beta_2) = (1, 0.5)$. Although the dependence of the number of iterations on the mesh size is removed for this preconditioner, it faces difficulties for absorbing boundary conditions and for high frequency problems due to the strong dependency of the number of iterations on ω. Nevertheless, some interesting three dimensional examples are presented in [RKE+07].

A relatively new preconditioner for the scalar and the vector valued wave equation is the sweeping preconditioner [EY11a, EY11b, TEY12] which is based on an approximate block $LDL^\top$ factorization of the discrete Helmholtz operator. In sweeping preconditioners the elimination process of the unknowns is done layer by layer, starting with an absorbing boundary layer. The preconditioner can be obtained by an approximation of the layer Schur complement matrix via the hierarchical matrix framework [EY11a], or it is represented by moving perfectly matched layers (PML) in the interior of the domain [EY11b, TEY12]. An application of the preconditioner corresponds for the latter case to the solution of a problem with a by one reduced spatial dimension which is realized either by a banded LU factorization (2D) or by a multifrontal method (3D). The preconditioner was till now just tested for a finite difference method and low order finite elements with PML boundary conditions and smoothly varying coefficients, and it remains rather unclear how it performs with other boundary conditions, coefficients with jumps and higher order discretizations. However, numerical results in the given references indicate that small iteration numbers can be reached which are almost independent of the mesh size and the frequency. Nevertheless, we should note that the setup times for both preconditioners are rather large, especially for large values of ω. This is even worsened by the fact that the structure of the preconditioner is sequential, and it is therefore rather difficult to benefit from a parallel architecture.

Much better suited for parallel computations are domain decomposition preconditioners. There, the original computational domain is partitioned into subdomains which may be much smaller than the original one. Domain decomposition preconditioners turn out to be very competitive if one is able to solve the underlying problems efficiently on these subdomains. When applying them, the original problem is reduced by solving it on the subdomains to an interface problem which is much smaller in size. Note that this can be done very efficiently in parallel. The fact that the interface problem often has better iterative properties than the full problem makes it, together with the reduced size of the interface problem, very attractive in order to solve the Helmholtz equation. There, the right choice of interface conditions is rather delicate. Taking the nodal values as unknowns, as it is done for the Laplace equation, turns out to have bad iteration properties, while transmission conditions, i.e. conditions on the impedance traces, lead to convergent schemes [Des91]. For completeness, we mention that some effort was put in finding better transmission conditions [GMN02]. They often turn out to be non local in nature, and local approximations are required.

One of the most popular domain decomposition methods for solving the Helmholtz equation is the FETI-H [FML00, TMF01] (finite element tearing and interconnecting -

Helmholtz) method which can be seen as an extension of the FETI method introduced in [FR91]. The main idea of the FETI method is to decompose the computational domain into non overlapping subdomains, which are treated separately, including their subdomain boundary. A set of Lagrangian multipliers guarantees the continuity across the interfaces. The resulting saddle point problem is solved iteratively via its dual problem containing just the introduced multipliers. This dual problem is frequently preconditioned by local Neumann or Dirichlet problems.

There are two major differences between the FETI-H method and a conventional FETI method. One is that in the FETI-H method singularities of the local subdomain problems are avoided by regularizing them with an interface mass matrix. This can be interpreted as equipping the local problems with one Dirichlet and one transmission interface condition. Furthermore, the resulting dual interface problem for the Lagrangian multipliers is preconditioned with an auxiliary coarse problem. With the help of this coarse problem the residual is orthogonalized in each iteration step with respect to a small set of carefully chosen plane waves on the interface. Numerical experiments from [FML00] indicate that the iteration number for the FETI-H method is at least independent from the mesh size and the number of subdomains. By adapting the number of plane waves, the number of iterations can be kept constant with increasing frequency. The drawback is that this leads to a larger coarse grid problem, and the cost per iteration grows.

Closely related to FETI-H is FETI-DPH [FATL05], a FETI-DP (FETI dual primal) method [FLP00] which is specialized for Helmholtz problems. In FETI-DP, a further development of FETI, the interface unknowns are divided into dual and primal unknowns. While continuity between the dual unknowns is still enforced by Lagrangian multipliers, the primal degrees of freedom, which usually are related to cross nodes, i.e. nodes belonging to three or more subdomains, are kept as global degrees of freedom, and they can be interpreted as coarse space components. Applying this technique, the local subdomain problem in FETI-DPH containing just inner and dual unknowns is non-singular, and a regularization by an interface mass matrix like for FETI-H is not needed. Because the coarse problem in FETI-DP does not contain any specific character of the Helmholtz equation, it is augmented in the FETI-DPH method with the help of a set of plane waves. By solving the larger coarse problem, the residual is orthogonalized additionally on the interface with respect to these plane waves. Finally, by applying a local Dirichlet preconditioner, scalability of the iterative method with respect to the problem size can be obtained. Numerical examples in [FATL05] indicate that the FETI-DPH is significantly faster than the FETI-

H method, although both methods have similar scalability properties concerning to mesh size, number of subdomains and wavenumber. The reference also includes some impressive large scale computations in three dimension.

Our work in the context of existing literature

Our work deals with domain decomposition preconditioners. They will be applied to a linear system of equations just for the facet unknowns which is obtained after eliminating the volume unknowns. Note that this linear system of equations is just related to the skeleton of a mesh, and a domain decomposition of the skeleton is induced by a decomposition of the underlying mesh. Since impedance traces are obtained from the facet functions by a simple transformation of variables (compare Definitions 5.1 and 5.9), transmission conditions on the interface in the sense of [Des91] can be enforced by guaranteeing the same value of the facet unknowns of different subdomains on the subdomain interface. Thus, our mixed hybrid formulation allows in a natural way for appropriate transmission conditions.

In this chapter Schwarz preconditioners and a BDDC preconditioner are adapted to the system of equations for the skeleton variables, and a new Robin type domain decomposition preconditioner is introduced. In order to obtain efficient solvers, these preconditioners are combined with a Krylov space solver. Because our problem is complex symmetric, the method of choice would be a GMRES iteration. There, the vector $\boldsymbol{x} \in K_n$ which minimizes the euclidean norm of the residual is selected as approximation. The GMRES method requires storing the full basis of the Krylov space which makes it more and more expensive with increasing iteration number. Due to this performance argument a conjugate gradient (CG) iteration was used, although there exists no convergence theory for complex symmetric problems.

6.2 Static condensation

All the preconditioners we will introduce in the following sections are applied to a system of equations for the facet unknowns. Thus, the first step in the solution process is to eliminate via static condensation [AB85] the volume unknowns in the system of equations

$$\begin{pmatrix} B & D \\ D^\top & C \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{inner} \\ \boldsymbol{x}_{facet} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{pmatrix} \tag{6.2}$$

obtained by discretizing the Formulations 5.2 and 5.10, respectively. In (6.2) we already distinguish between the different roles the unknown play. In $\boldsymbol{x}_{inner}$ the volume unknowns u, $\boldsymbol{\sigma}$ and $\boldsymbol{E}$, $\boldsymbol{H}$, respectively, are collected, while the degrees of freedom supported only on the facets, u_F, σ_F and $\boldsymbol{E}_F$, $\boldsymbol{H}_F$, respectively, are contained in $\boldsymbol{x}_{facet}$. B collects the coupling entries among the volume degrees of freedom, D couples the volume degrees of freedom with the facet degrees of freedom and so on.

Elimination of the volume unknowns is now equivalent to building the Schur complement matrix S, i.e. the reduced system of equation reads as

$$S\boldsymbol{x}_{inner} = \boldsymbol{b} \qquad \text{with} \qquad S = C - D^\top B^{-1} D. \tag{6.3}$$

For this reduced system containing just the facet unknowns $\boldsymbol{x}_{facet}$ we will introduce our preconditioners. Note that constructing the Schur complement matrix requires the inversion of the matrix B. Due to the broken continuity across element interfaces in the mixed hybrid formulation the volume basis functions can be chosen such that they are supported only on one single element. Thus, there is no coupling between volume degrees of freedom belonging to different elements, and B is block diagonal with a block size equal to the number of volume unknowns on one element. Therefore, the inversion and static condensation, respectively, can be done cheaply element by element. After solving the Schur complement system, the solution on the elements can be reconstructed again on the element level by evaluating $\boldsymbol{x}_{inner} = -B^{-1}D\boldsymbol{x}_{facet}$.

Remark 6.1. *In an actual implementation the Schur complement is already calculated for each element matrix during the assembly procedure, and S is constructed by assembling these element matrix Schur complements. Thus, the matrices B and D are not needed, and they do not have to be stored which saves a lot of memory. Note that for such an implementation the element matrices have to be recalculated when the volume solution is reconstructed.*

Remark 6.2. *Finally we want to remark that static condensation for the mixed hybrid system fits well into the concept of domain decomposition which was shortly explained in the introduction to this chapter. Here, the elements can be interpreted as subdomains, and computing the Schur complement is now equivalent to solve a subdomain or element problem, respectively. The resulting interface system of equations is the Schur complement equation. The constraints fixing the incoming and outgoing impedance traces can be obtained by a simple change of variables on the facet.*

6.3 Schwarz preconditioners

The theory for this type of domain decomposition preconditioner goes back to the pioneering work of Schwarz [Sch70] from the 19th century. Schwarz methods are based on a division of the finite element spaces into a set of possibly overlapping subspaces. One possibility for such a division is to define the subspaces via coarse grids which leads to multilevel methods. We will use in the following a splitting induced by a division of the computational domain into subdomains. For a detailed discussion on these preconditioners we recommend the books [TW05] and [SBG96].

6.3.1 Additive Schwarz preconditioners

However, independent of how the splitting is induced, we assume, in order to describe the basics of Additive Schwarz preconditioners, a set of closed subspaces $\{W_i, i = 1, \dots, N\}$ of the finite element space W. In the following we consider the finite element problem

$$a(u, v) = f(v) \qquad \forall v \in W \tag{6.4}$$

with a positive definite bilinear form $a : W \times W \to \mathbb{R}$ and the linear form $f : W \to \mathbb{R}$. This equation is equivalent to a linear operator equation in the dual space W^* of the space W,

$$\mathcal{A}u = F \tag{6.5}$$

with $u \in W$. Denoting the dual product in $W^* \times W$ by $\langle\, , \rangle$, $F \in W^*$ is defined via $\langle F, v\rangle = f(v)$ for $v \in W$ and the operator $\mathcal{A} : W \to W^*$ by $\langle \mathcal{A}u, v\rangle = a(u, v)$ for all $v \in W$.

The operator $\mathcal{A}_i : W_i \to W_i^*$ corresponding to $\mathcal{A}$ on a subspace W_i is given via

$$\langle \mathcal{A}_i w, \phi\rangle = a(w, \phi) \qquad \forall \phi \in W_i. \tag{6.6}$$

On each subspace W_i, we assume a positive definite bilinear form $c_i : W_i \times W_i \to \mathbb{R}$ which should approximate there the bilinear form a. This leads in operator notation to the operator $\mathcal{C}_i : W_i \to W_i^*$,

$$\langle \mathcal{C}_i \psi, \phi\rangle = c_i(\psi, \phi) \qquad \text{for } \psi, \phi \in W_i. \tag{6.7}$$

The operator $\mathcal{C}_i$ approximates $\mathcal{A}_i$, or in other words, $\mathcal{C}_i^{-1}$ should approximately invert $\mathcal{A}_i$ and provide a local solution on W_i. With the help of the orthogonal projection $\mathcal{Q}_i : W^* \to W_i^*$ onto the subspace W_i^*, i.e. for $w \in W^*$

$$\langle \mathcal{Q}_i w, \phi \rangle = \langle w, \phi \rangle \qquad \forall \phi \in W_i, \tag{6.8}$$

we are able to search for the local solution $u_i \in W_i$ via

$$c_i(u_i, \phi) = \langle \mathcal{C}_i u_i, \phi \rangle = \langle \mathcal{Q}_i \mathcal{A} u, \phi \rangle \qquad \forall \phi \in W_i. \tag{6.9}$$

Thus, by

$$u_i := \mathcal{T}_i u \qquad \text{with} \qquad \mathcal{T}_i = \mathcal{C}_i^{-1} \mathcal{Q}_i \mathcal{A}. \tag{6.10}$$

the local solution operator $\mathcal{T}_i$ is prescribed. In an Additive Schwarz method the solution operator $\mathcal{T}$ of the whole problem is assumed to be the sum of the local solution operators,

$$\mathcal{T} = \sum_{i=1}^{N} \mathcal{T}_i = \sum_{i=1}^{N} \mathcal{C}_i^{-1} \mathcal{Q}_i \mathcal{A} =: \mathcal{C}^{-1} \mathcal{A}$$

which gives rise to the definition of the Additive Schwarz preconditioner

$$\mathcal{C}^{-1} = \sum_{i=1}^{N} \mathcal{C}_i^{-1} \mathcal{Q}_i.$$

With the help of the variational projector $\mathcal{P}_i : W \to W_i$ given by

$$a(u, \phi) = a(\mathcal{P}_i u, \phi) \qquad \forall \phi \in W_i \tag{6.11}$$

for $u \in W$, we conclude from the relation

$$\langle \mathcal{A} u, \phi \rangle = a(u, \phi) = a(\mathcal{P}_i u, \phi) = \langle \mathcal{A}_i \mathcal{P}_i u, \phi \rangle \qquad \forall \phi \in W_i$$

that $\mathcal{A}_i \mathcal{P}_i u$ is the orthogonal projection of $\mathcal{A} u$, i.e.

$$\mathcal{Q}_i \mathcal{A} = \mathcal{A}_i \mathcal{P}_i \tag{6.12}$$

and consequently $\mathcal{T} = \sum_{i=1}^{N} \mathcal{C}_i^{-1} \mathcal{A}_i \mathcal{P}_i$. Note that because of the positive definiteness of the bilinear form the variational projector $\mathcal{P}_i$ is well defined.

Remark 6.3. *If the local problem is solved exactly which means $\mathcal{C}_i = \mathcal{A}_i$, the solution operator $\mathcal{T}$ is the sum of the variational projectors, $\mathcal{T} = \sum_{i=1}^{N} \mathcal{P}_i$. Thus, if T, C, A and P_i are the matrix representations of the operators $\mathcal{T}, \mathcal{C}, \mathcal{A}$ and $\mathcal{P}_i$, the iteration matrix reads as*

$$I - C^{-1}A = I - T = I - \sum_{i=1}^{N} P_i.$$

The preconditioner corresponds to a block Jacobi preconditioner where one block in the preconditioner matrix C contains the coupling entries among the degrees of freedom of one subspace.

Remark 6.4. *In the formal presentation of the Additive Schwarz preconditioner we neglected, in order to keep it simple, an overlapping splitting of the space W into the subspaces W_i. If we allow for an overlapping splitting, prolongation operators $\mathcal{R}_i : W_i \to W$ have to be introduced, such that $W = \sum_{i=1}^{N} \mathcal{R}_i W_i$. The preconditioner reads then as*

$$\mathcal{C}^{-1} = \sum_{i=1}^{N} \mathcal{R}_i \mathcal{C}_i^{-1} \mathcal{R}_i^{\top}$$

with the adjoint operator $\mathcal{R}_i^{\top}$.

As it was already argued, in order to get good convergence rates of iterative solvers, the condition number $\kappa(C^{-1}A)$ of the matrix $C^{-1}A$, where C and A are the matrix representations of the operators $\mathcal{C}$ and $\mathcal{A}$, has to be small. For example the preconditioned CG iteration converges with a rate of $\frac{\sqrt{\kappa(C^{-1}A)}-1}{\sqrt{\kappa(C^{-1}A)}+1}$. For approximating $\kappa(C^{-1}A)$ of the Additive Schwarz preconditioner with an exact subdomain solver, i.e. $\mathcal{C}_i = \mathcal{A}_i$, the following Lemma (compare Lemma 2.5 in [TW05]) is important.

Lemma 6.5 (Additive Schwarz Lemma)**.** *Let $\mathcal{C}^{-1}$ be the inverse of a self adjoint and positive definite operator $\mathcal{C} : W \to W^*$, then*

$$\langle \mathcal{C}u, u \rangle = \inf_{u = \sum_{u_i \in W_i} \mathcal{R}_i u_i} \|u_i\|_a^2 \qquad \textit{with } \|v\|_a := \sqrt{a(v,v)}$$

Note that we now allow for overlapping spaces W_i. Thus, the prolongation operator from Remark 6.4 was used. If we are able to find constants γ_1 and γ_2 which fulfill

$$\gamma_1 \|u\|_a^2 \leq \inf_{u = \sum_{u_i \in W_i} \mathcal{R}_i u_i} \|u_i\|_a^2 \leq \gamma_2 \|u\|_a^2,$$

the Additive Schwarz lemma gives spectral equivalence of C and A, and the condition number can be estimated by

$$\kappa(C^{-1}A) \leq \frac{\gamma_2}{\gamma_1}.$$

Bounds for γ_1 and γ_2 can be found in [TW05] or [BS08]. The constant γ_2 is called constant of stable splitting, and if only a finite number of spaces $\mathcal{R}_i V_i$ overlap, the constant γ_1 is of the order of $\mathcal{O}(1)$.

6.3.2 Multiplicative Schwarz preconditioners

The Multiplicative Schwarz preconditioner is closely related to the Additive Schwarz preconditioner described in the last section. For a detailed discussion, we refer, apart from the beforehand mentioned books [TW05] and [SBG96], to [BZ00]. In this section we will follow the presentation of [BPWX91].

We will again stick to the notation of the last section. Thus, we consider the finite element problem $a(u, v) = f(v)$ from (6.4) which corresponds to the operator equation $\mathcal{A}u = F$. We assume a splitting of the space W into N subspaces W_i, and the restriction of the operator $\mathcal{A}$ to W_i we call $\mathcal{A}_i$ (compare(6.6)). Furthermore, the operator $\mathcal{C}_i$ is an approximation of $\mathcal{A}_i$ and by $\mathcal{T}_i$ we define the local solution operator from (6.10). The orthogonal and variational projectors we denote as in the equations (6.8) and (6.11) as $\mathcal{Q}_i$ and $\mathcal{P}_i$.

We start by describing the application of the Multiplicative Schwarz preconditioner. Thus, finding an approximation $\tilde{u} \in W$ for some initial guess $u_0 \in W$ corresponds to the steps

1) Set $v_0 = u_0$.

2) Compute for $i = 1, \ldots, N$ v_i by

$$v_i = v_{i-1} + \mathcal{C}_i^{-1}\mathcal{Q}_i(F - \mathcal{A}v_{i-1}).$$

3) Set $\tilde{u} = v_N$.

If we denote the error in the iteration step i as $e_i = u - v_i$, it is easy to verify for $i \geq 1$

that

$$\begin{aligned} e_i &= (\mathcal{I} - \mathcal{C}_i^{-1}\mathcal{Q}_i\mathcal{A})\ e_{i-1} \\ &\overset{(6.12)}{=} (\mathcal{I} - \mathcal{C}_i^{-1}\mathcal{A}_i\mathcal{P}_i)\ e_{i-1} \\ &= (\mathcal{I} - \mathcal{T}_i)\ e_{i-1} \end{aligned}$$

with the identity $\mathcal{I}$. Consequently, we get

$$u - \tilde{u} = (\mathcal{I} - \mathcal{T}_N)\dots(\mathcal{I} - \mathcal{T}_1)(u - u_0) =: \prod_{i=1}^{N}(\mathcal{I} - \mathcal{T}_i)(u - u_0),$$

and the solution operator $\mathcal{T}$ reads as

$$\mathcal{T} := \mathcal{I} - \Big(\prod_{i=1}^{N}(\mathcal{I} - \mathcal{T}_i)\Big)$$

Remark 6.6. *If the local problem is solved exactly, thus $\mathcal{C}_i = \mathcal{A}_i$, $\mathcal{T}_i$ can be exchanged by the variational projector $\mathcal{P}_i$ and $(\mathcal{I} - \mathcal{T}) = \prod_{i=1}^{N}(\mathcal{I} - \mathcal{P}_i)$. For the matrix representations A, C, T, P_i of the operators $\mathcal{A}, \mathcal{C}, \mathcal{T}, \mathcal{P}_i$ the iteration matrix is*

$$I - C^{-1}A = I - T = \prod_{i=1}^{N}(I - P_i).$$

Note that for non overlapping subspaces the Multiplicative Schwarz preconditioner matches a block Gauss Seidel preconditioner with blocks corresponding to the subspaces.

Remark 6.7. *We should remark that the preconditioner just introduced is not symmetric and therefore not suitable for a preconditioned CG (PCG) iteration. In order to get a symmetric version, one combines the presented iteration with an iteration using a "backward" numbering of the spaces, i.e.*

$$(\mathcal{I} - \mathcal{T}) = \Big(\prod_{i=N}^{1}(\mathcal{I} - \mathcal{T}_i)\Big)\Big(\prod_{i=1}^{N}(\mathcal{I} - \mathcal{T}_i)\Big).$$

Concerning convergence results, in [TW05] it is shown under the assumption of a stable splitting of W, local stability of a and a strengthened Cauchy Schwarz inequality that the norm of the error propagation operator

$$\mathcal{E} := \mathcal{I} - \mathcal{T},$$

which is the operator corresponding to the iteration matrix, can be bounded from above by a constant smaller than one, i.e.

$$\|\mathcal{E}\|_{\mathcal{A}}^2 := \sup_{u \in W} \frac{\langle \mathcal{A}\mathcal{E}u, \mathcal{E}u \rangle}{\langle \mathcal{A}u, u \rangle} \; < \; 1.$$

Thus, for a positive definite problem a Richardson iteration equipped with a Multiplicative Schwarz preconditioner converges. For the symmetric version from Remark 6.6 it can be shown that the preconditioner does not perform worse than the corresponding Additive method, although it is much better in many applications.

6.3.3 Schwarz methods for the mixed hybrid formulation

We will use Schwarz Methods as domain decomposition preconditioners in a CG solver for the Schur complement problem (6.3). Thus, the space corresponding to W is $U^F \times V^F$ in the Helmholtz case and $X^F \times Y^F$ in the vectorial case. Because the underlying problem is an element interface and surface problem, respectively, a possible choice for the subspaces is induced by a decomposition of the element interfaces. We will use in the numerical results section two different choices for the subspaces W_i. One choice is to split the skeleton into single facets, and consequently W_i contains just functions which are supported only on the facet $F_i \in \mathcal{F}$. Because basis functions in W can be chosen such that they are supported only on one single facet, non overlapping blocks can be obtained. Collecting all the basis functions which are supported on facets of the element $T_i \in \mathcal{T}$ in the subspace W_i leads to overlapping blocks.

We should finally remark that the Additive Schwarz preconditioner can be easily implemented in a parallel code. There, the computational domain is split into subdomains, and each processor owns the degrees of freedom of one subdomain, or more precisely the degrees of freedom on facets of elements belonging to one subdomain. For Multiplicative Schwarz preconditioners parallelization is much more complicated.

6.4 The BDDC preconditioner

This section is devoted to the BDDC (balanced domain decomposition by constraints) preconditioner introduced by Dohrmann [Doh03a, Doh03b]. It can be seen as a further development of BDD [Man93] algorithms and fits into the context of balancing Neumann Neumann methods [TW05] with a coarse component representing primal constraints. The basic idea is to divide the computational domain into subdomains and to remove in parallel by building a Schur complement matrix the inner degrees of freedom on the subdomain. The resulting system of equations for the interface unknowns contains just a small number of primal variables which are the only global variables. These primal variables are nodal variables, or they represent constraints which should prevent floating of the subdomains, like mean values on the interface facets.

In the description of the BDDC preconditioner we will follow the work of [LW06].

6.4.1 Theoretical framework

Because we want to formulate the BDDC preconditioner in the context of block Cholesky factorization, we take a closer look onto this method.

Block Cholesky elimination

Therefore, we consider a symmetric positive definite block matrix which can be decomposed as

$$\begin{pmatrix} B & D^\top \\ D & C \end{pmatrix} = \begin{pmatrix} I_B & \\ DB^{-1} & I_C \end{pmatrix} \begin{pmatrix} B & \\ & C - DB^{-1}D^\top \end{pmatrix} \begin{pmatrix} I_B & B^{-1}D^\top \\ & I_C \end{pmatrix} \tag{6.13}$$

where I_B and I_C are unity matrices of the dimension of B and C, and $C - DB^{-1}D^\top$ is the Schur complement matrix we will denote as S. Note that S is positive definite as well. Inverting the block matrix leads to

$$\begin{aligned} \begin{pmatrix} B & D^\top \\ D & C \end{pmatrix}^{-1} &= \begin{pmatrix} I_B & -B^{-1}D^\top \\ & I_C \end{pmatrix} \begin{pmatrix} B^{-1} & \\ & S^{-1} \end{pmatrix} \begin{pmatrix} I_B & \\ -DB^{-1} & I_C \end{pmatrix} \\ &= \begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \Phi S^{-1} \Phi^\top, \qquad \text{with } \Phi = \begin{pmatrix} -B^{-1}D^\top \\ I_C \end{pmatrix}. \end{aligned} \tag{6.14}$$

In the following discussion we have a problem in mind which is divided into subproblems induced by spitting the computational domain into subdomains. The matrix A represents apart from the inner degrees of freedom on the subdomain the dual ones on the interfaces, and it can be written as the direct sum of its subdomain contributions. The C block represents the primal unknowns on the interface which are global. The matrix Φ can be interpreted as the extensions of the canonical basis on the set of the primal unknowns to the local degrees of freedom of the B block. Thus, Φ extends any coarse function to the subdomains.

Primal and dual degrees of freedom

In order to describe the BDDC preconditioner, we choose a positive definite problem on a computational domain Ω where at least on some part of the boundary Γ_D Dirichlet boundary conditions are imposed. This domain is divided into N subdomains Ω_i. The interface of these subdomains we denote by $\Sigma := \left(\bigcup_{i\neq j} \partial\Omega_i \cap \partial\Omega_j\right) \setminus \Gamma_D$. On each subdomain Ω_i we can assemble for the local solution vector $u^{(i)}$ subdomain contributions $A^{(i)}$, $f^{(i)}$ to the global system matrix A and right hand side f,

$$A^{(i)} = \begin{pmatrix} A_{II}^{(i)} & A_{\Sigma I}^{(i)\top} \\ A_{\Sigma I}^{(i)} & A_{\Sigma\Sigma}^{(i)} \end{pmatrix}, \quad f^{(i)} = \begin{pmatrix} f_I^{(i)} \\ f_\Sigma^{(i)} \end{pmatrix}, \quad u^{(i)} = \begin{pmatrix} u_I^{(i)} \\ u_\Sigma^{(i)} \end{pmatrix}.$$

We already distinguish between inner and interface unknowns by the subscripts I and Σ. Note that the interface unknowns belong to several processors. With the help of the restriction operators $R_\Sigma^{(i)}$ which restricts a vector v_Σ with interface entries to a vector $v_\Sigma^{(i)}$ just containing entries which belong to the interface unknowns on Ω_i, we get the global system of equations

$$\begin{pmatrix} A_{II}^{(1)} & & & \widetilde{A}_{\Sigma I}^{(1)\top} \\ & \ddots & & \vdots \\ & & A_{II}^{(N)} & \widetilde{A}_{\Sigma I}^{(N)\top} \\ \widetilde{A}_{\Sigma I}^{(1)} & \dots & \widetilde{A}^{(N)} & \widetilde{A}_{\Sigma\Sigma} \end{pmatrix} \begin{pmatrix} u_I^{(1)} \\ \vdots \\ u_I^{(N)} \\ u_\Sigma \end{pmatrix} = \begin{pmatrix} f_I^{(1)} \\ \vdots \\ f_I^{(N)} \\ \widetilde{f}_\Sigma \end{pmatrix}$$

with

$$\widetilde{A}_{\Sigma I}^{(i)} := R_\Sigma^{(i)\top} A_{\Sigma I}^{(i)}, \qquad \widetilde{A}_{\Sigma\Sigma} := \sum_{i=1}^{N} R_\Sigma^{(i)\top} A_{\Sigma\Sigma}^{(i)} R_\Sigma^{(i)}, \qquad \widetilde{f}_\Sigma := \sum_{i=1}^{N} R_\Sigma^{(i)\top} f_\Sigma^{(i)}.$$

By eliminating the inner unknowns on each subdomain, i.e. by building the subdomain Schur complement

$$S^{(i)} := A^{(i)}_{\Sigma\Sigma} - A^{(i)}_{\Sigma I} A^{(i)-1}_{II} A^{(i)\top}_{\Sigma I}, \qquad g^{(i)}_{\Sigma} = f^{(i)}_{\Sigma} - A^{(i)}_{\Sigma I} A^{(i)-1}_{II} f^{(i)}_{I},$$

the problem can be reduced to the interface

$$\Big(\sum_{i=1}^{N} R^{(i)\top}_{\Sigma} S^{(i)} R^{(i)}_{\Sigma} \Big) u_{\Sigma} = \sum_{i=1}^{N} R^{(i)\top}_{\Sigma} g^{(i)}_{\Sigma}.$$

It would be tempting to precondition this system with the sum of the inverses of the local Schur complement. But for subdomains which are not located at Γ_D inverting $S^{(i)}$ corresponds to solving a Neumann problem, and consequently $S^{(i)}$ is singular. This drawback can be eliminated by posing additional constraints, like fixing the mean value of u on interface facets. In [LW06] the authors show that including such constraints corresponds to fixing several unknowns when a change of variables is applied. These fixed interface unknowns, called primal unknowns, we will treat as global ones and they get the subscript Π. The other interface unknowns, the dual degrees of freedom which get the subscript Δ, we consider to be local.

Thus, if we introduce a restriction operator $R^{(i)}_{\Pi}$, which restricts the whole set of primal degrees of freedom u_{Π} to the primal degrees of freedom supported on the subdomain Ω_i, our system of equation reads as

$$\begin{pmatrix} A^{(1)}_{II} & A^{(1)\top}_{\Delta I} & & & & \widetilde{A}^{(1)\top}_{\Pi I} \\ A^{(1)}_{\Delta I} & A^{(1)}_{\Delta\Delta} & & & & \widetilde{A}^{(1)\top}_{\Pi\Delta} \\ & & \ddots & & & \vdots \\ & & & A^{(N)}_{II} & A^{(N)\top}_{\Delta I} & \widetilde{A}^{(N)\top}_{\Pi I} \\ & & & A^{(N)}_{\Delta I} & A^{(N)}_{\Delta\Delta} & \widetilde{A}^{(N)\top}_{\Pi\Delta} \\ \widetilde{A}^{(N)}_{\Pi I} & \widetilde{A}^{(N)}_{\Pi\Delta} & \dots & \widetilde{A}^{(N)}_{\Pi I} & \widetilde{A}^{(N)}_{\Pi\Delta} & \widetilde{A}_{\Pi\Pi} \end{pmatrix} \begin{pmatrix} u^{(1)}_{I} \\ u^{(1)}_{\Delta} \\ \vdots \\ u^{(N)}_{I} \\ u^{(N)}_{\Delta} \\ u_{\Pi} \end{pmatrix} = \begin{pmatrix} f^{(1)}_{I} \\ f^{(1)}_{\Delta} \\ \vdots \\ u^{(N)}_{I} \\ u^{(N)}_{\Delta} \\ \sum_{i=1}^{N} R^{(i)\top}_{\Pi} f^{(i)}_{\Pi} \end{pmatrix} \tag{6.15}$$

with

$$\widetilde{A}^{(i)}_{\Pi I} := R^{(i)\top}_{\Pi} A^{(i)}_{\Pi I}, \qquad \widetilde{A}^{(i)}_{\Pi\Delta} := R^{(i)\top}_{\Pi} A^{(i)}_{\Pi\Delta}, \qquad \widetilde{A}_{\Pi\Pi} := \sum_{i=1}^{N} R^{(i)\top}_{\Pi} A^{(i)}_{\Pi\Pi} R^{(i)}_{\Pi}.$$

Note that for a solution $u^{(i)}_{\Delta} = u^{(j)}_{\Delta}$ for the commonly owned unknowns is required. For the BDDC preconditioner this system is solved. This leads in fact to a continuous solution for the primal variables, but in general $u^{(i)}_{\Delta} \neq u^{(j)}_{\Delta}$ for the unknowns supported on Ω_i and Ω_j.

Continuity is then regained by taking a weighted average.

The preconditioner

Based on this notation and additional restriction operators, we are able to define the BDDC preconditioner. So, $R_{\Delta}^{(i)}$ restricts the interface unknowns to the dual unknowns of the domain Ω_i, and R_{Σ} is the direct sum of the $R_{\Sigma}^{(i)}$. Thus, applying R_{Σ} to an interface vector v_{Σ} leads to $\left(v_{\Sigma}^{(1)}, \ldots, v_{\Sigma}^{(N)}\right)^{\top}$. Furthermore, $R_{D\Sigma}^{(i)}$ is the scaled version of $R_{\Sigma}^{(i)}$, i.e. if a degree of freedom is apart from Ω_i supported on n other subdomain, the corresponding vector entry is divided by $(n+1)$ when it is taken. Finally, $R_{D\Sigma}$ is the direct sum of the $R_{D\Sigma}^{(i)}$.

Our system of equations, i.e. the linear system for the interface unknowns obtained from (6.15) by eliminating the inner unknowns, is of a block diagonal form comparable to (6.13). The dual degrees of freedom correspond as already indicated to the unknowns related to the block B and the primal ones to the degrees of freedom of the block C. Therefore, it makes sense to define a preconditioner with a structure comparable to (6.14). Thus, the BDDC preconditioner can be written as

$$C_{BDDC}^{-1} = R_{D\Sigma}^{\top}(T_{sub} + T_0)R_{D\Sigma}$$

with a subdomain correction T_{sub} for the dual unknowns, which corresponds to the first summand in (6.14). Because T_{sub} can be applied subdomain wise, the actual residual is split into the subdomain contributions by $R_{D\Sigma}$ and $R_{D\Sigma}^{\top}$ averages the dual degrees of freedom in order to regain continuity. Thus, when applying T_{sub} the inverse of the block corresponding to the dual unknowns of the subdomain, i.e. the inverse of the Schur complement

$$S_{\Delta}^{(i)} := A_{\Delta\Delta}^{(i)} - A_{\Delta_I}^{(i)} A_{II}^{(i)-1} A_{\Delta I}^{(i)\top}$$

has to be applied to the $u_{\Delta}^{(i)}$. With the help of the restriction operators this gives

$$T_{sub} = S_{\Delta}^{-1} := \sum_{i=1}^{N} R_{\Delta}^{(i)\top} S_{\Delta}^{(i)-1} R_{\Delta}^{(i)}.$$

The corse grid correction T_0 corresponds to the term $\Phi S^{-1} \Phi^{\top}$ in (6.14). Because the operator $R_{D\Sigma}$ decomposes a residual into subdomain contributions, Φ can be split as well into subdomain contributions. This distributed version of Φ we call $\Psi := \left(\Psi^{(1)}, \ldots, \Psi^{(N)}\right)^{\top}$.

Because not each primal degree of freedom is supported on a subdomain Ω_i the contributions $\Psi^{(i)}$ can be calculated according to the definition of Φ in (6.14) locally on the subdomain by exchanging the identity matrix with the operator $R_\Pi^{(i)}$ which provides zero columns for not supported primal degrees of freedom. By inserting into the definition, we get

$$\Psi^{(i)} = \begin{pmatrix} \Psi_\Delta^{(i)} \\ R_\Pi^{(i)} \end{pmatrix} = \begin{pmatrix} -\begin{pmatrix} 0 & I_\Delta^{(i)} \end{pmatrix} \begin{pmatrix} A_{II}^{(i)} & A_{\Delta I}^{(i)\top} \\ A_{\Delta I}^{(i)} & A_{\Delta\Delta}^{(i)} \end{pmatrix}^{-1} \begin{pmatrix} A_{\Pi I}^{(i)\top} \\ A_{\Pi\Delta}^{(i)\top} \end{pmatrix} R_\Pi^{(i)} \\ R_\Pi^{(i)} \end{pmatrix}.$$

Here $I_\Delta^{(i)}$ is an identity matrix with a dimension of the number of dual unknowns on Ω_i. Note that for primal degrees of freedom not supported on the subdomain $\Psi^{(i)}$ has a zero column.

The matrix S in (6.14) represents the Schur complement matrix $S_{\Pi\Pi}$ with respect to the primal degrees of freedom, and it can be obtained by adding the local Schur complements $S_{\Pi\Pi}^{(i)}$ on the subdomains, i.e.

$$\begin{aligned} S_{\Pi\Pi} &= \sum_{i=1}^{N} R_\Pi^{(i)\top} S_{\Pi\Pi}^{(i)} R_\Pi^{(i)} \\ \text{with} \quad S_{\Pi\Pi}^{(i)} &= A_{\Pi\Pi}^{(i)} - \begin{pmatrix} A_{\Pi I}^{(i)} & A_{\Pi\Delta}^{(i)} \end{pmatrix} \begin{pmatrix} A_{II}^{(i)} & A_{\Delta I}^{(i)\top} \\ A_{\Delta I}^{(i)} & A_{\Delta\Delta}^{(i)} \end{pmatrix}^{-1} \begin{pmatrix} A_{\Pi I}^{(i)\top} \\ A_{\Pi\Delta}^{(i)\top} \end{pmatrix}. \end{aligned}$$

Collecting this, we get for the corse grid correction

$$T_0 = \Psi S_{\Pi\Pi}^{-1} \Psi^\top$$

and

$$C_{BDDC}^{-1} = R_{D\Sigma}^\top S_\Delta^{-1} R_{D\Sigma} + R_{D\Sigma}^\top \Psi S_{\Pi\Pi}^{-1} \Psi^\top R_{D\Sigma}.$$

Remark 6.8. *If all interface degrees of freedom are chosen to be primal, the preconditioner is the inverse of the system matrix, and a direct solver is created.*

Remark 6.9. *The subdomain correction S_Δ^{-1} is the sum of the inverses of the local Schur complements $S_\Delta^{(i)}$. Thus, for parallel computations where each subdomain is treated by one processor, S_Δ^{-1} can be applied in parallel subdomain by subdomain and communication between processors is just needed during partitioning or averaging with $R_{D\Sigma}$ and $R_{D\Sigma}^\top$. The situation is different for the corse grid correction. Because S_Π^{-1} is the inverse of the sum of the local Schur complements $S_\Pi^{(i)}$, communication is needed during the application.*

Therefore, a small number of primal unknowns is desirable.

We have to note that the BDDC algorithm is closely related to the FETI-DP [FLP00] method. In [MDT05] it was proven that these two methods have for the same primal constraints equal non zero eigenvalues which leads to the same convergence rates. How to choose these primal constraints is still widely discussed and has a big impact onto the convergence rate. In [MT01] and [KWD02] the authors showed that for positive definite self adjoint problems in two and three dimensions, respectively, the condition number of the FETI-DP preconditioner, and consequently of the BDDC method as well, can be bounded polylogarithmic, i.e.

$$\kappa(C_{BDDC}^{-1}A) \leq c\Big(1 + \log\left(\frac{H}{h}\right)\Big)^2$$

for a certain set of primal constraints. Here c is a constant independent of the mesh size h and the subdomain size denoted by H.

6.4.2 The BDDC preconditioner for the mixed hybrid formulation

For applying the BDDC preconditioner to the problems in the Formulations 5.2 and 5.10 we will change the bilinear forms consistently. In order to motivate this change, we consider the variational formulation 5.2 of Helmholtz problem on the interval $[0, 1]$ with constant coefficients, $\epsilon = \mu = 1$. The computational domain $[0, 1]$ is meshed by the points $0 = x_0, \ldots, x_N = 1$ into N elements. Note that the facets are the grid points and the functions u_F and σ_F are just scalars. On the facets the facet normal $\boldsymbol{n}_F$ is considered to be 1. When changing the facet variables according to

$$\begin{aligned} g^+ &= u_F + \sigma_F \\ g^- &= u_F - \sigma_F, \end{aligned}$$

g^+ represents a right going trace and g^- a left going trace. After this change of variables and the elimination of the volume unknowns for $\beta = 1$ and $\alpha = 0$ the equations at the grid point x_j in the interior of the domain have the form

$$\begin{pmatrix} 0 & 0 \\ \gamma^+ & 0 \end{pmatrix} \begin{pmatrix} g_{j-1}^+ \\ g_{j-1}^- \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} g_j^+ \\ g_j^- \end{pmatrix} + \begin{pmatrix} 0 & \gamma^- \\ 0 & 0 \end{pmatrix} \begin{pmatrix} g_{j+1}^+ \\ g_{j+1}^- \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

with phase factors γ^+ and γ^-. Here the new facet unknowns g_j^+ and g_j^- represent the evaluations of g^+ and g^- in x_j. For this simple setting we can see that the right going trace in x_j is just a phase factor times the right going trace in x_{j-1}, and the left going trace is the left going trace from x_{j+1} multiplied with a phase factor. Note that a forward Gauss Seidel step propagates an incoming wave from the left through the domain and a backward Gauss Seidel iteration a wave coming from the right, if a correct numbering of the degrees of freedom is assumed. Thus, the symmetrized Multiplicative Schwarz preconditioner provides a direct solver. Furthermore, we notice that the system can be brought into a diagonal dominant form by exchanging rows.

In order to mimic such a row exchange, we add stabilizing terms to the bilinear form in Formulation 5.2, i.e. we search for $\tilde{u} = (u, \boldsymbol{\sigma}, u_F, \sigma_F) \in U \times \widetilde{V} \times U_F \times V_F$ such that

$$B_{s\Omega}(\tilde{u}, \tilde{v}) + B_{s\Gamma}(\tilde{u}, \tilde{v}) + B_{s\gamma}(\tilde{u}, \tilde{v}) = F_s(\tilde{v}) \tag{6.16}$$

for all $\tilde{v} = (v, \boldsymbol{\tau}, v_F, \tau_F) \in U \times \widetilde{V} \times U_F \times V_F$, with the additional term

$$B_{s\gamma}(\tilde{u}, \tilde{v}) = \sum_{T \in \mathcal{T}} \gamma \Big(\big\langle (\boldsymbol{n}_T \cdot \boldsymbol{n}_F) \sigma_F, v_F \big\rangle_{\partial T \setminus \Gamma} + \big\langle u_F, (\boldsymbol{n}_T \cdot \boldsymbol{n}_F) \tau_F \big\rangle_{\partial T \setminus \Gamma} \Big), \qquad \gamma \in \mathbb{C}. \tag{6.17}$$

To the variational formulation 5.10 of the vector valued wave equation the term

$$\sum_{T \in \mathcal{T}} \gamma \Big(\big\langle (\boldsymbol{n}_T \cdot \boldsymbol{n}_F) \boldsymbol{H}_F, \boldsymbol{e}_F \big\rangle_{\partial T \setminus \Gamma} + \big\langle \boldsymbol{E}_F, (\boldsymbol{n}_T \cdot \boldsymbol{n}_F) \boldsymbol{h}_F \big\rangle_{\partial T \setminus \Gamma} \Big), \qquad \gamma \in \mathbb{C}. \tag{6.18}$$

is added. The parameter $\gamma \in \mathbb{C}$ is a tuning parameter we choose based on numerical experiments. For the Helmholtz equation we made good experience with $\gamma = -0.5 - 0.1i$, for the vector valued wave equation $\gamma = 0.5$ was taken. These additional terms are just added for inner facets, and because of the different sign of $\boldsymbol{n}_T \cdot \boldsymbol{n}_F$ for the two neighboring elements, they cancel out when the global system of equations is assembled. Thus, the problem does not change. But for domain decomposition preconditioners, which are based on submatrices assembled just on a subdomain the situation changes. These additional terms do not cancel out in the submatrices for degrees of freedom located on the interface with other subdomains.

We use a BDDC preconditioner for this modified problem which is reduced by static condensation to a facet problem. The computational domain is divided into subdomains, and the degrees of freedom on facets which just belong to one subdomain are considered to be primal as well as the low order degrees of freedom on interface facets. The high order degrees of freedom on interface facets are the dual ones. This choice leads to a large global system, the primal system which consists of weakly coupled subdomain blocks due to the

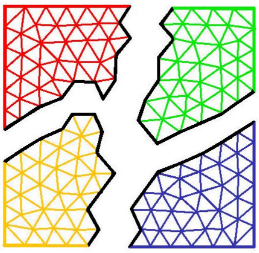

Figure 6.1: The domain is divided into subdomains. An Application of A_i^{-1} corresponds to a solution for all degrees of freedoms on the subdomain Ω_i, whereas by applying $\tilde{A}_i^{-1}$ we just solve for unknowns on inner (colored) facets.

missing high order unknowns at the interface.

6.5 A Robin type domain decomposition preconditioner

Like the BDDC preconditioner, the new Robin type domain decomposition (RDD) preconditioner will be applied to the facet system of equations $Ax = f$ obtained from the Formulations 5.2 and 5.10, respectively, by eliminating the volume degrees of freedom and adding the stabilization terms (6.17) and (6.18). We should mention that we made good experience by choosing the tuning parameter as in the BDDC case.

Before describing the preconditioner, we introduce some notations. We assume, as for the other preconditioners, that the computational domain is divided into N subdomains Ω_i (compare Fig. 6.1). For each subdomain a matrix A_i representing the subdomain problem is subassembled, and the matrix A is obtained by adding these submatrices. By $\tilde{A}_i$ we denote the block of A_i which corresponds to degrees of freedom on inner facets, i.e. facets which just belong to the domain Ω_i (the colored facets in Fig. 6.1). The operator $R^{(i)}$ restricts a vector to the components corresponding to these inner degrees of freedom of the domain Ω_i. The operator $R_D^{(i)}$ provides a weighted restriction to the domain Ω_i, i.e. when applying it, a vector entry is divided by the number of subdomains to which the corresponding degree of freedom belongs to. Note that an application of the prolongation $R_D^{(i)\top}$ results again in a division for the interface degrees of freedom. Thus, by summing up over all subdomains, a mean value on the interface can be created.

Using this notations, a preconditioner step for finding an approximate solution $\tilde{x}$ to the system $Ax = f$ reads as

(1) $y_0 = 0,$

(2) $y_1 = y_0 + \sum_{i=1}^{N} R^{(i)\top} \tilde{A}_i^{-1} R^{(i)} (f - Ay_0),$

(3) $y_2 = y_1 + \sum_{i=1}^{N} R_D^{(i)\top} A_i^{-1} R_D^{(i)} (f - Ay_1),$

(4) $y_3 = y_2 + \sum_{i=1}^{N} R^{(i)\top} \tilde{A}_i^{-1} R^{(i)} (f - Ay_2),$

(5) $\tilde{x} = y_3.$

First, in step (2), we solve the system of equations exactly for the degrees of freedom on the inner facets under the constraint that the solution on the interface is zero. Step (3) provides an update for the interface solution by solving the problem exactly subdomain by subdomain. A continuous interface solution is constructed by averaging the different subdomain solutions. Finally, in step (4) the solution is updated such that the system of equations is solved again exactly for the degrees of freedom on inner facets. Note that the interface solution remains unchanged.

Remark 6.10. *The RDD-preconditioner can also be introduced with the notations used for Schwarz preconditioners. The bilinear form representing the Schur complement system is defined on the facet space* $W := U_F \times V_F$ *(or* $X_F \times Y_F$ *in the vector valued case), and it is denoted by* a*. The subspace of* W *containing the functions which are supported on the subdomain* Ω_i *is denoted by* W_i*, and in* $\tilde{W}_i$ *functions supported only on inner facets of the domain* Ω_i *are collected. The operator representation of the restriction matrix* $R_D^{(i)}$ *is called* $\mathcal{R}_D^{(i)} : W \to W_i$*. Thus, when applying it to any function in* W*, the function is restricted to the domain* Ω_i*, and its values on the interface facets are divided by the number of neighboring subdomains. Furthermore,* $\mathcal{R}^{(i)} : W \to \tilde{W}_i$ *is the restriction operator corresponding to the matrix* $R^{(i)}$*, and by* $\mathcal{R}^{(i)\top}$ *and* $\mathcal{R}^{(i)\top}$ *the prolongation operators are denoted. Additionally, we use that the bilinear form* a *can be decomposed into the contributions* a_i *of the subdomains* Ω_i*, i.e.* $a = \sum_{i=1}^{N} a_i$*.*

Based on this, we define the variational projector $\mathcal{P}_D^{(i)}$ *via* $\mathcal{P}_D^{(i)} = \mathcal{R}_D^{(i)\top} \hat{\mathcal{P}}_D^{(i)}$ *with the projector* $\hat{\mathcal{P}}_D^{(i)} : W \to W_i$ *and*

$$a_i(\hat{\mathcal{P}}_D^{(i)} u, \phi) = a(u, \mathcal{R}_D^{(i)\top} \phi) \qquad \forall \phi \in W_i.$$

In the same way the variational projector $\mathcal{P}^{(i)}$ *with* $\mathcal{P}^{(i)} = \mathcal{R}^{(i)\top}\hat{\mathcal{P}}^{(i)}$ *can be introduced. Here,* $\hat{\mathcal{P}}^{(i)} : W \to \tilde{W}_i$ *is given via*

$$a_i(\hat{\mathcal{P}}^{(i)}u, \phi) = a(u, \mathcal{R}^{(i)\top}\phi) \qquad \forall \phi \in \tilde{W}_i.$$

If the operator $\mathcal{A}$ *corresponds to the bilinear form* a*, and* $\mathcal{I}$ *is the identity, the error propagation operator* $\mathcal{E}$ *of the RDD-preconditioner reads as*

$$\mathcal{E} = \mathcal{I} - \mathcal{C}_{RDD}^{-1}\mathcal{A} = \Big(\mathcal{I} - \sum_{i=1}^{N} \mathcal{P}^{(i)}\Big)\Big(\mathcal{I} - \sum_{i=1}^{N} \mathcal{P}_D^{(i)}\Big)\Big(\mathcal{I} - \sum_{i=1}^{N} \mathcal{P}^{(i)}\Big).$$

Remark 6.11. *Because we solve always exact for the degrees of freedom on inner facets, both sets of facet degrees of freedoms* u_F *and* σ_F *(or* $\boldsymbol{E}_F$ *and* $\boldsymbol{H}_F$ *in the time harmonic case) are not needed there anymore, and the problem can be formulated just by using* u_F*. On the interface both types of unknowns are still necessary in order to fix continuity conditions of the impedance traces across the interface and to guarantee convergence of the iterative solver. Nevertheless, neglecting one type of facet unknowns on inner facets safes a lot of degrees of freedom in an actual calculation.*

6.6 Numerical results

The numerical examples of this section were calculated with the MPI-parallel finite element code Netgen/Ngsolve of Schöberl (see *http://sourceforge.net/projects/ngsolve* or [Sch97]) on a Dell R-910 Server (4 Xeon E7 CPUs with 10 cores a 2.2 GHz, 512 GB RAM).

6.6.1 Comparison of the preconditioners

In the following, we want to compare the preconditioners, if not said differently, for a simple model problem in two dimensions. There, we solve the Helmholtz equation (Formulation 5.2) on a square $\Omega = [-1, 1]^2$. An incoming wave from above with Gaussian amplitude is fixed by $g(x, 1) = \exp(-10x^2)$ an $g = 0$ else.

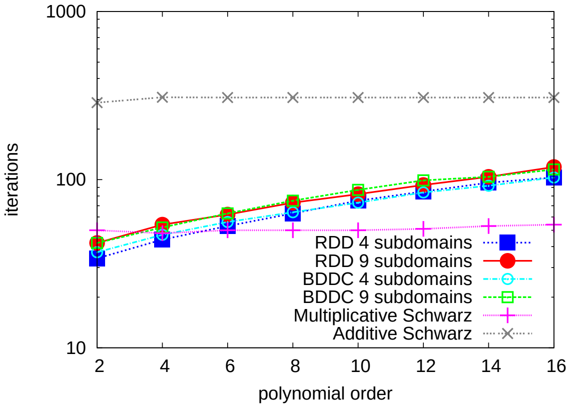

Figure 6.2: Number of iterations versus polynomial order for different preconditioners

Dependence of the number of iterations on the polynomial order

With the help of Figure 6.2, which plots the number of iterations against the polynomial order, we try to study this dependence for different preconditioners in a PCG method. For this figure the wavelength λ was chosen to be 0.2. As mesh size we took $h = 0.1$, which is enough to resolve the solution at polynomial order two. According to the plot, the BDDC and the RDD preconditioner show similar features, and the number of iterations grows with growing polynomial order. Note that an increase in the number of subdomains leads to a small increase in the iterations as well. The situation is different for Schwarz preconditioners with blocks related to elements. There, the iterations seem to be almost independent of the polynomial order. Nevertheless, it is advantageous to take a BDDC or RDD preconditioner. They can be used in parallel codes, and for the relevant polynomial orders the number of iterations is comparable to the Multiplicative Schwarz preconditioner. The Additive Schwarz preconditioner needs too many iterations to be competitive.

Dependence of the number of iterations on the mesh size and wavelength

In the Tables 6.1 - 6.3 the iteration numbers for the Multiplicative Schwarz, the BDDC and the RDD preconditioner for different wavelengths and mesh sizes are given. For all calculations the polynomial order was kept constant to four. The cells with a gray background belong to a setting where the number of degrees of freedom is to small to resolve the solution, and therefore, the iteration numbers are not representative.

$h \setminus \lambda$	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$\frac{1}{4}$	31	30	30	49	101	6	5
$\frac{1}{8}$	59	64	59	60	109	292	6
$\frac{1}{16}$	118	118	126	121	118	310	731
$\frac{1}{32}$	250	214	218	233	224	228	
$\frac{1}{64}$	518	460	406	432	449	467	510

Table 6.1: Iteration numbers of the Multiplicative Schwarz preconditioner ($p = 4$) for different mesh sizes and wavelength

$h \setminus \lambda$	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$\frac{1}{4}$	45	49	60	78	227	24	24
$\frac{1}{8}$	51	48	56	73	113	320	22
$\frac{1}{16}$	56	50	49	59	80	161	400
$\frac{1}{32}$	63	57	48	49	65	85	223
$\frac{1}{64}$	66	62	56	50	50	74	101

Table 6.2: Iteration numbers of the BDDC preconditioner using 9 subdomains ($p = 4$) for different mesh sizes and wavelength

$h \setminus \lambda$	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
$\frac{1}{4}$	78	65	61	64	355	25	26
$\frac{1}{8}$	95	84	71	70	84	430	24
$\frac{1}{16}$	123	96	83	73	72	116	504
$\frac{1}{32}$	154	125	101	84	74	74	171
$\frac{1}{64}$	202	164	127	111	89	82	89

Table 6.3: Iteration numbers of the RDD preconditioner using 9 subdomains ($p = 4$) for different mesh sizes and wavelength

While the number of iterations seems to stay constant in wavelength or frequency, respectively, for the Multiplicative Schwarz preconditioner (Table 6.1), it is indirect proportional to the mesh size. Thus, if the mesh size is divided by two, the number of iterations doubles. This indicates that a fixed number of iterations is needed to propagate the information of an incoming wave across one element.

In order to describe the dependence of the iterations on the mesh size and the wavelength for the BDDC preconditioner (compare Table 6.2), we choose a setting with $h = \lambda$. Due to a polynomial order of four, such a choice of parameters is close to the resolution limit of the oscillatory solution. Increasing the degrees of freedom per wavelength either

subdoms\\λ	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
4	65	63	56	50	47	59	78
9	66	62	56	50	50	74	101
16	66	64	57	51	50	81	116
25	68	64	61	51	52	88	129
36	66	68	60	51	52	96	144

Table 6.4: Iteration numbers of the BDDC preconditioner for different wavelengths and numbers of subdomains. The polynomial order was 4 and the mesh size $\frac{1}{64}$.

subdoms\\λ	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
4	169	134	107	86	73	63	66
9	202	164	127	111	89	82	89
16	240	184	150	123	104	93	105
25	281	236	187	150	122	110	121
36	365	259	205	164	139	127	143

Table 6.5: Iteration numbers of the RDD preconditioner for different wavelengths and numbers of subdomains. The polynomial order was 4 and the mesh size $\frac{1}{64}$.

by decreasing the mesh size or by increasing the wavelength leads first to less iterations. A further increase of the unknowns per wavelength causes afterwards growing iteration counts. Although, for a large wavelength and a small mesh size the RDD preconditioner needs much more iterations than the BDDC preconditioner (compare Table 6.3), it gets more and more competitive if the number of degrees of freedom per wavelength is reduced. Close to the resolution limit of the solution the RDD preconditioner needs in fact fewer iterations than the BDDC. One reason for this behavior could be the different structure of the two solvers. While the RDD preconditioner allows just for local corrections, the BDDC solver benefits additionally from a coarse grid solution. For a decreasing wavelength, the solution gets more and more oscillatory, and the coarse grid correction, which provides communication across the whole domain, loses its importance.

Dependence of the iterations on the number of subdomains

The number of iterations of the BDDC and the RDD preconditioner is also influenced by the number of subdomains the computational domain is divided into. In the Tables 6.4 and 6.5 and in Figure 6.3 iteration counts of these two preconditioners for different wavelengths and numbers of subdomains are provided. In the corresponding experiments

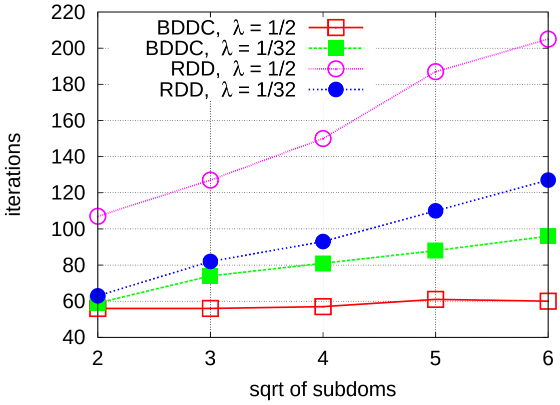

Figure 6.3: Number of iterations plotted versus the square root of the number of subdomains for the BDDC and the RDD preconditioner. The polynomial order was 4 and $h = \frac{1}{64}$.

the mesh size was kept constant to $\frac{1}{64}$ and the polynomial order to four. For the RDD preconditioner the number of iteration grows with the number of subdomains. Figure 6.3 indicates that this growth is proportional to the square root of the number of subdomains, i.e. to the number of subdomains in one spatial direction. Thus a constant number of iterations is needed to propagate the input data across one subdomain. The situation is slightly different for the BDDC preconditioner. While it shows the same features for small wavelengths, i.e. for settings close to the resolution limit, the iterations stay almost constant for large wavelengths. A reason for this is that for less oscillatory solution the BDDC preconditioner benefits from its coarse grid correction.

Comparison in computational times

Apart from iteration numbers computational times are interesting, when a preconditioner is applied to a specific problem. For our model problem with $p = 4$, $h = \frac{1}{64}$ Table 6.6 presents the timings of one iteration and the setup of our preconditioners. The RDD and the BDDC preconditioner were used in parallel, where each subdomain was assigned to one processor. At the first glance one can see that the Multiplicative Schwarz preconditioner is not competitive at all. While for a similar setting one RDD iteration is just slightly faster than one BDDC iteration, the time required for the set up process of the RDD preconditioner is almost by a factor of ten smaller. Consequently, the RDD preconditioner

	subdoms	setup(sec.)	time per iteration(sec.)
BDDC	4	14.38	0.493
	9	8.39	0.280
	16	4.25	0.201
	25	4.8	0.171
RDD	4	2.91	0.538
	9	0.95	0.236
	16	0.52	0.151
	25	0.44	0.118
MS			2.44

Table 6.6: Timings for a setting with $h = \frac{1}{64}$ and $p = 4$

is considerably faster for problems where both preconditioners show a comparable number of iterations.

Selection of the subdomains

For many practical applications the number of iterations is very sensitive to how the subdomains are chosen. In order to demonstrate this, we take a scattering example with the unit square as computational domain. First, a wave of wavelength 0.01 with Gaussian amplitude, which is injected from the top, is scattered at a square (compare Figure 6.4, top left). In a second example a cavity is included in the scatterer. This cavity is connected via a small channel to the exterior region (compare Figure 6.4 bottom left). We can see from the zoom in Figure 6.4 that in the cavity a resonance is excited via this channel. For the computations a polynomial order $p = 10$ was used for the meshes of Figure 6.4. Table 6.7 gives the iteration numbers and computational times of different preconditioners for the scattering problem at the square. The RDD and the BDDC preconditioner were based on five subdomains. If we choose for the cavity problem the same number of subdomains as for the squared scatterer, and if the whole cavity is contained in one subdomain, the exact local solver of the RDD and the BDDC preconditioner takes care of internal reflections, and the whole cavity is solved exactly in each iteration. Table 6.8 shows that the number of iterations stays for these two preconditioners approximately constant and that the computational times grow slightly due to the bigger number of unknowns. If a subdomain division is used which splits the cavity among different processors or subdomains, respectively, the preconditioners can not cope with the internal reflections anymore, which leads to an enormous growth in iterations and computational time. This is indicated by

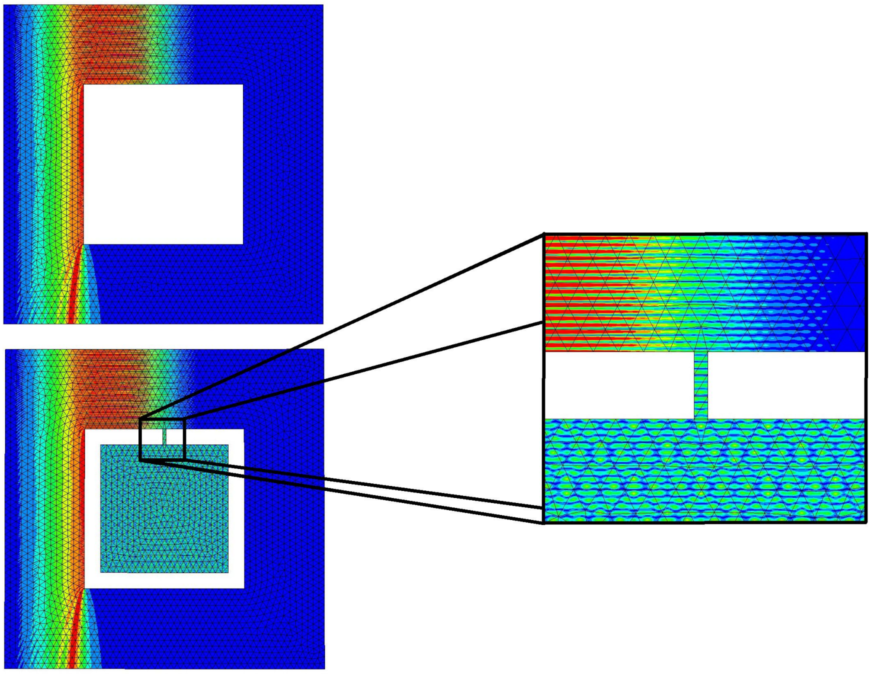

Figure 6.4: Geometry and absolute value of the solution u for a scattering problem at a square (top left) and of a square with a cavity included (bottom left); The right hand plot shows a zoom to a channel which connects the cavity with the exterior region.

	its.	time(sec.)
BDDC	52	14.9
RDD	43	9.1
MS	102	88.9
AS	575	186

Table 6.7: Iteration numbers and computational times for our preconditioners using 5 subdomains for the scattering problem at a square

	its.	time(sec.)
BDDC	50	18.7
RDD	39	10.4
MS	1612	1720
AS	$>10^5$	>1h

Table 6.8: Iteration numbers and computational times for our preconditioners using 5 subdomains for the cavity problem with the cavity included in one domain

	its.	time(sec.)
BDDC	203	45.9
RDD	350	69.0
MS	1612	1720
AS	$>10^5$	>1h

Table 6.9: Iteration numbers and computational times for our preconditioners using 6 subdomains for the cavity problem with the cavity divided among two of them

p	λ	λ/D	its.	$\|u-u_h\|_{L^2}/\|u\|_{L^2}$
3	0.05	40	119	0.42
	0.0667	30	124	0.067
	0.08	25	129	0.025
4	0.04	50	125	0.40
	0.05	40	122	0.073
	0.0667	30	127	0.012
5	0.04	50	117	0.082
	0.05	40	121	0.014
	0.0667	30	132	0.0026

Table 6.10: Iteration numbers of the RDD preconditioner (74 subdomains) and the relative error of the solution for different polynomial orders and wavelengths. As computational domain a sphere with diameter $D = 2$, meshed by 233000 elements ($h = 0.04$) was used.

polynomial order p	3	4	5
volume dofs in mio. ($u + \boldsymbol{\sigma}$)	4.8+20.2	8.3+34.3	13.3+53.3
facet dofs in mio. ($u_F + \sigma_F$)	4.8+4.8	7.3+7.3	10.2+10.2
assembling (sec)	57	220	714
setup (sec)	143	421	1160
time per iteration(sec)	3.8	8.6	17.3

Table 6.11: Computational times and number of unknowns for the settings used in Table 6.10. The computations were done in parallel on 75 processors

the numbers in Table 6.9. There, six subdomains were used, and the cavity is divided among two of them. Note that this effect is weaker for the BDDC preconditioner, which benefits from its corse grid solver. Finally, we remark that the Schwarz preconditioners, which just solve exactly on the element level, suffer a lot from these reflections. They are not suited for such a problem.

6.6.2 Large scale examples for the Helmholtz equation

We want to demonstrate the efficiency of our method for three dimensional problems with a small ratio of wavelength to domain size with an example where a plane wave is the analytic solution. As computational domain a sphere of diameter $D = 2$ was used which was meshed by 233000 elements of maximal size $h = 0.04$. The resulting linear system of equations was solved with a RDD preconditioner based on 74 subdomains. For different polynomial orders and wavelengths the numbers of iterations are given together with the relative error

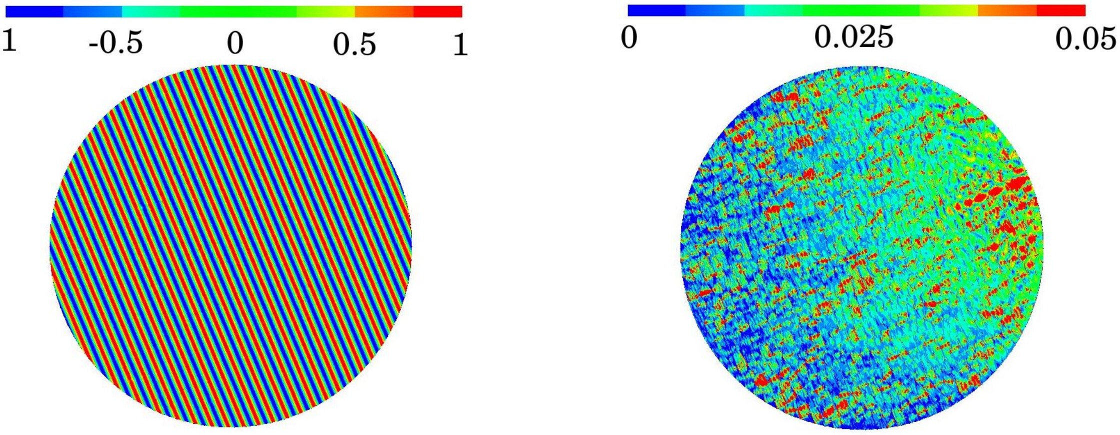

Figure 6.5: Real part of the solution u (left) and $|u - u_h|$ (right) for the model problem used in the Tables 6.10 and 6.11 ($h = 0.04$, $p = 3$, $\lambda = 0.08$).

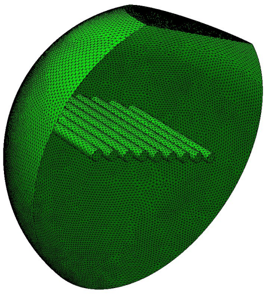

Figure 6.6: Geometry of a grating cut parallel to the xy plane.

$\|u - u_h\|_{L^2(\Omega)}/\|u\|_{L^2(\Omega)}$ of the approximate solution u_h in Table 6.10. Figure 6.5, which shows the approximate solution for $p = 3$ and $\lambda = 0.8$ (left) as well as its error $|u - u_h|$ (right), demonstrates that the error grows into the direction of propagation of the wave, i.e. from bottom left to top right. For completeness Table 6.11 lists the number of unknowns for the different settings of Table 6.10 together with the time needed for assembling the linear system of equations and the timings of the RDD preconditioner. Thus, finding an approximation of our problem for 50 wavelength per domain with a relative error of 0.08 takes about 33.5 minutes. For the computation about 66.6 millions of volume unknowns and 20.4 millions of facet unknowns are necessary.

As a second three dimensional example we calculate the total field of the Helmholtz problem for a grating shown in Figure 6.6. The grating consists of rods with a distance of 0.14 placed in a spherical domain of diameter two. Thus, assuming a wave incoming from

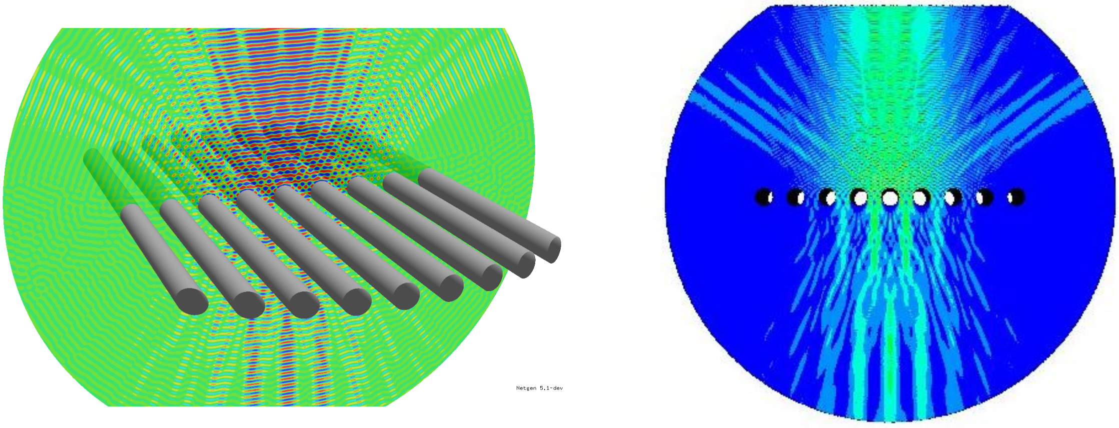

Figure 6.7: Real part of the solution (left) and its absolute value (right) for a wave scattered at the grating of Figure 6.6. The computations were done on Vienna Scientific Cluster 2 (VSC2).

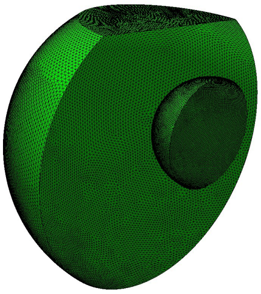

Figure 6.8: A small sphere with $\epsilon = 2.5$ placed in the computational domain with $\epsilon = 1$. The computational domain is cut parallel to the xy plane.

the top with Gaussian amplitude and wavelength 0.025 corresponds to an effective domain size of 80 wavelengths. For this setting Figure 6.7 shows the real part of the solution and its absolute value, i.e. the diffraction pattern. In the calculation the underlying mesh (compare Figure 6.6) had about 1.61 million elements with a maximal mesh size of 0.021. Selecting a polynomial order of $p = 4$ results in approximately 288.8 million volume unknowns (56.5 mio. for u and 232.3 mio. for $\boldsymbol{\sigma}$) and 98.0 million facet unknowns (49.0 mio. for both, u_F and σ_F). Using 1200 subdomains, the assembly of the matrix took 58 seconds and the setup of the RDD preconditioner 33 seconds. The problem was solved in 12.9 minutes with 399 iterations. These results have been achieved using the Vienna Scientific Cluster 2 (VSC2).

Finally, we consider a scattering problem at a spherical obstacle. For this purpose a

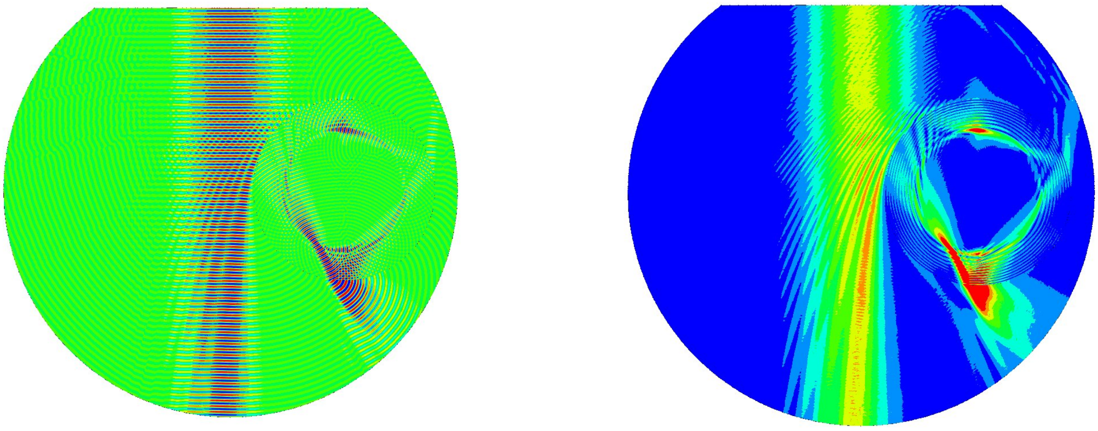

Figure 6.9: The total field u (left) and its absolute value (right) for a wave scattered at a small sphere (compare Figure 6.8). The computations were done on Vienna Scientific Cluster 2 (VSC2).

sphere of diameter 0.8 with $\epsilon = 2.5$ was placed (not exactly in the center) in a spherical domain with diameter two and $\epsilon = 1$. Choosing an incoming wave of Gaussian amplitude and wavelength 0.025 results in an effective size of 80 wavelengths for the computational domain and 50 wavelengths for the obstacle. The mesh in Figure 6.8 of about 2.67 million elements leads together with polynomial order $p = 4$ to 479 million volume unknowns (94 mio. for u and 385 mio. for $\boldsymbol{\sigma}$) and to 161 million facet unknowns (80.5 mio. for both, u_F and σ_F). Using the Vienna Scientific Cluster 2 (VSC2), 58.3 seconds were needed for the assembly procedure and 28.1 seconds for setting up the RDD preconditioner on 2000 processors. The solution, plotted in Figure 6.9 was obtained after 3726 iterations in 2.0 hours. The reason for this large number of iterations is that by the large refractive index of the scatterer a cavity like object is created, which causes a standing wave close to the surface of the scatterer (compare Figure 6.9). Because METIS, which is used in Ngsolve to partition the computational domain, divides the scatterer among several processors, the preconditioner suffers as discussed above from such internal reflections.

6.6.3 Large scale examples for the vector valued wave equation

As for the Helmholtz equation, we start for the vector valued wave equation with an example, where the incoming data $\boldsymbol{g}$ was chosen such that the exact solution is a plane wave $\boldsymbol{E} = (0,0,1)^\top e^{\boldsymbol{ikx}}$ with $\boldsymbol{k} = \frac{2\pi}{\lambda}(\cos(\phi), \sin(\phi), 0)^\top$. As computational domain a cube of side length two was taken. In all computations this cube is divided into about

p	λ	its.	$\|\boldsymbol{E}-\boldsymbol{E}_h\|_{L^2}/\|\boldsymbol{E}\|_{L^2}$
2	0.133	237	0.35
	0.2	388	0.073
	0.25	483	0.032
3	0.1	198	0.33
	0.133	326	0.068
	0.2	538	0.011

Table 6.12: Iteration numbers of the RDD preconditioner (69 subdomains) and the relative error of the solution for different polynomial orders and wavelengths. As computational domain a cube side length two, meshed by 66000 elements ($h = 0.08$), was used.

polynomial order p	2	3
volume dofs in mio. $(\boldsymbol{E}+\boldsymbol{H})$	2.0+4.0	4.0+7.0
facet dofs in mio. $(\boldsymbol{E}_F+\boldsymbol{H}_F)$	1.6+1.6	2.7+2.7
assembling (sec)	16.5	73
setup (sec)	231	1054
time per iteration(sec)	2.8	8.3

Table 6.13: Computational times and number of unknowns for the settings used in Table 6.12. The computations were done in parallel on 70 processors

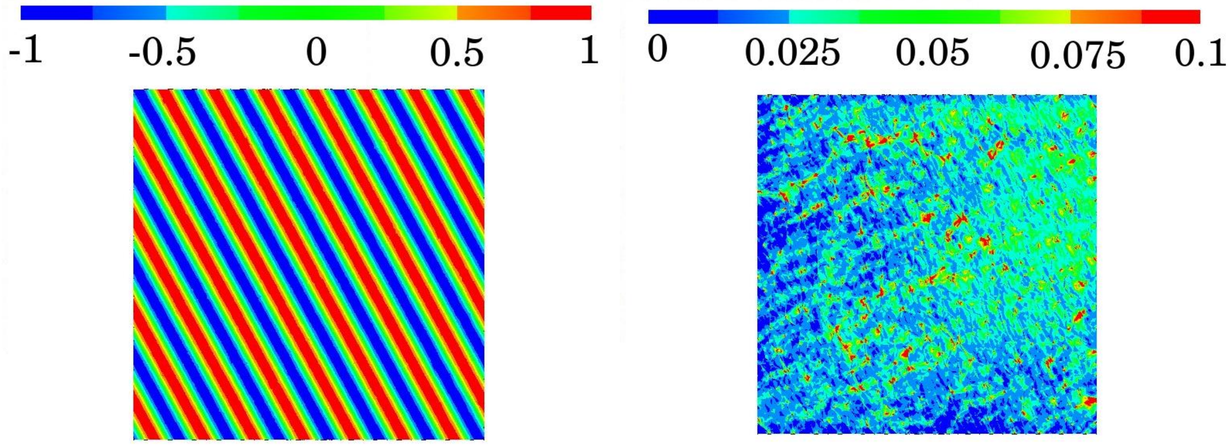

Figure 6.10: Real part of E_z (left) and $|\boldsymbol{E}-\boldsymbol{E}_h|$ (right) for the model problem used in the Tables 6.12 and 6.13 ($h = 0.08$, $p = 2$, $\lambda = 0.25$).

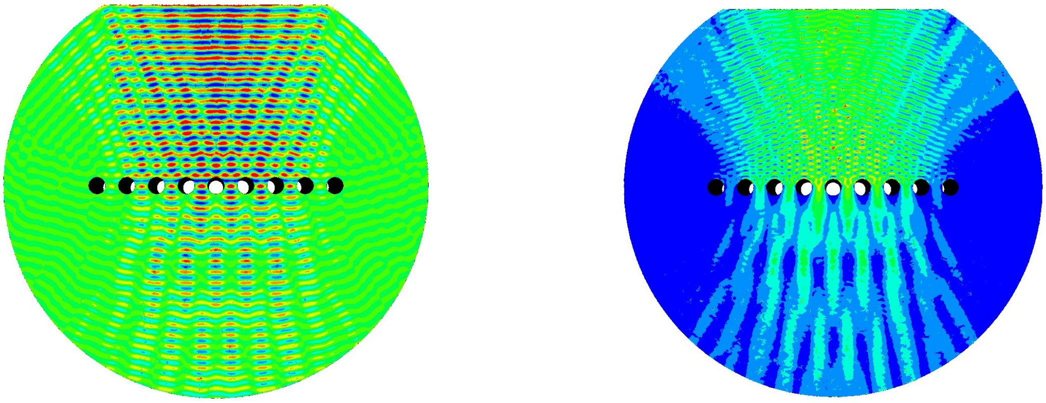

Figure 6.11: Real part of the field component parallel to the rods (left) and its absolute value (right). The computations were done on Vienna Scientific Cluster 2 (VSC2).

66000 elements with a maximal mesh size of $h = 0.08$. For this problem, the iteration numbers of a RDD preconditioner based on 69 subdomains together with the relative error $\|\boldsymbol{E} - \boldsymbol{E}_h\|_{L^2(\Omega)}/\|\boldsymbol{E}\|_{L^2(\Omega)}$ for different polynomial orders p and wavelengths λ are given in Table 6.12. In Figure 6.10 the solution (left), more precisely $\mathrm{Re}(E_z)$, is plotted together with the error $|\boldsymbol{E} - \boldsymbol{E}_h|$ for $\lambda = 0.25$ and $p = 2$. The Figure indicates that the error grows into the direction of propagation of the wave. For completeness Table 6.13 presents the number of unknowns as well as the computation times for the different settings used in Table 6.12. For example, the computation of $\boldsymbol{E}_h$ for $p = 3$ and $\lambda = 0.133$, which corresponds to 26 wavelengths in the space diagonal, takes 1.06 hours, and a relative error of 0.068 can be reached. For this calculation 11.0 millions of volume unknowns and 5.4 millions of facet unknowns are necessary.

As a second example, the vector valued wave equation is solved for the grating from Figure 6.6 (left). Thus, the size of the computational domain was two, and the spacing of the rods was 0.14. An incoming wave from the top polarized perpendicular to the rods with a Gaussian shaped amplitude and wavelength 0.05 was assumed. Note that this corresponds to an effective domain size of 40 wavelengths. The computational domain was meshed by 419000 elements, which results for a polynomial order of three in 69.2 million volume unknowns (25.2 mio. for u and 44.0 mio. for $\boldsymbol{\sigma}$) and 34.4 million facet unknowns (17.2 mio. for both, u_F and σ_F). In Figure 6.11 the absolute value of the solution (left) and the the real part of its component perpendicular to the rods (right) are plotted. The solution was obtained on Vienna Scientific Cluster 2 (VSC2) with an RDD preconditioner based on 1500 subdomains after 1071 iterations in 19.5 minutes by using 1500 processors.

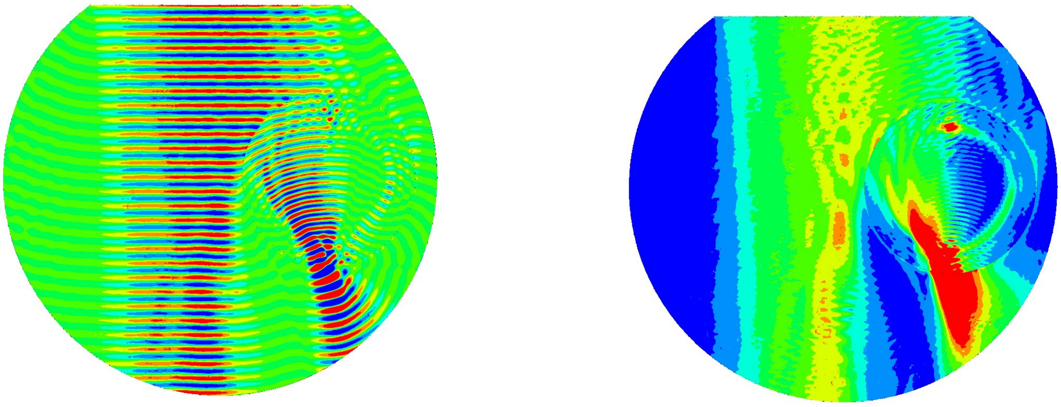

Figure 6.12: The solution on a cross section parallel to the xy plane; The left hand plot shows the real part of $\boldsymbol{E}_z$ and the right hand plot the absolute value of $\boldsymbol{E}$. The computations were done on Vienna Scientific Cluster 2 (VSC2).

For the assembly of the matrix and the setup of the preconditioner 12.7 seconds and 44.5 seconds, respectively, were necessary.

We conclude the numerical results section by solving a scattering problem comparable to the one from Figure 6.8 for the vector valued wave equation. Thus we consider a spherical obstacle of diameter 0.8 and $\epsilon = 2.5$ placed not exactly in the center of the computational domain with diameter 2 and $\epsilon = 1$. Assuming an incoming Gaussian shaped beam polarized parallel to the z axis with wavelength 0.066 leads to effective sizes of 30 wavelengths for the domain and 19 wavelengths for the obstacle. The underlying mesh consisted of 58000 elements, which corresponds for a polynomial order of three to 73.5 million volume unknowns (26.7 for $\boldsymbol{E}$ and 46.8 for $\boldsymbol{H}$) and 36.0 million facet unknowns (18.0 for both, $\boldsymbol{E}_F$ and $\boldsymbol{H}_F$). Figure 6.12 shows the real part of $\boldsymbol{E}_z$ and the absolute value of $\boldsymbol{E}$ on a cross section parallel to the xy plane. Solving this problem with a RDD-preconditioner based on 1500 subdomains requires because of internal reflections in the scatterer 2941 iterations. These iterations were done on Vienna Scientific Cluster 2 (VSC2) with 1500 processors in 1.0 hours. The assembly and the set up of the preconditioner took 13.3 seconds and 45.7 seconds, respectively.

Chapter 7

Hardy space infinite elements for the mixed hybrid formulation

When solving the Helmholtz equation on an unbounded domain with finite elements, the computational domain has to be restricted onto a finite domain, and appropriate boundary conditions have to be imposed, which let outgoing waves cross freely without or with only little reflections. These boundary conditions, called transparent boundary conditions, have to replace somehow the Sommerfeld radiation condition, which is the correct radiation condition at infinity.

In the previous chapters we implemented transparent boundary conditions via Robin type boundary conditions, also known as first order absorbing boundary conditions. These boundary conditions are only free from reflections for a certain angle of incidence and wavelength. In the literature there exists a large variety of methods to realize transparent boundary conditions with less or without reflections. One option is infinite elements [Ast00, DG98], which are based on a radial expansion following the asymptotic behavior of Hankel functions of the first kind. The main idea of the approach is to use just the first N Hankel functions together with approximations for the coefficients obtained by a finite element discretization at the domain boundary. Apart from boundary integral or boundary element techniques [HW08] and high order local approaches [Giv04, GK95], the perfectly matched layer technique, also known as complex scaling, [Sim79, Ber94] is widely used. There, the computational domain is surrounded by a perfectly matched absorbing layer. This layer is free of reflections at the interface, independent of the frequency and direction of the wave, and the wave decays exponentially with respect to the distance from the interface.

In the following, we will use Hardy space infinite elements (HSIE), first introduced in

[Nan08, HN09, NS11], in order to realize transparent boundary conditions. The method is based on the pole condition from [Sch98a, Sch02], where outgoing and incoming waves are distinguished by the position of the singularities of the Laplace transform with respect to the radial direction. When applying the HSIE method, the exterior domain is divided into infinite elements, and the radial component of the solution is transformed via a certain transformation to the Hardy space $H^+(D)$. A Galerkin method on the complex unit disc D approximates now the exterior solution of the wave equation.

The basics of HSIE are explained in Section 7.1 with the help of a one dimensional example. In Section 7.2 HSIE are brought into the new context of the mixed hybrid formulation from Chapter 5. The discrete problem, or more precisely, the choice of basis functions and the resulting structure of the element matrix is discussed in Section 7.3 .

7.1 Hardy space infinite element method in one dimension

In order to introduce the basic concepts of the HSIE method from a practical point of view, we take the one dimensional example from [HN09] where the Helmholtz equation

$$\begin{aligned} -u''(r) - \kappa^2 p(r) u(r) &= 0 \qquad \text{for } r \geq 0, && (7.1)\\ u'(0) &= g, && (7.2)\\ \lim_{r\to\infty} \big(\partial_r u(r) - i\omega u(r)\big) &= 0 && (7.3) \end{aligned}$$

with the wavenumber $\kappa \in \mathbb{C}$, $\mathrm{Re}(\kappa) > 0$, $g \in \mathbb{C}$ and $p \in L^\infty(\mathbb{R}^+)$ with $p = 1$ for $r \geq a$ is solved by using Hardy space boundary conditions. A very detailed discussion of this problem can be also found in [Nan08]. We should note that the Sommerfeld radiation condition (7.3) guarantees well posedness of the problem and that it is equivalent to a first order absorbing boundary condition in one dimension. Thus, for the problem under consideration simpler methods than HSIE exist. Because p is constant for $r \geq a$ we can decompose the solution u into an exterior solution $u_E(r) := u(a + r)$ for $r \geq 0$ and an interior part $u_I(r) := u(r)$ for $r < a$. For the exterior part we obtain

$$u_E(r) = C_1 e^{i\kappa r} + C_2 e^{-i\kappa r} \qquad \text{with } C_1 + C_2 = u_I(a). \tag{7.4}$$

Because the term $C_2 e^{-i\kappa r}$ represents an incoming wave, which is suppressed by the radiation condition, $C_2 = 0$ and $u_E(r) = u_I(a)e^{i\kappa r}$.

The pole condition, which is essential for HSIE, distinguishes between outgoing and incoming waves via the Laplace transform

$$\big(\mathcal{L}f\big)(s) = \int_0^\infty e^{-sr} f(r)\, dr,$$

initially defined for $\mathrm{Re}(s) > 0$. Transforming the exterior solution (7.4) results in

$$\big(\mathcal{L}u_E\big)(s) = \frac{C_1}{s - i\kappa} + \frac{C_2}{s + i\kappa}, \qquad \mathrm{Re}(s) \geq \mathrm{Im}(\kappa),$$

which has a holomorphic extension to $\mathbb{C}\backslash\{i\kappa, -i\kappa\}$. This gives rise to the definition of in- and outgoing waves. We call a solution u_E outgoing if its Laplace transform $\mathcal{L}u_E$ has no poles with negative imaginary part, and it is called incoming if the poles have no positive imaginary part. This condition can be also formulated with the help of Hardy spaces, which we define as

Definition 7.1. *(See [NS11]) Let* $P^-_{\kappa_0} = \{s \in \mathbb{C} : \mathrm{Im}(s/\kappa_0) < 0\}$ *be the half plane below the line* $\kappa_0\mathbb{R}$ *through the origin and* $\kappa_0 \in \mathbb{C}$*. The Hardy space* $H^-(P^-_{\kappa_0})$ *is the space of all functions* f *that are holomorphic in* $P^-_{\kappa_0}$*, such that*

$$\int_{\mathbb{R}} |f(\kappa_0 x - \kappa_0 i\epsilon)|^2 dx$$

is uniformly bounded for $\epsilon > 0$*.*

Definition 7.2. *(See [NS11]) The Hardy space* $H^+(D)$ *is the space of all functions* f *that are holomorphic in* $D = \{s \in \mathbb{C} : |s| < 1\}$*, such that*

$$\|f\|_{H^+(S^1)} := \lim_{r\to 1} \int_0^{2\pi} \left|f(re^{it})\right|^2 dt$$

is bounded.

Furthermore, due to the boundedness, any function f from the spaces $H^-(P^-_{\kappa_0})$ and $H^+(D)$, respectively, can be identified with its boundary value in L^2 and vice versa. Note that these Hardy spaces equipped with the L^2-norm $\|\cdot\|_{H^+(S^1)}$ of the boundary function are Hilbert spaces [Dur70].

These Hardy spaces give us the opportunity to define outgoing waves alternatively. A solution u is said to be outgoing if the holomorphic extension of the Laplace transform of its exterior part is in the space $H^-(P^-_{\kappa_0})$, where κ_0 with $\mathrm{Re}(\kappa_0) > 0$ can be interpreted as tuning parameter. Therefore, requiring that the holomorphic extension of the Laplace transform of u_E from (7.4) is in $H^-(P^-_{\kappa_0})$ leads directly to $C_2 = 0$. For simplicity reasons, the symbol $\mathcal{L}$ will replace the expression "holomorphic extension of the Laplace transform".

Because no convenient orthonormal basis is available in $H^-(P^-_{\kappa_0})$, $\mathcal{L}f$ is mapped from $H^-(P^-_{\kappa_0})$ to $H^+(D)$ by the Möbius transform

$$\mathcal{M}_{\kappa_0} : H^-(P^-_{\kappa_0}) \to H^+(D) \qquad \text{with} \qquad \big(\mathcal{M}_{\kappa_0} f\big)(z) := f\Big(i\kappa_0 \frac{z+1}{z-1}\Big)\frac{1}{z-1},$$

which is up to a factor $\sqrt{2\kappa_0}$ unitary.

Transforming the analytical solution in the exterior domain, $u_E(r) = u_0 e^{i\kappa r}$, with $u_0 = u_I(a)$ to $H^+(D)$, we obtain

$$\hat{u}(z) := \big(\mathcal{M}_{\kappa_0}\mathcal{L}u_E\big)(z) = \frac{u_0}{i\kappa_0(z+1) - i\kappa(z-1)} = \frac{u_0}{i(\kappa+\kappa_0)}\,\frac{1}{1-\frac{\kappa-\kappa_0}{\kappa+\kappa_0}z}.$$

Thus, using a monomial basis $z^0, \dots, z^n$ in order to approximate $\hat{u}$ leads to an exponential convergence in n. But from the identity

$$2i\kappa_0\hat{u}(1) = u_0, \tag{7.5}$$

it becomes obvious that using a monomial basis for $\hat{u}$ leads to coupling between all coefficients and u_0, i.e. the solution in the interior domain, which is not really desirable. Therefore, we take the ansatz

$$\hat{u}(z) = \frac{1}{2i\kappa_0}\big(u_0 + (z-1)\mathscr{U}(z)\big) =: \frac{1}{i\kappa_0}\mathcal{T}_-(u_0, \mathscr{U})(z),$$

with $\mathcal{T}_- : \mathbb{C} \times H^+(D) \to H^+(D)$. If we have an outgoing function f where $\mathcal{M}_{\kappa_0}\mathcal{L}f$ can be brought into a similar form as u_E, i.e. $\hat{f} = \frac{1}{i\kappa_0}\mathcal{T}_-(f_0, \mathscr{F})$, we obtain for

$$\begin{aligned} \mathcal{M}_{\kappa_0}\mathcal{L}f' &= \mathcal{M}_{\kappa_0}\big(s(\mathcal{L}f)(s) - f_0\big) = i\kappa_0\frac{z+1}{z-1}\frac{f_0 + (z-1)\mathscr{F}}{2i\kappa_0} - \frac{f_0}{z-1} \\ &= \frac{1}{2}\big(f_0 + (z+1)\mathscr{F}\big) =: \mathcal{T}_+(f_0, \mathscr{F}), \end{aligned}$$

with $\mathcal{T}_+ : \mathbb{C} \times H^+(D) \to H^+(D)$. This leads us to the transformation of the spatial derivative ∂_x,

$$\mathcal{M}_{\kappa_0}\mathcal{L}(\partial_x f) = \mathcal{T}_+(f_0, \mathscr{F}) = i\kappa_0 \mathcal{T}_+\mathcal{T}_-^{-1}\big(\mathcal{M}_{\kappa_0}\mathcal{L}f\big) =: \hat{\partial}_x\big(\mathcal{M}_{\kappa_0}\mathcal{L}f\big), \tag{7.6}$$

where the inverse of the operator $\mathcal{T}_-$ can be understood as a decomposition of $\mathcal{M}_{\kappa_0}\mathcal{L}f$ into its boundary value f_0 and a function with zero boundary values. Note that the existence of this inverse can be shown easily. Furthermore, we are going to use for a function f with $\mathcal{M}_{\kappa_0}\mathcal{L}f \in H^+(D)$ and an appropriate test function g the identity

$$\int_0^\infty f(r)g(r)\,dr = -2i\kappa_0\,A\big(\mathcal{M}_{\kappa_0}\mathcal{L}f, \mathcal{M}_{\kappa_0}\mathcal{L}g\big) \tag{7.7}$$

with

$$A\big(\hat{f}, \hat{g}\big) = \frac{1}{2\pi}\int_{S^1} \hat{f}(\overline{z})\hat{g}(z)\,|dz|, \qquad \text{for } \hat{f}, \hat{g} \in H^+(D),$$

and S^1 as the complex unit circle. Now we are in the position to derive a variational formulation in order to solve the Helmholtz equation numerically. Equation (7.1) leads for an appropriate test function v to the weak form

$$\int_0^a (u_I'v_I' - \kappa^2 p u_I v_I)dr + \int_0^\infty (u_E'v_E' - \kappa^2 u_E v_E)dr = g v_I(0).$$

Transforming the second integral with $\mathcal{M}_{\kappa_0}\mathcal{L}$ to S^1 results in a variational equation for $u_I \in H^1([0,a])$ and $\mathscr{U} \in H^+(D)$, i.e.

$$\begin{aligned}\int_0^a (u_I'v_I' - \kappa^2 p u_I v_I)dx - 2i\kappa_0 A\big(\mathcal{T}_+(u_0, \mathscr{U}),\, \mathcal{T}_+(v_0, \mathscr{V})\big)&\\ + \frac{2i\kappa^2}{\kappa_0}\, A\big(\mathcal{T}_-(u_0, \mathscr{U}),\, \mathcal{T}_-(v_0, \mathscr{V})\big) &= g v_I(0)\end{aligned}$$

for all $v_I \in H^1([0,a])$ and $\mathscr{V} \in H^+(D)$. Note that the two domains couple via the boundary degree of freedom $u_0 = u_I(a)$. The interior domain can be discretized with standard H^1 finite elements and $\mathscr{U}$ is approximated by the monomial basis $z^0, \ldots z^N$. The matrix representation $T_\pm^N \in \mathbb{C}^{(N+2)\times(N+2)}$ of the operator $\mathcal{T}_\pm\big|_{\mathbb{C}\times P^N} : \mathbb{C} \times P^N \to P^{N+1}$, with the

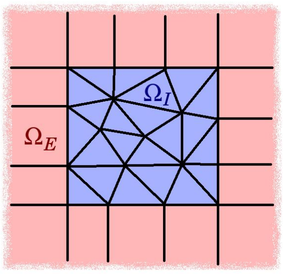

Figure 7.1: Geometry of the infinite problem.

polynomials P^N up to order N is consequently

$$T_{\pm}^N = \frac{1}{2}\begin{pmatrix} 1 & \pm 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \pm 1 \\ 0 & & & 1 \end{pmatrix}. \tag{7.8}$$

Thus, the coefficient vector $\underline{\hat{u}} \in \mathbb{C}^{N+2}$ of the $(N+1)$-order polynomial $\hat{u} = \mathcal{T}_{\pm}(u_0, \mathscr{U})$ can be obtained via $\underline{\hat{u}} = T_{\pm}^N(u_0, \underline{\mathscr{U}})$ from the coefficient vector $\underline{\mathscr{U}} \in \mathbb{C}^{N+1}$ of the polynomial $\mathscr{U}$. Because the monomials z^j are orthogonal with respect to A, i.e. $A(z^i, z^j) = \delta_{ij}$, the element matrix for the exterior domain reads as

$$M = -2i\kappa_0 T_+^{N\top} T_+^N + \frac{2i\kappa^2}{\kappa_0} T_-^{N\top} T_-^N.$$

Here again, the degree of freedom related to the first row and column, respectively, is the boundary degree of freedom of the interior domain.

7.2 The mixed hybrid formulation in the exterior domain

In this section we are going to state a variational formulation for solving a scattering problem in two dimensions, i.e. we are interested in the field u_{scat} scattered by an obstacle for a known incoming field u_{in}. Thus, we need to solve the Helmholtz equation on an unbounded domain Ω. As for the one dimensional problem from above, the computational domain can be divided into a bounded interior domain Ω_I, which is for simplicity reasons assumed to be rectangular, and an exterior domain Ω_E surrounding Ω_I (compare Figure 7.1) with an interface called Γ. While Ω_I is supposed to contain the obstacle, the exterior domain Ω_E is assumed to have constant material parameters ϵ and μ, and the incoming wave u_{in} solves the Helmholtz equation there.

As shown in Figure 7.1, the interior domain Ω_I is meshed by a finite element triangulation $\mathcal{T}$. The triangulation of Ω_E, we call $\mathcal{H}$, contains infinite strips, which can be seen as the extensions to elements of $\mathcal{T}$ next to the interface Γ. Thus, there are no hanging nodes on Γ. For completeness, at the four corner points of Γ right-angled triangles infinite in two directions have to be added to $\mathcal{H}$. The set of all infinite facets we call $\mathcal{F}_E$, and the finite facets are collected in $\mathcal{F}$.

7.2.1 The formal variational formulation

Now, we are in the position to state the variational equation for the scattering problem with HSIE. Its formal variational formulation is for $\tilde{u} = (u, \boldsymbol{\sigma}, u_F, \sigma_F)$

$$B_{s\Omega}(\tilde{u}, \tilde{v}) = F_{in}(\tilde{v}), \tag{7.9}$$

where the bilinear form $B_{s\Omega}$ was already defined in Formulation 5.2, with the difference, that now $\Omega = \Omega_I \cup \Omega_E$. Thus $\mathcal{T}$ in Formulation 5.2 has to be exchanged by $\mathcal{T} \cup \mathcal{H}$, i.e.

$$\begin{aligned} B_{s\Omega}(\tilde{u}, \tilde{v}) \; := \; \sum_{T \in \mathcal{T} \cup \mathcal{H}} \Big[& \big(i\omega\epsilon u, v\big)_T - \big(\operatorname{div} \boldsymbol{\sigma}, v\big)_T - \big(u, \operatorname{div}, \boldsymbol{\tau}\big)_T \\ & - \big(i\omega\mu\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_T + \big\langle u_F, \boldsymbol{\tau} \cdot \boldsymbol{n}_T \big\rangle_{\partial T} + \big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_T, v_F \big\rangle_{\partial T} \\ & + \beta \big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_F - \sigma_F, \boldsymbol{\tau} \cdot \boldsymbol{n}_F - \tau_F \big\rangle_{\partial T} + \alpha \big\langle u - u_F, v - v_F \big\rangle_{\partial T} \Big]. \end{aligned}$$

The linear form F_{in} reads as

$$\begin{aligned} F_{in}(\tilde{v}) \; := \; \sum_{T \in \mathcal{H}} \Big[& \big\langle u_{in}, \boldsymbol{\tau} \cdot \boldsymbol{n}_T \big\rangle_{\partial T \cap \Gamma} - \big\langle \boldsymbol{\sigma}_{in} \cdot \boldsymbol{n}_T, v_F \big\rangle_{\partial T \cap \Gamma} \\ & - \beta \big\langle \boldsymbol{\sigma}_{in} \cdot \boldsymbol{n}_F, \boldsymbol{\tau} \cdot \boldsymbol{n}_F - \tau_F \big\rangle_{\partial T \cap \Gamma} - \alpha \big\langle u_{in}, v - v_F \big\rangle_{\partial T \cap \Gamma} \Big], \end{aligned}$$

with $\boldsymbol{\sigma}_{in} = \frac{1}{i\omega\mu} \operatorname{grad} u_{in}$. Note that $B_{s\Omega}$ can be written as

$$B_{s\Omega}(\tilde{u}, \tilde{v}) = B_{s\Omega_I}(\tilde{u}, \tilde{v}) + B_{s\Omega_E}(\tilde{u}, \tilde{v}),$$

where $B_{s\Omega_I}$ and $B_{s\Omega_E}$ are defined according to $B_{s\Omega}$ with triangulations $\mathcal{T}$ and $\mathcal{H}$, respectively.

In order to explain this variational formulation, we remark that the variational equation in Formulation 5.2 is obtained from

$$B_{s\Omega}(\tilde{u},\tilde{v}) - \left\langle \boldsymbol{\sigma}\cdot\boldsymbol{n}_{\partial\Omega}, v_F \right\rangle_\Gamma = 0$$

by inserting absorbing boundary conditions. Because the incoming wave u_{in} solves the Helmholtz equation in Ω_E, the scattered field u_{scat} has to be a solution there as well. This gives rise to the variational formulations

$$\begin{aligned} B_{s\Omega_I}(\tilde{u},\tilde{v}) - \left\langle \boldsymbol{\sigma}\cdot\boldsymbol{n}_{\partial\Omega_I}, v_F \right\rangle_\Gamma &= 0 && \text{on } \Omega_I && (7.10)\\ B_{s\Omega_E}(\tilde{u}_{scat},\tilde{v}) - \left\langle \boldsymbol{\sigma}_{scat}\cdot\boldsymbol{n}_{\partial\Omega_E}, v_F \right\rangle_\Gamma &= 0 && \text{on } \Omega_E, && (7.11)\end{aligned}$$

where $\tilde{u}_{scat} = (u_{scat}, \boldsymbol{\sigma}_{scat}, u_{F,scat}, \sigma_{F,scat})$ contains the functions related to the scattered field. Using that on Γ the facet functions u_F ,σ_F, $u_{F,scat}$ and $\sigma_{F,scat}$ represent u, $\boldsymbol{\sigma}\cdot\boldsymbol{n}_F$, u_{scat} and $\boldsymbol{\sigma}_{scat}\cdot\boldsymbol{n}_F$, respectively, results in

$$\begin{aligned} u_{F,scat} &= u_F - u_{in} && \text{on } \Gamma\\ \sigma_{F,scat} &= \sigma_F - \boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_F && \text{on } \Gamma.\end{aligned}$$

Inserting this into (7.11) and adding (7.11) and (7.10) leads together with $-\boldsymbol{\sigma}_{scat}\cdot\boldsymbol{n}_{\partial\Omega_E} - \boldsymbol{\sigma}\cdot\boldsymbol{n}_{\partial\Omega_I} = \boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_{\partial\Omega_E}$ to (7.9).

Remark 7.3. *In the interior domain we solve for the total solution, thus $u|_{\Omega_I}$ and $\boldsymbol{\sigma}|_{\Omega_I}$ represent the total scalar and the total flux field, whereas in the exterior domain $u|_{\Omega_E}$ and $\boldsymbol{\sigma}|_{\Omega_E}$ stand for the scattered field and its flux field. The same distinction has to be made for the Lagrangian multipliers. While u_F and σ_F can be interpreted as the values of the total field u and the normal component of the total flux field $\boldsymbol{\sigma}$ on facets in $\mathcal{F}$, these two unknowns represent the scattered field u_{scat} and the normal component of the scattered flux field on the infinite Hardy element facets.*

7.2.2 The finite element spaces

In order to define the finite element spaces and to state the variational formulation, the integrals over the infinite elements of the triangulation $\mathcal{H}$ have to be transformed, like in the

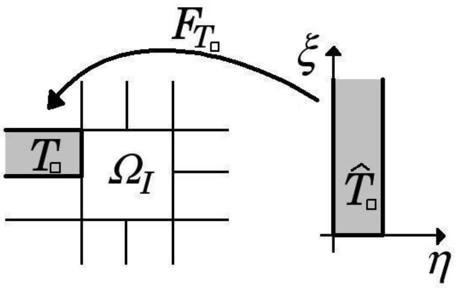

Figure 7.2: Geometry of the infinite problem.

one dimensional example, to integrals in Hardy spaces. For this purpose, the same technique as in [NHSS11], where a variational equation for time harmonic Maxwell's equations with HSIE is introduced, is used.

As shown in Figure 7.2, we assume that each infinite strip $T_\square$ of width h_T is obtained by an affine transformation $F_{T_\square}$ of the element $\hat{T}_\square = I_T \times [0,\infty)$ with $I_T = [0, h_T]$ (or $\hat{T}_\triangle = [0,\infty) \times [0,\infty)$ and $F_{T_\triangle}$ for infinite triangles $T_\triangle$). By $F_{T_\square}$ the infinite direction is mapped onto the ξ-direction. Thus, $F_{T_\square}$ corresponds to a translation, combined with a rotation by a multiple of $\frac{\pi}{2}$, and its Jacobian J is independent of ξ and η with $\det J = 1$. Furthermore, we need the transformation rules for surface integrals under the mapping $F_{T_\square}$, i.e.

Lemma 7.4. *Let $T \subset \mathbb{R}^3$ such that $T = F_T(\hat{T})$ with Jacobian matrix J_T. The outer normals to T and $\hat{T}$ are called $\boldsymbol{n}_T$ and $\boldsymbol{n}_{\hat{T}}$, and they transform according to Corollary 4.8. Furthermore, it is assumed that $f \in H^1(T)$ and $\boldsymbol{u}, \boldsymbol{v} \in H(\mathrm{div}, T)$ are obtained via an H^1- and $H(\mathrm{div})$-conforming transformation, respectively, (compare Lemma 4.10 and 4.14) from $f_{\hat{T}}$ and $\boldsymbol{u}_{\hat{T}}, \boldsymbol{v}_{\hat{T}}$. Defining $J_n := |\det J_T| \| J_T^{-\top} \boldsymbol{n}_{\hat{T}} \|$ the surface integrals transform as*

$$\begin{aligned}
\int_{\partial T} f \, ds &= \int_{\partial \hat{T}} f_{\hat{T}} \, J_n \, d\hat{s}, \\
\int_{\partial T} (\boldsymbol{v} \cdot \boldsymbol{n}_T) \, f \, ds &= \mathrm{sign}(\det J) \int_{\partial \hat{T}} (\boldsymbol{v}_{\hat{T}} \cdot \boldsymbol{n}_{\hat{T}}) f_{\hat{T}} \, d\hat{s} \\
\int_{\partial T} (\boldsymbol{u} \cdot \boldsymbol{n}_T)(\boldsymbol{v} \cdot \boldsymbol{n}_T) \, ds &= \int_{\partial \hat{T}} (\boldsymbol{u}_{\hat{T}} \cdot \boldsymbol{n}_{\hat{T}})(\boldsymbol{v}_{\hat{T}} \cdot \boldsymbol{n}_{\hat{T}}) \, J_n^{-1} \, d\hat{s}.
\end{aligned}$$

For a proof see [Mon03], section 3.9.

In order to define the space for the scalar field u, which is an L^2 function in Ω_I, we investigate the mass integral over an infinite strip $T_\square$

$$\big(u, v\big)_{T_\square} = \int_0^{h_T} \int_0^\infty (u \circ F_{T_\square})(v \circ F_{T_\square}) |\det J| \, d\xi d\eta.$$

By applying the Laplace and the Möbius transform with respect to ξ-direction, we get for $\eta \in [0, h]$

$$\begin{aligned} \hat{u}(\eta, \bullet) &:= \mathcal{M}_{\kappa_0}\mathcal{L}\big(u \circ F_{T_\square}(\eta, \bullet)\big) \in H^+(D), \\ \hat{v}(\eta, \bullet) &:= \mathcal{M}_{\kappa_0}\mathcal{L}\big(v \circ F_{T_\square}(\eta, \bullet))\big) \in H^+(D), \end{aligned}$$

and consequently by using (7.7)

$$\big(u, v\big)_{T_\square} = -2i\kappa_0 \int_0^{h_T} A\big(\hat{u}(\eta, \bullet), \hat{v}(\eta, \bullet)\big) d\eta =: -2i\kappa_0 A_{T_\square}(\hat{u}, \hat{v}).$$

Applying the rotation $F_{T_\triangle}$ to the mass integral for an infinite triangle $T_\triangle$ reads as

$$\big(u, v\big)_{T_\triangle} = \int_0^\infty \int_0^\infty (u \circ F_{T_\triangle})(v \circ F_{T_\triangle}) |\det J| \, d\xi d\eta.$$

Note that $\det J = 1$. Taking the Laplace and Möbius transform of u and v with respect to both coordinate axis, i.e considering

$$\hat{\hat{u}}(*, \hat{\xi}) := \mathcal{M}_{\kappa_0}\mathcal{L}\hat{u}(*, \hat{\xi}) \in H^+(D) \tag{7.12}$$

for all $\hat{\xi} \in S^1$, and defining

$$\begin{aligned} \int_0^\infty \int_0^\infty u(\eta, \xi) v(\eta, \xi) \, d\xi d\eta &= -4\kappa_0^2 A_{T_\triangle}(\hat{\hat{u}}, \hat{\hat{v}}) \\ \text{with} \qquad A_{T_\triangle}(\hat{\hat{u}}, \hat{\hat{v}}) &:= \frac{1}{4\pi^2} \int_{S^1} \int_{S^1} \hat{\hat{u}}(\overline{\hat{\eta}}, \overline{\hat{\xi}}) \, \hat{\hat{v}}(\hat{\eta}, \hat{\xi}) \, |d\hat{\xi}||d\hat{\eta}|, \end{aligned}$$

we get

$$\big(u, v\big)_{T_\triangle} = -4\kappa_0^2 A_{T_\triangle}(\hat{\hat{u}}, \hat{\hat{v}}).$$

This leads us to the definition of the space where the scalar field u is belonging to,

$$\begin{aligned} U_H \;:=\; & \Big\{ v : \Omega \to \mathbb{C},\; v|_T \in L^2(T) \quad \forall T \in \mathcal{T}, \\ & \qquad |A_{T_\Box}(\hat{v},\hat{v})| < \infty \quad \forall T_\Box \in \mathcal{H}, \quad |A_{T_\triangle}(\hat{v},\hat{v})| < \infty \quad \forall T_\triangle \in \mathcal{H} \Big\}. \end{aligned} \tag{7.13}$$

In order to find an appropriate space for the vector valued field $\boldsymbol{\sigma}$, the integrals $\big(\boldsymbol{\sigma},\boldsymbol{\sigma}\big)_T$ and $\big(\operatorname{div}\boldsymbol{\sigma},\operatorname{div}\boldsymbol{\sigma}\big)_T$ for infinite elements T need to be bounded. Following the transformation rules for $H(\operatorname{div},T)$-functions from Lemma 4.14, we get for the Laplace and Möbius transform of $\boldsymbol{\sigma}_{\hat{T}_\Box}$

$$\hat{\boldsymbol{\sigma}}_{\hat{T}_\Box}(\eta,\bullet) := (\det J) J^{-1} \begin{pmatrix} \mathcal{M}_{\kappa_0}\mathcal{L}\big(\sigma_1 \circ F_{T_\Box}(\eta,\bullet)\big) \\ \mathcal{M}_{\kappa_0}\mathcal{L}\big(\sigma_2 \circ F_{T_\Box}(\eta,\bullet)\big) \end{pmatrix} \qquad \text{for } \eta \in I_T. \tag{7.14}$$

Using this, we obtain for strips $T_\Box$ and a rotation $F_{T_\Box}$, i.e. $\det J = 1$ and $J^\top J = I$,

$$\begin{aligned} \big(\boldsymbol{\sigma},\boldsymbol{\tau}\big)_{T_\Box} &= \int_0^{h_T}\int_0^\infty (\boldsymbol{\sigma}\circ F_{T_\Box})\cdot(\boldsymbol{\tau}\circ F_{T_\Box})\,|\det J| d\xi d\eta \\ &= \int_0^{h_T}\int_0^\infty \boldsymbol{\sigma}_{\hat{T}_\Box}\cdot\boldsymbol{\tau}_{\hat{T}_\Box}\, d\xi d\eta = -2i\kappa_0 A_{IT_\Box}(\hat{\boldsymbol{\sigma}}_{\hat{T}_\Box},\hat{\boldsymbol{\tau}}_{\hat{T}_\Box}), \end{aligned} \tag{7.15}$$

where

$$A_{IT_\Box}(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}) := \sum_{j=1}^{2}\int_0^{h_T} A\big(\hat{u}_j(\eta,\bullet),\hat{v}_j(\eta,\bullet)\big)\,d\eta.$$

Applying the transformation rules for the divergence, we obtain for the second integral

$$\begin{aligned} \big(\operatorname{div}\boldsymbol{\sigma},\operatorname{div}\boldsymbol{\tau}\big)_{T_\Box} &= \int_0^{h_T}\int_0^\infty (\operatorname{div}\boldsymbol{\sigma}\circ F_{T_\Box})(\operatorname{div}\boldsymbol{\tau}\circ F_{T_\Box})|\det J|\, d\xi d\eta \\ &= \int_0^{h_T}\int_0^\infty \frac{1}{|\det J|}(\operatorname{div}_{\xi\eta}\boldsymbol{\sigma}_{\hat{T}_\Box})(\operatorname{div}_{\xi\eta}\boldsymbol{\tau}_{\hat{T}_\Box})\, d\xi d\eta \\ &= -2i\kappa_0 A_{T_\Box}(\widehat{\operatorname{div}}\hat{\boldsymbol{\sigma}}_{\hat{T}_\Box},\widehat{\operatorname{div}}\hat{\boldsymbol{\tau}}_{\hat{T}_\Box}), \end{aligned} \tag{7.16}$$

with

$$\widehat{\operatorname{div}}\hat{\boldsymbol{u}} := (\partial_\eta\otimes\operatorname{id})(\hat{\boldsymbol{u}})_1 + (\operatorname{id}\otimes\hat{\partial}_\xi)(\hat{\boldsymbol{u}})_2,$$

where id denotes the identity. Note that for the derivative in ξ-direction the transformed derivative $\hat{\partial}_\xi$ is needed, while the original derivative ∂_η is still used in η direction. For an infinite triangle $T_\triangle$, similar calculations lead to

$$\begin{aligned} \big(\boldsymbol{\sigma},\boldsymbol{\tau}\big)_{T_\triangle} &= -4\kappa_0^2 A_{IT_\triangle}\big(\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle},\hat{\hat{\boldsymbol{\tau}}}_{\hat{T}_\triangle}\big) && (7.17)\\ \big(\operatorname{div}\boldsymbol{\sigma},\operatorname{div}\boldsymbol{\tau}\big)_{T_\triangle} &= -4\kappa_0^2 A_{T_\triangle}(\widehat{\widehat{\operatorname{div}}}\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle},\widehat{\widehat{\operatorname{div}}}\hat{\hat{\boldsymbol{\tau}}}_{\hat{T}_\triangle}), && (7.18) \end{aligned}$$

with

$$\begin{aligned} \widehat{\widehat{\operatorname{div}}}\hat{\hat{\boldsymbol{u}}} &:= (\hat{\partial}_\eta\otimes\mathrm{id})(\hat{\hat{\boldsymbol{u}}})_1 + (\mathrm{id}\otimes\hat{\partial}_\xi)(\hat{\hat{\boldsymbol{u}}})_2,\\ A_{IT_\triangle}(\hat{\hat{\boldsymbol{u}}},\hat{\hat{\boldsymbol{v}}}) &:= \sum_{j=1}^{2} A_{T_\triangle}(\hat{\hat{u}}_j,\hat{\hat{v}}_j) \end{aligned}$$

and $\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle}$ according to (7.14) together with (7.12). Using this, we get for the space of $\boldsymbol{\sigma}$

$$\begin{aligned} V_H := \Big\{&\boldsymbol{\sigma} : \Omega\to\mathbb{C}^2,\ \boldsymbol{\sigma}|_T\in H(\operatorname{div},T)\quad \forall T\in\mathcal{T}, && (7.19)\\ &|A_{T_\square}(\widehat{\operatorname{div}}\hat{\boldsymbol{\sigma}}_{\hat{T}_\square},\widehat{\operatorname{div}}\hat{\boldsymbol{\sigma}}_{\hat{T}_\square})|<\infty\,,|A_{IT_\square}(\hat{\boldsymbol{\sigma}}_{\hat{T}_\square},\hat{\boldsymbol{\sigma}}_{\hat{T}_\square})|<\infty,\quad \forall T_\square\in\mathcal{H},\\ &|A_{T_\triangle}(\widehat{\widehat{\operatorname{div}}}\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle},\widehat{\widehat{\operatorname{div}}}\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle})|<\infty\,,|A_{IT_\triangle}(\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle},\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle})|<\infty,\quad \forall T_\triangle\in\mathcal{H}\Big\}. \end{aligned}$$

Finally, in order to define the space of the facet functions u_F and σ_F, respectively, the integral $\big\langle u_F,u_F\big\rangle_F$ for an infinite facet $F\in\mathcal{F}_E$ needs to be bounded. In the following, we use the parametrization $\gamma_F(\xi)=\boldsymbol{x}+\xi\boldsymbol{e}$, $\xi\in[0,\infty]$ for F with $\boldsymbol{x}$ as the starting point of the facet and $\boldsymbol{e}$ as its normalized directional vector. This leads to

$$\big\langle u_F,v_F\big\rangle_F=\int_0^\infty (u_F\circ\gamma_F)(v_F\circ\gamma_F)\,d\xi=-2i\kappa_0 A(\hat{u}_F,\hat{v}_F),$$

with the Laplace and Möbius transform $\hat{u}_F=\mathcal{M}_{\kappa_0}\mathcal{L}(u_F\circ\gamma_F)$, and the facet spaces read as

$$\begin{aligned} U_{HF}=V_{HF} := \Big\{&v : \mathcal{F}\cup\mathcal{F}_E\to\mathbb{C},\quad v|_F=L^2(F),\quad \forall F\in\mathcal{F},\\ &|A(\hat{v}_F,\hat{v}_F)|<\infty,\quad \forall F\in\mathcal{F}_E\Big\} && (7.20) \end{aligned}$$

7.2.3 The integrals involved

In order to evaluate the bilinear form $B_{s\Omega_E}$ for the exterior domain, the integrals involved need to be discussed. Thus, we get by taking the Laplace and Möbius transform for a strip

$T_\square$

$$\big(\operatorname{div}\boldsymbol{\sigma}, v\big)_{T_\square} = \int_0^{h_T}\int_0^\infty \frac{1}{\det J}\operatorname{div}\boldsymbol{\sigma}_{\hat{T}_\square}(v\circ F_{T_\square})|\det J|d\xi d\eta = -2i\kappa_0 A_{T_\square}\big(\widehat{\operatorname{div}\hat{\boldsymbol{\sigma}}}_{\hat{T}_\square}, \hat{v}\big),$$

and for an infinite triangle $T_\triangle$

$$\big(\operatorname{div}\boldsymbol{\sigma}, v\big)_{T_\triangle} = -4\kappa_0^2 A_{T_\triangle}\big(\widehat{\widehat{\operatorname{div}}}\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle}, \hat{\hat{v}}\big).$$

Next, we investigate the boundary integrals for the infinite elements. Because the Jacobian J of the transformations between the reference elements $\hat{T}$ and the element T is a rotational matrix, $J_n = |\det J|\|J^{-T}\boldsymbol{n}_{\hat{T}}\|$ needed for the transformation of surface integrals is one. After transforming a facet function w_F to the reference element, we will, in order to avoid confusions, distinguish between $w_{Fl} = w_F\circ F_{T_\square}|_{\eta=0}, w_{Fr} = w_F\circ F_{T_\square}|_{\eta=h_T}$ and $w_{Fb} = u_F\circ F_{T_\square}|_{\xi=0}$ for infinite strips $T_\square$. For infinite triangles just w_{Fl} and w_{Fb} are needed.

Following the transformation rules from Lemma 7.4, we get for

$$\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_T, v_F\big\rangle_{\partial T} = \big\langle\boldsymbol{\sigma}_{\hat{T}}\cdot\boldsymbol{n}_{\hat{T}}, v_F\circ F_T\big\rangle_{\partial\hat{T}}.$$

Note that if we exchange $\boldsymbol{n}_T$ by $\boldsymbol{n}_F$, we have to take account of the sign by a factor $\boldsymbol{n}_T\cdot\boldsymbol{n}_F$. For an infinite strip $T_\square$ we get

$$\begin{aligned}\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_{T_\square}, v_F\big\rangle_{\partial T_\square} &= \int_0^\infty \sigma_{\hat{T}_\square,\eta}(h_T,\xi)\,v_{Fr}(\xi)\,d\xi - \int_0^\infty \sigma_{\hat{T}_\square,\eta}(0,\xi)\,v_{Fl}(\xi)\,d\xi\\ &\quad - \int_0^{h_T}\sigma_{\hat{T}_\square,\xi}(\eta,0)v_{Fb}(\eta)\,d\eta\\ &= 2i\kappa_0\Big(A\big(\hat{\sigma}_{\hat{T}_\square,\eta}(0,\bullet),\hat{v}_{Fl}(\bullet)\big) - A\big(\hat{\sigma}_{\hat{T}_\square,\eta}(h_T,\bullet),\hat{v}_{Fr}(\bullet)\big)\\ &\qquad -\big\langle\hat{\sigma}_{\hat{T}_\square,\xi}(\bullet,1), v_{Fb}(\bullet)\big\rangle_{I_T}\Big),\end{aligned}$$

where $\sigma_{\hat{T}_\square,\xi}(\eta,0) = 2i\kappa_0\hat{\sigma}_{\hat{T}_\square,\xi}(\eta,1)$ was used. For an infinite triangle $T_\triangle$ a similar computation gives

$$\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_{T_\triangle}, v_F\big\rangle_{\partial T_\triangle} = -4\kappa_0^2\Big(A\big(\hat{\hat{\sigma}}_{\hat{T}_\triangle,\eta}(1,\bullet),\hat{v}_{Fl}(\bullet)\big) + A\big(\hat{\hat{\sigma}}_{\hat{T}_\triangle,\xi}(\bullet,1),\hat{v}_{Fb}(\bullet)\big)\Big).$$

Under the assumption that J is a rotational matrix, the integral $\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_F,\boldsymbol{\tau}\cdot\boldsymbol{n}_F\big\rangle_{\partial T}$ transforms as

$$\big\langle\boldsymbol{\sigma}\cdot\boldsymbol{n}_F,\boldsymbol{\tau}\cdot\boldsymbol{n}_F\big\rangle_{\partial T} = \big\langle\boldsymbol{\sigma}_{\hat{T}}\cdot\boldsymbol{n}_{\hat{T}},\boldsymbol{\tau}_{\hat{T}}\cdot\boldsymbol{n}_{\hat{T}}\big\rangle_{\partial\hat{T}}.$$

Taking the Laplace and Möbius transform leads to

$$\begin{aligned}
\left\langle \boldsymbol{\sigma}\cdot\boldsymbol{n}_{T_\square}, \boldsymbol{\tau}\cdot\boldsymbol{n}_{T_\square}\right\rangle_{\partial T_\square} &= -2i\kappa_0\Big(A\big(\hat{\sigma}_{\hat{T}_\square,\eta}(0,\bullet), \hat{\tau}_{\hat{T}_\square,\eta}(0,\bullet)\big) + A\big(\hat{\sigma}_{\hat{T}_\square,\eta}(h_T,\bullet), \hat{\tau}_{\hat{T}_\square,\eta}(h_T,\bullet)\big)\Big) \\
&\qquad -4\kappa_0^2\big\langle \hat{\sigma}_{\hat{T}_\square,\xi}(\bullet,1), \hat{\tau}_{\hat{T}_\square,\xi}(\bullet,1)\big\rangle_{I_T}, \\
\left\langle \boldsymbol{\sigma}\cdot\boldsymbol{n}_{T_\triangle}, \boldsymbol{\tau}\cdot\boldsymbol{n}_{T_\triangle}\right\rangle_{\partial T_\triangle} &= 8i\kappa_0^3\Big(A\big(\hat{\hat{\sigma}}_{\hat{T}_\triangle,\eta}(1,\bullet), \hat{\hat{\tau}}_{\hat{T}_\triangle,\eta}(1,\bullet)\big) + A\big(\hat{\hat{\sigma}}_{\hat{T}_\triangle,\xi}(\bullet,1), \hat{\hat{\tau}}_{\hat{T}_\triangle,\xi}(\bullet,1)\big)\Big).
\end{aligned}$$

For completeness, we give the following integrals, which are obtained by similar calculations

$$\begin{aligned}
\left\langle u, v\right\rangle_{\partial T_\square} &= -2i\kappa_0\Big(A\big(\hat{u}(0,\bullet), \hat{v}(0,\bullet)\big) + A\big(\hat{u}(h_T,\bullet), \hat{v}(h_T,\bullet)\big)\Big) \\
&\qquad -4\kappa_0^2\big\langle \hat{u}(\bullet,1), \hat{v}(\bullet,1)\big\rangle_{I_T}, \\
\left\langle u, v\right\rangle_{\partial T_\triangle} &= 8i\kappa_0^3\Big(A\big(\hat{\hat{u}}(1,\bullet), \hat{\hat{v}}(1,\bullet)\big) + A\big(\hat{\hat{u}}(\bullet,1), \hat{\hat{v}}(\bullet,1)\big)\Big), \\
\left\langle u, v_F\right\rangle_{\partial T_\square} &= -2i\kappa_0\Big(A\big(\hat{u}(0,\bullet), \hat{v}_{Fl}(\bullet)\big) + A\big(\hat{u}(h_T,\bullet), \hat{v}_{Fr}(\bullet)\big)\Big) - 2i\kappa_0\big\langle \hat{u}(\bullet,1), v_{Fb}(\bullet)\big\rangle_{I_T}, \\
\left\langle u, v_F\right\rangle_{\partial T_\triangle} &= 4\kappa_0^2\Big(A\big(\hat{\hat{u}}(1,\bullet), \hat{v}_{Fl}(\bullet)\big) + A\big(\hat{\hat{u}}(\bullet,1), \hat{v}_{Fb}(\bullet)\big)\Big), \\
\left\langle u_F, v_F\right\rangle_{\partial T_\square} &= -2i\kappa_0\Big(A\big(\hat{u}_{Fl}(\bullet), \hat{v}_{Fl}(\bullet)\big) + A\big(\hat{u}_{Fr}(\bullet), \hat{v}_{Fr}(\bullet)\big)\Big) + \big\langle u_{Fb}(\bullet), v_{Fb}(\bullet)\big\rangle_{I_T}, \\
\left\langle u_F, v_F\right\rangle_{\partial T_\triangle} &= -2i\kappa_0\Big(A\big(\hat{u}_{Fl}(\bullet), \hat{v}_{Fl}(\bullet)\big) + A\big(\hat{u}_{Fr}(\bullet), \hat{v}_{Fb}(\bullet)\big)\Big),
\end{aligned}$$

and concerning the right hand side we list the terms including volume functions

$$\begin{aligned}
\left\langle u_{in}, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\right\rangle_{\partial T_\square\cap\Gamma} &= -2i\kappa_0\big\langle (u_{in}\circ F_{T_\square})(\bullet,0), \hat{\tau}_{\hat{T}_\square,\xi}(\bullet,1)\big\rangle_{I_T}, \\
\left\langle \boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_F, \boldsymbol{\tau}\cdot\boldsymbol{n}_F\right\rangle_{\partial T_\square\cap\Gamma} &= -2i\kappa_0\Big\langle \big((\boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_{T_\square})\circ F_{T_\square}\big)(\bullet,0), \hat{\tau}_{\hat{T}_\square,\xi}(\bullet,1)\Big\rangle_{I_T}, \\
\left\langle u_{in}, v\right\rangle_{\partial T_\square\cap\Gamma} &= 2i\kappa_0\big\langle (u_{in}\circ F_{T_\square})(\bullet,0), \hat{v}(\bullet,1)\big\rangle_{I_T}.
\end{aligned}$$

7.3 The finite elements

In order to solve the variational formulation of the last section, we discuss the set of finite elements needed to discretize the problem.

7.3.1 The finite element basis functions

Because of the special form of the infinite elements and their transformation to the reference element, it is useful to take a tensor product ansatz for the volume basis functions on the reference element. Thus, a basis function for $\boldsymbol{\sigma}_T$ has the form

$$\boldsymbol{\sigma}_{\hat{T}}(\eta,\xi) = \begin{pmatrix} \sigma^1_\eta(\eta)\,\sigma^1_\xi(\xi) \\ \sigma^2_\eta(\eta)\,\sigma^2_\xi(\xi) \end{pmatrix},$$

where the superscript 1 is used for the η component of $\boldsymbol{\sigma}_{\hat{T}}$ and 2 for the ξ component. By taking the Laplace and Möbius transform we obtain on the strip $T_\square$ and the triangle $T_\triangle$

$$\hat{\boldsymbol{\sigma}}_{\hat{T}_\square} = \begin{pmatrix} \sigma^1_\eta\,\hat{\sigma}^1_\xi \\ \sigma^2_\eta\,\hat{\sigma}^2_\xi \end{pmatrix} \qquad \text{and} \qquad \hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle} = \begin{pmatrix} \hat{\sigma}^1_\eta\,\hat{\sigma}^1_\xi \\ \hat{\sigma}^2_\eta\,\hat{\sigma}^2_\xi \end{pmatrix},$$

respectively. We will first concentrate onto the infinite triangle. Because $\boldsymbol{\sigma}$ is element wise in $H(\mathrm{div})$, we construct Raviart-Thomas like elements. Thus, if we take the ansatz from the one dimensional model problem for the η-part of the first component $\hat{\sigma}^1_\eta \in H^+(D)$, $\hat{\sigma}^1_\eta = \frac{1}{i\kappa_0}\mathcal{T}_-(\sigma^1_{\eta 0}, \mathscr{S}^1_\eta)$, i.e. we decompose the function into the boundary value $\mathbb{C} \supset \sigma^1_{\eta 0} = \sigma^1_\eta(0)$ and a function $\mathscr{S}^1_\eta \in H^+(D)$ with zero boundary values, the the ξ-part of the first component $\hat{\sigma}^1_\xi$ plays the role of a derivative. Consequently, we take the ansatz for $\hat{\sigma}^1_\xi = \mathcal{T}_+(\sigma^1_{\xi 0}, \mathscr{S}^1_\xi)$ with $\sigma^1_{\xi 0} \in \mathbb{C}$ and $\mathscr{S}^1_\xi \in H^+(D)$. The same argumentation for the second component and the infinite strip leads to

$$\hat{\boldsymbol{\sigma}}_{\hat{T}_\square} = \begin{pmatrix} \sigma^1_\eta\,\mathcal{T}_+(\sigma^1_{\xi 0}, \mathscr{S}^1_\xi) \\ \frac{1}{i\kappa_0}\,\sigma^2_\eta\,\mathcal{T}_-(\sigma^2_{\xi 0}, \mathscr{S}^2_\xi) \end{pmatrix} \quad \text{and} \quad \hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle} = \begin{pmatrix} \frac{1}{i\kappa_0}\,\mathcal{T}_-(\sigma^1_{\eta 0}, \mathscr{S}^1_\eta)\,\mathcal{T}_+(\sigma^1_{\xi 0}, \mathscr{S}^1_\xi) \\ \frac{1}{i\kappa_0}\,\mathcal{T}_+(\sigma^2_{\eta 0}, \mathscr{S}^2_\eta)\,\mathcal{T}_-(\sigma^2_{\xi 0}, \mathscr{S}^2_\xi) \end{pmatrix}.$$

When discretizing the problem, we use for the functions in $H^+(D)$, $\mathscr{S}^1_\xi(z), \mathscr{S}^1_\eta(z), \mathscr{S}^2_\xi(z)$ and $\mathscr{S}^2_\eta(z)$ a polynomial ansatz with the monomials $z^0, z^1, \ldots, z^N$ as basis functions. If the interior problem is discretized with Raviart Thomas elements of order p, the η-part of $\operatorname{div}\boldsymbol{\sigma}$ has to be a polynomial of order p, i.e. it has to be from $P^p(I_T)$. Thus, we take $\sigma^1_\eta \in P^{p+1}(I_T)$, and $\sigma^2_\eta \in P^p(I_T)$. As we will see later on, taking integrated Legendre polynomials as basis functions for σ^1_η and Legendre polynomials as basis functions for σ^2_η leads to sparse element matrices. For more information on Legendre and integrated Legendre Polynomials see Section 8.2.1. Thus, in total there are $(2p+3)(N+2)$ and $2(N+2)^2$, respectively, $\boldsymbol{\sigma}$-degrees on an element.

Using the product ansatz for a basis function of u, i.e. $u \circ F_T = u_\eta(\eta)\,u_\xi(\xi)$ and taking the Laplace and Möbius transform results in

$$\hat{u} = u_\eta\,\hat{u}_\xi \quad \text{for } T = T_\square, \qquad \text{and} \qquad \hat{\hat{u}} = \hat{u}_\eta\,\hat{u}_\xi \quad \text{for } T = T_\triangle.$$

Due to the fact that u represents the divergence of $\boldsymbol{\sigma}$, and the divergence of a $\boldsymbol{\sigma}$ basis function is

$$\begin{aligned}
\widehat{\mathrm{div}}\hat{\boldsymbol{\sigma}}_{\hat{T}_\square} &= \partial_n \sigma_\eta^1\, \mathcal{T}_+\big(\sigma_{\xi 0}^1, \mathscr{S}_\xi^1\big) + \sigma_\eta^2\, \mathcal{T}_+\big(\sigma_{\xi 0}^2, \mathscr{S}_\xi^2\big),\\
\widehat{\widehat{\mathrm{div}}}\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle} &= \mathcal{T}_+\big(\sigma_{\eta 0}^1, \mathscr{S}_\eta^1\big)\, \mathcal{T}_+\big(\sigma_{\xi 0}^1, \mathscr{S}_\xi^1\big) + \mathcal{T}_+\big(\sigma_{\eta 0}^2, \mathscr{S}_\eta^2\big)\, \mathcal{T}_+\big(\sigma_{\xi 0}^2, \mathscr{S}_\xi^2\big),
\end{aligned}$$

the ansatz

$$\begin{aligned}
\hat{u} &= u_\eta\, \mathcal{T}_+(u_{\xi 0}, \mathscr{U}_\xi) && \text{for } T = T_\square,\\
\hat{\hat{u}} &= \mathcal{T}_+(u_{\eta 0}, \mathscr{U}_\eta)\, \mathcal{T}_+(u_{\xi 0}, \mathscr{U}_\xi) && \text{for } T = T_\triangle
\end{aligned}$$

is appropriate. For $\mathscr{U}_\eta$ and $\mathscr{U}_\xi$ again the monomials up to order N are used as basis, and $u_\eta \in P^p(I_T)$ is expanded into a Legendre polynomial basis. This leads to $(p+1)(N+2)$ and $(N+2)^2$ degrees of freedom on $T_\square$ and $T_\triangle$, respectively.

For the facet functions, which represent the values of u and $\boldsymbol{\sigma} \cdot \boldsymbol{n}_T$ on the facet, we have to take

$$\hat{u}_F = \mathcal{T}_+(u_{F0}, \mathscr{U}_F), \qquad \text{and} \qquad \hat{\sigma}_F = \mathcal{T}(\sigma_{F0}, \mathscr{S}_F),$$

with a monomial basis up to order N for $\mathscr{U}_F$ and $\mathscr{S}_F$ in order to be consistent.

7.3.2 The element matrices

We finish the chapter about Hardy space infinite elements with a discussion on the element matrix obtained by using this set of basis functions.

Inserting the ansatz just discussed for $\boldsymbol{\sigma}$ and the same ansatz for $\boldsymbol{\tau}$ into the representation of the mass integral (7.15) and (7.17) leads to

$$\begin{aligned}
\big(\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_{T_\square} &= -2i\kappa_0 \big\langle \sigma_\eta^1, \tau_\eta^1 \big\rangle_{I_T}\, A\big(\mathcal{T}_+(\sigma_{\xi 0}^1, \mathscr{S}_\xi^1), \mathcal{T}_+(\tau_{\xi 0}^1, \mathscr{T}_\xi^1)\big)\\
&\quad + \frac{2i}{\kappa_0} \big\langle \sigma_\eta^2, \tau_\eta^2 \big\rangle_{I_T}\, A\big(\mathcal{T}_-(\sigma_{\xi 0}^2, \mathscr{S}_\xi^2), \mathcal{T}_-(\tau_{\xi 0}^2, \mathscr{T}_\xi^2)\big)\\
\big(\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_{T_\triangle} &= 4\, A\big(\mathcal{T}_-(\sigma_{\eta 0}^1, \mathscr{S}_\eta^1), \mathcal{T}_-(\tau_{\eta 0}^1, \mathscr{T}_\eta^1)\big) A\big(\mathcal{T}_+(\sigma_{\xi 0}^1, \mathscr{S}_\xi^1), \mathcal{T}_+(\tau_{\xi 0}^1, \mathscr{T}_\xi^1)\big)\\
&\quad + 4\, A\big(\mathcal{T}_+(\sigma_{\eta 0}^2, \mathscr{S}_\eta^2), \mathcal{T}_+(\tau_{\eta 0}^2, \mathscr{T}_\eta^2)\big) A\big(\mathcal{T}_-(\sigma_{\xi 0}^2, \mathscr{S}_\xi^2), \mathcal{T}_-(\tau_{\xi 0}^2, \mathscr{T}_\xi^2)\big).
\end{aligned}$$

The resulting element matrix just contains entries between degrees of freedoms which are associated to the same spatial component. For the triangular element these blocks correspond up to a constant to a Kronecker product of the matrices $T_-^{N\top} T_-^N$ and $T_+^{N\top} T_+^N$, which are both three diagonal. For infinite strips these blocks can be calculated due to the integral relations of Legendre and integrated Legendre polynomials as the Kronecker product of a band matrix and a diagonal matrix, respectively, with $T_+^{N\top} T_+^N$.

With similar arguments one can conclude that the element matrix entries corresponding to $\big(u, v\big)_T$ can be written again as a Kronecker product of a diagonal matrix and $T_+^{N\top} T_+^N$, respectively, with $T_+^{N\top} T_+^N$ for the elements $T_\Box$ and $T_\triangle$.

For the term coupling $\boldsymbol{\sigma}$ and u one gets

$$\begin{aligned}\big(\operatorname{div}\boldsymbol{\sigma}, v\big)_{T_\Box} &= -2i\kappa_0 \big\langle \partial_\eta \sigma^1_\eta, v_\eta \big\rangle_{I_T} A\big(\mathcal{T}_+(\sigma^1_{\xi 0}, \mathscr{S}_\xi), \mathcal{T}_+(v_{\xi 0}, \mathscr{V}_\xi)\big) \\ &\quad -2i\kappa_0 \big\langle \sigma^2_\eta, v_\eta \big\rangle_{I_T} A\big(\mathcal{T}_+(\sigma^2_{\xi 0}, \mathscr{S}_\xi), \mathcal{T}_+(v_{\xi 0}, \mathscr{V}_\xi)\big).\end{aligned}$$

Because the matrix representation for $\big\langle \partial_\eta \sigma^1_\eta, v_\eta \big\rangle_{I_T}$ and $\big\langle \sigma^2_\eta, v_\eta \big\rangle_{I_T}$ are diagonal matrices, the resulting element matrix blocks are again of the form $D \otimes T_+^{N\top} T_+^N$ with a diagonal matrix D. For infinite triangles D has to be exchanged by $-4\kappa_0^2 T_+^{N\top} T_+^N$.

Representatively for the boundary integrals we investigate $\big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_T, v_F \big\rangle_{\partial T}$. The other boundary integrals have similar properties concerning their element matrix contribution. For this integral we get

$$\begin{aligned}\big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_{T_\Box}, v_F \big\rangle_{\partial T_\Box} &= 2i\kappa_0 \; \sigma^1_\eta(0) \; A\big(\mathcal{T}_+(\sigma^1_{\xi 0}, S^1_\xi), \mathcal{T}_+(v_{Fl0}, \mathscr{V}_{Fl})\big) \\ &\quad -2i\kappa_0 \; \sigma^1_\eta(h_T) \; A\big(\mathcal{T}_+(\sigma^1_{\xi 0}, S^1_\xi), \mathcal{T}_+(v_{Fr0}, \mathscr{V}_{Fr})\big) \\ &\quad -\big\langle \sigma^2_\eta, u_{Fb} \big\rangle_{I_T} \sigma^2_{\xi 0}, \\ \big\langle \boldsymbol{\sigma} \cdot \boldsymbol{n}_{T_\triangle}, v_F \big\rangle_{\partial T_\triangle} &= 2i\kappa_0^2 \; \sigma^1_{\eta 0} \; A\big(\mathcal{T}_+(\sigma^1_{\xi 0}, S^1_\xi), \mathcal{T}_+(v_{Fl0}, \mathscr{V}_{Fl})\big) \\ &\quad 2i\kappa_0^2 \; \sigma^2_{\xi 0} \; A\big(\mathcal{T}_+(\sigma^2_{\eta 0}, S^2_\eta), \mathcal{T}_+(v_{Fb0}, \mathscr{V}_{Fb})\big).\end{aligned}$$

Note that when using integrated Legendre polynomials for σ^1_η the boundary values $\sigma^1_\eta(0)$ and $\sigma^1_\eta(h_T)$ are zero for polynomial orders larger than one. Consequently facet functions on infinite facets couple just to volume basis functions of the strip $T_\Box$ containing the low order polynomials in the η-component. The corresponding coupling blocks are proportional to $T_+^{N\top} T_+^N$. Facet basis functions of boundary facets on Γ couple only to the $\xi-$component of volume basis functions containing the constant $\sigma^2_{\xi 0}$ via diagonal coupling blocks. On infinite triangles it follows from a similar argumentation that the element matrix contribution of the integral is also sparse.

Thus, choosing the basis functions as described above leads to sparse element matrices, especially for high polynomial orders.

Chapter 8

A tensor product implementation of the mixed hybrid finite element method

This chapter is devoted to an optimized version of the mixed hybrid formulation in two dimensions which we already published in [HHS10]. The method makes use of high order finite elements and is therefore suitable for the Helmholtz equation with high wave numbers. Starting point is the mixed hybrid formulation 5.2. As already mentioned, the volume degrees of freedom u and $\boldsymbol{\sigma}$ can be eliminated element by element, and the system of equations is reduced to a much smaller system containing just the introduced Lagrangian multipliers u_F and σ_F. The reduction of the unknown functions to the element interfaces is especially of interest for large frequencies ω, where it is well known that the number of unknowns to resolve the solution grows due to the pollution error [Ihl98, IB97] faster than $O(\omega^2)$ in two dimensions.

The elimination of the volume degrees of freedom is simplified by using an eigenfunction basis for u and σ on each element. For such a basis the coupling blocks for the inner degrees of freedom are sparse, more precisely these blocks are three by three block diagonal. Therefore, they can be inverted cheaply, even for high polynomial orders up to thousand. For rectangular elements the two dimensional eigenvalue problem decouples into two one dimensional problems, which considerably simplifies determining the basis functions. If the materials are uniform, these eigenvalue problems have be solved only once for each edge length and polynomial order.

From the literature it is well known that h refinement [Cia78] and p refinement [Sch98b] can achieve only algebraic convergence, while hp refinement [AP02, Dem06, Sch98b] leads to exponential convergence. In order to benefit from hp-refinement or to mesh arbitrary shaped objects, a rectangular mesh with hanging nodes has to be introduced.

That this concept is not only applicable to the Helmholtz equation under standard boundary conditions of Dirichlet, Neumann or Robin type, we show by using the tensor product ansatz also for Hardy space infinite elements from Chapter 7. In this case an eigenvalue problem for the infinite direction with the dimension of the Hardy space has to be solved additionally. The basis functions on the Hardy element are now constructed by combining the eigenfunctions of the infinite facet with the eigenfunctions of the boundary facet.

After stating in Section 8.1 the problem formulation and the discrete function spaces, Section 8.2 presents the eigenfunction basis and the underlying one dimensional eigenvalue problem. The consequences of this eigenfunction basis for the linear system of equations are analyzed in Section 8.3. In the Sections 8.4 and 8.5, respectively, the influence of Hanging nodes onto the global matrix and how Hardy space infinite elements can be combined with the eigenfunction approach is discussed. The chapter is concluded by a numerical examples section.

8.1 Preliminaries

For the beginning we consider the mixed Helmholtz equation under absorbing boundary conditions which results for a triangulation $\mathcal{T}$ in solving the mixed hybrid formulation 5.2. In the following, we assume that μ is constant on the domain and that ϵ is piecewise constant, i.e. ϵ is at least constant on each element.

8.1.1 The underlying mixed hybrid formulation

By exchanging the Dirichlet and Neumann traces u_F and σ_F with incoming and outgoing impedance traces, an equivalent formulation which fits into the context of the ultra weak variational formulation can be obtained. Thus, we define inspired by g_I^β and g_O^β in Lemma 5.5 the incoming and outgoing wave contributions G_I, G_O for one element via

$$\begin{aligned} \sigma_F &= (\boldsymbol{n}_T \cdot \boldsymbol{n}_F)\sqrt{\frac{\tilde{\epsilon}}{\mu}}\big(G_O - G_I\big) && (8.1) \\ u_F &= G_O + G_I. && (8.2) \end{aligned}$$

Because the coefficient ϵ can jump across element interfaces, it is not uniquely defined there. We therefore introduce $\tilde{\epsilon}$ as the mean of the values of ϵ on the two neighboring edges. If there is a jump in ϵ, G_O and G_I do not approximate the physical incoming and outgoing waves. For any element T the outgoing wave G_O is the incoming wave G_I of the neighboring element and vice versa. Due to reflections at element interfaces this is not the case for physical waves in the presence of a jumping coefficient ϵ. Note that because of the difference in sign of $\boldsymbol{n}_T \cdot \boldsymbol{n}_F$ and the exchange of G_I and G_O for two neighboring elements, the Lagrangian multiplier σ_F has the same sign independent of the element. If the element boundary is in Γ, $\boldsymbol{n}_T \cdot \boldsymbol{n}_F$ is assumed to be positive, and G_I and G_O equal the incoming and outgoing waves of the domain, respectively. Replacing u_F and σ_F by G_O and G_I in Formulation 5.2 and neglecting the damping term, i.e. $\alpha = 0$, leads to

Formulation 8.1. *Find* $\big(u, \boldsymbol{\sigma}, G_I, G_O\big) =: \tilde{u} \in U \times \widetilde{V} \times U_F \times U_F$*, such that for all* $\tilde{v} := \big(v, \boldsymbol{\tau}, g_I, g_O\big) \in U \times \widetilde{V} \times U_F \times U_F$

$$B_I(\tilde{u},\tilde{v}) + B_{IF}(\tilde{u},\tilde{v}) + B_{IF}(\tilde{v},\tilde{u}) + B_F(\tilde{u},\tilde{v}) + B_\Gamma(\tilde{u},\tilde{v}) = F(\tilde{v}),$$

with the bilinear form for the volume degrees of freedom

$$\begin{aligned} B_I(\tilde{u},\tilde{v}) \;:=\; \sum_{T\in\mathcal{T}} \Big[& i\omega\epsilon\big(u,v\big)_T - \big(\operatorname{div}\boldsymbol{\sigma}, v\big)_T - \big(u, \operatorname{div}\boldsymbol{\tau}\big)_T \\ & -i\omega\mu\big(\boldsymbol{\sigma},\boldsymbol{\tau}\big)_T + \beta\big\langle \boldsymbol{\sigma}\cdot\boldsymbol{n}_T, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T}\Big], \end{aligned}$$

the bilinear forms for the facet degrees of freedom

$$\begin{aligned} B_F(\tilde{u},\tilde{v}) \;&:=\; \sum_{T\in\mathcal{T}} \beta\big\langle \tilde{\epsilon}/\mu\,(G_O - G_I), g_O - g_I\big\rangle_{\partial T}, \\ B_\Gamma(\tilde{u},\tilde{v}) \;&:=\; -\big\langle \sqrt{\tilde{\epsilon}/\mu}(G_O + G_I), g_O + g_I\big\rangle_\Gamma, \end{aligned}$$

the bilinear form coupling volume functions and facet functions

$$B_{IF}(\tilde{u},\tilde{v}) := \sum_{T\in\mathcal{T}} \Big[\Big\langle \Big(1 - \beta\sqrt{\tilde{\epsilon}/\mu}\Big)\, G_O, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\Big\rangle_{\partial T} + \Big\langle \Big(1 + \beta\sqrt{\tilde{\epsilon}/\mu}\Big)\, G_I, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\Big\rangle_{\partial T}\Big],$$

and the linear functional

$$F(\tilde{v}) := -\big\langle \sqrt{\tilde{\epsilon}/\mu}\; g, g_O + g_I\big\rangle_\Gamma.$$

We remark again that the formulation is complex symmetric and that for a choice of $\beta = \sqrt{\tilde{\epsilon}/\mu}$ there is no coupling between the flux field $\boldsymbol{\sigma}$ and G_O. Thus, the solution in the element depends only on the incoming waves G_I. Furthermore, if there is no jump in ϵ across an element interface, the physical incoming wave and G_I coincide. Consequently, physically incoming waves determine the volume solution on the element.

8.1.2 The discrete finite element spaces

First, we recall the definitions of the finite element spaces involved in Formulation 8.1,

$$\begin{aligned} U &= L^2(\Omega), \\ \widetilde{V} &= \left\{ \boldsymbol{\tau} \in \left(L^2(\Omega)\right)^2 \;:\; \boldsymbol{\tau}\big|_T \in H(\mathrm{div}, T) \quad \forall T \in \mathcal{T} \right\}, \\ U_F &= L^2(\mathcal{F}). \end{aligned}$$

Until now no assumptions onto the mesh were made. As we will see in the following, a rectangular mesh with equally sized elements T simplifies the calculation. We will denote such an element T of width h_x and height h_y by

$$T = I_x^T \times I_y^T = \left[q_x^T, q_x^T + h_x^T\right] \times \left[q_y^T, q_y^T + h_y^T\right], \tag{8.3}$$

where (q_x^T, q_y^T) is the lower left vertex of the rectangle. For rectangular meshes, a tensor product basis for the discrete L^2-conforming space U_{hp} can be used, i.e.

$$U_{hp} = \prod_{T \in \mathcal{T}} P^{p_x}(I_x^T) \otimes P^{p_y}(I_y^T).$$

Here, $P^p(I)$ denotes the set of all polynomials on the interval I with an order less or equal to p. In U_{hp} the polynomial order p_x is chosen for x direction and polynomial order p_y for y direction.

The discrete counterpart V_{hp} of the broken Raviart Thomas space V reads as

$$V_{hp} = \prod_{T \in \mathcal{T}} \left(P^{p_x+1}(I_x^T) \otimes P^{p_y}(I_y^T)\right) \times \left(P^{p_x}(I_x^T) \otimes P^{p_y+1}(I_y^T)\right).$$

The polynomial orders are chosen such that the element wise divergence of a function from V_{hp} is in the space U_{hp}. Considering one element T, the trace of any function u in U_{hp} and the normal trace $\boldsymbol{\tau} \cdot \boldsymbol{n}_T$ for any $\boldsymbol{\tau} \in V_{hp}$ are polynomials of order p_x for horizontal facets

(facets parallel to the x axis) and polynomials of order p_y for vertical facets (facets parallel to the y axis), respectively. This gives rise to the definition of the discrete facet space

$$U_{F,hp} = \prod_{F \in \mathcal{F}} P^{p_F}(F),$$

where p_F equals p_x for horizontal facets and $p_F = p_y$ for vertical ones.

8.2 The eigenfunction basis

This section introduces a discrete eigenfunction basis of the space $U_{hp} \times V_{hp} \times U_{F,hp} \times U_{F,hp}$. By such an eigenfunction basis the contribution of the bilinear form $B_I(\tilde{u}, \tilde{v})$ to the system matrix is reduced to a three by three block diagonal matrix. Consequently, the interior degrees of freedom can be eliminated cheaply on the element level, even for high polynomial orders, and a considerably smaller system has to be solved for the incoming and outgoing wave degrees of freedom. Additionally, the element matrices in the assembly procedure can be calculated a priory, and costly numerical integration for high order polynomials is avoided.

Using an unstructured mesh, the number of degrees of freedom on an element of order p is of order p^2 which results in a computational cost of $O(p^6)$ for solving the eigenvalue problem. Such a construction would be much more costly than solving the original problem. On rectangles, the two dimensional eigenvalue problem decouples into two one dimensional eigenvalue problems. These one dimensional eigenvalue problems can be solved in $O(p^3)$ operations where p is p_x and p_y, respectively, which makes an eigenfunction basis competitive.

8.2.1 Legendre and integrated Legendre polynomials

In order to solve the eigenvalue problem introduced below, we expand the eigenfunctions into integrated Legendre polynomials. This leads because of nice orthogonality properties to a sparse matrix eigenvalue problem. Although, there are several different sets of orthogonal polynomials, integrated Legendre polynomials were chosen because of the L^2-orthogonality relation of their derivative and the possibility to evaluate them in a stable manner via a three term recursion. Consequently, we will use this subsection for a short introduction on Legendre and integrated Legendre polynomials. For a detailed discussion on orthogonal polynomials see [Sze39].

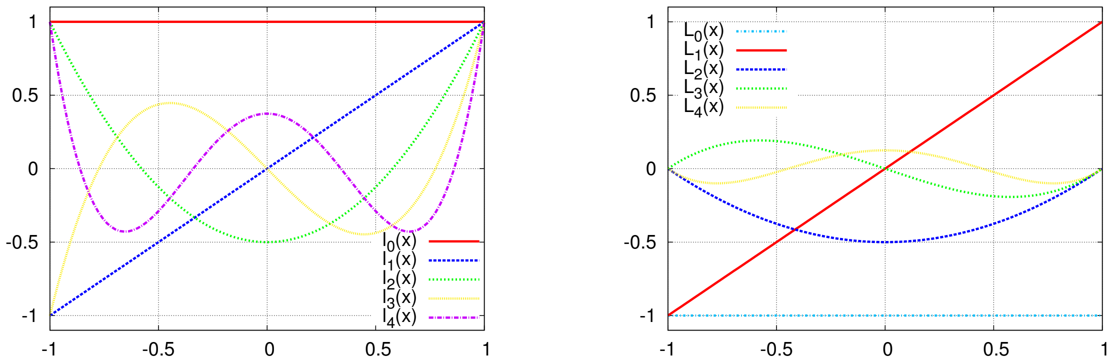

Figure 8.1: The Legendre polynomials $l_i(x)$ (left) and the integrated Legendre polynomials $L_j(x)$ (right)

Legendre polynomials

The Legendre polynomials which we denote as $l_i(x)$ for $0 \le i \le p$ span up the polynomial space $P^p([-1,1])$, and they are defined according to Rodriguez formulation as

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left[(x^2-1)^n \right]$$

or alternatively as

$$P_n(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(\frac{1+x}{2} \right)^{n-k} \left(\frac{1-x}{2} \right)^k .$$

In Figure 8.1 (left) the Legendre polynomials up to order 4 are plotted. The fact that the Legendre polynomials fulfill the three term recursion

$$\begin{aligned} (i+1)\, l_{i+1}(x) &= (2i+1)\, x\, l_i(x) - i\, l_{i-1}(x), \\ (x^2-1) \frac{d}{dx} l_i(x) &= ix\, l_i(x) - i\, l_{i-1}(x), \end{aligned}$$

together with $l_0(x) = 1$ and $l_1(x) = x$ allows a stable and fast evaluation of the polynomials and their derivative. The major advantage of the Legendre polynomials is their L^2-orthogonality on the interval $[-1,1]$, i.e.

$$\int_{-1}^{1} l_n(x) l_m(x)\, dx = \frac{2}{2n+1}\, \delta_{nm}.$$

Apart from this, the i-th Legendre polynomial has i zeros in the interval $[-1,1]$, and it is an odd function if i is odd, and for even i the polynomial $l_i(x)$ is even, too. The polynomials are normalized such that $l_i(1) = 1$ for all i and consequently $l_i(-1) = (-1)^i$.

Integrated Legendre polynomials

With the help of the Legendre polynomials the integrated Legendre polynomials $L_i(x)$ are defined as

$$L_i(x) = \int_{-1}^{x} l_{i-1}(s)\,ds \qquad \text{for } i \geq 2.$$

For completeness we add $L_0(x) = -1$ and $L_1(x) = x$. Figure 8.1 (right) shows the first five integrated Legendre polynomials. Taking into account that we can write any $L_i(x)$ as a linear combination of just two Legendre polynomials,

$$L_{i+1}(x) = \frac{1}{2i+1}\, l_{i+1}(x) - \frac{1}{2i+1}\, l_{i-1}(x) \qquad \text{for } i \geq 1$$

leads again to a stable evaluation of these polynomials, even for high polynomial orders. Although, the integrated Legendre polynomials do not form a set of orthogonal polynomials, the last expression shows that $L_i(x)$ is L^2-orthogonal on $[-1,1]$ to almost all other $L_j(x)$. More precisely,

$$\begin{aligned}
\int_{-1}^{1} L_n(x)L_n(x)\,dx &= \begin{cases} \frac{2}{2n+1} & \text{for } n = 0,1 \\ \frac{4}{(2n-3)(2n-1)(2n+1)} & \text{else} \end{cases} \\
\int_{-1}^{1} L_n(x)L_{n+2}(x)\,dx &= \frac{-2}{(2n-1)(2n+1)(2n+3)} \\
\int_{-1}^{1} L_n(x)L_m(x)\,dx &= 0 \qquad\qquad \text{for } |n-m| \neq 0,2.
\end{aligned}$$

From the definition of the L_j it is obvious that their derivatives are the l_{j-1} and therefore orthogonal. Again, integrated Legendre polynomials are even or odd if the index is even or odd, respectively. Furthermore, they are, apart from L_0 and L_1, zero on the interval boundaries.

8.2.2 The eigenvalue problem

Now, we can define the one dimensional eigenvalue problem, needed for the eigenfunction basis. This discrete eigenvalue problem reads as

Problem 8.2 (The discrete eigenvalue problem). *Find eigenfunctions* $0 \neq \phi_j(s) \in P^{p+1}(I)$ *with the interval* $I = [q, q+h]$ *and the corresponding eigenvalues* $\lambda_j \in \mathbb{C}$ *for* $1 \leq j \leq (p+2)$, *such that*

$$\int_I \phi_j' \varphi' \, ds = \lambda_j \mathcal{B}_h(\phi_j, \varphi) \qquad \forall \varphi \in P^{p+1}(I)$$

with

$$\mathcal{B}_h(\phi_j, \varphi) = \int_I i\omega\mu\phi_j\varphi \, ds - \beta\big(\phi_j(q)\varphi(q) + \phi_j(q+h)\varphi(q+h)\big)$$

and the orthogonality relation $\mathcal{B}_h(\phi_j, \phi_k) = \delta_{jk}$.

Note that one eigenfunction is the constant function belonging to the eigenvalue zero. For notational reasons this eigenfunction will get the index $p+2$ in the following. We should also mention that the eigenvalue problem just depends on the edge length, the polynomial order and the constants β and μ, and it is independent of the coefficient ϵ. Thus, it has to be solved just once for each edge length and polynomial order, independent of the element position.

This eigenvalue problem can be solved by expanding the discrete eigenfunctions into an integrated Legendre polynomial basis $L_I^i(s)$ transformed to the interval $I = [q, q+h]$. The superscript i indicates the polynomial order, i.e. $i \leq p+1$. With the help of the orthogonality relations from above the matrix entries can be computed without numerical integration, and a sparse and well conditioned matrix eigenvalue problem is obtained. In the following, the vector $\boldsymbol{\phi}_j$ is the coefficient vector of the eigenfunction ϕ_j with respect to an integrated Legendre polynomial basis. We obtain the following matrix eigenvalue problem by using orthogonality relations.

Problem 8.3 (The matrix eigenvalue problem). *Find* $\mathbf{0} \neq \boldsymbol{\phi}_j \in \mathbb{C}^{p+2}$ *and* $\lambda_j \in \mathbb{C}$, *such that*

$$D\,\boldsymbol{\phi}_j = \lambda_j M \boldsymbol{\phi}_j,$$

and the orthogonality relation $\boldsymbol{\phi}_j^\top M \boldsymbol{\phi}_i = \delta_{ij}$ *holds. The elements of the diagonal matrix* $D \in \mathbb{C}^{(p+2)\times(p+2)}$ *are*

$$D_{jj} = \frac{4}{h\,(2j-3)},$$

and the nonzero entries of the matrix $M \in \mathbb{C}^{(p+2)\times(p+2)}$ *are*

$$\begin{aligned} M_{jj} &= \begin{cases} \frac{i\omega\mu h}{(2j-3)^2(2j-1)} - 2\beta & \textit{for } j = 1,2 \\ \frac{2i\omega\mu h}{(2j-5)(2j-3)(2j-1)} & \textit{else} \end{cases} \\ M_{j\,j+2} &= M_{j+2\,j} = \frac{-i\omega\mu h}{(2j-3)(2j-1)(2j+1)} \qquad \textit{for } 1 \le j \le p. \end{aligned}$$

Because of the orthogonality relations of integrated Legendre polynomials and the fact that they can be divided into odd and even functions, the eigenvalue problem splits into an odd and an even problem of only half of the size of the original one, and therefore, the whole problem is faster to solve.

For the eigenfunctions of the discrete eigenvalue Problem 8.2 we obtain

$$\begin{aligned} \phi_j(s) &= \sum_{k=0}^{p+1} \left(\boldsymbol{\phi}_j\right)_{k+1} L_I^k(s), \\ \phi_j'(s) &= \sum_{k=0}^{p+1} \left(\boldsymbol{\phi}_j\right)_{k+1} L_I^{k\prime}(s) = \frac{2}{h}\sum_{k=0}^{p} \left(\boldsymbol{\phi}_j\right)_{k+2} l_I^k(s), \end{aligned} \tag{8.4}$$

where $l_I^k(s)$ represents the Legendre polynomial l_k of order k transformed to the interval I.

8.2.3 The basis functions

After having solved the eigenvalue problems, we are in the position to define the basis functions. The eigenfunction basis for the spaces U_{hp} and V_{hp} reads as

Definition 8.4 (The volume basis functions)**.** *Let* T *be a rectangle in* $\mathcal{T}$ *according to (8.3), then for* $1 \le j \le p_x + 2$ *and* $1 \le k \le p_y + 2$ *the basis functions of the space* U_{hp} *are defined as*

$$v_{jk}(x,y) = \begin{cases} \phi_j'(x)\varphi_k'(y) & \textit{for } (x,y) \in T \\ 0 & \textit{else} \end{cases}$$

and the basis functions for V_{hp} *as*

$$\begin{aligned} \boldsymbol{\tau}_{jk}^x(x,y) &= \begin{cases} \left(\phi_j(x)\varphi_k'(y), 0\right)^\top & \textit{for } (x,y) \in T \\ \mathbf{0} & \textit{else} \end{cases} \\ \boldsymbol{\tau}_{jk}^y(x,y) &= \begin{cases} \left(0, \phi_j'(x)\varphi_k(y)\right)^\top & \textit{for } (x,y) \in T \\ \mathbf{0} & \textit{else.} \end{cases} \end{aligned}$$

The functions ϕ_j *are eigenfunctions of Problem 8.2 with* $p = p_x$, $I = I_x^T$ *and* λ_j^x *as eigenvalues. The* φ_k *are eigenfunctions to the eigenvalues* λ_j^y *of the same problem with* $p = p_y$ *and* $I = I_y^T$.

In this definition all basis functions containing the derivative of the constant eigenfunction, namely v_{jk} for $j = p_x + 2$ or $k = p_x + 2$, $\boldsymbol{\tau}_{jk}^x$ for $k = p_y + 2$ and $\boldsymbol{\tau}_{jk}^y$ for $j = p_x + 2$ are zero. We keep them for notational convenience, but they can be omitted in an actual implementation.

The basis functions for the facet degrees of freedom can be defined similarly,

Definition 8.5 (The facet basis functions)**.** *Let* $F \in \mathcal{F}$ *be a facet of length* h *and polynomial order* p_F*. The basis functions for* $U_{F,hp}$ *are defined via*

$$g_{O\,j} = g_{I\,j} = \begin{cases} \psi_j' & \text{on } F, \\ 0 & \text{else}, \end{cases}$$

for $1 \leq j \leq p_F + 1$*. Here the* ψ_j *are eigenfunctions of Problem 8.2 to the non zero eigenvalues for* $p = p_F$ *and* $I = F$.

Again the constant eigenfunction ψ_{p_F+2} is not needed because of $\psi_{p_F+2}' = 0$. On a rectangular element T, p_F and F are equal to p_x and I_x^T if no hanging nodes are present. Thus, for horizontal facets the ψ_j coincide with the ϕ_j from the definition of the volume basis functions. For vertical edges ψ_j equals φ_j.

8.3 The linear system of equations

In this section we will examine the system matrix obtained by using the eigenfunction basis from above. Discretizing Formulation 8.1 leads to a system of equations of the form (compare section 6.2)

$$\begin{pmatrix} A & D \\ D^\top & C \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_{inner} \\ \boldsymbol{u}_{facet} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{f} \end{pmatrix}, \tag{8.5}$$

where $\boldsymbol{u}_{inner}$ is the coefficient vector for the inner degrees of freedom, and $\boldsymbol{u}_{facet}$ contains the facet unknowns. Therefore, the submatrix A couples the volume degrees of freedom with themselves, the block C connects the facet degrees of freedom with each other and the coupling entries between this two types of unknowns are stored in D. By A_T and D_T the element matrices collected in A and D are denoted.

8.3.1 The coupling of the volume basis functions

First, we start to discuss the structure of the matrix block A, which represents the bilinear form $B_I(\tilde{u}, \tilde{v})$, for the eigenfunction basis from above. Because each interior basis function is supported only on one element, A is block diagonal with blocks equal to the element matrices A_T, and it is therefore sufficient just to investigate A_T which we write as

$$A_T = \begin{pmatrix} A^T_{uu} & A^T_{u\boldsymbol{\sigma}_x} & A^T_{u\boldsymbol{\sigma}_y} \\ A^T_{\boldsymbol{\sigma}_x u} & A^T_{\boldsymbol{\sigma}_x\boldsymbol{\sigma}_x} & A^T_{\boldsymbol{\sigma}_x\boldsymbol{\sigma}_y} \\ A^T_{\boldsymbol{\sigma}_y u} & A^T_{\boldsymbol{\sigma}_y\boldsymbol{\sigma}_x} & A^T_{\boldsymbol{\sigma}_y\boldsymbol{\sigma} x} \end{pmatrix}. \tag{8.6}$$

Here, A^T_{uu} contains the coupling elements between the basis functions v_{jk}, $A^T_{u\boldsymbol{\sigma}_x}$ collects the entries between the basis functions v_{jk} and the functions $\boldsymbol{\tau}^x_{lm}$ and so on. Note that all the blocks are of the dimension $(p_x+2)(p_y+2) \times (p_x+2)(p_y+2)$. In the following we will use for notational convenience

$$\left(A^T_{uu}\right)_{jk,mn} = \left(A^T_{uu}\right)_{j(p_y+2)+k,m(p_y+2)+n} \tag{8.7}$$

for all block matrices.

Lemma 8.6. *In the eigenfunction basis of Definition 8.4 the element matrix A_T of an element T is block diagonal with three by three blocks.*

Proof. This result follows directly by inserting the the definitions of the basis functions into B_I and using the orthogonality relations. We first investigate the matrix A^T_{uu},

$$\begin{aligned}
\left(A^T_{uu}\right)_{jk,mn} &= B_I\big((v_{mn}, \mathbf{0}, 0, 0), (v_{jk}, \mathbf{0}, 0, 0)\big) &&= i\omega\epsilon\big(v_{jk}, v_{mn}\big)_T \\
&= i\omega\epsilon \int_{I^T_x} \phi'_j\phi'_m \, dx \int_{I^T_y} \varphi'_k\varphi'_n \, dy &&= i\omega\epsilon\lambda^x_j\lambda^y_k \, \mathcal{B}_{h^T_x}(\phi_j, \phi_m) \, \mathcal{B}_{h^T_y}(\varphi_k, \varphi_n) \\
&= i\omega\epsilon\lambda^x_j\lambda^y_k\delta_{jm}\delta_{kn},
\end{aligned}$$

and thus, A^T_{uu} is diagonal. The block $A^T_{u\boldsymbol{\sigma}_x}$ is also diagonal, according to

$$\begin{aligned}
\left(A^T_{u\boldsymbol{\sigma}_x}\right)_{jk,mn} &= B_I\big((0\boldsymbol{\tau}^x_{mn}, 0, 0), (v_{jk}, \mathbf{0}, 0, 0)\big) = -\big(v_{jk}, \operatorname{div} \boldsymbol{\tau}^x_{mn}\big)_T \\
&= \int_{I^h_x} \phi'_j\phi'_m \, dx \int_{I^T_y} \varphi'_k\varphi'_n \, dy \\
&= -\lambda^x_j\lambda^y_k\delta_{jm}\delta_{kn}.
\end{aligned}$$

By similar computations and symmetry arguments we obtain diagonality of $A^T_{u\boldsymbol{\sigma}_y}$, $A^T_{\boldsymbol{\sigma}_x u}$ and $A^T_{\boldsymbol{\sigma}_y u}$. The coupling block between the $\boldsymbol{\tau}^x_{jk}$ reads as

$$\begin{aligned}
\big(A^T_{\boldsymbol{\sigma}_x\boldsymbol{\sigma}_x}\big)_{jk,mn} &= B_I\big((0,\boldsymbol{\tau}^x_{mn},0,0)(0,\boldsymbol{\tau}^x_{jk},0,0)\big) \\
&= -i\omega\mu\big(\boldsymbol{\tau}^x_{jk},\boldsymbol{\tau}^x_{mn}\big)_T + \beta\big\langle\boldsymbol{\tau}^x_{jk}\cdot\boldsymbol{n}_T,\boldsymbol{\tau}^x_{mn}\cdot\boldsymbol{n}_T\big\rangle_{\partial T} \\
&= -i\omega\mu\int_{I^T_x}\phi_j\phi_m\,dx\int_{I^T_y}\varphi'_k\varphi'_n\,dy \quad +\beta\,\phi_j(q^T_x)\,\phi_m(q^T_x)\int_{I^T_y}\varphi'_k\varphi'_n\,dy \\
&\qquad +\beta\,\phi_j(q^T_x+h^T_x)\,\phi_m(q^T_x+h^T_x)\int_{I^T_y}\varphi'_k\varphi'_n\,dy \\
&= -\mathcal{B}_{h^T_x}(\phi_j,\phi_m)\int_{I^T_y}\varphi'_k\varphi'_n\,dy = -\lambda^y_k\,\delta_{kn}\delta_{jm}.
\end{aligned}$$

An equivalent computation results into diagonality of $A^T_{\boldsymbol{\sigma}_y\boldsymbol{\sigma}_y}$. Finally,

$$\begin{aligned}
\big(A^T_{\boldsymbol{\sigma}_x\boldsymbol{\sigma}_y}\big)_{jk,mn} &= B_I\big((0,\boldsymbol{\tau}^y_{mn},0,0),(0,\boldsymbol{\tau}^x_{jk},0,0)\big) \\
&= -i\omega\mu\big(\boldsymbol{\tau}^x_{jk},\boldsymbol{\tau}^y_{mn}\big)_T + \beta\big\langle\boldsymbol{\tau}^x_{jk}\cdot\boldsymbol{n}_T,\boldsymbol{\tau}^y_{mn}\cdot\boldsymbol{n}_T\big\rangle_{\partial T} \\
&= 0,
\end{aligned}$$

and similar arguments hold for $A^T_{\boldsymbol{\sigma}_y\boldsymbol{\sigma}_x}$. Because all the nine coupling blocks are either zero or diagonal, the degrees of freedom can be reordered such that A_T is three by three block diagonal. □

Static condensation of the volume degrees of freedom corresponds to an inversion of the matrix A which is according to this result equivalent to the inversion of 3×3 matrices, and therefore, it can be done cheaply.

Note that the basis functions v_{jk}, $\boldsymbol{\tau}^x_{jk}$ and $\boldsymbol{\tau}^y_{jk}$ are zero if $j = p_x+2$ or $k = p_y+2$. Thus, their coefficients are set to zero by definition. If $j = p_x+2$ or $k = p_y+2$ the corresponding 3×3 block is singular and it has just one diagonal element belonging to the basis function $\boldsymbol{\tau}^x_{jk}$ and $\boldsymbol{\tau}^y_{jk}$, respectively, which has to be inverted.

8.3.2 The coupling of the facet basis functions

In this subsection, we give the matrix representation C of the bilinear form $B_F + B_\Gamma$, which describes the coupling between the facet degrees of freedom. Because each facet basis function is supported just on one single facet, there is no coupling between unknowns belonging to two different facets, and C is block diagonal with blocks corresponding to the

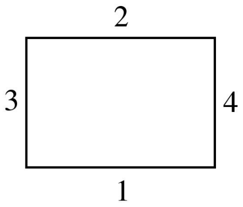

Figure 8.2: The local numbering

facets. The block coupling the unknowns of the facet F with themselves is denoted by C_F.

When dealing just with facets, the expressions in and outgoing waves, G_I and G_O, which are related to a certain element, are not suitable anymore. We will rather change to left and right going waves for vertical facets and down and up going waves for horizontal facets. Up or right going waves are denoted as G_+ and down or left going waves by G_-. Using a local edge numbering as indicated in Figure 8.2, G_+ corresponds to G_I for an element T where the facet has local edge numbering 1 or 3, and for local edge numbering 2 or 4 it corresponds to G_O. For G_- the situation is reversed.

By reordering the degrees of freedom, the matrix C can be brought according to the following Lemma into a two by two block diagonal form.

Lemma 8.7. *Using the eigenfunction basis of Definition 8.5 the matrix block C^F for a facet F is*

$$C_F = \begin{pmatrix} C^F_{G_+G_+} & C^F_{G_+G_-} \\ C^F_{G_-G_+} & C^F_{G_-G_-} \end{pmatrix} = \begin{pmatrix} \alpha_1 \operatorname{diag}(\lambda_k^F) & \alpha_2 \operatorname{diag}(\lambda_k^F) \\ \alpha_2 \operatorname{diag}(\lambda_k^F) & \alpha_3 \operatorname{diag}(\lambda_k^F) \end{pmatrix},$$

where the four blocks of C^F are from $\mathbb{C}^{(p_F+1)\times(p_F+1)}$, λ_k^F, $1 \le k \le p_F+1$ are the eigenvalues of Problem 8.2 solved on F and

$$\alpha_1 = \alpha_3 = 2\beta\frac{\tilde{\epsilon}}{\mu}, \qquad \alpha_2 = -2\beta\frac{\tilde{\epsilon}}{\mu}, \qquad \textit{for } F \subset \Gamma$$

$$\alpha_1 = \alpha_3 = \beta\frac{\tilde{\epsilon}}{\mu} - \sqrt{\frac{\tilde{\epsilon}}{\mu}}, \qquad \alpha_2 = -\beta\frac{\tilde{\epsilon}}{\mu} - \sqrt{\frac{\tilde{\epsilon}}{\mu}}, \qquad \textit{else.}$$

Proof. Because, g_O, g_I and g_+, g_-, respectively, are from the same set of basis functions, the integral $\langle g_{\bullet,j}, g_{*,k}\rangle_F$ with $\bullet, * \in \{I, O\}$ or $\{+,-\}$, respectively, can be written as

$$\langle g_{\bullet j}, g_{*k}\rangle_F = \int_F \psi_j' \psi_k' \, ds = \lambda_k^F \delta_{jk}.$$

Thus, the contribution of the term $\beta\tilde{\epsilon}/\mu\langle G_O - G_I, g_O - g_I\rangle_{\partial T} = \beta\tilde{\epsilon}/\mu\langle G_+ - G_-, g_+ - g_-\rangle_{\partial T}$ from an element T where F is a boundary facet to C_F is $\beta\tilde{\epsilon}/\mu\,\mathrm{diag}(\lambda_k^F)$ for $C^F_{G_+G_+}$ and $C^F_{G_-G_-}$, while it is $-\beta\tilde{\epsilon}/\mu\,\mathrm{diag}(\lambda_k^F)$ for $C^F_{G_+G_-}$ and $C^F_{G_-G_+}$. Because each inner facet has two neighboring elements which contribute to C^F, the proof is completed for inner facets. The boundary integral $-\sqrt{\tilde{\epsilon}/\mu}\langle G_O + G_I, g_O + g_I\rangle_\Gamma = -\sqrt{\tilde{\epsilon}/\mu}\langle G_+ + G_-, g_+ + g_-\rangle_\Gamma$ adds for boundary facets F the matrices $-\sqrt{\tilde{\epsilon}/\mu}\,\mathrm{diag}(\lambda_k^F)$ to each of the four blocks of C_F. $\square$

8.3.3 The coupling between facet and volume degrees of freedom

Finally, we examine the matrix representation D of the bilinear form B_{IF}, or more precisely the corresponding element matrices D_T, coupling the inner degrees of freedom of T to the facet degrees of freedom of its boundary facets.

In the following, we will use the same local edge numbering as above (compare Figure 8.2). The element matrix D_T is again of block matrix structure, and for these blocks we use a similar notation as introduced in the last section. Thus, the block $D^{T,m}_{\boldsymbol{\sigma}_x G_O}$ contains the coupling between the volume basis functions $\boldsymbol{\tau}^x_{jk}$ of the element T and the facet basis functions g_{Ol} of its boundary facet with local number m, and

$$\left(D^{Tm}_{\boldsymbol{\sigma}_x G_O}\right)_{jk,l} = \left(D^{Tm}_{\boldsymbol{\sigma}_x G_O}\right)_{j(p_y+2)+k,l} = B_{IF}\big((0,\mathbf{0},0,g_{Ol}),(0,\boldsymbol{\tau}^x_{jk},0,0)\big).$$

Note that there is no coupling between the scalar volume basis functions v_{jk} and the facet basis functions. On horizontal edges with $m = 1,2$ the product $\boldsymbol{\tau}^x_{jk}\cdot\boldsymbol{n}_T$ is zero and consequently $B^{Tm}_{\boldsymbol{\sigma}_x,\bullet} = 0$ with $\bullet = G_O, G_I$. For the coupling entries between the $\boldsymbol{\tau}^y_{jk}$ and the l-th facet basis function we get

$$\left(D^{Tm}_{\boldsymbol{\sigma}_y\bullet}\right)_{jk,l} = s_k\left(1 \pm \beta\sqrt{\frac{\tilde{\epsilon}}{\mu}}\right)\int_{I^m}\phi'_j\psi'_l\,dx \tag{8.8}$$

with the $+$ sign for $\bullet = G_I$ and the $-$ sign for $\bullet = G_O$. The factor s_k contains the evaluation of the function φ_k at the edge, more precisely $s_k = -\varphi_k(q^T_y)$ for $m = 1$ and $s_k = \varphi_k(q^T_y + h^T_y)$ for $m = 2$. We should mention that, because of the expansion of the eigenfunctions into integrated Legendre polynomials, where just the zero and first order polynomial have values different from zero at q^T_x and $q^T_x + h^T_x$, the eigenfunction φ_k is simple to evaluate at these points. Thus, s_k is just a linear combination of the first two components of the eigenvector $\boldsymbol{\varphi}_k$ of the matrix eigenvalue problem.

Besides this, it is important to mention that for uniform meshes, i.e. if no hanging nodes are present, ϕ_j and ψ_l are solutions of the same eigenvalue problem, and consequently

$$\int_{I^m} \phi_j' \psi_l' \, dx = \lambda_j^x \, \delta_{jl}.$$

The matrix block is sparse.

On vertical edges with $m = 3, 4$ the only non zero coupling elements are the elements between the basis functions $\boldsymbol{\tau}_{jk}^x$ and the facet basis functions. We obtain

$$\left(D_{\boldsymbol{\sigma}_x \bullet}^{Tm}\right)_{jk,l} = s_j \left(1 \pm \beta \sqrt{\frac{\tilde{\epsilon}}{\mu}}\right) \int_{I^m} \varphi_k' \psi_k' \, dy = s_j \left(1 \pm \beta \sqrt{\frac{\tilde{\epsilon}}{\mu}}\right) \lambda_k^y \delta_{kl}.$$

Again the + sign is taken for $\bullet = G_I$ and $-$ for $\bullet = G_O$, $s_j = -\phi_j(q_x^T)$ for $m = 3$ and $s_j = \phi_j(q_x^T + h_x^T)$ for $m = 4$, respectively. Following the same arguments, s_j can be evaluated from the first two components of the corresponding eigenvector $\boldsymbol{\phi}_j$ of the matrix eigenvalue problem.

8.3.4 Solving the system of equations

When solving the system of equations from (8.5), we first eliminate the interior degrees of freedom (compare Section 6.2) and solve the resulting system for facet degrees of freedom

$$S \boldsymbol{u}_{facet} = \boldsymbol{f} \qquad \text{with} \qquad S = C - D^\top A^{-1} D.$$

First, let us take a closer look onto the structure of the Schur complement matrix S, especially onto $D^\top A^{-1} D$. As already mentioned, the volume basis functions are supported on just one element. Thus there is no coupling between degrees of freedom belonging to different elements in A and consequently in A^{-1}. The matrix D just contains entries between inner degrees of freedom of an element and the facet degrees of freedom of its boundary facets. Therefore, the term $D^\top A^{-1} D$ and consequently S has just coupling blocks between facets which are boundary facets of the same element. Thus, an inner facet couples apart from itself to two parallel facets, and to four perpendicular facets, while boundary facets just couple to one parallel facet and to two perpendicular ones. From the structure of the matrices A and D it becomes obvious that coupling blocks between parallel edges are two by two block diagonal, while coupling blocks between perpendicular edges are full.

We solve the Schur complement system with a preconditioned conjugate gradient method (PCG), using the complex symmetric inner product $\boldsymbol{x}^\top \boldsymbol{y}$. Although this method works well for our numerical examples, there exists no rigorous convergence analysis. As a preconditoner we use an Additive Schwarz block preconditoner P,

$$P = \sum_i \left(P^{F_i}\right)^{-1}$$

where each block P^{F_i} contains the coupling entries of the Schur complement matrix S between unknowns belonging to the facet F_i. So, the P^{F_i} are the diagonal blocks of S, and they can be written as block matrix,

$$P^{F_i} = \begin{pmatrix} P^{F_i}_{G_+G_+} & P^{F_i}_{G_+G_-} \\ P^{F_i}_{G_-G_+} & P^{F_i}_{G_-G_-} \end{pmatrix},$$

where each of the four blocks is diagonal. We are going to show this for $P^{F_i}_{G_+G_+}$ if F_i is an horizontal facet. For the other blocks and facet types the calculation is the same. The block $P^{F_i}_{G_+G_+}$ contain apart from the matrix $C^{F_i}_{G_+G_+}$ contributions from $D^\top A^{-1} D$ via the elements T_a above the facet and T_b below the facet, where G_+ equals G_I and G_O, respectively. Because we know from Subsection 8.3.3 that D just couples wave degrees of freedom of horizontal facets to $\boldsymbol{\sigma}_y$ degrees of freedom of the neighboring element, we get

$$P^{F_i}_{G_+G_+} = C^{F_i}_{G_+G_+} + \left(D^{T_a 1}_{\boldsymbol{\sigma}_y G_I}\right)^\top \left(A^{T_a}\right)^{-1}_{\boldsymbol{\sigma}_y \boldsymbol{\sigma}_y} D^{T_a 1}_{\boldsymbol{\sigma}_y G_I} + \left(D^{T_b 2}_{\boldsymbol{\sigma}_y G_O}\right)^\top \left(A^{T_b}\right)^{-1}_{\boldsymbol{\sigma}_y \boldsymbol{\sigma}_y} D^{T_b 2}_{\boldsymbol{\sigma}_y G_O}.$$

For boundary facets just one of the last two summands has to be considered, which one depends on if the neighboring element is above or below the facet. Taking the structure of the involved matrices from the Subsections 8.3.1 - 8.3.3, i.e.

$$\begin{aligned} \left(C^F_{G_+G_+}\right)_{jk} &= c^F_j \, \delta_{jk}, \\ \left(D^{Tm}_{\sigma_y \bullet}\right)_{jk,l} &= d^T_{jk} \, \delta_{jl}, \\ \left((A_T)^{-1}_{\boldsymbol{\sigma}_y \boldsymbol{\sigma}_y}\right)_{jk,mn} &= a^T_{jk} \, \delta_{jm} \delta_{kn} \end{aligned}$$

with complex numbers $c_j^F, d_{jk}^T, a_{jk}^T$ and inserting them, we obtain

$$\begin{aligned}
\left(P^{F_i}_{G_+G_+}\right)_{jk} &= \left(C^{F_i}_{G_+G_+}\right)_{jk} + \sum_{pqmn} \left(D^{T_a 1}_{\boldsymbol{\sigma}_y G_I}\right)_{pq,j} \left((A_{T_a})^{-1}_{\boldsymbol{\sigma}_y\boldsymbol{\sigma}_y}\right)_{pq,mn} \left(D^{T_a 1}_{\boldsymbol{\sigma}_y G_I}\right)_{mn,k} \\
&\quad + \sum_{pqmn} \left(D^{T_b 2}_{\boldsymbol{\sigma}_y G_O}\right)_{pg,j} \left((A_{T_b})^{-1}_{\boldsymbol{\sigma}_y\boldsymbol{\sigma}_y}\right)_{pq,mn} \left(D^{T_b 2}_{\boldsymbol{\sigma}_y G_O}\right)_{mn,k} \\
&= c_j^{F_i}\delta_{jk} + \sum_{pqmn} (d_{pq}^{T_a}\delta_{pj})\,(a_{pq}^{T_a}\delta_{pm}\delta_{qn})\,(d_{mn}^{T_a}\delta_{mk}) \\
&\quad + \sum_{pqmn} (d_{pq}^{T_b}\delta_{pj})\,(a_{pq}^{T_b}\delta_{pm}\delta_{qn})\,(d_{mn}^{T_b}\delta_{mk}) \\
&= c_j^{F_i}\delta_{jk} + \sum_{qn} d_{jq}^{T_a} a_{jq}^{T_a} d_{kn}^{T_a}\ \delta_{jk}\delta_{qn} + \sum_{qn} d_{jq}^{T_b} a_{jq}^{T_b} d_{kn}^{T_b}\ \delta_{jk}\delta_{qn} \\
&= \left(c_j^{F_i} + \sum_n d_{jn}^{T_a} a_{jn}^{T_a} d_{jn}^{T_a} + \sum_n d_{jn}^{T_b} a_{jn}^{T_b} d_{jn}^{T_b}\right)\ \delta_{jk}.
\end{aligned}$$

Consequently, $P^{F_i}_{G_+G_+}$ is diagonal.
Hence, the blocks P^{F_i} of the preconditioner are two by two block diagonal and therefore cheap to invert.

8.4 Hanging nodes

The uniform mesh approach presented in the last section has several disadvantages. For example in order to get a good approximation of an arbitrary domain, a very fine mesh is needed if the domain is meshed by uniform rectangles. In regions where the material parameters a constant and big elements with high polynomial orders are suitable, small rectangles have to be used. Furthermore refinement to specific points of the mesh, where a singularity of the solution is expected is not possible.

8.4.1 The mesh

One way to overcome this difficulty is to introduce a mesh which allows for hanging nodes. For such a rectangular mesh the volume degrees of freedom can be eliminated with the same technique introduced in the last section. The main difference is that using an eigenfunction basis for a mesh with hanging nodes, which contains different sized rectangles, requires the solution of an eigenvalue problem for each edge length and polynomial order. Consequently, it is convenient to have just a small number of different sized elements. This can be achieved

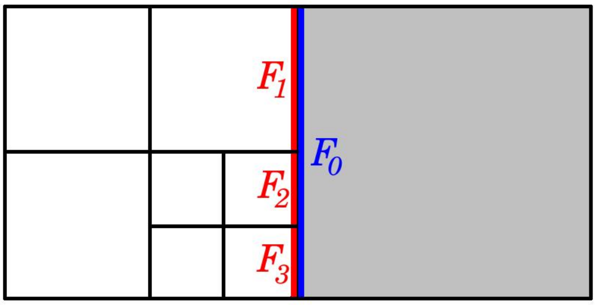

Figure 8.3: A mesh with hanging nodes

by the following refinement strategy. If an element needs to be refined, it is divided into four equally sized elements with polynomial orders p_x and p_y, which are just half of the polynomial orders of the original element. Following this, the total number of unknowns stays approximately constant. If the refinement process is started from an uniform mesh with edge length h_x and h_y, an element obtained after k refinement steps has the size $(1/2)^k h_x \times (1/2)^k h_y$. Thus, for a mesh obtained after n refinement steps at most $(n+1)$ eigenvalue problems have to be solved for each spacial direction.

In Figure 8.3 a simple mesh after a few refinement steps is illustrated. When an element is refined, each of its long boundary facets in $\mathcal{F}$ is exchanged by the two new ones. Thus the shaded element in Figure 8.3 has instead of the blue facet F_0 the three smaller boundary facets F_1, F_2 and F_3 printed in red. Like in the uniform mesh case, the facet basis functions of these facets are derived according to Definition 8.5 by solving the eigenvalue Problem 8.2.

8.4.2 The structure of the system matrix

If we use the same definitions for the basis functions as in the uniform mesh case, i.e. the Definitions 8.4 and 8.5, the structures of the matrix representations of B_I and B_F, namely A and C, do not change, but for the bilinear form B_{IF} the situation is different. Now, the coupling blocks which couple the volume degree of freedom of an element to the facet unknowns of its boundary edges are full. In the following, we will demonstrate this for a horizontal hanging node edge with local number f of the element $T := [q_x^T, q_x^T + h_x^T] \times [q_y^T, q_y^T + h_y^T]$. We assume that this hanging node facet is obtained after n refinement steps starting with the element edge. Thus, its length is $h_f = \left(\frac{1}{2}\right)^n h_x^T$.

In the uniform mesh case, as well as in the hanging node case, the entries of the coupling matrix D follow equation (8.8). But now the functions ϕ_j and ψ_l are obtained by solving the eigenvalue problem 8.2 for different mesh sizes and polynomial orders. Consequently, they are not orthogonal anymore.

In the following, we will focus on the integral $\int_{I^f} \phi_j' \psi_l' \, dx$. One possibility to evaluate this integral is numerical integration, which is costly because all the eigenfunctions have to be evaluated at the integration points. We are going to present another option.

Therefore, we assume that the facet f of length $h_f = h_n$ and $p_f = p_n$ is obtained after n refinement steps from $I_0 = I_x^T$ with $h_0 = h_x^T$ and $p_0 = p_x$. After refinement step k we obtain by refining the facet I_{k-1} a facet $I_k = [q_k, q_k + h_k]$ of mesh size $h_k = \frac{1}{2} h_{k-1}$, polynomial order $p_k = \frac{1}{2} p_{k-1}$, and the vertex q_k equals either q_{k-1} or $q_{k-1} + h_k$. Additionally we introduce transformation matrices $E \in \mathbb{C}^{(p_0+1)\times(p_0+1)}$ and $\widetilde{E} \in \mathbb{C}^{(p_n+1)\times(p_n+1)}$ transforming the eigenfunctions functions ϕ_j and ψ_l into Legendre polynomials. Thus, E and $\widetilde{E}$ contain the eigenvectors $\boldsymbol{\phi}_j$ and $\boldsymbol{\psi}_l$ of the corresponding matrix eigenvalue problem, i.e. $E_{jk} = (\boldsymbol{\phi}_j)_{k+1}$ and $\widetilde{E}_{lk} = (\boldsymbol{\psi}_l)_{k+1}$. Note that the first component in the eigenvectors is not needed, because it belongs to the constant integrated Legendre polynomial which has a vanishing derivative. Together with equation (8.4) the integral under consideration can be transformed to integrals of Legendre polynomials,

$$\int_{I_f} \phi_i' \psi_j' \, dx = \frac{4}{h_n h_0} \sum_{m=1}^{p_0+1} \sum_{q=1}^{p_n+1} E_{im} \widetilde{E}_{jq} \int_{I_n} l_{I_0}^{m-1} l_{I_n}^{q-1} \, dx.$$

By introducing the transformation matrices $T_k^+, T_k^- \in \mathbb{R}^{(p_{k-1}+1)\times(p_k+1)}$ with

$$\begin{aligned} (T_k^+)_{ij} &= (2j-1) \int_0^1 l_{[-1,1]}^{i-1} l_{[0,1]}^{j-1} \, dx, \\ (T_k^-)_{ij} &= (2j-1) \int_0^1 l_{[-1,1]}^{i-1} l_{[-1,0]}^{j-1} \, dx, \end{aligned}$$

the Legendre polynomial $l_{I_{k-1}}^i$ can be replaced by the Legendre polynomials $l_{I_k}^j$ on the interval I_k obtained by refining I_{k-1},

$$l_{I_{k-1}}^{i-1} = \sum_{j=1}^{p_k+1} \frac{2j-1}{h_k} \left(\int_{I_k} l_{I_{k-1}}^{i-1} l_{I_k}^{j-1} \, dx \right) l_{I_k}^{j-1} = \sum_{j=1}^{p_k+1} (T_k^{\pm})_{ij} \, l_{I_k}^{j-1}.$$

Here, we take T_k^- if I_k is the left hand half of I_{k-1}, i.e. $q_k = q_{k-1}$ and T_k^+ else. Doing this for each refinement step, the polynomials $l_{I_0}^{m-1}$ can be expressed in terms of $l_{I_n}^{s-1}$, and we end up with

$$\int_{I_f} \phi_i' \psi_j' \, dx = \frac{4}{h_0 h_n} \sum_{s=1}^{p_n+1} \sum_{q=1}^{p_n+1} \Big(E \prod_{k=1}^{n} T_k^{\pm} \Big)_{is} \widetilde{E}_{jq} \int_{I_n} l_{I_n}^{s-1} l_{I_n}^{q-1} \, dx.$$

Finally, the orthogonality of the Legendre polynomials results in

$$\int_{I_f} \phi_i' \psi_j' \, dx = \left(E \Big(\prod_{k=1}^{n} T_k^{\pm} \Big) M \widetilde{E} \right)_{ij},$$

where the diagonal matrix $M \in \mathbb{R}^{(p_n+1)\times(p_n+1)}$ has the entries $M_{ii} = \frac{4}{(2i-1)h_0}$.

For vertical edges the computation is completely similar. Because these coupling blocks are full matrices, the coupling blocks of the Schur complement matrix between different boundary facets of one element are also full.

8.4.3 The preconditioner

In order to solve the resulting Schur complement system, we use again a PCG solver with an Additive Schwarz block preconditioner P containing blocks P^{F_i} which are constructed in a similar fashion as in the uniform mesh case.

If the facet F_i, to which the block P^{F_i} belongs to, is a facet where at least one of the endpoints is a hanging node, the facet to facet coupling block of the Schur complement matrix is full, and therefore, especially for high polynomial orders, expensive to invert. In this case, the block P^{F_i} is constructed by treating F_i as a facet where none of the endpoints is a hanging node, and where the two perhaps fictitious surrounding elements T_1 and T_2 share the polynomial order with the facets. Thus, the eigenfunctions ϕ_j for the horizontal edge (or φ_k if the edge is vertical) needed to construct the volume basis functions of T_1 and T_2 equal the set of eigenfunctions ψ_l of the basis functions of F_i. Consequently, the blocks in D_{T_1} and D_{T_2} which couple to the volume degrees of freedom of the hypothetical elements T_1 and T_2 are again sparse. A calculation of P_{F_i} with the formulas from Section 8.3.4 results in a two by two block diagonal matrix.

If none of the endpoints of F_i is a hanging node, the corresponding diagonal block in the Schur complement matrix is already two by two block diagonal, and it is taken for P^{F_i}.

8.5 Hardy space infinite elements

In the previous sections an absorbing boundary condition was used in order to reduce the infinite domain to a finite computational domain. In the following, we discuss how HSIE, which were introduced in Chapter 7, can be adapted such that they fit into the setting described above. For this purpose we stick to the notation from Section 7.2 with the difference that the interior domain is meshed by rectangles and not, as shown in Figure 7.1, by triangles.

8.5.1 The mixed hybrid formulation

Starting from Formulation (7.9), we obtain by a change from the Dirichlet and Neumann traces u_F and σ_F on the facet to incoming and outgoing waves G_I and G_O, as described in (8.1) and (8.2), an equivalent formulation, i.e.

Formulation 8.8. *Find* $\big(u, \boldsymbol{\sigma}, G_I, G_O\big) =: \tilde{u} \in U_H \times V_H \times U_{HF} \times U_{HF}$ *with* U_H, V_H, U_{HF} *defined in (7.13),(7.19), (7.20), such that for all* $\tilde{v} := \big(v, \boldsymbol{\tau}, g_I, g_O\big) \in U_H \times V_H \times U_{HF} \times U_{HF}$

$$\widetilde{B}_I(\tilde{u},\tilde{v}) + \widetilde{B}_{IF}(\tilde{u},\tilde{v}) + \widetilde{B}_{IF}(\tilde{v},\tilde{u}) + \widetilde{B}_F(\tilde{u},\tilde{v}) = F_{in}(\tilde{v}).$$

The bilinear forms $\widetilde{B}_I$, $\widetilde{B}_{IF}$ *and* $\widetilde{B}_F$ *are obtained from* B_I, B_{IF} *and* B_F *in Formulation 8.1 by exchanging the triangulation* $\mathcal{T}$ *by* $\mathcal{T} \cup \mathcal{H}$, *and*

$$\begin{aligned} F_{in}(\tilde{v}) \;:=\; \sum_{T\in\mathcal{H}} \Big[& \big\langle u_{in} - \beta\boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_T, \boldsymbol{\tau}\cdot\boldsymbol{n}_T\big\rangle_{\partial T\cap\Gamma} \\ & -\big\langle \Big(1-\beta\sqrt{\tilde{\epsilon}/\mu}\Big)\,\boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_T, g_O\big\rangle_{\partial T\cap\Gamma} \\ & -\big\langle \Big(1+\beta\sqrt{\tilde{\epsilon}/\mu}\Big)\,\boldsymbol{\sigma}_{in}\cdot\boldsymbol{n}_T, g_I\big\rangle_{\partial T\cap\Gamma}\Big]. \end{aligned}$$

Thus, Hardy space boundary conditions are implemented by neglecting the boundary term B_Γ in Formulation 8.1 and adding therefore the infinite elements surrounding the domain to the sums in the bilinear forms B_I, B_{IF} and B_F. Consequently, the discussion concerning the basis functions and the element matrices from the previous sections is still valid for the elements in the interior domain, and we can restrict ourselves onto the exterior domain and the Hardy elements, respectively.

Note that in this formulation the damping term is neglected, thus $\alpha = 0$. Furthermore, ϵ is assumed to be constant in the exterior domain Ω_E, which leads to $\tilde{\epsilon} = \epsilon$.

8.5.2 The eigenfunction basis

As in Chapter 7 the infinite element is transformed by a displacement combined with a rotation F_T to the reference strip and reference triangle $\hat{T}$, respectively. Transforming the volume functions $\boldsymbol{\sigma}$ and u to the reference element we obtain $\boldsymbol{\sigma}_{\hat{T}} = (\det J)J^{-1}\boldsymbol{\sigma} \circ F_T$ and $u \circ F_T$. Again, we take the ansatz

$$\hat{\boldsymbol{\sigma}}_{\hat{T}_\square} = \begin{pmatrix} \sigma_\eta^1 \otimes \hat{\sigma}_\xi^1 \\ \sigma_\eta^2 \otimes \hat{\sigma}_\xi^2 \end{pmatrix} \qquad \text{or} \qquad \hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}_\triangle} = \begin{pmatrix} \hat{\sigma}_\eta^1 \otimes \hat{\sigma}_\xi^1 \\ \hat{\sigma}_\eta^2 \otimes \hat{\sigma}_\xi^2 \end{pmatrix},$$

$$\hat{u} = u_\eta \otimes \hat{u}_\xi \qquad \text{or} \qquad \hat{\hat{u}} = \hat{u}_\eta \otimes \hat{u}_\xi .$$

for basis functions describing the Möbius and Laplace transform of $\boldsymbol{\sigma}_{\hat{T}}$ and u with respect to the infinite coordinate axis, i.e. $\hat{\boldsymbol{\sigma}}_{\hat{T}}$ and $\hat{u}$ (or $\hat{\hat{\boldsymbol{\sigma}}}_{\hat{T}}$ and $\hat{\hat{u}}$ for infinite triangles). For the functions σ_η^1, σ_η^2 and u_η the polynomial eigenfunction basis derived from the eigenvalue problem 8.2 is used. If we want to generalize the approach used for rectangles in Ω_I to infinite elements, a discrete eigenvalue problem needs to be solved also for the infinite variable ξ. Additionally, as we will see below, an eigenfunction basis allows for a cheap elimination of the volume degrees of freedom. For such a basis this elimination corresponds to the inversion of three by three matrices, while a monomial basis for the infinite variable requires the inversion of matrices with the dimension $3(N+2)$ where N is the maximal order of the monomials.

The eigenvalue problem

The discrete eigenvalue problem for the Möbius and Laplace transform of functions related to the infinite axis reads as

Problem 8.9. *Find the eigenfunctions* $0 \neq \Psi_j(z) \in W := \{v \in H^+(D) : v(z) \in P^{N+1}\}$ *and the corresponding eigenvalues* $\gamma_j \in \mathbb{C}$ *for* $1 \leq j \leq (N+2)$*, such that*

$$2i\kappa_0\, A\big(\hat{\partial}_z \Psi_j(z), \hat{\partial}_z \Psi(z)\big) = \gamma_j \mathcal{C}_N(\Psi_j(z), \Psi(z)) \qquad \forall \Psi \in W$$

with $A(\,,\,)$ *from (7.7),*

$$\mathcal{C}_N(\Psi_j(z), \Psi(z)) = -2\omega\mu\kappa_0\, A(\Psi_j(z), \Psi(z)) - 4\beta\kappa_0^2 \Psi_j(1)\, \Psi(1)$$

and the orthogonality relation $\mathcal{C}_N(\Psi_j, \Psi_k) = \delta_{jk}$.
Note that the derivative $\hat{\partial}_z$ has to be understood in the sense of (7.6), i.e. $\hat{\partial}_z = i\kappa_0 \mathcal{T}_+ \mathcal{T}_-^{-1}$. We remark additionally that the evaluation of the Möbius and Laplace transform at one, $\Psi_j(1)$, corresponds according to the relation (7.5) to an evaluation of the untransformed function at zero.

Using for a function $\Psi \in W$ the ansatz

$$\Psi(z) = i\kappa_0 \mathcal{T}_-(\psi_0, Q) = 2i\kappa_0\big(\psi_0 + (z-1)Q(z)\big), \qquad \text{with } Q(z) = \sum_{i=0}^{N} Q_i z^i$$

and $\psi_0 \in \mathbb{C}$, we obtain for the eigenvalue problem

$$2i\kappa_0 A\big(\mathcal{T}_+(\psi_{oj}, Q_j), \mathcal{T}_+(\psi_0, Q)\big) = \frac{2\omega\mu}{\kappa_0} A\big(\mathcal{T}_-(\psi_{oj}, Q_j), \mathcal{T}_-(\psi_0, Q)\big) + \beta\psi_{0j}\psi_0.$$

This leads us to the matrix eigenvalue problem for the coefficient vector $\mathbf{\Psi}_j = (\psi_{j0}, Q_{j0}, \ldots, Q_{jN})^T \in \mathbb{C}^{N+2}$ of the function Ψ_j.

Problem 8.10. *Find* $\mathbf{0} \neq \mathbf{\Psi}_j \in \mathbb{C}^{N+2}$ *and the eigenvalues* $\gamma_j \in \mathbb{C}$ *such that*

$$2i\kappa_0 T_+^{N\top} T_+^N \mathbf{\Psi}_j = \gamma_j M \mathbf{\Psi}_j$$

together with the orthogonality relation $\mathbf{\Psi}_j^\top M \mathbf{\Psi}_k = \delta_{jk}$ *is fulfilled. The matrices* M, T_+^N *are in* $\mathbb{C}^{(N+2)\times(N+2)}$ *with*

$$M := \frac{2\omega}{\kappa_0} T_-^{N\top} T_-^N + \beta D,$$

with $T_\pm^N$ *from (7.8). All entries of* D *are except for* $D_{11} = 1$ *zero, i.e.* $D_{ij} = \delta_{i1}\delta_{j1}$.
Because the matrix T_+^N is regular, this generalized eigenvalue problem can be easily transformed to a complex symmetric eigenvalue problem.

The basis functions

With the help of these eigenfunctions and the solution of the eigenvalue problem 8.2 we are able to define the volume basis functions for the discrete spaces approximating U_H and V_H.

Definition 8.11. *(The volume basis functions on infinite elements) Let* $T_\square$ *and* $T_\triangle$ *be an infinite strip and an infinite triangle, respectively, in* $\mathcal{H}$. *Then, for* $1 \leq j \leq p+2$ *and* $1 \leq k, l \leq N+2$ *the Möbius and Laplace transform of a function from the discretized* *space for* U_H *is represented on a reference element of* $T_\Box$ *and* $T_\triangle$*, respectively, by the basis functions*

$$\begin{aligned} \hat{v}_{jk} &= \phi_j' \otimes \hat{\partial}_\xi \Xi_k && \textit{on } \hat{T}_\Box \\ \hat{\hat{v}}_{lk} &= \hat{\partial}_\eta \Theta_l \otimes \hat{\partial}_\xi \Xi_k && \textit{on } \hat{T}_\triangle. \end{aligned}$$

The Möbius and Laplace transform of functions from the discrete space approximating V_H *is spanned on the reference element of* $T_\Box$ *and* $T_\triangle$*,respectively, by the basis functions*

$$\begin{aligned} \hat{\boldsymbol{\tau}}^1_{\hat{T}_\Box jk} &= \big(\phi_j \otimes \hat{\partial}_\xi \Xi_k,\, 0\big)^\top && \textit{and} && \hat{\boldsymbol{\tau}}^2_{\hat{T}_\Box jk} = \big(0,\, \phi_j' \otimes \Xi_k\big)^\top, \\ \hat{\boldsymbol{\tau}}^1_{\hat{T}_\triangle lk} &= \big(\Theta_l \otimes \hat{\partial}_\xi \Xi_k,\, 0\big)^\top && \textit{and} && \hat{\boldsymbol{\tau}}^2_{\hat{T}_\triangle lk} = \big(0,\, \hat{\partial} \Theta_l \otimes \Xi_k\big)^\top. \end{aligned}$$

Here, Ξ_k *and* Θ_l *are eigenfunctions to the eigenvalues* γ_j *and* γ_k *of Problem 8.9. The functions* ϕ_j *are obtained by solving Problem 8.2 with polynomial order* p *and* $I = [0, h_T]$*, where* h_T *is the width of the strip. For strips with the infinite axis parallel to the* y *axis,* $p = p_x$ *and* $p = p_y$ *else.*

Note that for an eigenfunction $\Psi(z) = \frac{1}{2i\kappa_0}(\psi_0 + (z-1)Q(z))$ of Problem 8.9 with the coefficients $\psi_0, Q_0, \dots, Q_N$, the derivative $\hat{\partial}_z \Psi$ is of the form $\hat{\partial}_z \Psi(z) = \frac{1}{2}(\psi_0 + (z+1)Q(z))$. Because the eigenfunction ϕ_j for $j = p+2$ of Problem 8.2 is constant and corresponds to the eigenvalue zero, the basis functions v_{jk} and $\hat{\boldsymbol{\tau}}^2_{\hat{T}_\Box j,k}$ are zero for all k. We will keep them in the following for notational convenience. As basis functions on the infinite facets we take

Definition 8.12. *(The basis functions on infinite facets) Let* $F \in \mathcal{F}$ *be an infinite facet. Then the discrete space approximating* U_{HF} *is spanned by the basis functions* $g_{Ij} = g_{Oj}$ *with the Möbius and Laplace transform*

$$\hat{g}_{Ij}(\xi) = \hat{g}_{Oj}(\xi) = \begin{cases} \hat{\partial}_\xi \Psi_j(\xi) & \textit{on } F, \\ 0 & \textit{else,} \end{cases}$$

for $1 \le j \le N+2$*. The function* Ψ_j *is an eigenfunction of Problem 8.9.*

8.5.3 The linear system of equations

Now, we are in the position to investigate the influence of Hardy space infinite elements onto the structure of the system matrix. We will see in the following that the entries of the system matrix contain apart from constants just the precomputed eigenvalues. Thus, numerical integration is not needed which makes the assembly procedure very competitive.

The coupling between volume basis functions

The definition of the basis functions allows, as in the absorbing boundary condition case, for a cheap elimination of all volume unknowns in the interior and in the exterior domain, i.e. an inversion of the matrix A in (8.5). It is easy to see that the matrix A is still block diagonal with blocks A_T related to the elements $T \in \mathcal{T}$. As we will see in Lemma 8.13 these blocks are for Hardy space infinite elements, like for the rectangular elements, three by three block diagonal. In the following, we will stick to the notation from (8.6) and (8.7).

Lemma 8.13. *In the eigenfunction basis of Definition 8.11, the element matrix A_T of an infinite element T is block diagonal with three by three blocks.*

Proof. We will proof the lemma first for infinite strips $T_\square$. Using the notation for integrals over infinite facets from (8.6) we get by inserting the test functions into $\widetilde{B}_I$

$$\begin{aligned}\big(A^{T_\square}_{uu}\big)_{jk,mn} &= i\omega\epsilon\big(v_{jk}, v_{mn}\big)_{T_\square} = 2\omega\kappa_0\epsilon\big\langle\phi'_j, \phi'_m\big\rangle_{I_T} A\big(\hat{\partial}_\xi\Xi_k, \hat{\partial}_\xi\Xi_n\big)\\ &= 2\omega\kappa_0\epsilon\,\mathcal{B}_{h_T}(\phi_j, \phi_m)\,\mathcal{C}_N(\Xi_k, \Xi_n)\\ &= 2\omega\kappa_0\epsilon\lambda_j\gamma_k\delta_{jm}\delta_{kn},\end{aligned}$$

and $A^{T_\square}_{uu}$ is diagonal. The diagonality of $A^{T_\square}_{u\boldsymbol{\sigma}^1}$ follows by

$$\begin{aligned}\big(A^{T_\square}_{u\boldsymbol{\sigma}^1}\big)_{jk,mn} &= -\big(v_{jk}, \operatorname{div}\boldsymbol{\tau}^1_{mn}\big)_{T_\square} = 2i\kappa_0\big\langle\phi'_j, \phi'_m\big\rangle_{I_T} A\big(\hat{\partial}_\xi\Xi_k, \hat{\partial}_\xi\Xi_n\big)\\ &= 2i\kappa_0\lambda_j\gamma_k\delta_{jm}\delta_{kn}.\end{aligned}$$

A similar calculation together with symmetry arguments leads to the diagonality of $A^{T_\square}_{u\boldsymbol{\sigma}^2}$, $A^{T_\square}_{\boldsymbol{\sigma}^1 u}$ and $A^{T_\square}_{\boldsymbol{\sigma}^2 u}$. Furthermore, we obtain

$$\begin{aligned}\big(A^{T_\square}_{\boldsymbol{\sigma}^2\boldsymbol{\sigma}^2}\big)_{jk,mn} &= -i\omega\mu\big(\boldsymbol{\tau}^2_{jk}, \boldsymbol{\tau}^2_{mn}\big)_{T_\square} + \beta\big\langle\boldsymbol{\tau}^2_{jk}\cdot\boldsymbol{n}_{T_\square}, \boldsymbol{\tau}^2_{mn}\cdot\boldsymbol{n}_{T_\square}\big\rangle_{\partial T_\square}\\ &= -2\omega\mu\kappa_0\big\langle\phi'_j, \phi'_m\big\rangle_{I_T} A\big(\Xi_k, \Xi_n\big) - 4\kappa_0^2\beta\big\langle\phi'_j, \phi'_m\big\rangle_{I_T}\big(\Xi_k\Xi_n\big)\big|_1\\ &= \big\langle\phi'_j, \phi'_m\big\rangle_{I_T}\mathcal{C}_N(\Xi_k, \Xi_n) \quad = \lambda_j\,\delta_{jm}\delta_{kn},\end{aligned}$$

and by a similar argumentation $\big(A^{T_\square}_{\boldsymbol{\sigma}^1\boldsymbol{\sigma}^1}\big)_{jk,mn} = -2i\kappa_0\gamma_k\,\delta_{jm}\delta_{kn}$. Thus, these two blocks are diagonal as well. Finally, the orthogonality of the $\boldsymbol{\tau}^1_{jk}$ and $\boldsymbol{\tau}^2_{mn}$ leads immediately to

$A^{T_\square}_{\boldsymbol{\sigma}^1\boldsymbol{\sigma}^2} = A^{T_\square}_{\boldsymbol{\sigma}^2\boldsymbol{\sigma}^1} = 0$. Because all nine coupling blocks are diagonal or zero, a reordering of the degrees of freedom leads to a three by three block diagonal matrix $A_{T_\square}$.

For infinite triangles the proof follows the same strategy. $\square$

We should mention that the three by three blocks in $A_{T_\square}$ corresponding to the basis functions $\boldsymbol{\tau}^1_{jk}, \boldsymbol{\tau}^2_{jk}$ and v_{jk} which contain the constant eigenfunction ϕ_j, i.e. $j = p+2$ degenerate to a scalar. Because of the three by three block matrix structure elimination of the volume degrees of freedom can be done cheaply just by inverting these three by three matrices.

The coupling between facet basis functions

The coupling elements of facet basis functions are stored in the block C from (8.5), which is the matrix representation of the bilinear form $\widetilde{B}_F$. If we stick to one facet F we change the notation as previously from in and outgoing waves G_I, G_O to up and down or right and left going waves G_+, G_-, whether the facet is horizontal or vertical. Thus, G_+ is the outgoing wave G_O for the element placed below or left of the facet, and G_+ equals G_I for the other element. For G_- the situation is vice versa. With the same argumentation as for interior facets, it is sufficient just to consider the matrix blocks C_F related to any facet F.

Lemma 8.14. *Using the eigenfunction basis of Definition 8.5 and 8.12 the matrix block C_F for a facet F is*

$$C_F = \begin{pmatrix} C^F_{G_+G_+} & C^F_{G_+G_-} \\ C^F_{G_-G_+} & C^F_{G_-G_-} \end{pmatrix} = \begin{pmatrix} a\,\mathrm{diag}(\rho_k) & -a\,\mathrm{diag}(\rho_k) \\ -a\,\mathrm{diag}(\rho_k) & a\,\mathrm{diag}(\rho_k) \end{pmatrix},$$

where the four blocks of C_F are from $\mathbb{C}^{n\times n}$, and $1 \le k \le n$. For finite facets F, ρ_k are the eigenvalues λ_k of Problem 8.2 solved on F with $n = p+1$ and $a = 2\beta\tilde{\epsilon}/\mu$. For infinite facets ρ_k equals the eigenvalue γ_k of Problem 8.9 for $n = N+2$ and $a = -4i\kappa_0\beta\tilde{\epsilon}/\mu$.

Proof. Because g_I, g_O and g_+, g_-, respectively, are both described by the same set of basis functions, we get for $\bullet, * \in \{I, O\}$ or $\{+,-\}$

$$\left\langle g_{\bullet j}, g_{*k}\right\rangle_F = \begin{cases} \left\langle \psi'_j, \psi'_k\right\rangle_F & = \lambda_k\delta_{jk} \quad \text{for } F \text{ finite,} \\ -2i\kappa_0\, A(\hat{\partial}_\xi\Psi_j, \hat{\partial}_\xi\Psi_k) & = -2i\kappa_0\gamma_k\delta_{jk} \quad \text{for } F \text{ infinite.} \end{cases}$$

Thus, the contribution of $\widetilde{B}_F$ coming from one neighboring element T of F to the blocks $C^F_{G_+G_+}, C^F_{G_-G_-}$ is $\beta\tilde{\epsilon}/\mu\,\mathrm{diag}(\lambda_k)$ and $-2i\kappa_0\beta\tilde{\epsilon}/\mu\,\mathrm{diag}(\gamma_k)$, respectively. For $C^F_{G_+G_-}, C^F_{G_+G_-}$

it is $-\beta\tilde{\epsilon}/\mu\,\mathrm{diag}(\lambda_k)$ or $2i\kappa_0\beta\tilde{\epsilon}/\mu\,\mathrm{diag}(\gamma_k)$. The fact that each facet is surrounded by two neighboring elements completes the proof. $\square$

A reordering of the degrees of freedom leads to a two by two block diagonal matrix.

The coupling between facet and volume degrees of freedom

Facet and volume degrees of freedom couple via the bilinear form $\tilde{B}_{IF}$. Therefore, its matrix representation D (compare (8.5)), or more precisely the element matrix D_T for an infinite element T is investigated. This element matrix has a block structure and we denote the blocks coupling the basis functions $\boldsymbol{\tau}^1_{jk}$ with the facet basis functions g_{Om} of a facet placed left (l), right (r) or at the bottom (b) of the reference element $\hat{T}$ by $D^{Tl}_{\boldsymbol{\sigma}^1 G_O}$, $D^{Tr}_{\boldsymbol{\sigma}^1 G_O}$ and $D^{Tb}_{\boldsymbol{\sigma}^1 G_O}$, respectively.

If we stick to the infinite reference strip $\hat{T}_\square$, $\boldsymbol{\tau}^2_{jk}\cdot\boldsymbol{n}_{T_\square}$ is zero on infinite facets and $\boldsymbol{\tau}^1_{jk}\cdot\boldsymbol{n}_{T_\square}$ on the finite ones. Therefore $\boldsymbol{\tau}^2$- degrees of freedom do not couple to facet unknowns of infinite facets and $\boldsymbol{\tau}^1$-degrees of freedom not to facet unknowns of a finite facet. Thus, the non zero coupling blocks are according to the definition of the basis functions 8.5, 8.11 and 8.12

$$\begin{aligned}
\big(D^{T\square l}_{\boldsymbol{\sigma}^1\bullet}\big)_{jk,m} &= 2i\kappa_0\big(1\pm\beta\sqrt{\tilde{\epsilon}/\mu}\big)\,\phi_j(0)\,A(\hat{\partial}_\xi\Xi_k,\hat{\partial}_\xi\Psi_m)\\
&= 2i\kappa_0\big(1\pm\beta\sqrt{\tilde{\epsilon}/\mu}\big)\,\phi_j(0)\,\gamma_k\delta_{km},\\
\big(D^{T\square r}_{\boldsymbol{\sigma}^1\bullet}\big)_{jk,m} &= -2i\kappa_0\big(1\pm\beta\sqrt{\tilde{\epsilon}/\mu}\big)\,\phi_j(h_T)\,\gamma_k\delta_{km},\\
\big(D^{T\square b}_{\boldsymbol{\sigma}^2\bullet}\big)_{jk,m} &= -2i\kappa_0\big(1\pm\beta\sqrt{\tilde{\epsilon}/\mu}\big)\,\Xi_k(1)\,\lambda_j\delta_{jm},
\end{aligned}$$

where the sign $-$ is taken for $\bullet = G_O$ and $+$ for $\bullet = G_I$. Note that $\phi_j(0)$ and $\phi_j(h_T)$ are just a linear combination of the first two components of the basis functions vector representation, and $\Xi_k(1)$ is proportional to the first component of the eigenvector associated to Ξ_k.

For completeness we list up the nonzero coupling blocks for the infinite triangle

$$\begin{aligned}
\big(D^{T\square l}_{\boldsymbol{\sigma}^1\bullet}\big)_{jk,m} &= -4\kappa_0^2\big(1\pm\beta\sqrt{\tilde{\epsilon}/\mu}\big)\,\Theta_j(1)\,\gamma_k\delta_{km},\\
\big(D^{T\square b}_{\boldsymbol{\sigma}^2\bullet}\big)_{jk,m} &= -4\kappa_0^2\big(1\pm\beta\sqrt{\tilde{\epsilon}/\mu}\big)\,\Xi_k(1)\,\gamma_j\delta_{jm}.
\end{aligned}$$

8.5.4 The preconditioner

From the sparsity pattern of the blocks A, C and D we can conclude that adding Hardy space infinite elements and infinite facets, respectively, does not change the structure of the Schur complement matrix $S = C - D^\top A^{-1} D$. Thus, coupling blocks between parallel facets in S are still two by two block diagonal, even if the facets are infinite, and between perpendicular facets these blocks are full.

This encourages us to use for the PCG iteration the same Additive Schwarz preconditioner P introduced for the problem with absorbing boundary conditions,

$$P = \sum_i \left(P^{F_i}\right)^{-1}.$$

Here P^{F_i} collects all the coupling entries from the Schur complement matrix S between degrees of freedom of the finite or infinite facet F_i. Consequently, P^{F_i} are the diagonal blocks of S, and their inversion is equivalent to the inversion of two by two matrices.

8.6 Numerical results

The numerical results presented in this section were computed, if not said differently, on a computational domain $\Omega = (0,2)^2$. An incoming wave from the left with a Gaussian shaped amplitude of $u_{in}(0,y) = \exp\left(-\frac{(y-1)^2}{0.1}\right)$ is prescribed. On the other boundaries $u_{in} = 0$.

8.6.1 Absorbing boundary conditions

For the following computations first order absorbing boundary conditions were used. Thus, the input data g has to be chosen as $g = 2u_{in}$. These results we already published in [HHS10].

An optimal choice for β

This section is started by studying the dependence of the number of iterations needed for solving the Schur complement system via a PCG iteration with the preconditioner from Subsection 8.3.4, on the stabilization parameter β. First, we consider a constant value $\epsilon = 1$, $\mu = 1$ and the angular frequency was chosen to be $\omega = 10\pi$. The domain Ω is divided into 4×4 equally sized elements with a polynomial order of 50 in each spatial

β	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1	2	3	4	5
$\epsilon = \frac{1}{16}$	95	89	81	76	66	54	44	40	44
$\epsilon = \frac{1}{4}$	120	105	94	78	59	41	51	55	58
$\epsilon = 1$	140	121	102	73	35	54	79	88	103
$\epsilon = 4$	108	86	62	34	65	99	151	194	239
$\epsilon = 16$	50	34	47	70	131	317	597	919	1366

Table 8.1: iteration numbers for different values of ϵ and β

direction, which is enough to get a good resolution of the wave. The number of iterations required for different values of ϵ and β is given in Table 8.6.1. The results show that the iteration number is very sensitive to the choice of β for large values of ϵ and that the optimal choice would be $\beta = \sqrt{\mu/\epsilon}$, which fits well into the discussion of Subsection 5.2.1. There, we remarked that for this choice of β the local problem on the element level, which is needed to be solved in order to eliminate the volume unknowns, corresponds to a Helmholtz problem with absorbing boundary conditions on the element boundary, and Lemma 5.8 guarantees energy conservation. From our problem Formulation 8.1 we can see that for such a choice of β the volume solution just depends on the incoming waves G_I and that there is only indirect coupling to the outgoing waves G_O.

Choosing a good value for β is much more difficult if ϵ is not constant on the domain. Following the discussion in the last paragraph, an obvious option would be to choose $\beta = \sqrt{\mu/\epsilon}$ for each element separately. Because the eigenvalue problem 8.2 depends on β, it has to be solved for each element size and each β which is expensive and compensates the advantages of the method presented in this chapter. Furthermore, if ϵ jumps across a facet, a different eigenfunction basis is used on the two neighboring elements which leads to full coupling blocks in the Schur complement for this facet. Therefore, a global choice of β is desirable.

In order to examine the dependence of the iteration number on β for a non constant ϵ, we assume for our computational domain $\Omega = (0,2)^2$

$$\epsilon(\boldsymbol{x}) = \begin{cases} \epsilon_{per} & \text{for } \boldsymbol{x} \in (1, \frac{5}{3}) \times (\frac{1}{3}, 1), \\ 0 & \text{else,} \end{cases}$$

for different values of ϵ_{per} and $\mu = 1$. A uniform mesh of 6 × 6 elements, by which the perturbation can be resolved, is used together with a polynomial order of 120 in order to approximate the wave for $\omega = 15\pi$. Figure 8.4 shows the iteration numbers for different values of ϵ_{per} and β. The plots show that the iteration number increases with growing ϵ_{per}.

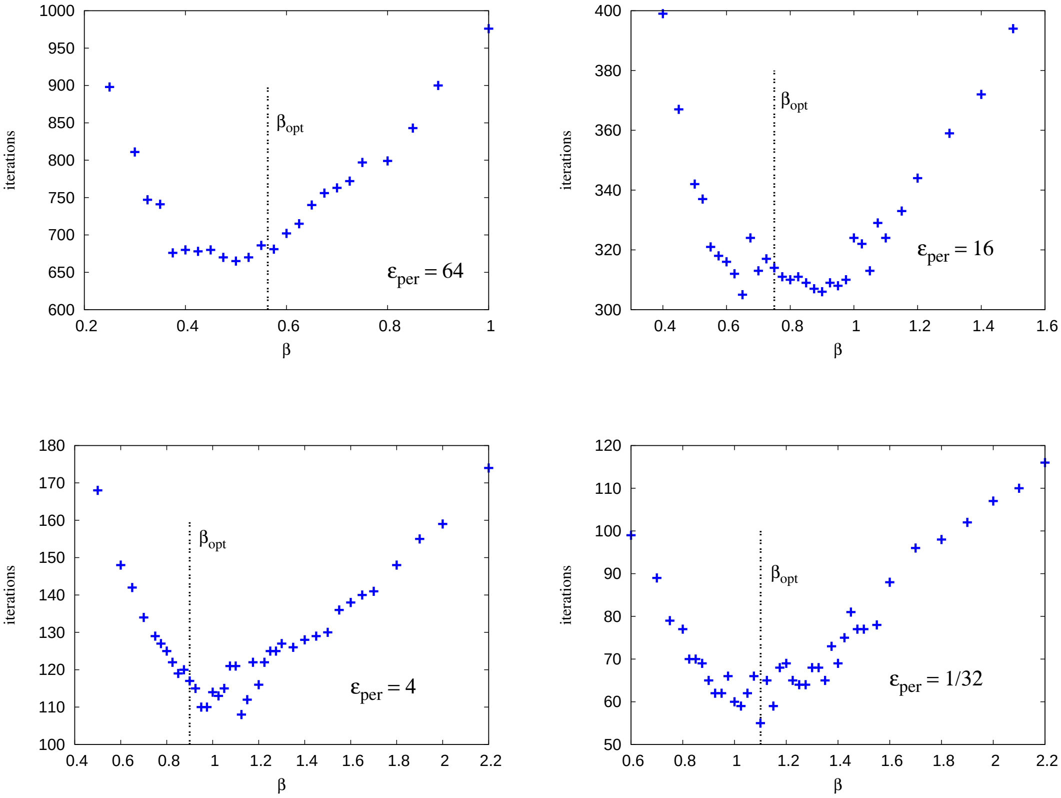

Figure 8.4: Number of iterations depending on β for different values of ϵ in the perturbed region, i.e. $\epsilon_{per} = 64$ (top left), $\epsilon_{per} = 16$ (top right), $\epsilon_{per} = 4$ (bottom left) and $\epsilon_{per} = \frac{1}{32}$ (bottom right)

.

Motivated by the results with constant ϵ, we choose a global β as

$$\beta_{opt} = \sqrt{\mu/\bar{\epsilon}}$$

with $\bar{\epsilon}$ as an effective ϵ of the domain. We made good experience by taking for $\sqrt{\bar{\epsilon}}$ the mean value of $\sqrt{\epsilon}$ on the domain Ω, i.e.

$$\sqrt{\bar{\epsilon}} = \frac{1}{|\Omega|} \int_{\Omega} \sqrt{\epsilon} dx.$$

The value for the corresponding β_{opt} is marked by a dashed line in Figure 8.4. From the

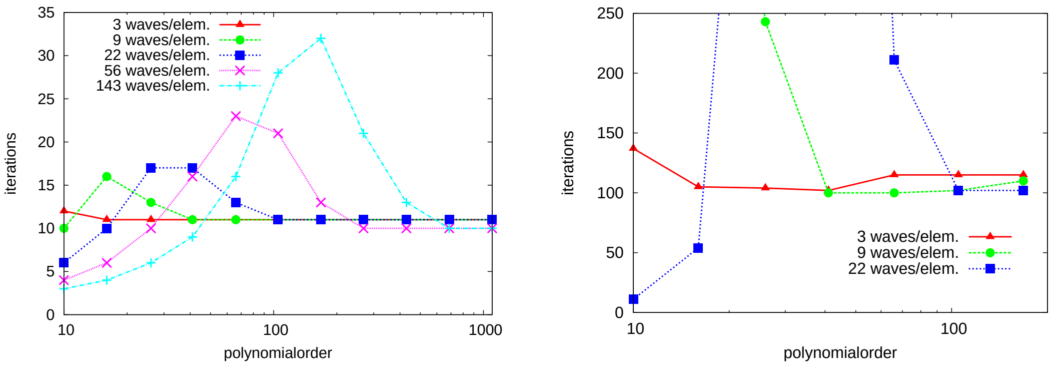

Figure 8.5: For the red, green, blue, magenta and cyan lines ω is chosen such that the size of one element is 3, 9, 22, 56 and 143 wavelength, respectively. For the left hand plot the mesh was 2 times 2 elements, and consequently $\omega = 6\pi$, 18π, 44π, 106π and 286π. For 12 times 12 elements in the right hand plot we took $\omega = 36\pi$, 108π and 264π.

numerical results one can see that β_{opt} is a good choice for the stabilizing parameter and it gives almost an optimal iteration number.

Dependence of the iteration number on frequency, polynomial order and mesh size

Now, the number of iterations is studied for different polynomial orders and angular frequencies ω. In the following calculations ϵ, μ and β were chosen as one. In Figure 8.5 the number of iterations is plotted against the polynomial order for different frequencies ω, or more precisely number of waves per element. The two plots presented there were calculated for two different uniform meshes consisting of 2×2 elements (left) and 12×12 elements (right). From [Ain04] we know that about at least three till four unknowns per wavelength are needed to resolve the wave. From Figure 8.5 we can conclude that if the polynomial order is much to small to resolve the solution, we get small iteration counts. For a growing polynomial order the number of iterations increases rapidly, or convergence of the solver is lost. If the polynomial order is chosen such that the wave can be resolved, thus, it is between three and four times the number of waves per element, the number of iterations reaches again a minimum. The number of iterations at this minimum seems to be almost independent of the frequency, but it depends on the mesh size. For 2×2 elements, or $h = 1$ 10 iterations are needed, while 12×12 elements or $h = \frac{1}{6}$ results in about 100 iterations. A further increase in the polynomial order leads just to a small growth of the number of

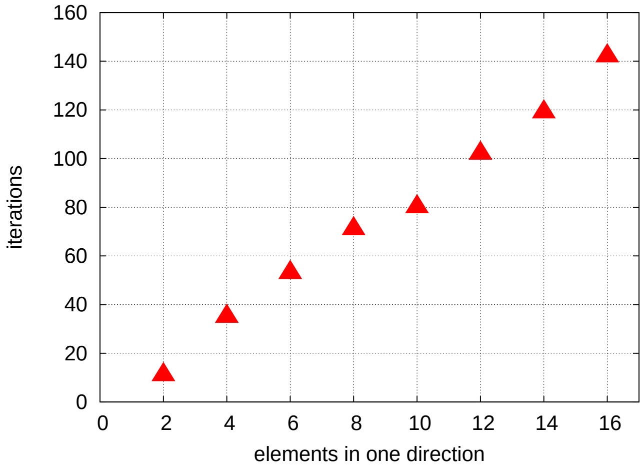

Figure 8.6: Number of iterations for different mesh sizes. The frequency was chosen, such that the size of one element is 22 wavelengths

iterations.

Next, we study the dependence of the iteration number on the mesh size, or more precisely for different uniform meshes. Therefore, ω was taken such that each element is of the size of 22 wavelengths, i.e. $\omega = 44\pi/h$, and the polynomial order was chosen as 105, which corresponds to about five unknowns per wavelength, and it is therefore large enough to resolve the solution. Figure 8.6 shows that for such a setting the number of iterations is proportional to the number of elements in one spacial direction. One interpretation of this fact would be that a fixed number of iterations is needed to propagate the input data given by u_{in} on the boundary through one element. According to Figure 8.5 this number of iterations seems to be almost independent of the frequency if the wave is resolved. Based on this observation, the "speed of propagation" should be also frequency independent.

Computational times

Apart from iteration counts, computational times are an issue. In Figure 8.7 the computational times on an Intel 2GHz processor are plotted against the polynomial order for two different uniform meshes with 4×4 or 8×8 elements. The angular frequency is chosen for both meshes such that the element has a size of nine wavelengths. Thus, a polynomial order of about thirty is needed to resolve the wave. The plot shows that the solution time

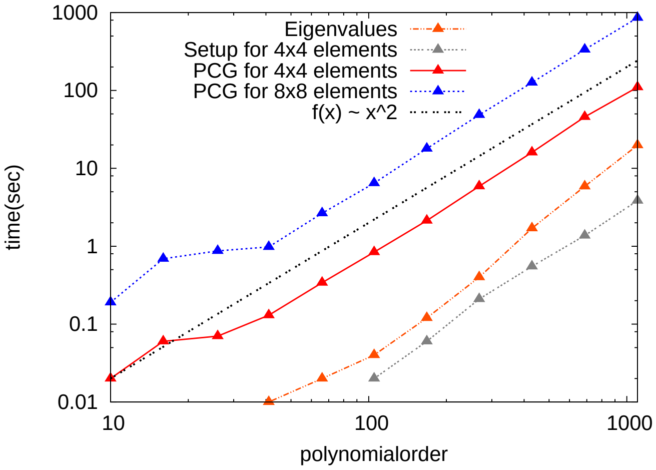

Figure 8.7: computational times for a uniform mesh of 4 × 4 and 8 × 8 elements

for both meshes, and the setup of the system of equations is of order p^2. Taking into account that the number of iterations grows just slightly with the polynomial order, we can conclude that one iteration is at least of complexity p^2. For polynomial orders less than 30 the wave can not be resolved, and a higher iteration number is needed, which perturbs the $\mathcal{O}(p^2)$ behavior in the solution times. Although, solving the eigenvalue problem is of higher complexity, the computational time needed for this is negligible compared to the solution process for the relevant polynomial orders.

Hanging nodes

In order to study the influence of hanging nodes onto the number of iterations, we generate a hanging node mesh by refining a uniform mesh to the point $P = (1.51, 1.01)$ according to the following strategy. We start the procedure by meshing the computational domain $\Omega = (0, 2)^2$ by one single element (refinement level zero). In each refinement step the element containing the point P is divided into four equally sized elements with half the polynomial order. Because of approximation properties the size of neighboring elements should not differ too much, i.e. we want to avoid arbitrary level hanging nodes. Therefore, an element is also refined if P is in a distance of ch, where h is the edge length of the element. The global constant c defines implicitly the mesh grading. Figure 8.8 shows for $c = 1$ a mesh of refinement level ten which contains 61 elements.

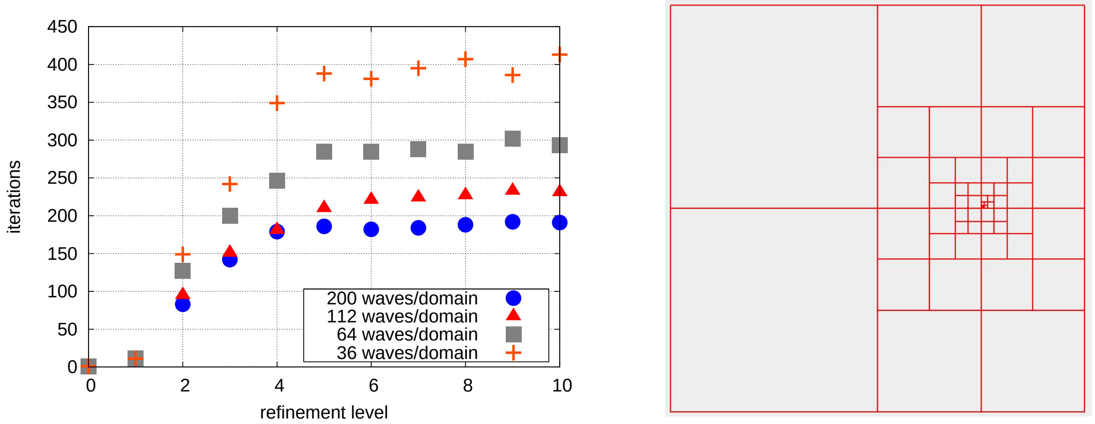

Figure 8.8: number iterations for different refinement levels and wavelengths (left) and the mesh with hanging nodes of refinement level 10 (right)

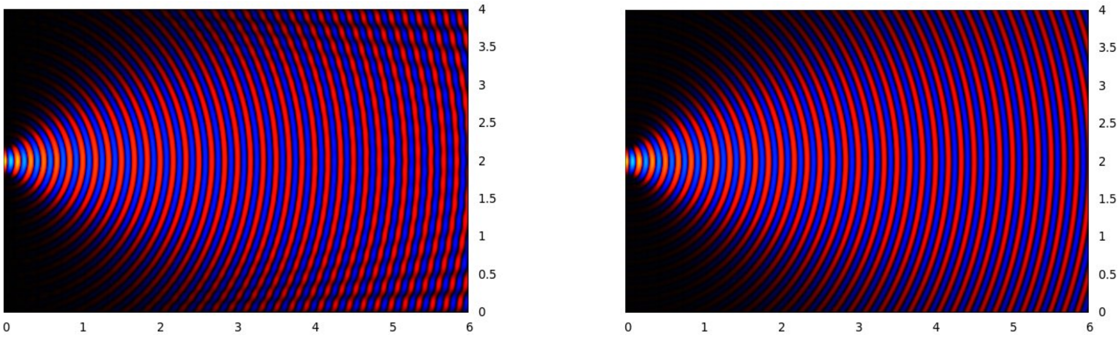

Figure 8.9: Solution of the Helmholtz equation with a first order absorbing boundary condition (left) and with a Hardy space boundary condition (right)

For such meshes the plot in Figure 8.8, where the number of iterations is plotted against the refinement level for different angular frequencies, was computed. For a small refinement level the number of iterations grows rapidly, while it stays approximately constant for refinement levels higher than five. Contrary to the uniform mesh case, where the number of iterations seemed to be frequency independent for a polynomial order large enough, for hanging nodes meshes it depends on the frequency. An increase in the angular frequency, which corresponds to an increase in the number of waves per domain, leads to a growing iteration number.

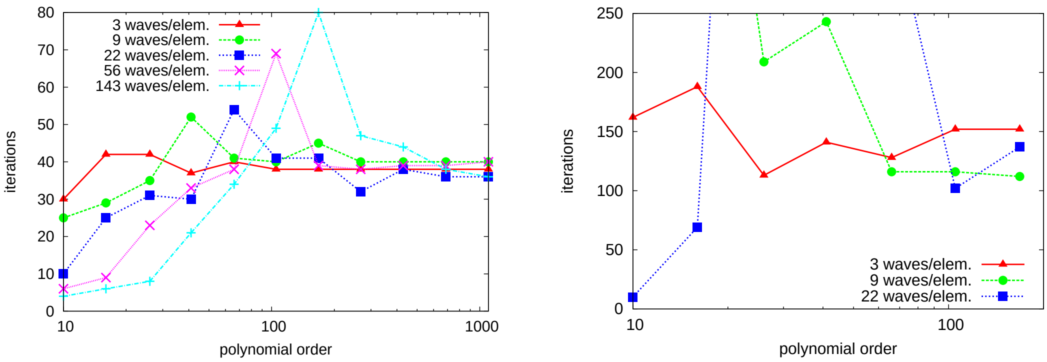

Figure 8.10: For the red, green, blue, magenta and cyan lines ω is chosen such that the size of one element is 3, 9, 22, 56 and 143 wavelength, respectively. The underlying mesh consisted of 2×2 (left) and 12×12 elements (right) with infinite Hardy space elements of order 10.

8.6.2 Hardy space infinite elements

We want to start this section with an example which shows the advantages of Hardy space infinite elements compared to a first order absorbing boundary condition.

Comparing the boundary conditions

In Figure 8.9 the Helmholtz equation is solved on $\Omega = (0,6) \times (0,4)$ where an incident beam with a Gaussian shape $u_{in} = \exp\left(-\frac{(y-2)^2}{0.1}\right)$ is prescribed via absorbing boundary conditions (left) and Hardy space infinite elements (right). As underlying mesh, a mesh of squared elements with an edge length of one was taken, and for $\omega = 12\pi$, which is equivalent to 6 wavelengths per element, a polynomial order of 30 was chosen. As already argued, absorbing boundary conditions are only exact for one single angle of incidence. If the angle of incidence does not exactly match with the angle of incidence expected by the boundary condition reflections appear. This can be seen in the left hand plot of Figure 8.9. There, reflections appear on the top right and bottom right corner of the domain, which leads to an interference pattern. The right hand plot, calculated with a Hardy space infinite elements, is free of reflections on the boundary and does not show an interference pattern.

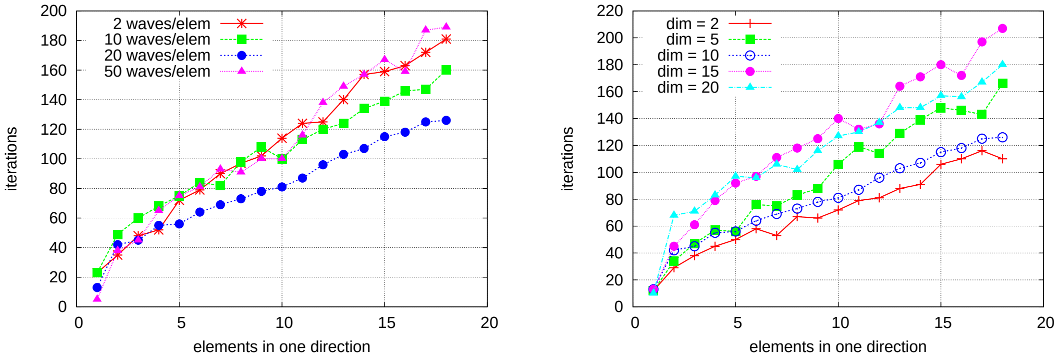

Figure 8.11: Iteration counts for different mesh sizes. For the left hand plot the dimension of the Hardy space is kept constant to 10 and the frequency is changed, while in the right hand plot different dimensions of the Hardy space are used for 20 waves per element.

Dependence on the polynomial order

We continue the numerical results section by repeating the calculations for Figure 8.5 with Hardy space infinite elements instead of an absorbing boundary condition. In Figure 8.10 the number of iterations for solving our model problem on $\Omega = (0,2)^2$ with input data $u_{in}(x,y) = \exp\left(-\frac{(y-1)^2}{0.1}\right)$ for $x = 0$ is plotted against the polynomial order for different angular frequencies and two different uniform meshes containing 2×2 elements (left) and 12×12 elements (right). The order of the Hardy elements was kept constant to 10.

Although, solving a problem with Hardy space infinite elements requires for our preconditioner more iterations, we observe by comparing Figure 8.5 and 8.10 that the major features of the plots are the same for both types of boundary conditions. Choosing a polynomial order much to small to resolve the solution requires a small number of iterations. Increasing the polynomial order leads first to a fast growth in the iteration number. If the polynomial order is high enough to resolve the wave, the number of iterations decreases rapidly, and a minimum is reached for about four till five unknowns per wavelength. A further increase in the polynomial order leads just to a small increase for the iteration counts. Furthermore, we can see that a finer mesh requires more iterations.

Dependence on the element size and the dimension of the Hardy space

In order to investigate the dependence of the iteration counts on the mesh size for our model problem, the number of iterations is plotted in Figure 8.11 against the number of elements in one spatial direction. In the left hand plot the angular frequency, or more precisely the number of waves per element, is changed, and the dimension of the Hardy space is kept constant to ten. Whereas the right hand plot shows the results for different dimensions of the Hardy space and 20 waves per element. The polynomial order was chosen such that there were five unknowns per wavelength, which is enough to get a good resolution of the wave.

While we concluded from Figure 8.6 that the number of iterations is proportional to the number of elements in one spatial direction for absorbing boundary conditions, Figure 8.11 indicates still a linear dependence of the iteration counts on the mesh size, but the straight line is shifted along the iteration number axis. From the plots we can see that the slope of a straight line trying to fit the numerical data is about the same, independent of frequency or the dimension of the Hardy elements. Therefore increasing the number of elements in one spatial direction by one increases the iteration counts by a constant number. Thus, the PCG solver needs a constant number of iterations in order to transport the data fixed by the incoming wave on the boundary across one element. Furthermore, exchanging absorbing boundary conditions by Hardy space infinite elements increases the number of iterations, which can be seen by the shift of the straight line fitting the numerical data to higher iteration numbers. The right hand plot in Figure 8.11 indicates that this shift grows with a growing dimension of the Hardy elements.

Large scale examples

To conclude the results section, we want to present a large scale problem. We assume that an incoming beam with a Gaussian shaped amplitude is scattered at different obstacles. In order to prescribe a beam coming from the left of the computational domain $\Omega := (0,4) \times (0,3)$, we use $u_{in} = \exp\left(-\frac{(y-y_0)^2}{0.05}\right)$ at $x = 0$ and $u_{in} = 0$ else. As obstacle an equilateral triangle with side length 2.3 and a circle of diameter 2 were chosen. For the triangle y_0 was chosen to be 1.5, while we took $y_0 = 2$ for the circle. The shapes of these obstacles are visualized in green in the two meshes of the computational domain in Figure 8.12. In the red area of the mesh we assumed ϵ to be one, and the obstacle consists of a material with $\epsilon = 2$. This parameter leads for $\omega = 120\pi$ to an effective size of the obstacles of 195 and 170 wavelengths, respectively.

The meshes in Figure 8.12 are obtained from one single element (refinement level zero) by eight refinement steps towards the boundary of the obstacle. In each refinement step an element with edge length h and polynomial order p is divided into four equally sized elements of polynomial order $\frac{p}{2}$, if the boundary crosses the element or if the distance

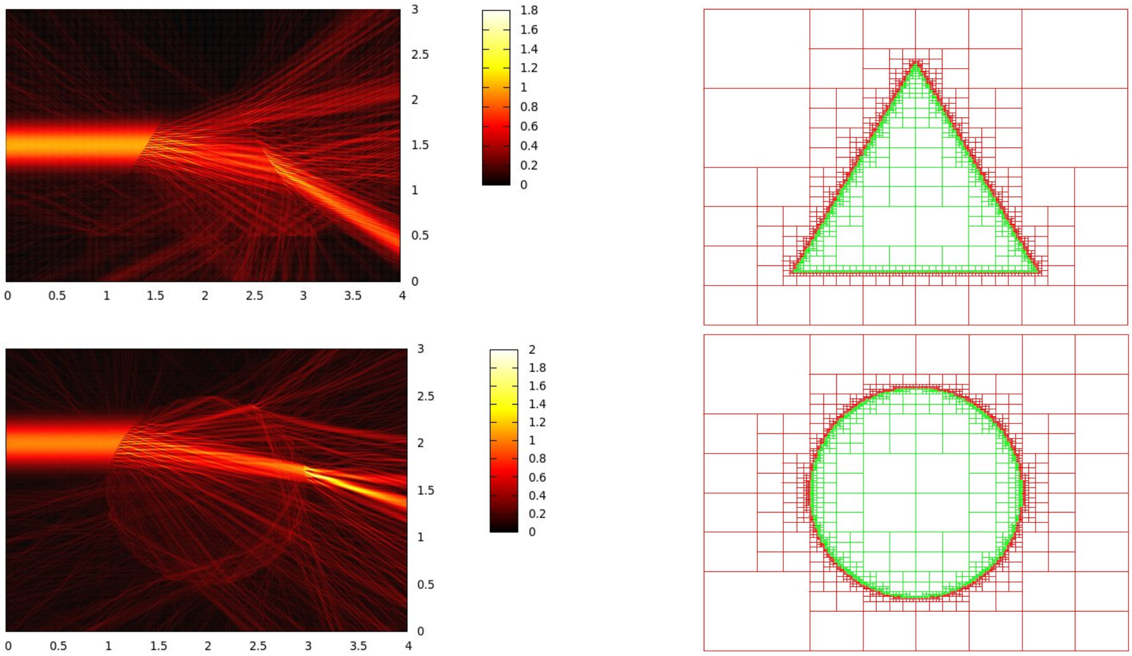

Figure 8.12: Amplitude of the total field for scattering at a triangular (top left) and a circular (bottom left) obstacle. The right hand side shows the underlying meshes. For $\omega = 120\pi$ the size of the domain corresponds to 240 times 180 wavelengths ($\epsilon = 1$). Considering $\epsilon = 2$ in the obstacles leads to an effective side length of the triangle of 195 wavelengths and to an effective diameter of the circle of 170 wavelengths.

between element and boundary is less than ch with a global constant c. Taking $c = 0.1$ results in 2506 elements for the mesh with the circular obstacle and in 2440 elements for the problem with the triangular scatterer. As polynomial order we took for the element of refinement level zero 1500 for each spatial direction. Thus, the largest elements appearing in the two meshes are of level two with order 375, and the smallest elements which are of level eight have polynomial order six. Non reflecting boundary conditions were realized by Hardy space infinite elements with dimension five.

In the left hand plots the absolute value of the total field u is plotted for the described settings. For both settings the eigenvalue problem was solved on an Intel 2.5GHz processor in 0.9 seconds, and for assembling the matrices 5.5 seconds for the circular obstacle and 6.7 seconds for the triangular one were needed. The by far most time consuming part is the iterative solver. The problem with the triangular object was solved with 2622 iterations in 3.6 hours, and for the other problem 9427 iterations and 10 hours were needed. One reason for the high iteration numbers is that the obstacles form small cavities, and the PCG solver suffers from internal reflections.

Chapter 9

Simulation of diffraction gratings

In the last chapters we have described methods realizing transparent boundaries. The easiest to implement is the absorbing boundary condition. But it causes reflections, as soon as an outgoing wave hits the boundary at a slightly different angle as expected (compare Figure 8.9). This advantages can be compensated at the cost of more implementation effort for example by Hardy space infinite elements. There, the region outside the domain is meshed by semiinfinite elements, where a special set of basis function is used to approximate the solution. In this chapter we are going to present a realization of transparent boundary conditions which is well suited for scattering problems in periodic media also called gratings. This method we already published in [HSSZ09].

Diffraction gratings are of big interest in physics. While in the past gratings were mainly used in spectroscopy, a large variety of applications appeared during the last decades. Nowadays gratings are important in nanotechnology as anti reflection coatings or grating couplers and in the extreme ultra violet technology. For such structures the far field, i.e. the field in a large distance above or below the grating, consists of a finite number of plane waves or modes with directions which are well known [BW99, Rei05] and which only depend on the period of the structure. In order to determine the intensities of these waves, Maxwell's equations have to be solved. For a detailed discussion on gratings we recommend the books of Petit [Pet80] and Nevière and Popov [NP03].

One of the most frequently used methods to deal with gratings is the rigorous coupled wave (RCW) method, see [Li97, MG81, NP03], where the computational domain is divided into parallel layers, and a plane wave expansion of the solution is used in each slice. Due to the special structure a fast application of transfer operators is possible. But, because of the slicing, the method faces difficulties approximating curvilinear geometries. Apart from finite difference methods [NCMC71, Vin78, Pet80], integral equation methods [Pet80, PMM97] are another possibility to calculate the reflection pattern of a periodic structure. There, an induced surface distribution on the boundary of the optical element, described by the integral form of the Helmholtz equation, is used to compute the solution at any point in the space. This method is not restricted to an infinite periodic system and can be used also for general scattering problems. An alternative are boundary variation methods [BR93, BR96] which make use of the analytic dependence of the diffracted fields on variations of the surface of the diffractive element. When using such a method, the derivatives of the efficiency with respect to the grating height is calculated via a recursion formula. For completeness, we have to mention the curvilinear coordinate method [CDCM82, Gar99] which is applicable for multi coated gratings, i.e. periodic structures of multiple layers with equally shaped interfaces. The method is based on a transformation which transforms the whole system to a system of parallel planes with constant coefficients.

In our method (compare [HSSZ09]) we use finite elements to solve the diffraction problem, see also [Abb91, Bao97, BD00, EHS02, Sch04, BCW05, SZB$^+$07]. Here, the near field domain is divided into simple elements, and the unknown field is approximated by piecewise polynomials which satisfy continuity conditions on the element interfaces. This approach is very flexible when modeling complicated domains. According to the theory of Bloch-Floquet the whole periodic grating can be reduced onto one unit cell by stating quasi periodic boundary conditions [Flo83, Blo28, Kuc01]. A difficult task is to model the semiinfinite domains above and below the grating. Commonly used are the perfectly matched layer (PML) technique [BCW05, SZB$^+$07] or transparent boundary conditions [Abb91, Bao97, BD00, EHS02, Sch04] derived from integral equation methods. The first one can be seen as a complex continuation of the computational domain which is free of reflections at the interface and damps out the solution. Truncating this continued exterior domain to a finite domain causes only small reflections which can be well controlled. The second method mentioned above is derived by integral equation methods, and it can be efficiently implemented on uniform grids by the fast Fourier transform.

The approach we present in this chapter uses propagating and evanescent waves in order to describe the field in the semiinfinite far field domain. The novelty is that these two different approximations for the near and the far field are coupled via the method of Nitsche [Nit71, Ste98, HHL03], which is very similar to discontinuous Galerkin methods. For further reading on discontinuous Galerkin methods in general see [ABCM02, CKS00] and [HPSS05, PSM02] for discontinuous Galerkin for Maxwell's equations. Continuity between these two different approximation spaces is now enforced directly by the variational formulation itself. Furthermore, planar layers modeled by the transfer matrix method [NP03, Rei05] fit well into this concept.

This chapter is organized as follows. In Section 9.1 the diffraction problem is described, and the Rayleigh expansion is stated. Section 9.2 deals with gratings periodic in one direction. There, the variational formulation for the finite element domain under quasi periodic boundary conditions is stated. The waves describing the far field are coupled via the method of Nitsche to the near field approximation. In Section 9.3 this approach is generalized to the full Maxwell case with gratings periodic in two spatial directions. Section 9.4 deals with the modeling of planar layers, and the chapter is finished with numerical examples.

9.1 Preliminaries

In this section we derive the far field solution. Therefore, we introduce the Rayleigh expansion.

9.1.1 The geometry

Before starting, we have to introduce some notations. We consider structures in $\mathbb{R}^3$ which are at least periodic in x-direction with period d_x. For biperiodic gratings an additional periodicity is assumed in y-direction with period d_y. Due to the periodicity material parameters like the electric permittivity ϵ and the magnetic permeability μ are invariant under translation into x-direction and in the biperiodic case into y-direction as well,

$$\begin{aligned} \epsilon(x + nd_x, y + md_y, z) &= \epsilon(x, y, z) \\ \mu(x + nd_x, y + md_y, z) &= \mu(x, y, z) \end{aligned}$$

with $n, m \in \mathbb{Z}$. Non-orthogonal periodicity can be treated similarly. The near field domain is according to Figure 9.1 defined as $\Omega^I := \big\{(x, y, z) \,:\, a < z < b\big\}$ with $a < b$, $\Omega^+ := \big\{(x, y, z) \,:\, z > b\big\}$ and $\Omega^- := \big\{(x, y, z) \,:\, z < a\big\}$ we call the semiinfinite far field domains above and below the grating and $\Omega^\pm = \Omega^+ \cap \Omega^-$.

Due to the periodicity, the infinite computational domain Ω^I can be restricted to the unit cell Ω^I_p (compare Figure 9.1). The restrictions of Ω^+ and Ω^- we call Ω^+_p and Ω^-_p. The interfaces Γ^+ and Γ^- we define as $\Gamma^+ := \overline{\Omega}^I_p \cap \overline{\Omega}^+_p$ and $\Gamma^- := \overline{\Omega}^I_p \cap \overline{\Omega}^-_p$. The periodic

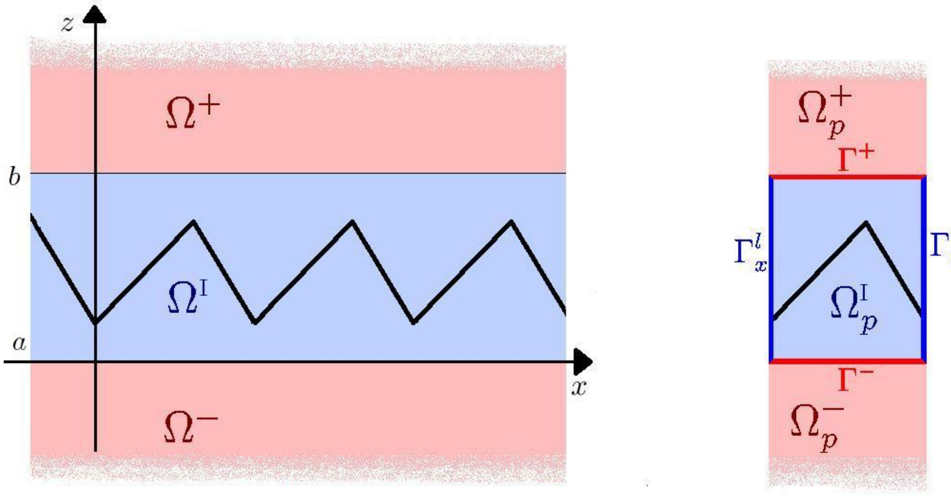

Figure 9.1: Geometry of the problem

boundaries perpendicular to the x-axis are called Γ_x^l and Γ_x^r, and in the biperiodic case Γ_y^l and Γ_y^r describe the boundaries orthogonal to the y-axis.

In the following, we use the superscript $+$ and $-$ for parameters to indicate where they are defined. In the interior domain Ω^I we assume ϵ and μ to be piecewise constant. The permeability μ is regarded as positive and real, while ϵ can be complex valued with a positive real part. The imaginary part of ϵ models absorption and emission of the material, respectively. Above and below the grating ϵ and μ are constant with the values ϵ^+, μ^+ and ϵ^-, μ^-. In Ω^+ the permittivity is absorption free and therefore real.

9.1.2 The Rayleigh expansion

In absence of currents and free charges, we have to solve the vector valued wave equation

$$\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl}\boldsymbol{E}\right) - \omega^2\epsilon\boldsymbol{E} = 0 \tag{9.1}$$

in order to compute the electric field.

Biperiodic gratings

Above and below the grating, where ϵ and μ are constant, this equation reduces to

$$\operatorname{curl}\big(\operatorname{curl}\boldsymbol{E}\big) - (k^{\pm})^2\boldsymbol{E} = 0 \qquad \text{in } \Omega^{\pm}.$$

The absolute value of the wave vector $k^{\pm} = \omega\nu^{\pm}$ includes the refractive index $\nu^{\pm} = \sqrt{\epsilon^{\pm}\mu^{\pm}}$. Note that the square root is chosen such that $\mathrm{Im}(\nu^{\pm}) \geq 0$ and that for $\mathrm{Im}(\nu^{\pm}) = 0$ the real part $\mathrm{Re}(\nu^{\pm})$ is positive. The fundamental solution of this equation is

$$\boldsymbol{E}(\boldsymbol{x}) = \boldsymbol{A}e^{i(\alpha x+\beta y+\gamma z)}$$

with the restriction that

$$\alpha^2 + \beta^2 + \gamma^2 = (k^{\pm})^2 \qquad \text{with } \alpha, \beta, \gamma \in \mathbb{C}.$$

The parameters α, β, γ can be interpreted as the components of a wave vector $\boldsymbol{k}^{\pm}$ who indicates the direction of the plane wave of constant amplitude $\boldsymbol{A}$. The divergence freeness of the electric field leads to $\boldsymbol{k}^{\pm} \cdot A = 0$. Thus, the amplitude is perpendicular to the direction of propagation.

We assume an incoming electric field with the shape of a plane wave and a wave vector

$$\boldsymbol{k}^{+} = (\alpha_0^{+}, \beta_0^{+}, \gamma_0^{+})^{\top} = k^{+}\big(\sin\theta\cos\phi, \sin\theta\sin\phi, -\cos\theta\big)^{\top}. \tag{9.2}$$

Here $\theta \in [0, \frac{\pi}{2}]$ and $\phi \in [0, 2\pi]$ are the angles of incidence. According to the theory of Bloch-Floquet [Flo83, Blo28, Kuc01] the total electric field is quasi periodic, i.e.

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{x} + d_x\boldsymbol{e}_x) &= \rho_x\boldsymbol{E}(\boldsymbol{x}), \\ \boldsymbol{E}(\boldsymbol{x} + d_y\boldsymbol{e}_y) &= \rho_y\boldsymbol{E}(\boldsymbol{x}) \end{aligned}$$

with the canonical basis vectors $\boldsymbol{e}_x, \boldsymbol{e}_y$ and the factors ρ_x, ρ_y fixed by the incoming plane wave. Due to this quasi periodicity we get further restrictions for wave vectors of plane waves above and below the grating,

$$\alpha_n^{\pm} = \alpha_0^{+} + n\frac{2\pi}{d_x}, \tag{9.3}$$

$$\beta_m^{\pm} = \beta_0^{+} + m\frac{2\pi}{d_y}, \tag{9.4}$$

$$\gamma_{nm}^{\pm} = \sqrt{(k^{\pm})^2 - (\alpha_n^{\pm})^2 - (\beta_m^{\pm})^2}, \tag{9.5}$$

where $n, m \in \mathbb{Z}$. Note that the coefficients $\alpha_n^{\pm}$ and $\beta_m^{\pm}$ are real numbers if the material is not absorbing, but for higher orders n and m, $\gamma_{nm}^{\pm}$ is complex. To define the $\gamma_{nm}^{\pm}$ uniquely we take the value with $\mathrm{Re}(\gamma_{nm}^{\pm}) > 0$ if the real part is not zero, else we take γ with

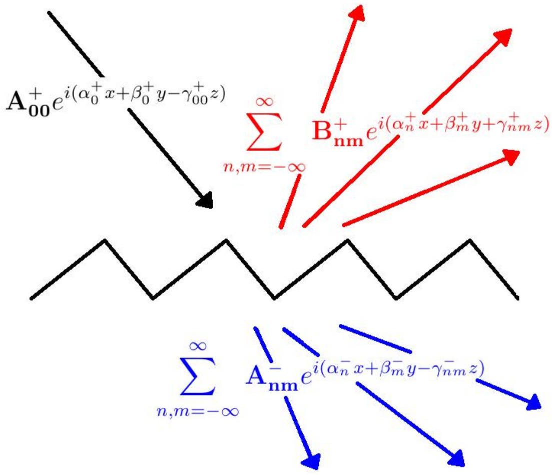
$\mathbf{A}_{\mathbf{00}}^{+} e^{i(\alpha_0^+ x+\beta_0^+ y-\gamma_{00}^+ z)}$
$\sum_{n,m=-\infty}^{\infty} \mathbf{B}_{\mathbf{nm}}^{+} e^{i(\alpha_n^+ x+\beta_m^+ y+\gamma_{nm}^+ z)}$
$\sum_{n,m=-\infty}^{\infty} \mathbf{A}_{\mathbf{nm}}^{-} e^{i(\alpha_n^- x+\beta_m^- y-\gamma_{nm}^- z)}$

Figure 9.2: The Rayleigh expansion

$\mathrm{Im}(\gamma_{nm}^{\pm}) > 0$. An imaginary part in $\gamma_{nm}^{\pm}$ results in an evanescent mode.

Collecting everything leads to the Rayleigh expansion of the electric field in the far field region $\Omega^{\pm}$

$$\boldsymbol{E}(\boldsymbol{x}) = \sum_{mn,=-\infty}^{\infty} \boldsymbol{A}_{nm}^{\pm} e^{i(\alpha_n^{\pm} x+\beta_m^{\pm} y-\gamma_{nm}^{\pm} z)} + \sum_{mn,=-\infty}^{\infty} \boldsymbol{B}_{nm}^{\pm} e^{i(\alpha_n^{\pm} x+\beta_m^{\pm} y+\gamma_{nm}^{\pm} z)} \tag{9.6}$$

with constant amplitudes $\boldsymbol{A}_{nm}^{\pm}$ and $\boldsymbol{B}_{nm}^{\pm}$. Note that by the first sum waves are described which are propagating into negative z-direction, and the second sum consists of waves propagating in positive z-direction. The only wave propagating towards the grating is the incident wave in Ω^{+} with wave vector $\boldsymbol{k}^{+}$ and a known amplitude $\boldsymbol{A}_{00}^{+}$. Thus, all coefficients $\boldsymbol{B}_{nm}^{-}$ and $\boldsymbol{A}_{nm}^{+}$ with the exception of $\boldsymbol{A}_{00}^{+}$ are zero (compare Figure 9.2). The coefficients $\boldsymbol{B}_{nm}^{+}$ and $\boldsymbol{A}_{nm}^{-}$ are unknown and need to be calculated. Summarizing this, the electric field in $\Omega^{\pm}$ is

$$\boldsymbol{E}(\boldsymbol{x}) = \begin{cases} \boldsymbol{E}_{in}(\boldsymbol{x}) + \boldsymbol{E}^{+}(\boldsymbol{x}) & \text{in } \Omega^{+} \\ \boldsymbol{E}^{-}(\boldsymbol{x}) & \text{in } \Omega^{-} \end{cases}$$

with

$$\begin{aligned}
\boldsymbol{E}_{in}(\boldsymbol{x}) &= \boldsymbol{A}^+_{00} e^{i(\alpha_0^+ x + \beta_0^+ y - \gamma_{00}^+ z)}, && (9.7)\\
\boldsymbol{E}^+(\boldsymbol{x}) &= \sum_{n,m=-\infty}^{\infty} \boldsymbol{B}^+_{nm} e^{i(\alpha_n^+ x + \beta_m^+ y + \gamma_{nm}^+ z)}, && (9.8)\\
\boldsymbol{E}^-(\boldsymbol{x}) &= \sum_{n,m=-\infty}^{\infty} \boldsymbol{A}^-_{nm} e^{i(\alpha_n^- x + \beta_m^- y - \gamma_{nm}^- z)}. && (9.9)
\end{aligned}$$

From (2.26) and (2.27) we obtain for the transmission coefficient $\mathscr{R}_{nm}$, defined as the fraction of incoming energy reflected into the mode of order n, m above the grating, and the transmission coefficient $\mathscr{T}_{nm}$, which describes the fraction of energy transmitted into the mode of order n, m below the grating,

$$\mathscr{R}_{nm} = \frac{|\boldsymbol{B}^+_{nm}|^2 \mathrm{Re}(\overline{\gamma}^+_{nm})}{|\boldsymbol{A}^+_{00}|^2 \mathrm{Re}(\overline{\gamma}^+_{00})} \qquad \text{and} \qquad \mathscr{T}_{nm} = \frac{|\boldsymbol{A}^-_{nm}|^2 \mu^+ \mathrm{Re}(\overline{\gamma}^-_{nm})}{|\boldsymbol{A}^+_{00}|^2 \mu^- \mathrm{Re}(\overline{\gamma}^+_{00})}.$$

One dimensional gratings

For the special case of a one dimensional grating periodic in x-direction with an incoming wave propagating in the xz-plane, i.e. the wave vector is given by

$$\boldsymbol{k}^+ = (\alpha_0^+, 0, \gamma_0^+)^\top = k^+(\sin\theta, 0 - \cos\theta)^\top,$$

the whole problem is invariant under translation into y-direction. Therefore, the electromagnetic fields decouple into TE and TM modes (compare section 2.1.4). Thus, we can consider instead of the vector valued wave equation (9.1) the Helmholtz equation

$$\operatorname{div}\left(\frac{1}{\tilde{\mu}} \operatorname{grad} u\right) + \omega^2 \tilde{\epsilon} u = 0,$$

where $u = E_y, \tilde{\epsilon} = \epsilon$ and $\tilde{\mu} = \mu$ in the TE-case, and $u = H_y, \tilde{\epsilon} = \mu$ and $\tilde{\mu} = \epsilon$ for TM modes. Following the same argumentation as for biperiodic gratings we obtain for the far field solution

$$u(\boldsymbol{x}) = \begin{cases} u_{in}(\boldsymbol{x}) + u^+(\boldsymbol{x}) & \text{in } \Omega^+, \\ u^-(\boldsymbol{x}) & \text{in } \Omega^- \end{cases}$$

with

$$
\begin{aligned}
u_{in}(\boldsymbol{x}) &= A_0^+ e^{i(\alpha_0^+ x - \gamma_0^+ z)}, && (9.10)\\
u^+(\boldsymbol{x}) &= \sum_{n=-\infty}^{\infty} B_n^+ e^{i(\alpha_n^+ x + \beta_n^+ z)}, && (9.11)\\
u_-(\boldsymbol{x}) &= \sum_{n=-\infty}^{\infty} A_n^- e^{i(\alpha_n^- x - \gamma_n^- z)}. && (9.12)
\end{aligned}
$$

Note that scalar amplitude A_0^+ of the incoming wave is known and

$$\alpha_n^\pm = \alpha_0^+ + n\frac{2\pi}{d_x} \quad \text{and} \quad \gamma_n^\pm = \sqrt{(k^\pm)^2 - (\alpha_n^\pm)^2} \qquad \text{for } n \in \mathbb{Z}.$$

Like in the biperiodic case reflection and transmission coefficients can be computed via

$$\mathscr{R}_n = \frac{|B_n^+|^2 \mathrm{Re}(\overline{\gamma}_n^+)}{|A_0^+|^2 \mathrm{Re}(\overline{\gamma}_0^+)} \qquad \text{and} \qquad \mathscr{T}_n = q\,\frac{|A_n^-|^2 \mu^+ \mathrm{Re}(\overline{\gamma}_n^-)}{|A_0^+|^2 \mu^- \mathrm{Re}(\overline{\gamma}_0^+)},$$

where $q = 1$ in the TE case and for TM modes $q = \frac{\epsilon^+}{\epsilon^-}$.

9.2 Problem formulations and discretizations for one dimensional gratings

In the last section we stated a plane wave ansatz for the solution of the Helmholtz equation in the far field domain $\Omega^\pm$ with the unknown coefficients A_n^-, B_m^+, respectively. These coefficients can be calculated by solving the wave equation in the near field domain Ω^I together with coupling conditions at the interface to the exterior domains.

9.2.1 The classical formulation

According to the theory of Bloch Floquet the solution has to be quasi periodic, i.e. $u(\boldsymbol{x} + d_x \boldsymbol{e}_x) = \rho_x u(\boldsymbol{x})$, where the factor ρ_x is defined by the quasi periodicity of the incoming wave. Thus, ρ_x equals $e^{i\alpha_0^+ d_x}$. The fact that the normal component of the flux field has to fulfill the same periodic constraint leads to the quasi periodic boundary conditions

$$u^I(\boldsymbol{x} + d_x \boldsymbol{e}_x) = \rho_x u^I(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Gamma^l_x \tag{9.13}$$

$$\frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma^l_x} \cdot \operatorname{grad} u^I(\boldsymbol{x} + d_x \boldsymbol{e}_x) = \rho_x \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma^l_x} \cdot \operatorname{grad} u^I(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Gamma^l_x, \tag{9.14}$$

where $\boldsymbol{n}_{\Gamma^l_x}$ is the outer normal vector on Γ^l_x, and the superscript I implies that u is the solution in Ω^I_p.

On the interfaces the solution on Ω^I_p has to match the solution in $\Omega^\pm_p$, and their fluxes have to be normal continuous, i.e.

$$u_I(\boldsymbol{x}) = u_{in}(\boldsymbol{x}) + u^+(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Gamma^+, \tag{9.15}$$

$$\frac{1}{\tilde{\mu}^+} \boldsymbol{n}_{\Gamma^+} \cdot \operatorname{grad} u^I(\boldsymbol{x}) = \frac{1}{\tilde{\mu}^+} \boldsymbol{n}_{\Gamma^+} \cdot \operatorname{grad} \big(u_{in}(\boldsymbol{x}) + u^+(\boldsymbol{x})\big) \qquad \forall \boldsymbol{x} \in \Gamma^+, \tag{9.16}$$

$$u_I(\boldsymbol{x}) = u^-(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Gamma^- \tag{9.17}$$

$$\frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma^-} \cdot \operatorname{grad} u^I(\boldsymbol{x}) = \frac{1}{\tilde{\mu}^-} \boldsymbol{n}_{\Gamma^-} \cdot \operatorname{grad} u^-(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Gamma^-. \tag{9.18}$$

Summarizing this, we end up with the classical problem formulation.

Formulation 9.1 (classical formulation)**.** *Find* $u : \Omega^I_p \to \mathbb{C}$*, such that*

$$\operatorname{div}\left(\frac{1}{\tilde{\mu}} \operatorname{grad} u^I\right) + \omega^2 \epsilon u^I = 0 \qquad \textit{in } \Omega^I_p \tag{9.19}$$

together with the quasi periodic boundary conditions (9.13) and (9.14) on Γ^l_x *and* Γ^r_x*, the interface conditions (9.15) - (9.18) on* $\Gamma^\pm$ *and the far field ansatz (9.10) - (9.12) is fulfilled.*

9.2.2 The weak formulation

Now, we are able to state the variational formulation for the near field domain Ω^I_p. Taking the scalar Helmholtz equation, multiplying it with a test function, integrating it over the computational domain, and integrating by parts, we get the variational formulation

$$B^I_1(u^I, v) - \big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\partial\Omega^I_p} \cdot \operatorname{grad} u^I, \overline{v} \big\rangle_{\partial\Omega^I_p} = 0 \qquad \forall v \in H^1(\Omega^I_p),$$

where $u^I \in H^1(\Omega_p^I)$. The bilinear form B_1^I is given by

$$B_1^I(u,v) := \big(\frac{1}{\tilde{\mu}}\operatorname{grad} u, \operatorname{grad}\overline{v}\big)_{\Omega_p^I} - \omega^2\tilde{\epsilon}\big(u,\overline{v}\big)_{\Omega_p^I}.$$

Note that we still have to take care of the quasi periodic boundary conditions (9.13) and (9.14) and the interface conditions (9.15) - (9.18).

9.2.3 The incorporation of quasi periodic boundary conditions

The next step is to incorporate the quasi periodic boundary conditions (9.13) and (9.14). Therefore, we restrict the test function space $H^1(\Omega_p^I)$ to the quasi periodic space

$$Q_p^I := \Big\{ v \in H^1(\Omega_p^I) \ : \ v(\boldsymbol{x} + d_x \boldsymbol{e}_x) = \rho_x v(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \Gamma_x^l \Big\}.$$

Note that the quasi periodic constraint (9.13) is an essential one, and it is directly incorporated into the space. We should mention that a conforming finite element discretization of the space Q_p^I requires matching meshes on the two periodic boundaries. The quasi periodicity of the flux (9.14) is, as we will see in the following lemma, included in the weak sense. Thus, it is a natural constraint.

Lemma 9.2. *Any function $u^I \in Q_p^I$ satisfying the variational equation*

$$B_1^I(u^I,v) - \big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma^\pm} \cdot \operatorname{grad} u^I, \overline{v} \big\rangle_{\Gamma^\pm} \qquad \forall v \in Q_p^I$$

fulfills the Helmholtz equation (9.19) together with the boundary conditions (9.13) and (9.14).

Proof. The condition (9.13) is fulfilled by the definition of the periodic space Q_p^I. Integration by parts of the first term in $B_1^I(u^I,v)$ yields

$$\big(- \operatorname{div}\left(\frac{1}{\tilde{\mu}}\operatorname{grad} u^I\right) - \omega^2\tilde{\epsilon}u^I, \overline{v}\big)_{\Omega_p^I} + \big\langle \frac{1}{\tilde{\mu}}\boldsymbol{n}_\Gamma \cdot \operatorname{grad} u^I, \overline{v} \big\rangle_{\Gamma_x^l \cup \Gamma_x^r} = 0 \qquad \forall v \in Q_p^I,$$

where $\boldsymbol{n}_\Gamma$ is the outer normal on Γ_x^l and Γ_x^r, respectively. Choosing test functions with $v|_{\Gamma_x^l \cup \Gamma_x^r} = 0$ directly leads to

$$-\operatorname{div}\left(\frac{1}{\tilde{\mu}}\operatorname{grad} u^I\right) - \omega^2\tilde{\epsilon}u^I = 0 \qquad \text{on } \Omega_p^I.$$

By transforming the integral over Γ_x^r to an integral over Γ_x^l and by using the quasi periodicity of the test function together with $\boldsymbol{n}_{\Gamma_x^l} = -\boldsymbol{n}_{\Gamma_x^r}$ we obtain

$$\begin{aligned} 0 &= \Big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma_x^l} \cdot \operatorname{grad} u^I, \overline{v} \Big\rangle_{\Gamma_x^l} + \Big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma_x^r} \cdot \operatorname{grad} u^I(\boldsymbol{x} + d_x \boldsymbol{e}_x), \overline{v}(\boldsymbol{x} + d_x \boldsymbol{e}_x) \Big\rangle_{\Gamma_x^l} \\ &= \Big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma_x^l} \cdot \operatorname{grad} u^I, \overline{v} \Big\rangle_{\Gamma_x^l} - \Big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma_x^l} \cdot \operatorname{grad} u^I(\boldsymbol{x} + d_x \boldsymbol{e}_x), \overline{\rho}_x \overline{v}(\boldsymbol{x}) \Big\rangle_{\Gamma_x^l} \\ &= \Big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma_x^l} \cdot \operatorname{grad} u^I(\boldsymbol{x}) - \overline{\rho}_x \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma_x^l} \cdot \operatorname{grad} u^I(\boldsymbol{x} + d_x \boldsymbol{e}_x) . \overline{v} \Big\rangle_{\Gamma_x^l}, \end{aligned}$$

This results with the help of $\overline{\rho}_x = \frac{1}{\rho_x}$ in condition (9.14). □

Remark 9.3. *The term*

$$\Big\langle \frac{1}{\tilde{\mu}} \boldsymbol{n}_{\Gamma^\pm} \cdot \operatorname{grad} u^I, \overline{v} \Big\rangle_{\Gamma^\pm}$$

requires Neumann boundary data $\mu^{-1} \boldsymbol{n}_{\Gamma^\pm} \cdot \operatorname{grad} u^I$. Integral equation methods apply here the Dirichlet to Neumann (DtN) map $\mathcal{T}$ to replace the boundary term by

$$\big\langle \mathcal{T}(u^I), \overline{v} \big\rangle_{\Gamma^\pm}.$$

In [EHS02, Sch04] this DtN operator is realized efficiently on uniform grids by the fast Fourier transform. Our approach uses the DtN operator provided by the Rayleigh expansion.

9.2.4 Incorporation of the interface conditions by Nitsche's method

Finally, the interface conditions have to be included into the variational formulation. The outgoing plane waves u^+ and u^- together with the known incoming plane wave u_{in} have to be coupled to the polynomial solution u^I in the interior domain.

For notational reasons we define the wave vectors of outgoing waves in Ω^+ as $\boldsymbol{k}_n^+ := (\alpha_n^+, \gamma_n^+)^\top$, and the wave vectors of outgoing waves in Ω^- we call $\boldsymbol{k}_n^- := (\alpha_n^-, \gamma_n^-)^\top$. The wave vector of the incoming wave we denote by $\boldsymbol{k}^+$.

Upon this, the spaces wherein we search for the solutions u^+ and u^- are defined as

$$\begin{aligned} Q^+ &:= \Big\{ v \,:\, v = \sum_{n=-\infty}^{\infty} B_n^+ e^{i\boldsymbol{k}_n^+ \boldsymbol{x}},\ B_n^+ \in \mathbb{C} \Big\}, \\ Q^- &:= \Big\{ v \,:\, v = \sum_{n=-\infty}^{\infty} A_n^- e^{i\boldsymbol{k}_n^- \boldsymbol{x}},\ A_n^- \in \mathbb{C} \Big\}. \end{aligned}$$

Using the Rayleigh expansion, we can construct for any function $v^+ \in Q^+$,

$$v^+ = \sum_{n=-\infty}^{\infty} B_n^+ e^{i\boldsymbol{k}_n^+ \boldsymbol{x}},$$

the DtN operator $\mathcal{T}$ as

$$\mathcal{T}(u^+) := \frac{i}{\tilde{\mu}^+} \sum_{n=-\infty}^{\infty} (\boldsymbol{n}_{\Gamma^+} \cdot \boldsymbol{k}_n^+) B_n^+ e^{i\boldsymbol{k}_n^+ \boldsymbol{x}}.$$

For $v^- \in Q^-$ and u_{in} the operator can be determined similarly. Furthermore, let $[u]_\pm$ be the jump of u at the interface $\Gamma^\pm$, i.e. $[u]_\pm := u^I - u^\pm$.

Incorporating the interface conditions (9.15)-(9.18) into our variational formulation with Nitsche's method [Nit71, Ste98, HHL03], or similarly with a discontinuous Galerkin method [ABCM02, CKS00], yields together with the fact that the incoming wave is known to our problem formulation.

Formulation 9.4 (weak formulation)**.** *Find $\boldsymbol{u} := (u^I, u^+, u^-) \in Q_p^I \times Q^+ \times Q^-$ such that*

$$B_1(\boldsymbol{u}, \boldsymbol{v}) = F_1(\boldsymbol{v}) \qquad \forall \boldsymbol{v} := (v^I, v^+, v^-) \in Q_p^I \times Q^+ \times Q^-$$

with the bilinear form

$$\begin{aligned} B_1(\boldsymbol{u}, \boldsymbol{v}) &:= B_1^I(u^I, v^I) + \big\langle \mathcal{T}(u^\pm), \overline{v}^\pm \big\rangle_{\Gamma^\pm} + \big\langle \mathcal{T}(u^\pm), [\overline{v}]_\pm \big\rangle_{\Gamma^\pm} \\ &\quad + \big\langle [u]_\pm, \mathcal{T}(\overline{v}^\pm) \big\rangle_{\Gamma^\pm} + \eta \big\langle [u]_\pm, [\overline{v}]_\pm \big\rangle_{\Gamma^\pm} \end{aligned}$$

and a properly chosen stabilizing parameter η. The linear form containing the incident wave $u_{in} = A_0^+ e^{i\boldsymbol{k}^+ \boldsymbol{x}}$ reads as

$$F_1(\boldsymbol{v}) = -\big\langle \mathcal{T}(u_{in}), v^I \big\rangle_{\Gamma^+} + \big\langle u_{in}, \mathcal{T}(v^+) \big\rangle_{\Gamma^+} + \eta \big\langle u_{in}, [v]_+ \big\rangle_{\Gamma^+}.$$

Next we state consistency of the weak problem formulation.

Lemma 9.5. *The exact solution* $\boldsymbol{u} = (u^I, u^+, u^-)$ *of the classical Formulation 9.1 is a solution of Formulation 9.4.*

Proof. Inserting the exact solution (u^I, u^+, u^-) into $B_1^I(\boldsymbol{u}, \boldsymbol{v})$ results together with the interface conditions (9.15) and (9.17), i.e. $[u]_+ = u_{in}$ on Γ^+ and $[u]_- = 0$ on Γ_-, in

$$\begin{aligned} B_1(\boldsymbol{u}, \boldsymbol{v}) - F_1(\boldsymbol{v}) &= B_1^I(u^I, v^I) + \big\langle \mathcal{T}(u^\pm), \overline{v}^\pm \big\rangle_{\Gamma^\pm} + \big\langle \mathcal{T}(u^\pm), [\overline{v}]_\pm \big\rangle_{\Gamma^\pm} \\ &\quad + \big\langle \mathcal{T}(u_{in}), v^I \big\rangle_{\Gamma^+}, \\ B_1(\boldsymbol{u}, \boldsymbol{v}) - F_1(\boldsymbol{v}) &= B_1^I(u^I, v^I) + \big\langle \mathcal{T}(u^+ + u_{in}), v^I \big\rangle_{\Gamma^+} + \big\langle \mathcal{T}(u^-), v^I \big\rangle_{\Gamma^-}. \end{aligned}$$

Note that according to the definition of the operator $\mathcal{T}$ and the interface conditions (9.16) and (9.18) $\mathcal{T}(u^+ + u_{in}) = \tilde{\mu}^{-1} \boldsymbol{n}_{\Gamma^+} \cdot \operatorname{grad} u^I$ and $\mathcal{T}(u^-) = \tilde{\mu}^{-1} \boldsymbol{n}_{\Gamma^-} \cdot \operatorname{grad} u^I$. Thus,

$$B_1(\boldsymbol{u}, \boldsymbol{v}) - F_1(\boldsymbol{v}) = B_1^I(u^I, v^I) + \big\langle \tilde{\mu}^{-1} \boldsymbol{n}_{\Gamma^\pm} \cdot \operatorname{grad} u^I, v^I \big\rangle_{\Gamma^\pm}$$

which is zero according to Lemma 9.2. □

Remark 9.6. *(i) In the method described in Formulation 9.4 the Rayleigh coefficients* B_n^+ *and* A_n^- *appear as additional unknowns in the variational equation.*

(ii) For elliptic problems it is well known that the stabilizing parameter η *has to be chosen sufficiently large, i.e.* $\eta \approx \frac{1}{h}$, *to obtain an elliptic bilinear form on the finite element space. For our indefinite problem we made good experience by taking* $\eta = i\omega$, *which represents an impedance boundary condition for the near field domain.*

(iii) Using this formulation, the gradient of the plane waves $\mathcal{T}(u^\pm)$ *has to be evaluated instead of the gradient of the interior solution* $\boldsymbol{n}_{\Gamma^\pm} \cdot \operatorname{grad} u^I$ *when assembling the system matrix. Because* $\mathcal{T}(u^\pm)$ *is equivalent to the multiplication of the inner product of normal and wave vector with a plane wave, an evaluation of the DtN map is equivalent to a function evaluation and therefore cheaper.*

(iv) Except to the term $\big\langle \mathcal{T}(u^\pm), v^\pm \big\rangle_{\Gamma^\pm}$ *the bilinear form is hermitian. This integral of two plane waves over the interface leads because of*

$$\int_0^{d_x} e^{i(\alpha_0 + n\frac{2\pi}{d_x})x} e^{-i(\alpha_0 + m\frac{2\pi}{d_x})x} dx = \delta_{nm} d_x$$

just to diagonal entries in the system matrix. Thus, the stiffness matrix can be decomposed into a hermitian and a diagonal matrix.

9.3 Problem formulations and discretizations for biperiodic gratings

In this section we provide the main points of the corresponding problem formulation for the biperiodic grating. For a more detailed discussion see [HSSZ09].

For biperiodic gratings, the electric field $\boldsymbol{E}$, which we call $\boldsymbol{E}^I$ in Ω_p^I, and the magnetic field, i.e. $(i\omega\mu)^{-1}\operatorname{curl}\boldsymbol{E}$ have to be tangential continuous across interfaces. Furthermore, on the periodic boundaries $\Gamma_x^l, \Gamma_x^r, \Gamma_y^l, \Gamma_y^r$ the tangential components of these two fields have to be quasi periodic with factors $\rho_x = e^{i\alpha_0^+ d_x}$ and $\rho_y = e^{i\beta_0^+ d_y}$ obtained by the incoming wave (9.7). This leads us to

Formulation 9.7 (the classical formulation)**.** *Find a vector valued function* $\boldsymbol{E}^I : \Omega_p^I \to \mathbb{C}^3$ *which satisfies*

$$\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl}\boldsymbol{E}^I\right) - \omega^2\epsilon\boldsymbol{E}^I = 0 \tag{9.20}$$

under the quasi periodic boundary conditions

$$\begin{aligned}
\boldsymbol{n}_{\Gamma_x^l} \times \boldsymbol{E}^I(\boldsymbol{x} + d_x\boldsymbol{e}_x) &= \rho_x \boldsymbol{n}_{\Gamma_x^l} \times \boldsymbol{E}^I(\boldsymbol{x}) && \forall \boldsymbol{x} \in \Gamma_x^l, && (9.21)\\
\frac{1}{\mu}\boldsymbol{n}_{\Gamma_x^l} \times \operatorname{curl}\boldsymbol{E}^I(\boldsymbol{x} + d_x\boldsymbol{e}_x) &= \rho_x \frac{1}{\mu}\boldsymbol{n}_{\Gamma_x^l} \times \operatorname{curl}\boldsymbol{E}^I(\boldsymbol{x}) && \forall \boldsymbol{x} \in \Gamma_x^l, && (9.22)\\
\boldsymbol{n}_{\Gamma_y^l} \times \boldsymbol{E}^I(\boldsymbol{x} + d_y\boldsymbol{e}_y) &= \rho_y \boldsymbol{n}_{\Gamma_y^l} \times \boldsymbol{E}^I(\boldsymbol{x}) && \forall \boldsymbol{x} \in \Gamma_y^l, && (9.23)\\
\frac{1}{\mu}\boldsymbol{n}_{\Gamma_y^l} \times \operatorname{curl}\boldsymbol{E}^I(\boldsymbol{x} + d_y\boldsymbol{e}_y) &= \rho_y \frac{1}{\mu}\boldsymbol{n}_{\Gamma_y^l} \times \operatorname{curl}\boldsymbol{E}^I(\boldsymbol{x}) && \forall \boldsymbol{x} \in \Gamma_y^l, && (9.24)
\end{aligned}$$

and the interface conditions

$$\begin{aligned}
\boldsymbol{n}_{\Gamma^+} \times \boldsymbol{E}^I &= \boldsymbol{n}_{\Gamma^+} \times (\boldsymbol{E}_{in} + \boldsymbol{E}^+) && \textit{on } \Gamma^+, && (9.25)\\
\frac{1}{\mu}\boldsymbol{n}_{\Gamma^+} \times \operatorname{curl}\boldsymbol{E}^I &= \frac{1}{\mu^+}\boldsymbol{n}_{\Gamma^+} \times \operatorname{curl}(\boldsymbol{E}_{in} + \boldsymbol{E}^+) && \textit{on } \Gamma^+, && (9.26)\\
\boldsymbol{n}_{\Gamma^-} \times \boldsymbol{E}^I &= \boldsymbol{n}_{\Gamma^-} \times \boldsymbol{E}^- && \textit{on } \Gamma^+, && (9.27)\\
\frac{1}{\mu}\boldsymbol{n}_{\Gamma^-} \times \operatorname{curl}\boldsymbol{E}^I &= \frac{1}{\mu^-}\boldsymbol{n}_{\Gamma^-} \times \operatorname{curl}\boldsymbol{E}^- && \textit{on } \Gamma^+, && (9.28)
\end{aligned}$$

together with the ansatz (9.7) -(9.9).

We get by testing (9.20) with a vector valued test function and carrying out integration by parts

$$B_2^I(\boldsymbol{E}^I, \boldsymbol{e}^I) - \big\langle \frac{1}{\mu} \operatorname{curl} \boldsymbol{E}^I, \boldsymbol{n}_{\partial\Omega_p^I} \times \boldsymbol{e}^I \big\rangle_{\partial\Omega_p^I} = 0 \qquad \forall \boldsymbol{e}^I \in H(\operatorname{curl}, \Omega_p^I),$$

with $\boldsymbol{E}^I \in H(\operatorname{curl}, \Omega_p^I)$ and the bilinear form for the internal domain Ω_p^I

$$B_2^I(\boldsymbol{E}^I, \boldsymbol{e}^I) := \big(\frac{1}{\mu} \operatorname{curl} \boldsymbol{E}^I, \operatorname{curl} \boldsymbol{e}^I\big)_{\Omega_p^I} - \big(\omega^2 \epsilon \boldsymbol{E}^I, \boldsymbol{e}^I\big)_{\Omega_p^I}.$$

The quasi periodic boundary conditions can be again implemented by incorporating the conditions (9.21) and (9.23) directly into the function space, i.e. restricting $H(\operatorname{curl}, \Omega_p^I)$ to

$$Y_p^I := \Big\{ \boldsymbol{v} \in H(\operatorname{curl}, \Omega_p^I) \ : \ \rho_x \boldsymbol{n}_{\Gamma_x^l} \times \boldsymbol{v}|_{\Gamma_x^l} = \boldsymbol{n}_{\Gamma_x^l} \times \boldsymbol{v}|_{\Gamma_x^r}, \quad \rho_y \boldsymbol{n}_{\Gamma_y^l} \times \boldsymbol{v}|_{\Gamma_y^l} = \boldsymbol{n}_{\Gamma_y^l} \times \boldsymbol{v}|_{\Gamma_y^r} \Big\},$$

and by neglecting the boundary integral on the periodic boundaries.

Lemma 9.8. *Any function $\boldsymbol{E}^I \in Y_p^I$ satisfying the variational equation*

$$B_2^I(\boldsymbol{E}^I, \boldsymbol{e}^I) - \big\langle \frac{1}{\mu} \operatorname{curl} \boldsymbol{E}^I, \boldsymbol{n}_{\Gamma^\pm} \times \boldsymbol{e}^I \big\rangle_{\Gamma^\pm} = 0 \qquad \forall \boldsymbol{e}^I \in Y_p^I$$

fulfills the differential equation (9.20) together with the quasi periodic boundary conditions (9.25) -(9.28).

The proof of this lemma is similar to the proof of Lemma 9.2.

In order to incorporate the interface conditions, we have to define the plane wave spaces and to introduce the DtN map $\boldsymbol{\mathcal{T}}$. With the wave vectors $\boldsymbol{k}_{nm}^+ := (\alpha_n^+, \beta_m^+, \gamma_{nm}^+)$, $\boldsymbol{k}_{nm}^- := (\alpha_n^-, \beta_m^-, \gamma_{nm}^-)$ and $\boldsymbol{k}^+$ for the incoming wave, the plane wave spaces on Ω^+ and Ω^- read as

$$\begin{aligned} Y^+ &:= \Big\{ \boldsymbol{v} \ : \ \boldsymbol{v} = \sum_{n,m=-\infty}^{\infty} \boldsymbol{B}_{nm}^+ e^{i\boldsymbol{k}_{nm}^+ \boldsymbol{x}}, \quad \boldsymbol{B}_{nm}^+ \in \mathbb{C}^3, \quad \boldsymbol{B}_{nm}^+ \cdot \boldsymbol{k}_{nm}^+ = 0 \Big\}, \\ Y^- &:= \Big\{ \boldsymbol{v} \ : \ \boldsymbol{v} = \sum_{n,m=-\infty}^{\infty} \boldsymbol{A}_{nm}^- e^{i\boldsymbol{k}_{nm}^- \boldsymbol{x}}, \quad \boldsymbol{A}_{nm}^- \in \mathbb{C}^3, \quad \boldsymbol{A}_{nm}^- \cdot \boldsymbol{k}_{nm}^- = 0 \Big\}. \end{aligned}$$

For any $\boldsymbol{v}^+ \in Y^+$,

$$\boldsymbol{v}^+ = \sum_{n,m=-\infty}^{\infty} \boldsymbol{B}_{nm}^+ e^{i\boldsymbol{k}_{nm}^+ \boldsymbol{x}},$$

the DtN map is defined via

$$\boldsymbol{\mathcal{T}}(\boldsymbol{v}^+) = \frac{i}{\mu^+} \sum_{n,m=-\infty}^{\infty} \boldsymbol{k}^+_{nm} \times \boldsymbol{B}^+_{nm} e^{i\boldsymbol{k}^+_{nm}\boldsymbol{x}}.$$

Similar definitions exist for functions in Y^- and $\boldsymbol{E}_{in}$. Additionally, we make use of the tangential jump $[\boldsymbol{v}]_\tau := \boldsymbol{n}_{\Gamma^\pm} \times \boldsymbol{v}^I - \boldsymbol{n}_{\Gamma^\pm} \times \boldsymbol{v}^\pm$. This leads us to the problem formulation with the interface conditions incorporated by Nitsche's method.

Formulation 9.9 (the weak formulation)**.** *Find* $\boldsymbol{u} := (\boldsymbol{E}^I, \boldsymbol{E}^+, \boldsymbol{E}^-) \in Y^I_p \times Y^+ \times Y^-$ *such that*

$$B_2(\boldsymbol{u}, \boldsymbol{v}) = F_2(\boldsymbol{v}) \qquad \forall \boldsymbol{v} := (\boldsymbol{e}^I, \boldsymbol{e}^+, \boldsymbol{e}^-) \in Y^I_p \times Y^+ \times Y^-$$

with the bilinear form

$$\begin{aligned} B_2(\boldsymbol{u}, \boldsymbol{v}) \ := \ & B^I_2(\boldsymbol{E}^I, \boldsymbol{e}^I) - \big\langle \boldsymbol{\mathcal{T}}(\boldsymbol{E}^\pm), \boldsymbol{n}_{\Gamma^\pm} \times \boldsymbol{e}^\pm \big\rangle_{\Gamma^\pm} - \big\langle \boldsymbol{\mathcal{T}}(\boldsymbol{E}^\pm), [\boldsymbol{e}]_\tau \big\rangle_{\Gamma^\pm} \\ & - \big\langle [\boldsymbol{E}]_\tau, \boldsymbol{\mathcal{T}}(\boldsymbol{e}^\pm) \big\rangle_{\Gamma^\pm} + \eta \big\langle [\boldsymbol{E}]_\tau, [\boldsymbol{e}]_\tau \big\rangle_{\Gamma^\pm} \end{aligned}$$

with a properly chosen stabilizing parameter η. *The linear form reads with* $\boldsymbol{E}_{in} = \boldsymbol{A}^+_{00} e^{i\boldsymbol{k}^+\boldsymbol{x}}$ *as*

$$F_2(\boldsymbol{v}) := \big\langle \boldsymbol{\mathcal{T}}(\boldsymbol{E}_{in}), \boldsymbol{n}_{\Gamma^+} \times \boldsymbol{e}^I \big\rangle_{\Gamma^+} - \big\langle \boldsymbol{n}_{\Gamma^+} \times \boldsymbol{E}_{in}, \boldsymbol{\mathcal{T}}(\boldsymbol{e}^+) \big\rangle_{\Gamma^+} + \eta \big\langle \boldsymbol{n}_{\Gamma^+} \times \boldsymbol{E}_{in}, [\boldsymbol{e}]_\tau \big\rangle_{\Gamma^+}.$$

For this formulation consistency can be shown in the same way as in the proof of Lemma 9.5. We should also mention that Remark 9.6 applies for this problem formulation as well.

9.4 Modeling of layers

In many practical application are often planar layers between the semiinfinite substrate and the periodic structure. The simplest way to take these layers into account is to add them to the computational domain Ω^I_p. This can result in a large increase of the number of unknowns in the case of many and (or) thick substrate layers, which is not desirable. A different possibility to model such layers is to add them via the boundary condition, i.e. they are part of the semiinfinite domain Ω^-_p. In this case, we have to allow for waves entering the computational domain Ω^I_p from below (B_F in Figure 9.3), caused by reflections of the outgoing waves on Γ^- (A_F in Figure 9.3). The aim of this section is, to express the Rayleigh coefficients of the incoming waves on Γ_- by the outgoing ones via the transfer

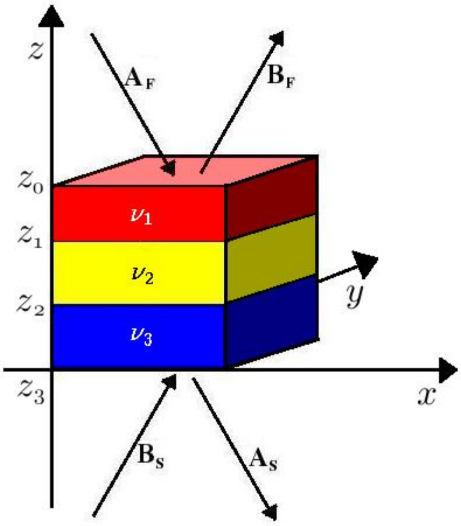

Figure 9.3: A stack of substrate layers

matrix method [NP03, Rei05]. By doing this, the substrate layers can be incorporated directly into the function spaces.

In the following, we assume that the layers are parallel to the xy plane (compare Figure 9.3), and the j-th layer with thickness $d_j = z_{j-1} - z_j$ is enclosed by the planes $z = z_{j-1}$ and $z = z_j$ with $z_{j-1} > z_j$. Note that Γ_- is the plane $z = z_0$. We consider μ to be constant on the whole domain and ϵ and consequently ν are constant at least on each substrate layer with values ϵ_j, ν_j.

Because the polarization of a wave does not change when it is reflected, it makes sense to decompose the field into a TE and a TM component. For the described setting the solution in one layer can be expressed as in (9.6) via the Rayleigh expansion. Knowing that in any layer the wave of order i,j is reflected on the interfaces either into the wave moving in opposite z-direction of the same order, or it is transmitted to the order i, j wave of the neighboring layer, it is sufficient to consider each order individually. Thus, we can restrict ourself to one up and one down going wave.

9.4.1 TE-modes

For TE modes the electric field is parallel to the layers, and we will denote it in the j-th layer as

$$E_{\|}(x, y, z) = A_j e^{i(\alpha_j x + \beta_j y - \gamma_j z)} + B_j e^{i(\alpha_j x + \beta_j y + \gamma_j z)}.$$

With the help of this representation the electric field and its normal flux at the boundaries of the layer can be described by

$$\left.\begin{pmatrix} E_{\|} \\ \frac{\partial E_{\|}}{\partial z} \end{pmatrix}\right|_{z=z_{j-1}} = \begin{pmatrix} 1 & 1 \\ -i\gamma_j & i\gamma_j \end{pmatrix} \begin{pmatrix} A_j e^{i(\alpha_j x+\beta_j y-\gamma_j z_{j-1})} \\ B_j e^{i(\alpha_j x+\beta_j y+\gamma_j z_{j-1})} \end{pmatrix} \tag{9.29}$$

$$\left.\begin{pmatrix} E_{\|} \\ \frac{\partial E_{\|}}{\partial z} \end{pmatrix}\right|_{z=z_j} = \begin{pmatrix} 1 & 1 \\ -i\gamma_j & i\gamma_j \end{pmatrix} \begin{pmatrix} \kappa & 0 \\ 0 & \frac{1}{\kappa} \end{pmatrix} \begin{pmatrix} A_j e^{i(\alpha_j x+\beta_j y-\gamma_j z_{j-1})} \\ B_j e^{i(\alpha_j x+\beta_j y+\gamma_j z_{j-1})} \end{pmatrix} \tag{9.30}$$

with $\kappa := e^{i\gamma_j d_j}$. From this equations we get the propagation matrix for layer j

$$P_j := \begin{pmatrix} 1 & 1 \\ -i\gamma_j & i\gamma_j \end{pmatrix} \begin{pmatrix} \kappa & 0 \\ 0 & \frac{1}{\kappa} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -i\gamma_j & i\gamma_j \end{pmatrix}^{-1}$$

which connects the field and its normal flux above and below the layer,

$$\left.\begin{pmatrix} E_{\|} \\ \frac{\partial E_{\|}}{\partial z} \end{pmatrix}\right|_{z_j} = P_j \left.\begin{pmatrix} E_{\|} \\ \frac{\partial E_{\|}}{\partial z} \end{pmatrix}\right|_{z_{j-1}} .$$

Considering a stack of m layers with interfaces at $z_0, z_1, \ldots, z_m$ and using that on interfaces the tangential electric field (here $E_{\|}$) and its normal flux are continuous, we get by simple matrix multiplication

$$\left.\begin{pmatrix} E_{\|} \\ \frac{\partial E_{\|}}{\partial z} \end{pmatrix}\right|_{z_0} = \prod_{j=1}^{m} P_j \left.\begin{pmatrix} E_{\|} \\ \frac{\partial E_{\|}}{\partial z} \end{pmatrix}\right|_{z_m} . \tag{9.31}$$

Combining this with (9.29) and(9.30), which connect the electric field with its Rayleigh coefficients, we obtain a system of the form

$$\begin{pmatrix} A_F \\ B_F \end{pmatrix} = M \begin{pmatrix} A_S \\ 0 \end{pmatrix} . \tag{9.32}$$

Note that because of no incoming waves from below the stack B_S was considered to be zero. Elimination of A_S leads to the desired relation between A_F and B_F, i.e. $B_F = \vartheta A_F = \frac{M_{21}}{M_{11}} A_F$.

9.4.2 TM-modes

For TM-modes we repeat this calculations for the magnetic field $\boldsymbol{H}$, which is parallel to the layers. Here the situation is almost similar. The only difference is that now the tangential component of the magnetic field $H_{\|}$ and $\frac{1}{\omega^2\nu_j^2}\frac{\partial H_{\|}}{\partial z}$ have to be continuous across interfaces. Thus, $\frac{H_{\|}}{\partial z}$ jumps on layer interfaces. To take care of these jumps, the propagation relation corresponding to (9.31) has to be modified by introducing an interface matrix

$$D_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\nu_i^2}{\nu_j^2} \end{pmatrix},$$

and we obtain

$$\left.\begin{pmatrix} H_{\|} \\ \frac{\partial H_{\|}}{\partial z} \end{pmatrix}\right|_{z_0} = P_1 D_{12} P_2 D_{23} \dots P_m \left.\begin{pmatrix} H_{\|} \\ \frac{\partial H_{\|}}{\partial z} \end{pmatrix}\right|_{z_m}.$$

This relation leads again to an equation of the form (9.32).

Remark 9.10. *Calculating the factor ϑ via $\frac{M_{21}}{M_{11}}$ as suggested above can lead in the case of absorbing layers, where one has to deal with large numbers in M, to a reduced accuracy in ϑ. One possibility to avoid this is to start with an arbitrary number for A_S in (9.32), for example $A_S = 1$, and to multiply the resulting vector with one propagation matrix after the other. Because for the calculation of ϑ just the ratio between A_F and B_F is needed, the vector can be normalized after each multiplication, and large numbers are avoided.*

Remark 9.11. *If we use plane wave basis functions in the space Y^- which are either TE or TM, the factor ϑ linking an up going degree of freedom with the corresponding down going degree of freedom can be calculated as just described. Substrate layers can be implemented then by replacing B_F by ϑA_F during the assembly procedure.*

9.5 Numerical examples

The numerical examples presented in this section were calculated with the finite element code Netgen/Ngsolve of Schöberl (see *http://sourceforge.net/projects/ngsolve* or [Sch97]), and the linear system of equations were solved with the direct solver "PARDISO" [SG04, SG06].

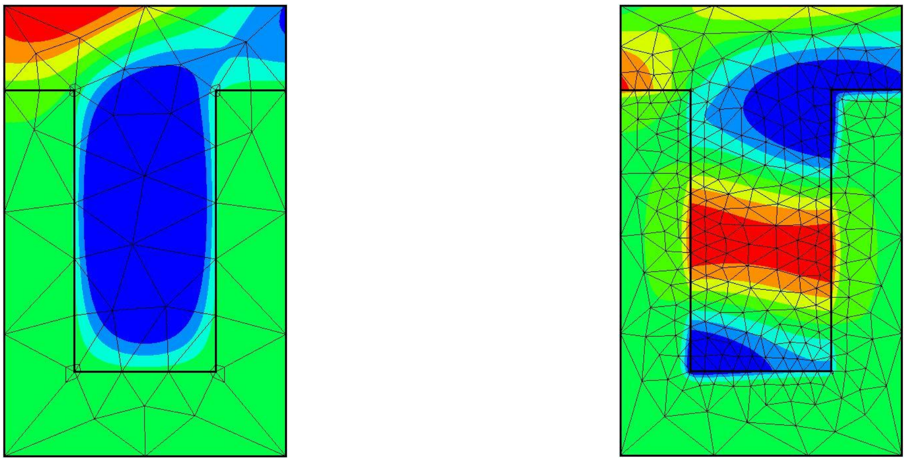

Figure 9.4: The real part of the electric field for TE modes(left) and the magnetic field for TM modes (right) of a lamellar grating. The wave vector of the incoming wave points into the direction of $(\sin(30°), -\cos(30°))^\top$.

9.5.1 A lamellar grating as benchmark problem

We start the numerical results section by studying a two dimensional benchmark problem from the literature [Gar99, BCW05]. There, a lamellar grating of period 1μm with groove-with 0.5μm and a groove-depth of 1μm is taken. The material above the grating is assumed to be vacuum with a refractive index of one, while the grating itself is highly absorbing with a refractive index of $0.22+6.71i$. For such a structure an incoming wave with an angle of incidence of 30° to normal incidence is considered.

The left hand plot of Figure 9.4 shows the real part of the electric field for an incoming TE wave. In order to get a good approximation of the solution, geometric refinement towards the corner points of the grating was needed. There, in each refinement step these vertices are cut off for the actual mesh with a geometric refinement factor of 0.125. For geometric refinement of order one and polynomial order 12, which leads for the mesh of Figure 9.4 to 15849 unknowns, the fraction of intensity reflected into the direction of order -1 is with 0.7342789 in very good agreement with [Gar99, BCW05].

In the TM case, due to singularities, a finer mesh and geometric refinement of order five was needed to get the same accuracy. Additionally, the high absorption in the substrate leads to a strongly decaying solution, as it is shown in the right hand plot of Figure 9.4, and a refinement of the mesh towards the interface is necessary. The calculation in this plot was done for polynomial order 6, which results in 23122 unknowns, and an intensity for the zero order reflection of 0.8484817 was obtained. This is again close to the results

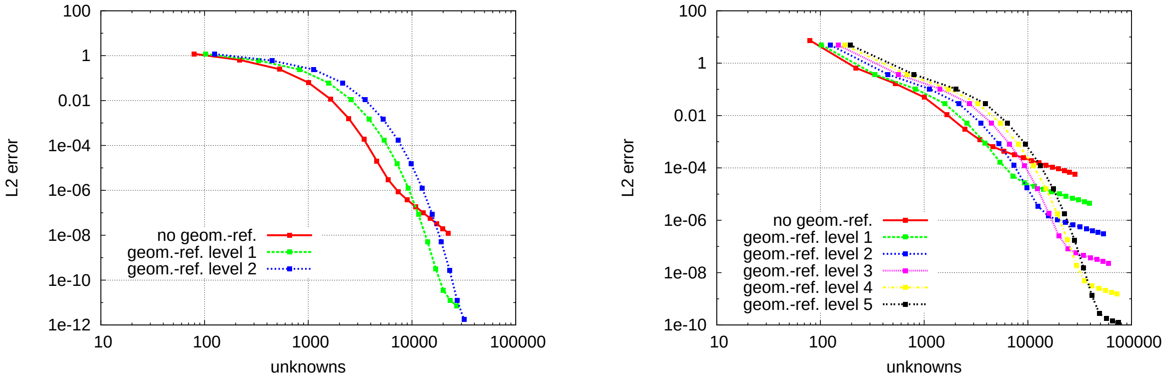

Figure 9.5: L^2-error as a function of unknowns for different geometric refinement levels in the lamellar grating example for the TE case (left) and the TM case (right)

of [Gar99, BCW05].

In Figure 9.5 we investigate for both polarizations the dependence of the solution on the number of geometric refinement levels towards the corner points of the grating profile and on the number degrees of freedom. The number of unknowns is varied for a fixed mesh by changing the polynomial order. In this example the L^2-error was calculated via the L^2-norm on $\Gamma^+ \cup \Gamma^-$ of the difference of the actual far field solution and a reference solution of higher accuracy. The plotted curves show some typical features. For small polynomial orders, in the preasymptotic range, exponential convergence of the solution can be reached. A further increase in the polynomial order slows because of the fixed mesh the convergence rate down to algebraic convergence. While, according to Figure 9.5, for TE modes one level of geometric refinement is sufficient, for TM modes geometric refinement has due to singularities at the corner points of the grating profile a large influence onto the error of the solution.

Finally, we examine the dependence of the solution on the distance of the boundaries $\Gamma^\pm$ from the grating. In Figure 9.6 the error in the intensity of the zero order reflection is plotted against the number of plane waves used in the calculation for different distances (in μm). The plot shows that the bigger the distance between the grating and the artificial boundary $\Gamma^\pm$, the less plane waves are needed to get an accurate solution for the far field. When cutting the unit cell in a small distance from the grating, the local field at the grating has still a large influence onto the solution on $\Gamma^\pm$, and a large number of evanescent waves is needed to describe it correctly. For big distances the influence of the evanescent waves decreases, and the solution at $\Gamma^\pm$ is dominated by propagating waves.

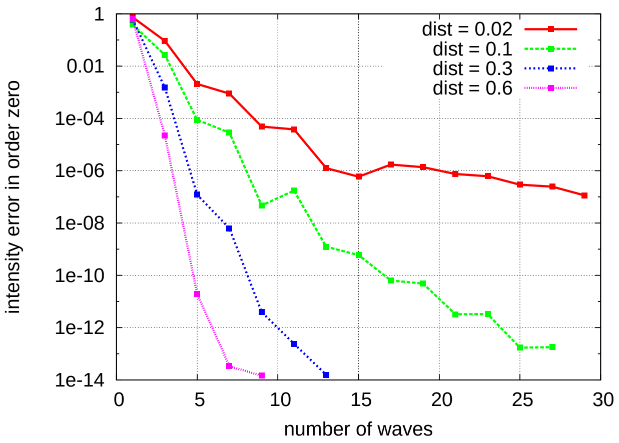

Figure 9.6: Error depending on the number of plane waves for different distances of the boundary $\Gamma^{\pm}$ from the surface of the grating

9.5.2 A large unit cell

This example is based on the same lamellar grating as above, with the difference that a much larger unit cell was taken. In order to show that the presented approach is able to treat two dimensional problems with a very small ratio of wavelength to period of the grating, the unit cell was chosen to consist of 100 periods of the lamellar grating, while the wavelength was kept constant. Thus, the ratio of wavelength to period is 400. For this setting about 400 plane waves are needed to describe the far field correctly. Figure 9.7 shows the real part of the electric field on a part of the super cell. By using 18000 elements of polynomial order 8, which corresponds to about 710000 unknowns, the reflected intensities could be calculated up to five digits. The calculation was done on a 2GHz Intel processor within 530 seconds.

9.5.3 A biperiodic grating

We will finish this section with a three dimensional example, a grating periodic in x and y-direction with periods of 0.6μm. A sketch of the unit cell is given in Figure 9.8. There, a strongly absorbing silicon substrate with refractive index $4.76+5i$ is covered by a 1 μm thick insulating SiO_2 layer with refractive index 1.5 (green in Figure 9.8), which is incorporated as substrate layer. The SiO_2 is coated by a weakly absorbing layer (blue) of thickness

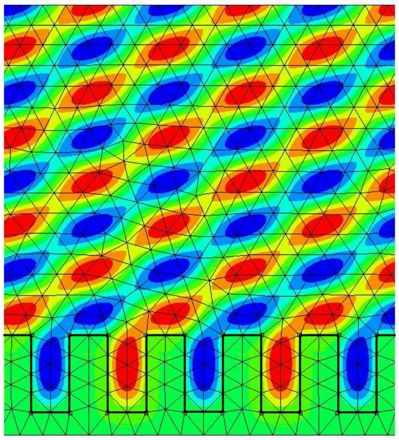

Figure 9.7: Real part of the electric field for a super cell of the lamellar grating. The incoming wave has a TE polarization and propagates into the direction of $(\sin(30°), -\cos(30°))^{\top}$.

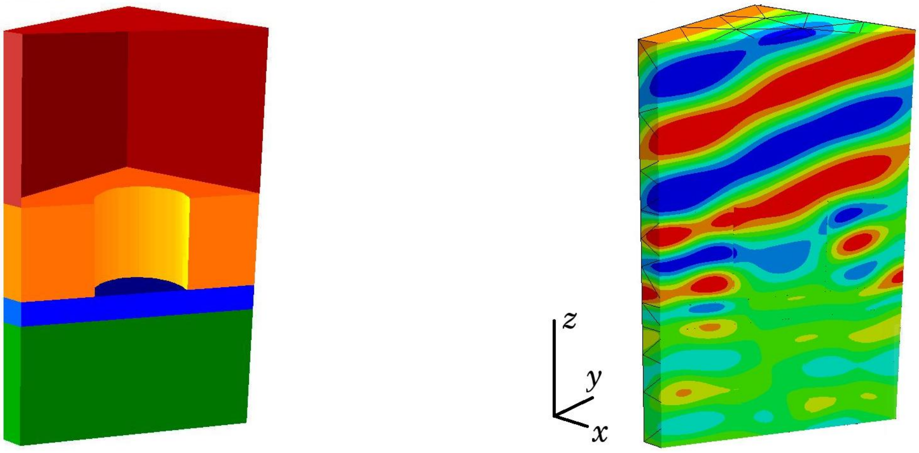

Figure 9.8: Geometry of the unit cell and real part of the x component of the electric field for the biperiodic grating. The incoming wave is polarized parallel to the xz-plane, and it propagates into the direction of $(\sin(30°), 0, -\cos(30°))^{\top}$.

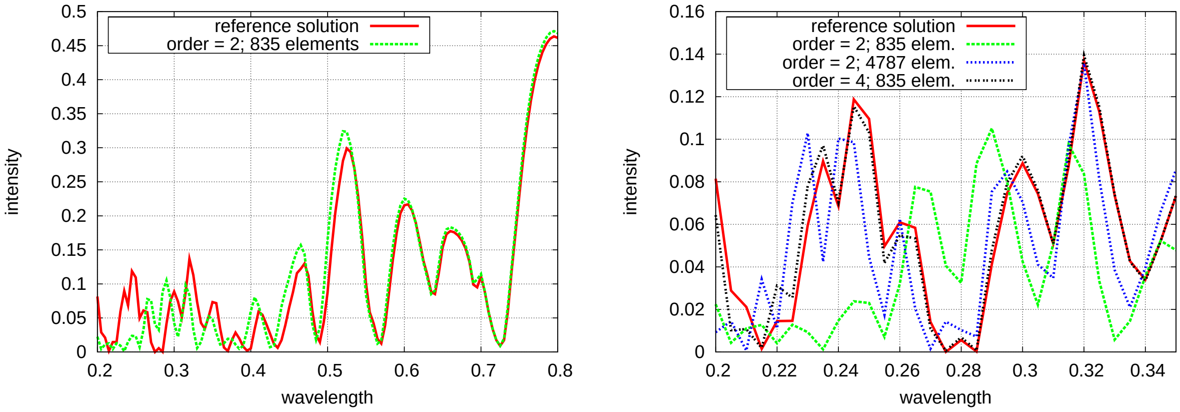

Figure 9.9: Zero order intensity as a function of the wavelength compared with a reference solution for the biperiodic grating

0.08μm with refractive index $2.62{+}0.48i$ and a photo resist (orange) of thickness 0.3μm and refractive index $1.68{+}0.003i$. Into this photo resist cylindric holes of radius 0.15μm are etched periodically. The refractive index above this structure and in the holes is taken as one.

For such a structure an incoming wave propagating parallel to the xz-plane with an electric field polarized in the plane of incidence and an angle of incidence of 30° is assumed. The electric field, which has to be tangential continuous across interfaces, can be naturally approximated by Nédélec finite elements, which provide tangential continuous basis functions on element interfaces. Figure 9.8 shows the real part of the x-component of the resulting electric field for a wavelength of 0.3μm. There, it can be clearly seen that in regions with a higher refractive index, like the photo resist, the wavelength is smaller, and the complex part of the refractive index in the absorbing layer (blue) leads to a damping out of the field.

For such a setting, the intensity of the zero order reflection was calculated for different wavelengths, polynomial orders and meshes. In Figure 9.9 this intensity is plotted against the wavelength, and the results are compared with a solution obtained by an RCW code [Bis01]. For the left hand plot, the underlying mesh of the finite element solution (green) consisted of 835 elements with polynomial order two, which corresponds to about 9100 unknowns. From the good agreement of the finite element solution with the RCW solution in the large wavelength region, we can conclude that already small polynomial orders lead to a good approximation of the solution.

For small wavelengths the situation is different. Therefore, we investigate in the right hand plot of Figure 9.9 two possibilities to get a better approximation. One option is to refine the mesh for a constant polynomial order. The blue line was calculated for a finer mesh consisting of 4787 elements with polynomial order two, which results in 48000 unknowns. By using this strategy just some features of the reference solution (red) can be resolved. Another approach is to increase the polynomial order while the mesh remains unchanged. Repeating the calculation for the original mesh of 835 elements with an increased polynomial order of four leads for approximately the same number of unknowns as above to the black line. The black line agrees much better with the reference solution, and for wavelength larger than 0.3μm they are almost identical.

Chapter 10

Summary and Outlook

Results and Conclusions

The main goal of the thesis is to develop new solvers for the Helmholtz equation and the vector valued wave equation, which are applicable to problems with large wave numbers, or more precisely to problems with a small ratio of wavelength to domain size.

The foundation of all our solvers is the consistent mixed hybrid formulation for the Helmholtz equation and the vector valued wave equation, respectively, presented in Chapter 5. This formulation allows to reduce the original problem to a problem posed on the skeleton of an underlying mesh. Consequently, the number of degrees of freedom is reduced significantly. For the mixed hybrid formulation we were able to prove existence and uniqueness of the solution in the continuous case. In addition, it was shown that the continuous problem as well as the discrete problem is energy conserving.

The facet degrees of freedom of this hybrid formulation are strongly related to impedance traces. Therefore, it allows in a natural way to find for domain decomposition solvers appropriate interface conditions, and convergent methods can be obtained. Chapter 6 was dedicated to such solvers. There, we discussed how to combine additive Schwarz and multiplicative Schwarz solvers as well as the BDDC preconditioner with our formulation in order to get convergent schemes. Furthermore, a new Robin type domain decomposition preconditioner, based on direct solvers for the subdomain problems is introduced. The two latter can be easily implemented in parallel environments, and they are well suited for efficiently solving large problems with a small ratio of wavelength to domain size. This was demonstrated by numerical examples.

In Chapter 8 we discuss an optimized implementation of the mixed hybrid formulation based on eigenfunctions for the two dimensional Helmholtz equation. This approach allows for discretizations with polynomial orders up to thousands. Numerical examples demonstrate the efficiency of this solution strategy for large regions with constant material parameters, especially in the case of large wave numbers. A drawback of the method is that it is limited to rectangular meshes, which requires the introduction of hanging nodes for complicated structures, and which therefore complicates the modeling process.

Throughout the major part of the thesis robin boundary conditions are used as transparent boundary conditions. In Chapter 7 it was demonstrated for the two dimensional Helmholtz equation, that the mixed hybrid formulation can also be combined with Hardy space infinite elements, a more powerful method in order to realize transparent boundaries. Furthermore, as discussed in Chapter 8, it is possible to adapt these elements such that they can be used together with the optimized implementation based on eigenfunctions.

Finally, the simulation of periodic structures for the Helmholtz and the vector valued wave equation was discussed in Chapter 9. The presented approach combines the advantages of the finite element method, which is used to approximate the near field, with a physical description of the solution in the semi infinite far field region, a plane wave expansion. The originality of the method lies in the way the coupling of these two approximations is realized, namely the method of Nitsche. Numerical results indicate that the performance of our approach is more than competitive with existing methods. This is due to the flexibility of *hp* finite elements.

Future Work

Some numerical results (compare Figures 5.1 and 5.3) indicate that the solution of the mixed hybrid problem converges for both, the Helmholtz equation and the vector valued wave equation with optimal order. A point of further investigation would be to verify these convergence rates theoretically. Therefore, a detailed error analysis of the hybrid formulation is necessary.

The preconditioners presented in Chapter 6 are based on robin type boundary conditions of the underlying problem. Thus, a topic of further research would be to examine these solvers together with problem settings containing more accurate realizations of transparent boundary conditions like perfectly matched layers or the presented Hardy space infinite elements.

Besides this, a generalization of the optimized approach from Chapter 8 based on discrete one dimensional eigenfunctions to higher dimensions, as well as an application of this technique to the time harmonic Maxwell case deserves further investigation. While such a generalization seems to be straight forward in the Helmholtz case, it is not clear, if a competitive realization of the mixed hybrid formulation for the vector valued wave equation using a discrete eigenfunction basis is possible.

A further promising topic is to use the hybridization technique together with domain decomposition solvers for the simulation of diffraction gratings. This would simplify solving problems containing periodic structures a lot, and computations for a small ratio of wavelength to period are possible. The challenging point will be to find efficient preconditioners. Since the plane wave degrees of freedom couple to all facet degrees of freedom on the boundary above and below the grating, respectively, the structure of the system matrix is not comparable to the structure of a matrix obtained from a wave type problem with absorbing boundary conditions.

Bibliography

[AB85] D.N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates. *RAIRO Model. Math. Anal. Numer.*, 19(1):7–32, 1985.

[Abb91] T. Abbound. *Étude Mathématique et Numérique de quelques Problèmes de Diffraction d'ondes éléctromagnetiques.* Phd thesis, Ecole Polytechnique Palaiseau, 1991.

[ABCM02] D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[AC03] M. Ainsworth and J. Coyle. Hierarchic finite element bases on unstructured tetrahedral meshes. *Internat. J. Numer. Methods Engrg.*, 58(14):2103–2130, 2003.

[Ada75] R.A. Adams. *Sobolev Spaces*, volume 65 of *Pure and Applied Mathematics*. Academic Press, 1975.

[Ain04] M. Ainsworth. Discrete dispersion relation for *hp*-version finite element approximation at high wave number. *SIAM J. Numer. Anal.*, 42(2):553–575, 2004.

[AP02] M. Ainsworth and K. Pinchedez. *hp*- approximation theory for BDFM and RT finite elements on quadrilaterals. *SIAM J. Numer. Anal.*, 40(6):2047–2068, 2002.

[Ast00] R.J. Astley. Infinite elements for wave problems: a review of current formulations and an assessment of accuracy. *Int. J. Numer. Meth. Engrg.*, 49(7):951–976, 2000.

[Bao97] G. Bao. Variational approximation of Maxwell's equations in biperiodic structures. *SIAM J. Appl. Math.*, 57(2):364–381, 1997.

[BCW05] G. Bao, Z. Chen, and H. Wu. Adaptive finite-element method for diffraction gratings. *J. Opt. Soc. Amer. A*, 22(6):1106–1114, 2005.

[BD00] G. Bao and D.C. Dobson. On the scattering by a biperiodic structure. *Proc Amer. Math. Soc.*, 128(9):2715–2723, 2000.

[BDM85] F. Brezzi, J. Douglas, and L.D. Marini. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.*, 47(2):217–235, 1985.

[Ber94] J.P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114(2):185–200, 1994.

[BF91] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer, Berlin, 1991.

[BGT83] A. Bayliss, C.I. Goldstein, and E. Turkel. An iterative method for the Helmholtz equation. *J. Comput. Phys.*, 49(3):443–457, 1983.

[Bis01] J. Bischoff. *Beiträge zur theoretischen und experimentellen Untersuchung der Lichtbeugung an mikrostrukturierten Mehrschichtsystemen*. Habilitation thesis, TU Ilemnau, 2001.

[Blo28] F. Bloch. Über die Quantenmechanik von Elektronen in Kristallgittern. *Z. Phys.*, 52:555–600, 1928.

[BPWX91] J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comp.*, 57(195):1–21, 1991.

[BR93] O.P. Bruno and F. Reitich. Numerical solution of diffraction problems: A method of variation of boundaries. *J. Opt. Soc. Amer. A*, 10(6):1168–1175, 1993.

[BR96] O.P. Bruno and F. Reitich. Calculation of electromagnetic scattering via boundary variations and analytic continuation. *Appl. Comput. Electromagn. Soc. J.*, 11(1):17–31, 1996.

[Bra03] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.* Springer, Berlin, 3 edition, 2003.

[BS08] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics.* Springer, New York, 3 edition, 2008.

[BW99] M. Born and E. Wolf. *Principles of Optics.* Cambridge University Press, Cambridge, 7 edition, 1999.

[BZ00] J.H. Bramble and X. Zhang. *The analysis of multigrid methods*, volume 7 of *Handbook of Numerical Analysis.* North-Holland, Amsterdam, 2000.

[CCS06] J. Carrero, B. Cockburn, and D. Schötzau. Hybridized globally divergence-free LDG methods. Part I: The Stokes problem. *Math. Comp.*, 75(254):533–563, 2006.

[CD98] O. Cessenat and B. Despres. Application of an Ultra Weak Variational Formulation of elliptic PDEs to the two-dimensional Helmholtz problem. *SIAM J. Numer. Anal.*, 35(1):255–299, 1998.

[CDCM82] J. Chandezon, M.D. Dupuis, G. Cornet, and D. Maystre. Multicoated gratings: A differential formalism applicable in the entire optical region. *J. Opt. Soc. Amer.*, 72(7):839–846, 1982.

[CG04] B. Cockburn and J. Gopalakrishnan. A characterization of hybrid mixed methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 42(1):283–301, 2004.

[CG05a] B. Cockburn and J. Gopalakrishnan. Error analysis of variable degree mixed methods for elliptic problems via hybridization. *Math. Comp.*, 74(252):1653–1677, 2005.

[CG05b] B. Cockburn and J. Gopalakrishnan. Incompressible finite elements via hybridization. Part I: The Stokes system in two space dimensions. *SIAM J. Numer. Anal.*, 43(4):1627–1650, 2005.

[CG05c] B. Cockburn and J. Gopalakrishnan. Incompressible finite elements via hybridization. Part II: The Stokes system in three space dimensions. *SIAM J. Numer. Anal.*, 43(4):1651–1672, 2005.

[Cia78] P.G. Ciarlet. *The finite element method for elliptic problems.* North-Holland, Amsterdam, 1978.

[CK98] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*, volume 93 of *Applied Mathematical Sciences.* Springer, Berlin, 2 edition, 1998.

[CKS00] L. Cockburn, G. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin Methods: Theory Computation and Applications.* Springer, Berlin, 2000.

[Dem06] L. Demkowicz. *Computing with hp-adaptive finite elements, Volume I, One and two dimensional Maxwell problems.* Chapman & Hall / CRC Press, Boca Raton (FL), 2006.

[Des91] B. Després. *Méthodes de décomposition de domains pour les problèms de propagation d'ondes en régime harmonique.* Phd thesis, Université Paris IX Dauphine, 1991.

[DG98] L. Demkowicz and K. Gerdes. Convergence of the infinite element methods for the Helmholtz equation in separable domains. *Numer. Math.*, 79(1):11–42, 1998.

[Doh03a] C.R. Dohrmann. A preconditioner for substructuring based on constrained energy minimization. *SIAM J. Sci. Comput.*, 25(1):246–258, 2003.

[Doh03b] C.R. Dohrmann. A study of domain decomposition preconditioners. Technical Report SAND2003-4391, Sandia National Laboratories, 2003.

[Dur70] L. Duren. *Theory of H^p spaces*, volume 38 of *Pure and Applied Mathematics.* Academic Press, San Diego (CA), 1970.

[dV65] B. Fraejis de Veubeke. Displacement and equilibrium models in the finite element method. In O.C. Zienkiewicz and G. Holister, editors, *Stress Analysis*, chapter 9, pages 145–197. John Wiley & Sons, New York, 1965.

[EG12] O.G. Ernst and M.J. Gander. Why it is difficult to solve the Helmholtz problems with classical iterative methods. 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 325–363. Springer, 2012.

[EHS02] J. Elschner, R. Hinder, and G. Schmidt. Finite element solution of conical diffraction problems. *Adv. Comput. Math.*, 16(2-3):139–156, 2002.

[Erl08] Y.A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Arch. Comput. Methods Eng.*, 15(1):37–66, 2008.

[EVO04] Y.A. Erlangga, C. Vuik, and C.W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Appl. Numer. Math.*, 50(3-4):409–425, 2004.

[EY11a] B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: Hierarchical matrix representation. *Comm. Pure Appl. Math.*, 64(5):697–735, 2011.

[EY11b] B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: Moving perfectly matched layers. *Multiscale Model. Simul.*, 9(2):686–710, 2011.

[FATL05] C. Farhat, P. Avery, R. Tezaur, and J. Li. FETI-DPH: A dual-primal domain decomposition method for acoustic scattering. *J. Comput. Acoustics*, 13(3):499–524, 2005.

[FHH03] C. Farhat, I. Harari, and U. Hetmaniuk. A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime. *Comput. Methods Appl. Mech. Engrg.*, 192(11-12):1389–1419, 2003.

[Flo83] G. Floquet. Sur les équations différentielles linéaries à coefficients périodiques. *Ann. Sci École Norm. Sup.*, 12:47–88, 1883.

[FLP00] C. Farhat, M. Lesoinne, and K. Pierson. A scalable dual-primal domain decomposition method. *Numer. Linear Algebra Appl.*, 7(7-8):687–714, 2000.

[FML00] C. Farhat, A. Macedo, and M. Lesoinne. A two-level domain decomposition method for the iterative solution of high frequency exterior Helmholtz problems. *Numer. Math.*, 85(2):283–308, 2000.

[FR91] C. Farhat and F.X. Roux. A method of finite element tearing and interconnecting and its parallel solution algorithm. *Internat. J. Numer. Meth. Engrg.*, 32(6):1205–1227, 1991.

[FW09] X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM J. Numer. Anal.*, 47(4):2872–2896, 2009.

[Gar99] G. Garnet. Reformulation of the lamellar grating problem through the concept of adaptive spatial resolution. *J. Opt. Soc. Amer. A*, 16(10):2510–2516, 1999.

[Giv04] D. Givoli. High-order local non-reflecting boundary conditions: A review. *Wave Motion*, 39(4):319–326, 2004.

[GK95] M.J. Grote and J.B. Keller. Exact nonreflecting boundary conditions for the time dependent wave equation. *SIAM J. Appl. Math.*, 55(2):280–297, 1995.

[GMN02] M.J. Gander, F. Magoulès, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60, 2002.

[GR86] V. Girault and P.A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer, Berlin, 1986.

[HHL03] A. Hansbo, P. Hansbo, and M.G. Larson. A finite element method on composite grids based on Nitsche's method. *ESAIM:Math. Model. Numer: Anal.*, 37(3):495–514, 2003.

[HHS10] A. Hannukainen, M. Huber, and J. Schöberl. A mixed hybrid finite element method for the Helmholtz equation. *J. Mod. Opt.*, 58(5-6):424–437, 2010.

[HN09] T. Hohage and L. Nannen. Hardy space infinite elements for scattering and resonance problems. *SIAM J. Numer. Anal.*, 47(2):972–996, 2009.

[HPS11] M. Huber, A. Pechstein, and J. Schöberl. Hybrid domain decomposition solvers for the Helmholtz and the time harmonic Maxwell's equation (to appear). In *Domain Decomposition Methods in Science and Engineering XX, Lecture Notes in Computational Science and Engineering*. Springer, 2011.

[HPSS05] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Interior penalty method for the indefinite time-harmonic Maxwell equations. *Numer. Math.*, 100(3):485–518, 2005.

[HS12] M. Huber and J. Schöberl. Hybrid domain decomposition solvers for the Helmholtz equation (submitted). In *Domain Decomposition Methods in Science* *and Engineering XXI, Lecture Notes in Computational Science and Engineering*. Springer, 2012.

[HSSZ09] M. Huber, J. Schöberl, A. Sinwel, and S. Zaglmayr. Simulation of diffraction in periodic media with a coupled finite element and plane wave approach. *SIAM J. Sci. Comput.*, 31(2):1500–1517, 2009.

[HT95] I. Harari and E. Turkel. Accurate finite difference methods for time harmonic wave propagation. *J. Comput. Phys.*, 119(2):252–270, 1995.

[HW08] G.C. Hsiao and W.L. Wendland. *Boundary integral equations*, volume 164 of *Applied Mathematical Sciences*. Springer, Berlin, 2008.

[IB97] F. Ihlenburg and I. Babuska. Finite element solution of the Helmholtz equation with high wave number Part II: *hp*-version of the FEM. *SIAM J. Numer. Anal.*, 34(1):315–358, 1997.

[Ihl98] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer, New York, 1998.

[Jac99] D.J. Jackson. *Classical Electrodynamics*. John Wiley & Sons, New York, 3 edition, 1999.

[Kuc01] P. Kuchment. The mathematics of photonic crystals. In G. Bao, L. Cowsar, and W. Masters, editors, *Mathematical Modeling in Optical Science*, volume 22 of *Frontiers in Applied Mathematics*, pages 207–272. Society for Industrial and Applied Mathematics, 2001.

[KWD02] A. Klawonn, O.B. Widlund, and M. Dryja. Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients. *SIAM J. Numer. Anal.*, 40(1):159–179, 2002.

[Li97] L. Li. New formulation of the fourier modal method for crossed surface-relief gratings. *J. Opt. Soc. Amer. A*, 14(10):2758–2767, 1997.

[LL84] E.D. Landau and E.F. Lifshitz. *Electrodynamics of Continuous Media*, volume 8 of *Course of Theoretical Physics*. Pergamon Press, Oxford, 2 edition, 1984.

[LP11] S. Lanteri and R. Perrussel. An implizit hybridized discontinuous Galerkin method for time-domain Maxwell's equations. Rapport de recherche RR-7578, INRIA, Université de Nice Sophia Antipolis, 2011.

[LW06] J. Li and O.B. Widlund. FETI-DP, BDDC, and block Cholesky methods. *Int. J. Numer. Meth. Engrg.*, 66(2):250–271, 2006.

[Man93] J. Mandel. Balancing domain decomposition. *Comm. Numer. Methods Engrg.*, 9(3):233–241, 1993.

[MB96] J.M. Melenk and I. Babuska. The partition of unity finite element method: Basic theory and applications. *Comput. Methods Appl. Mech. Engrg.*, 139(1-4):289–314, 1996.

[McL00] W. McLean. *Strongly elliptic systems and boundary integral methods.* Cambridge University Press, Cambridge, 2000.

[MDT05] J. Mandel, C.R. Dohrmann, and R. Tezaur. An algebraic theory for primal and dual substructuring methods by constraints. *Appl. Numer. Math.*, 54(2):167–193, 2005.

[Mel95] J.M. Melenk. *On Generalized Finite Element Methods.* Phd thesis, University of Maryland, 1995.

[MG81] M.G. Moraham and T.K. Gaylord. Rigorous coupled-wave analysis of planar-grating diffraction. *J. Opt. Soc. Amer.*, 71(7):811–818, 1981.

[Mon03] P. Monk. *Finite Element Methods for Maxwell's Equations.* Oxford University Press, Oxford, 2003.

[MSS10] P. Monk, A. Sinwel, and J. Schöberl. Hybridizing Raviart-Thomas elements for the Helmholtz equation. *Electromagnetics*, 30(1):149–176, 2010.

[MT01] J. Mandel and R. Tezaur. On the convergence of a dual-primal substructuring method. *Numer. Math.*, 88(3):543–558, 2001.

[MW99] P. Monk and D.Q. Wang. A least-squares methods for the Helmholtz equation. *Comput. Meth. Appl. Mech. Engrg.*, 175(1-2):121–136, 1999.

[N8́0] J.C. Nédélec. Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 35(3):315–341, 1980.

[N8́6] J.C. Nédélec. A new family of mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 50(1):57–81, 1986.

[N0́1] J.C. Nédélec. *Acoustic and electromagnetic scattering*, volume 144 of *Applied Mathematical Sciences.* Springer, Berlin, 2001.

[Nan08] L. Nannen. *Hardy-Raum Methoden zur numerischen Lösung von Streu- und Resonanzproblemen auf unbeschränkten Gebieten.* Phd thesis, Georg-August-Universität zu Göttingen, 2008.

[NCMC71] M. Nevière, G. Cerutti-Maori, and M. Cadhilac. Sur une nouvelle méthode de résolution de problème de la diffraction d'une onde plane par un réseau infiniment conducteur. *Opt. Commun.*, 3(1):48–52, 1971.

[NHSS11] L. Nannen, T. Hohage, A. Schädle, and J. Schöberl. High order curl-conforming Hardy space infinite elements for exterior Maxwell problems. *arXiv:1103.2288v1*, 27 pages, 2011.

[Nit71] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1970/71.

[NP03] E. Nevière and E. Popov. *Light Propagation in periodic Media: Differential Theory and Design.* Marcel Dekker, New York, 2003.

[NPC11] N.C. Nguyen, J. Peraire, and B. Cockburn. Hybridizable discontinuous Galerkin methods for the time harmonic Maxwell's equations. *J. Comput. Phy.*, 230(19):7151–7175, 2011.

[NS11] L. Nannen and A. Schädle. Hardy space infinite elements for Helmholtz-type problems with unbounded inhomogeneities. *Wave Motion*, 48(2):116–129, 2011.

[Pet80] R. Petit, editor. *Electromagnetic Theory of Gratings*, volume 22 of *Topics in Current Physics.* Springer, New York, 1980.

[PM94] C. Le Potier and R. Le Martret. Finite volume solution for Maxwells equation's in nonsteady mode. *La Recherche Aérospatiale*, 5:329–342, 1994.

[PMM97] D.W. Prather, M.S. Mirotznik, and J.N. Mait. Boundary integral methods applied to the analysis of diffractive optical elements. *J. Opt. Soc. Amer. A*, 14(1):34–43, 1997.

[PSM02] I. Perugia, D. Schötzau, and P. Monk. Stabilized interior penalty methods for the time-harmonic Maxwell equations. *Comput. Meth. Appl. Mech. Engrg.*, 191(41-42):4675–4697, 2002.

[Rei05] G.A. Reider. *Photonik: Eine Einführung in die Grundlagen.* Springer, Wien, 2 edition, 2005.

[RKE$^+$07] C.D. Riyanti, A. Kononov, Y.A. Erlangga, C. Vuik, C.W. Oosterlee, R.-E. Plessix, and W.A. Mulder. A parallel multigrid-based preconditioner for the 3d heterogeneous high frequency Helmholtz equation. *J.Comput. Phys.*, 224(1):431–448, 2007.

[RT77] P.A. Raviart and J.M. Thomas. A mixed finite element method for second order elliptic problems. In I. Galligani and E. Magenes, editors, *Mathematical Aspects of Finite Element Methods*, pages 292–315. Springer, 1977.

[SBG96] B.F. Smith, P.F. Bjorstad, and W.D. Gropp. *Domain Decomposition: Parallel multilevel algorithms for elliptic partial differential equations.* Cambridge University Press, Cambridge, 1996.

[Sch70] H.A. Schwarz. Über einen Grenzübergang durch alternierendes Verfahren. *Vierteljahresschrift der Naturforschenden Gesellschaft in Zürich*, 15:272–286, 1870.

[Sch97] J. Schöberl. NETGEN - an advanced front 2D/3D-mesh generator based on abstract rules. *Comput. Vis. Sci.*, 1(1):41–52, 1997.

[Sch98a] F. Schmidt. An alternative derivation of the exact DtN-map on a circle. ZIB-Report SC-98-32, Zuse Institute Berlin (ZIB), 1998.

[Sch98b] C. Schwab, editor. *p- and hp- finite element methods: Theory and applications in solid and fluid mechanics.* Oxford University Press, New York, 1998.

[Sch02] F. Schmidt. *Solution of interior-exterior Helmholtz-type problems based on the pole condition concept: Theory and algorithms.* Habilitation thesis, Free University Berlin, 2002.

[Sch04] G. Schmidt. Electromagnetic scattering by periodic structures. *J. Math. Sci.*, 124(6):5390–5406, 2004.

[SG04] O. Schenk and K. Gärtner. Solving unsymmetric sparse systems of linear equations with PARDISO. *J. Future Generation Comput. Syst.*, 20(3):475–487, 2004.

[SG06] O. Schenk and K. Gärtner. On fast factorization pivoting methods for sparse symmetric indefinite systems. *Electron. Trans. Numer. Anal.*, 23:158–179, 2006.

[Sim79] B. Simon. The definition of molecular resonance curves by the method of exterior complex scaling. *Phys. Lett. A*, 71(2-3):211–214, 1979.

[Sin09] A. Sinwel. *A New Family of Mixed Finite Elements for Elasticity.* Phd thesis, Johannes Kepler Universität, 2009.

[Ste98] R. Stenberg. Mortaring by a method of J.A. Nitsche. In S. Idelsohn, E. Onate, and E. Dvorkin, editors, *Computational Mechanics, New Trends and Applications.* International Center of Nnumerical Methods Engineering, 1998.

[SZ05] J. Schöberl and S. Zaglmayr. High order Nédélec elements for local complete sequence properties. *COMPEL*, 24(2):374–384, 2005.

[SZB$^+$07] A. Schädle, L. Zschiedrich, S. Burger, R. Klose, and F. Schmidt. Domain decomposition method for Maxwell's equations: scattering of periodic structures. *J. Comput. Phys.*, 226(1):477–493, 2007.

[Sze39] G. Szegö. *Orthogonal Polynomials*, volume 23 of *Colloquium Publications.* American Mathematical Society, Rhode Island, USA, 1939.

[TEY12] P. Tsuji, B. Engquist, and L. Ying. A sweeping preconditioner for time-harmonic Maxwell's equations with finite elements. *J. Comput. Phys.*, 231(9):3770–3783, 2012.

[TF06] R. Tezaur and C. Farhat. Tree-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems. *Internat. J. Numer. Methods Engrg.*, 66(5):796–815, 2006.

[TMF01] R. Tezaur, A. Macedo, and C. Farhat. Iterative solution of large-scale acoustic scattering problems with multiple right hand-sides by a domain decomposition method with Lagrange multipliers. *Internat. J. Numer. Methods Engrg.*, 51(10):1175–1193, 2001.

[TW05] A. Toselli and O. Widlund. *Domain Decomposition Methods - Algorithms and Theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer, Berlin, 2005.

[vdV03] H. van der Vorst. *Iterative Krylov methods for large linear systems*, volume 13 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2003.

[Vin78] P. Vincent. A finite difference-method for dielectric and conducting crossed gratings. *Opt. Commun.*, 26(3):293–296, 1978.

[Zag06] S. Zaglmayr. *High Order Finite Element Methods for Electromagnetic Field Computation*. Phd thesis, Johannes Kepler Universität, 2006.

Lebenslauf

Persönliche Daten

Name: Martin Huber

Geburtstag: 23. Dezember 1979

Geburtsort: Linz (Österreich)

Nationalität: Österreich

Ausbildung

09/1986–07/1990	Volksschule Oberkappel
09/1990–05/1999	Gymnasium Untergriesbach
05/1999	Abitur am Gymnasium Untergriesbach
09/2000–07/2006	Studium der Technischen Physik an der Johannes Kepler Universität Linz
07/2006	Diplomprüfung in Technische Physik
08/2006–03/2008	Doktoratsstudium an der Johannes Kepler Universität Linz
04/2006–08/2011	Doktoratsstudium an der RWTH Aachen
seit 08/2011	Doktoratsstudium an der TU Wien

Beruflicher Werdegang

07/2006–04/2008	Research Assistent am Johann Radon Institut for Computational and Applied Mathematics (RICAM) der Österreichischen Akademie der Wissenschaften in Linz
04/2008–09/2010	Wissenschaftlicher Mitarbeiter am Institut MathCCES der RWTH Aachen
10/2010–01/2013	Universitätsassistent am Institut für Analysis und Scientific Computing der TU Wien