DISSERTATION

A Hybrid Discontinuous Galerkin Method with Impedance Traces for the Helmholtz 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, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Ing. Dipl.-Ing. Michael Leumüller
[redacted]

Diese Dissertation haben begutachtet:

1. **Professeur associé Xavier Claeys, Ph.D.**
   Unité de Mathématiques Appliquées, ENSTA Paris
2. **Prof. Dr. Stefan A. Sauter**
   Institut für Mathematik, Universität Zürich
3. **Univ.Prof. Dipl.-Ing. Dr.techn. Joachim Schöberl**
   Institut für Analysis und Scientific Computing, TU Wien

Wien, am 6. Februar 2025

Kurzfassung

Die Helmholtz-Gleichung mit absorbierenden Randbedingungen ist Gegenstand zahlreicher Forschungen. Sie beschreibt die zeit-harmonische Wellenausbreitung. Die Entwicklung von Simulationsmethoden stellt aufgrund des oszillatorischen Verhaltens der Lösungen der Helmholtz-Gleichung eine Herausforderung dar. Wird die Finite Elemente Methode angewendet, so ist eine feine Netzgröße erforderlich, was bei vielen Wellenlängen im Rechengebiet zu großen Systemen linearer Gleichungen führt, die gelöst werden müssen. Direkte Löser können eingesetzt werden, um diese Probleme zu lösen, aber der Speicherbedarf solcher Löser ist zu hoch und steigt zu schnell mit der Netzfeinheit. Vor konditionierte iterative Löser bieten eine Antwort auf diese Probleme, und für koerzive (elliptische) Probleme wurden solche mit großem Erfolg entwickelt und angewendet, z. B. Mehrgitterverfahren. Lokale Glättungsschritte in Kombination mit direkten Grobgitterkorrekturen führen zu sehr zufriedenstellenden Ergebnissen. Leider ist die Helmholtz-Gleichung nicht koerzive und zeigt ein nicht lokales Verhalten; solche iterativen Löser können daher nicht angewendet werden. Die Idee, Gebietszerlegungsvorkonditionierer zusammen mit einem minimal-residualen iterativen Löser zu verwenden, wurde entwickelt. Die Helmholtz-Gleichung muss auf Teilgebieten gelöst werden, und daher müssen Vorkonditionierer, die auf Gebietszerlegung basieren, absolut stabil sein, im Sinne davon, dass die lokalen Vorkonditionierungsprobleme stets eindeutig lösbar sind. Es wurde gezeigt, dass diskontinuierliche Galerkin-Methoden eine lokale Stabilitätseigenschaft aufweisen. Durch die Einführung von diskontinuierlichen Methoden wird das ohnehin schon große System linearer Gleichungen erheblich vergrößert, da die koppelnden Unbekannten dupliziert werden. Die hybriden diskontinuierlichen Galerkin-Methoden mit statischen Kondensationsfähigkeiten wurden entwickelt, um dieses Problem zu entschärfen. Alle Volumenunbekannten werden auf Skelettunbekannte reduziert, wodurch ein System linearer Gleichungen nur für diese Skelettunbekannten entsteht. Die Eigenschaft, statische Kondensation anwenden zu können, ist für die Helmholtz-Gleichung höchst nicht trivial, kann jedoch direkt aus einer lokalen Stabilitätseigenschaft der hybriden Methoden abgeleitet werden. Nicht alle hybriden diskontinuierlichen Galerkin-Methoden sind für Gebietszerlegungsvorkonditionierer im Kontext der Helmholtz-Gleichung geeignet. Die Übergangsbedingungen zwischen Teilgebieten sind entscheidend. Der Schwerpunkt dieser Dissertation liegt auf einer hybriden diskontinuierlichen Galerkin-Methode, die diese günstigen Eigenschaften aufweist. Sie ist lokal absolut stabil, weist optimale Konvergenzraten in Bezug auf die Netzgröße auf, und iterative Löser mit Vorkonditionierern basierend auf Gebietszerlegungskonzepten können angewendet werden. In dieser Arbeit wird die Stabilitäts- und Fehleranalyse der hybriden diskontinuierlichen Galerkin-Methode durchgeführt. Darüber hinaus werden die günstigen Eigenschaften der Methode im Hinblick auf iterative Löser durch numerische Simulationen hervorgehoben.

Abstract

The Helmholtz equation with absorbing boundary conditions has been the focus of much research. It describes a time-harmonic wave propagation. Due to the oscillatory behaviour of solutions for the Helmholtz equation, developing simulation methods is challenging. If the finite element method (FEM) is applied, then a fine mesh size is required leading for many wavelengths in the computational domain to large systems of linear equations, which need to be solved. Direct solver can be applied to solve these problems, but the memory consumption of such solvers is too demanding and increases too quickly with respect to the mesh size. Preconditioned iterative solvers are an answer to these issues, and for coercive (elliptic) problems, such have been developed and applied to great success, e.g. multi-grid. Local smoothing steps, in combination with direct coarse grid corrections, lead to very satisfactory results. Sadly, the Helmholtz equation is not coercive and exhibits a non-local behaviour; these iterative solvers can not be applied. The idea of using domain decomposition preconditioners with some minimal residual iterative solver has been devised. The Helmholtz equation needs to be solved on subdomains and therefore preconditioners based upon domain decomposition need to be absolutely stable in the sense that the local preconditioning problems are always uniquely solvable. It has been shown that discontinuous Galerkin (DG) methods exhibit a local stability property, leading to stably solvable problems on subdomains. By introducing DG methods, the already large system of linear equations is substantially increased due to the duplication of coupling unknowns. The hybrid discontinuous Galerkin (HDG) methods with static condensation capabilities have been developed to counteract this issue. All volume unknowns are condensed to skeleton unknowns, leading to a system of linear equations only for these skeleton unknowns. The number of skeleton unknowns is for small polynomial degrees larger than the number of unknowns for conforming spaces, but HDG methods exhibit less unknowns as DG methods. The property of being able to apply static condensation is highly non-trivial for the Helmholtz equation, but can be directly derived from a local stability property of HDG methods. Not all HDG methods are suitable for domain decomposition preconditions in the context of the Helmholtz equation. The transmission conditions between subdomains are crucial such that they represent impedance traces. The focus of this dissertation is exactly on a HDG method exhibiting all of these favourable properties. It is locally absolutely stable, exhibits optimal convergence rates with respect to the mesh size, and iterative solvers with preconditioners based on domain decomposition concepts can be applied. In this work, the rigorous stability and error analysis of the HDG method is carried out, which is a novelty. Additionally, the favourable properties concerning iterative solvers of the method are highlighted by large-scale numerical simulations.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 6. Februar 2025

Michael Leumüller

Contents

1. Introduction **1**
- 1.1. The Helmholtz Equation . . . 2
  - 1.1.1. Numerical Methods . . . 4

2. Finite Element Method **7**
- 2.1. Weak Derivatives and Weak Formulation . . . 7
  - 2.1.1. Weak Derivative . . . 7
  - 2.1.2. Weak Formulation . . . 8
- 2.2. Existence and Uniqueness of Weak Solutions . . . 9
  - 2.2.1. Coercive Sesquilinearform . . . 9
  - 2.2.2. Inf-Sup Stability . . . 10
- 2.3. Existence and Uniqueness for the Helmholtz Equation with Robin Boundary Conditions . . . 10
  - 2.3.1. The Dual Helmholtz Equation . . . 12
- 2.4. Discretisation . . . 13
  - 2.4.1. Triangulation and Mesh . . . 14
  - 2.4.2. Polynomial Spaces . . . 15
- 2.5. Quasi-optimality, Best Approximation and Error Estimates . . . 17
  - 2.5.1. Projection Based Error Estimation . . . 20
- 2.6. Solvers . . . 22
  - 2.6.1. Direct Solvers . . . 22
  - 2.6.2. Iterative Solvers . . . 22
  - 2.6.3. Static Condensation . . . 23
- 2.7. Discontinuous Galerkin Method . . . 25

3. Hybrid Discontinuous Galerkin Method **29**
- 3.1. Deriving HDG Formulations . . . 29
- 3.2. State of the Art of HDG Methods for the Helmholtz Equation with Robin Boundary Conditions . . . 31
- 3.3. The HDG-Formulation with Impedance Traces . . . 40
  - 3.3.1. Favourable Properties of B . . . 45
  - 3.3.2. Discrete Absolute Stability . . . 46
  - 3.3.3. Pre-Asymptotic Error Estimates for the Pressure and Jumps . . . 51
  - 3.3.4. Pre-Asymptotic Error Estimate for the Flux . . . 55
  - 3.3.5. Asymptotic Error Estimates for the HDG-Formulation . . . 61
  - 3.3.6. Static Condensation . . . 64
  - 3.3.7. Lowest Order Discretisation . . . 67

4. Numerical Results **73**
4.1. Convergence Rates . . . 73
4.1.1. Plane Waves in 2D . . . 73
4.1.2. Plane Waves in 3D . . . 75
4.2. Lowest Order Error Rates . . . 78
4.3. Dispersion and Dissipation . . . 83
4.3.1. H^1 and HDG-Comparison . . . 86
4.3.2. Wave Number Experiments . . . 86
4.4. Iterative Solver and Preconditioners . . . 89
4.4.1. Block-Jacobi Preconditioner . . . 92
4.4.2. Gauss-Seidel Preconditioner . . . 97
4.4.3. Sweeping Preconditioner . . . 98
4.4.4. Non-overlapping Domain Decomposition Preconditioner . . . 99
4.4.5. Stabilisation Parameters α and β . . . 101
4.4.6. Computational Costs . . . 103
4.5. Complementary Numerical Examples . . . 110
4.5.1. Heterogeneous Materials . . . 110
4.5.2. Scattering on Spheres . . . 112

A. Analysis of the HDG-Formulation with Brezzi-Douglas-Marini Spaces **113**
A.1. The interpolation P for $\mathcal{BDM}$ . . . 115

List of Figures **123**

List of Tables **127**

Bibliography **129**

1. Introduction

The Helmholtz equation is a pivotal partial differential equation in physics and engineering, frequently encountered in wave propagation, acoustics, electromagnetics, and quantum mechanics. In fields like acoustics and electromagnetics, the Helmholtz equation models how sound and electromagnetic waves propagate through different media. For instance, it helps to describe the behaviour of acoustic pressure fields or electromagnetic fields in a uniform medium. In quantum mechanics, the time-independent Schrödinger equation is a variant of the Helmholtz equation used to describe the spatial distribution of a particle's wave function within a potential field. The equation is also relevant in analysing the vibration modes of physical objects, such as the resonant frequencies of a drumhead or membrane vibrations. The solutions depend on the boundary conditions and domain geometry. Analytically solving the Helmholtz equation is often impractical, which gives rise to the necessity of numerical methods. This work focuses on the Helmholtz equation with impedance boundary conditions and is an extension of the work published in [LS23].

For high wave numbers, the FEM necessitates a very fine mesh, leading to large systems of linear equations. The computational costs associated with direct solvers increase significantly in these cases. Furthermore, the indefinite structure of the Helmholtz equation means that iterative solvers typically used for elliptic problems are not applicable. While domain decomposition preconditioners show promise [CCJP20, CP22, CCP22, cla23], they require inverting sub-blocks of finite element system matrices.

Discontinuous spaces can be utilised to avoid this problem, resulting in absolutely stable local methods, as demonstrated in [FX13]. In [MPS13], a DG method for the Helmholtz equation has been proposed and analysed. Notably, the existence of a discrete solution is proven without requiring a resolution condition for the discrete space, and quasi-optimality has been established in the asymptotic regime. Additional studies on DG methods have provided explicit pre-asymptotic estimates, such as those in [FW09, FW11, FX13, Wu14]. The analysis relies on carefully chosen test functions within DG spaces. This technique cannot be applied to conforming finite element (FE) spaces.

A disadvantage of DG methods compared to standard FEM is the increased number of coupling unknowns. The HDG method has been developed to address this issue. By applying the Schur complement, it results in a smaller system of linear equations for the coupling unknowns. This approach requires the invertibility of element matrices, which is closely related to local absolute stability. In [LCQ17], an absolutely stable HDG method for Maxwell's equations was analysed using a new technique based on L^2-projections, demonstrating discrete absolute stability.

DG and HDG methods rely on the stabilisation of jumps over facets, typically dependent on mesh size, originating from the analysis of the elliptic Poisson equation. In contrast, experiments in [Hub13, HPS13, HS14] suggest that iterative solvers for the Helmholtz equation require mesh size independent stabilisations, representing impedance traces between

subdomains. In [GM11], an HDG method with such a stabilisation has been analysed. The technique used is based on results from [CGS10] and utilises special projections tailored to the HDG-formulation employed in the study. The authors were able to demonstrate optimal convergence rates in the asymptotic range.

The main contribution of this work is the analysis of HDG formulations based upon the formulation introduced in [MSS10] and further investigated in [Hub13, HPS13, HS14]. The authors examined various promising iterative solvers, highlighting the favourable properties of this HDG-formulation. Numerical experiments in these studies indicate that solvers require a second variable per facet to represent the flux, similar to the discontinuous Petrov-Galerkin method in [DGMZ12, GMO14]. This work establishes the analytical foundation for the HDG method with two variables per facet, as introduced in [MSS10]. It employs a stabilisation that is independent of mesh size, domain, and wave number, which is known as a priori for all Helmholtz problems.

Structure of this Work In the following Section 1.1, the origins and historical context of the Helmholtz equation are first discussed. This is followed by an overview of well established numerical methods and their applications across various domains.

Chapter 2 delves into the FEM in detail, encompassing the theoretical framework of weak formulations and established theories concerning existence and uniqueness. Additionally, it introduces the discrete polynomial spaces used, culminating in error estimates. The chapter concludes with a brief, preliminary overview of solvers tailored to the Helmholtz equation.

The subsequent Chapter 2.7, dedicated to DG methods, serves as a preparatory stage for Chapter 3, the highlight of this study. Chapter 3 presents the HDG method and includes the detailed analysis. Emphasis is placed on stability, absolute stability, general error estimates, and asymptotic error behaviours, with minor attention given to analytic dispersion and dissipation analysis.

In Chapter 4, the theoretical findings established earlier are validated through numerical simulations. This chapter documents additional intriguing numerical experiments, including convergence tests, dispersion and dissipation data analysis, regularity experiments, and explorations of iterative solver schemes.

1.1. The Helmholtz Equation

The Helmholtz equation with a fixed wave number $\kappa > 0$

$$-\Delta u = \kappa^2 u \tag{1.1}$$

is the time-independent form of the wave equation and was first introduced by its name giver Hermann von Helmholtz in his book "Die Thermodynamik chemischer Vorgänge" [vH82] published in 1882. The one-dimensional wave equation in the form of the partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2},$$

with the fixed non-negative real coefficient c representing the propagation speed of the wave, has been first postulated by Jean-Baptiste le Rond d'Alembert in 1747 and published in "Recherches sur la courbe que forme une corde tenduë mise en vibration" [d'A47] and ten years later, Leonhard Euler discovered the wave equation in three space dimensions

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u$$

in "De vibratione chordarum exercitatio" [Eul49]. The homogenous Helmholtz equation (1.1) is, in general, defined in the whole domain and describes wave propagation. Fundamental solutions are, for example, plane waves

$$u(\mathbf{x}) = e^{j\mathbf{k}\cdot\mathbf{x}}$$

with the imaginary unit $j := \sqrt{-1}$ and the wave vector

$$\mathbf{k} \in \mathbb{R}^d, \qquad |\mathbf{k}| = \kappa,$$

in the spacial space $\mathbb{R}^d$ of dimension d. Note that contrary to the time domain, where the solution function is real-valued, in the frequency domain, it is complex-valued. In the early 20th century the Helmholtz equation with varying coefficients $\mu > 0$ and $\varepsilon > 0$

$$-\operatorname{div}(\mu^{-1}\nabla u) - \kappa^2 \varepsilon u = f$$

and an excitation f was introduced. Explicit solutions for this equation are hard to find and only exist for simplified problems. With the rise of numerics due to the increased computational resources, algorithmic methods to calculate discrete approximations have come into focus, but the unbounded space $\mathbb{R}^d$ is complicated, and most methods could not cope with such a domain. Therefore, truncation of the unbounded domain to a bounded domain Ω came into focus. At that point, it was essential to clarify which kind of condition needed to be applied for a reasonable solution. Generally, the Helmholtz equation permits incoming as well as outgoing propagating waves, meaning that energy constantly flows into the system or leaves the system. The solution with outgoing energy was chosen as appropriate, which is reflected by the Sommerfeld radiation condition

$$\lim_{|\mathbf{x}|\to\infty} |\mathbf{x}|^{\frac{d-1}{2}} \left(\frac{\partial}{\partial |\mathbf{x}|} - j\kappa \right) u(\mathbf{x}) = 0$$

published by Sommerfeld 1912 in [Som12]. In combination with this radiation condition, the existence of unique solutions was finally established, but it was wholly unsuitable for numerical methods.

Absorbing Boundary Condition of First Kind The Helmholtz equations are stated on the whole space $\mathbb{R}^3$ with a suitable radiation condition. For numerical simulations, this unbounded domain is unfeasible. An approach is to only consider a truncated bounded domain Ω, which contains all inhomogeneities and excitations. Then, the question arises of what to do on the introduced new boundary $\partial\Omega$. There, a suitable transparent boundary

needs to be introduced, which facilitates incoming and outgoing waves; therefore, reflecting boundary conditions are unsuitable. A numerically easily describable absorbing boundary condition is the absorbing boundary condition of first kind (ABC), also called Robin boundary condition

$$\nabla u \cdot \mathbf{n} - j\kappa u = g \qquad \text{on } \partial\Omega,$$

with the inhomogeneity g representing an impinging wave and the outward pointing normal vector $\mathbf{n}$. Literature regarding this topic was published in 1977 by Engquist and Majda in [EM77].

Boundary Value Problem In this work the following boundary value problem (BVP) representing the Helmholtz equation with an ABC is considered.

Definition 1 (Helmholtz Boundary Value Problem with Robin Boundary Conditions)**.** *Consider a bounded domain $\Omega \subset \mathbb{R}^3$ with a unique outward pointing normal vector field $\mathbf{n}$ on $\partial\Omega$. For a given wave number $\kappa > 0$ and spatially dependent physical parameters $\mu(\mathbf{x}) > 0$ and $\varepsilon(\mathbf{x}) > 0$, as well as a given volume excitation $f(\mathbf{x})$ and boundary excitation $g(\mathbf{x})$ find a scalar field $u(\mathbf{x})$ satisfying*

$$-\operatorname{div}(\mu^{-1}\nabla u) - \kappa^2 \varepsilon u = f \qquad \text{in } \Omega, \tag{1.2a}$$

$$\sqrt{\mu^{-1}}\nabla u \cdot \mathbf{n} - j\kappa\sqrt{\varepsilon} u = g \qquad \text{on } \partial\Omega. \tag{1.2b}$$

This work focuses on studies of this BVP. Additionally to these formulations, there are other equivalent formulations. The mixed formulation is introduced by defining the additional variable $\sigma := \frac{1}{j\kappa\mu}\nabla u$ leading to:

Definition 2 (Mixed Helmholtz Problem with Robin boundary conditions)**.** *Consider a bounded domain $\Omega \subset \mathbb{R}^3$ with a unique normal vector field $\mathbf{n}$ on $\partial\Omega$. For a given wave number $\kappa > 0$ and possibly spatially dependent physical parameters $\mu(\mathbf{x})$ and $\varepsilon(\mathbf{x})$, as well as an given volume excitation $f(\mathbf{x})$ and boundary excitation $g(\mathbf{x})$ find a scalar field $u(\mathbf{x})$ and a vector field $\sigma(\mathbf{x})$ so that they satisfy*

$$j\kappa\mu\sigma - \nabla u = 0 \qquad \text{in } \Omega, \tag{1.3a}$$

$$-\operatorname{div}(\sigma) + j\kappa\varepsilon u = \frac{1}{j\kappa} f \qquad \text{in } \Omega, \tag{1.3b}$$

$$\sqrt{\mu}\sigma \cdot \mathbf{n} - \sqrt{\varepsilon} u = \frac{1}{j\kappa} g \qquad \text{on } \partial\Omega. \tag{1.3c}$$

The scalar field is usually referred to as the pressure, and the vector field is the flux.

1.1.1. Numerical Methods

Various numerical methods for solving the Helmholtz equation have been established over the decades. This section will give a short incomplete overview.

Finite Difference Method The finite difference method (FDM) was already known by Leonhard Euler in the 18th century. It is suitable for solving problems within a bounded domain, which is divided into regular cubes, with each corner representing an unknown. These unknowns couple to the coefficients in the Taylor expansion of the solution up to a finite degree. The Laplacian is discretised using a finite difference approximation based on these Taylor coefficients, forming a so-called stencil. Incorporating transparent boundary conditions is more complex. A straightforward approach is to use a Robin boundary condition. Higher-order representations of boundary conditions are also available. The method's advantage is its ability to handle inhomogeneous materials, although it is limited by the requirement for a structured grid, which also constrains the geometry. Nevertheless, it is easy to implement and results in a sparse system matrix. Accuracy can be improved either by decreasing the grid spacing or by using a larger stencil. For a more detailed explanation, see "Finite Difference Methods for Ordinary and Partial Differential Equations" by Randall J. LeVeque (2007) [LeV07].

Boundary Element Method The boundary element method (BEM) is suitable for complex domains with homogeneous materials. This method relies on an integral equation defined solely on the boundaries. To handle an unbounded domain, it is truncated to a bounded domain such that all excitations and objects are enclosed within it. The method uses a Dirichlet-to-Neumann operator on the truncation boundary. This operator connects the Dirichlet data to the respective Neumann data of an outgoing solution. The integral equation is then discretised using functions with local support, typically piecewise polynomials. However, due to the non-local kernel in the integral equations, all degrees of freedom corresponding to these functions are coupled, resulting in a dense system of linear equations. Although the resulting matrix may be small, the numerical effort required for direct solving is considerable. Matrix compression techniques have been developed to address this issue. These techniques are based on the observation that functions far apart couple similarly, allowing these matrix blocks to be combined. This reduces both the assembly time and the solving time. The accuracy of BEM can be adjusted by changing the number of functions on the boundaries. For an in-depth exploration, refer to "Boundary Element Methods: Fundamentals and Applications" by Stefan A. Sauter and Christoph Schwab [SS11].

Finite Element Method Since the FEM is the main focus of this work and has its dedicated section, this will be a concise overview. FEM can handle complex domains and heterogeneous materials because it is based on functions, typically piecewise polynomials with local support. Similar to the FDM, high-order solutions are achievable. The local nature of the scheme results in a sparse system matrix. Various transparent boundary conditions can be applied to the unbounded domain, such as ABC or higher-order conditions, perfectly matched layers [Ber94], and infinite elements [Wes20]. Specifically, infinite elements provide discretisation without truncation but require somewhat homogeneous material at some truncation boundary. For literature on FEM, especially concerning the Helmholtz equation, refer to [IB95, IB97]. For applications to the similar Maxwell's equations, see Monk's book [Mon03].

2. Finite Element Method

In this chapter, the FEM, as well as all mathematical requirements and general conclusions are stated. To be able to start, the notion of weak formulations of partial differential equations needs to be clarified.

In further work, many integrals which can be represented by using scalar products are used. Therefore, the following short notation is defined.

Definition 3. *For a subset $M \subset \mathbb{R}^d$ and complex, possibly vector-valued fields $u(\mathbf{x}) \in \mathbb{C}^l, v(\mathbf{x}) \in \mathbb{C}^l$ with $l \in \mathbb{N}$ the integral short notation is defined as*

$$(u, v)_M := \int_M u(\mathbf{x}) \cdot \overline{v(\mathbf{x})} d\mathbf{x}.$$

The second argument in the scalar product is complex conjugated, and the notation will be used for volume as well as boundary terms indicated by the used subset.

2.1. Weak Derivatives and Weak Formulation

The Helmholtz equation stated in the strong form in Definition 1 requires differentiability twice, meaning $w \in C^2(\Omega)$ and furthermore the physical parameters need to satisfy $\varepsilon \in C(\overline{\Omega}), \mu^{-1} \in C^1(\overline{\Omega})$. The Robin boundary condition additionally requires that $w \in C^1(\overline{\Omega})$. A generalisation of a strong derivative is the so-called weak derivative.

2.1.1. Weak Derivative

For a strong derivative, there holds for smooth functions with zero boundary value $\phi \in C_0^\infty(\Omega)$ due to the product rule and therefore partial integration

$$(f', \phi)_\Omega = -(f, \phi')_\Omega.$$

This property is used to define the weak derivative.

Definition 4 (Weak Derivative)**.** *Let f be an integrable function, then the integrable function v is called its weak derivative, if and only if there holds*

$$(v, \phi)_\Omega = -(f, \phi')_\Omega$$

for all $\phi \in C_0^\infty(\Omega)$.

In the following, the weak derivative v of f will be naturally denoted by f' although it does not represent a classical derivative any more and can only be considered applied in integral form. For weak derivatives, there is the rule of partial integration, also called the

Gauss-integration formula. Consider an integrable function f such that the weak divergence exists, then there holds

$$\int_{\partial\Omega} f \cdot \mathbf{n} d\mathbf{s} = \int_{\Omega} \operatorname{div}(f) d\mathbf{x}$$

and further for weakly derivable functions u and weakly derivable vector functions $\mathbf{v}$ there holds the rule of partial integration

$$(u, \mathbf{v} \cdot \mathbf{n})_{\partial\Omega} = (\nabla u, \mathbf{v})_{\Omega} + (u, \operatorname{div} \mathbf{v})_{\Omega}.$$

With weak derivatives the following well-established spaces are introduced.

Definition 5. *Assume Ω to be a Lipschitz domain, then the following spaces can be defined:*

$$L^2(\Omega) := \{u : \|u\|_{L^2(\Omega)} := \sqrt{(u,u)_{\Omega}} < \infty\},$$
$$H^1(\Omega) := \{u : u \in L^2(\Omega), \nabla u \in [L^2(\Omega)]^d\}.$$

The associated norm is given by

$$\|u\|_{H^1(\Omega)} := \sqrt{(u,u)_{\Omega} + (\nabla u, \nabla u)_{\Omega}}.$$

2.1.2. Weak Formulation

With the weak derivative, it is possible to state the Helmholtz equation in integral form, which is also called its weak formulation. Assuming that the Helmholtz equation holds in the strong form then multiplying Equation (1.2a) with an integrable and derivable function v and integrating over the domain leads to

$$-(\operatorname{div}(\mu^{-1}\nabla u), v)_{\Omega} - \kappa^2(\varepsilon u, v)_{\Omega} = (f, v)_{\Omega}.$$

Further assuming that v is weakly derivable, partial integration on the first term leads to

$$(\mu^{-1}\nabla u, \nabla v)_{\Omega} - \kappa^2(\varepsilon u, v)_{\Omega} - (\mu^{-1}\nabla u \cdot \mathbf{n}, v)_{\partial\Omega} = (f, v)_{\Omega}.$$

The naturally occurring boundary term is replaced by the Robin boundary condition (1.2b) giving

$$(\mu^{-1}\nabla u, \nabla v)_{\Omega} - \kappa^2(\varepsilon u, v)_{\Omega} - j\kappa \left(\sqrt{\frac{\varepsilon}{\mu}} u, v\right)_{\partial\Omega} = (f, v)_{\Omega} + \left(\frac{1}{\sqrt{\mu}} g, v\right)_{\partial\Omega} \tag{2.1}$$

for all integrable and derivable functions v. Interestingly, the left hand side of the equation is now a complex symmetric sesquilinear form and the right hand side is an anti-linear form. This is a very favourable property due to the existing theory for FEM. Additionally, it can be seen that before, the necessary derivatives were solely on the solution u and for v, only integrateablity was necessary. After partial integration, these roles have changed. Now, u and v require the same regularity and live in the same space of integrable functions. The solution does not require $C^2(\Omega)$ any more. That is the reason why it is called the weak formulation, and u is only the weak solution of the Helmholtz equation. The sesquilinear form above is probably the most classic weak formulation of the Helmholtz equation with Robin boundary condition and the FEM, including the established theory, will be illustrated in this example.

2.2. Existence and Uniqueness of Weak Solutions

In this section, the question of the existence and uniqueness of weak solutions for weak formulations is answered. Standard techniques are shortly motivated, highlighted and expanded on. Considering the following general setup, further requirements will be stated on demand. Let X and Y be some Banach spaces such that the sesquilinear form

$$B : X \times Y \mapsto \mathbb{C}$$

is well defined in the sense of continuity in the arguments

$$|B(u, v)| \leq C_c \|u\|_X \|v\|_Y,$$

with the continuity constant $C_c > 0$. For an anti-linear form f in the space of continuous anti-linear functionals is $u \in X$ a weak solution if and only if

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

for all $v \in Y$. When the spaces X and Y coincide, then the well-established theory for coercive, also called elliptic sesquilinear forms, may be applied.

2.2.1. Coercive Sesquilinearform

Consider a continuous sesquilinear form

$$B : X \times X \mapsto \mathbb{C}.$$

A SLF is called coercive if there exists a constant $\gamma > 0$ so that

$$\gamma \|u\|_X^2 \leq |B(u, u)|.$$

The estimate goes exactly in the other direction than the continuity. There holds the following existence and uniqueness result.

Theorem 6 (Lax-Milgram)**.** *Let B be a continuous and coercive sesquilinear form on the Hilbert space X. For each continuous anti-linear form f on X there exists a unique solution $u \in X$ of the problem*

$$B(u, v) = f(v) \qquad \forall v \in X.$$

There holds the stability estimate

$$\|u\|_X \leq \frac{1}{\gamma} \sup_{v \in X \setminus \{0\}} \frac{|f(v)|}{\|v\|_X}.$$

Proof. See [Alt12]. □

2.2.2. Inf-Sup Stability

The following technique has weaker requirements for the existence and uniqueness of a solution. The coercivity is weakened into the following inf-sup conditions

$$\inf_{u\in X\setminus\{0\}} \sup_{v\in Y\setminus\{0\}} \frac{|B(u,v)|}{\|u\|_X\|v\|_Y} \geq \alpha_U > 0, \tag{2.2a}$$

$$\inf_{v\in Y\setminus\{0\}} \sup_{u\in X\setminus\{0\}} \frac{|B(u,v)|}{\|u\|_X\|v\|_Y} \geq \alpha_E > 0. \tag{2.2b}$$

Equation (2.2a) implies the uniqueness of a solution $u \in X$ of the problem

$$B(u,v) = f(v) \qquad \forall v \in Y$$

for a given continuous sesquilinear form B on the Banach spaces X, Y and continuous anti-linear form f on Y. The second Equation (2.2b) implies the surjectivity of the operator defined by $< B(u), v >:= B(u,v)$.

Theorem 7. *Let B be a continuous sesquilinear form on the Banach space $X \times Y$. For each continuous anti-linear form f on Y there exists a unique solution $u \in X$ of the problem*

$$B(u,v) = f(v) \qquad \forall v \in Y$$

if and only if the two inf-sup-*conditions* (2.2a) *and* (2.2b) *are satisfied and then there holds the stability estimate*

$$\|u\|_X \leq \frac{1}{\alpha} \sup_{v\in Y\setminus\{0\}} \frac{|f(v)|}{\|v\|_Y}.$$

2.3. Existence and Uniqueness for the Helmholtz Equation with Robin Boundary Conditions

In Section 2.2, an overview of the theoretical existence and uniqueness theory has been given in the context of Hilbert spaces. Now, it is interesting in which specific space the solution of the Helmholtz equation should be found. Motivated by the weak formulation (2.1) the solution needs to be square integrable, and also its gradient needs to be square integrable. Additionally, the trace on the domain boundary needs to be square and integrable. For this boundary term, u needs to be in $H^s(\Omega)$ with $s > 1/2$ at least, which is satisfied by $u \in H^1(\Omega)$. The following to the H^1 norm equivalent norm is more suited for the Helmholtz equation, because it incorporates the wave number.

Definition 8.

$$\|u\|_{H^1_\kappa(\Omega)} := \sqrt{\kappa^2(u,u)_\Omega + (\nabla u, \nabla u)_\Omega}.$$

For the Helmholtz Equation with homogeneous Robin Boundary Conditions on Lipschitz domains, the following wave number explicit stability has been proven in [EM12, Theorem 2.4].

Theorem 9. *Let Ω be a bounded Lipschitz domain. Then there exists $C(\Omega) > 0$ (independent of κ) such that for $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ the solution $u \in H^1(\Omega)$ of (1.2a),(1.2b) satisfies*

$$\|u\|_{H^1_\kappa(\Omega)} \leq C(\Omega) \left(\kappa^2 \|g\|_{L^2(\partial\Omega)} + \kappa^{5/2} \|f\|_{L^2(\Omega)} \right).$$

For convex or star-shaped smooth domains, better estimates are possible, which are reflected by the following inf-sup-conditions taken from [EM12, Theorem 2.5].

Theorem 10. *Let Ω be a bounded Lipschitz domain. Then there exists $C > 0$ (independent of κ) such that the sesquilinear form, referred to as B, of (2.1) satisfies*

$$\inf_{u \in H^1(\Omega)\setminus\{0\}} \sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{\Re B(u,v)}{\|u\|_{H^1_\kappa(\Omega)} \|v\|_{H^1_\kappa(\Omega)}} \geq C\kappa^{-7/2}.$$

Furthermore, for every $f \in (H^1(\Omega))'$ and $g \in H^{-1/2}(\partial\Omega)$ the problem (2.1) is uniquely solvable, and its solution $u \in H^1(\Omega)$ satisfies the a priori bound

$$\|u\|_{H^1_\kappa(\Omega)} \leq C\kappa^{7/2} \left(\|f\|_{(H^1(\Omega))'} + \|g\|_{H^{-1/2}(\partial\Omega)} \right).$$

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

$$\inf_{u \in H^1(\Omega)\setminus\{0\}} \sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{\Re B(u,v)}{\|u\|_{H^1_\kappa(\Omega)} \|v\|_{H^1_\kappa(\Omega)}} \geq C\kappa^{-1}.$$

The analysis is based upon the following Rellich identities.

Definition 11 (Rellich Identities)**.** *For any function $v \in H^2(\Omega)$ on a Lipschitz domain Ω there hold*

$$\begin{aligned} d\|v\|^2_{L^2(\Omega)} + 2\Re(v, \mathbf{x}\cdot\nabla v)_\Omega &= (\mathbf{x}\cdot\mathbf{n}, |v|^2)_{\partial\Omega}, \\ (d-2)\|\nabla v\|^2_{L^2(\Omega)} + 2\Re(\nabla v, \nabla(\mathbf{x}\cdot\nabla v))_\Omega &= (\mathbf{x}\cdot\mathbf{n}, |\nabla v|^2)_{\partial\Omega}. \end{aligned}$$

The proof can be found in [FW09, Lemma 4.1]. Considering the test function $v = \mathbf{x}\cdot\nabla u$ assumed to be in $H^1(\Omega)$ and with the Rellich identities above, it is possible to show for the Helmholtz solution the following stability.

Theorem 12. *Assume Ω to be a star-shaped Lipschitz domain, then there hold the following additional stability estimates*

$$\begin{aligned} \|u\|_{H^1_\kappa(\Omega)} &\leq C(\Omega) \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)} \right), \\ |u|_{H^2(\Omega)} := \sqrt{(\nabla^2 u, \nabla^2 u)_\Omega} &\leq C(\Omega) \left((1+\kappa) \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)} \right) + \|g\|_{H^{1/2}(\partial\Omega)} \right), \end{aligned}$$

where $\nabla^2 u$ denotes the Hessian matrix of u.

For the proof see [Mel95] or [MPS13, Remark 2.6]. Over the years, further results have been established with the help of the special test function $\mathbf{x} \cdot \nabla u$.

The stability and regularity for the Helmholtz equation with Robin boundary conditions can generally be written as:

Definition 13. *The Helmholtz equation with Robin boundary conditions is H^s-regular if for all $0 \leq t \leq s$ there exist constants $C_{f,t}(\Omega, \kappa) > 0$ and $C_{g,t}(\Omega, \kappa) > 0$ so that there holds*

$$|u|_{H^t(\Omega)} \leq C_{f,t}(\Omega, \kappa)\|f\|_{L^2(\Omega)} + C_{g,t}(\Omega, \kappa)\|g\|_{L^2(\partial\Omega)}.$$

2.3.1. The Dual Helmholtz Equation

As was already introduced, apart from the primal Helmholtz equation, there also exists the mixed formulation by introducing a vector field $\sigma := \frac{1}{j\kappa}\nabla u$. It turns out it is possible to eliminate the scalar field u and end up with a formulation purely in σ. This formulation will be called the dual Helmholtz equation. Considering Equation (1.3c) on the boundary in the mixed formulation and multiplying it with suitable vector-valued functions $\tau \cdot \mathbf{n}$ gives

$$\left(\sqrt{\frac{\mu}{\varepsilon}}\sigma \cdot \mathbf{n}, \tau \cdot \mathbf{n}\right)_{\partial\Omega} - (u, \tau \cdot \mathbf{n})_{\partial\Omega} = \frac{1}{j\kappa}\left(\frac{1}{\sqrt{\varepsilon}}g, \tau \cdot \mathbf{n}\right)_{\partial\Omega}.$$

Applying the Green integration rule gives

$$(u, \tau \cdot \mathbf{n})_{\partial\Omega} = (\nabla u, \tau)_\Omega + (u, \operatorname{div}\tau)_\Omega$$

and replacing the gradient via Equation (1.3a) and u with Equation (1.3b) leads to

$$\begin{aligned}(\nabla u, \tau)_\Omega &= j\kappa(\mu\sigma, \tau)_\Omega,\\ (u, \operatorname{div}\tau)_\Omega &= \frac{1}{j\kappa}\left(\frac{1}{\varepsilon}\operatorname{div}\sigma, \operatorname{div}\tau\right)_\Omega - \frac{1}{\kappa^2}\left(\frac{1}{\varepsilon}f, \operatorname{div}\tau\right)_\Omega.\end{aligned}$$

By combining these terms the weak dual formulation is

$$\begin{aligned}\frac{j}{\kappa}\left(\frac{1}{\varepsilon}\operatorname{div}\sigma, \operatorname{div}\tau\right)_\Omega - j\kappa(\mu\sigma, \tau)_\Omega + \left(\sqrt{\frac{\mu}{\varepsilon}}\sigma \cdot \mathbf{n}, \tau \cdot \mathbf{n}\right)_{\partial\Omega}&\\ = -\frac{1}{\kappa^2}\left(\frac{1}{\varepsilon}f, \operatorname{div}\tau\right)_\Omega + \frac{1}{j\kappa}\left(\frac{1}{\sqrt{\varepsilon}}g, \tau \cdot \mathbf{n}\right)_\Omega.&\end{aligned}$$

A suitable space such that the sesquilinear form is well defined requires the divergence in $L^2(\Omega)$ and also the normal trace in $L^2(\partial\Omega)$.

Definition 14. *Assume Ω to be a Lipschitz domain, then the following space can be defined*

$$H(\operatorname{div}, \Omega) := \{\sigma : \sigma \in [L^2(\Omega)]^d, \operatorname{div}(\sigma) \in L^2(\Omega)\}.$$

Note that for functions $\sigma \in H(\operatorname{div}, \Omega)$, the normal component on the boundary is only in $H^{-1/2}(\partial\Omega)$. The appropriate space for the dual weak formulation is

$$\sigma \in H(\operatorname{div}, \Omega) \cap H^s(\Omega)$$

with $s > 1/2$.

2.4. Discretisation

The results in Section 2.3 give existence and uniqueness of solution and stability, but they do not give any idea on how this solution actually can be calculated. Searching for a strong solution is only possible in some special cases, and in this work, the focus is on the numerical calculation of an approximation to the solution. First, the space $H^1(\Omega)$ has infinite dimension. A numerical scheme in this space will be more than challenging. Therefore the approach of the FEM is to use a finite-dimensional subspace of X and Y

$$X_h \subset X, \qquad Y_h \subset Y,$$

with $\dim X_h < \infty$ and $\dim Y_h < \infty$. Choosing bases of X_h and Y_h

$$\operatorname{span}(u_1, \ldots, u_N) = X_h, \qquad \operatorname{span}(v_1, \ldots, v_M) = Y_h,$$

the discrete problem is

$$B(u_h, v_h) = f(v_h)$$

for all $v_h \in Y_h$. Due to the linearity, the requirement for all $v_h \in Y_h$ can be replaced by the requirement on the base vectors v_l

$$B(u_h, v_l) = f(v_l)$$

for all $v_l, l = 1, \ldots, M$. Representing u_h as a linear combination of base vectors

$$u_h = \sum_{i=1}^{N} x_i u_i$$

with coefficients $x_i \in \mathbb{C}$ and then inserting into the problem above yields

$$B(u_h, v_l) = B\left(\sum_{i=1}^{N} x_i u_i, v_l\right) = \sum_{i=1}^{N} B(u_i, v_l) x_i = f(v_l).$$

Writing the coefficients $B_{l,i} := B(u_i, v_l)$ into a matrix and $f_l := f(v_l)$ into a vector then leads to a system of linear equations with the representation

$$Bx = f.$$

This is the basic idea of FEM. A very important property of it is the following so-called Galerkin orthogonality. Let u_h be the solution of the discretised problem, then there holds

$$B(u - u_h, v_h) = f(v_h) - f(v_h) = 0, \qquad \forall v_h \in Y_h.$$

Therefore FEM is a Galerkin method. In the case of the existence of the continuous solution u, it is not given that the discrete solution u_h also exists, which correlates to the invertibility of the matrix B. In the following lines, a summary of the possible reasons for the existence of discrete solutions is given.

Coercive Weak Formulation The coercivity and continuity of the sesquilinear form in the space X directly transfers into the discrete subspace X_h. Therefore, the discrete solution exists and is unique.

Discrete Inf-Sup-Condition Contrary to coercivity, the inf-sup-conditions are not directly transferable to discrete subspaces and need to be proven separately. An advantage is that the subspace has finite dimension, and if the dimensions of X_h and Y_h are equal, then only existence or uniqueness needs to be shown because, for a square linear finite-dimensional problem, one implies the other. A way to derive the discrete inf-sup-condition from the continuous inf-sup-condition is to construct a Fortin operator, which is defined as the following.

Definition 15 (Fortin Operator)**.** *A linear operator $F_h : X \mapsto X_h$ is called Fortin operator if for any $u \in X$ there holds*

$$\begin{aligned} B(F_h u, v_h) &= B(u, v_h) && \forall v_h \in Y_h, \\ \|F_h u\|_X &\leq C_F \|u\|_X, \end{aligned}$$

with a positive constant $C_F > 0$.

The idea of the proof goes along the lines

$$\begin{aligned} \inf_{v_h \in Y_h \setminus \{0\}} \sup_{u_h \in X_h \setminus \{0\}} \frac{|B(u_h, v_h)|}{\|u_h\|_X \|v_h\|_Y} &\geq \inf_{v_h \in Y_h \setminus \{0\}} \sup_{u \in X \setminus \{0\}} \frac{|B(F_h u, v_h)|}{\|F_h u\|_X \|v_h\|_Y} \\ &\geq \inf_{v_h \in Y_h \setminus \{0\}} \sup_{u \in X \setminus \{0\}} \frac{|B(u, v_h)|}{C_F \|u\|_X \|v_h\|_Y} \\ &\geq \inf_{v \in Y \setminus \{0\}} \sup_{u \in X \setminus \{0\}} \frac{|B(u, v)|}{C_F \|u\|_X \|v\|_Y} \geq \frac{\alpha_E}{C_F} > 0. \end{aligned}$$

2.4.1. Triangulation and Mesh

The next question which needs to be answered is what discrete subspace will be used in the FEM. The domain Ω is split into non-overlapping simplices T, very often triangles in 2D and tetrahedra in 3D. The set containing all simplices will be denoted by $\mathcal{T}$ and there holds

$$\cup_{T \in \mathcal{T}} T = \overline{\Omega}, \qquad \forall T_k, T_l \in \mathcal{T} : k \neq l \Rightarrow \mathring{T}_k \cap \mathring{T}_l = \emptyset.$$

The simplices are also called finite elements, from which the method has its name. This is the simplest form of a triangulation leading to a mesh. The first and most prominent question should be what kind of restrictions are given for domains, and the very pleasant answer is that good mesh generators like NetGen [Sch97] can cope with very complicated domains. Curved elements are possible, as well as a wide variety of shapes and forms. One crucial property, which is often required and will also be used in this work is shape regularity. This means that there do not exist arbitrarily small angles in elements, and they can be bounded from below by a positive constant. The boundary of an element consists

either of three edges in 2D or four faces in 3D which will be called facets and denoted by F:

$$\partial T = \cup_{i=1}^{d+1} F_i, \qquad d := \dim(\Omega). \tag{2.3}$$

The set of all facets is $\mathcal{F}$. The set of facets on the domain boundary $\partial\Omega$ is defined as

$$\mathcal{F}_O := \{F \in \mathcal{F} : \operatorname{codim}(F \cap \partial\Omega) = 1\}$$

and the set of all interior facets is

$$\mathcal{F}_I := \mathcal{F} \setminus \mathcal{F}_O.$$

Two different elements $T_k, T_l \in \mathcal{T}$ are adjacent if and only if there exists an interior facet $F_{k,l} \in \mathcal{F}_I$ so that

$$T_k \cap T_l = F_{k,l}.$$

2.4.2. Polynomial Spaces

The idea of FEM is to use piecewise polynomials of a certain polynomial degree on each element T as a discrete approximation space. In most cases, this is insufficient because the discrete space needs to be a subspace of the continuous space. Therefore, certain continuities need to be satisfied. For the different spaces, they are the following:

$$\begin{aligned} L^2(\Omega) &\Rightarrow \text{ discontinuous functions allowed,} \\ H^1(\Omega) &\Rightarrow \text{ continuous,} \\ H(\operatorname{div}, \Omega) &\Rightarrow \text{ normal component continuous over facets.} \end{aligned}$$

The local polynomial spaces are glued together according to the required continuities. For a detailed explanation of the construction of the following discrete spaces, see [SZ05]. Therein, a rigorous explanation, introduction and construction of the spaces below are given. A polynomial space can usually be characterised on a mesh by specifying a discrete space and functionals, which both depend on the polynomial degree $p \in \mathbb{N}$.

Nodal FEM Spaces

The polynomial space of order $p \geq 1$ for $H^1(\Omega)$ will be denoted as $\mathcal{P}^p(\Omega)$ where

$$\mathcal{P}^p(\Omega) \subset H^1(\Omega) \qquad \text{and} \qquad \mathcal{P}^p(\Omega)|_T = \mathcal{P}^p(T), \forall T \in \mathcal{T}$$

holds for all elements $T \in \mathcal{T}$. For the gradient, there follows

$$\nabla v|_T \in [\mathcal{P}^{p-1}(T)]^d \qquad \forall v \in \mathcal{P}^p(\Omega), \forall T \in \mathcal{T}.$$

Due to the required continuity, only polynomial degrees larger than one are possible and for the lowest order case $p = 1$, each basis function spanning the discrete space can be associated with a unique vertex in the mesh, giving the space its name.

Raviart-Thomas and Brezzi-Douglas-Marini Spaces

For the vector-valued discrete space of $H(\mathrm{div}, \Omega)$, there will be two possible spaces introduced. The first space is the Brezzi-Douglas-Marini space $\mathcal{BDM}^p(\Omega)$ for the order $p \geq 0$ with the property

$$\mathcal{BDM}^p(\Omega) \subset H(\mathrm{div}, \Omega)$$

and on each element it spans the vector valued polynomial space

$$\mathcal{BDM}^p(\Omega)|_T = [\mathcal{P}^p(T)]^d, \qquad \forall T \in \mathcal{T}.$$

For the divergence, there holds

$$\mathrm{div}(\tau)|_T \in \mathcal{P}^{p-1}(T), \qquad \forall \tau \in \mathcal{BDM}^p(\Omega), \forall T \in \mathcal{T}.$$

The second space is the Raviart-Thomas space $\mathcal{RT}^p(\Omega)$ of order $p \geq 0$ with the property

$$\mathcal{RT}^p(\Omega) \subset H(\mathrm{div}, \Omega).$$

On each element T, the function τ in this discrete space has the form

$$\tau = \mathbf{p} + \mathbf{x}q, \qquad \mathbf{p} \in [\mathcal{P}^p(T)]^d, \qquad q \in \mathcal{P}^p(T).$$

It is larger than the $\mathcal{BDM}^p(\Omega)$ space but smaller than $\mathcal{BDM}^{p+1}(\Omega)$, because

$$\mathcal{BDM}^p(\Omega) \subset \mathcal{RT}^p(\Omega) \subset \mathcal{BDM}^{p+1}(\Omega).$$

It is constructed such that there holds for the divergence

$$\mathrm{div}(\tau) \in \mathcal{P}^p(T) \qquad \forall \tau \in \mathcal{RT}^p(\Omega), \forall T \in \mathcal{T}.$$

It even holds that the divergence spans the whole element-wise scalar polynomial space $\mathcal{P}^p(T)$.

Legendre Spaces

The discrete space $\mathcal{Q}^p(\Omega)$ of order $p \geq 0$ for $L^2(\Omega)$ has no required continuities therefore local discontinuous polynomial spaces in the sense of

$$q \in \mathcal{P}^p(T) \qquad \forall q \in \mathcal{Q}^p(T)$$

are used. Although any polynomials could theoretically be used as a basis, for well conditioned discretisation matrices, it is advantageous to use L^2-orthogonal polynomials, which are in one dimension exactly the name-giving Legendre polynomials.

2.5. Quasi-optimality, Best Approximation and Error Estimates

When the question of existence and uniqueness of continuous solution u and discrete solution u_h is answered, then, the final question of convergence remains in the sense of bounding the error

$$\|u - u_h\|_Z \leq ?$$

in some Z-norm, which may be the X-norm, but does not need to be. A numerical method is called optimal if there holds

$$\|u - u_h\|_Z = \inf_{v_h \in X_h} \|u - v_h\|_Z.$$

The right hand side is the best possible approximation of the solution u in the chosen discrete subspace and Z-norm, and the equality means that the discrete solution is already the best possible approximation. If this property does not hold with an equality sign, but rather with a positive constant $C > 0$ so that

$$\|u - u_h\|_Z \leq C \inf_{v_h \in X_h} \|u - v_h\|_V$$

then the method is only quasi-optimal, which is sufficient in most cases and additionally, the norm on the right hand side might be different for the Z-norm. Left hand side and right hand side are considered in different norms. Very often, the left norm is a weaker norm than the right norm.

Next, the right hand side of the quasi-optimality estimates needs to be bounded. The discrete space X_h satisfies some kind of approximation property, which naturally also depends on the properties of the continuous solution u

$$\inf_{v_h \in X_h} \|u - v_h\|_V \leq C(h)|u|_W$$

with some on the discrete space dependent constant $C(h) > 0$ and another possibly different W-seminorm. If the continuous solution is stable in the W-seminorm

$$|u|_W \leq C_{stab}\|f\|_{Y'}$$

then the best approximation can be further bounded by

$$\inf_{v_h \in X_h} \|u - v_h\|_V \leq C(h)C_{stab}\|f\|_{Y'}.$$

Note that the stability constant is usually independent of mesh parameters but not physical parameters specific to the Helmholtz problem, such as material coefficients and the domain. Convergence of a discrete method is now considered the following way. Assuming there exists a series of discrete spaces $(X_{h_i})_{i=1,\ldots,\infty}$ so that

$$\lim_{i \to \infty} C(h_i) = 0$$

then the numerical scheme converges. This property is, for most methods, insufficient information and the explicit form of the constant $C(h)$ is quite important.

In the following, the quasi-optimality is stated for the introduced standard theories.

Coercive sesquilinear form Coercive sesquilinear forms automatically satisfies the following quasi-optimality due to the coercivity, Galerkin-orthogonality and continuity

$$\begin{aligned}\gamma\|u-u_h\|_X^2 &\leq |B(u-u_h,u-u_h)| = |B(u-u_h,u-v_h)+B(u-u_h,v_h-u_h)|\\ &= |B(u-u_h,u-v_h)| \leq C_c\|u-u_h\|_X\|u-v_h\|_X\end{aligned}$$

resulting in

$$\|u-u_h\|_X \leq \frac{C_c}{\gamma}\inf_{v_h\in X_h}\|u-v_h\|_X.$$

Quite remarkably, coercive SLFs are optimal in the induced B-norm because in this norm, the continuity constant and the coercivity constant are equal to one.

The theory can be relaxed a little by accepting different norms for the continuity estimate giving

$$\gamma\|u-u_h\|_X^2 \leq C_c\|u-u_h\|_Y\|u-v_h\|_Z.$$

In this setting, the following connection between the X-norm and Y-norm is required

$$\|u-u_h\|_Y \leq C_{XY}\|u-u_h\|_X.$$

This means that the Y-norm must be weaker than the X-norm, giving

$$\|u-u_h\|_X \leq \frac{C_cC_{XY}}{\gamma}\inf_{v_h\in X_h}\|u-v_h\|_Z.$$

Inf-Sup Stable sesquilinear form Discretely inf-sup stable sesquilinear forms are quasi-optimal by

$$\|u-u_h\|_X \leq \|u-v_h\|_X + \|u_h-v_h\|_X$$

and considering

$$\alpha_U \leq \inf_{q_h\in X_h\setminus\{0\}}\sup_{w_h\in Y_h\setminus\{0\}}\frac{|B(q_h,w_h)|}{\|q_h\|_X\|w_h\|_Y} \leq \sup_{w_h\in Y_h\setminus\{0\}}\frac{|B(u_h-v_h,w_h)|}{\|u_h-v_h\|_X\|w_h\|_Y}$$

giving with the Galerkin orthogonality and continuity

$$\alpha_U\|u_h-v_h\|_X \leq \sup_{w_h\in Y_h\setminus\{0\}}\frac{|B(u_h-v_h,w_h)|}{\|w_h\|_Y} = \sup_{w_h\in Y_h\setminus\{0\}}\frac{|B(u-v_h,w_h)|}{\|w_h\|_Y} \leq C_c\|u-v_h\|_X.$$

Combining with the term above leads to

$$\|u-u_h\|_X \leq \left(1+\frac{C_c}{\alpha_U}\right)\inf_{v_h\in X_h}\|u-v_h\|_X.$$

In the same fashion as for the coercive case, the norm on the right hand side can be generalised into

$$\|u-u_h\|_X \leq \inf_{v_h\in X_h}\left(\|u-v_h\|_X + \frac{C_c}{\alpha_U}\|u-v_h\|_Z\right).$$

Note that the continuous inf-sup conditions are not required, just the existence of a continuous solution, then the discrete condition already gives quasi-optimality.

Schatz Argument The continuous Helmholtz equation falls into the inf-sup stable regime if Robin boundary conditions are prescribed. To the author's knowledge, the discretised form of Equation (2.1) does not satisfy a discrete inf-sup condition. Another technique tailored to the Helmholtz equation is required: the Schatz argument. It was first introduced by Alfred H. Schatz in 1974 in "An observation concerning Ritz-Galerkin methods with indefinite bilinear forms" [Sch74]. Let us consider the following sesquilinear form in this small part

$$B(u,v) := (\nabla u, \nabla v)_\Omega - \kappa^2 (u,v)_\Omega.$$

It is well known that the negative L^2 term in the sesquilinear form is the issue. Considering the new sesquilinear form

$$B_+(u,v) := B(u,v) + 2\kappa^2 (u,v)_\Omega$$

then this sesquilinear form is coercive and therefore

$$\|u-u_h\|^2_{H^1_\kappa(\Omega)} = |B_+(u-u_h, u-u_h)| \leq |B(u-u_h, u-u_h)| + \kappa^2 \|u-u_h\|^2_{L^2(\Omega)}.$$

For the first term, we use the Galerkin orthogonality to replace u_h with an arbitrary v_h and then apply continuity

$$|B(u-u_h, u-u_h)| = |B(u-u_h, u-v_h)| \leq C_c \|u-u_h\|_{H^1_\kappa(\Omega)} \|u-v_h\|_{H^1_\kappa(\Omega)}.$$

The second term requires the following duality argument, also called the Aubin-Nitsche trick. Consider the solution of the adjoint Helmholtz problem

$$B(v,w) = (v, u-u_h)_\Omega,$$

with the error $u-u_h$ as the excitation, then by setting $u-u_h$ as test function v this leads to

$$\|u-u_h\|^2_{L^2(\Omega)} = B(u-u_h, w).$$

The Galerkin orthogonality of $u-u_h$ enables pushing in an arbitrary discrete function, for example some kind of interpolant $I_h(w)$ and applying the continuity again gives

$$\|u-u_h\|^2_{L^2(\Omega)} \leq C_c \|u-u_h\|_{H^1_\kappa(\Omega)} \|w - I_h(w)\|_{H^1_\kappa(\Omega)}.$$

The second term represents an approximation term of the adjoint solution. Under the assumption of H^2-regularity and stability, there holds

$$\|w - I_h(w)\|_{H^1_\kappa(\Omega)} \leq Ch|w|_{H^2(\Omega)} \leq C(\Omega,\kappa) h \|u-u_h\|_{L^2(\Omega)}$$

giving

$$\kappa \|u-u_h\|_{L^2(\Omega)} \leq C_c C(\Omega,\kappa) \kappa h \|u-u_h\|_{H^1_\kappa(\Omega)}.$$

Combining the results together and if $C_c C(\Omega) C(1+\kappa)\kappa h < 1$ holds, then the right hand side L^2-term can be absorbed into the left hand side leading to

$$\|u - u_h\|_{H^1_\kappa(\Omega)} \leq \frac{C_c}{1 - C_c^2 C(\Omega,\kappa)^2 \kappa^2 h^2} \inf_{v_h \in X_h} \|u - v_h\|_{H^1},$$

the quasi-optimality with a pollution effect and a certain mesh resolution condition for the asymptotic regime. This technique was first developed in another context. The Poisson equation is coercive, and therefore, $\mathcal{O}(h^p)$ convergence order in the H^1 norm can be expected if the solution is sufficiently smooth, but simulations showed that the L^2-error converged with a higher order of $\mathcal{O}(h^{p+1})$. This can also be seen in the estimate for the L^2-error above. It also converges with an order higher. Another issue can be seen in this analysis. The existence of a unique discrete solution is non-trivial, and the analysis above only covers the existence in the asymptotic range, which depends on the problem-specific, possibly unknown constants.

Polution Effect The analysis by the Schatz argument above also highlights another fascinating behaviour of the Helmholtz equation, the pollution effect. The Whittaker-Nyquist–Shannon sampling theorem, first published by Edmund Taylor Whittaker in 1915 in the article "On the functions which are represented by the expansions of the interpolation-theory" [Whi15], states that to resolve a sine wave a certain number of points is required. In FEM, this corresponds to a certain mesh size and polynomial degree, which needs to satisfy an indirect correlation with the wave number $h^p \approx \kappa^{-1}$. The approximation error of FEM satisfies exactly this relation, and now one might expect that the FEM solution also satisfies this. This means that if the wave number is increased and in the same proportion, the mesh size is decreased such that $\kappa^p h = \mathcal{O}(1)$ the FEM solution accuracy should stay the same. This is not the case, with increasing wave number, the mesh size needs to satisfy a harder requirement $\kappa^2 h^p = \mathcal{O}(1)$. Jens Markus Melenk relaxed this condition under consideration of h, p-FEM analysis in his thesis [Mel95]. Still, the pollution effect remains and is compensated by increasing polynomial degree. This also implies that for high-frequency simulations, a small mesh size is required, leading to large system matrices. One might think that there could be a numerical regime such that the pollution effect is eliminated, but the work by Babuška and Sauter, "Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Number?" [BS97] gives an answer. For one dimension, it is possible to consider FEM without pollution, but for all higher dimensions, there is always pollution. The name of the game in FEM formulations for Helmholtz is, how high is the pollution effect, and how efficiently can the possibly large system of linear equations be solved?

2.5.1. Projection Based Error Estimation

In the last couple of years, another strategy has been established. The projection-based error estimation, see for example [CGS10]. Only the existence of a continuous unique solution is required, and then an appropriate projection $P(u)$ of this solution into the discrete space is used to establish convergence by using

$$\|u - u_h\|_X \leq \|u - P(u)\|_X + \|P(u) - u_h\|_X.$$

The first term is independent of the discrete solution, and only some kind of approximation property of the projection is required

$$\|u - P(u)\|_X \leq C(h)\|u\|_Z$$

and the second term has the advantage that $P(u) - u_h$ lives in the discrete space X_h and therefore it can be analysed by using techniques which only hold in finite-dimensional space. Considering, for example, the Aubin-Nitsche trick for the projected error

$$B(v, w) = (v, P(u) - u_h)_\Omega \qquad \forall v \in X$$

and now choosing $v = P(u) - u_h \in X_h$ leads to

$$B(P(u) - u_h, w) = \|P(u) - u_h\|^2_{L^2(\Omega)}.$$

This might not seem to be helpful, but after introducing the same projection for the adjoint solution w there holds

$$\|P(u) - u_h\|^2_{L^2(\Omega)} = B(P(u) - u_h, w - P(w)) + B(P(u) - u_h, P(w)).$$

Usually, at some point, continuity estimates are required, but due to the projection, the first term may be estimated by weaker norms V and W

$$B(P(u) - u_h, w - P(w)) \leq \|P(u) - u_h\|_V \|w - P(w)\|_W.$$

Now, the term for the adjoint problem can usually be bounded with the stability constant by

$$\|w - P(w)\|_W \leq C(h)\|w\|_Z \leq C(h)C_{stab}\|P(u) - u_h\|_{L^2(\Omega)}.$$

There are two possibilities for the other part. Either the V-norm can be bounded by the L^2-norm

$$\|P(u) - u_h\|_V \leq C\|P(u) - u_h\|_{L^2(\Omega)},$$

then an asymptotic result like in the Schatz argument pops up, or, which will be the case in this work, it can be directly controlled by

$$\|P(u) - u_h\|_V \leq C\|u - P(u)\|_V.$$

Going back to the initial equation and looking at the second term and realising that due to the Galerkin orthogonality and $P(w) \in X_h$ there holds

$$B\left(P(u) - u_h, P(w)\right) = B(P(u) - u, P(w)).$$

Similarly to the first part, a sharper continuity estimate and the stability of the adjoint problem bounds this term. After combining these findings, the projected L^2-error can be bounded by the approximation error of the projection P.

In this work, this technique is explored and applied to the discretisation of the Helmholtz equation.

2.6. Solvers

This section gives a short introduction to possible solvers for FEM and the discrete Helmholtz equation. Solvers can, in general, be split into two categories: direct and iterative methods.

2.6.1. Direct Solvers

Direct solvers are so called because they solve a linear equation $Ax = y$ by applying the inverse A^{-1} in some numerical fashion and the resulting solution x is exact up to numerical calculation errors, for example round of errors. A variety of direct solvers have been established, and they are based upon clever factorisations of the system matrix to minimise computation time and memory costs. For FEM system matrices, special factorisations are required. The FEM matrix is sparse, meaning it only requires $\mathcal{O}(n_{DoF})$ memory, where n_{DoF} denotes the number degrees of freedom (DoF). If a factorisation would require an almost full matrix, then the memory costs would scale with $\mathcal{O}(n_{DoF}^2)$ substantially worse for large problems, therefore, factorisations which limit the fill-in are required. A great example is the PARDISO [DCDBK+16, VCKS17, KFS18] , or the Sparse Cholesky factorisation [rei71, duf83, duf17]. Memory usage and computation time depend not only on the system size but also on the sparsity pattern. Although direct solvers are already optimised for sparse matrices, the computation cost still increases too fast, such that the virtual memory runs out or the simulation time is unfeasibly long.

2.6.2. Iterative Solvers

The second class of solvers are iterative, meaning the system of linear equations $Ax = y$ is not solved exactly, but rather, after each step, an approximation x_n is calculated. After n steps when a certain accuracy $\|x - x_n\|_X \leq \varepsilon$ is reached, the iteration is stopped, and the approximation is used. The big advantage of iterative solvers is that the time and memory-consuming direct inverse A^{-1} is not required. For sparse matrices all steps have almost linear costs; for example, the usual memory consumption is $\mathcal{O}(n_{DoF} \log(n_{DoF}))$. In that regard, iterative solvers are better for large systems of linear equations. For small problems, iterative solvers might have too much overhead computations, so that they are slower than direct solvers. One very well-established iterative solver should be mentioned, the conjugate gradient (CG) method developed by Magnus Hestenes and Eduard Stiefel in 1952 [HS+52]. It is based upon orthogonal, optimal descending directions, and the theory is well established for coercive, hermitian problems. It is also a direct solver because without considering rounding errors, it will stop after n_{DoF} steps for certain. In most cases, it is used iteratively by stopping after a few iterations. The high-performance version of the CG uses a preconditioner. Instead of solving the original system of linear equations, a preconditioner C is considered, and the equivalent system $CAx = Cy$ is solved. The convergence speed of preconditioned conjugate gradient (PCG) is dictated by the spectrum of CA. The eigenvalues of $I - CA$ should be as small as possible. Therefore, the preconditioner should approximate the system matrix inverse $C \approx A^{-1}$. If the inverse is chosen as the preconditioner, the PCG stops after a single step. If the preconditioner is the identity I, PCG is just CG. The goal is to find a preconditioner C such that the eigenvalues of

$I - CA$ are small, leading to fast rates of convergence, and at the same time, the memory consumption and computation time for C should be in the same range as the application of A itself. The very powerful multigrid preconditioner was established by Achi Brandt 1977 in his paper "Multi-Level Adaptive Solutions to Boundary-Value Problems" [Bra77]. It has been proven to be optimal for coercive, hermitian problems.

The Helmholtz equation is neither coercive nor hermitian. Only complex symmetry can be established, such that effective iterative solvers and preconditioners established for coercive problems are not applicable and fail. In recent years, a lot of research has gone into finding powerful iterative solvers for "indefinite" problems such as the Helmholtz equation. A very nice compendium of such solvers can be found in Clemens Pechsteins "A Unified Theory of Non-overlapping Robin-Schwarz Methods: Continuous and Discrete, Including Cross Points" [Pec22]. In this work, the generalised minimal residual (GMRES) algorithm is applied with preconditioners based on domain decomposition in combination with Schur complement, which will be introduced in the following, to solve the discrete Helmholtz problem.

2.6.3. Static Condensation

Static condensation is a technique to reduce the size of a system of equations by condensing out internal DoF before solving the large system. The technique was formally introduced by John H. Argyris [Zum65] and Ray W. Clough [Clo60] in the late 1960s and early 1970s. Their work provided the mathematical foundation for static condensation, also known as the Guyan reduction [Guy65], named after R.J. Guyan, who independently developed a similar method. Consider a high-order discretisation of a partial differential equation with FEM, then the unknowns correspond to local basis functions, which only couple with neighbours. FEM spaces are constructed in a way so that they correspond to boundary functions or volume functions, and the volume functions only couple with each other element-wise and with boundary functions defined on the element boundary. Eliminating all volume unknowns in the linear equation by a Schur complement can be done locally and, therefore, quite efficiently. This reduces the number of unknowns substantially for higher order as the number of volume DoF is large. As usual, coercive problems also lead to coercive local problems, and therefore, static condensation is applicable. Given the Schatz Argument, this is problematic for the Helmholtz equation. Local problems correspond to Dirichlet problems, which have a real spectrum, and if only a single sub-system has κ as a resonance frequency, then the whole process of static condensation fails. It is easily seen that this method is for standard FEM discretisations for Helmholtz unstable. The problem goes even further; preconditioners often rely on local corrections and/or smoothing processes, and these may also be unstable. All these issues can be rectified by more involved methods based on the FEM. If static condensation can be applied to a formulation, then the iterative solver operates on the reduced system of linear equations.

Schur Complement

As the method of focus of this work is used with static condensation, the concept of the Schur complement, developed by its name giver Issai Schur [Sch17], is essential and will be

briefly introduced. The following text is taken from [LHS22]. Assume a system of linear equations

$$Mx = f$$

with the following block structure

$$M := \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad x := (x_1, x_2)^\top, \qquad f := (f_1, f_2)^\top.$$

A Schur complement is applied to solve the system of linear equations. It is defined as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & BD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}C & I \end{pmatrix}$$

and therefore the inverse of M can be written as

$$M^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix},$$

if D is invertible. The matrix $A - BD^{-1}C$ is always invertible if the matrices M and D are invertible. Using this approach, the system of linear equations can be solved in the three steps given below.

Preprocessing The first step is the form

$$\begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix} f = \begin{pmatrix} f_1 \\ \tilde{f}_2 \end{pmatrix} =: \tilde{f}.$$

As can be seen, only f_2 is altered.

Solving The second step is given by

$$\begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \tilde{f} = \begin{pmatrix} x_1 \\ \tilde{x}_2 \end{pmatrix} =: \tilde{x}.$$

Instead of one large system of linear equations, smaller systems of linear equations can be considered in parallel.

Postprocessing The third step consists of

$$\begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \hat{x} = \begin{pmatrix} x_1 \\ \tilde{x}_2 \end{pmatrix} = x.$$

Note that the whole concept of Schur complements can only be applied if a regular submatrix block D exists. Additionally, it is possible to use multiple Schur complements for one system of linear equations and if they do not couple with each other, then they can be done in parallel, which is exactly the case for the method in this work.

2.7. Discontinuous Galerkin Method

In deriving the weak form of the Helmholtz equation, see (2.1), a partial integration was essential in the process. This partial integration is only valid for sufficiently smooth functions, namely continuous functions which holds for the $H^1(\Omega)$ space and therefore also, the used discrete Nodal FE space must satisfy this requirement. This chapter is dedicated to answering the question of what happens if the requirement of continuity is dropped and rather reinforced in a weak sense in the sesquilinear form itself.

Starting with Equation (1.2a) and considering a mesh $\mathcal{T}$. After dropping the continuity of the test space over element boundaries ∂T, there holds the rule of partial integration

$$\left(-\operatorname{div}\left(\mu^{-1}\nabla u\right), v\right)_T = (\mu^{-1}\nabla u, \nabla v)_T - (\mu^{-1}\nabla u \cdot \mathbf{n}, v)_{\partial T}$$

on each element $T \in \mathcal{T}$ for test functions $v \in H^s_{pw}(\mathcal{T})$ in the piece-wise space

$$H^s_{pw}(\mathcal{T}) := \prod_{T\in\mathcal{T}} H^s(T), \qquad\qquad s > \frac{3}{2}.$$

Combining the boundary integrals on inner facets F_I and considering the continuity

$$\mu_+^{-1}\nabla u_+ \cdot \mathbf{n}_+ + \mu_-^{-1}\nabla u_- \cdot \mathbf{n}_- = 0,$$

with adjacent elements and variables thereon indicated by the subscript $+$ and $-$, leads to

$$\begin{aligned}
&(\mu_+^{-1}\nabla u_+ \cdot \mathbf{n}_+, v_+)_{F_I} + (\mu_-^{-1}\nabla u_- \cdot \mathbf{n}_-, v_-)_{F_I} \\
&= \frac{1}{2}\big((\mu_+^{-1}\nabla u_+ \cdot \mathbf{n}_+, v_+)_{F_I} - (\mu_-^{-1}\nabla u_- \cdot \mathbf{n}_-, v_+)_{F_I} \\
&\quad +(\mu_-^{-1}\nabla u_- \cdot \mathbf{n}_-, v_-)_{F_I} - (\mu_+^{-1}\nabla u_+ \cdot \mathbf{n}_+, v_-)_{F_I}\big) \\
&= \frac{1}{2}\left(\mu_+^{-1}\nabla u_+ \cdot \mathbf{n}_+ - \mu_-^{-1}\nabla u_- \cdot \mathbf{n}_-, v_+ - v_-\right)_{F_I} \\
&= (\{\mu^{-1}\nabla u\}_{\mathbf{n}_+}, [v]_+)_{F_I}
\end{aligned}$$

with

$$[v]_+ := v_+ - v_-$$

the jump at the boundary and

$$\{\mu^{-1}\nabla u\}_{\mathbf{n}_+} := \frac{1}{2}(\mu_+^{-1}\nabla u_+ + \mu_-^{-1}\nabla u_-) \cdot \mathbf{n}_+$$

the mean value of the flux. The remaining boundary term on $\partial\Omega$ relates to

$$(\mu^{-1}\nabla u \cdot \mathbf{n}, v)_{\partial\Omega} = j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}u, v\right)_{\partial\Omega} + \left(\frac{1}{\sqrt{\mu}}g, v\right)_{\partial\Omega}$$

due to the Robin boundary condition in Equation (1.2b).

With this definition, the weak formulation looks like

$$\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u,\nabla v)_T-\kappa^2(\varepsilon u,v)_T-\sum_{F_I\in\mathcal{F}_I}\left(\left\{\mu^{-1}\nabla u\right\}_{\mathbf{n}_+},[v]_+\right)_{F_I}-j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}u,v\right)_{\partial\Omega}$$
$$=(f,v)_\Omega+\left(\frac{1}{\sqrt{\mu}}g,v\right)_{\partial\Omega}.$$

To this point, only the continuity of the normal flux was used. This formulation is not symmetric, which is a desirable property, but terms can be added which are zero for continuous solutions u, giving

$$\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u,\nabla v)_T-\kappa^2(\varepsilon u,v)_T$$
$$-\sum_{F_I\in\mathcal{F}}\left(\left\{\mu^{-1}\nabla u\right\}_{\mathbf{n}_+},[v]_+\right)_{F_I}+\left([u]_+,\left\{\mu^{-1}\nabla v\right\}_{\mathbf{n}_+}\right)_{F_I}+j\alpha\left([u]_+,[v]_+\right)_{F_I}-j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}u,v\right)_{\partial\Omega}$$
$$=(f,v)_\Omega+\left(\frac{1}{\sqrt{\mu}}g,v\right)_{\partial\Omega}.$$

The first added term is for the purpose of symmetry and zero because the solution is continuous over inner facets. The second term is to make the weak formulation stable and consistent because of the continuity of the solution. The stabilisation term has an additional interesting feature as it has an imaginary coefficient because α is assumed to be positive. This will be an important aspect of the introduced formulation in this work, as can be seen in the analysis later, but just to whet the appetite, by choosing test and trial functions as the same and looking at the imaginary part, then it follows that, first, the trace on the domain boundary is directly controlled by the sesquilinear form and secondly, also jumps on inner facets are controlled. The required solution space such that the weak formulation is well defined is the $H^s_{pw}(\mathcal{T})$ for $s>3/2$, because then there holds for all $u\in H^s_{pw}(\mathcal{T})$

$$u|_T\in L^2(T),\qquad \nabla u|_T\in[L^2(T)]^d,\qquad \forall T\in\mathcal{T}$$

and

$$u|_F\in L^2(F),\qquad \nabla u|_F\cdot\mathbf{n}\in L^2(F),\qquad \forall F\in\mathcal{F}.$$

As can be seen, not only u and its gradient are square integrable, but also their traces on facets are square integrable. Assuming a solution just in $H^1(\Omega)$ is not sufficient any more for this DG formulation. This is a property very well known for DG and also HDG formulations, as will be seen later.

It is possible to generate a DG-formulation which supports a Schur complement, but the biggest disadvantage of DG methods remains. Due to the discontinuous space, the unknowns increase substantially, and this also includes coupling degrees of freedom. Computational costs are high. Hybrid methods have been developed to circumvent this issue.

In this section, the introduced DG-formulation should only be considered as an example to illustrate the DG-method. Many other such formulations have been developed with a

multitude of different properties and initially, this method was developed for time-domain problems. The first origin goes back to the work by Reed and Hill in 1973 [RH73]. Since then, it has been further developed and also applied to elliptic, parabolic and hyperbolic problems.

3. Hybrid Discontinuous Galerkin Method

In this chapter, the HDG method is introduced. As was seen in the previous chapters, standard FEM usually does not lead to formulations which can facilitate Schur complement for the Helmholtz problem. For DG methods, there are formulations which satisfy static condensation properties but suffer from the drawback that many more DoF are introduced, which also leads to a large remaining system of linear equations after static condensation. The introduced HDG methods in this chapter are exactly tailored towards enabling static condensation while having a somewhat small number of remaining so-called skeleton unknowns.

The idea of HDG methods is not only to introduce discontinuous spaces, but to add additional facet spaces and let all element unknowns only couple with their attached facet unknowns on element boundaries. This leads to a formulation where the different element unknowns do not couple directly but rather over the facet space. The facet space is supported on all facets of the mesh, which form the skeleton of the mesh. All unknowns of the facet spaces are the skeleton DoF. To put it into the in the previous chapter established context, the goal is to apply a Schur complement to the HDG-formulation so that only a smaller system of linear equations needs to be solved for the skeleton DoF, or equally the iterative solver will only operate on skeleton DoF.

3.1. Deriving HDG Formulations

This section gives a first glance at how HDG formulations are derived. In the following the solution u of the Helmholtz problem is assumed to be in $H^s(\Omega), s > 3/2$. The pressure u as well as the flux $\mu^{-1}\nabla u \cdot \mathbf{n}$ are assumed to be continuous over inner facets $F \in \mathcal{F}_I$. The initial weak formulation for Equation (1.2a) on a mesh $\mathcal{T}$ looks like

$$\sum_{T\in\mathcal{T}} (\mu^{-1}\nabla u, \nabla v)_T - \kappa^2(\varepsilon u, v)_T - (\mu^{-1}\nabla u \cdot \mathbf{n}, v)_{\partial T} = (f, v)_\Omega$$

with discontinuous test functions $v \in H^s_{pw}(\mathcal{T})$ and $s > 3/2$. Because of the normal continuity of the flux, we can add

$$0 = (\mu_+^{-1}\nabla u_+ \cdot \mathbf{n}_+ + \mu_-^{-1}\nabla u_- \cdot \mathbf{n}_-, \hat{v})_{F_I}$$

with an arbitrary function $\hat{v} \in L^2(\mathcal{F})$ in the space

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

leading to

$$\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u,\nabla v)_T-\kappa^2(\varepsilon u,v)_T-(\mu^{-1}\nabla u\cdot\mathbf{n},v-\hat{v})_{\partial T}-(\mu^{-1}\nabla u\cdot\mathbf{n},\hat{v})_{\partial\Omega}=(f,v)_\Omega.$$

Contrary to DG methods, where the element test functions are used to reinforce the continuity in a weak sense, in HDG methods, new variables on the skeleton are introduced, leading to, at first glance, even more unknowns. The additional mixed term on the element boundary compensates for the respective terms appearing in the sum over elements, due to

$$\begin{aligned}0=\sum_{F_I\in\mathcal{F}_I}(\mu_+^{-1}\nabla u_+\cdot\mathbf{n}_++\mu_-^{-1}\nabla u_-\cdot\mathbf{n}_-,\hat{v})_{F_I}&=\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u\cdot\mathbf{n},\hat{v})_{\partial T\setminus\partial\Omega}\\&=\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u\cdot\mathbf{n},\hat{v})_{\partial T}-(\mu^{-1}\nabla u\cdot\mathbf{n},\hat{v})_{\partial\Omega}.\end{aligned}$$

Next, by defining $\hat{u}:=u$ the symmetry terms

$$0=(u-\hat{u},\mu^{-1}\nabla v\cdot\mathbf{n})_{\partial T}$$

and the stabilisation terms

$$0=j\alpha(u-\hat{u},v-\hat{v})_{\partial T}$$

can be added, which are zero because of the continuity of u over inner facets. The facet variable $\hat{u}\in L^2(\mathcal{F})$ is well defined due to the considered regularity of $u\in H^s(\Omega)$ with $s>3/2$. For the domain boundary term the Robin boundary condition in Equation (1.2b) is used

$$(\mu^{-1}\nabla u\cdot\mathbf{n},\hat{v})_{\partial\Omega}=j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}u,\hat{v}\right)_{\partial\Omega}+\left(\frac{1}{\sqrt{\mu}}g,\hat{v}\right)_{\partial\Omega}=j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}\hat{u},\hat{v}\right)_{\partial\Omega}+\left(\frac{1}{\sqrt{\mu}}g,\hat{v}\right)_{\partial\Omega},$$

resulting in

$$\begin{aligned}\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u,\nabla v)_T-\kappa^2(\varepsilon u,v)_T-(\mu^{-1}\nabla u\cdot\mathbf{n},v-\hat{v})_{\partial T}-(u-\hat{u},\mu^{-1}\nabla v\cdot\mathbf{n})_{\partial T}\\-j\alpha(u-\hat{u},v-\hat{v})_{\partial T}-j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}\hat{u},\hat{v}\right)_{\partial\Omega}=(f,v)_\Omega+\left(\frac{1}{\sqrt{\mu}}g,\hat{v}\right)_{\partial\Omega}.\end{aligned}$$

Introducing the short notation $[u]:=u-\hat{u}$ on facets gives

$$\begin{aligned}\sum_{T\in\mathcal{T}}(\mu^{-1}\nabla u,\nabla v)_T-\kappa^2(\varepsilon u,v)_T-(\mu^{-1}\nabla u\cdot\mathbf{n},[v])_{\partial T}-([u],\mu^{-1}\nabla v\cdot\mathbf{n})_{\partial T}\\-j\alpha([u],[v])_{\partial T}-j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}\hat{u},\hat{v}\right)_{\partial\Omega}=(f,v)_\Omega+\left(\frac{1}{\sqrt{\mu}}g,\hat{v}\right)_{\partial\Omega}.\end{aligned}$$

Now, the only coupling of element DoFs is through facet DoFs. Therefore, the Schur complement of the HDG-formulation is smaller than that of the DG method. This HDG-formulation is very similar to the DG-formulation introduced for the Helmholtz equation.

Over corners and edges of the mesh, no continuity of the facet variables are required. Discontinuous polynomials are sufficient as a discrete space

$$\hat{v}_h \in \mathcal{P}^p_{pw}(\mathcal{F}) := \prod_{F\in\mathcal{F}} \mathcal{P}^p(F) \subset L^2(\mathcal{F}).$$

The weak formulation for this HDG-formulation reads:

Definition 16. *For given $f \in L^2(\Omega), g \in L^2(\partial\Omega)$, positive, spatially dependent physical parameters $\mu(\mathbf{x}) \in L^\infty(\Omega)$, $\varepsilon(\mathbf{x}) \in L^\infty(\Omega)$ and $\kappa > 0$ find $u \in H^s_{pw}(\mathcal{T})$, $\hat{u} \in L^2(\mathcal{F})$, with $s > 3/2$, such that*

$$B_{prim,\alpha}(u,\hat{u};v,\hat{v}) = (f,v)_\Omega + \left(\frac{1}{\sqrt{\mu}}g,\hat{v}\right)_{\partial\Omega}$$

holds for all $v \in H^s_{pw}(\mathcal{T})$, $\hat{v} \in L^2(\mathcal{F})$ with the sesquilinear form

$$B_{prim,\alpha}(u,\hat{u};v,\hat{v}) := \sum_{T\in\mathcal{T}} (\mu^{-1}\nabla u, \nabla v)_T - \kappa^2(\varepsilon u, v)_T$$
$$-(\mu^{-1}\nabla u\cdot\mathbf{n},[v])_{\partial T} - ([u],\mu^{-1}\nabla v\cdot\mathbf{n})_{\partial T} - j\alpha([u],[v])_{\partial T} - j\kappa\left(\sqrt{\frac{\varepsilon}{\mu}}\hat{u},\hat{v}\right)_{\partial\Omega}.$$

The jumps are defined as

$$[u] := u - \hat{u}, \qquad \text{on } \partial T$$

with the positive stabilisation parameter $\alpha > 0$.

The steps above are the common way to derive consistent HDG formulations for the Helmholtz equation. As will be seen later, this is by far not the only possible consistent formulation. A large variety of different formulations exist with different properties.

3.2. State of the Art of HDG Methods for the Helmholtz Equation with Robin Boundary Conditions

HDG methods are a recent development. The idea was to take the desirable properties of DG methods and increase the computational efficiency. As was already stated, the major advantage in this regard is the applicability of a Schur complement. In 2005, the general concept of HDG methods took shape in the paper "A locally conservative LDG method for the incompressible Navier-Stokes equations" [CKS05] by Bernardo Cockburn, Guido Kanschat, and Dominik Schötzau. The following gives an overview of established DG and HDG methods for the Helmholtz equation.

DG-Formulation by Melenk, Parsania and Sauter The following DG formulation was published by Jens Markus Melenk, Asieh Parsania, and Stefan A. Sauter in their paper "General DG-Methods for Highly Indefinite Helmholtz Problems" [MPS13].

Definition 17. *Assume an abstract finite-dimensional space $S \subset H^s_{pw}(\Omega)$, with $s > 3/2$. Find $u \in S$ such that, for all $v \in S$,*

$$a_{\mathcal{T}}(u,v) - \kappa^2 (u,v)_\Omega = (f,v)_\Omega - \sum_{F \in \mathcal{F}_O} \left(\delta \frac{1}{j\kappa} g, \nabla v \cdot \mathbf{n} \right)_F + ((1-\delta)g, v)_F$$

where $a_{\mathcal{T}}$ is the DG-sesquilinear form on $S \times S$ defined by

$$\begin{aligned} a_{\mathcal{T}}(u,v) := & \sum_{T \in \mathcal{T}} (\nabla u, \nabla v)_T \\ & - \sum_{F \in \mathcal{F}_I} \frac{1}{2} (u_+ \mathbf{n}_+ + u_- \mathbf{n}_-, \nabla v_+ + \nabla v_-)_F + \frac{1}{2} (\nabla u_+ + \nabla u_-, v_+ \mathbf{n}_+ + v_- \mathbf{n}_-)_F \\ & - \sum_{F \in \mathcal{F}_O} (\delta u, \nabla v \cdot \mathbf{n})_F + (\nabla u \cdot \mathbf{n}, \delta v)_F \\ & - \frac{1}{j\kappa} \sum_{F \in \mathcal{F}_I} (\beta (\nabla u_+ \cdot \mathbf{n}_+ + \nabla u_- \cdot \mathbf{n}_-), \nabla v_+ \cdot \mathbf{n}_+ + \nabla v_- \cdot \mathbf{n}_-)_F - \frac{1}{j\kappa} \sum_{F \in \mathcal{F}_O} (\delta \nabla u \cdot \mathbf{n}, \nabla v \cdot \mathbf{n})_F \\ & + j\kappa \sum_{F \in \mathcal{F}_I} (\alpha (u_+ - u_-), v_+ - v_-)_F + j\kappa \sum_{F \in \mathcal{F}_O} ((1-\delta)u, v)_F. \end{aligned}$$

It is complex and symmetric, and the functions α, β and δ are positive. They showed consistency of the formulation and continuity as well as coercivity of $a_{\mathcal{T}}(u,v) + \kappa^2 (u,v)_\Omega$ under the conditions that

$$\alpha = \mathcal{O}\left(\frac{p^2}{\kappa h} \right), \qquad \beta = \mathcal{O}\left(\frac{\kappa h}{p} \right), \qquad \delta = \mathcal{O}\left(\frac{\kappa h}{p} \right)$$

for piecewise polynomial spaces as S. By further assuming a star-shaped domain, they proved that the formulation is asymptotically quasi-optimal, but the far more interesting aspect was that they showed the unconditional unique solvability by employing the special discrete test function

$$v = \nabla u \cdot \mathbf{x}.$$

As already seen in the continuous stability results in the chapter above, this test function gives for the continuous case wave number explicit stability results. Interestingly, this is not only a continuous test function, but for DG-spaces also a valid discrete test function, which leads to absolute stability. For general conforming FEM-spaces such as the Nodal space, this does not hold, which highlights the advantageous properties of using discontinuous formulations. For further information on the special test function, the publications [Mel95, EM12, MS22] are recommended. This DG-formulation only has one major drawback regarding iterative solvability. The stabilisation parameters are mesh size and polynomial degree dependent. The questions are: Can this formulation be adapted into an HDG-formulation? Is it possible to use only wave number dependent stabilisation coefficients? Does a stable Schur complement exist?

Coercive Formulation by Andrea Moiola and Euan A. Spence The special test function in the previous paragraph gave rise to the question of whether the Helmholtz Equation is sign-indefinite and therefore has the requirement of a Schatz argument with the entailing requirement of a resolution condition. Andrea Moiola and Euan A. Spence answered this question in their work "Is the Helmholtz Equation Really Sign-Indefinite?" [MS14]. They considered

$$V := \{v : v \in H^1(\Omega), \Delta v \in L^2(\Omega), v \in H^1(\partial\Omega), \nabla v \cdot \mathbf{n} \in L^2(\partial\Omega)\}$$

as their continuous Ansatz and test space. The discrete subspace would also need to satisfy these conditions and the corresponding necessary continuities. They considered the sesquilinear form

$$\begin{aligned} b(u,v) := (\nabla u, \nabla v)_\Omega + \kappa^2 (u,v)_\Omega + \left(Mu + \frac{1}{3\kappa^2} Lu, Lv\right)_\Omega \\ -j\kappa(u, Mv)_{\partial\Omega} - \left(\mathbf{x}\cdot\nabla u - j\kappa\beta u + \frac{d-1}{2}u, \nabla v\cdot\mathbf{n}\right)_{\partial\Omega} \\ -\kappa^2(\mathbf{x}\cdot\mathbf{n}u, v)_{\partial\Omega} + (\mathbf{x}\cdot\mathbf{n}\nabla u, \nabla v)_{\partial\Omega} \end{aligned}$$

on $V \times V$ with the right hand anti-linear form

$$G(v) := \left(f, Mv - \frac{1}{3\kappa^2} Lv\right)_\Omega + (g, Mv)_{\partial\Omega}$$

with β an arbitrary real constant

$$Lu := \Delta u + \kappa^2 u, \qquad \text{and} \qquad Mu := \mathbf{x}\cdot\nabla u - j\kappa\beta u + \frac{d-1}{2}u.$$

They showed that this formulation is consistent and also the remarkable property of coercivity. An integral part plays the special test function, which somewhat can already be seen by the form of their sesquilinear form. Therefore, for star-shaped domains, there exist coercive formulations for the Helmholtz equation with Robin boundary conditions.

Lowest Order DG-Formulation by Feng and Wu The authors Xiaobing Feng and Haijun Wu published the paper [FW09] in which they introduced and analysed the DG-sesquilinear form

$$a_h(u,v) := b_h(u,v) + j\left(J_0(u,v) + J_1(u,v) + L_1(u,v)\right)$$

on the space

$$E := \prod_{T\in\mathcal{T}} H^2(T),$$

where

$$\begin{aligned} b_h(u,v) := \sum_{T\in\mathcal{T}} (\nabla u, \nabla v)_T \\ -\frac{1}{2}\sum_{F\in\mathcal{F}_I} \left((\nabla u_+ + \nabla u_-)\cdot\mathbf{n}_+, v_+ - v_-\right)_F + \sigma\left(u_+ - u_-, (\nabla u_+ + \nabla u_-)\cdot\mathbf{n}_+\right)_F, \end{aligned}$$

$$J_0(u,v) := \sum_{F\in\mathcal{F}_I} \frac{\gamma_0}{h}(u_+ - u_-, v_+ - v_-)_F,$$

$$J_1(u,v) := \sum_{F\in\mathcal{F}_I} \gamma_1 h(\nabla u_+ \cdot \mathbf{n}_+ + \nabla u_- \cdot n_-, \nabla v_+ \cdot \mathbf{n}_+ + \nabla v_- \cdot n_-)_F,$$

$$L_1(u,v) := \sum_{F\in\mathcal{F}_I} \sum_{i=1}^{d-1} \frac{\beta_1}{h} \left((\nabla u_+ - \nabla u_-)\cdot\tau_i, (\nabla v_+ - \nabla v_-)\cdot\tau_i\right)_F,$$

with the h-independent real number σ and positive numbers $\gamma_0, \gamma_1, \beta_1$ representing stabilisations. The L_1 is a stabilisation of the tangential jumps of the gradient with the unit tangential vectors τ_i. They mainly considered the case of $\sigma = 1$ for which the sesquilinear form is complex symmetric. The main result of their work is the absolute stability of the formulation and explicit tracking of the pre-asymptotic stability constants as well. The proof is based upon applying the special test function $\nabla u \cdot \mathbf{n}$, and they showed their results under the assumption of a star-shaped domain for the polynomial degree $p = 1$.

High Order DG-Formulation by Feng and Wu After their analysis of the lowest order DG-formulation, see previous paragraph, they published the adaptation towards a higher order formulation in the paper [FW11]. They adapted their previous formulation into the sesquilinear form

$$a_h^p(u,v) := b_h(u,v) + j\left(L_1(u,v) + \sum_{i=0}^{p} J_i(u,v)\right)$$

on the space

$$E^p := \prod_{T\in\mathcal{T}} H^{p+1}(T),$$

where

$$b_h(u,v) := \sum_{T\in\mathcal{T}} (\nabla u, \nabla v)_T$$
$$-\frac{1}{2}\sum_{F\in\mathcal{F}_I} \left((\nabla u_+ + \nabla u_-)\cdot n_+, v_+ - v_-\right)_F + \sigma\left(u_+ - u_-, (\nabla u_+ + \nabla u_-)\cdot\mathbf{n}_+\right)_F,$$

$$L_1(u,v) := \sum_{F\in\mathcal{F}_I} \sum_{l=1}^{d-1} \frac{\beta_1 p}{h} \left((\nabla u_+ - \nabla u_-)\cdot\tau_l, (\nabla v_+ - \nabla v_-)\cdot\tau_l\right)_F,$$

$$J_0(u,v) := \sum_{F\in\mathcal{F}_I} \frac{\gamma_0 p}{h}(u_+ - u_-, v_+ - v_-)_F,$$

$$J_i(u,v) := \sum_{F\in\mathcal{F}_I} \gamma_i \frac{h}{p}^{2i-1} \left(\frac{\partial^i u_+}{\partial \mathbf{n}_+^i} + \frac{\partial^i u_-}{\partial \mathbf{n}_-^i}, \frac{\partial^i v_+}{\partial \mathbf{n}_+^i} + \frac{\partial^i v_-}{\partial \mathbf{n}_-^i}\right)_F,$$

with the h-independent real number σ and positive numbers γ_i, β_1 representing stabilisations. The L_1 is a stabilisation of the tangential jumps of the gradient with the unit

tangential vectors τ_i and J_i is the stabilisation of the jump of the i-th normal derivative. The main result of their work is the absolute stability of the formulation and explicit tracking of the pre-asymptotic stability constants, this time considering the higher polynomial order p. As can be seen, their stabilisation parameters are mesh size and polynomial degree dependent. The interesting aspect of this formulation is the stabilisation of the tangential flux as well as the stabilisation of the higher-order normal moments on facets.

Local DG-Methods by Feng and Xing Xiaobing Feng went on and developed with Yulong Xing mixed formulations, [FX13]. They used discrete spaces

$$V_h := \mathcal{P}^r_{pw}(\mathcal{T}), \qquad \Sigma_h := \left[\mathcal{P}^l_{pw}\right]^d$$

and the following first formulation: Find $(u_h, \sigma_h) \in V_h \times \Sigma_h$ such that

$$A_h(u_h, \sigma_h; v_h, \tau_h) = F(v_h, \tau_h)$$

for all $(v_h, \tau_h) \in V_h \times \Sigma_h$ where

$$\begin{aligned} A_h(u_h, \sigma_h; v_h, \tau_h) &:= (\sigma_h, \nabla v_h)_\Omega - \kappa^2 (u_h, v_h)_\Omega + j\kappa (u_h, v_h)_{\partial\Omega} \\ &- \sum_{F \in \mathcal{F}_I} \left(\frac{1}{2}(\nabla u_{h,+} + \nabla u_{h,-}) - j\beta(u_{h,+}\mathbf{n}_+ + u_{h,-}\mathbf{n}_-), v_{h,+}\mathbf{n}_+ + v_{h,-}\mathbf{n}_- \right) \\ &- \sum_{F \in \mathcal{F}_I} j\delta(\nabla u_{h,+} \cdot \mathbf{n}_+ + \nabla u_{h,-} \cdot \mathbf{n}_-, \tau_{h,+} \cdot \mathbf{n}_+ + \tau_{h,-} \cdot \mathbf{n}_-) \\ &+ \frac{1}{2} \sum_{F \in \mathcal{F}_I} (u_{h,+}\mathbf{n}_+ + u_{h,-}\mathbf{n}_-, \tau_{h,+} + \tau_{h,-}) \\ &+ (\sigma_h, \tau_h)_\Omega - (\nabla u_h, \tau_h)_\Omega, \end{aligned}$$

and

$$F(v_h, \tau_h) := (f, v_h)_\Omega + (g, v_h)_{\partial\Omega}.$$

They showed for lowest order discretisation $r = 1$ and $l = 1$ the local absolute stability of this formulation. The analysis is based on the test functions $\nabla u \cdot \mathbf{x}$, but they developed a pre-asymptotic discrete inf-sup-condition. Their established analysis can also cope with mesh size independent stabilisation parameters δ and β. In their analysis, inverse estimates are necessary, which only hold for discrete functions, and therefore, only a discrete pre-asymptotic inf-sup condition has been considered. They also analysed a very similar second

formulation with the sesquilinear form

$$\begin{aligned} B_h(u_h,\sigma_h;v_h,\tau_h) &:= (\sigma_h,\nabla v_h)_\Omega - \kappa^2(u_h,v_h)_\Omega + j\kappa(u_h,v_h)_{\partial\Omega} \\ &- \sum_{F\in\mathcal{F}_I}\left(\frac{1}{2}(\sigma_{h,+}+\tau_{h,-}) - j\beta(u_{h,+}\mathbf{n}_+ + u_{h,-}\mathbf{n}_-), v_{h,+}\mathbf{n}_+ + v_{h,-}\mathbf{n}_-\right) \\ &- \sum_{F\in\mathcal{F}_I} j\delta(\sigma_{h,+}\cdot\mathbf{n}_+ + \tau_{h,-}\cdot\mathbf{n}_-, \tau_{h,+}\cdot\mathbf{n}_+ + \tau_{h,-}\cdot\mathbf{n}_-) \\ &+ \frac{1}{2}\sum_{F\in\mathcal{F}_I}(u_{h,+}\mathbf{n}_+ + u_{h,-}\mathbf{n}_-, \tau_{h,+}+\tau_{h,-}) \\ &+ (\sigma_h,\tau_h)_\Omega - (\nabla u_h,\tau_h)_\Omega \end{aligned}$$

where in some terms ∇u_h is replaced by σ_h. A crucial aspect of the analysis was the star-shaped domain again.

CIP-FEM by Zhu and Wu The CIP-FEM is not a discontinuous method. Nonetheless in their two papers [Wu14, ZW12] Lingxue Zhu and Haijun Wu established and analysed the following conforming formulation for the Helmholtz equation on the space

$$V := H^1(\Omega) \cap \prod_{T\in\mathcal{T}} H^2(T)$$

with the weak formulation

$$\begin{aligned} (\nabla u,\nabla v)_\Omega - \kappa^2(u,v)_\Omega + j\kappa(u,v)_{\partial\Omega} \\ + \sum_{F\in\mathcal{F}_I} j\gamma\frac{h}{p^2}(\nabla u_+\cdot\mathbf{n}_+ + \nabla u_-\cdot\mathbf{n}_-, \nabla v_+\cdot\mathbf{n}_+ + \nabla v_-\cdot\mathbf{n}_-)_F \\ = (f,v)_\Omega + (g,v)_{\partial\Omega}. \end{aligned}$$

It resembles the classic weak formulation of the Helmholtz equation, but with an additional stabilisation term for the normal jump of the flux on inner facets. With this stabilisation, they were able to perform pre-asymptotic error analysis by applying the special test function $\nabla u\cdot\mathbf{x}$ on a star-shaped domain. In the asymptotic regime, the results are the same as for the classical Schatz argument, and then they extend the error estimates onto an area where the established theory does not give estimates. Additionally, they showed that this formulation with a conforming space is also absolutely stable. The stabilisation itself is mesh size and polynomial degree dependent, which makes it unsuitable for the iterative solver proposed in this work.

HDG-Formulation by Chen, Lu and Xu The authors Huangxin Chen, Peipei Lu and Xuejen Xu published an absolutely stable HDG-formulation for the Helmholtz equation

[CLX13]. The formulation itself is for the mixed problem and looks like

$$\begin{aligned}
\sum_{T\in\mathcal{T}} (j\kappa\sigma_h, \tau_h)_T - (u_h, \operatorname{div}\tau_h)_T + (\hat{u}_h, \tau_h\cdot\mathbf{n})_{\partial T} &= 0,\\
\sum_{T\in\mathcal{T}} (j\kappa u_h, v_h)_T - (\sigma_h, \nabla v_h)_T + (\sigma_h\cdot\mathbf{n}, v_h)_{\partial T} + \alpha(u_h-\hat{u}_h, v_h)_{\partial T} &= (f, v_h)_\Omega,\\
-(\sigma_h\cdot\mathbf{n}, \hat{v}_h)_{\partial\Omega} - \alpha(u_h-\hat{u}_h, \hat{v}_h)_{\partial\Omega} + (\hat{u}_h, \hat{v}_h)_{\partial\Omega} &= (g, \hat{v}_h)_{\partial\Omega},\\
\sum_{T\in\mathcal{T}} (\sigma_h\cdot\mathbf{n}, \hat{v}_h)_{\partial T\setminus\partial\Omega} + \alpha(u_h-\hat{u}_h, \hat{v}_h)_{\partial T\setminus\partial\Omega} &= 0,
\end{aligned}$$

on the spaces

$$\sigma_h, \tau_h \in \left[\mathcal{P}^p_{pw}(\mathcal{T})\right]^d, \qquad u_h, v_h \in \mathcal{P}^p_{pw}(\mathcal{T}), \qquad \hat{u}_h, \hat{v}_h \in \mathcal{P}^p(\mathcal{F})$$

with the stabilisation parameter

$$\alpha := \mathcal{O}\left(\frac{p}{\kappa h}\right).$$

They show absolute stability as well as optimal convergence rates for the linear case of spaces. The interesting aspect is an analysis which is based on inserting projections, which alleviates the star-shaped restriction and only a regularity and stability assumption on the continuous adjoint problem remains. A very similar hybrid method for Maxwell's equations was introduced by Xiaobing Feng, Peipei Lu and Xuejun Xu in [FLX16] and also analysed similarly.

HDG-Formulation by Zhu and Wu The formulation considered by Bingxin Zhu and Haijun Wu is the same as the formulation in the previous paragraph, with the one small but very impactful difference that the stabilisation parameter is considered as

$$\alpha := \mathcal{O}(\kappa),$$

wave number dependent and mesh size independent. They proved for the linear case error estimates as well as the absolute stability of the formulation, by employing elliptic projections [ZW20].

HDG-Formulation by Griesmaier and Monk The paper [GM11], by Roland Griesmaier and Peter Monk, covers a very similar weak formulation

$$\begin{aligned}
\sum_{T\in\mathcal{T}} (j\kappa\sigma_h, \tau_h)_T - (u_h, \operatorname{div}\tau_h)_T + (\hat{u}_h, \tau_h\cdot\mathbf{n})_{\partial T} &= 0,\\
\sum_{T\in\mathcal{T}} (j\kappa u_h, v_h)_T - (\sigma_h, \nabla v_h)_T + (\sigma_h\cdot\mathbf{n}, v_h)_{\partial T} + \alpha(u_h-\hat{u}_h, v_h)_{\partial T} &= (f, v_h)_\Omega,\\
(\hat{u}_h, \hat{v}_h)_{\partial\Omega} &= (g, \hat{v}_h)_{\partial\Omega},\\
\sum_{T\in\mathcal{T}} (\sigma_h\cdot\mathbf{n}, \hat{v}_h)_{\partial T\setminus\partial\Omega} + \alpha(u_h-\hat{u}_h, \hat{v}_h)_{\partial T\setminus\partial\Omega} &= 0
\end{aligned}$$

on piecewise polynomial spaces as in the previous two paragraphs. The main difference in the formulation is the consideration of Dirichlet boundary conditions instead of Robin boundary conditions. They carried out an error and stability analysis explicitly tracing the stability parameter α, with the ability to choose it wave number or mesh size dependent. They did not cover the absolute stability of the formulation, which is not possible due to the choice of boundary conditions and rather showed optimal convergence rates, under sub-optimal stabilisation, for high-order discretisation. The key aspect was the usage of a special projection tailored towards the formulation itself which was previously established for elliptic problems by Bernard Cockburn, Jayadeep Gopalakrishnan and Raytcho Lazarov in their work [CGL09] and further refined by Bernhard Cockburn, Jayadeep Gopalakrishnan and Francisco-Javier Sayas in [CGS10]. Later, Francisco-Javier Sayas published a compendium about this projection method for multiple formulations in his paper [Say13].

HDG-Formulation for Maxwell's Equations by Lu, Chen and Qiu The time-harmonic Maxwell's equations have a similar but more complex structure as the Helmholtz equation. In the paper [LCQ17], Peipei Lu, Huangxin Chen and Weifeng Qiu consider a weak formulation similar to the formulation in [FLX16] with a mesh size-dependent stabilisation parameter. Their major contribution is the establishment of an error analysis, which covers an explicit stability and error estimation of the whole convergence process. They do not need a resolution condition for their results and only rely on a stability and regularity assumption on the adjoint problem. The most noteworthy addition is the consideration of L^2-projections in their estimations, which lead them to their results for high-order discretisations.

The previously mentioned contributions to the field of simulations for the Helmholtz equation, were purely towards stability and error analysis to facilitate unique solvability as well as optimality of the formulations. For the next methods fast-solving strategies exist or are being developed.

Optimized Schwarz Methods by Claeys, Collino, Joly and Parolin The optimized Schwarz methods studied by Xavier Claeys, Francis Collino, Patrick Joly and Emile Parolin in [CCJP20, CP22, CCP22, cla23] are based upon a domain decomposition approach. They use the standard weak formulation for conforming nodal elements, see Equation (2.1), but split the domain into non-overlapping subdomains $\Omega := \cup_i \Omega_i$. Instead of a conforming space on the whole domain, the Helmholtz equation is discretised with spaces that are only conforming on these subdomains, in the sense of $X_h := \prod_i H^1(\Omega_i)$. The weak formulation is adapted onto the subdomains to facilitate the discontinuous space. The continuity is reinforced by an additional operator on subdomain boundaries. The methods are tailored towards iterative methods based on domain decomposition and the authors prove the solvability of the subdomain problems and the convergence of the iterative methods under the assumption of a discrete inf-sup condition.

The DPG-Method by Demkowicz and Gopalakrishnan The discontinuous Petrov Galerkin methods introduced by Leszek Demkowicz and Jayadeep Gopalakrishnan in [DG10, DG11,

DGN12, GMO14, DGMZ12] has very beneficial properties. In a Petrov Galerkin method, the Ansatz and test space in the weak formulation are different. For the DPG method, the Ansatz space consists of piecewise polynomials, the same as for HDG formulations, but the test space consists of local solutions of Helmholtz equations with the polynomial Ansatz functions as the right hand side. This leads to a least squares method, for which a solution always exists. Therefore, the existence and optimality analysis is very short. The more interesting part is the necessity of a discontinuous PG method. It would also be possible to use a conforming, continuous Ansatz space of piece-wise polynomials, but then the test space would have non-local support, and the discretisation matrix would not be sparse any more. On facets, the method uses two different hybrid variables, one representing the pressure and the other the normal component of the flux. In the method, a Schur complement is applied, and the remaining system of linear equations on the hybrid variables is still a coercive problem. The second hybrid variable is necessary for a mesh size independently stable Schur complement. This coercivity leads to the second favourable property, as efficient iterative solvers and pre-conditioners for elliptic problems can be applied. Examples of such solving strategies are developed by Jacob Badger, Stefan Henneking, Socratis Petrides and Leszek Demkowicz, see for example [BHPD23] and further works.

HDG-Formulation by Monk, Schöberl and Sinwel Peter Monk, Joachim Schöberl and Astrid Sinwel proposed an HDG-formulation with two facet variables in [MSS10]. The focus of their work was on the development of solving strategies with this formulation. It uses stabilisations, which are mesh size and polynomial degree independent and only may depend on the given wave number. Extensive research into solving strategies was carried out by Martin Huber and Joachim Schöberl. in [Hub13, HPS13, HS14]. A missing part was the stability and error analysis for the formulation, which is exactly the main focus of the present work. Therefore, the specific details of the formulation will be covered later.

HDG-Formulation by Modave and Chaumont-Frelet A very similar formulation to the one in the previous paragraph was published by Axel Modave and Théophile Chaumont-Frelet in [MCF23]. Instead of pressure and flux as hybrid variables, they consider the very natural left and right side impedance traces. With this introduction, they can write the Schur complement as an exchange operation on these impedance traces and can prove contractive properties as well as unique solvability. Their focus lies on numerical studies of the condition of the Schur complement, and they highlight the necessity of a second hybrid variable.

This formulation is absolutely stable due to the second facet variable, but it has the major disadvantage that it is only optimal if h-dependent stabilisation parameters α and β are chosen, which makes iterative solves based on domain decomposition unsuitable. For the wave number-dependent stabilisation, the formulation does not converge with optimal rates.

3.3. The HDG-Formulation with Impedance Traces

This section considers the following mixed Helmholtz problem with an impedance boundary condition.

Definition 18 (Mixed Helmholtz Problem with Robin boundary conditions)**.** *Consider a bounded domain $\Omega \subset \mathbb{R}^3$ with a unique normal vector field $\mathbf{n}$ on $\partial\Omega$. For a given wave number $\kappa > 0$, volume excitation $f(\mathbf{x})$ and boundary excitation $g(\mathbf{x})$ find a scalar field $u(\mathbf{x})$ and a vector field $\sigma(\mathbf{x})$ so that they satisfy*

$$j\kappa\sigma - \nabla u = 0 \qquad \text{in } \Omega, \tag{3.3a}$$

$$-\operatorname{div}(\sigma) + j\kappa u = \frac{1}{j\kappa} f \qquad \text{in } \Omega, \tag{3.3b}$$

$$\sigma \cdot \mathbf{n} - u = \frac{1}{j\kappa} g \qquad \text{on } \partial\Omega. \tag{3.3c}$$

A major part of this section has been published in [LS23] and is therefore closely related. Some parts were reformulated, and a new section regarding the asymptotic analysis has been added. In this section, the stability and error analysis for the HDG-formulation is carried out.

For the analysis, the following adjoint problem is crucial.

Definition 19 (Adjoint Mixed Helmholtz Problem)**.** *For given $\kappa > 0$ and $f \in L^2(\Omega)$, let (ϕ, w) be the solution of the BVP*

$$j\kappa\phi + \nabla w = 0 \qquad \text{in } \Omega, \tag{3.4a}$$

$$\operatorname{div}(\phi) + j\kappa w = \frac{1}{j\kappa} f \qquad \text{in } \Omega, \tag{3.4b}$$

$$\phi \cdot \mathbf{n} - w = 0 \qquad \text{on } \partial\Omega. \tag{3.4c}$$

The adjoint BVP has a homogenous Robin boundary condition and a volume excitation f. The existence and uniqueness of the continuous solution have been proven in [Mel95, Proposition 8.1.3].

Remark 20. *The mixed adjoint problem is equivalent to*

$$-\Delta w - \kappa^2 w = f \qquad \text{in } \Omega,$$
$$\nabla w \cdot \mathbf{n} + j\kappa w = 0 \qquad \text{on } \partial\Omega.$$

To derive the weak formulation, the standard approaches are used. Equation (3.3a) is multiplied with complex piecewise test functions τ and (3.3b) is multiplied with complex piecewise functions v, then the gradient is partially integrated, leading to the equations

$$\sum_{T\in\mathcal{T}} j\kappa(\sigma,\tau)_T + (u, \operatorname{div}\tau)_T - (u, \tau\cdot\mathbf{n})_{\partial T} = 0,$$

$$\sum_{T\in\mathcal{T}} (\operatorname{div}\sigma, v)_T - j\kappa(u,v)_T = \frac{j}{\kappa}(f,v)_\Omega.$$

Under the assumption of a continuous solution u the facet variable can be defined as $\hat{u} = u$, giving

$$(u, \tau \cdot \mathbf{n})_{\partial T} = (\hat{u}, \tau \cdot \mathbf{n})_{\partial T}$$

and for a normal continuous solution σ the zero element boundary term

$$0 = \sum_{F_I \in \mathcal{F}_I} (\sigma_+ \cdot \mathbf{n}_+, \hat{v})_{F_I} + (\sigma_- \cdot \mathbf{n}_-, \hat{v})_{F_I} = \sum_{T \in \mathcal{T}} (\sigma \cdot \mathbf{n}, \hat{v})_{\partial T} - (\sigma \cdot \mathbf{n}, \hat{v})_{\partial\Omega}$$

is introduced to symmetrise the formulation. The appearing boundary term is replaced by the Robin boundary condition in (3.3c) giving

$$(\sigma \cdot \mathbf{n}, \hat{v})_{\partial\Omega} = (\hat{u}, \hat{v})_{\partial\Omega} + \frac{1}{j\kappa}(g, \hat{v})_{\partial\Omega}.$$

Adding all equations together gives the formulation

$$\sum_{T \in \mathcal{T}} j\kappa(\sigma, \tau)_T + (u, \operatorname{div} \tau)_T + (\operatorname{div} \sigma, v)_T - j\kappa(u, v)_T$$
$$-(\hat{u}, \tau \cdot \mathbf{n})_{\partial T} - (\sigma \cdot \mathbf{n}, \hat{v})_{\partial T} + (\hat{u}, \hat{v})_{\partial\Omega} = \frac{j}{\kappa}(f, v)_\Omega + \frac{j}{\kappa}(g, \hat{v})_{\partial\Omega}.$$

Finally, the two stabilisation terms

$$0 = \sum_{T \in \mathcal{T}} \alpha(u - \hat{u}, v - \hat{v})_{\partial T},$$
$$0 = -\sum_{T \in \mathcal{T}} \beta(\sigma \cdot \mathbf{n} - \hat{\sigma}_\mathbf{n}, \tau \cdot \mathbf{n} - \hat{\tau}_\mathbf{n})_{\partial T}$$

are added, leading to the following formulation.

Definition 21 (HDG-Formulation)**.** *For given $f \in L^2(\Omega), g \in L^2(\partial\Omega)$ and $\kappa > 0$ find $\sigma \in H_{pw}(\operatorname{div})(\mathcal{T}) \cap [H^s_{pw}(\mathcal{T})]^d$, $u \in H^s_{pw}(\mathcal{T})$, $\hat{u} \in L^2(\mathcal{F})$, $\hat{\sigma}_\mathbf{n} \in L^2(\mathcal{F})$, with $s > 1/2$, such that*

$$B(\sigma, \hat{\sigma}_\mathbf{n}, u, \hat{u}; \tau, \hat{\tau}_\mathbf{n}, v, \hat{v}) = \frac{j}{\kappa}(f, v)_\Omega + \frac{j}{\kappa}(g, \hat{v})_{\partial\Omega} \tag{3.7}$$

holds for all $\tau \in H_{pw}(\operatorname{div})(\mathcal{T}) \cap [H^s_{pw}(\mathcal{T})]^d$, $v \in H^s_{pw}(\mathcal{T})$, $\hat{v} \in L^2(\mathcal{F})$, $\hat{\tau}_\mathbf{n} \in L^2(\mathcal{F})$ with the sesquilinear form

$$B(\sigma, \hat{\sigma}_\mathbf{n}, u, \hat{u}; \tau, \hat{\tau}_\mathbf{n}, v, \hat{v}) := \sum_{T \in \mathcal{T}} \Big(j\kappa(\sigma, \tau)_T + (u, \operatorname{div} \tau)_T + (\operatorname{div} \sigma, v)_T - j\kappa(u, v)_T$$
$$-(\hat{u}, \tau \cdot \mathbf{n})_{\partial T} - (\sigma \cdot \mathbf{n}, \hat{v})_{\partial T} + \alpha([u], [v])_{\partial T} - \beta([\![\sigma]\!], [\![\tau]\!])_{\partial T} \Big) + (\hat{u}, \hat{v})_{\partial\Omega}.$$

The jumps are defined as

$$[u] := u - \hat{u}, \qquad [\![\sigma]\!] := \sigma \cdot \mathbf{n} - \hat{\sigma}_\mathbf{n}, \qquad \text{on } \partial T$$

and the positive stabilisation parameters are $\alpha > 0$, $\beta > 0$.

This HDG-formulation was first introduced in [MSS10]. The subscript for the facet variables $\hat{\sigma}_{\mathbf{n}}$ and $\hat{\tau}_{\mathbf{n}}$ indicates that the variables are scalar-valued and their sign changes with respect to the adjacent elements and their normal vector.

The newly used discontinuous, complex-valued spaces are defined by:

$$H(\mathrm{div})(T) := \{\sigma \in [L^2(T)]^d : \mathrm{div}\,\sigma \in L^2(T)\},$$
$$H_{pw}(\mathrm{div})(\mathcal{T}) := \prod_{T\in\mathcal{T}} H(\mathrm{div})(T).$$

The respective discrete spaces of polynomial degree $p \in \mathbb{N}_0$ are

$$\mathcal{RT}^p(T) \subset H(\mathrm{div})(T) \cap H^s(T),$$
$$\mathcal{RT}^p_{pw}(\mathcal{T}) := \prod_{T\in\mathcal{T}} \mathcal{RT}^p(T) \subset H_{pw}(\mathrm{div})(\mathcal{T}) \cap H^s_{pw}(\mathcal{T}),$$
$$\mathcal{P}^p_{pw}(\mathcal{F}) := \prod_{F\in\mathcal{F}} \mathcal{P}^p(F) \subset L^2(\mathcal{F}),$$
$$\mathcal{P}^p_{pw}(\mathcal{T}) := \prod_{T\in\mathcal{T}} \mathcal{P}^p(T) \subset L^2(\Omega).$$

For the discrete, discontinuous compound space, the short notation

$$\mathcal{X}_h := \mathcal{RT}^p_{pw}(\mathcal{T}) \times \mathcal{P}^p_{pw}(\mathcal{T}) \times \mathcal{P}^p_{pw}(\mathcal{F}) \times \mathcal{P}^p_{pw}(\mathcal{F})$$

is used.

Remark 22. *An interesting property of the HDG-formulation is the choice of the h-independent parameters $\alpha = \mathcal{O}(1)$ and $\beta = \mathcal{O}(1)$. The reason for the choices in this study is due to the favourable properties of iterative solvers and their preconditioning established in [Hub13, HPS13, HS14].*

For methods with static condensation, the sub-problems defined on individual elements must be uniquely and stably solvable. The formulation in Definition 21 satisfies that condition, and the following discrete absolute stability result holds.

Theorem 23. *For a given wave number $\kappa > \kappa_0 > 0$, stabilisation parameters $\alpha = \mathcal{O}(1)$, $\beta = \mathcal{O}(1)$ and a star-shaped domain, assuming H^2-regularity of the adjoint Helmholtz problem and a polynomial degree $p \geq 1$, there holds for the discrete solution of the HDG-formulation (3.7)*

$$\sum_{T\in\mathcal{T}} \alpha\|[u_h]\|^2_{L^2(\partial T)} + \beta\|[\![\sigma_h]\!]\|^2_{L^2(\partial T)} + \frac{1}{2}\|\hat{u}_h\|^2_{L^2(\partial\Omega)} \leq \frac{1}{\kappa}\|f\|_{L^2(\Omega)}\|u_h\|_{L^2(\Omega)} + \frac{1}{2\kappa^2}\|g\|^2_{L^2(\partial\Omega)},$$

$$\sum_{T\in\mathcal{T}} \kappa\|\sigma_h\|^2_{L^2(T)} \leq \sum_{T\in\mathcal{T}} \kappa\|u_h\|^2_{L^2(T)} + \frac{2}{\kappa}\|f\|_{L^2(\Omega)}\|u_h\|_{L^2(\Omega)} + \frac{1}{\kappa^2}\|g\|^2_{L^2(\partial\Omega)},$$

$$\|u_h\|_{L^2(\Omega)} \leq C\left(\left(1+\kappa h+\kappa^3h^3\right)\|f\|_{L^2(\Omega)} + \sqrt{1+\kappa h+\kappa^3h^3}\,\frac{1}{\sqrt{\kappa}}\|g\|_{L^2(\partial\Omega)}\right)$$

with a constant $C > 0$ independent of κ and h.

This theorem implies the existence and uniqueness of the discrete solution for the HDG-formulation without a resolution condition on the discrete space. The proof uses a similar approach as introduced in [LCQ17] for Maxwell's equations. Two major differences are that only a single facet space is used in the mentioned work, and the stabilisation parameter therein is h-dependent. The method in [LCQ17] is related to the case of choosing $\alpha \approx \frac{1}{h}$, $\beta = 0$ and omitting the second facet variable in the HDG-formulation (3.7). The main idea of the proof is to use an Aubin-Nitsche technique along with L^2-projections of σ and u so that volume terms vanish and only facet terms remain. The real part of the HDG formulation controls those facet terms. With similar techniques and arguments, the following quasi-best approximation for jumps and the pressure are proven with the short notations

$$\eta(\sigma) := \inf_{\tau_h \in \mathcal{RT}^p_{pw}(\mathcal{T})} \left(\|\sigma - \tau_h\|^2_{L^2(\Omega)} + h^2 \|\nabla(\sigma - \tau_h)\|^2_{L^2(\Omega)} \right),$$
$$\eta(u) := \inf_{v_h \in \mathcal{P}^p_{pw}(\mathcal{T})} \left(\|u - v_h\|^2_{L^2(\Omega)} + h^2 \|\nabla(u - v_h)\|^2_{L^2(\Omega)} \right).$$

Theorem 24. *Assuming H^2-regularity of the Helmholtz problem, there exists a constant $C > 0$, independent of κ, h, α, β, such that*

$$\sum_{T \in \mathcal{T}} \alpha \|[u - u_h]\|^2_{L^2(\partial T)} + \beta \|[\![\sigma - \sigma_h]\!]\|^2_{L^2(\partial T)} + \|u - \hat{u}_h\|^2_{L^2(\partial\Omega)} \leq \frac{C}{h} \left(\left(\frac{2}{\alpha} + \beta \right) \eta(\sigma) + 2\alpha \eta(u) \right).$$

Considering the approximation properties of the discrete spaces, the convergence rates below directly follow.

Corollary 25. *Assuming H^{p+2}-regularity of the Helmholtz problem with a polynomial degree of p, there exists a constant $C > 0$, independent of κ, h, α, β, such that*

$$\sum_{T \in \mathcal{T}} \alpha \|[u - u_h]\|^2_{L^2(\partial T)} + \beta \|[\![\sigma - \sigma_h]\!]\|^2_{L^2(\partial T)} + \|u - \hat{u}_h\|^2_{L^2(\partial\Omega)} \leq C h^{2p+1} \left(\left(\frac{2}{\alpha} + \beta \right) \|\sigma\|^2_{H^{p+1}(\Omega)} + 2\alpha \|u\|^2_{H^{p+1}(\Omega)} \right).$$

These convergence rates are quasi-optimal in the mesh size and wave number. The first approximation result for the error in the pressure on elements is the following.

Theorem 26. *For a given wave number $\kappa > \kappa_0 > 0$, stabilisation parameters $\alpha = \mathcal{O}(1)$, $\beta = \mathcal{O}(1)$ and a star-shaped domain, assuming H^2-regularity of the Helmholtz problem and a polynomial degree $p \geq 1$, there exists a constant $C > 0$ independent of κ, h such that*

$$\|u - u_h\|^2_{L^2(\Omega)} \leq (1 + C(\kappa^2 + \kappa^4 h^2)) \left(\eta(\sigma) + \eta(u) \right).$$

With according approximation properties, there follows again:

Corollary 27. *For a given wave number $\kappa > \kappa_0 > 0$, stabilisation parameters $\alpha = \mathcal{O}(1)$, $\beta = \mathcal{O}(1)$ and a star-shaped domain, assuming H^2-regularity of the adjoint problem and a polynomial degree $p \geq 1$ and H^{p+2}-regularity of the Helmholtz problem there exists a constant $C > 0$, independent of κ, h, such that*

$$\|u - u_h\|_{L^2(\Omega)} \leq (1 + C(\kappa + \kappa^2 h))h^{p+1} \left(\|\sigma\|_{H^{p+1}(\Omega)} + \|u\|_{H^{p+1}(\Omega)}\right).$$

This result is sub-optimal in the wave number, as it predicts an enduring pollution effect even for small mesh sizes. In numerical experiments, this behaviour is not seen if polynomial spaces of order $p \geq 1$ are used. Only for the lowest order case $p = 0$ this holds. To further study this phenomenon a dispersion and dissipation analysis for the HDG formulation in one dimension has been carried out.

Due to the h-independent stabilisation, the quasi-best approximation for the flux cannot be straightforwardly derived. For the following result, a more refined technique based on projections is needed.

Theorem 28. *For a given wave number $\kappa > \kappa_0 > 0$, stabilisation parameters $\alpha = \mathcal{O}(1)$, $\beta = \mathcal{O}(1)$ and a star-shaped domain, assuming H^2-regularity of the Helmholtz problem and a polynomial degree $p \geq 1$, there exists a constant $C > 0$ independent of κ, h such that*

$$\|\sigma - \sigma_h\|^2_{L^2(\Omega)} \leq (1 + C(\kappa^2 + \kappa^4 h^2)) \left(\eta(\sigma) + \eta(u)\right).$$

Corollary 29. *For a given wave number $\kappa > \kappa_0 > 0$, stabilisation parameters $\alpha = \mathcal{O}(1)$, $\beta = \mathcal{O}(1)$ and a star-shaped domain, assuming H^2-regularity of the adjoint problem, H^{p+2}-regularity of the Helmholtz problem and a polynomial degree $p \geq 1$, there exists a constant $C > 0$ independent of κ, h such that*

$$\|\sigma - \sigma_h\|_{L^2(\Omega)} \leq (1 + C(\kappa + \kappa^2 h))h^{p+1} \left(\|\sigma\|_{H^{p+1}(\Omega)} + \|u\|_{H^{p+1}(\Omega)}\right).$$

In the same way, as for the pressure, this result is sub-optimal. Asymptotically, there holds the following optimal result.

Theorem 30. *For a given wave number $\kappa > \kappa_0 > 0$, stabilisation parameters $\alpha = \mathcal{O}(1)$, $\beta = \mathcal{O}(1)$ and a star-shaped domain, assuming H^2-regularity of the Helmholtz problem and a polynomial degree $p \geq 1$, there exist positive constants C and c_1 independent of κ, h such that under the assumption*

$$\kappa(\kappa h)^p < c_1$$

holds the quasi-best approximation

$$\|\sigma - \sigma_h\|^2_{L^2(\Omega)} + \|u - u_h\|^2_{L^2(\Omega)} \leq C \left(\eta(\sigma) + \eta(u)\right).$$

Assuming the higher regularity H^{p+2} of the Helmholtz problem leads to the convergence rate

$$\|\sigma - \sigma_h\|^2_{L^2(\Omega)} + \|u - u_h\|^2_{L^2(\Omega)} \leq Ch^{2p+2} \left(|\sigma|^2_{H^{p+1}} + |u|^2_{H^{p+1}}\right).$$

The contribution of [GM11], which is based upon the techniques developed in [CGS10], motivates the proof of this theorem. A comprehensive and thorough explanation of the later work can be found in [Say13]. In these studies, HDG methods with h-independent stabilisation are analysed. The HDG-formulation in [GM11] is similar to the formulation in this work. The first major difference is that only a single facet variable with a comparable α-stabilisation is used, and the second is that existence, uniqueness and optimal convergence rates are proven under the assumption of a resolution condition. Proving optimal rates for h-independently stabilised HDG-formulation is challenging, because element boundary terms need to be estimated via an inverse estimate with volume terms leading to sub-optimal rates. The idea is to use a, to the HDG-formulation tailored, projection into the discrete space, so that boundary terms vanish. For that purpose, a special projection has been established and analysed in [CGS10]. In this study, due to the second β-stabilisation, that projection cannot be applied. A generalisation needs to be established for the proof.

The general approach applies the Aubin-Nitsche trick, which is commonly used to analyse Helmholtz formulations, then interpolations are introduced, and continuity estimates are carried out. The most interesting parts are the interpolations as well as the combination of well-established techniques.

3.3.1. Favourable Properties of B

In the following lemma two equivalent forms of B are shown. They are based on applying partial integration and collecting the resulting element boundary terms together.

Lemma 31. *The sesquilinear form B is equivalent to the following two sesquilinear forms in the sense of*

$$B(\sigma,\hat{\sigma}_{\mathbf{n}},u,\hat{u};\tau,\hat{\tau}_{\mathbf{n}},v,\hat{v}) = B_I(\sigma,\hat{\sigma}_{\mathbf{n}},u,\hat{u};\tau,\hat{\tau}_{\mathbf{n}},v,\hat{v}) = B_{II}(\sigma,\hat{\sigma}_{\mathbf{n}},u,\hat{u};\tau,\hat{\tau}_{\mathbf{n}},v,\hat{v}),$$

where they are defined as

$$\begin{aligned} B_I(\sigma,\hat{\sigma}_{\mathbf{n}},u,\hat{u};\tau,\hat{\tau}_{\mathbf{n}},v,\hat{v}) := \sum_{T\in\mathcal{T}} \Big(& j\kappa(\sigma,\tau)_T - (\nabla u,\tau)_T + (\operatorname{div}\sigma,v)_T - j\kappa(u,v)_T \\ & + ([u],\tau\cdot\mathbf{n}+\alpha[v])_{\partial T} - (\llbracket\sigma\rrbracket,\beta\llbracket\tau\rrbracket+\hat{v})_{\partial T}\Big) + (\hat{u}-\hat{\sigma}\cdot\mathbf{n},\hat{v})_{\partial\Omega}, \end{aligned}$$

and

$$\begin{aligned} B_{II}(\sigma,\hat{\sigma}_{\mathbf{n}},u,\hat{u};\tau,\hat{\tau}_{\mathbf{n}},v,\hat{v}) := \sum_{T\in\mathcal{T}} \Big(& j\kappa(\sigma,\tau)_T + (u,\operatorname{div}\tau)_T - (\sigma,\nabla v)_T - j\kappa(u,v)_T \\ & + (\sigma\cdot\mathbf{n}+\alpha[u],[v])_{\partial T} - (\beta\llbracket\sigma\rrbracket+\hat{u},\llbracket\tau\rrbracket)_{\partial T}\Big) + (\hat{u},\hat{v}-\hat{\tau}\cdot\mathbf{n})_{\partial\Omega}. \end{aligned}$$

Proof. Partial integration gives

$$(u,\operatorname{div}\tau)_T = (u,\tau\cdot\mathbf{n})_{\partial T} - (\nabla u,\tau)_T,$$

and additionally introducing

$$0 = (\hat{\sigma}\cdot\mathbf{n}_+,\hat{v})_{F_I} + (\hat{\sigma}\cdot\mathbf{n}_-,\hat{v})_{F_I}$$

implies the equivalence between the sesquilinear forms. □

The sesquilinear form B satisfies the following weak coercivity and Garding inequality.

Lemma 32. *For the sesquilinear form B there hold*

$$-\Re B(\sigma, \hat{\sigma}_{\mathbf{n}}, u, \hat{u}; \sigma, \hat{\sigma}_{\mathbf{n}}, -u, -\hat{u}) = \sum_{T\in\mathcal{T}} \alpha\|[u]\|^2_{\partial T} + \beta\|[\![\sigma]\!]\|^2_{\partial T} + \|\hat{u}\|^2_{\partial\Omega},$$
$$\Im B(\sigma, \hat{\sigma}_{\mathbf{n}}, u, \hat{u}; \sigma, \hat{\sigma}_{\mathbf{n}}, u, \hat{u}) = \sum_{T\in\mathcal{T}} \kappa\|\sigma\|^2_T - \kappa\|u\|^2_T.$$

Proof. Inserting the specified test and trial functions and considering the real or imaginary part directly leads to the stated results. □

The interesting aspect is that the real part controls the skeleton terms and the imaginary part the volume terms. And these two estimates are decoupled from each other.

3.3.2. Discrete Absolute Stability

Usually, for finite element discretisations of the Helmholtz equation, a resolution condition is required to prove the existence and uniqueness of discrete solutions as well as quasi-optimality. These results hold in the asymptotic regime. Several publications have established the existence and uniqueness of DG and HDG methods without such a requirement, e.g. [FW09, FW11, FX13, MPS13, MS14, Wu14, FLX16]. The idea has been to use the discrete test function $\mathbf{x}\cdot\nabla u_h$, which is contained in the discrete, discontinuous space and leads to positive volume terms. The disadvantage is that the pre-asymptotic and the asymptotic analysis require different techniques and that the domain is restricted to be star-shaped.

A new approach based upon L^2-projections, which leads to results viable in the pre-asymptotic as well as the asymptotic regime, has been developed in [LCQ17]. In this section, the technique therein is applied to the HDG-formulation (3.7), with the additional consideration of a second facet variable and β-stabilisation as well as h-independent α-stabilisation.

The following stability estimates are purely a result of the fact that the HDG-formulation satisfies a weak coercivity for element boundary terms and a Guarding-inequality. Then, just continuity estimates for the right hand side terms are used. This lemma and the proof of it are similar to [LCQ17, Lemma 3.1].

Lemma 33. *For the discrete solution of (3.7) there holds*

$$\|\sigma, \hat{\sigma}_n, u, \hat{u}\|^2_\partial + \frac{1}{2}\|\hat{u}_h\|^2_{\partial\Omega} \leq \frac{1}{\kappa}\|f\|_\Omega\|u_h\|_\Omega + \frac{1}{2\kappa^2}\|g\|^2_{\partial\Omega}, \tag{3.8}$$
$$\sum_{T\in\mathcal{T}} \kappa\|\sigma_h\|^2_T - \kappa\|u_h\|^2_T \leq \frac{2}{\kappa}\|f\|_\Omega\|u_h\|_\Omega + \frac{1}{\kappa^2}\|g\|^2_{\partial\Omega} \tag{3.9}$$

with the element boundary norm defined by

$$\|\sigma, \hat{\sigma}_{\mathbf{n}}, u, \hat{u}\|^2_\partial := \sum_{T\in\mathcal{T}} \alpha\|[u_h]\|^2_{\partial T} + \beta\|[\![\sigma_h]\!]\|^2_{\partial T}.$$

Proof. Considering the real part in Lemma 32 leads to

$$\sum_{T\in\mathcal{T}} \alpha\|[u_h]\|_{\partial T}^2 + \beta\|[\![\sigma_h]\!]\|_{\partial T}^2 + \|\hat{u}_h\|_{\partial\Omega}^2 = \Re\frac{j}{\kappa}(f,u_h)_\Omega + \Re\frac{j}{\kappa}(g,\hat{u}_h)_{\partial\Omega}$$

directly implying (3.8). Similarly, taking the imaginary part in Lemma 32 yields

$$\sum_{T\in\mathcal{T}} \kappa\|\sigma_h\|_T^2 - \kappa\|u_h\|_T^2 = \Im\frac{j}{\kappa}(f,u_h)_\Omega + \Im\frac{j}{\kappa}(g,\hat{u}_h)_{\partial\Omega},$$

and with (3.8), which implies $\|\hat{u}_h\|_{\partial\Omega}^2 \leq \frac{2}{\kappa}\|f\|_\Omega\|u_h\|_\Omega + \frac{1}{\kappa^2}\|g\|_{\partial\Omega}^2$, gives (3.9). □

As can be seen the stability only depends on bounding u_h by the excitation. In the following theorem, u_h will be bounded by f and g, which concludes the stability analysis of the HDG-formulation. The proof is based upon [LCQ17, Lemma 3.2].

Theorem 34. *Assuming H^s-regularity of the adjoint Helmholtz problem with $s > 3/2$ and a polynomial degree $p \geq s-1$, there holds for the discrete solution of the HDG-formulation (3.7)*

$$\sum_{T\in\mathcal{T}} \alpha\|[u_h]\|_{L^2(\partial T)}^2 + \beta\|[\![\sigma_h]\!]\|_{L^2(\partial T)}^2 + \frac{1}{2}\|\hat{u}_h\|_{L^2(\partial\Omega)}^2 \leq \frac{1}{\kappa}\|f\|_{L^2(\Omega)}\|u_h\|_{L^2(\Omega)} + \frac{1}{2\kappa^2}\|g\|_{L^2(\partial\Omega)}^2,$$

$$\sum_{T\in\mathcal{T}} \kappa\|\sigma_h\|_{L^2(T)}^2 \leq \sum_{T\in\mathcal{T}} \kappa\|u_h\|_{L^2(T)}^2 + \frac{2}{\kappa}\|f\|_{L^2(\Omega)}\|u_h\|_{L^2(\Omega)} + \frac{1}{\kappa^2}\|g\|_{L^2(\partial\Omega)}^2,$$

$$\begin{aligned}\|u_h\|_{L^2(\Omega)} \leq{}& \left(C(\Omega,\kappa,\alpha,\beta,h)^2 + 2C_{f,0}(\Omega,\kappa)\kappa\right)\frac{\|f\|_{L^2(\Omega)}}{\kappa} \\ &+ \left(\sqrt{2}C(\Omega,\kappa,\alpha,\beta,h) + 2\sqrt{C_{f,0}(\Omega,\kappa)\kappa}\right)\frac{\|g\|_{L^2(\partial\Omega)}}{\kappa}\end{aligned}$$

with the constant

$$C(\Omega,\kappa,\alpha,\beta,h)^2 := \left(\left(\frac{2}{\alpha}+\beta\right)C_1^2(\Omega)h^{2s-3} + 2\alpha C_2^2(\Omega)\kappa^2 h^{2s-1}\right)C_{f,s}^2(\Omega,\kappa)$$

and the stability constants of the adjoint Helmholtz problem $C_{f,0}(\Omega,\kappa)$, $C_{f,s}(\Omega,\kappa)$ in Definition 13, as well as interpolation constants $C_1(\Omega)$ and $C_2(\Omega)$, see Lemma 38.

Proof. Using an Aubin-Nitsche trick for the adjoint problem, by considering as excitation $f = j\kappa u_h$, yields in combination with the adjoint consistency, see Lemma 35,

$$\|u_h\|_\Omega^2 = B(\sigma_h,\hat{\sigma}_{\mathbf{n},h},u_h,\hat{u}_h;\phi,\phi_{\mathbf{n}},w,w),$$

with $(\sigma_h,\hat{\sigma}_{\mathbf{n},h},u_h,\hat{u}_h)$ as test functions. At this point in standard theory, the continuity of the sesquilinear form is used. As this is not applicable in this case, projections are

introduced. Adding and subtracting the element L^2-projections Π and facet L^2-projections Π_F, see Definition 37, leads to

$$\begin{aligned}\|u_h\|_\Omega^2 = &B(\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u_h, \hat{u}_h; \phi - \Pi\phi, \phi_{\mathbf{n}} - \Pi_F\phi_{\mathbf{n}}, w - \Pi w, w - \Pi_F w)\\ &+ B(\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u_h, \hat{u}_h; \Pi\phi, \Pi_F\phi_{\mathbf{n}}, \Pi w, \Pi_F w).\end{aligned}$$

For the first term, the following continuity estimate holds due to the nature of the used projections and by applying the Cauchy-Schwarz inequality to the identity in Lemma 39

$$\begin{aligned}&B(\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u_h, \hat{u}_h; \phi - \Pi\phi, \phi_{\mathbf{n}} - \Pi_F\phi_{\mathbf{n}}, w - \Pi w, w - \Pi_F w)\\ &\leq \|\sigma_h, \hat{\sigma}_{h,\mathbf{n}}, u_h, \hat{u}_h\|_\partial \left(\sum_{T\in\mathcal{T}} \left(\frac{2}{\alpha} + \beta\right) \|(\phi - \Pi\phi)\cdot\mathbf{n}\|_{\partial T}^2 + 2\alpha\|w - \Pi w\|_{\partial T}^2 \right)^{\frac{1}{2}}.\end{aligned}$$

Due to the assumed regularity of the adjoint solution, the approximation properties of the projections, see Lemma 38, and the stability of the Helmholtz equation, see Definition 13, there further holds

$$\begin{aligned}&\sum_{T\in\mathcal{T}} \left(\frac{2}{\alpha} + \beta\right) \|(\phi - \Pi\phi)\cdot\mathbf{n}\|_{\partial T}^2 + 2\alpha\|w - \Pi w\|_{\partial T}^2\\ &\quad\leq \left(\frac{2}{\alpha} + \beta\right) C_1^2(\Omega) h^{2s-3} \|\phi\|_{H^{s-1}(\Omega)}^2 + 2\alpha C_2^2(\Omega) h^{2s-1} \|w\|_{H^s(\Omega)}^2\\ &\quad\leq \left(\left(\frac{2}{\alpha} + \beta\right) C_1^2(\Omega) \frac{h^{2s-3}}{\kappa^2} + 2\alpha C_2^2(\Omega) h^{2s-1} \right) C_{f,s}^2(\Omega,\kappa)\kappa^2 \|u_h\|_{L^2(\Omega)}^2.\end{aligned}$$

The leading constant will get the short notation

$$C(\Omega,\kappa,\alpha,\beta,h)^2 = \left(\left(\frac{2}{\alpha} + \beta\right) C_1^2(\Omega) \frac{h^{2s-3}}{\kappa^2} + 2\alpha C_2^2(\Omega) h^{2s-1} \right) C_{f,s}^2(\Omega,\kappa)\kappa^2.$$

Finally, only the element boundary norm needs to be bounded with Lemma 33

$$\begin{aligned}\|\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u, \hat{u}\|_\partial &\leq \left(\frac{1}{\kappa}\|f\|_\Omega \|u_h\|_\Omega + \frac{1}{2\kappa^2}\|g\|_{\partial\Omega}^2 \right)^{1/2}\\ &\leq \frac{C(\Omega,\kappa,\alpha,\beta,h)}{2\kappa}\|f\|_\Omega + \frac{1}{2C(\Omega,\kappa,\alpha,\beta,h)}\|u_h\|_\Omega + \frac{1}{\sqrt{2}\kappa}\|g\|_{\partial\Omega}.\end{aligned}$$

An absorption argument will be needed.

For the second term there holds

$$B(\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u_h, \hat{u}_h; \Pi\phi, \Pi_F\phi_{\mathbf{n}}, \Pi w, \Pi_F w) = \frac{j}{\kappa}(f, \Pi w)_\Omega + \frac{j}{\kappa}(g, \Pi_F w)_{\partial\Omega},$$

because the projected adjoint solution is in the discrete space $\mathcal{X}_h$. Using the Cauchy-Schwarz inequality and further applying the continuity of the L^2-projection, the adjoint regularity in Definition 13 and Lemma 36 yields

$$\frac{1}{\kappa}\|f\|_\Omega\|\Pi w\|_\Omega + \frac{1}{\kappa}\|g\|_{\partial\Omega}\|\Pi_F w\|_{\partial\Omega} \leq \left(C_{f,0}(\Omega,\kappa)\|f\|_\Omega + \sqrt{\frac{C_{f,0}(\Omega,\kappa)}{\kappa}}\|g\|_{\partial\Omega} \right)\|u_h\|_\Omega.$$

Combining these estimates leads to

$$\|u_h\|_\Omega \leq \frac{C(\Omega,\kappa,\alpha,\beta,h)^2}{2\kappa}\|f\|_\Omega + \frac{1}{2}\|u_h\|_\Omega + \frac{C(\Omega,\kappa,\alpha,\beta,h)}{\sqrt{2}\kappa}\|g\|_{\partial\Omega} \\ + C_{f,0}(\Omega,\kappa)\|f\|_\Omega + \sqrt{\frac{C_{f,0}(\Omega,\kappa)}{\kappa}}\|g\|_{\partial\Omega}$$

concluding the proof. □

In the following, the main stability theorem is proven.

Proof of Theorem 23. On a star-shaped domain with H^2-regularity, the constants in the previous theorem are

$$s = 2, \qquad C_{f,0} = \mathcal{O}(1), \qquad C_{f,2} = \mathcal{O}(\kappa),$$

which leads for a positive wave number directly to the stated result. □

The following lemma establishes the consistency of the adjoint Helmholtz problem in Definition 19.

Lemma 35 (Adjoint Consistency)**.** *The adjoint solution* (ϕ, w) *of* (3.4a) - (3.4c)*, satisfies*

$$B(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h; \phi, \phi\cdot\mathbf{n}, w, w) = \frac{j}{\kappa}(v_h, f)_\Omega,$$

for all $(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \in \mathcal{X}_h$.

Proof. The HDG-formulation is complex symmetric in the sense of

$$B(\overline{\tau}, \overline{\hat{\tau}_{\mathbf{n}}}, \overline{v}, \overline{\hat{v}}; \overline{\sigma}, \overline{\hat{\sigma}_{\mathbf{n}}}, \overline{w}, \overline{\hat{w}}) = B(\sigma, \hat{\sigma}_{\mathbf{n}}, u, \hat{u}; \tau, \hat{\tau}_{\mathbf{n}}, v, \hat{v}).$$

Additionally, the conjugated adjoint solution is exactly the solution of the mixed Helmholtz equation with $\overline{f}$ as volume excitation, therefore

$$B(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h; \phi, \phi\cdot\mathbf{n}, w, w) = B(\overline{\phi}, \overline{\hat{\phi}_{\mathbf{n}}}, \overline{w}, \overline{\hat{w}}; \overline{\tau_h}, \overline{\hat{\tau}_{\mathbf{n},h}}, \overline{v_h}, \overline{\hat{v}_h}) = \frac{j}{\kappa}(\overline{f}, \overline{v_h})_\Omega = \frac{j}{\kappa}(v_h, f)_\Omega.$$

□

Lemma 36. *For the solution* (ϕ, w) *of* (3.4a) - (3.4c) *there holds*

$$\|w\|^2_{\partial\Omega} = \|\phi\cdot\mathbf{n}\|^2_{\partial\Omega} \leq \frac{C_{f,0}(\Omega,\kappa)}{\kappa}\|f\|^2_\Omega,$$

with the stability constant $C_{f,0}$.

Proof. The first equality immediately follows from (3.4c). Multiplying (3.4a) and (3.4b) by $\overline{\phi}$ and $\overline{w}$ respectively gives

$$j\kappa\|\phi\|_\Omega^2 + (\nabla w, \phi)_\Omega = 0,$$
$$j\kappa\|w\|_\Omega^2 + (\operatorname{div}\phi, w)_\Omega = -\frac{j}{\kappa}(f, w)_\Omega.$$

Partial integration of $(\nabla w, \phi)_\Omega$ leads to

$$j\kappa\|\phi\|_\Omega^2 - (w, \operatorname{div}\phi)_\Omega + (w, \phi\cdot\mathbf{n})_{\partial\Omega} = 0.$$

By conjugating and adding the equations the terms $(\operatorname{div}\phi, w)_\Omega$ cancel out and

$$j\kappa\|w\|_\Omega^2 - j\kappa\|\phi\|_\Omega^2 + (\phi\cdot\mathbf{n}, w)_{\partial\Omega} = -\frac{j}{\kappa}(f, w)_\Omega$$

remains. Using (3.4c) and only considering the real part in combination with the adjoint regularity implies

$$\|\phi\cdot\mathbf{n}\|_{\partial\Omega}^2 \leq \frac{1}{\kappa}\|f\|_\Omega\|w\|_\Omega \leq \frac{C_{f,0}(\Omega,\kappa)}{\kappa}\|f\|_\Omega^2.$$

□

The following lemma is essential for pre-asymptotic stability. It highlights the effect of introducing L^2-projections of σ and u defined by

Definition 37.

$$(\Pi\sigma, \tau_h)_T = (\sigma, \tau_h)_T \qquad \forall\tau_h \in \mathcal{RT}_{pw}^p(\mathcal{T}),$$
$$(\Pi u, v_h)_T = (u, v_h)_T \qquad \forall v_h \in \mathcal{P}_{pw}^p(\mathcal{T}),$$
$$(\Pi_F u, \hat{v}_h)_F = (u, \hat{v}_h)_F \qquad \forall\hat{v}_h \in \mathcal{P}_{pw}^p(\mathcal{F}),$$

into the SLF. The proof follows along the lines of [LCQ17, Lemma 3.2].

For the proof of the main result of this subsection, the following standard approximation properties of L^2-projections are required.

Lemma 38 (Approximation Properties of L^2-projections)**.** *Assuming H^{s+1}-regularity of ϕ, H^{t+1}-regularity of w, with $s, t \in \mathbb{R}_+$. If the polynomial degree of the discrete spaces satisfies $p \geq s \geq 0$, $p \geq t \geq 0$, then there exist constants $C_1(\Omega) > 0$, $C_2(\Omega) > 0$ independent of h so that*

$$\sum_{T\in\mathcal{T}}\|(\phi - \Pi\phi)\cdot\mathbf{n}\|_{\partial T}^2 \leq C_1^2(\Omega)h^{2s+1}\|\phi\|_{H^{s+1}}^2,$$
$$\sum_{T\in\mathcal{T}}\|w - \Pi w\|_{\partial T}^2 \leq C_2^2(\Omega)h^{2t+1}\|w\|_{H^{t+1}}^2.$$

Additionally, the best approximation properties

$$\|\phi - \Pi\phi\|_{L^2(\Omega)} = \inf_{\tau_h \in \mathcal{RT}^p_{pw}(\mathcal{T})} \|\phi - \tau_h\|_{L^2(\Omega)},$$
$$\|w - \Pi w\|_{L^2(\Omega)} = \inf_{v_h \in \mathcal{P}^p_{pw}(\mathcal{T})} \|w - v_h\|_{L^2(\Omega)},$$

$$\|\phi \cdot \mathbf{n} - \Pi\phi \cdot \mathbf{n}\|_{\partial T} \leq C \inf_{\tau_h \in \mathcal{RT}^p_{pw}(T)} (h^{-1/2}\|\phi - \tau_h\|_{L^2(T)} + h^{1/2}|\phi - \tau_h|_T),$$

$$\|w - \Pi w\|_{\partial T} \leq C \inf_{v_h \in \mathcal{P}^p_{pw}(T)} (h^{-1/2}\|w - v_h\|_{L^2(T)} + h^{1/2}|w - v_h|_T)$$

hold with a positive mesh size independent constant C.

Lemma 39. *Using the L^2-projection Π on elements and the L^2-projection Π_F on facets, there holds for arbitrary $\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h \in \mathcal{X}_h$*

$$\begin{aligned} &B(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h; \phi - \Pi\phi, \phi \cdot \mathbf{n} - \Pi_F\phi \cdot \mathbf{n}, w - \Pi w, w - \Pi_F w) \\ &\quad = \sum_{T \in \mathcal{T}} ([v_h], (\phi - \Pi\phi) \cdot \mathbf{n} + \alpha(w - \Pi w))_{\partial T} - \beta([\![\tau_h]\!], (\phi - \Pi\phi) \cdot \mathbf{n})_{\partial T}. \end{aligned}$$

Proof. The form B_I in Lemma 31 will be used. Due to the L^2-projection properties the terms

$$\begin{aligned} j\kappa(\tau_h, \phi - \Pi\phi)_T = 0, (\operatorname{div} \tau_h, w - \Pi w)_T = 0, -j\kappa(v_h, w - \Pi w)_T = 0, \\ -(\nabla v_h, \phi - \Pi\phi)_T = 0, -([\![\tau_h]\!], -\beta(\phi - \Pi_F\phi) \cdot \mathbf{n} + w - \Pi_F w)_{\partial T} = 0, \\ -\alpha([v_h], w - \Pi_F w)_{\partial T} = 0, (\hat{v}_h - \hat{\tau}_h \cdot \mathbf{n}, w - \Pi_F w)_{\partial\Omega} = 0 \end{aligned}$$

vanish. Incorporating these changes yields

$$\begin{aligned} &B(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h; \phi - \Pi\phi, \phi_{\mathbf{n}} - \Pi_F\phi_{\mathbf{n}}, w - \Pi w, w - \Pi_F w) \\ &\quad = \sum_{T \in \mathcal{T}} ([v_h], (\phi - \Pi\phi) \cdot \mathbf{n} + \alpha(w - \Pi w))_{\partial T} - \beta([\![\tau_h]\!], (\phi - \Pi\phi) \cdot \mathbf{n})_{\partial T}. \end{aligned}$$

□

3.3.3. Pre-Asymptotic Error Estimates for the Pressure and Jumps

In this section, Theorem 24 and Theorem 42 are proven. The analysis is similar to the proof of stability in the last section and leans on [LCQ17]. The continuous and discrete solutions satisfy

$$B(\sigma, \sigma_{\mathbf{n}}, u, u, \tau_h; \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) = \frac{j}{\kappa}(f, v_h)_\Omega + \frac{j}{\kappa}(g, \hat{v}_h)_{\partial\Omega},$$
$$B(\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u_h, \hat{u}_h; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) = \frac{j}{\kappa}(f, v_h)_\Omega + \frac{j}{\kappa}(g, \hat{v}_h)_{\partial\Omega},$$

for all $(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \in \mathcal{X}_h$. In the case of coercive weak formulations, optimal convergence rates are proven with coercivity, Galerkin orthogonality and continuity. The Helmholtz equation does not satisfy coercivity therefore, other techniques need to be established. Usually, a Schatz argument is applied to derive asymptotic results, but these depend on a resolution condition. The L^2-projections circumvent this necessity.

For the analysis, the following projected errors, contained in the discontinuous space $\mathcal{X}_h$, are needed.

Definition 40.

$$e_\sigma := \Pi\sigma - \sigma_h, \qquad e_u := \Pi u - u_h, \qquad e_{\hat{u}} := \Pi_F u - \hat{u}_h, \qquad e_{\hat{\sigma}} := \Pi_F \sigma \cdot \mathbf{n} - \hat{\sigma}_{\mathbf{n},h}. \tag{3.10}$$

The usage of projections leads to a splitting of error estimates into e.g.

$$\|u - u_h\|_\Omega = \|u - \Pi u + \Pi u - u_h\|_\Omega \leq \|u - \Pi u\|_\Omega + \|\Pi u - u_h\|_\Omega.$$

The first part only depends on the approximation properties of the projection. The second part holds the advantage that e_u is a viable choice as a discrete test function. Therefore, if the projection has optimal approximation properties and if the projected error has an optimal convergence rate, then the discrete solution has an optimal rate as well.

For the projected errors, the Galerkin orthogonality does not hold, but they satisfy the following discrete weak formulation.

Lemma 41. *The projected errors in (3.10) satisfy the weak formulation*

$$\begin{aligned} &B(e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \\ &\quad = -\sum_{T\in\mathcal{T}} (\sigma\cdot\mathbf{n} - \Pi\sigma\cdot\mathbf{n}, [v_h])_{\partial T} + \alpha(u - \Pi u, [v_h])_{\partial T} - \beta(\sigma\cdot\mathbf{n} - \Pi\sigma\cdot\mathbf{n}, [\![\tau_h]\!])_{\partial T}, \end{aligned}$$

for all $(\tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \in \mathcal{X}_h$.

Proof. There holds due to the Galerkin orthogonality

$$\begin{aligned} 0 &= B(\sigma, \sigma_\mathbf{n}, u, u; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) - B(\sigma_h, \hat{\sigma}_{\mathbf{n},h}, u_h, \hat{u}_h; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \\ &= B(\sigma - \Pi\sigma, \sigma_\mathbf{n} - \Pi_F\sigma\cdot\mathbf{n}, u - \Pi u, u - \Pi_F u; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \\ &\quad - B(e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \end{aligned}$$

and with a similar argument as in Lemma 39, there follows

$$\begin{aligned} &B_{II}(\sigma - \Pi\sigma, \sigma_\mathbf{n} - \Pi_F\sigma\cdot\mathbf{n}, u - \Pi u, u - \Pi_F u; \tau_h, \hat{\tau}_{\mathbf{n},h}, v_h, \hat{v}_h) \\ &\qquad = \sum_{T\in\mathcal{T}} ((\sigma - \Pi\sigma)\cdot\mathbf{n} + \alpha(u - \Pi u), [v_h])_{\partial T} - \beta((\sigma - \Pi\sigma)\cdot\mathbf{n}, [\![\tau_h]\!])_{\partial T}. \end{aligned}$$

□

With the previous lemma, Theorem 24 can be shown.

Proof of Theorem 24. The proof is similar to [LCQ17, Lemma 4.2]. Setting $\tau_h := e_\sigma, \hat{\tau}_{\mathbf{n},h} := e_{\hat{\sigma}}, v_h := -e_u, \hat{v}_h := -e_{\hat{u}}$ in Lemma 41 and considering the real part according to Lemma 32 gives

$$\begin{aligned}\|e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}\|_\partial^2 + \|e_{\hat{u}}\|_{\partial\Omega}^2 \\ = -\Re \sum_{T\in\mathcal{T}} (\sigma\cdot\mathbf{n} - \Pi\sigma\cdot\mathbf{n} + \alpha(u-\Pi u), [e_u])_{\partial T} - \beta(\Pi\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}, [\![e_\sigma]\!])_{\partial T}.\end{aligned}$$

The right hand side can be estimated by

$$\begin{aligned}&\left|\Re \sum_{T\in\mathcal{T}} (\sigma\cdot\mathbf{n} - \Pi\sigma\cdot\mathbf{n} + \alpha(u-\Pi u), [e_u])_{\partial T} - \beta(\Pi\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}, [\![e_\sigma]\!])_{\partial T}\right| \\ &\leq \left(\sum_{T\in\mathcal{T}} \left(\frac{2}{\alpha}+\beta\right)\|\Pi\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}\|_{\partial T}^2 + 2\alpha\|\Pi u - u\|_{\partial T}^2\right)^{\frac{1}{2}} C\|e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}\|_\partial\end{aligned}$$

and due to the projection properties, there holds

$$\sum_{T\in\mathcal{T}} \left(\frac{2}{\alpha}+\beta\right)\|\Pi\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}\|_{\partial T}^2 + 2\alpha\|\Pi u - u\|_{\partial T}^2 \leq \frac{C}{h}\left(\left(\frac{2}{\alpha}+\beta\right)\eta(\sigma) + 2\alpha\eta(u)\right)$$

with positive constants $C > 0$, concluding the proof. □

The previous result states quasi-optimal convergence rates for jumps. In the following step, Theorem 42 is proven, similarly to [LCQ17, Lemma 4.3], by applying an Aubin-Nitsche trick with the right hand side $f = j\kappa e_u$.

Theorem 42. *Assuming H^s-regularity of the adjoint problem with $s > 3/2$ and H^2-regularity of the Helmholtz problem, there exists a constant $C > 0$, independent of κ, h, α, β, such that*

$$\|u-u_h\|_{L^2(\Omega)}^2 \leq \eta(u) + \left(C(\Omega,\kappa,\alpha,\beta,h) + C(\Omega,\kappa,h)\right)^2 \frac{C}{h}\left(\left(\frac{2}{\alpha}+\beta\right)\eta(\sigma) + 2\alpha\eta(u)\right)$$

with the constant

$$C(\Omega,\kappa,h)^2 := 2\left(C_1(\Omega)h^{2s-3} + C_2(\Omega)\kappa^2 h^{2s-1}\right) C_{f,s}^2(\Omega,\kappa)$$

and the adjoint stability constants $C_{f,s}(\Omega,\kappa)$ in Definition 13 as well as interpolation constants $C_1(\Omega)$ and $C_2(\Omega)$, see Lemma 38.

Proof. Using an Aubin-Nitsche technique for the adjoint problem and inserting projections Π and Π_F, as well as applying Lemma 39 and Lemma 41, yields

$$\begin{aligned}\|e_u\|_\Omega^2 =& B(e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}; \phi, \phi_{\mathbf{n}}, w, w) \\ =& B(e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}; \phi - \Pi\phi, \phi_{\mathbf{n}} - \Pi_F\phi_{\mathbf{n}}, w - \Pi w, w - \Pi_F w) \\ &+ B(e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}; \Pi\phi, \Pi_F\phi_{\mathbf{n}}, \Pi w, \Pi_F w) \\ =& \sum_{T\in\mathcal{T}} ([e_u], (\phi-\Pi\phi)\cdot\mathbf{n} + \alpha(w-\Pi w))_{\partial T} - \beta([\![e_\sigma]\!], (\phi-\Pi\phi)\cdot\mathbf{n})_{\partial T} \\ &+ \sum_{T\in\mathcal{T}} ((\Pi\sigma-\sigma)\cdot\mathbf{n} + \alpha(\Pi u - u), [\Pi w])_{\partial T} - \beta(\Pi\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}, [\![\Pi\phi]\!])_{\partial T}.\end{aligned}$$

The first term can be estimated in the same fashion as in the proof for the absolute stability

$$\sum_{T\in\mathcal{T}}([e_u],(\phi-\Pi\phi)\cdot\mathbf{n}+\alpha(w-\Pi w))_{\partial T}-\beta([\![e_\sigma]\!],(\phi-\Pi\phi)\cdot\mathbf{n})_{\partial T}$$
$$\leq C(\Omega,\kappa,\alpha,\beta,h)\|e_\sigma,e_{\hat{\sigma}},e_u,e_{\hat{u}}\|_\partial$$

and the second term has the same structure, therefore, estimating the terms leads to

$$\sum_{T\in\mathcal{T}}((\Pi\sigma-\sigma)\cdot\mathbf{n}+\alpha(\Pi u-u),[\Pi w])_{\partial T}-\beta(\Pi\sigma\cdot\mathbf{n}-\sigma\cdot\mathbf{n},[\![\Pi\phi]\!])_{\partial T}$$
$$\leq\left(\sum_{T\in\mathcal{T}}\left(\frac{2}{\alpha}+\beta\right)\|(\sigma-\Pi\sigma)\cdot\mathbf{n}\|^2_{\partial T}+2\alpha\|u-\Pi u\|^2_{\partial T}\right)^{\frac{1}{2}}\|\Pi\phi,\Pi_F\phi\cdot\mathbf{n},\Pi w,\Pi_F w\|_\partial.$$

The jumps of the adjoint problem can be bounded by

$$\|\Pi\phi,\Pi_F\phi\cdot\mathbf{n},\Pi w,\Pi_F w\|^2_\partial\leq 2\left(C_1(\Omega)h^{2s-3}+C_2(\Omega)\kappa^2h^{2s-1}\right)C^2_{f,s}(\Omega,\kappa)\|e_u\|^2_\Omega.$$

Introducing the short notation

$$C(\Omega,\kappa,h)^2:=2\left(C_1(\Omega)h^{2s-3}+C_2(\Omega)\kappa^2h^{2s-1}\right)C^2_{f,s}(\Omega,\kappa)$$

and inserting both estimates in the Aubin-Nitsche trick gives

$$\begin{aligned}\|e_u\|^2_\Omega\leq&C(\Omega,\kappa,\alpha,\beta,h)\|e_\sigma,e_{\hat{\sigma}},e_u,e_{\hat{u}}\|_\partial\\&+C(\Omega,\kappa,h)\left(\sum_{T\in\mathcal{T}}\left(\frac{2}{\alpha}+\beta\right)\|(\sigma-\Pi\sigma)\cdot\mathbf{n}\|^2_{\partial T}+2\alpha\|u-\Pi u\|^2_{\partial T}\right)^{\frac{1}{2}}\|e_u\|_\Omega\\\leq&\left(C(\Omega,\kappa,\alpha,\beta,h)+C(\Omega,\kappa,h)\right)\sqrt{\frac{C}{h}}\left(\left(\frac{2}{\alpha}+\beta\right)\eta(\sigma)+2\alpha\eta(u)\right)^{\frac{1}{2}}\|e_u\|_\Omega\end{aligned}$$

which concludes the proof. □

Proof of Theorem 26. On a star-shaped domain with H^2-regularity, the constants in the previous theorem are

$$s=2,\qquad C_{f,0}=\mathcal{O}(1),\qquad C_{f,2}=\mathcal{O}(\kappa),$$

which leads for a positive wave number directly to the stated result. □

The applied Aubin-Nitsche technique is commonly used to prove the superconvergence of the projected error e_u. For the HDG-formulation, this is partially true in the pre-asymptotic analysis. The additional rate, generated due to the reasoning of the adjoint regularity, compensates for the otherwise sub-optimal rate by estimating boundary through volume terms.

First Error Estimate for the Flux

Due to Lemma 32, the following error estimation for the flux is possible:

$$\sum_{T\in\mathcal{T}} \kappa\|e_\sigma\|_T^2 = \sum_{T\in\mathcal{T}} \kappa\|e_u\|_T^2 + \Im B(e_\sigma, e_{\hat{\sigma}_\mathbf{n}}, e_u, e_{\hat{u}}; e_\sigma, e_{\hat{\sigma}_\mathbf{n}}, e_u, e_{\hat{u}}).$$

Due to the Galerkin orthogonality, there holds

$$\begin{aligned}&\Im B(e_\sigma, e_{\hat{\sigma}_\mathbf{n}}, e_u, e_{\hat{u}}; e_\sigma, e_{\hat{\sigma}_\mathbf{n}}, e_u, e_{\hat{u}})\\ &\qquad = \Im B(\Pi\sigma - \sigma, \Pi_F\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}, \Pi u - u, \Pi_F u - u; e_\sigma, e_{\hat{\sigma}_\mathbf{n}}, e_u, e_{\hat{u}}).\end{aligned}$$

With a similar argument as in Lemma 39, there follows

$$\begin{aligned}&B(\Pi\sigma - \sigma, \Pi_F\sigma\cdot\mathbf{n} - \sigma\cdot\mathbf{n}, \Pi u - u, \Pi_F u - u; e_\sigma, e_{\hat{\sigma}_\mathbf{n}}, e_u, e_{\hat{u}})\\ &\quad = \sum_{T\in\mathcal{T}} ((\Pi\sigma - \sigma)\cdot\mathbf{n} + \alpha(\Pi u - u), [e_u])_{\partial T} - \beta((\Pi\sigma - \sigma)\cdot\mathbf{n}, [\![e_\sigma]\!])_{\partial T}\\ &\quad \le \left(\sum_{T\in\mathcal{T}} \left(\frac{2}{\alpha} + \beta\right)\|(\sigma - \Pi\sigma)\cdot\mathbf{n}\|_{\partial T}^2 + 2\alpha\|u - \Pi u\|_{\partial T}^2\right)^{\frac{1}{2}} \|e_\sigma, e_{\hat{\sigma}}, e_u, e_{\hat{u}}\|_\partial.\end{aligned}$$

All parts can be estimated in the same manner as above, leading to

$$\sum_{T\in\mathcal{T}} \kappa\|e_\sigma\|_T^2 \le \left(1 + (C(\Omega,\kappa,\alpha,\beta,h) + C(\Omega,\kappa,h))^2\,\kappa\right)\frac{C}{h}\left(\left(\frac{2}{\alpha} + \beta\right)\eta(\sigma) + 2\alpha\eta(u)\right).$$

This result is sub-optimal with respect to the mesh size and the wave number.

3.3.4. Pre-Asymptotic Error Estimate for the Flux

To prove optimal convergence rates for the flux, with respect to the mesh size, the issue due to h-independent stabilisation needs to be overcome. In [CGS10], a method based upon a specially devised projection is established, and in [GM11] this method is applied to an HDG-formulation of the Helmholtz equation to asymptotically show optimal rates. Removing the second facet variable $\hat{\sigma}_\mathbf{n}$ and the β-stabilisation in (3.7) would lead to the formulation therein. Therefore, similar steps as in [GM11] lead to the desired result, but they differ due to the β-stabilisation. For $\beta = 0$, the projection introduced in this work falls back to the projection in [CGS10].

The next step towards an error bound for the flux is to introduce a suitable projection.

Lemma 43. *Let P be the interpolation*

$$P : (\sigma, u) \mapsto (P_\sigma(\sigma,u), P_{\hat{\sigma}_\mathbf{n}}(\sigma,u), P_u(\sigma,u), P_{\hat{u}}(\sigma,u)) \in \mathcal{X}_h$$

defined by

$$(P_\sigma, \tau_h)_T = (\sigma, \tau_h)_T \qquad \tau_h \in [\mathcal{P}_{pw}^{p-1}(\mathcal{T})]^d, \tag{3.11a}$$

$$(P_u, v_h)_T = (u, v_h)_T \qquad v_h \in \mathcal{P}_{pw}^{p}(\mathcal{T}), \tag{3.11b}$$

$$\left(\frac{1}{\alpha} + \beta\right)((\sigma - P_{\sigma,+}) \cdot \mathbf{n}_+, \mu_h)_{\partial T} = -\left(u - P_{u,+} - \frac{\alpha\beta}{2}(P_{u,+} - P_{u,-}), \mu_h\right)_{\partial T} \qquad \mu_h \in \mathcal{P}_{pw}^{p}(\mathcal{F}_I), \tag{3.11c}$$

$$(\sigma \cdot \mathbf{n} - P_\sigma \cdot \mathbf{n}, \mu_h)_{F_O} = -\alpha(u - P_u, \mu_h)_{F_O} \qquad \mu_h \in \mathcal{P}_{pw}^{p}(\mathcal{F}_O), \tag{3.11d}$$

$$P_{\hat{\sigma}_\mathbf{n}}(\sigma, u) := \{P_\sigma(\sigma, u)\} \cdot \mathbf{n} \qquad \text{on } F_I, \tag{3.11e}$$

$$P_{\hat{\sigma}_\mathbf{n}}(\sigma, u) := P_\sigma(\sigma, u) \cdot \mathbf{n} \qquad \text{on } F_O,$$

$$P_{\hat{u}}(\sigma, u) := \{P_u(\sigma, u)\} + \frac{1}{2\alpha}[P_\sigma(\sigma, u)]_\mathbf{n} \qquad \text{on } F_I,$$

$$P_{\hat{u}}(u) := \Pi_F u \qquad \text{on } F_O.$$

With projected errors defined as

$$e_\sigma^P := P_\sigma(\sigma, u) - \sigma_h, \quad e_u^P := P_u(\sigma, u) - u_h, \quad e_{\hat{\sigma}}^P := P_{\hat{\sigma}_\mathbf{n}}(\sigma, u) - \hat{\sigma}_{\mathbf{n},h}, \quad e_{\hat{u}}^P := P_{\hat{u}}(\sigma, u) - \hat{u}_h$$

and the following short notations for mean values and jumps

$$\{P_\sigma(\sigma, u)\} := \frac{1}{2}\left(P_{\sigma,+}(\sigma, u) + P_{\sigma,-}(\sigma, u)\right),$$

$$\{P_u(\sigma, u)\} := \frac{1}{2}\left(P_{u,+}(\sigma, u) + P_{u,-}(\sigma, u)\right),$$

$$[P_\sigma(\sigma, u)]_\mathbf{n} := P_{\sigma,+}(\sigma, u) \cdot \mathbf{n}_+ + P_{\sigma,-}(\sigma, u) \cdot \mathbf{n}_-,$$

there holds

$$\kappa \sum_{T \in \mathcal{T}} \|e_\sigma^P\|_T^2 = \kappa \sum_{T \in \mathcal{T}} \|e_u^P\|_T^2 - \Re(\sigma - P_\sigma(\sigma, u), e_\sigma^P)_T. \tag{3.13}$$

Proof. The main idea is a repetition of the arguments in the proof of Lemma 39. To shorten the notation, the arguments (σ, u) will be omitted in the proof. Applying the Galerkin orthogonality and inserting the projection yields considering the imaginary part

$$\begin{aligned} 0 =& \Im B(\sigma - \sigma_h, \sigma \cdot \mathbf{n} - \hat{\sigma}_{\mathbf{n},h}, u - u_h, u - \hat{u}_h; e_\sigma^P, e_{\hat{\sigma}}^P, e_u^P, e_{\hat{u}}^P) \\ =& \Im B(e_\sigma^P, e_{\hat{\sigma}}^P, e_u^P, e_{\hat{u}}^P; e_\sigma^P, e_{\hat{\sigma}}^P, e_u^P, e_{\hat{u}}^P) \\ &+ \Im B_{II}(\sigma - P_\sigma, \sigma \cdot \mathbf{n} - P_{\hat{\sigma}_\mathbf{n}}, u - P_u, u - P_{\hat{u}}; e_\sigma^P, e_{\hat{\sigma}}^P, e_u^P, e_{\hat{u}}^P). \end{aligned}$$

Looking at the element boundary terms of B_{II} on inner facets leads to

$$\begin{aligned} ((\sigma - P_\sigma) \cdot \mathbf{n} + \alpha[u - P_u], [e_u^P])_{F_I} &= ((\sigma - P_\sigma) \cdot \mathbf{n} + \alpha(u - P_u - u + P_{\hat{u}}), [e_u^P])_{F_I} \\ &= \left((\sigma - P_\sigma) \cdot \mathbf{n} + \alpha\left(u - P_u - u + \{P_u\} + \frac{1}{2\alpha}[P_\sigma]_\mathbf{n}\right), [e_u^P]\right)_{F_I} \\ &= \left(\sigma \cdot \mathbf{n} - \{P_\sigma\} \cdot \mathbf{n} - \frac{\alpha}{2}(P_{u,+} - P_{u_-}), [e_u^P]\right)_{F_I} = 0. \end{aligned}$$

For the other terms, starting with

$$\begin{aligned}\left(\beta[\![\sigma - P_\sigma]\!] + u - P_{\hat{u}}, [\![e_\sigma^P]\!]\right)_{F_I} &= \left(\beta(\sigma - P_\sigma - \sigma + P_{\hat{\sigma}_\mathbf{n}}) \cdot \mathbf{n} + u - P_{\hat{u}}, [\![e_\sigma^P]\!]\right)_{F_I} \\ &= \left(-\frac{\beta}{2}[P_\sigma]_\mathbf{n} + u - \{P_u\} - \frac{1}{2\alpha}[P_\sigma]_\mathbf{n}, [\![e_\sigma^P]\!]\right)_{F_I} \\ &= \left(-\left(\frac{1}{2\alpha} + \frac{\beta}{2}\right)[P_\sigma]_\mathbf{n} + u - \{P_u\}, [\![e_\sigma^P]\!]\right)_{F_I} = 0.\end{aligned}$$

The terms on the outer boundary are

$$\begin{aligned}((\sigma - P_\sigma) \cdot \mathbf{n} + \alpha[u - P_u], [e_u^P])_{F_O} - (\beta[\![\sigma - P_\sigma]\!] + u - P_{\hat{u}}, [\![e_\sigma^P]\!])_{F_O} + (u - P_{\hat{u}}, e_{\hat{u}}^P - e_{\hat{\sigma}}^P)_{F_O} \\ = ((\sigma - P_\sigma) \cdot \mathbf{n} + \alpha(u - P_u), [e_u^P])_{F_O} = 0.\end{aligned}$$

All boundary terms vanish, as well as the volume terms, except for the flux, which concludes the proof. □

Without a β-stabilisation P would exactly be the interpolation in [CGS10, GM11, Say13]. On the domain boundary, the projection has to have the same properties as the projection in [CGS10].

The Interpolation P

For P, existence and uniqueness, as well as the projection property and a suitable approximation property, need to be proven. The definition represents a square linear system of equations. Therefore, the uniqueness of the projection automatically implies its existence.

The interpolation P_u is decoupled from σ. The properties for P_u are proven in the following lemma.

Lemma 44 (Projection Decoupling)**.** *The projection P_u is uniquely defined by*

$$(P_u(\sigma, u), v_h)_T = (u, v_h)_T \qquad \forall v_h \in \mathcal{P}_{pw}^p(\mathcal{T})$$

and only depends on u, therefore $P_u(\sigma, u) = P_u(u)$. For $u \in H^{p+1}(T)$ there holds

$$\|u - P_u(u)\|_T \leq Ch^{p+1}|u|_{H^{p+1}(T)},$$

with a constant $C > 0$ independent of h, κ, α, β.

Proof. Equation (3.11b) implies that $P_u(\sigma, u)$ is the L^2-projection Πu. □

The proof for P_σ is more involved and similar to the analysis in [CGS10, Say13].

Lemma 45. *The space $\mathcal{P}_\perp^p(T)$ is defined by*

$$\mathcal{P}_\perp^p(T) := \{u \in \mathcal{P}^p(T) : (u, v)_T = 0, \forall v \in \mathcal{P}^{p-1}(T)\}.$$

If $u \in P_\perp^p(T)$ satisfies $u = 0$ on a facet of T, then $u \equiv 0$ on the whole element.

Proof. See [CGS10, Lemma A.1] and [Say13, Lemma 2.1]. □

Lemma 46. *The space $\mathcal{RT}^p_\perp(T)$ is defined by*

$$\mathcal{RT}^p_\perp(T) := \{\sigma \in \mathcal{RT}^p(T) : (\sigma,\tau)_T = 0, \forall \tau \in [\mathcal{P}^{p-1}(T)]^d\}.$$

Assume $\sigma \in \mathcal{RT}^p_\perp(T)$ then there holds

$$\|\sigma\|_T \leq C h^{\frac{1}{2}} \|\sigma \cdot \mathbf{n}\|_{\partial T},$$

with a constant $C > 0$ independent of h, κ, α, β.

Proof. The proof is similar to [CGS10, Proposition A.3] and [Say13, Lemma 2.1]. First, it is shown that the boundary term is a norm on the space $\mathcal{RT}^p_\perp(T)$. Assuming $\sigma \cdot \mathbf{n} = 0$ on ∂T then there holds

$$\|\operatorname{div}\sigma\|^2_T = (\sigma \cdot \mathbf{n}, \operatorname{div}\sigma)_{\partial T} - (\sigma, \nabla \operatorname{div}\sigma)_T = 0,$$

because $\nabla(\operatorname{div}(\sigma))|_T \in [\mathcal{P}^{p-1}(T)]^d$. According to [Say13, Proposition 2.3] this implies $\sigma \in [\mathcal{P}^p(T)]^d$ and by splitting σ into

$$\sigma = \sum_{i=1}^{d-1} \sigma \cdot \mathbf{n}_i$$

there holds $\sigma \cdot \mathbf{n}_i \in \mathcal{P}^p_\perp(T)$ as well as $\sigma \cdot \mathbf{n}_i = 0$ on the facet F_{I_i} corresponding to the normal vector $\mathbf{n}_i$. Then Lemma 45 implies that $\sigma \cdot \mathbf{n}_i$ vanishes on the whole element and therefore $\sigma = 0$. The estimate is proven by a standard scaling argument. □

With this lemma, the uniqueness and approximation property of P_σ can be proven.

Lemma 47. *Assuming H^1-regularity of σ and u, there exists a constant $C > 0$, independent of κ, h, α, β, so that*

$$\|\sigma - P_\sigma(\sigma,u)\|^2_\Omega \leq C\left(\eta(\sigma) + \frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right).$$

Proof. The proof is an adaptation of the approach in [CGS10, Proposition A.3]. Consider the standard element-wise $\mathcal{RT}$-interpolant satisfying

$$\begin{aligned} (\mathcal{RT}(\sigma), \tau_h)_T &= (\sigma, \tau_h)_T && \forall \tau_h \in [\mathcal{P}^{p-1}_{pw}(\mathcal{T})]^d, \\ ((\sigma - \mathcal{RT}(\sigma)) \cdot \mathbf{n}, \mu_h)_F &= 0 && \forall \mu_h \in \mathcal{P}^p_{pw}(\mathcal{F}) \end{aligned}$$

and define $\delta_\sigma := \mathcal{RT}(\sigma) - P_\sigma(\sigma,u)$. Due to (3.11a) and (3.11c - 3.11d) there holds for δ_σ

$$(\delta_\sigma, \tau_h)_T = 0 \qquad \tau_h \in [\mathcal{P}^{p-1}_{pw}(\mathcal{T})]^d, \tag{3.15a}$$

$$\begin{aligned} \left(\frac{1}{\alpha}+\beta\right)(\delta_\sigma \cdot \mathbf{n}, \mu_h)_{\partial T \cap F_I} &= -\left(u - P_u - \frac{\alpha\beta}{2}[u - P_u], \mu_h\right)_{\partial T \cap F_I} && \mu_h \in \mathcal{P}^p_{pw}(\mathcal{F}_I), \\ \left(\frac{1}{\alpha}+\beta\right)(\delta_\sigma \cdot \mathbf{n}, \mu_h)_{\partial T \cap F_O} &= -(1+\alpha\beta)(u - P_u, \mu_h)_{\partial T \cap F_O} && \mu_h \in \mathcal{P}^p_{pw}(\mathcal{F}_O). \end{aligned}$$

The idea is to choose $\mu_h = \delta_\sigma \cdot \mathbf{n}$ in the equations above and to estimate facet terms by

$$\begin{aligned}|(u - P_u, \delta_\sigma \cdot \mathbf{n})_F| &\leq Ch^{-\frac{1}{2}} \|u - P_u\|_F \|\delta_\sigma\|_T \\ &\leq Ch^{-1}(\|u - P_u\|_T + h|u - P_u|_{H^1(T)})\|\delta_\sigma\|_T.\end{aligned}$$

Rewriting it in the fashion of quasi-optimality by inserting an arbitrary v_h leads to

$$h|u - P_u|_{H^1(T)} \leq h|u - v_h|_{H^1(T)} + h|v_h - P_u|_{H^1(T)}.$$

A discrete inverse estimate gives

$$h|v_h - P_u|_{H^1(T)} \leq C_{inv}\|v_h - P_u\|_T \leq C_{inv}(\|v_h - u\|_T + \|u - P_u\|_T).$$

Note that due to the property of the L^2-projection, there holds

$$\|u - P_u\|_T = \inf_{w_h \in \mathcal{P}^p(T)} \|u - w_h\|_T \leq \|u - v_h\|_T$$

with the arbitrary v_h from above. Combining them together gives

$$\begin{aligned}|(u - P_u, \delta_\sigma \cdot \mathbf{n})_F| &\leq Ch^{-1}(\|u - P_u\|_T + h|u - P_u|_{H^1(T)})\|\delta_\sigma\|_T \\ &\leq Ch^{-1}((1 + 2C_{inv})\|u - v_h\|_T + h|u - v_h|_{H^1(T)})\|\delta_\sigma\|_T.\end{aligned}$$

The constant C changes in each line but stays independent of h, κ, α, β. Due to (3.15a) Lemma 46 can be applied yielding

$$\begin{aligned}\left(\frac{1}{\alpha} + \beta\right)\|\delta_\sigma\|_T^2 &\leq \left(\frac{1}{\alpha} + \beta\right) Ch\|\delta_\sigma \cdot \mathbf{n}\|_{\partial T}^2 \\ &= Ch((u - P_u, \delta_\sigma \cdot \mathbf{n})_{\partial T \setminus \partial\Omega} + \frac{\alpha\beta}{2}([u - P_u], \delta_\sigma \cdot \mathbf{n})_{\partial T \setminus \partial\Omega} \\ &\quad + (1 + \alpha\beta)(u - P_u, \delta_\sigma \cdot \mathbf{n})_{\partial T \cap \partial\Omega}) \\ &\leq C(1 + \alpha\beta)(\|u - v_h\|_\Omega + h|u - v_h|_{H^1(\Omega)})\|\delta_\sigma\|_T.\end{aligned}$$

Note that only an element patch is required, and due to the shape regularity, a finite overlap is given, which can be inserted into the leading constant. This gives

$$\|\delta_\sigma\|_\Omega \leq C\frac{1 + \alpha\beta}{\alpha^{-1} + \beta}\sqrt{\eta(u)}.$$

Note that the element wise Raviart-Thomas interpolant also satisfies the following best approximation.

Lemma 48. *For $\sigma \in [H^s(\Omega)]^d$ with $s > \frac{1}{2}$ there holds*

$$\|\sigma - \mathcal{R}\mathcal{T}\sigma\|_{L^2(\Omega)} \leq C \inf_{\tau_h \in \mathcal{R}\mathcal{T}^p_{pw}(\mathcal{T})} \left(\|\sigma - \tau_h\|_{L^2(\Omega)} + h^s \sum_{T \in \mathcal{T}} |\sigma - \tau_h|_{H^s(T)}\right).$$

Proof. The proof uses the continuity of the $\mathcal{RT}$-interpolant in combination with the invariance of polynomials giving

$$\begin{aligned}\|\sigma - \mathcal{RT}\sigma\|_{L^2(\Omega)} &\leq \inf_{\tau_h \in \mathcal{RT}^p_{pw}(\mathcal{T})} \|\sigma - \tau_h\|_{L^2(\Omega)} + \|\tau_h - \mathcal{RT}\sigma\|_{L^2(\Omega)} \\ &= \inf_{\tau_h \in \mathcal{RT}^p_{pw}(\mathcal{T})} \|\sigma - \tau_h\|_{L^2(\Omega)} + \|\mathcal{RT}(\tau_h - \sigma)\|_{L^2(\Omega)}.\end{aligned}$$

Then the continuity estimate from Chapter 16, Theorem 16.6 in [EG21] are applied to conclude. □

With this best approximation the proof can be finished with

$$\|\sigma - P_\sigma(\sigma, u)\|_\Omega \leq C \left(\eta(\sigma) + \frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right)^{1/2}.$$

□

With this result, the error estimate for the flux can be finalised.

Theorem 49. *Assuming H^s-regularity of the adjoint problem with $s > 3/2$ and H^2-regularity of the Helmholtz problem, there exists a constant $C(\kappa, \alpha, \beta) > 0$, independent of h, such that*

$$\begin{aligned}\|\sigma - \sigma_h\|^2_{L^2(\Omega)} \leq & C \left(\eta(\sigma) + \frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right) \\ & + \left(C(\Omega, \kappa, \alpha, \beta, h) + C(\Omega, \kappa, h)\right)^2 \frac{2C}{h}\left(\left(\frac{2}{\alpha} + \beta\right)\eta(\sigma) + 2\alpha\eta(u)\right).\end{aligned}$$

Proof. According to Lemma 43 there holds

$$\begin{aligned}\kappa \sum_{T\in\mathcal{T}} \|e^P_\sigma\|^2_T &= -\Im\Big(\sum_{T\in\mathcal{T}} j\kappa(\sigma - P_\sigma(\sigma,u), e^P_\sigma)_T - j\kappa\|e^P_u\|^2_T\Big) \\ &\leq \kappa \sum_{T\in\mathcal{T}} |(\sigma - P_\sigma(\sigma,u), e^P_\sigma)_T| + \|e^P_u\|^2_T \\ &\leq \kappa \sum_{T\in\mathcal{T}} \|\sigma - P_\sigma(\sigma,u)\|_T \|e^P_\sigma\|_T + \|e^P_u\|^2_T.\end{aligned}$$

Applying Young's inequality, Lemma 47 for $\sigma - P_\sigma$ and Theorem 42 for e^P_u yields

$$\begin{aligned}\kappa \sum_{T\in\mathcal{T}} \|e^P_\sigma\|^2_T &\leq \kappa \sum_{T\in\mathcal{T}} \|\sigma - P_\sigma(\sigma,u)\|^2_T + 2\|e^P_u\|^2_T \\ &\leq \kappa C\left(\eta(\sigma) + \frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right) \\ &\quad + 2\kappa\left(C(\Omega,\kappa,\alpha,\beta,h) + C(\Omega,\kappa,h)\right)^2 \frac{C}{h}\left(\left(\frac{2}{\alpha}+\beta\right)\eta(\sigma) + 2\alpha\eta(u)\right).\end{aligned}$$

Considering

$$\|\sigma - \sigma_h\|_\Omega \leq \|\sigma - P_\sigma(\sigma,u)\|_\Omega + \|e^P_\sigma\|_\Omega$$

and applying Lemma 47 concludes the proof. □

Although this error estimation for the flux is quasi-optimal with respect to the mesh size, it is sub-optimal in the wave number, as it would predict a pollution effect.

3.3.5. Asymptotic Error Estimates for the HDG-Formulation

The in Lemma 43 introduced projection can be used to derive asymptotically optimal convergence rates, with respect to the mesh size and wave number, in a similar fashion to the analysis in [GM11].

An Aubin-Nitsche-like trick for the adjoint problem is necessary, and using the interpolation of the adjoint solution there holds.

Lemma 50. *The adjoint solution satisfies with the interpolation P the weak formulation*

$$B(e_\sigma^P, e_{\hat{\sigma}}^P, e_u^P, e_{\hat{u}}^P; \phi - P_\phi, w - P_w, \phi \cdot \mathbf{n} - P_{\hat{\phi}_\mathbf{n}}, w - P_{\hat{w}}) = j\kappa \sum_{T \in \mathcal{T}} (e_\sigma^P, \phi - P_\phi)_T.$$

Proof. The proof is straightforward, the same as in Lemma 43. □

Theorem 51. *Assuming H^s-regularity of the adjoint problem with $s \geq 2$ and H^2-regularity of the Helmholtz problem, there exists a constant $C(\Omega, \kappa, h, \alpha, \beta) > 0$ and a constant $C > 0$ independent of κ, h, α, β, such that*

$$\begin{aligned}\left(\frac{1}{2} - 2CC^2(\Omega, \kappa, h, \alpha, \beta)\kappa^2 h^{2\min\{s-1,p\}}\right) \|e_\sigma^P\|_\Omega^2 \\ \leq C\left(\frac{1}{2} + \kappa^2 h^{2\min\{s-1,p\}}\left(C^2(\Omega, \kappa, h, \alpha, \beta) + C_{f,\min\{s,p+1\}}^2(\Omega, \kappa)\right)\right) \\ \left(\eta(\sigma) + \frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right).\end{aligned}$$

Proof. Considering the adjoint problem with the right hand side $j\kappa^2 e_u^P$ leads to

$$\begin{aligned}\kappa\|e_u^P\|_\Omega^2 =& B(e_\sigma^P, e_u^P, e_{\hat{\sigma}}^P, e_{\hat{u}}^P; \phi, w, \phi \cdot \mathbf{n}, w) \\ =& B_I(e_\sigma^P, e_u^P, e_{\hat{\sigma}}^P, e_{\hat{u}}^P; \phi - P_\phi, w - P_w, \phi \cdot \mathbf{n} - P_{\hat{\phi}_\mathbf{n}}, w - P_{\hat{w}}) \\ &+ B(e_\sigma^P, e_u^P, e_{\hat{\sigma}}^P, e_{\hat{u}}^P; P_\phi, P_w, P_{\hat{\phi}_\mathbf{n}}, P_{\hat{w}}),\end{aligned}$$

after inserting the projection. The first part satisfies due to the previous lemma

$$\begin{aligned}\left|B(e_\sigma^P, e_u^P, e_{\hat{\sigma}}^P, e_{\hat{u}}^P; \phi - P_\phi, w - P_w, \phi \cdot \mathbf{n} - P_{\hat{\phi}_\mathbf{n}}, w - P_{\hat{w}})\right| &= \left|j\kappa \sum_{T \in \mathcal{T}} (e_\sigma^P, \phi - P_\phi)_T\right| \\ &\leq \kappa\|e_\sigma^P\|_\Omega\|\phi - P_\phi\|_\Omega\end{aligned}$$

and the second part satisfies for arbitrary $\psi_h \in \mathcal{P}^{p-1}(\mathcal{T})$

$$\begin{aligned}\left|B(e_\sigma^P, e_u^P, e_{\hat{\sigma}}^P, e_{\hat{u}}^P; P_\phi, P_w, P_{\hat{\phi}_\mathbf{n}}, P_{\hat{w}})\right| &= \left|-j\kappa \sum_{T \in \mathcal{T}} (\sigma - P_\sigma, P_\phi)_T\right| \\ &= \left|-j\kappa \sum_{T \in \mathcal{T}} (\sigma - P_\sigma, P_\phi - \psi_h)_T\right| \\ &\leq \kappa\|\sigma - P_\sigma\|_\Omega\|P_\phi - \psi_h\|_\Omega.\end{aligned}$$

The projection error of the adjoint solution can be estimated due to Lemma 47 by

$$\|\phi - P_\phi\|_\Omega^2 \leq C\left(\eta(\phi) + \frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(w)\right).$$

The constant C will be implicitly reused and redefined in the course of the proof, but always remains independent of κ, h, α and β. The necessary bounds for the best approximation parts of the adjoint solutions are

$$\eta(w) \leq Ch^{2\min\{s,p+1\}}|w|^2_{H^{\min\{s,p+1\}}(\Omega)} \leq CC^2_{f,\min\{s,p+1\}}(\Omega,\kappa)\kappa^4 h^{2\min\{s,p+1\}}\|e_u^P\|_\Omega^2$$

and

$$\eta(\phi) \leq \frac{C}{\kappa^2}h^{2\min\{s-1,p+1\}}|w|^2_{H^{\min\{s,p+2\}}(\Omega)} \leq CC^2_{f,\min\{s,p+2\}}(\Omega,\kappa)\kappa^2 h^{2\min\{s-1,p+1\}}\|e_u^P\|_\Omega^2.$$

The second approximation term of the adjoint solution can be estimated by

$$\|P_\phi - \psi_h\|_\Omega \leq \|P_\phi - \phi\|_\Omega + \|\phi - \psi_h\|_\Omega.$$

The first part of this estimate is already shown and by considering the L^2-projection as ψ_h there holds for the second term

$$\|\phi - \psi_h\|_\Omega \leq \frac{C}{\kappa}h^{\min\{s-1,p\}}|w|_{H^{\min\{s,p+1\}}(\Omega)} \leq CC_{f,\min\{s,p+1\}}(\Omega,\kappa)\kappa h^{\min\{s-1,p\}}\|e_u^P\|_\Omega.$$

With the constant

$$\begin{aligned}C^2(\Omega,\kappa,h,\alpha,\beta) := {}& C^2_{f,\min\{s,p+2\}}(\Omega,\kappa)h^{2(\min\{s-1,p+1\}-\min\{s-1,p\})}\\ &+\frac{1+\alpha\beta}{\alpha^{-1}+\beta}C^2_{f,\min\{s,p+1\}}(\Omega,\kappa)\kappa^2h^2\end{aligned}$$

the estimates can be combined into

$$\begin{aligned}\kappa\|e_u^P\|_\Omega^2 \leq C\kappa^2h^{2\min\{s-1,p\}}\Bigg(&C^2(\Omega,\kappa,h,\alpha,\beta)\kappa\|e_\sigma^P\|_\Omega^2\\ &+\left(C^2(\Omega,\kappa,h,\alpha,\beta)+C^2_{f,\min\{s,p+1\}}(\Omega,\kappa)\right)\kappa\|\sigma-P_\sigma\|_\Omega^2\Bigg).\end{aligned} \tag{3.16}$$

Additionally there holds

$$\begin{aligned}\kappa\|e_\sigma^P\|_\Omega^2 - \kappa\|e_u^P\|_\Omega^2 &= \Im B(e_\sigma^P,e_u^P,e_{\hat\sigma}^P,e_{\hat u}^P;e_\sigma^P,e_u^P,e_{\hat\sigma}^P,e_{\hat u}^P)\\ &\leq \kappa\|\sigma-P_\sigma\|_\Omega\|e_\sigma^P\|_\Omega \leq \frac{\kappa}{2}\|\sigma-P_\sigma\|_\Omega^2 + \frac{\kappa}{2}\|e_\sigma^P\|_\Omega^2.\end{aligned} \tag{3.17}$$

Merging both estimates and using Lemma 47 gives

$$\begin{aligned}&\left(\frac{1}{2} - 2CC^2(\Omega,\kappa,h,\alpha,\beta)\kappa^2h^{2\min\{s-1,p\}}\right)\kappa\|e_\sigma^P\|_\Omega^2\\ &\leq \left(\frac{1}{2} + C\kappa^2h^{2\min\{s-1,p\}}\left(C^2(\Omega,\kappa,h,\alpha,\beta)+C^2_{f,\min\{s,p+1\}}(\Omega,\kappa)\right)\right)\kappa\|\sigma-P_\sigma\|_\Omega^2\\ &\leq C\left(\frac{1}{2} + C\kappa^2h^{2\min\{s-1,p\}}\left(C^2(\Omega,\kappa,h,\alpha,\beta)+C^2_{f,\min\{s,p+1\}}(\Omega,\kappa)\right)\right)\\ &\qquad\kappa\left(\eta(\sigma)+\frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right).\end{aligned}$$

□

Additionally, there is a faster convergence rate for the projected error of the pressure.

Corollary 52. *Asymptotically, the projected error of the pressure* $\|e_u^P\|_\Omega$ *converges with a rate of*

$$t = \min\{s-1, p\}$$

faster than the projected error in the flux $\|e_\sigma^P\|_\Omega$.

Proof. Due to (3.16) there holds

$$\kappa\|e_u^P\|_\Omega^2 \leq C\kappa^2 h^{2\min\{s-1,p\}} \Bigg(C^2(\Omega,\kappa,h,\alpha,\beta)\kappa\|e_\sigma^P\|_\Omega^2 + \left(C^2(\Omega,\kappa,h,\alpha,\beta) + C^2_{f,\min\{s,p+1\}}(\Omega,\kappa)\right)\kappa\|\sigma - P_\sigma\|_\Omega^2 \Bigg).$$

□

As can be seen in the asymptotic range, the convergence rate is gained for the projected error in the pressure.

Proof of Theorem 30. On a star-shaped domain with H^2-regularity the theorie in [MPS13] can be applied to derive a better approximation of the adjoint solution compared to the results in Theorem 51. In [MPS13, Theorem 4.8] a splitting of the adjoint solution for an excitation f into $w = w_{H^2} + w_{\mathcal{A}}$ with an H^2-regular and an analytic part is used. They proof that there exists a positive constant $C > 0$ independent of κ, h and a discrete function v_{H^2} so that

$$\kappa\left(\kappa\|w_{H^2} - v_{H^2}\|_{L^2(\Omega)} + \|\nabla(w_{H^2} - v_{H^2})\|_{L^2(\Omega)}\right) \leq C\left(\kappa h + \kappa^2 h^2\right)\|f\|_{L^2(\Omega)}$$

holds. Additionally, they show that there exists a discrete function $v_{\mathcal{A}}$ so that

$$\kappa\left(\kappa\|w_{\mathcal{A}} - v_{\mathcal{A}}\|_{L^2(\Omega)} + \|\nabla(w_{\mathcal{A}} - v_{\mathcal{A}})\|_{L^2(\Omega)}\right) \leq C\left(1+\kappa h\right)\left(h^p + \kappa(\kappa h)^p\right)\|f\|_{L^2(\Omega)},$$
$$|w_{\mathcal{A}} - v_{\mathcal{A}}|_{H^2(\Omega)} \leq \frac{C}{\kappa h}\left(h^p + \kappa(\kappa h)^p\right)\|f\|_{L^2(\Omega)}$$

hold. With these estimates, the adjoint approximation constants can be improved. The best approximation $\eta(w)$ can be bounded by

$$\eta(w) \leq C\left(\kappa^2 h^2 + \kappa^4 h^4 + \kappa^6 h^6 + (1+\kappa h)^2(1+\kappa^2 h^2)(h^p + \kappa(\kappa h)^p)^2\right)\|e_u^P\|_\Omega^2$$

and the best approximation $\eta(\phi)$ by

$$\eta(\phi) \leq C\left(\kappa^2 h^2 + \kappa^4 h^4 + (1+\kappa h)^2(h^p + \kappa(\kappa h)^p)^2\right)\|e_u^P\|_\Omega^2.$$

The last piece is the estimate of the best approximation property for the adjoint solution by a piecewise polynomial of degree $p-1$ by

$$\|\phi - \psi_h\|_\Omega \leq C\left(\kappa h + \kappa^2 h^2 + (1+\kappa h)\left(h^{p-1} + \kappa(\kappa h)^{p-1}\right)\right)\|e_u^P\|_\Omega.$$

With the constants

$$C_p^2(\kappa, h) := C\left(\kappa^2 h^2 + \kappa^4 h^4 + \kappa^6 h^6 + (1+\kappa h)^4 (h^p + \kappa(\kappa h)^p)^2\right)$$

and

$$C_{p-1}^2(\kappa, h) := C\left(\kappa^2 h^2 + \kappa^4 h^4 + (1+\kappa h)^2\left(h^{p-1} + \kappa(\kappa h)^{p-1}\right)^2\right)$$

the estimates can be combined into

$$\kappa\|e_u^P\|_\Omega^2 \leq C_p^2(\kappa, h)\kappa(\|e_\sigma^P\|_\Omega^2 + \|\sigma - P_\sigma\|_\Omega^2) + C_{p-1}^2(\kappa, h)\kappa\|\sigma - P_\sigma\|_\Omega^2.$$

Connecting this equation with (3.17) leads to

$$\left(\frac{1}{2} - C_p^2(\kappa, h)\right)\kappa\|e_\sigma^P\|_\Omega^2 \leq \left(\frac{1}{2} + C_p^2(\kappa, h) + C_{p-1}^2(\kappa, h)\right)\kappa\|\sigma - P_\sigma\|_\Omega^2.$$

The requirement for the asymptotic convergence therefore is

$$2C_p^2(\kappa, h) < 1.$$

Under the reasonable assumption $\kappa h \leq \mathcal{O}(1)$, the requirement falls back to

$$\kappa(\kappa h)^p < \mathcal{O}(1)$$

which leads to the resolution requirement

$$h = \mathcal{O}\left(\kappa^{-\frac{p+1}{p}}\right).$$

This implies the compensation of the pollution effect by high order finite elements. □

3.3.6. Static Condensation

In Schur complement element matrices need to be invertible, see Section 2.6.3, which represent local problems. In this section, only discrete formulations are considered, but the subscript h is omitted to improve the readability. The HDG-formulation in Definition 21 can be separated by testfunctions giving

$$\sum_{T\in\mathcal{T}} j\kappa(\sigma, \tau)_T + (u, \operatorname{div}\tau)_T - (\hat{u}, \tau\cdot\mathbf{n})_{\partial T} - \beta(\sigma\cdot\mathbf{n}, \tau\cdot\mathbf{n})_{\partial T} + \beta(\hat{\sigma}_\mathbf{n}, \tau\cdot\mathbf{n})_{\partial T} = 0,$$

$$\sum_{T\in\mathcal{T}} (\operatorname{div}\sigma, v)_T - j\kappa(u, v)_T + \alpha(u, v)_{\partial T} - \alpha(\hat{u}, v)_{\partial T} = \frac{j}{\kappa}(f, v)_\Omega,$$

$$\sum_{T\in\mathcal{T}} \beta(\sigma\cdot\mathbf{n}, \hat{\tau}_\mathbf{n})_{\partial T} - \beta(\hat{\sigma}_\mathbf{n}, \hat{\tau}_\mathbf{n})_{\partial T} = 0, \tag{3.19a}$$

$$\sum_{T\in\mathcal{T}} -(\sigma\cdot\mathbf{n}, \hat{v})_{\partial T} - \alpha(u, \hat{v})_{\partial T} + \alpha(\hat{u}, \hat{v})_{\partial T} + (\hat{u}, \hat{v})_{\partial\Omega} = \frac{j}{\kappa}(g, \hat{v})_{\partial\Omega}. \tag{3.19b}$$

Now, the question is which variables can be uniquely eliminated, depending on other variables.

Elimination of Volume Unknowns

First, the volume variables σ and u are considered. The submatrix blocks in the Schur complement, see Section 2.6.3, are associated with the following sesquilinear form parts and variables.

$$
\begin{aligned}
A &\equiv \sum_{T\in\mathcal{T}} -\beta(\hat{\sigma}_{\mathbf{n}}, \hat{\tau}_{\mathbf{n}})_{\partial T} + \alpha(\hat{u}, \hat{v})_{\partial T} + (\hat{u}, \hat{v})_{\partial\Omega},\\
B &\equiv \sum_{T\in\mathcal{T}} \beta(\sigma\cdot\mathbf{n}, \hat{\tau}_{\mathbf{n}})_{\partial T} - (\sigma\cdot\mathbf{n}, \hat{v})_{\partial T} - \alpha(u, \hat{v})_{\partial T},\\
C &\equiv \sum_{T\in\mathcal{T}} \beta(\hat{\sigma}_{\mathbf{n}}, \tau\cdot\mathbf{n})_{\partial T} - (\hat{u}, \tau\cdot\mathbf{n})_{\partial T} - \alpha(\hat{u}, v)_{\partial T},\\
D &\equiv \sum_{T\in\mathcal{T}} j\kappa(\sigma,\tau)_T + (u, \operatorname{div}\tau)_T + (\operatorname{div}\sigma, v)_T - j\kappa(u,v)_T - \beta(\sigma\cdot\mathbf{n}, \tau\cdot\mathbf{n})_{\partial T} + \alpha(u,v)_{\partial T},\\
x_1 &\equiv (\hat{\sigma}_{\mathbf{n}}, \hat{u})^\top,\\
x_2 &\equiv (\sigma, u)^\top,\\
f_1 &\equiv \frac{j}{\kappa}(g, \hat{v})_{\partial T},\\
f_2 &\equiv \frac{j}{\kappa}(f, v)_{\Omega}.
\end{aligned}
$$

Equations represented by

$$DX = -C$$

need to be solved, which are equivalent to solving the following weak formulation for σ and u

$$
\begin{aligned}
\sum_{T\in\mathcal{T}} j\kappa(\sigma,\tau)_T + (u, \operatorname{div}\tau)_T + (\operatorname{div}\sigma, v)_T - j\kappa(u,v)_T - \beta(\sigma\cdot\mathbf{n}, \tau\cdot\mathbf{n})_{\partial T} + \alpha(u,v)_{\partial T}&\\
= \sum_{T\in\mathcal{T}} -\beta(\hat{\sigma}_{\mathbf{n}}, \tau\cdot\mathbf{n})_{\partial T} + (\hat{u}, \tau\cdot\mathbf{n})_{\partial T} + \alpha(\hat{u}, v)_{\partial T}&
\end{aligned}
$$

for given $\hat{\sigma}_{\mathbf{n}}, \hat{u}$. These equations are element-wise independent, and they decouple into the equations

$$B_T(\sigma, u; \tau, v) = -\beta(\hat{\sigma}_{\mathbf{n}}, \tau\cdot\mathbf{n})_{\partial T} + (\hat{u}, \tau\cdot\mathbf{n})_{\partial T} + \alpha(\hat{u}, v)_{\partial T}$$

with the sesquilinear form

$$
\begin{aligned}
B_T(\sigma, u; \tau, v) := j\kappa(\sigma,\tau)_T + (u, \operatorname{div}\tau)_T + (\operatorname{div}\sigma, v)_T - j\kappa(u,v)_T&\\
-\beta(\sigma\cdot\mathbf{n}, \tau\cdot\mathbf{n})_{\partial T} + \alpha(u,v)_{\partial T}&
\end{aligned}
$$

on each element. In the following similar arguments as for the full HDG-formulation are used to prove unique solvability and stability estimates. There holds

$$-\Re B_T(\sigma, u; \sigma, -u) = \beta\|\sigma\cdot\mathbf{n}\|^2_{\partial T} + \alpha\|u\|^2_{\partial T} =: \|\sigma, u\|^2_{\partial T}$$

The traces can be controlled which is an essential property. In combination with estimating the right hand side the traces on the element boundary are bounded by

$$\|\sigma, u\|_{\partial T} \leq \left(\beta\|\hat{\sigma}_{\mathbf{n}}\|_{\partial T}^2 + \left(\frac{1}{\beta} + \alpha\right)\|\hat{u}\|_{\partial T}^2\right)^{1/2}.$$

Remark 53. *If the β-stabilisation would be omitted, then the crucial right hand side term $(\hat{u}, \sigma \cdot \mathbf{n})_T$ would require an h-dependent inverse estimate leading to a stability constant which degenerates with decreasing mesh size.*

Next, for the imaginary part of the sesquilinear form there holds

$$\Im B_T(\sigma, u; \sigma, u) = \kappa\|\sigma\|_T^2 - \kappa\|u\|_T^2.$$

Very similar to previous methods, an Aubin-Nitsche trick is used to control u, and again projections are introduced

$$\|u\|_T^2 = B_T(\sigma, u; \phi, w) = B_T(\sigma, u; \phi - \Pi\phi, w - \Pi w) + B_T(\sigma, u; \Pi\phi, \Pi w).$$

For the first part there holds after partial integration

$$\begin{aligned} B(\sigma, u; \phi - \Pi\phi, w - \Pi w) &= (u - \beta\sigma \cdot \mathbf{n}, (\phi - \Pi\phi) \cdot \mathbf{n})_{\partial T} + \alpha(u, w - \Pi w)_{\partial T} \\ &\leq \|\sigma, u\|_{\partial T}\left(\left(\frac{1}{\alpha} + \beta\right)\|(\phi - \Pi\phi) \cdot \mathbf{n}\|_{\partial T}^2 + \alpha\|w - \Pi w\|_{\partial T}^2\right)^{1/2} \\ &\leq C(\Omega, h, \alpha, \beta)\|\sigma, u\|_{\partial T}\|u\|_T \end{aligned}$$

and for the second part

$$\begin{aligned} B(\sigma, u; \Pi\phi, \Pi w) &= -\beta(\hat{\sigma}_{\mathbf{n}}, \Pi\phi \cdot \mathbf{n})_{\partial T} + (\hat{u}, \Pi\phi \cdot \mathbf{n})_{\partial T} + \alpha(\hat{u}, \Pi w)_{\partial T} \\ &\leq C(\Omega, \kappa)\left(\beta\|\hat{\sigma}_{\mathbf{n}}\|_{\partial T} + \|\hat{u}\|_{\partial T} + \alpha\|\hat{u}\|_{\partial T}\right)\|u\|_T. \end{aligned}$$

Combining both estimates gives

$$\begin{aligned} \|u\|_T \leq C(\Omega, h, \alpha, \beta)&\left(\beta\|\hat{\sigma}_{\mathbf{n}}\|_{\partial T}^2 + \left(\frac{1}{\beta} + \alpha\right)\|\hat{u}\|_{\partial T}^2\right)^{1/2} \\ &+ C(\Omega, \kappa)\left(\beta\|\hat{\sigma}_{\mathbf{n}}\|_{\partial T} + (1 + \alpha)\|\hat{u}\|_{\partial T}\right). \end{aligned}$$

With this estimate σ can be bounded. An interesting aspect is that both $\alpha = 0$, as well as $\beta = 0$, would be an issue for the Schur complement. This is contrary to the absolute stability of the full HDG-formulation, where only $\alpha = 0$ is an issue.

Elimination of Boundary Unknowns

A for the iterative solver irrelevant, but none the less interesting aspect is the elimination of the facet unknowns. Due to (3.19a) there holds

$$\sum_{T \in \mathcal{T}} \beta(\sigma \cdot \mathbf{n} - \hat{\sigma}_{\mathbf{n}}, \hat{\tau}_{\mathbf{n}})_{\partial T} = 0,$$

which can be considered facet-wise. The traces of the discrete spaces match such that the equations imply

$$\begin{aligned} \hat{\sigma}_{\mathbf{n}} &= \frac{1}{2}(\sigma_+ + \sigma_-) \cdot \mathbf{n} && \text{on } F_I, \\ \hat{\sigma}_{\mathbf{n}} &= \sigma \cdot \mathbf{n} && \text{on } F_O. \end{aligned}$$

On inner facets, the flux variable is exactly the mean value of the flux of adjacent elements. For the pressure facet unknowns, there holds

$$\sum_{T \in \mathcal{T}} -(\sigma \cdot \mathbf{n}, \hat{v})_{\partial T} - \alpha(u, \hat{v})_{\partial T} + \alpha(\hat{u}, \hat{v})_{\partial T} + (\hat{u}, \hat{v})_{\partial \Omega} = \frac{j}{\kappa}(g, \hat{v})_{\partial \Omega}$$

according to (3.19b). Combing terms on inner facets leads to

$$\begin{aligned} \hat{u} &= \frac{1}{2}(u_+ + u_-) + \frac{1}{2\alpha}(\sigma_+ \cdot \mathbf{n}_+ + \sigma_- \cdot \mathbf{n}_-) && \text{on } F_I, \\ (1+\alpha)\hat{u} &= \frac{1}{1+\alpha}\left(\alpha u + \sigma \cdot \mathbf{n} + \frac{j}{\kappa} g\right) && \text{on } F_O. \end{aligned}$$

On inner facets, the variable is a combination of the mean value of the pressure from adjacent elements and the jump of the flux.

These forms could be reinserted into the HDG-formulation and would lead to an equivalent DG formulation which produces the same solution as the HDG-formulation.

3.3.7. Lowest Order Discretisation

In the previous section, the FE order was assumed to be larger or equal to one ($p \geq 1$). This is also necessary to prove the asymptotic quasi-optimality of the method. The lowest order case ($p = 0$) for spaces is still interesting. The pre-asymptotic analysis is valid, which would suggest that the quasi-optimality constant deteriorates with the wave number, which can only hold if the pollution error does not display super convergence. Numerical simulations suggest exactly this behaviour, and in the following section a dispersion and dissipation analysis for the lowest order formulation on a one-dimensional domain is carried out to further support the claim.

Dispersion and Dissipation Analysis

Dispersion and dissipation analysis is based on the fact that in simulations for the Helmholtz equation pollution effects appear and the goal is to calculate the discrete wave number κ_h which exactly satisfies the discrete system. For a fixed wave number κ the homogenous Helmholtz equation on the whole domain $\Omega = \mathbb{R}$ is considered and it is assumed that the solution is a plane wave of the form

$$u(x) := e^{j\kappa_h x}.$$

This Ansatz is inserted into the formulation, and the discrete wave number is calculated.

Conforming H^1-Formulation Considering the conforming H^1-formulation

$$B(u;v) := (\nabla u, \nabla v)_\Omega - \kappa^2 (u,v)_\Omega + j\kappa (u,v)_{\partial\Omega}$$

with a structured mesh of size h, $x_i := ih$. The discrete solution is a linear combination

$$u_h := \sum_{i=0}^{N} u_i \phi_i$$

of the Hat functions ϕ_i and the discrete solution has to satisfy

$$(\nabla u_h, \nabla \phi_i)_\Omega - \kappa^2 (u_h, \phi_i)_\Omega = 0,$$

which leads to

$$\left(\nabla u_h, \frac{1}{h}\right)_{(x_{i-1},x_i)} - \left(\nabla u_h, \frac{1}{h}\right)_{(x_i,x_{i+1})} - \kappa^2 (u_h, \phi_i)_{(x_{i-1},x_{i+1})} = 0$$

after inserting hat functions as test functions. Evaluating these integrals gives

$$(\phi_i, \phi_i)_{(x_{i-1},x_{i+1})} = \frac{2h}{3}, \qquad (\phi_{i-1}, \phi_i)_{(x_{i-1},x_i)} = \frac{h}{6}, \qquad (\phi_{i+1}, \phi_i)_{(x_i,x_{i+1})} = \frac{h}{6},$$

and replacing them above leads to

$$-u_{i-1} + 2u_i - u_{i+1} - (\kappa h)^2 \frac{1}{6}(u_{i-1} + 4u_i + u_{i+1}) = 0.$$

Combining coefficients on different points x_i by the phase shift induced by κ_h gives

$$u_{i-1} = u_i e^{-i\kappa_h h}, \qquad u_{i+1} = u_i e^{i\kappa_h h}$$

and can be reinserted and simplified to

$$\cos(\kappa_h h) = \frac{6 - 2\kappa^2 h^2}{6 + \kappa^2 h^2}.$$

As can be seen, the discrete wavenumber does not coincide with κ, but it is real-valued and converges to κ with the square rate of

$$|\kappa - \kappa_h| \leq \mathcal{O}(h^2),$$

see also Figure 3.1.

Due to the zero imaginary parte there is no dissipation error, and the dispersion has the same convergence rate as the best approximation.

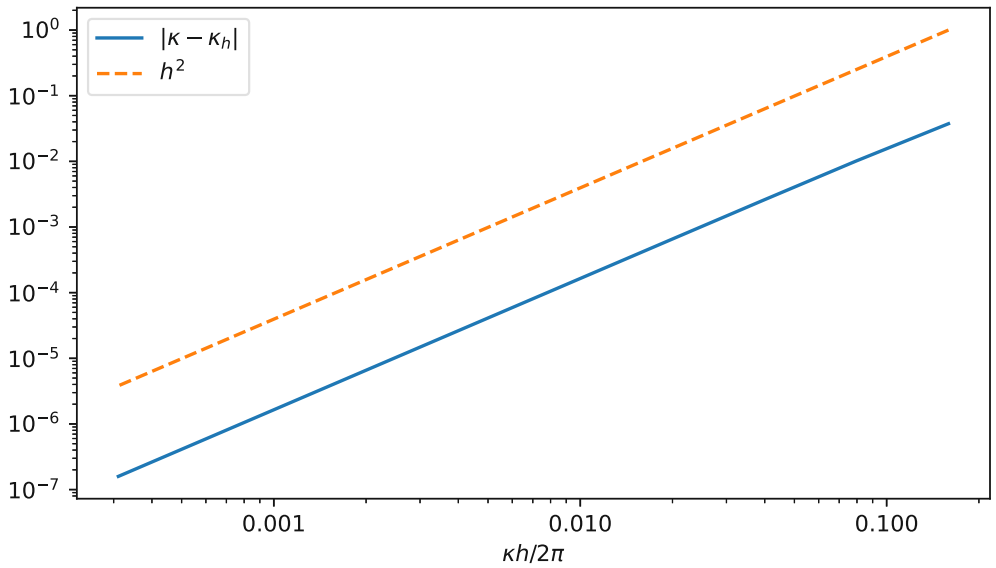


Figure 3.1.: Error of the discrete wave number κ_h of the H^1-formulation with respect to the mesh size h for the given wave number of $\kappa = 1$. Reference line of square convergence rate as a dashed line.

Mixed Formulation Applying the approach from above to the following mixed formulation

$$B(u;v) := j\kappa(\sigma,\tau)_\Omega + (u, \operatorname{div}\tau)_\Omega + (\operatorname{div}\sigma, v)_\Omega - j\kappa(u,v)_\Omega - (\sigma\cdot\mathbf{n}, \tau\cdot\mathbf{n})_{\partial\Omega}$$

with the lowest order Ansatz for flux and pressure

$$\sigma_h := \sum_{i=0}^{N} s_i\phi_i,$$
$$u_h := \sum_{i=1}^{N} u_i \mathbb{1}_{[x_{i-1},x_i]},$$

leads to the equations

$$j\kappa(\sigma_h,\phi_i)_\Omega + (u_h,\phi_i')_\Omega = 0,$$
$$\left(\operatorname{div}\sigma_h, \mathbb{1}_{[x_{i-1},x_i]}\right)_\Omega - j\kappa\left(u_h, \mathbb{1}_{[x_{i-1},x_i]}\right)_\Omega = 0,$$

which can be simplified into

$$j\kappa\left(s_{i-1}\frac{h}{6} + s_i\frac{2h}{3} + s_{i+1}\frac{h}{6}\right) + u_i - u_{i+1} = 0,$$
$$-s_{j-1} + s_j - j\kappa h u_j = 0$$

and further

$$jk\left(e^{-i\kappa_h h}\frac{h}{6}+\frac{2h}{3}+e^{i\kappa_h h}\frac{h}{6}\right)s_i+u_i(1-e^{i\kappa_h h})=0,$$
$$(1-e^{-i\kappa_h h})s_j-jk h u_j=0$$

resulting in

$$\cos(\kappa_h h)=\frac{6-2k^2h^2}{6+\kappa^2h^2},$$

the same form as for the conforming H^1-formulation. Therefore, the mixed formulation exhibits the same dispersion and dissipation as the classical H^1-formulation.

HDG-Formulation The DG-formulation equivalent to the HDG-formulation in one dimension is

$$\begin{aligned}B(.;.):=&\,jk(\sigma,\tau)_T+(\sigma',v)_T+(u,\tau')_T-jk(u,v)_T\\&-(\{u\},[\tau]_{\mathbf{n}})_{F_I}-([\sigma]_{\mathbf{n}},\{v\})_{F_I}+\alpha([u]_+,[v]_+)_{F_I}-\beta([\sigma]_{\mathbf{n}},[\tau]_{\mathbf{n}})_{F_I},\end{aligned}$$

with the discrete variables

$$\sigma_h:=\sum_{i=1}^{N}s_{i,l}\phi_{i,l}+s_{i,r}\phi_{i,r},$$
$$u_h:=\sum_{i=1}^{N}u_i\mathbb{1}_{[x_{i-1},x_i]}.$$

The discrete solution solves the equations

$$\begin{aligned}jk(\sigma_h,\phi_{i,l})_{(x_{i-1},x_i)}+(u_h,\phi'_{i,l})_{(x_{i-1},x_i)}+\{u_h\}_l\phi_{i,l}(x_{i-1})+\beta[\sigma_h]_{\mathbf{n},l}\phi_{i,l}(x_{i-1})&=0,\\jk(\sigma_h,\phi_{i,r})_{(x_{i-1},x_i)}+(u_h,\phi'_{i,r})_{(x_{i-1},x_i)}-\{u_h\}_r\phi_{i,r}(x_i)-\beta[\sigma_h]_{\mathbf{n},r}\phi_{i,r}(x_i)&=0,\\(\sigma'_h,1)_{(x_{i-1},x_i)}-jk(u_h,1)_{(x_{i-1},x_i)}-\frac{1}{2}[\sigma_h]_{\mathbf{n},l}-\frac{1}{2}[\sigma_h]_{\mathbf{n},r}+\alpha[u]_{+,l}+\alpha[u]_{+,r}&=0.\end{aligned}$$

After inserting the Ansatz, evaluating the integrals and simplifications remain

$$\begin{aligned}jk\left(s_{i,l}\frac{h}{3}+s_{i,r}\frac{h}{6}\right)-u_i+\frac{1}{2}(u_{i-1}+u_i)-\beta(s_{i,l}-s_{i-1,r})&=0,\\jk\left(s_{i,l}\frac{h}{6}+s_{i,r}\frac{h}{3}\right)+u_i-\frac{1}{2}(u_i+u_{i+1})-\beta(s_{i,r}-s_{i+1,l})&=0,\\-s_{i,l}+s_{i,r}-jk h u_i+\frac{1}{2}(s_{i,l}-s_{i-1,r})-\frac{1}{2}(s_{i,r}-s_{i+1,l})+\alpha(u_i-u_{i-1})+\alpha(u_i-u_{i+1})&=0.\end{aligned}$$

Introducing a harmonic Ansatz for the coefficients with a discrete exponential shift leads to the following system of linear equations in matrix form:

$$\begin{pmatrix}\frac{1}{2}(e^{-i\kappa_h h}-1) & \frac{jkh}{3}-\beta & \frac{jkh}{6}+\beta e^{-i\kappa_h h}\\ -\frac{1}{2}(e^{i\kappa_h h}-1) & \frac{jkh}{6}+\beta e^{i\kappa_h h} & \frac{jkh}{3}-\beta\\ -\alpha(e^{i\kappa_h h}+e^{-i\kappa_h h}-2-jkh) & \frac{1}{2}(e^{i\kappa_h h}-1) & -\frac{1}{2}(e^{-i\kappa_h h}-1)\end{pmatrix}\begin{pmatrix}u_i\\s_{i,l}\\s_{i,r}\end{pmatrix}=0.$$

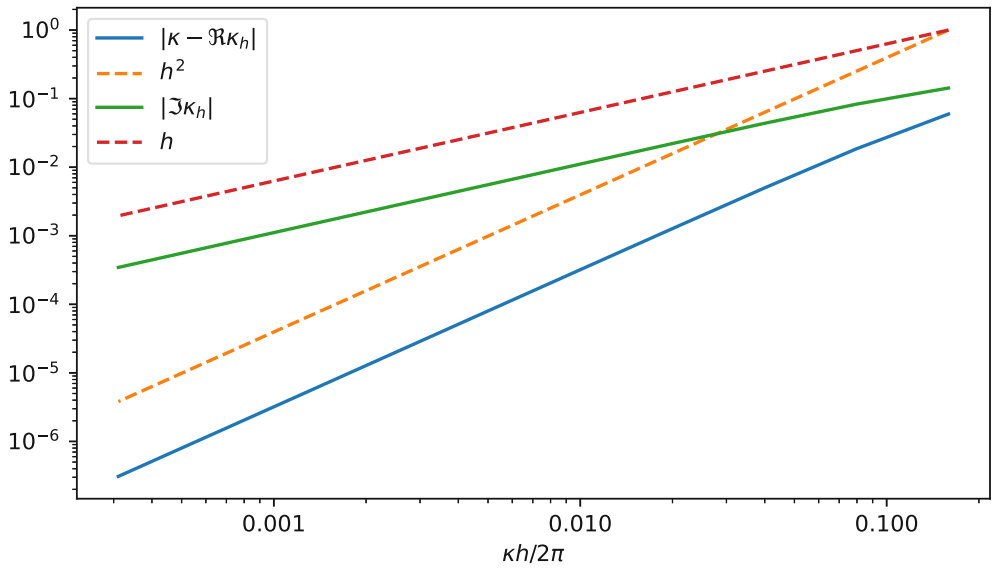


Figure 3.2.: Error in the imaginary as well as the real part of the discrete wave number κ_h for the HDG-formulation with respect to the mesh size h for the given wave number of $\kappa = 1$. Reference lines of the second-order convergence rate and the first-order convergence rate as dashed lines.

Possible discrete wave numbers are defined by a singular matrix. Therefore, a null point search for the determinant of the matrix is used to derive the defining equation.

$$\begin{aligned} &-2j\kappa h(4\alpha+1)\cos^2(\kappa_h h) \\ &+\left(-12\beta(2-\beta-2\alpha\beta)+(6+8\alpha(2\beta+1))j\kappa h-\frac{\alpha}{3}(4-2\beta)k^2h^2\right)\cos(\kappa_h h) \\ &-12\beta(2\alpha\beta-2+\beta)-2(6\alpha\beta^2+8\alpha\beta+3\beta^2+2)j\kappa h \\ &-2\alpha\left(\frac{4}{3}-4\beta+\frac{\beta}{3}\right)\kappa^2h^2-\frac{2\alpha}{3}\left(2+\frac{\beta}{2}\right)jk^3h^3 \\ &=0. \end{aligned}$$

As a further simplification, the stabilisation parameters of the DG-formulation are set to the, for simulations reasonable values, $\alpha = 1/2, \beta = 1$. The equation simplifies after some minor manipulation to

$$\cos(\kappa_h h) = \frac{1}{4\kappa h}\left(jk^2h^2 - 10\kappa h + 24j \pm \sqrt{-5\kappa^4h^4 + 20jk^3h^3 + 148\kappa^2h^2 - 672j\kappa h - 576}\right).$$

The error in the real part and the imaginary part of the discrete wave number can be found in Figure 3.2. The real part converges with a square rate

$$|\Re(\kappa - \kappa_h)| = \mathcal{O}(h^2)$$

meaning the dispersion converges at a very good rate, which would not explain the wave number-dependent asymptotic behaviour. For the imaginary part, on the other hand, there holds

$$|\Im(\kappa - \kappa_h)| = \mathcal{O}(h),$$

only linear convergence. The dissipation error has a lower convergence rate, which results in the asymptotic behaviour of the method.

4. Numerical Results

In this chapter, the favourable properties of the HDG-formulation in Definition 21 are highlighted by various numerical examples. The first part consists of simulations regarding the error estimates in Section 3.3. Additionally, a numerical dispersion and dissipation evaluation is given. The second part introduces the explored iterative solver strategies and highlights their behaviour through examples. All simulations shown in this chapter have been carried out with the open source FEM software Netgen/NGSolve [Sch97, Sch14], which readily provides all the necessary discrete spaces.

4.1. Convergence Rates

In the following the rates of convergence proclaimed in Section 3.3 are tested. The excitations were chosen such that the solutions of the considered Helmholtz equations are smooth plane waves. Additionally, on the considered domains, the solution of the adjoint Helmholtz problem is at least H^2-regular.

4.1.1. Plane Waves in 2D

In the 2D simulation the unit square $\Omega = [0,1] \times [0,1]$ is used as the computational domain. The plane wave

$$u(\mathbf{x}) := e^{j\mathbf{k}\cdot\mathbf{x}}$$

is the solution of the strong Helmholz equation

$$\begin{aligned} -\Delta u - \kappa^2 u &= 0 && \text{in } \Omega, \\ \nabla u \cdot \mathbf{n} + j\kappa u &= j(\mathbf{k}\cdot\mathbf{n} + \kappa)u && \text{on } \partial\Omega, \end{aligned}$$

for a wave vector $\mathbf{k} \in \mathbb{R}^2$ and the coresponding wave number $\kappa = \|\mathbf{k}\|$. For the simulation, a series of hierarchical, structured meshes is used to simulate the h-dependency of the convergence, see Figure 4.1. The wave vector has been chosen as

$$\mathbf{k} = 60 \begin{pmatrix} \cos(\pi/6) \\ \sin(\pi/6) \end{pmatrix}$$

such that the propagation direction of the plane wave does not coincide with the axis, which would lead to one-dimensional problems. Multiple errors which are covered in the analysis have been evaluated. Firstly, the L^2-error of the pressure $\|u - u_h\|_{L^2(\Omega)}$, secondly, the L^2-projected error of the pressure $\|\Pi u - u_h\|_{L^2(\Omega)}$ and thirdly, the best-approximation error of the pressure $\|u - \Pi u\|_{L^2(\Omega)}$. Regarding the flux σ the similar errors $\|\sigma - \sigma_h\|_{L^2(\Omega)}$, $\|\Pi\sigma - \sigma_h\|_{L^2(\Omega)}$, $\|\sigma - \Pi\sigma\|_{L^2(\Omega)}$ have been considered. In the following, the numerical results for different polynomial degrees are shown and discussed, starting with $p = 1$.

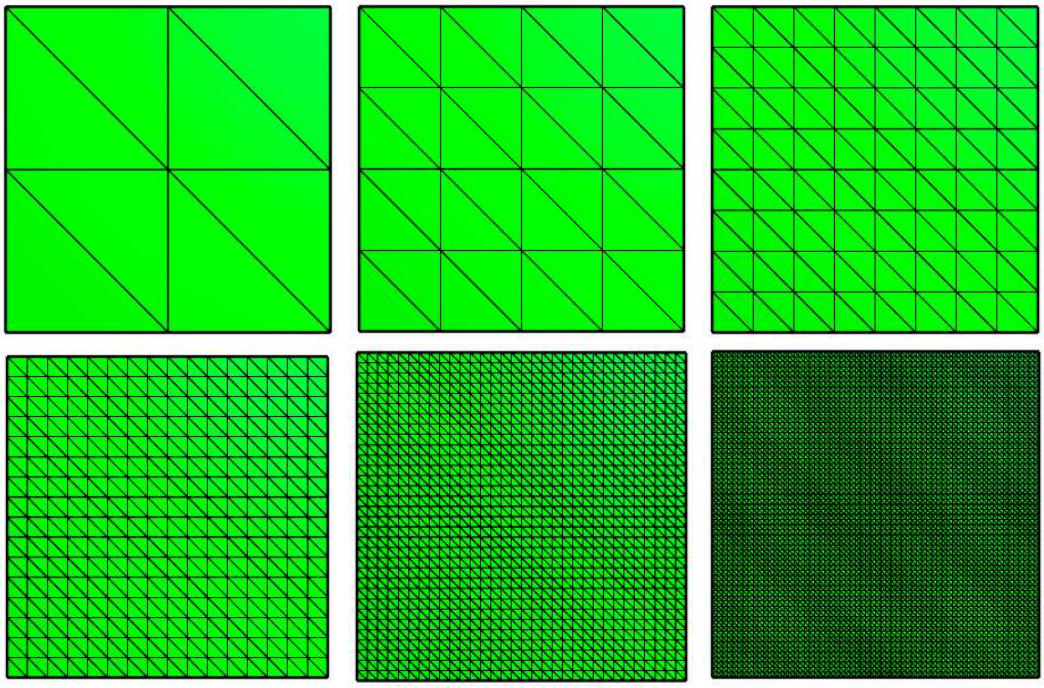


Figure 4.1.: First six meshes of the series of hierarchical 2D structured meshes for $N \in \{2, 4, 8, 16, 32, 64\}$ with $2N^2$ elements. The three additional meshes with $N \in \{128, 256, 512\}$ used in the simulations are omitted.

Polynomial Degree $p = 1$

In Figure 4.2a, the errors in the pressure are shown. For this polynomial degree, the established theory states the following asymptotic convergence rates for the errors in the pressure:

$$\begin{aligned}\|u - u_h\|_{L^2(\Omega)} &= \mathcal{O}\left(h^2\right), \\ \|\Pi u - u_h\|_{L^2(\Omega)} &= \mathcal{O}\left(h^3\right), \\ \|u - \Pi u\|_{L^2(\Omega)} &= \mathcal{O}\left(h^2\right).\end{aligned}$$

The simulation results of this example coincide with the theoretical estimates. Noteworthy is the additional convergence rate of the projected error and the catching up of the error itself to the best approximation error leading to the optimal asymptotic behaviour. The errors for the flux can be found in Figure 4.2b. Following the same line as for the pressure, the established theory states the following asymptotic convergence rates for the errors in the flux:

$$\begin{aligned}\|\sigma - \sigma_h\|_{L^2(\Omega)} &= \mathcal{O}\left(h^2\right), \\ \|\Pi\sigma - \sigma_h\|_{L^2(\Omega)} &= \mathcal{O}\left(h^2\right), \\ \|\sigma - \Pi\sigma\|_{L^2(\Omega)} &= \mathcal{O}\left(h^2\right).\end{aligned}$$

The simulation follows the theory where the major difference to the pressure estimates resides in the lower convergence rate of the projected error of the flux. No higher rate of convergence can be seen, and both the error, as well as the projected error asymptotically have the same rate as the best approximation.

Polynomial Degree $p = 2$

For a second-order discretisation, all errors are expected to asymptotically converge with third-order $\mathcal{O}\left(h^3\right)$ except the projected error of the pressure. The theory states the behaviour of

$$\|\Pi u - u_h\|_{L^2(\Omega)} = \mathcal{O}\left(h^{3+\min\{s-1,2\}}\right)$$

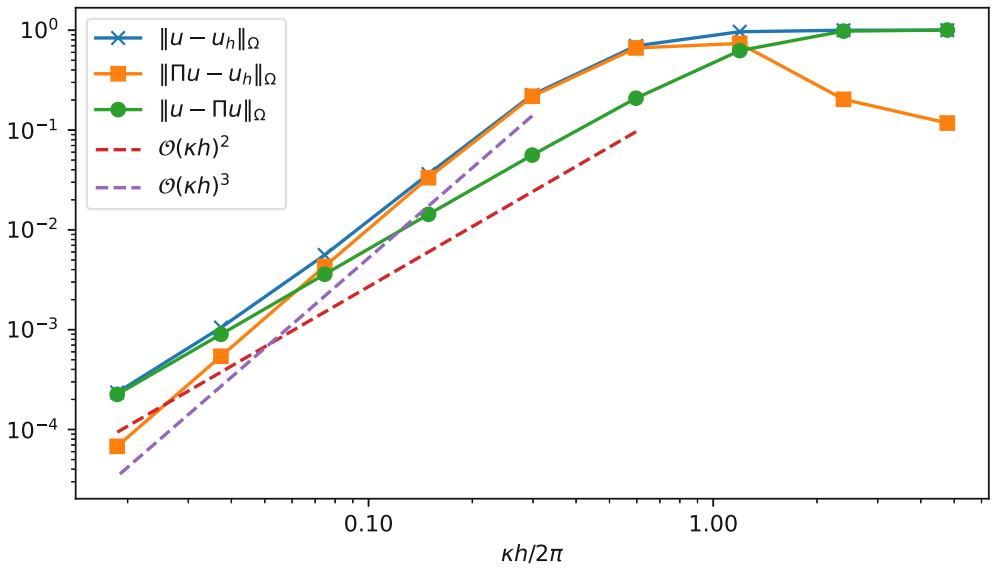
where s is the regularity of the adjoint Helmholtz equation. As one can see in the first-order simulations, the regularity is at least $s \geq 2$. Therefore, the projected error should have a convergence rate in the range between 4 and 5. In Figure 4.3a the errors of the pressure are shown and in Figure 4.3b the errors of the flux can be seen. The simulation result behaves as expected.

A variety of simulations with higher polynomial degrees have been carried out showing the expected asymptotic convergence rates.

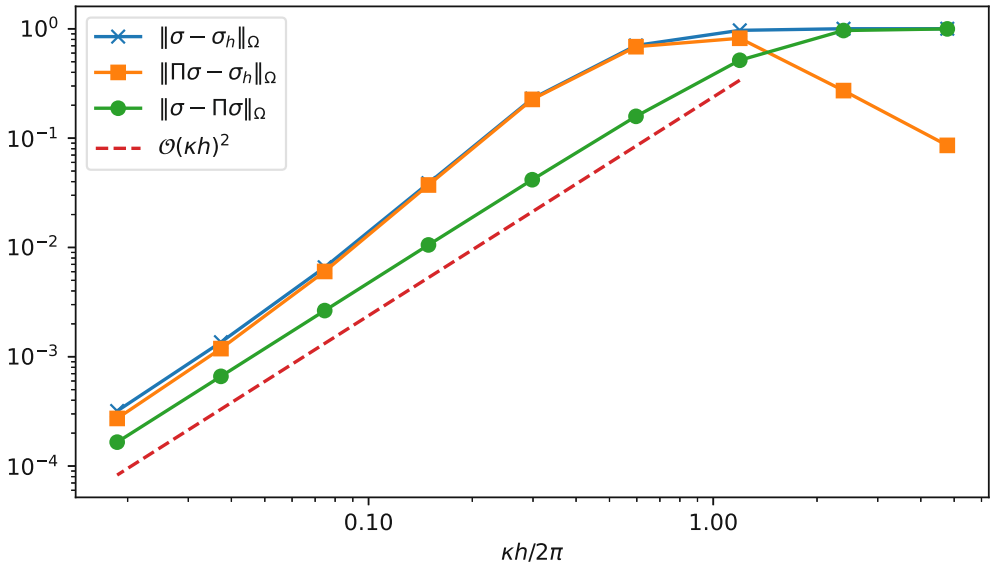
4.1.2. Plane Waves in 3D

Similarly to the 2D case, the unit cube $\Omega = [0,1] \times [0,1] \times [0,1]$ has been chosen as the computational domain for the 3D examples. The chosen plane wave is

$$u(\mathbf{x}) := e^{j\mathbf{k}\cdot\mathbf{x}}$$

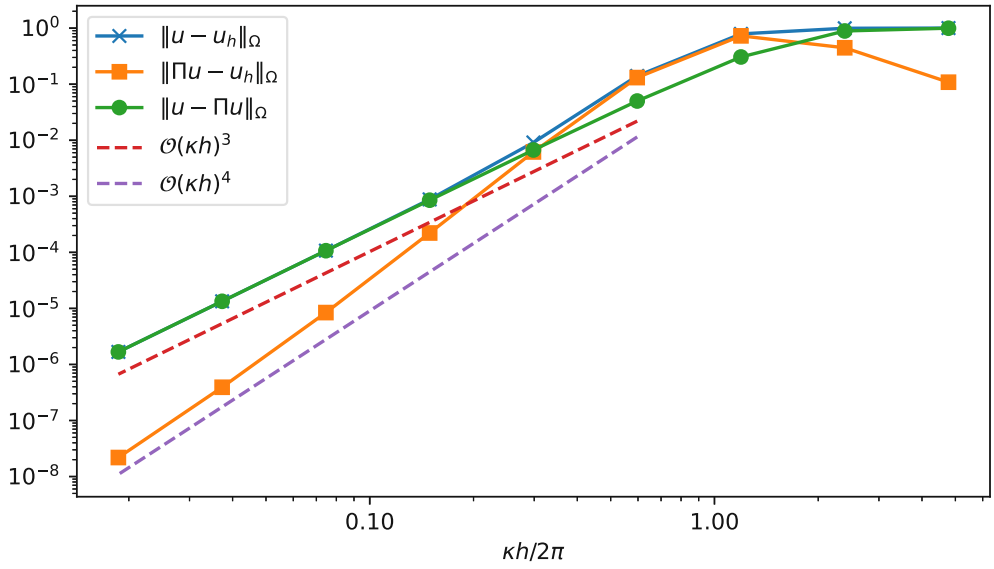


(a) The convergence rates of the errors in the pressure can be seen. Two dashed reference lines representing second and third-order convergence rates are included.

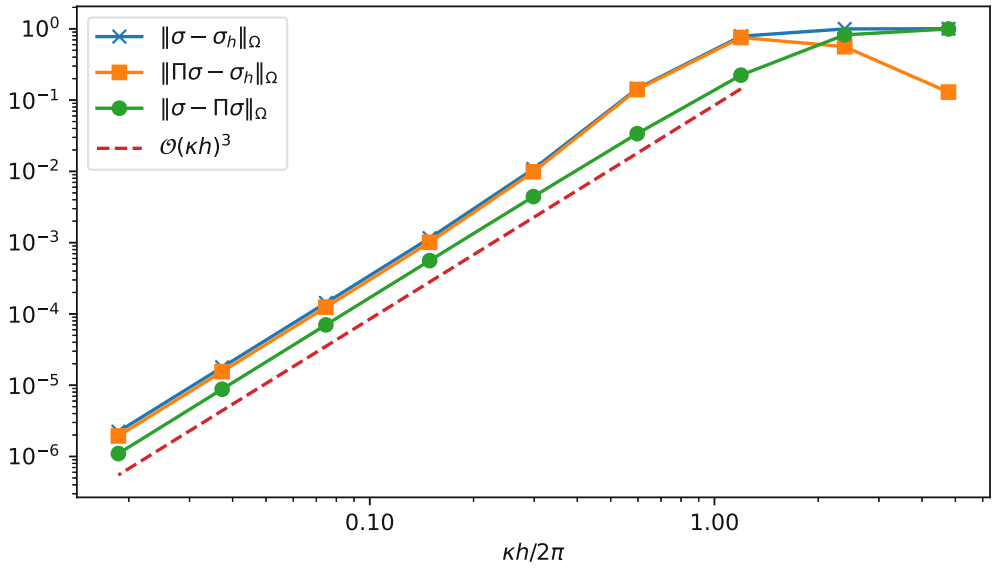


(b) The convergence rates of the errors in the flux can be seen. A dashed reference line representing the second-order convergence rate is included.

Figure 4.2.: The errors for the plane wave 2D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 1$.



(a) The convergence rates of the errors in the pressure can be seen. Two dashed reference lines representing third and fourth-order convergence rates are included.



(b) The convergence rates of the errors in the flux can be seen. A dashed reference line representing the third-order convergence rate is included.

Figure 4.3.: The errors for the plane wave 2D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 2$.

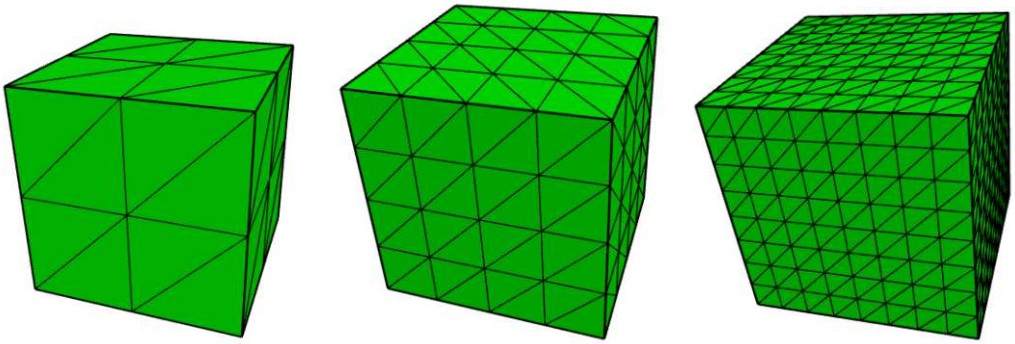


Figure 4.4.: The series of hierarchical, structured 3D meshes for $N \in \{2, 4, 8\}$ with $6N^3$ elements are shown.

which solves the strong Helmholtz equation

$$\begin{aligned} -\Delta u - \kappa^2 u &= 0 && \text{in } \Omega, \\ \nabla u \cdot \mathbf{n} + j\kappa u &= j(\mathbf{k} \cdot \mathbf{n} + \kappa)u && \text{on } \partial\Omega, \end{aligned}$$

for a wave vector $\mathbf{k} \in \mathbb{R}^3$ and the coresponding wave number $\kappa = \|\mathbf{k}\|$. A series of hierarchical, structured 3D meshes, shown partially in Figure 4.4, has been used. The non-axis parallel wave vector

$$\mathbf{k} = 5 \begin{pmatrix} \cos(\pi/6)\cos(\pi/5) \\ \sin(\pi/6)\cos(\pi/5) \\ \sin(\pi/5) \end{pmatrix}$$

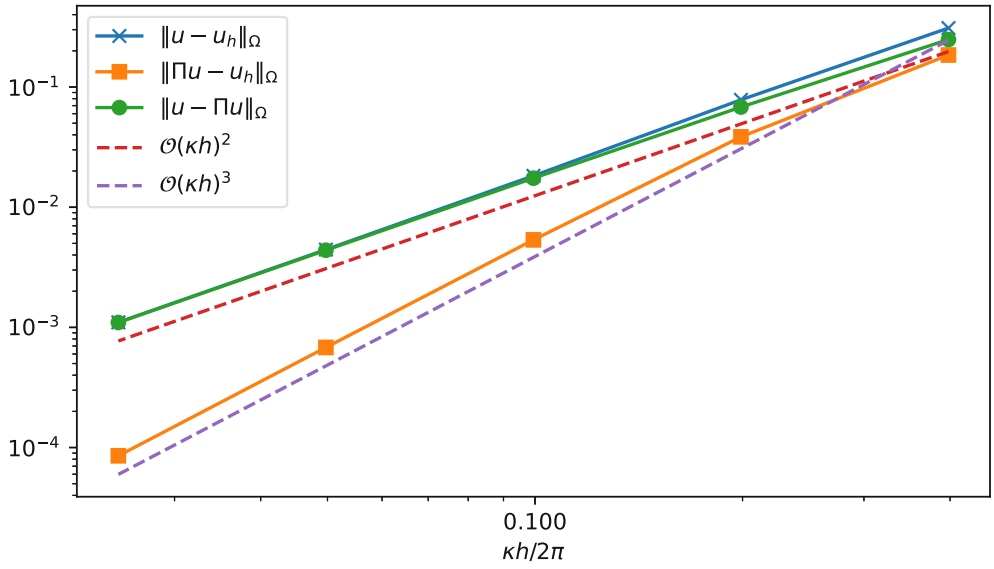
has been selected. The same errors for pressure and flux as in the 2D examples have been considered, and the results for simulations with various polynomial degrees are shown below. The expected rates of convergence coincide with the 2D case for the various errors. For the case of $p = 1$ polynomial degree, the convergence of the pressure can be found in Figure 4.5a and the convergence for the flux in Figure 4.5b. Due to the sharp increase of the problem size through refinement in 3D, only three data points are evaluated, reflecting the convergence behaviour in the asymptotic range. These results also coincide with the proven results.

The results for the second-degree discretisation, $p = 2$, can be found in Figure 4.6a and Figure 4.8b for pressure and flux, respectively. The results are similar to the 2D case with the accelerated convergence rate for the projected error of the pressure.

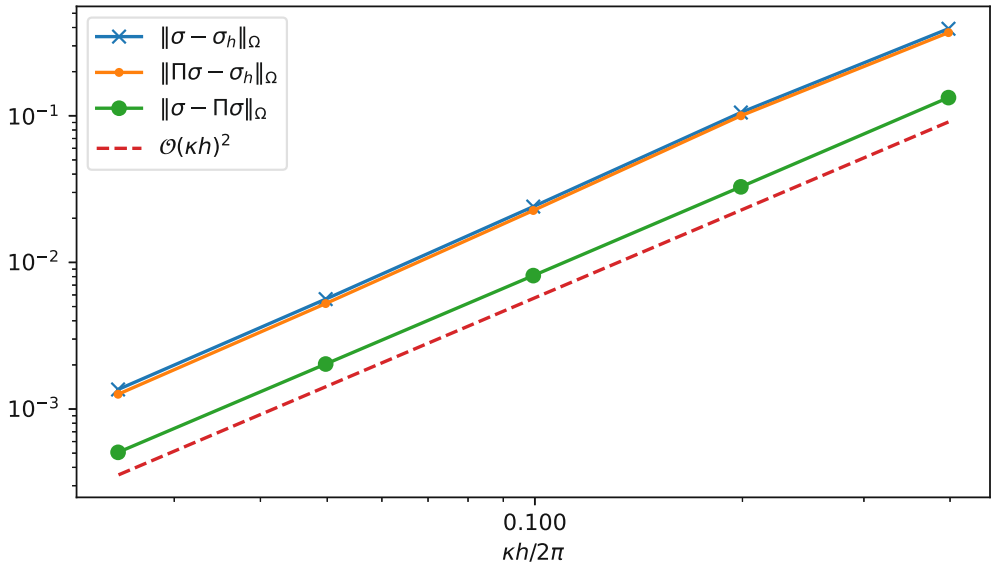
For a better view of the pre-asymptotic convergence range simulations with the higher wave number $\kappa = 20$ have been carried out and can be seen in Figure 4.7 for the polynomial degree $p = 1$ and in Figure 4.8 for $p = 2$.

4.2. Lowest Order Error Rates

In Section 3.3, the established asymptotic theory does not cover the lowest order case of $p = 0$. The absolute stability, as well as the error estimates for the non-asymptotic

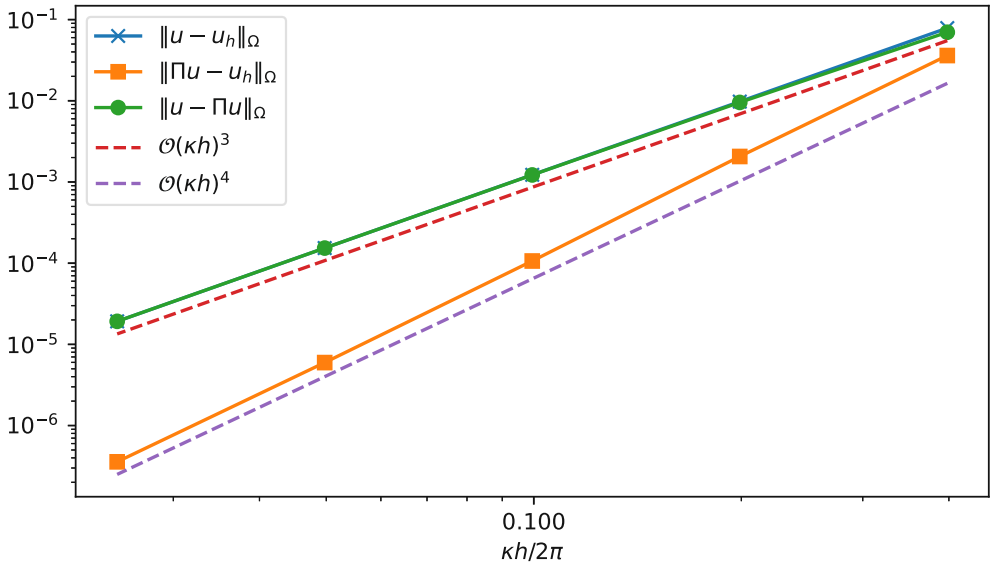


(a) The convergence rates of the errors in the pressure can be seen. Two dashed reference lines representing second and third-order convergence rates are included.

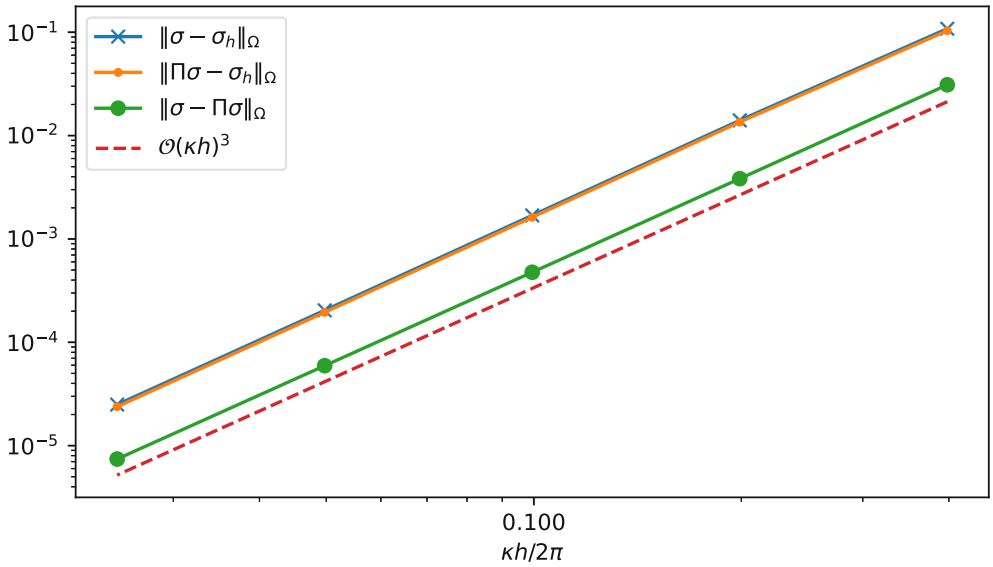


(b) The convergence rates of the errors in the flux can be seen. A dashed reference line representing the second-order convergence rate is included.

Figure 4.5.: The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 1$ and the wave number $\kappa = 5$.

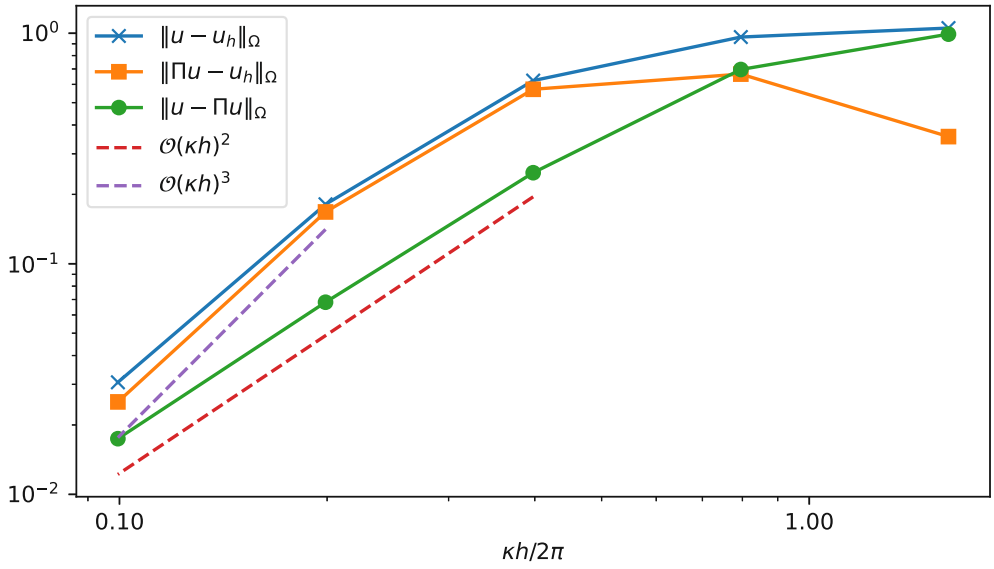


(a) The convergence rates of the errors in the pressure can be seen. Two dashed reference lines representing third and fourth-order convergence rates are included.

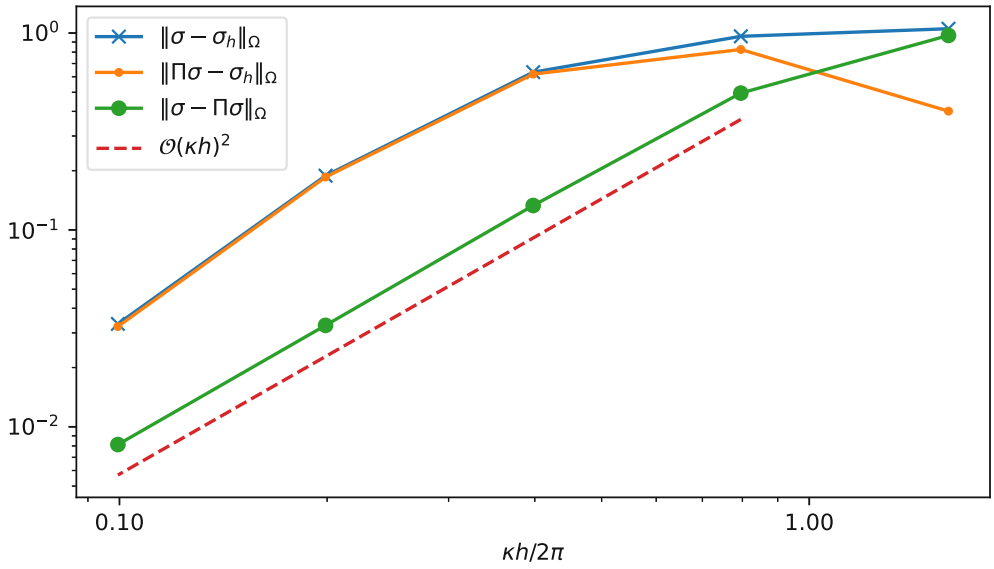


(b) The convergence rates of the errors in the flux can be seen. A dashed reference line representing the third-order convergence rate is included.

Figure 4.6.: The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 2$ and the wave number $\kappa = 5$.

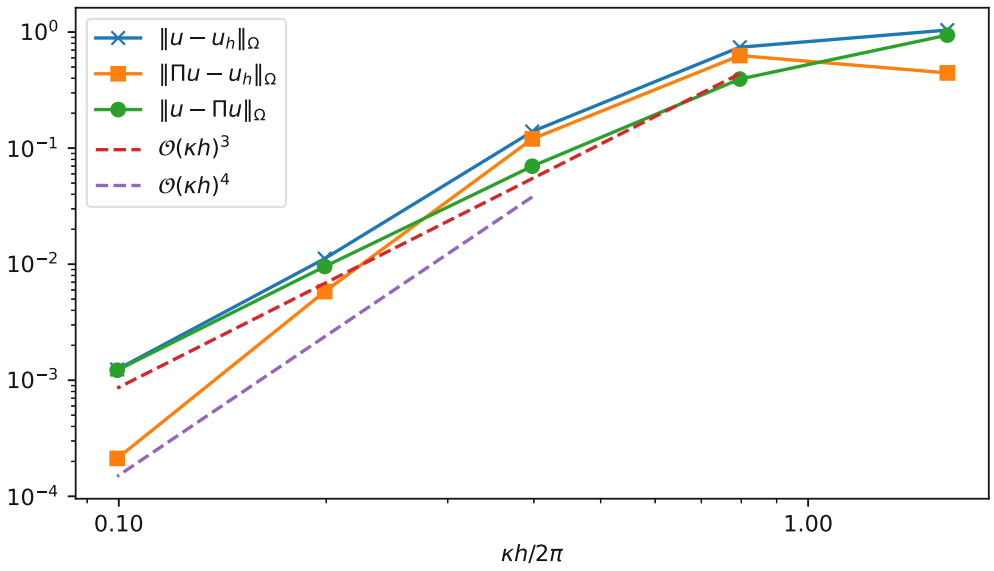


(a) The convergence rates of the errors in the pressure can be seen. Two dashed reference lines representing second and third-order convergence rates are included.

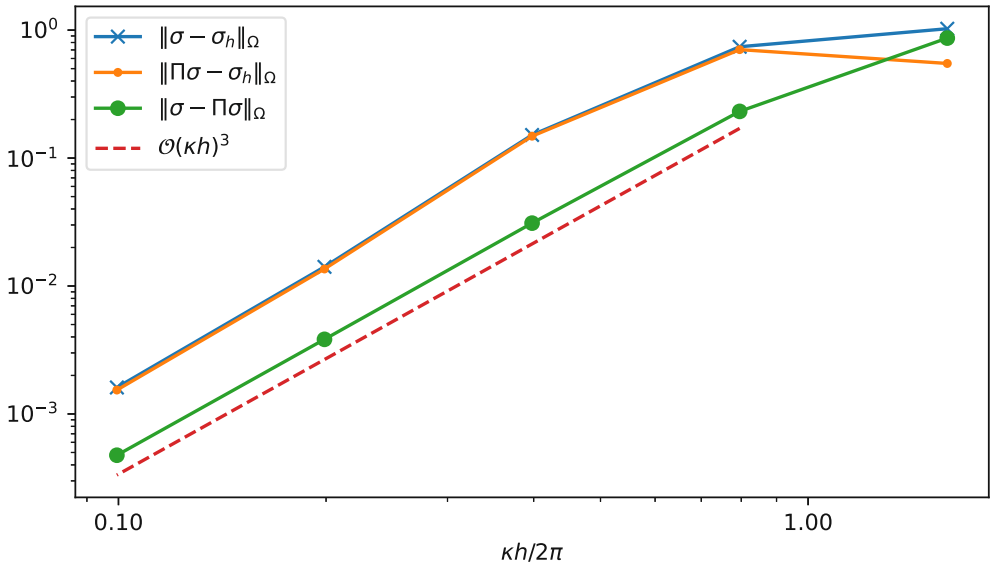


(b) The convergence rates of the errors in the flux can be seen. A dashed reference line representing the second-order convergence rate is included.

Figure 4.7.: The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 1$ and the wave number $\kappa = 20$.



(a) The convergence rates of the errors in the pressure can be seen. Two dashed reference lines representing third and fourth-order convergence rates are included.



(b) The convergence rates of the errors in the flux can be seen. A dashed reference line representing the third-order convergence rate is included.

Figure 4.8.: The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 2$ and the wave number $\kappa = 20$.

range, can be easily generalised for the lowest order case, but these estimates predict an error that deteriorates with the wave number. The analytical dispersion and dissipation analysis established in Subsection 3.3.7 also suggest this behaviour. In the following, the 2D plane wave example is used with the lowest order polynomial degree for multiple wave numbers underlining that the analysis is sharp for the lowest-order case. Simulations for wave numbers $\kappa \in \{10, 20, 40\}$ with the corresponding wave vectors

$$\mathbf{k} = \kappa \begin{pmatrix} \cos(\pi/6) \\ \sin(\pi/6) \end{pmatrix}$$

have been carried out on a series of structured meshes, see Figure 4.1. The errors of the pressure can be seen in Figure 4.9a. In the figure, the x-axis is scaled to the ratio of the mesh size to the wavelength such that the best approximation error for different wave numbers will overlap. This can be seen in the best approximation errors for the different chosen wave numbers. With this scaling, the wave number-dependency of the errors can easily be seen. The error of the pressure does not converge to the best possible approximation any more, and this effect deteriorates with increasing wave number. An optimal asymptotic result with respect to the wave number for lowest order polynomial degree is impossible. This is also indicated by the convergence rate of the projected error. For higher polynomial degrees, the higher convergence rate of the projected error is essential in the asymptotic analysis for the pollution to vanish. In the lowest order case, no higher convergence rate of the projected error is seen, and the error is always dominated by this pollution. The error of the flux follows the same reasoning, and the errors can be found in Figure 4.9b.

4.3. Dispersion and Dissipation

The concept of pollution is closely related to the dispersion and dissipation of a method. In the following, these two properties are numerically evaluated for the HDG-formulation and for the conforming H^1-formulation for the Helmholtz equation. The unit square in 2D $\Omega = [0, 1] \times [0, 1]$ is chosen as domain. The question is, how well can a plane wave

$$u_{\mathbf{k}}(\mathbf{x}) := e^{j\mathbf{k}\cdot\mathbf{x}},$$

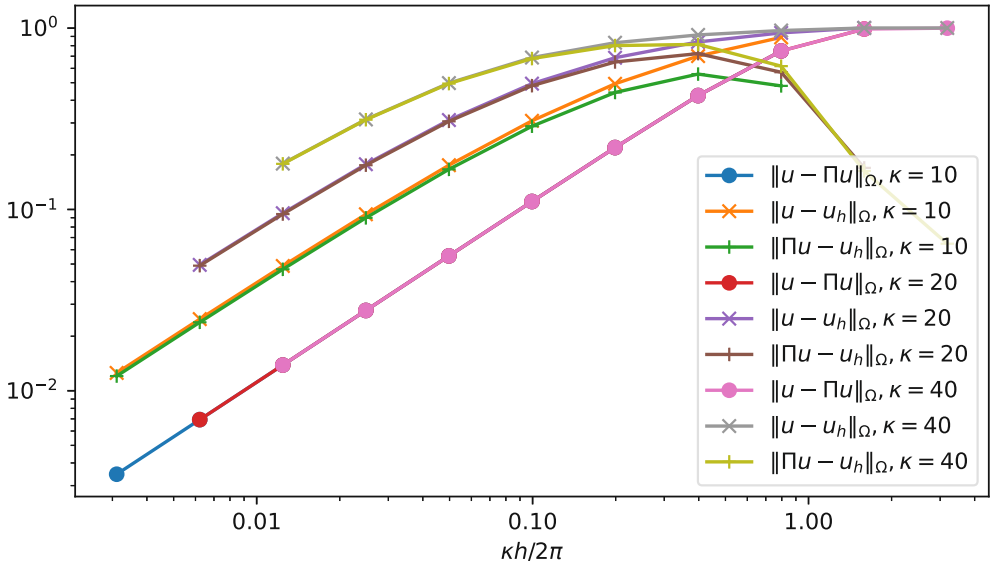
for a given wave vector $\mathbf{k} = (\mathbf{k}_x, \mathbf{k}_y)^\top \in \mathbb{R}^2$, be approximated by the numerical scheme. To this end, the eigenvalue problem, with a quasi-periodic side condition, is considered and the eigenvalue closest to $\kappa = \|\mathbf{k}\|$ represents the best approximation of that plane wave. For the unit square example, the quasi-periodicity means

$$u_{\mathbf{k}}(x = 1, y) = e^{j(\mathbf{k}_x + \mathbf{k}_y y)} = e^{j\mathbf{k}_x} e^{j\mathbf{k}_y y} = e^{j\mathbf{k}_x} u_{\mathbf{k}}(x = 0, y)$$

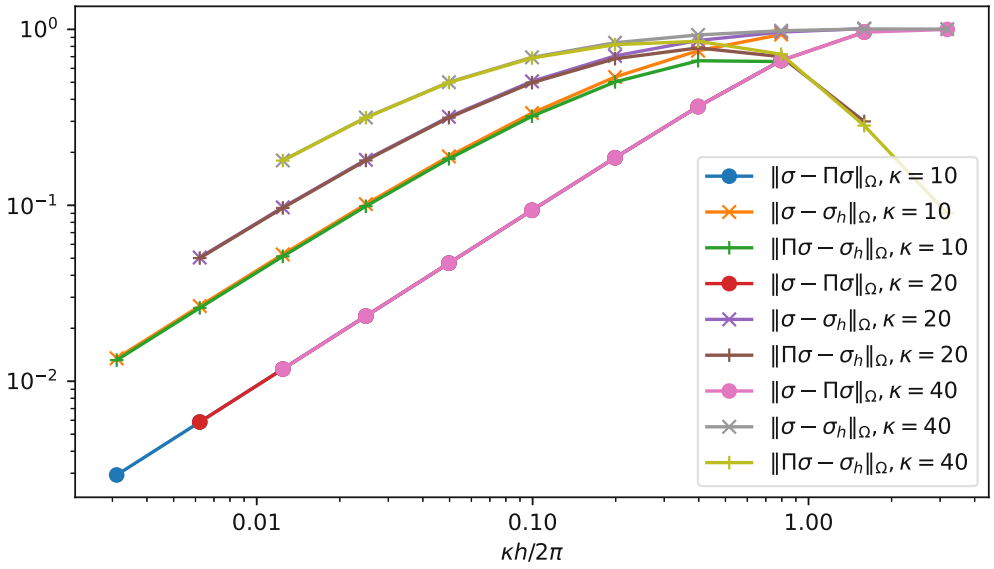
in x-direction connecting the left and right boundary and

$$u_{\mathbf{k}}(x, y = 1) = e^{j(\mathbf{k}_x x + \mathbf{k}_y)} = e^{j\mathbf{k}_x x} e^{j\mathbf{k}_y} = e^{j\mathbf{k}_y} u_{\mathbf{k}}(x, y = 0)$$

in y-direction for the bottom and top boundary. This quasi-periodicity is enforced with a Lagrange multiplier supported on the boundary $\partial\Omega$ and aligning with the trace of the



(a) The convergence rates of the errors in the pressure are shown.



(b) The convergence rates of the errors in the flux are shown.

Figure 4.9.: The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The results for wave numbers $\kappa \in \{10, 20, 40\}$ can be seen. The polynomial degree in the simulation has been chosen as $p = 0$.

respective formulation. By collecting the shift factors into the variable

$$\gamma := \begin{cases} e^{j\mathbf{k}_x} & \text{left,} \\ e^{j\mathbf{k}_y} & \text{bottom,} \\ -1 & \text{right,} \\ -1 & \text{top,} \end{cases}$$

the eigenvalue problem for the standard H^1-formulation is:

Definition 54 (H^1-Eigenvalue Problem)**.** *For a given wave vector* $\mathbf{k} \in \mathbb{R}^2$ *find* $(u, p, \lambda) \in H^1(\Omega) \times L^2(\partial\Omega, \textit{periodic}) \times \mathbb{C}$ *so that*

$$(\nabla u, \nabla v)_\Omega - \lambda^2 (u, v)_\Omega + (\gamma u, q)_{\partial\Omega} + (p, \gamma v)_{\partial\Omega} = 0$$

holds for all $(v, q) \in H^1(\Omega) \times L^2(\partial\Omega, \textit{periodic})$.

Only considering the term with the periodic test function q on the left-right leads to

$$0 = (\gamma u, q)_{left} + (\gamma u, q)_{right} = (\gamma_l u_l + \gamma_r u_r, q)_{l,r},$$

which implies

$$\gamma_l u_l + \gamma_r u_r = 0 \Leftrightarrow \gamma_l u_l = -\gamma_r u_r \Leftrightarrow e^{j\mathbf{k}_x} u_l = u_r.$$

Similarly, there holds for the bottom-top term

$$0 = (\gamma u, q)_{bottom} + (\gamma u, q)_{top} = (\gamma_b u_b + \gamma_t u_t, q)_{b,t},$$

implying

$$\gamma_b u_b + \gamma_t u_t = 0 \Leftrightarrow \gamma_b u_b = -\gamma_t u_t \Leftrightarrow e^{j\mathbf{k}_y} u_b = u_t.$$

Therefore, these equations imply the quasi-periodicity of a plane wave. For the HDG-formulation the eigenvalue problem is:

Definition 55 (HDG-Eigenvalue Problem)**.** *For a given wave vector* $\mathbf{k} \in \mathbb{R}^2$ *find* $\sigma \in H_{pw}(\mathrm{div})(\mathcal{T}) \cap [H^s_{pw}(\mathcal{T})]^d$, $u \in H^s_{pw}(\mathcal{T})$, $\hat{u} \in L^2(\mathcal{F})$, $\hat{\sigma}_\mathbf{n} \in L^2(\mathcal{F})$, $p \in L^2(\partial\Omega, \textit{periodic})$, $\lambda \in \mathbb{C}$, *with* $s > 1/2$ *so that*

$$B(\sigma, \hat{\sigma}_\mathbf{n}, u, \hat{u}; \tau, \hat{\tau}_\mathbf{n}, v, \hat{v}; \lambda) + (\gamma\hat{u}, q)_{\partial\Omega} + (p, \gamma\hat{v})_{\partial\Omega} = 0$$

holds for all $\tau \in H_{pw}(\mathrm{div})(\mathcal{T}) \cap [H^s_{pw}(\mathcal{T})]^d$, $v \in H^s_{pw}(\mathcal{T})$, $\hat{v} \in L^2(\mathcal{F})$, $\hat{\tau}_\mathbf{n} \in L^2(\mathcal{F})$, $q \in L^2(\partial\Omega, \textit{periodic})$ *with the sesquilinear form*

$$\begin{aligned} B(\sigma, \hat{\sigma}_\mathbf{n}, u, \hat{u}; \tau, \hat{\tau}_\mathbf{n}, v, \hat{v}; \lambda) := \sum_{T\in\mathcal{T}} & j\lambda(\sigma, \tau)_T + (u, \mathrm{div}\,\tau)_T + (\mathrm{div}\,\sigma, v)_T - j\lambda(u, v)_T \\ & -(\hat{u}, \tau\cdot\mathbf{n})_{\partial T} - (\sigma\cdot\mathbf{n}, \hat{v})_{\partial T} - \alpha([u], [v])_{\partial T} + \beta([\![\sigma]\!], [\![\tau]\!])_{\partial T}. \end{aligned}$$

It is noteworthy that in the used sesquilinear form, the Robin boundary condition is replaced by the quasi-periodicity condition. In the HDG-formulation, the facet variable is coupled contrary to the volume variable of the H^1-formulation.

Both eigenvalue problems are discretised with their appropriate spaces, and the Lagrange multiplier space is large enough to strongly enforce the quasi-periodicity. The simulations are carried out on structured meshes, see Figure 4.1. To get a full view of the dispersion and dissipation behaviour of the formulations, all possible propagation directions need to be explored, which will be numerically evaluated by a certain amount of snapshots for different wave vectors. Due to the symmetry of the problem with respect to the sign of the wave vector, only half of all directions need to be explored and due to the further symmetry of the mesh, only one-eighth would be necessary, but one-fourth is chosen. In the simulations, the used polynomial degrees, as well as the mesh, will always be specified and for a given wave number $\kappa > 0$, the wave vector will be chosen on one-fourth of the circle with a radius of the wave number, or more specific

$$\mathbf{k} = \kappa \begin{pmatrix} \cos\phi \\ \sin\phi \end{pmatrix}, \qquad \phi \in [0, \pi/2].$$

4.3.1. H^1 and HDG-Comparison

In the following, the dispersion and dissipation behaviour of the H^1- and HDG-formulation are compared. For the first comparison, a structured mesh with $N = 2$ and a wave number of $\kappa = 3$ is used. For the HDG-formulation, the polynomial degree $p = 0$ is used, and for the H^1-formulation, the polynomial degree is $p = 1$. This is reasonable because the gradient in the H^1-formulation is comparable to σ in the HDG-formulation. For thirty angles, both eigenvalue problems have been solved, and the eigenvalue κ_h closest to the prescribed wave number κ has been chosen as the closest approximation of the analytical plane wave by the respective discrete scheme. Then, the discrete wave vector is calculated by

$$\mathbf{k}_h := \frac{\kappa_h}{\kappa}\mathbf{k}.$$

In Figure 4.10, the dispersion of both methods can be seen for one-fourth of the propagation directions. The blue line represents the optimal dispersion without error. As can be seen, the HDG-formulation for the lowest order case has a favourable dispersion behaviour. For a better view of the dispersion, Figure 4.11a, which compares the real part of the wave numbers, is more useful. The real part dictates the dispersion, and the imaginary part the dissipation, which can be seen for both methods in Figure 4.11b. Unsurprisingly, the H^1-formulation does not have any dissipation. This is also easily explained by the Hermitian eigenvalue problem. On the contrary, the HDG-formulation has dissipation, which can also be seen in numerical simulations and is essential for the absolute stability of the method.

4.3.2. Wave Number Experiments

In this subsection, the dispersion and dissipation of the HDG-formulation for the wave numbers $\kappa \in \{0.75, 1.5, 3, 6\}$ is explored. As discretisation, the structured mesh with

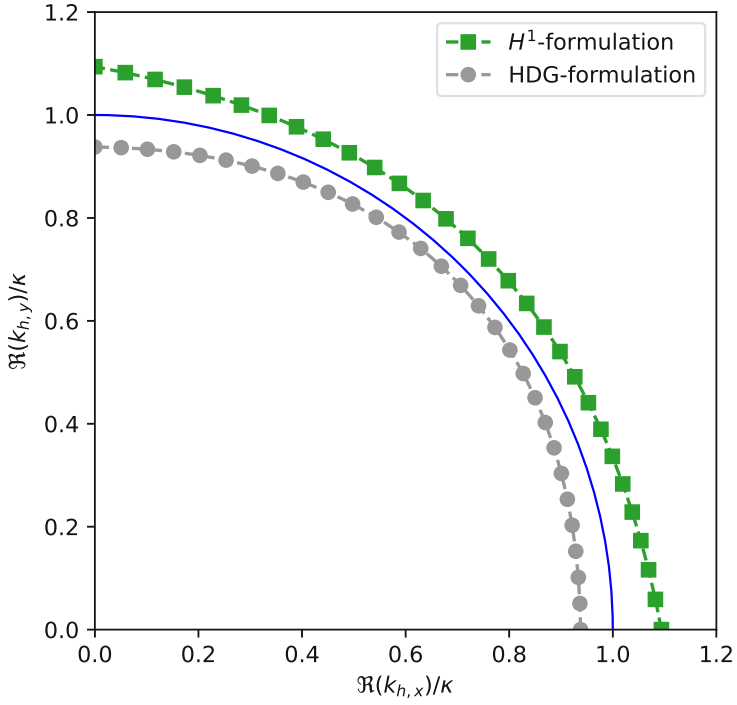
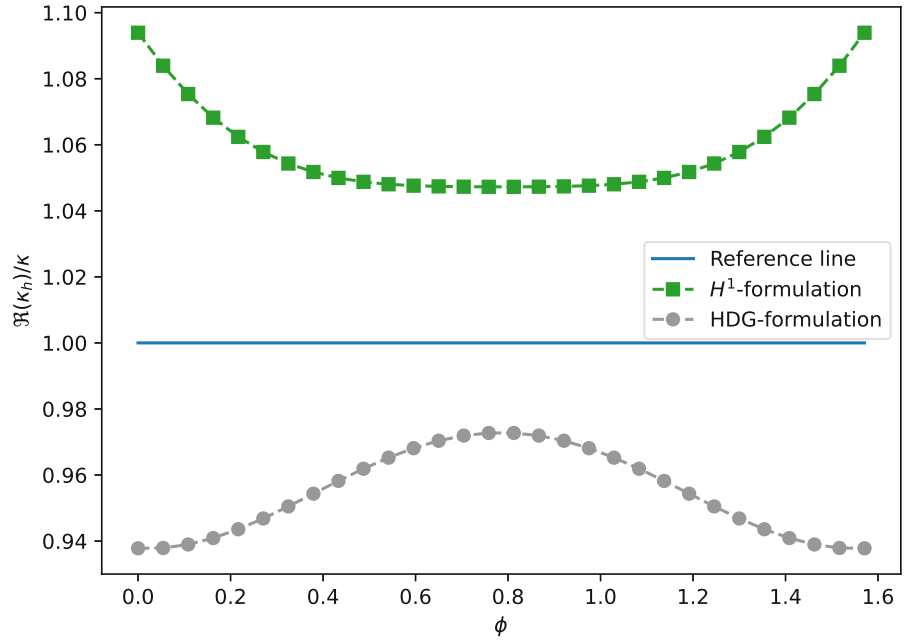
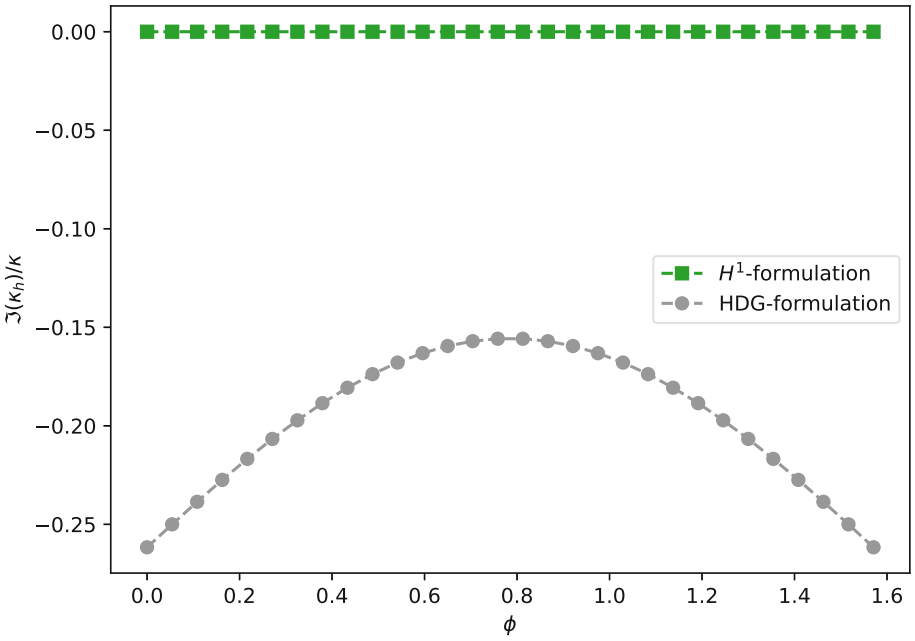


Figure 4.10.: Dispersion for plane waves with a wave number $\kappa = 3$ and angles $\phi \in [0, \pi/2]$ on a structured mesh with $N = 2$. The H^1-formulation in green with square markers and the HDG-formulation in grey with circle marks. The reference line with radius one is highlighted in blue.



(a) Dispersion. The reference line is highlighted in blue.



(b) Dissipation.

Figure 4.11.: Dispersion and dissipation for plane waves with a wave number $\kappa = 3$ and angles $\phi \in [0, \pi/2]$ on a structured mesh with $N = 2$. The H^1-formulation in green with square markers and the HDG-formulation in grey with circle marks.

$N = 2$ and the first-order polynomial degree $p = 1$ are used. The dispersion can be seen in Figure 4.12a and the dissipation in Figure 4.12b. As can be seen with increasing wave number, the dispersion and dissipation also increase on an identical discretisation. The more relevant experiment regarding pollution is when the ratio of unknowns per wavelength is kept constant for multiple wave numbers meaning $\kappa h = \mathcal{O}(1)$. Then, the analysis predicts a behaviour of the pollution error of $\mathcal{O}(\kappa)$ with respect to the wave number. In the following, this ratio is kept constant for multiple wave numbers, namely $\kappa h = 2$. For each wave number, the maximum pollution error over all angles

$$\|\kappa_h - \kappa\| := \max_{\phi \in [0,\pi/2]} |\kappa_h - \kappa|$$

is evaluated. The results for spaces with polynomial degree $p = 0$ can be seen in Figure 4.13a, for $p = 1$ in Figure 4.13b and for $p = 2$ in Figure 4.13c. It can be seen that for all polynomial degrees, the pollution error increases with the in the wave number expected rate.

4.4. Iterative Solver and Preconditioners

For similar HDG-methods, Martin Huber has already published results regarding iteratively solving with preconditioners in the papers [HPS13, HS14] and in his dissertation [Hub13].

The restarted, preconditioned GMRES algorithm is applied to solve the discretised HDG-formulation in Definition 21. The used FEM software Netgen/NGSolve [Sch97, Sch14] supports static condensation for HDG-methods. Additionally, Netgen/NGSolve contains a parallel implementation of the GMRES algorithm for which custom preconditioners can be developed. The GMRES solver is restarted after 20 iterations and the error is measured in the Euclidean norm on the space $\mathbb{C}^n$. The GMRES iteration is stopped after the relative error is less than 10^5.

In the following, four pre-conditioning strategies are introduced and discussed. They all share the common property that they only operate on the Schur complement after static condensation. Therefore, the preconditioner only operates on the hybrid facet unknowns $\hat{u}$ and $\hat{\sigma}$. Whenever intermediate solutions after some GMRES iterations are shown, the current facet unknowns are extended back onto the elements for the visualisation.

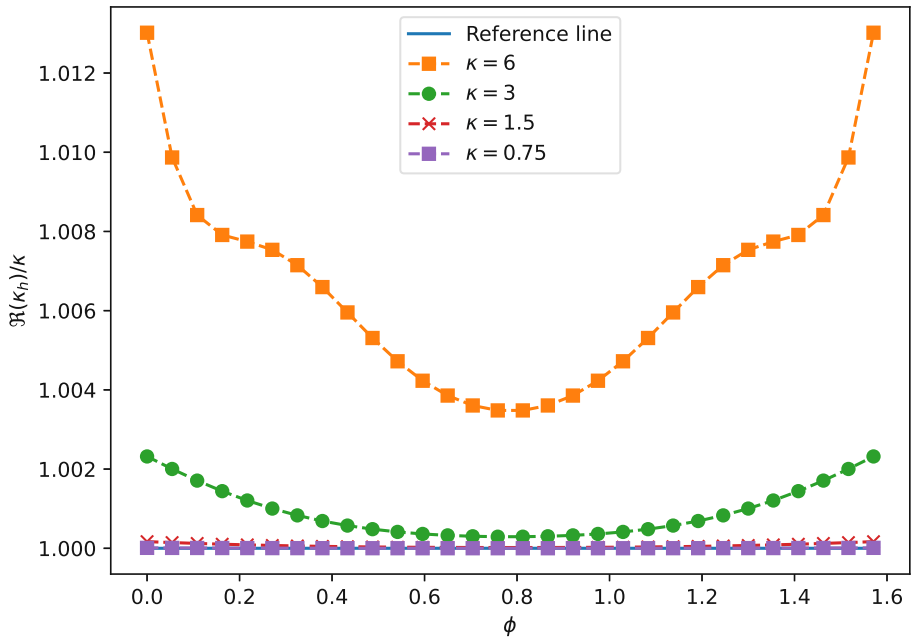
The goal is to solve the system of linear equations

$$Sx = y,$$

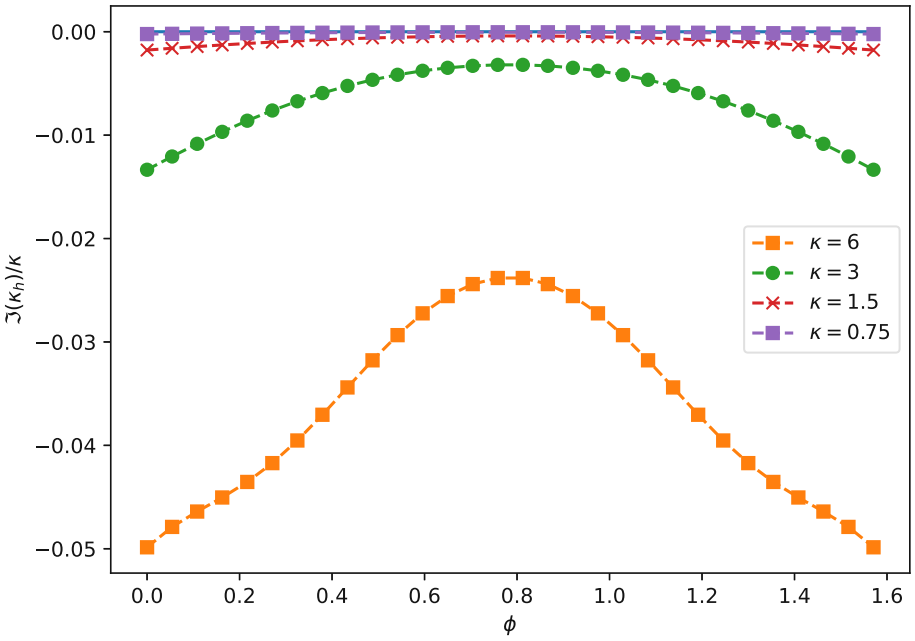
where $S \in \mathbb{C}^{n \times n}$ is the Schur complement of the discrete problem, $y \in \mathbb{C}^n$ is the discretised right hand side in combination with static condensation and $x \in \mathbb{C}^n$ is the discrete vector representing the hybrid facet variables.

To highlight the different preconditioners, the Helmholtz problem on the unit square $\Omega := [0,1]^2$ with the mesh in Figure 4.14 is chosen. The physical parameters are set to one, the wave number to $\kappa = 60$ and the excitations to

$$f(\mathbf{x}) := 0, \qquad g(\mathbf{x}) := \begin{cases} j\kappa e^{-10(y-1/2)^2} & \text{on the left boundary,} \\ 0 & \text{else.} \end{cases}$$

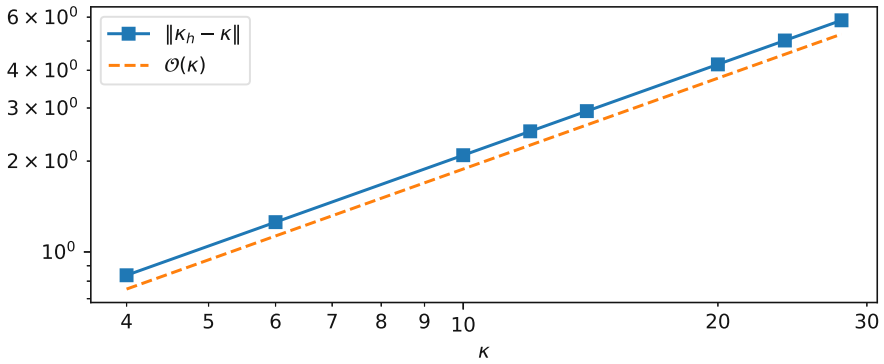


(a) Dispersion.

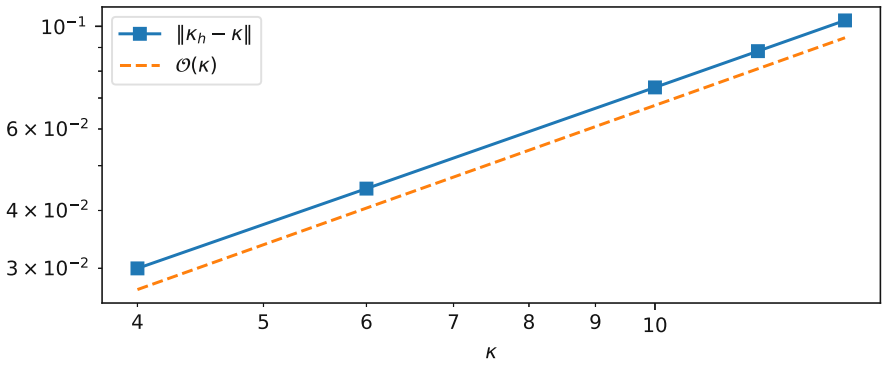


(b) Dissipation.

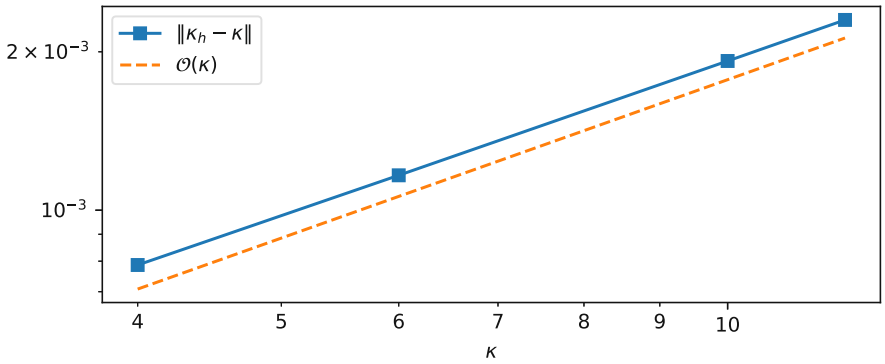
Figure 4.12.: Dispersion and dissipation of the HDG-formulation for plane waves with wave numbers $\kappa \in \{0.75, 1.5, 3, 6\}$ and angles $\phi \in [0, \pi/2]$ on a structured mesh with $N = 2$ and the polynomial degree $p = 1$.



(a) $p = 0$



(b) $p = 1$



(c) $p = 2$

Figure 4.13.: Pollution error of the HDG-formulation with respect to the wave number on structured meshes. The resolution is kept constant regarding $\kappa h = 2$ for spaces with varying polynomial degree p.

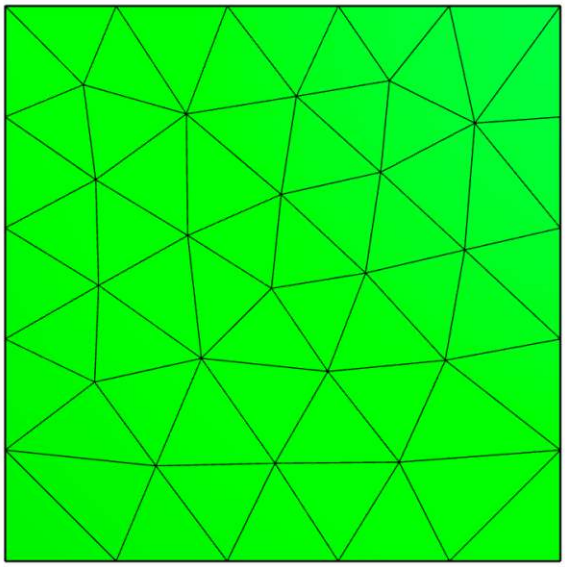


Figure 4.14.: The mesh of the unit square for the preconditioner introduction example is shown.

There is no volume excitation, and the Gaussian peak on the left boundary leads to an impinging wave from the left side centred in the middle of the boundary. Finally, the polynomial degree is set to $p = 6$.

4.4.1. Block-Jacobi Preconditioner

The first considered preconditioner is the Block-Jacobi preconditioner, which falls into the class of additive Schwarz preconditioners. A general additive Schwarz preconditioner has the following properties. Considering a finite set of restriction matrices $R_i \in \mathbb{R}^{m \times n}$ into m-dimensional subspaces and the complementary prolongations $R_i^\top \in \mathbb{R}^{n \times m}$ then the large matrix S can be restricted to the smaller subspaces by

$$S_i := R_i S R_i^\top .$$

Now, the additive Schwarz preconditioner can be written as

$$C_{AS}^{-1} := \sum_i R_i^\top S_i^{-1} R_i,$$

which is applied in each GMRES-iteration. Basically, instead of solving one large system of linear equations, many, in practice, very small ones are solved as an approximation. Here, it is also obvious why the local absolute stability of a formulation is crucial, because it asserts the invertibility of the local matrices S_i. The behaviour of the preconditioned iterative solver is greatly influenced by the specific choice of the subspaces. In the following, subspaces for 2D and 3D are suggested, but by no means does the author claim that these subspaces are the only possible or optimal choices.

Restriction Matrices in 2D

For 2D problems the subspaces are element boundary patches of facet unknowns. The mesh of a 2D domain consists of

$$n_{\mathcal{F}} := |\mathcal{F}|$$

facets and each facet F_i is uniquely identified by an index $i \in \{1, 2, \ldots, n_{\mathcal{F}}\}$. The polynomial degree p is fixed to some natural number. On each facet $F \in \mathcal{F}$ the discrete variables

$$\hat{u}_{h,F} := \hat{u}_h|_F \in \mathcal{P}^p(F), \qquad \hat{\sigma}_{\mathbf{n},h,F} := \hat{\sigma}_{\mathbf{n},h}|_F \in \mathcal{P}^p(F)$$

are used in the Schur complement of the HDG-formulation. The dimensions of these discrete spaces are

$$\dim(\mathcal{P}^p(F)) = p + 1, \qquad \forall F \in \mathcal{F}.$$

Therefore, on each facet each discrete variable $\hat{u}_{h,F}$ and $\hat{\sigma}_{\mathbf{n},h,F}$ can be associated to $p+1$ unknowns:

$$\begin{aligned} \hat{u}_{h,F} \quad &\widehat{=} \quad x_{u,F} := (x_{u,F,1}, x_{u,F,2}, \ldots, x_{u,F,p+1})^\top \in \mathbb{C}^{p+1}, \\ \hat{\sigma}_{\mathbf{n},h,F} \quad &\widehat{=} \quad x_{\sigma,F} := (x_{\sigma,F,1}, x_{\sigma,F,2}, \ldots, x_{\sigma,F,p+1})^\top \in \mathbb{C}^{p+1}. \end{aligned}$$

The dimensions of the Schur complement are

$$S \in \mathbb{C}^{2(p+1)n_{\mathcal{F}} \times 2(p+1)n_{\mathcal{F}}}$$

and for the solution vector $x \in \mathbb{C}^{2(p+1)n_{\mathcal{F}}}$ holds the representation

$$x = \left(x_{u,F_1}, x_{u,F_2}, \ldots, x_{u,F_{n_{\mathcal{F}}}}, x_{\sigma,F_1}, x_{\sigma,F_2}, \ldots, x_{\sigma,F_{n_{\mathcal{F}}}}\right)^\top.$$

The first half of x contains the unknowns associated to the variable $\hat{u}_h$ and the second half the unknowns associated to $\hat{\sigma}_{h,\mathbf{n}}$. The restriction R_{u,F_i} of $\hat{u}_h$ onto the facet F_i, $i \in \{1, 2, \ldots, n_{\mathcal{F}}\}$

$$R_{u,F_i}(\hat{u}_h) = \hat{u}_{h,F_i},$$

has the matrix representation

$$R_{u,i} := \left(0^{(p+1)\times(i-1)(p+1)}, I^{(p+1)\times(p+1)}, 0^{(p+1)\times(2n_{\mathcal{F}}-i)(p+1)}\right) \in \mathbb{R}^{(p+1)\times 2(p+1)n_{\mathcal{F}}},$$

where $I^{m\times m}$ is the identity matrix in $\mathbb{R}^{m\times m}$ and $0^{n\times m} \in \mathbb{R}^{n\times m}$ consists of zero entries. Similarly to $\hat{u}_h$, the restriction R_{σ,F_i} of $\hat{\sigma}_{\mathbf{n},h}$ onto the facet F_i, $i \in \{1, 2, \ldots, n_{\mathcal{F}}\}$

$$R_{\sigma,F_i}(\hat{\sigma}_{\mathbf{n},h}) = \hat{\sigma}_{\mathbf{n},h,F_i},$$

has the matrix representation

$$R_{\sigma,i} := \left(0^{(p+1)\times(n_{\mathcal{F}}+i-1)(p+1)}, I^{(p+1)\times(p+1)}, 0^{(p+1)\times(n_{\mathcal{F}}-i)(p+1)}\right) \in \mathbb{R}^{(p+1)\times 2(p+1)n_{\mathcal{F}}}.$$

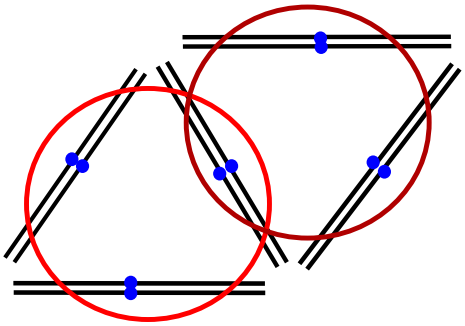


Figure 4.15.: Two element boundary patches of facet unknowns on a 2D skeleton mesh for the Jacobi and Gauss-Seidel preconditioners. The facet unknowns are highlighted with blue dots, and the patches with a red circles.

The number of elements $n_{\mathcal{T}}$ in the 2D mesh is

$$n_{\mathcal{T}} := |\mathcal{T}|.$$

For each element $T_l \in \mathcal{T}$, $l \in \{1, 2, \dots, n_{\mathcal{T}}\}$ its boundary consists of three facets $F_{i_{l,1}}$, $F_{i_{l,1}}$, $F_{i_{l,1}}$ which are uniquely defined by the indices $i_{l,1}, i_{l,1}, i_{l,1} \in \{1, 2, \dots, n_{\mathcal{F}}\}$. The restriction onto the facet unknowns of the element boundary patch for an element $T_l \in \mathcal{T}$ is

$$R_l := \sum_{m=1}^{3} R_{u,i_{l,m}} + R_{\sigma,i_{l,m}}$$

and the Block-Jacobi preconditioner takes the form

$$C_J^{-1} := \sum_{l=1}^{n_{\mathcal{T}}} R_l^\top S_l^{-1} R_l$$

with the matrices

$$S_l := R_l S R_l^\top .$$

A simple example of two patches can be seen in Figure 4.15. The patches are overlapping, which also leads to an overlapping Block-Jacobi preconditioner.

Restriction Matrices in 3D

On 3D domains, unknowns per facet are combined into blocks. The mesh consists of

$$n_{\mathcal{F}} := |\mathcal{F}|$$

facets which are numbered in ascending order. On each facet $F \in \mathcal{F}$ the discrete variables

$$\hat{u}_{h,F} := \hat{u}_h|_F \in \mathcal{P}^p(F), \qquad \hat{\sigma}_{\mathbf{n},h,F} := \hat{\sigma}_{\mathbf{n},h}|_F \in \mathcal{P}^p(F)$$

of polynomial degree p are used in the Schur complement of the HDG-formulation. The dimensions of these discrete spaces are

$$\dim(\mathcal{P}^p(F)) = \frac{(p+1)(p+2)}{2} =: n_p, \qquad \forall F \in \mathcal{F}.$$

With a similar notation as in the 2D case, each discrete variable $\hat{u}_{h,F}$ and $\hat{\sigma}_{\mathbf{n},h,F}$ can be associated to n_p unknowns per facet:

$$\begin{aligned} \hat{u}_{h,F} \quad &\widehat{=} \quad x_{u,F} := \left(x_{u,F,1}, x_{u,F,2}, \ldots, x_{u,F,n_p}\right)^\top \in \mathbb{C}^{n_p}, \\ \hat{\sigma}_{\mathbf{n},h,F} \quad &\widehat{=} \quad x_{\sigma,F} := \left(x_{\sigma,F,1}, x_{\sigma,F,2}, \ldots, x_{\sigma,F,n_p}\right)^\top \in \mathbb{C}^{n_p}. \end{aligned}$$

The dimensions of the Schur complement are

$$S \in \mathbb{C}^{2n_p n_\mathcal{F} \times 2n_p n_\mathcal{F}}$$

and for the solution vector $x \in \mathbb{C}^{2n_p n_\mathcal{F}}$ holds the representation

$$x = \left(x_{u,F_1}, x_{u,F_2}, \ldots, x_{u,F_{n_\mathcal{F}}}, x_{\sigma,F_1}, x_{\sigma,F_2}, \ldots, x_{\sigma,F_{n_\mathcal{F}}}\right)^\top.$$

The first half of x contains the unknowns associated to the variable $\hat{u}_h$ and the second half the unknowns associated to $\hat{\sigma}_{h,\mathbf{n}}$. The restriction R_{u,F_i} of $\hat{u}_h$ onto the facet F_i, $i \in \{1, 2, \ldots, n_\mathcal{F}\}$

$$R_{u,F_i}(\hat{u}_h) = \hat{u}_{h,F_i},$$

has the matrix representation

$$R_{u,i} := \left(0^{n_p \times (i-1)n_p}, I^{n_p \times n_p}, 0^{n_p \times (2n_\mathcal{F} - i)n_p}\right) \in \mathbb{R}^{n_p \times 2n_p n_\mathcal{F}}.$$

Similarly to $\hat{u}_h$, the restriction R_{σ,F_i} of $\hat{\sigma}_{\mathbf{n},h}$ onto the facet F_i, $i \in \{1, 2, \ldots, n_\mathcal{F}\}$

$$R_{\sigma,F_i}(\hat{\sigma}_{\mathbf{n},h}) = \hat{\sigma}_{\mathbf{n},h,F_i},$$

has the matrix representation

$$R_{\sigma,i} := \left(0^{n_p \times (n_\mathcal{F}+i-1)n_p}, I^{n_p \times n_p}, 0^{n_p \times (n_\mathcal{F}-i)n_p}\right) \in \mathbb{R}^{n_p \times 2n_p n_\mathcal{F}}.$$

With this definition the restriction onto the unknowns associated to a facet $F_i \in \mathcal{F}$, $i \in \{1, 2, \ldots, n_\mathcal{F}\}$ is

$$R_i := R_{u,i} + R_{\sigma,i}$$

and the Block-Jacobi preconditioner takes the form

$$C_J^{-1} := \sum_{i=1}^{n_\mathcal{F}} R_i^\top S_i^{-1} R_i$$

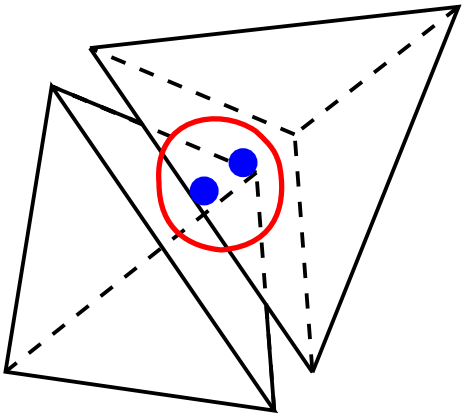
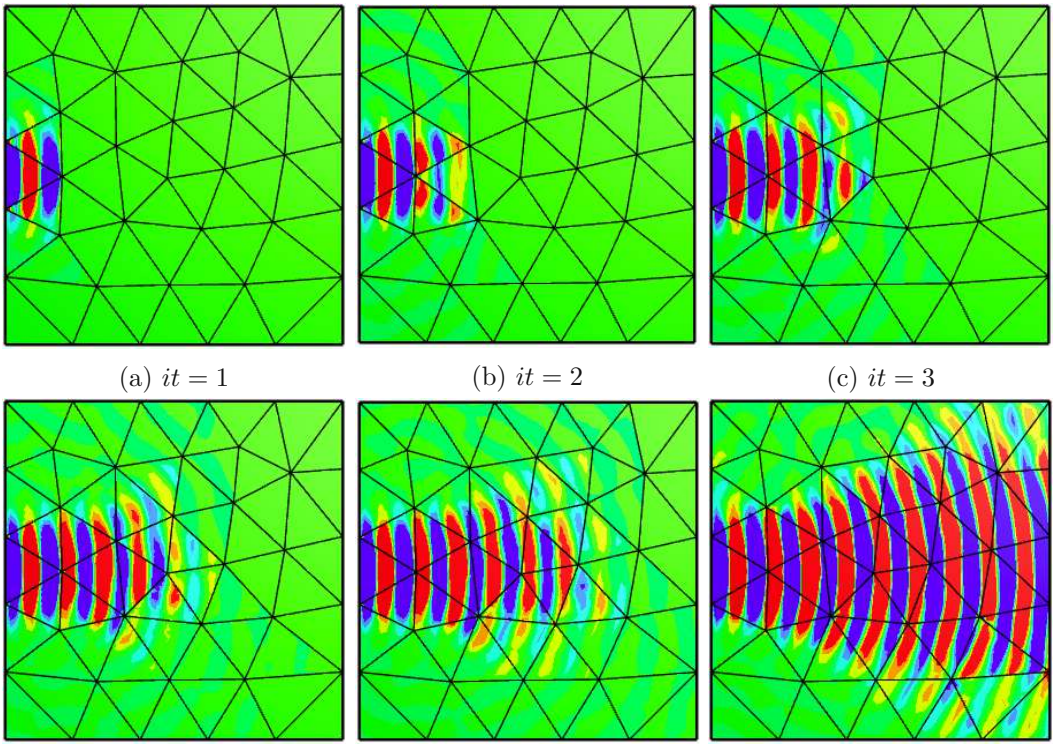


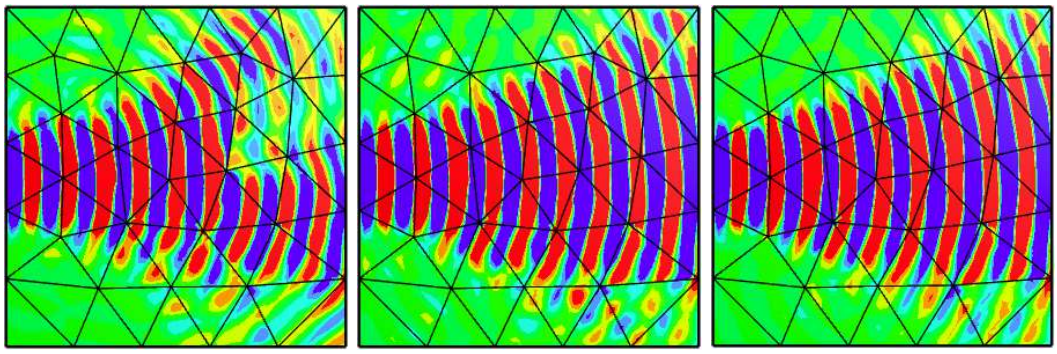
Figure 4.16.: One facet patch of unknowns on a 3D mesh for the Jacobi and Gauss-Seidel preconditioners. The facet unknowns are highlighted with blue dots, and the patch with a red circle.



(a) $it = 1$ (b) $it = 2$ (c) $it = 3$

(d) $it = 4$ (e) $it = 5$ (f) $it = 9$

Figure 4.17.: The behaviour of the GMRES method in combination with the Block-Jacobi preconditioner is shown.



(a) $it = 1$ (b) $it = 2$ (c) $it = 3$

Figure 4.18.: The behaviour of the GMRES method in combination with the Gauss-Seidel preconditioner is shown.

with the matrices

$$S_i := R_i S R_i^\top .$$

An example of a patch of facet unknowns can be seen in Figure 4.16. The facet patches are non-overlapping for 3D simulations.

The solutions after the first couple of iterations for this preconditioner applied to the above-introduced example can be seen in Figure 4.17. The solution is propagated through the domain, starting at the excitation boundary on the left towards the right side. This is also the expected behaviour because, in each iteration step, the excitation information can only be propagated from one element to neighbouring elements. In this case, this means it requires at least six iterations such that all elements receive some information about the propagating wave. At that stage, the method has not converged, and for this example, after nine iterations, the solution on the right side is not accurate. To achieve a relative error of 10^{-5} GMRES with restarts after 20 iterations requires 41 iterations. This example also gives a glimpse of the importance of the value of the stabilisation parameters α and β, because to be able to propagate plane waves, the sub-matrices must be discretisations of local Helmholtz problems with Robin boundary conditions.

4.4.2. Gauss-Seidel Preconditioner

The second preconditioner is the Block-Gauss-Seidel preconditioner based upon the same sub-blocks introduced with the Jacobi preconditioner. The major difference between these two is that the Gauss-Seidel algorithm is sequential, in the sense that the local inverses are applied multiplicatively one after another. How this behaviour looks like for the considered example can be seen in Figure 4.18. After just three iterations, the solution looks quite promising already. The Gauss-Seidel preconditioner requires less iterations to converge, namely just ten iterations for an accuracy of 10^{-5}. This preconditioner has a sweeping-like motion over the patches, and the order in which these patches are handled one after another is given by the ordering of the elements In one iteration, the wave is propagated

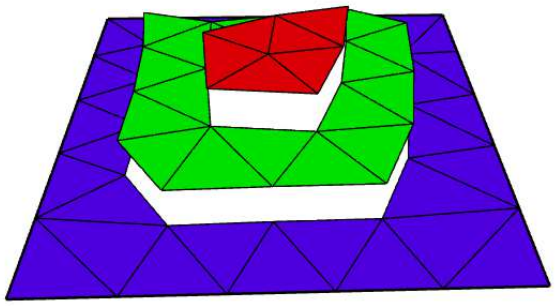
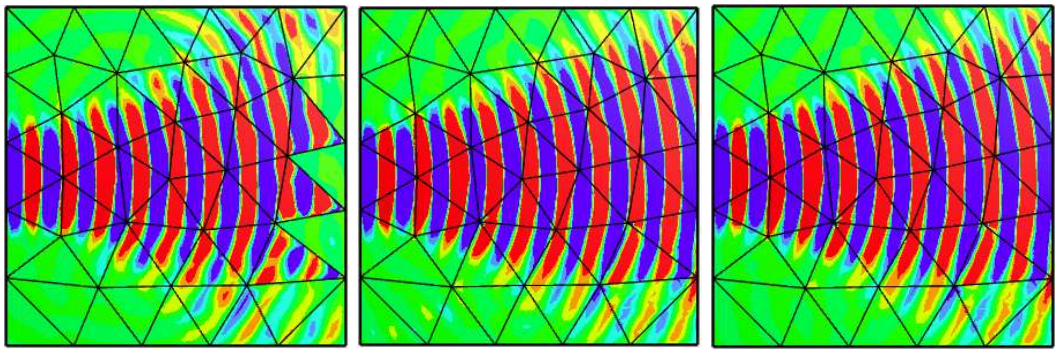


Figure 4.19.: The layers for the sweeping preconditioner on the domain of the introductory problem are shown.

along the elements in ascending order. It is noteworthy that for a 1D example with element numbering running from the left to the right, the Gauss-Seidel preconditioner performs a forward sweep through the domain, and if the preconditioner is additionally applied in the reverse direction so that the preconditioner is symmetric, then a forward-backwards sweep is performed. From the computational standpoint, this preconditioner has the disadvantage of being sequential, which is a hindrance for implementation on distributed server clusters. The positive aspect is the performance concerning the reduced iteration numbers compared to the Jacobi preconditioner.

4.4.3. Sweeping Preconditioner

The notion of a sweeping preconditioner for the Helmholtz equation has been focused by many authors, and multiple versions have been developed. Originally, Frédéric Nataf introduced the first idea of this type of pre-conditioning in his publication [Nat93]. The method has been the subject of ongoing research, for example by Björn Engquist and Lexing Ying in their publications [EY11a] and [EY11b]. In their approaches, they considered a 2D domain with a structured mesh. Then, they split the domain into non-overlapping vertical slices and proposed the idea of propagating the solution only on this smaller slice. As is natural, they faced the issue of ill-posed local problems representing Dirichlet problems. The answer was to introduce artificial transparent boundary conditions on the boundary of the layers, for example, perfectly matched layers (PMLs). This concept has been further studied by many other authors. The choice of layers for general problems, the choice of sweeping directions, forward vs forward-backwards sweeping, a mixture of horizontal and vertical sweeping or the introduction of additional diagonal sweeping has been explored, but they fall back to the same idea by Enquist and Ying: in that regard the sweeping preconditioner in this work falls exactly into the same category. The domain is split into non-overlapping layers, and then a multiplicative forward-backwards sweeping is carried out, with the difference that the HDG-formulation has the intrinsic property of local problems with Robin boundary conditions. The usage of PMLs requires the artificial introduction of absorbing layers, the HDG-formulation automatically incorporates these, which makes it easily applicable without further implementation effort. For the introductory example, the chosen layers are of onion shape and can be seen in Figure 4.19. The general idea is to combine



(a) $it = 1$ (b) $it = 2$ (c) $it = 3$

Figure 4.20.: The behaviour of the GMRES method in combination with the sweeping preconditioner is shown.

all boundaries with prescribed Robin boundary conditions and start layering originating from them. Then, each layer goes one element thickness deeper into the domain, finalising with the last layer being the core. Now, a forward-backwards sweep should transfer any prescribed boundary condition through the whole domain at least once, and the volume excitation is at least transferred in one direction. The behaviour of sweeping for the introductory example can be seen in Figure 4.20. After only two iterations, the solution seems already quite good, and the GMRES method converges after six iterations to a relative accuracy of 10^{-5}.

4.4.4. Non-overlapping Domain Decomposition Preconditioner

The fourth and last preconditioner considered in this work is a multiplicative non-overlapping domain decomposition preconditioner. The idea is the same as for the sweeping preconditioner, with the difference that subdomains with equal size regarding the number of associated unknowns are used instead of layers with a thickness of one element. The reason is that depending on the transparent boundary, the outermost and innermost layers of the sweeping preconditioner vary in size, so much such that calculating the required factorisation of the onto the outmost layer restricted system matrix is the bottleneck in the algorithm. To balance the computational cost, all subdomains are chosen of equal size, but with that, the question arises if the preconditioner is still performant. In Figure 4.21, the used subdomains for the preconditioner can be seen. The mesh is partitioned with PyMetis [KWH+22], a Python wrapper for the Metis graph paritioning software [KK97]. Each domain contains approximately the same number of unknowns (490, 504, 490) respectively, therefore, the factorisation effort of the subdomain system matrices are almost equal. In theory, all these factorisations can be done in parallel. The behaviour of the preconditioner in the first three iterations for the introductory example is shown in Figure 4.22. After one iteration, the borders of the three subdomains can be seen, and the solution does not have the same quality as after one step of the sweeping preconditioner above, but immediately after the second step, the solution already has quite reasonable accuracy. To finally converge towards a relative error of 10^{-5}, the preconditioner requires seven itera-

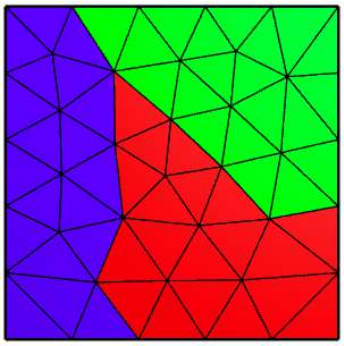
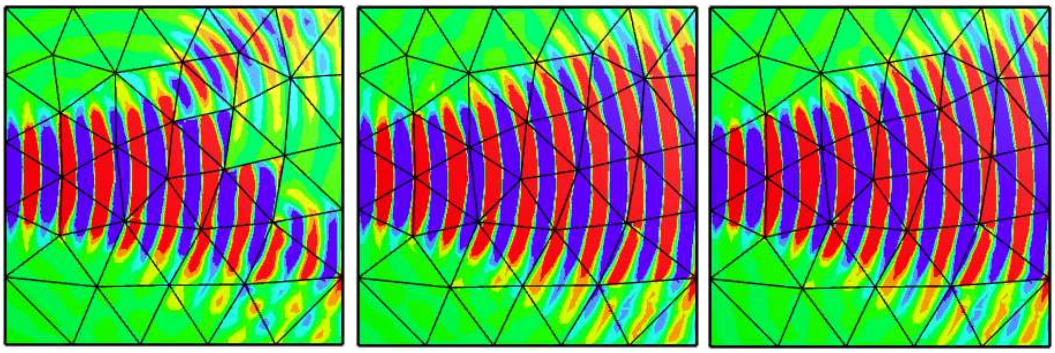


Figure 4.21.: The subdomains for the domain decomposition preconditioner on the domain of the introductory problem are shown.



(a) $it = 1$ (b) $it = 2$ (c) $it = 3$

Figure 4.22.: The behaviour of the GMRES method in combination with the domain decomposition preconditioner is shown.

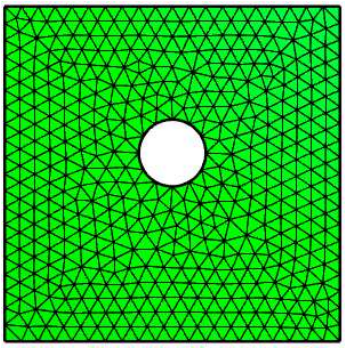


Figure 4.23.: The unit square domain with an off-centred circle as a scattering object can be seen.

tions. From the difference of two iterations between this preconditioner and the sweeping preconditioner, no conclusion about which one is more efficient can be drawn.

These are the four preconditioners considered in this work. The first conclusion, which can be directly drawn from the simple example, is that the actual iteration numbers are not a viable method to compare them. In the following experiments, the required iterations will be stated, but the emphasis is on the actual computational time. To be able to compare the different preconditioners simulations for the same benchmark are carried out on identical computer resources.

4.4.5. Stabilisation Parameters α and β

Throughout this work, the necessity of mesh size and polynomial degree independent stabilisation parameters $\alpha = \mathcal{O}(1)$ and $\beta = \mathcal{O}(1)$ has been claimed. In the following numerical experiment, this claim is put to the test. To this end, the unit square domain with a circle as a scattering object inside the domain is chosen, see Figure 4.23. The boundary condition on the circle is a homogenous Dirichlet boundary condition, and the outmost boundary is transparent with Robin boundary conditions and the excitations are set to

$$f(\mathbf{x}) := 0, \qquad g(\mathbf{x}) := \begin{cases} j\kappa e^{-10(y-1/2)^2} & \text{on the left boundary,} \\ 0 & \text{else,} \end{cases}$$

with a wave number $\kappa = 240$. The scattering circle is introduced to generate a solution that has reflections and multiple propagation directions throughout the domain. The problem is discretised with a polynomial degree of $p = 6$, and for this setting, the solution can be seen in Figure 4.24. The four previously introduced preconditioners are applied to this problem, but in the used HDG-formulation the stabilisation parameters α and β are varied. For the sweeping preconditioner, the domain is split into layers, which can be seen in Figure 4.25 and the subdomains for the domain decomposition preconditioner can be seen in Figure 4.26. An intuitive choice for the stabilisation with an ABC in mind might be $\alpha = 1$ and $\beta = 1$. Therefore, a parameter study has been carried out around this point.

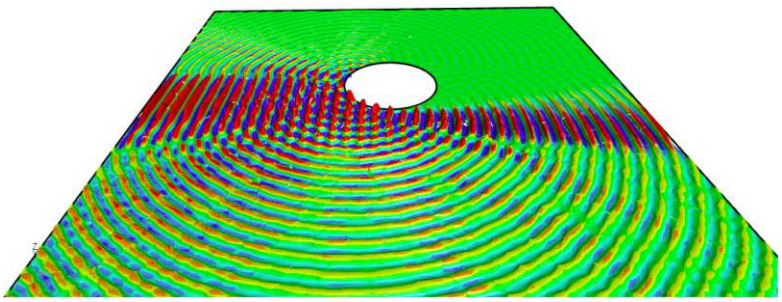


Figure 4.24.: The solution for the α-β test example can be seen. The view is tilted slightly to highlight the oscillating pressure.

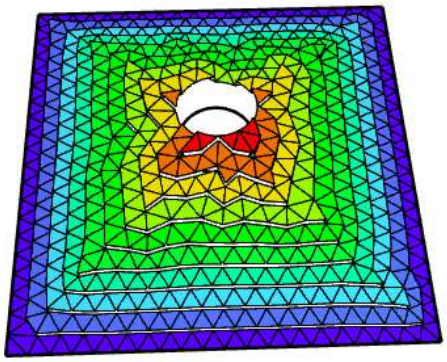


Figure 4.25.: The layers for the sweeping preconditioner on the domain of the α-β test problem are shown.

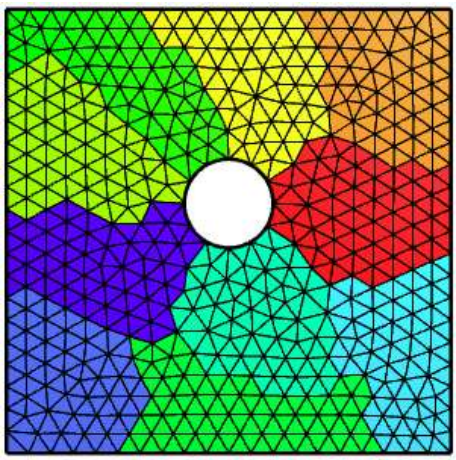


Figure 4.26.: The subdomains for the domain decomposition preconditioner on the domain of the α-β test problem are shown.

The stabilisation parameter α for the jumps of the pressure has been chosen in the interval

$$\alpha \in \left[\frac{1}{8}, \frac{15}{8}\right].$$

As a rule of thumb, the relation $\beta = \alpha^{-1}$ is used, which is equivalent to $\alpha\beta = 1$. A deviation from this relation is explored in the sense that β is chosen as

$$\beta := \frac{\delta}{\alpha} \qquad \text{with} \qquad \delta \in \left[\frac{1}{2}, \frac{3}{2}\right].$$

For multiple snapshots of α and $\delta = \alpha\beta$ the iteration numbers of all four methods have been evaluated and the results can be seen in Figure 4.27. The maximum number of iterations in the GMRES-algorithm with restart after 20 iterations was set to 300. Starting with the Gauss-Seidel iteration numbers, there can be seen a significant trench with respect to the parameter α. A deviation results in higher iteration numbers. So for this preconditioner, α should be chosen somewhere in the interval $[1/4, 3/4]$. In simulations, the choice $\alpha = 1/2$ and $\alpha\beta = 1$ usually performed very well. Next, a look at the iteration numbers for the Jacobi preconditioner reveals a similar dependency on α, but the choice for the Gauss-Seidel preconditioner also behaves very well in this case. The sweeping and the domain decomposition preconditioners seem to be more robust towards small deviations regarding the stabilisation parameters, but the Gauss-Seidel choice is also very good for these cases. The reasons above lead to the choice

$$\alpha := \frac{1}{2}, \qquad \beta := 2,$$

for the stabilisation parameters of the HDG-formulation, and when not further specified, those have been used for the simulations in this work.

4.4.6. Computational Costs

This subsection is dedicated to comparing the computational costs of the four preconditioners. As has already been stated, the number of iterations is not a sufficient indicator for a fast solver strategy; therefore also, the wall time and CPU time are tracked. The test example has the same setup as in the previous subsection, the unit square with a circle scatterer in the interior, see Figure 4.23. To be able to explore the behaviour of the preconditioners, multiple wavelengths have been considered for the underlying Helmholtz equation, more specifically

$$\kappa \in \{30, 60, 120, 240, 480, 960\}.$$

With different wave numbers also, the mesh size or the polynomial degree needs to be adapted. In this case, the polynomial degree was fixed for all wave numbers to $p = 6$, and the mesh size changed by the law

$$h = \frac{\kappa}{48}.$$

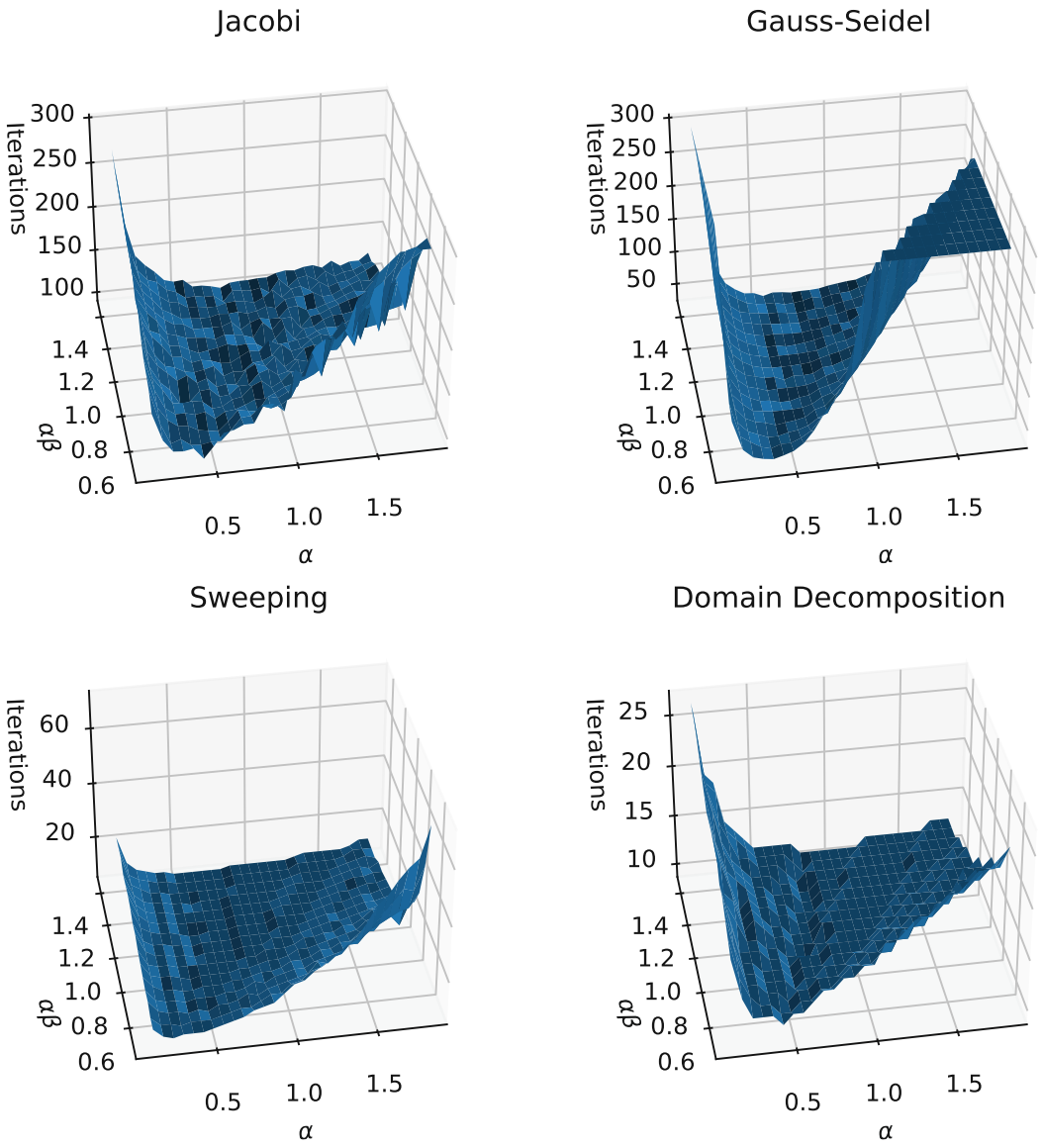
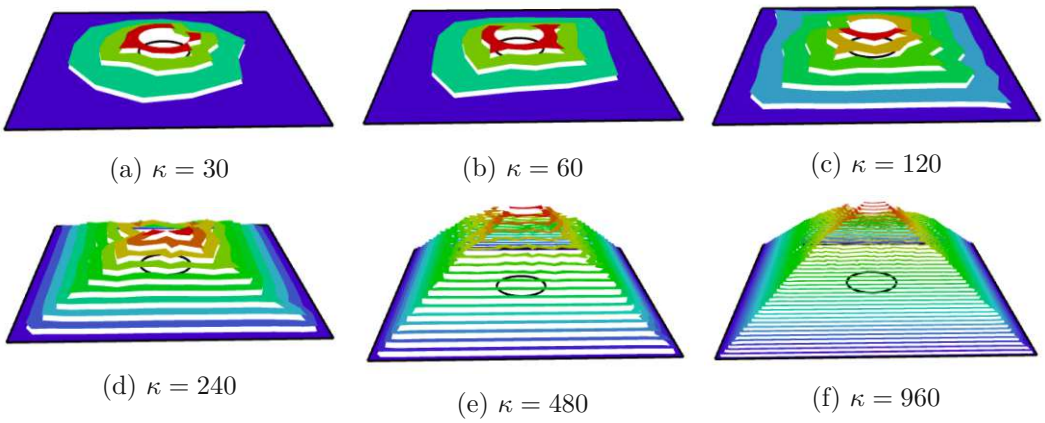


Figure 4.27.: The iteration numbers for the four considered preconditioners for different choices of the stabilisation parameters α and β can be seen. The maximum number of iterations was set to 300, which is an upper boundary in the graphs. On the top left are the iterations for Jacobi, on the top right Gauss-Seidel, on the bottom left, sweeping and on the bottom right, domain decomposition can be seen.



(a) $\kappa = 30$ (b) $\kappa = 60$ (c) $\kappa = 120$

(d) $\kappa = 240$ (e) $\kappa = 480$ (f) $\kappa = 960$

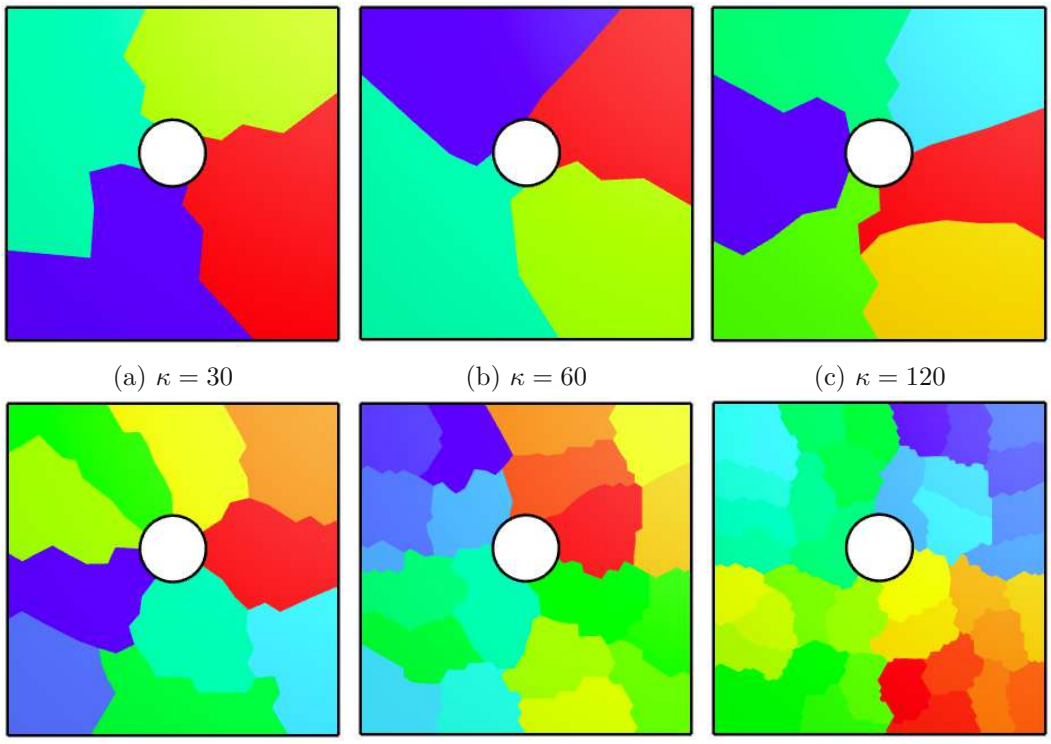
Figure 4.28.: The sweeping layers for the computational costs example with respect to the chosen wave numbers are shown.

The boundary condition on the circle was a homogenous Dirichlet boundary condition, and the outermost boundary is transparent with Robin boundary conditions and the excitations were set to

$$f(\mathbf{x}) := 0, \qquad g(\mathbf{x}) := \begin{cases} j\kappa e^{-10(y-1/2)^2} & \text{on the left boundary,} \\ 0 & \text{else.} \end{cases}$$

For each wave number/mesh size, layers for the sweeping preconditioner and subdomains for the domain decomposition preconditioner have been generated. The layers can be seen in Figure 4.28. The for the domain decomposition preconditioner used subdomains can be seen in Figure 4.29. For the lowest two wave numbers, the mesh is the same because the actual mesh size is dictated by the mesh around the circle. Starting with the third wave number, the wavelength determines the necessary mesh size in the domain and more layers/subdomains are required. To give a full view of the problem, the pressure of the solution with respect to the different wave numbers can be seen in Figure 4.30. The hardware architecture for the simulations was an AMD Ryzen 5 2600 Six-Core Processor with 3.40 GHz with underlying 16 GB virtual memory and a Windows 10 operating system. The simulations on this architecture for the described problems lead to the wall times, CPU times and iterations seen in Figure 4.31, Figure 4.32 and Figure 4.33 respectively.

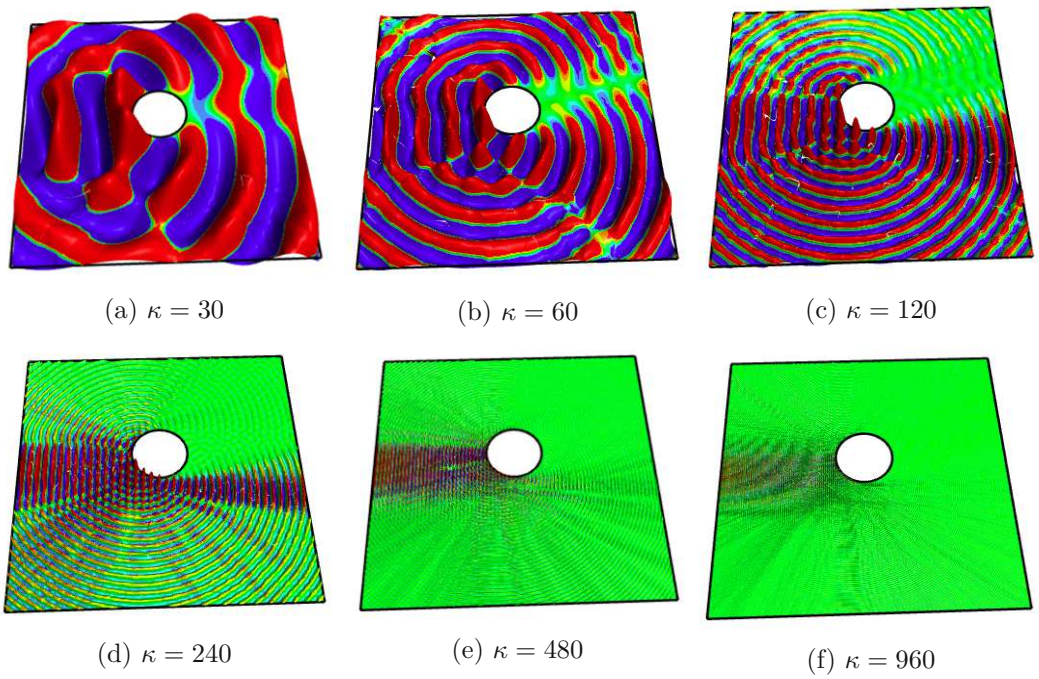
The wall time and CPU time behave similarly for the respective preconditioners, which is reasonable on a shared memory system and a highly parallelised software like Netgen/NGSolve. The simulations for the lowest two wave numbers almost overlap in the data points due to the same number of unknowns. Afterwards, the mesh size was halved for each subsequent wave number, which leads to the quick increase in the number of hybrid facet unknowns n_{DoF_s}. Comparing the sweeping and the domain decomposition preconditioner, the simulation times are almost equal for the most part, but for the highest frequency, the domain decomposition preconditioner takes a small lead. The Jacobi preconditioner requires more wall time to converge closely followed by the Gauss-Seidel preconditioner. Regarding the CPU time Jacobi and Gauss-Seidel are quick for the smallest two wave



(a) $\kappa = 30$ (b) $\kappa = 60$ (c) $\kappa = 120$

(d) $\kappa = 240$ (e) $\kappa = 480$ (f) $\kappa = 960$

Figure 4.29.: The subdomains for the computational costs example with respect to the chosen wave numbers are shown.



(a) $\kappa = 30$ (b) $\kappa = 60$ (c) $\kappa = 120$

(d) $\kappa = 240$ (e) $\kappa = 480$ (f) $\kappa = 960$

Figure 4.30.: The pressure of the solutions for the computational costs example with respect to the chosen wave numbers are shown.

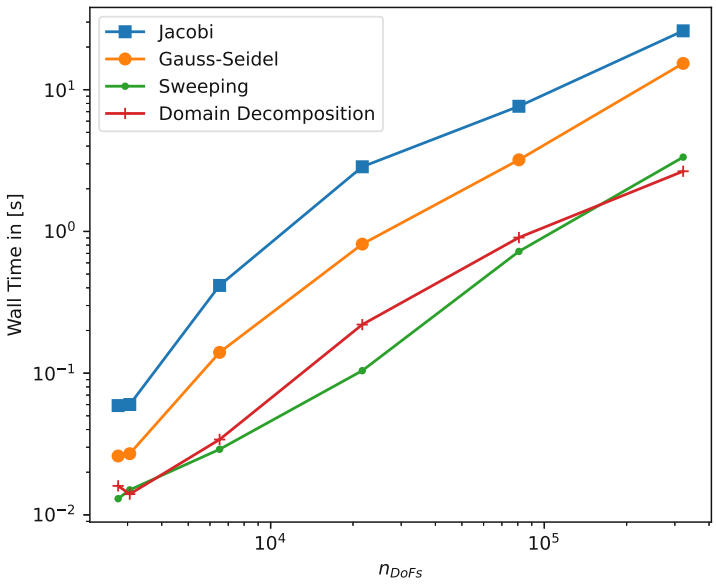


Figure 4.31.: The wall time of the four preconditioners with respect to the number of hybrid facet unknowns $n_{\mathrm{DoF}s}$ is shown.

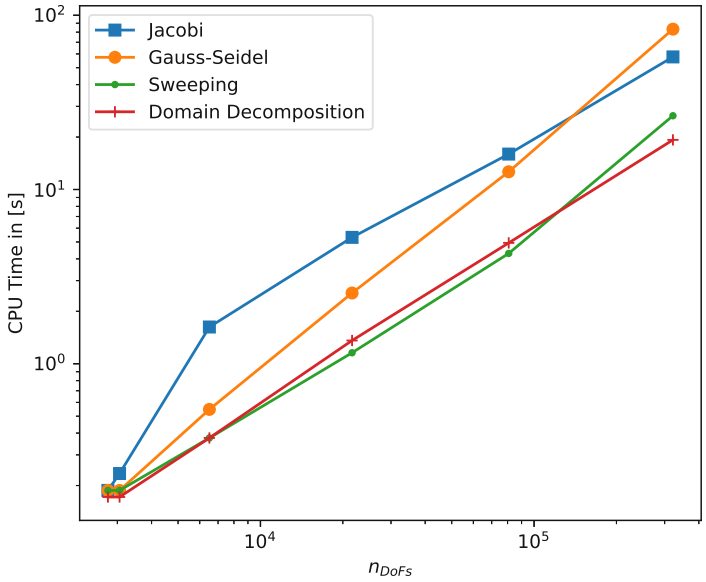


Figure 4.32.: The CPU time of the four preconditioners with respect to the number of hybrid facet unknowns n_{DoF_s} is shown.

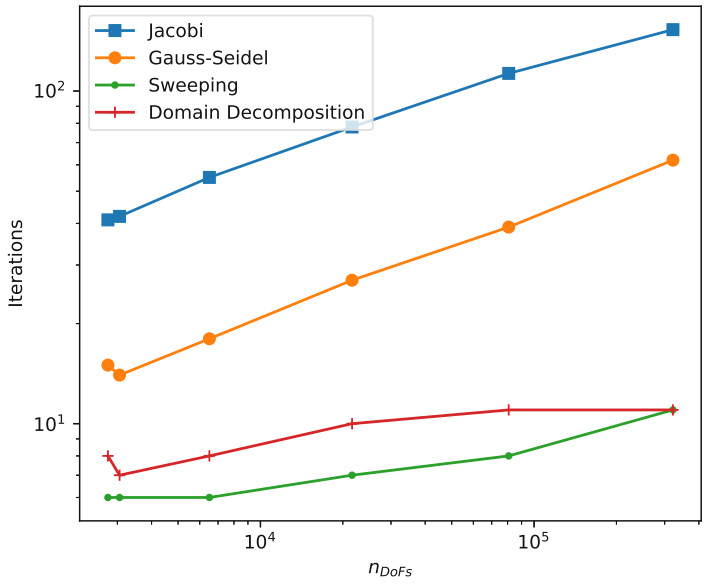


Figure 4.33.: The number of iterations of the four preconditioners with respect to the number of hybrid facet unknowns n_{DoF_s} is shown.

numbers but afterwards require more time than sweeping and domain decomposition. Additionally, for the Gauss-Seidel preconditioner it seems that the CPU time increases faster than for the other three.

Having a look at the required iteration numbers, Jacobi requires approximately twice as many iterations than Gauss-Seidel. For the sweeping and domain decomposition preconditioners, it may look as if the iterations are almost constant, which is not the case, but the iteration numbers increase with the number of subdomains.

Apart from iteration numbers and simulation times, memory is the third crucial resource. In the following, the costs for the preconditioners are roughly estimated. Jacobi has several sub-blocks that are equal to the number of elements n_E. Each block represents for a fixed polynomial degree $6(p+1)$ unknowns, and assuming that each local inverse requires a full $6(p+1) \times 6(p+1)$ matrix, this leads to

$$n_J := 36(p+1)^2 n_E$$

non-zero entries which need to be stored. Gauss-Seidel requires the same memory storage

$$n_{GS} := n_J.$$

For the domain decomposition preconditioner, this is more complicated to estimate. Splitting into equally sized subdomains gives approximately n_E/n_{dom} elements per subdomain. This leads to approximately $3(p+1)n_E/n_{dom}$ unknowns per subdomain. Which is a slight underestimation due to missed unknowns on the domain boundary. A sparse Cholesky factorisation requires around $27(p+1)^2 n_E/n_{dom} \log_2(n_E/n_{dom})$ non-zero entries per subdomain. Summing over all subdomains gives

$$n_{DD} := 27(p+1)^2 n_E \log_2(n_E/n_{dom})$$

non-zero entries. Estimating the sweeping preconditioner is the hardest because, for general problems, the largest layer size heavily depends on the computational domain, but for the sake of this simple argument, let us assume a square domain with a structured mesh. The number of layers is $n_{layers} = n_{edge}/2$, where n_{edge} is the number of element-edges along an edge of the square. The outer boundary consists of $16(n_{edge} - 1)$ facets. The next layer then has $16(n_{edge} - 3)$ facets, ending with the innermost layer which consists either of 5 or 16 facets. In a structured mesh there holds $n_E = 2n_{edge}^2$. To each facet $2(p+1)$ unknowns are associated, and when considering the memory requirements of a sparse Cholesky factorisation the over all consumption is given by summing over all layers:

$$n_{SW} := 96(p+1)^2 \sum_{i=1}^{\sqrt{n_E/2}} (i-1)\log_2(i-1) \approx 48(p+1)^2 n_E \log_2 n_E.$$

Jacobi and Gauss-Seidel are directly proportional to the number of elements. The memory consumptions of sweeping and domain decomposition are similar, with the difference that the logarithmic part for domain decomposition scales inverse proportional with respect to the number of subdomains. It is noteworthy that if the number of subdomains in the domain decomposition is chosen of the same magnitude as the number of elements in the

mesh, then the preconditioner is similar to the Gauss-Seidel preconditioner. In the end, if the memory is a bottleneck, then the Jacobi preconditioner can be applied to larger problems. The author would like to remind the reader that the numbers and results above are only derived from a single example and, therefore, can not be generalised. It should only give an intuition of the possible capabilities of the presented preconditioner versions.

4.5. Complementary Numerical Examples

This subsection highlights further capabilities of the proposed solvers through interesting, larger numerical examples, which can also be seen in [LS23].

4.5.1. Heterogeneous Materials

The analysis in this work only covers the case of constant material parameters, but the method can also cope with heterogeneous materials. In this numerical experiment, the following heterogeneous Helmholtz problem has been considered.

Definition 56 (Mixed Heterogeneous Helmholtz Problem)**.** *For given $\kappa > 0$ and $g \in L^2(\partial\Omega)$, let (σ, u) be the solution of the BVP*

$$\begin{aligned} j\kappa\sigma - \nabla u &= 0 && \text{in } \Omega, \\ -\operatorname{div}(\sigma) + j\kappa c u &= 0 && \text{in } \Omega, \\ \sigma \cdot \mathbf{n} + u &= g && \text{on } \partial\Omega, \end{aligned}$$

where $c(\mathbf{x}) > 0$ is a given positive, bounded, varying material coefficient.

The following 2D-examples have similar geometry and material coefficients as [GGS20, Experiment 6.5]. The domain consists of the square $\Omega = [-1, 1]^2$ with a circle of radius $r = 1/2$ as a penetrable obstacle centred in the middle. The wave number has been chosen as $\kappa = 100$, and the excitation was

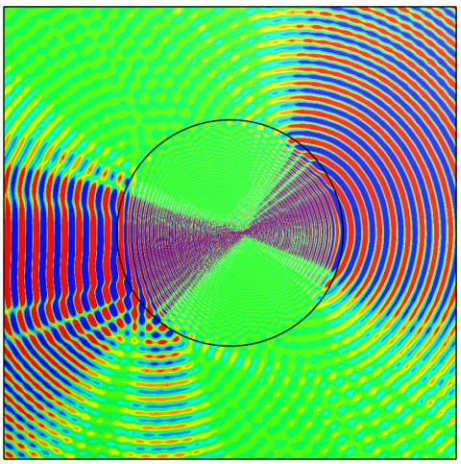
$$g(\mathbf{x}) = -10j\kappa e^{-20\left(y+\frac{1}{10}\right)^2}$$

on the left boundary. On the other outer boundaries, homogenous Robin boundary conditions were applied. The excitation is slightly offset from the axis of symmetry and represents an inflowing Gaussian peak. For the discrete FE spaces, a polynomial degree of $p = 4$ has been used, and the maximal mesh size was chosen as $h = 2\pi/8\kappa$. Two different material profiles were considered for the simulations, specifically

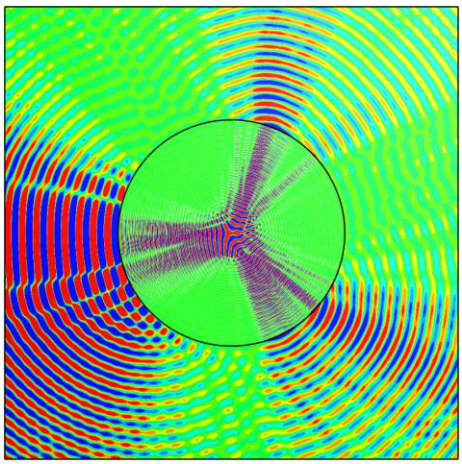
$$c_1(\mathbf{x}) := \begin{cases} 2\sqrt{x^2+y^2}c_{min} + \left(1 - 2\sqrt{x^2+y^2}\right) c_{max}, & \|\mathbf{x}\|_2 < \frac{1}{2}, \\ 1, & \text{otherwise}, \end{cases}$$

and

$$c_2(\mathbf{x}) := \begin{cases} \left(1 - 2\sqrt{x^2+y^2}\right) c_{min} + 2\sqrt{x^2+y^2}c_{max}, & \|\mathbf{x}\|_2 < \frac{1}{2}, \\ 1, & \text{otherwise}, \end{cases}$$



(a) Result for linearly decreasing material c_1 with respect to the radius inside the circle



(b) Result for linearly increasing material c_2 with respect to the radius inside the circle

Figure 4.34.: The real part $\Re(u_h)$ off the pressure is shown for simulations with the heterogeneous material coefficients c_1 and c_2. A Gaussian peak has been applied on the left boundary as an excitation.

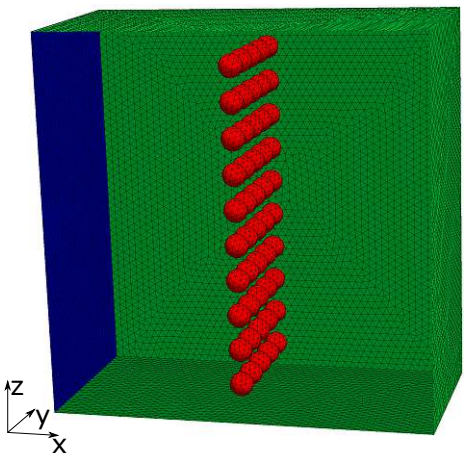
Table 4.1.: Number DoFs for the 2D simulation with heterogeneous materials

	σ_h	$\hat{\sigma}_{\mathbf{n},h}$	u_h	$\hat{u}_h$
Number DoFs	4 279 920	1 072 530	2 139 960	1 072 530

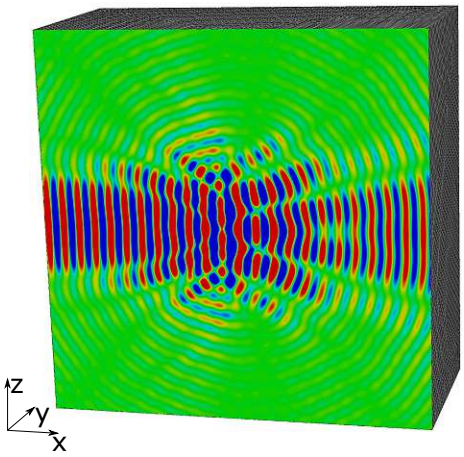
with the minimum $c_{min} = 1/50$ and the maximum $c_{max} = 50$. The first material c_1 is constant outside the circle and decreases linearly with respect to the radius. On the contrary, the second material c_2 increases linearly. In Figure 4.34a and 4.34b, the real part of the pressure can be seen for the simulations with material coefficients c_1 and c_2. Both simulations have been carried out on 8 cores, required approximately 14 GB of memory and the number of DoFs can be found in Table 4.1. After static condensation, a system of linear equations with the combined size of $\hat{\sigma}_{\mathbf{n},h}$ and $\hat{u}_h$ was solved with the Block-Jacobi preconditioner introduced in Section 4.4.1. In Table 4.2, the number of iterations and the computation times for both simulations are shown. The wall time reflects the duration of the iterative solving, and the processor time is the sum of the computation times of all cores combined.

Table 4.2.: Number of iterations and computation times for heterogeneous materials

	iterations	wall time in s	processor time in s	processor time per core in s
c_1	5 833	2 363	15 712	1 964
c_2	4 024	1 634	10 867	1 359



(a) The spherical scatterers can be seen in red, and the left excitation boundary is in blue.



(b) The real part $\Re(u_h)$ of the pressure is drawn on the plane of symmetry.

Figure 4.35.: A cut view along the x, z-plane of symmetry is shown for the 3D-example with spherical scatterers.

Table 4.3.: Number DoFs for the 3D scattering experiment

	σ_h	$\hat{\sigma}_{\mathbf{n},h}$	u_h	$\hat{u}_h$
Number DoFs	56 994 735	16 616 295	18 998 245	16 616 295

4.5.2. Scattering on Spheres

A 3D example with 100 spheres as scatterers treated as homogeneous Dirichlet boundary conditions has been considered. The domain is comprised of the cube $\Omega = [0,1]^3$ and on the plane of symmetry perpendicular to the x-axis, an array of scatterers is positioned. They are aligned on a 10×10 grid with an equidistant spacing of $\delta = 1/11$ in between midpoints. The radius $r = \delta/3$ of each sphere is identical. A cut view of the geometry and the solution can be seen in Figure 4.35. A wave number of $\kappa = 150$ was considered, and the material coefficient was constant $c = 1$. The excitation was the Gaussian peak

$$g(\mathbf{x}) = -10j\kappa e^{-100\left(\left(y-\frac{1}{2}\right)^2+\left(z-\frac{1}{2}\right)^2\right)}$$

on the left boundary. For the discrete space, a polynomial degree of $p = 4$ has been used, and the maximal mesh size was chosen as $h = 2\pi/2\kappa$. The simulation has been carried out on 16 cores, required approximately 153 GB of memory and the number of DoFs can be seen in Table 4.3. The solver with the Block-Jacobi preconditioner in Section 4.4.1 stopped after 417 iterations with a wall time of 8204 s.

A. Analysis of the HDG-Formulation with Brezzi-Douglas-Marini Spaces

In Section 3.3, the HDG-formulation was introduced and discretised with the use of Raviart-Thomas spaces. The Brezzi-Douglas-Marini spaces are also a viable option as a discrete subspace of $H(\mathrm{div})$. For the most part, all results hold in the same manner, and in this chapter, the differences will be highlighted and analysed.

The formulation itself does not change. The flux is approximated with the space

$$\mathcal{BDM}^p_{pw}(\mathcal{T}) := \Pi_{T\in\mathcal{T}} \subset H_{pw}(\mathrm{div}, \mathcal{T}) \cap H^s_{pw}(\mathcal{T}).$$

This discrete space is smaller than $\mathcal{RT}^p_{pw}(\mathcal{T})$. The space for the pressure and the facet variables stays the same. In this setting, all spaces are approximated with polynomials of exactly the degree p. The stability estimates and error bounds hold with the new space only the best approximation changes to

$$\eta(\sigma) := \inf_{\tau_h \in \mathcal{BDM}^p_{pw}(\mathcal{T})} \left(\|\sigma - \tau_h\|^2_{L^2(\Omega)} + h^2 \|\nabla(\sigma - \tau_h)\|^2_{H^1(\Omega)} \right).$$

The space also leads to the following changed definition of the L^2-projection of σ, see Definition 37,

$$(\Pi\sigma, \tau_h)_T = (\sigma, \tau_h)_T \qquad \forall \tau_h \in \mathcal{BDM}^p_{pw}(\mathcal{T}).$$

The result in Lemma 39 still holds because the projection orthogonalities are still strong enough. All other steps in the analysis for the absolute stability and the pre-asymptotic error for the pressure are not impacted.

The really interesting change appears in the pre-asymptotic error estimate for the flux. The in Lemma 43 defined interpolation P needs to be altered. One would expect that, because the space of the flux gets smaller also, the projection properties for this part need to be weaker, but that is not the case. Rather, the projection of the pressure is weakened in the following way.

Lemma A.1 (Interpolation P for $\mathcal{BDM}^p_{pw}(\mathcal{T})$)**.** *The interpolation P is defined by*

$$\begin{aligned}
(P_\sigma, \tau_h)_T &= (\sigma, \tau_h)_T && \tau_h \in [\mathcal{P}^{p-1}(\mathcal{T})]^d,\\
(P_u, v_h)_T &= (u, v_h)_T && v_h \in \mathcal{P}^{p-1}(\mathcal{T}),\\
\left(\frac{1}{\alpha} + \beta\right)((\sigma - P_\sigma)\cdot\mathbf{n}, \mu_h)_{\partial T} &= \left(u - P_u + \frac{\alpha\beta}{2}[u - P_u], \mu_h\right)_{\partial T} && \mu_h \in \mathcal{P}^p(\mathcal{F}_I),\\
(\sigma\cdot\mathbf{n} - P_\sigma\cdot\mathbf{n}, \mu_h)_{F_O} &= \alpha(u - P_u, \mu_h)_{F_O} && \mu_h \in \mathcal{P}^p(\mathcal{F}_O),\\
P_{\hat{\sigma}_\mathbf{n}}(\sigma, u) &:= \{P_\sigma(\sigma, u)\}_\mathbf{n} && \text{on } F_I,\\
P_{\hat{\sigma}_\mathbf{n}}(\sigma, u) &:= P_\sigma(\sigma, u)\cdot\mathbf{n} && \text{on } F_O,\\
P_{\hat{u}}(\sigma, u) &:= \{P_u(\sigma, u)\} - \frac{1}{2\alpha}[P_\sigma(\sigma, u)]_\mathbf{n} && \text{on } F_I,\\
P_{\hat{u}}(u) &:= \Pi_F u && \text{on } F_O.
\end{aligned}$$

First, it is remarkable that this is still a square system of linear equations and the similarity with the projection established in [CGS10, Say13] used in [GM11], secondly, the following very similar result to Lemma 43 holds

Lemma A.2. *With the interpolation P, there holds*

$$\kappa\sum_{T\in\mathcal{T}}\|e^P_\sigma\|^2_T = \kappa\sum_{T\in\mathcal{T}}\|e^P_u\|^2_T - \kappa\Re(\sigma - P_\sigma, e^P_\sigma)_T + \kappa\Re(u - P_u, e^P_u)_T.$$

Proof. The proof is based on the form of

$$B_{II}(\sigma - P_\sigma, \sigma\cdot\mathbf{n} - P_{\hat{\sigma}_\mathbf{n}}, u - P_u, u - P_{\hat{u}}; e^P_\sigma, e^P_{\hat{\sigma}}, e^P_u, e^P_{\hat{u}}).$$

All boundary terms vanish because the interpolation properties are the same as for the original interpolation. For the volume terms, there holds

$$(u - P_u, \operatorname{div} e^P_\sigma) = 0,$$

because now $\operatorname{div}(e^P_\sigma)|_T \in \mathcal{P}^{p-1}(T)$ holds and for the other mixed term

$$(\sigma - P_\sigma, \nabla e^P_u)_T = 0,$$

because $\nabla e^P_u \in [\mathcal{P}^{p-1}(T)]^d$. There only remains

$$j\kappa(\sigma - P_\sigma, e^P_\sigma)_T - j\kappa(u - P_u, e^P_u)_T.$$

□

The main differences are the two remaining terms which can be adressed by applying Young inequalities leading to:

Corollary A.3. *With the interpolation P, there holds*

$$\kappa\sum_{T\in\mathcal{T}}\|e^P_\sigma\|^2_T = 3\kappa\sum_{T\in\mathcal{T}}\|e^P_u\|^2_T + \kappa\|\sigma - P_\sigma\|^2_T + \kappa\|u - P_u\|^2_T.$$

As long as P_σ and P_u are quasi-optimal interpolations and if for the projected error e_u^P also quasi-optimality holds then it also holds for e_σ^P. This is also the only remaining aspect because all other results fall together afterwards.

Remark A.4. *The static condensation holds for the $\mathcal{BDM}$-space in the same way as for the $\mathcal{RT}$-space.*

A.1. The interpolation P for $\mathcal{BDM}$

The existence and uniqueness, as well as the quasi-optimality of the interpolation P in the case of $\mathcal{BDM}$ spaces, will be shown in this section. A similar projection was established in [CGS10, Say13]. The first important aspect is that interpolation corresponds to a square system of linear equations, then in the following established approximation estimates directly give uniqueness and, therefore, also the existence of the interpolation.

The proof mainly follows the ideas in [CGS10, Say13] and is adapted. The following lemma is directly transferred from [CGS10, Lemma A.1].

Lemma A.5. *Let F be any face of an element T. The trace map*

$$\gamma_F : \mathcal{P}^p_\perp(T) \to \mathcal{P}^p(F) \qquad \text{defined by} \qquad \gamma_F(p) = p|_F$$

is a bijection. Moreover,

$$\|p\|_T \leq Ch^{1/2}\|p\|_F \qquad \forall p \in \mathcal{P}^p_\perp(T).$$

Similarly, to the case of $\mathcal{RT}$-spaces, the interpolation for the pressure P_u can be decoupled from the interpolation of the flux P_σ. The following result is an adaptation of [CGS10, Proposition A.1].

Proposition A.6. *On each element $T \in \mathcal{T}$, the component P_u satisfies*

$$(P_u, v_h)_T = (u, v_h)_T \qquad v_h \in \mathcal{P}^{p-1}(T), \tag{A.1}$$

$$\left(\frac{1}{\alpha} + \beta\right)(\operatorname{div}\sigma, w_h)_T = \left(u - P_u + \frac{\alpha\beta}{2}[u - P_u], w_h\right)_{\partial T \setminus \partial\Omega} + (1 + \alpha\beta)(u - P_u, w_h)_{\partial T \cap \partial\Omega} \qquad w_h \in \mathcal{P}^p_\perp(T). \tag{A.2}$$

Proof. The statement directly follows from

$$\begin{aligned}((\sigma - P_\sigma)\cdot\mathbf{n}, w)_{\partial T} &= (\operatorname{div}(\sigma - P_\sigma), w)_T + (\sigma - P_\sigma, \nabla w)_T \\ &= (\operatorname{div}(\sigma - P_\sigma), w)_T \\ &= (\operatorname{div}\sigma, w)_T\end{aligned}$$

which holds because $\nabla w|_T \in [\mathcal{P}^{p-1}(T)]^d$ and therefore $(\sigma - P_\sigma, \nabla w)_T = 0$, as well as $w|_T \in \mathcal{P}^p_\perp(T)$, $\operatorname{div}(P_\sigma)|_T \in \mathcal{P}^{p-1}(T)$ and therefore $(\operatorname{div} P_\sigma, w)_T = 0$. □

As seen below, the interpolation P_u is uniquely defined by the equations above. With these results, the approximation property can be proven in the same manner as in [CGS10, Proposition A.2].

Proposition A.7. *Assuming the H^s-regularity of the Helmholtz equation, then*

$$\sum_{T\in\mathcal{T}} \|u-P_u\|_T \leq C\left(1+\sqrt{2(1+\alpha\beta)}\right) h^{\min\{s,p+1\}}|u|_{H^{\min\{s,p+1\}}(\Omega)} + 2C\left(\frac{1}{\alpha}+\beta\right) h^{\min\{s-1,p+1\}}|\operatorname{div}\sigma|_{H^{\min\{s-2,p\}}(\Omega)}.$$

Proof. For the analysis, the L^2-projection will be used, and the estimate will be split into two parts by

$$\|u-P_u\|_T \leq \|u-\Pi u\|_T + \|\delta_u\|_T$$

with $\delta_u := P_u - \Pi u$. For the first part, the estimates are trivial because of the approximation properties of the L^2-projection. Due to (A.1) there holds $\delta_u \in \mathcal{P}^p_\perp(\mathcal{T})$ and due to (A.2)

$$\begin{aligned}\left(\delta_u + \frac{\alpha\beta}{2}[\delta_u], w_h\right)_{\partial T\setminus\partial\Omega} + (1+\alpha\beta)(\delta_u, w_h)_{\partial T\cap\partial\Omega} &= -\left(\frac{1}{\alpha}+\beta\right)(\operatorname{div}\sigma, w_h)_T \\ &+ \left(u-P_u+\frac{\alpha\beta}{2}[u-P_u], w_h\right)_{\partial T\setminus\partial\Omega} \\ &+ (1+\alpha\beta)(u-P_u, w_h)_{\partial T\cap\partial\Omega}\end{aligned}$$

for each element. Next, $w_h = \delta_u$ will be chosen for each element. In the work [CGS10], they were able to analyse and estimates the projection on each element separately. This is not possible for the projection in this section because, due to the jumps, it has a non-local behaviour. The idea is to choose $w_h = \delta_u$ on each element. If terms from adjacent elements are considered, then it can be seen that they combine the following way

$$\begin{aligned}([\delta_{u_+}], \delta_{u_+})_{F_I} + ([\delta_{u_-}], \delta u_-)_{F_I} &= (\delta_{u_+} - \delta_{u_-}, \delta_{u_+})_{F_I} + (\delta_{u_-} - \delta_{u_+}, \delta u_-)_{F_I} \\ &= (\delta_{u_+} - \delta_{u_-}, \delta_{u_+} - \delta_{u_-})_{F_I} = \|[\delta_u]\|^2_{F_I}.\end{aligned}$$

Therefore, adjacent terms form a positive norm of the jump again, which is a favourable property leading to

$$\begin{aligned}&\sum_{T\in\mathcal{T}} \|\delta_u\|^2_{\partial T\setminus\partial\Omega} + (1+\alpha\beta)\|\delta_u\|^2_{\partial T\cap\partial\Omega} + \sum_{F_I\in\mathcal{F}_I}\frac{\alpha\beta}{2}\|[\delta_u]\|^2_{F_I} \\ &= \sum_{T\in\mathcal{T}} (u-\Pi u, \delta_u)_{\partial T\setminus\partial\Omega} + (1+\alpha\beta)(u-\Pi u, \delta_u)_{\partial T\cap\partial\Omega} \\ &\quad - \left(\frac{1}{\alpha}+\beta\right)(\operatorname{div}\sigma, \delta_u)_T + \sum_{F_I\in\mathcal{F}_I}\frac{\alpha\beta}{2}([u-\Pi u], [\delta_u])_{F_I}.\end{aligned}$$

The following simple estimates bound the right hand side terms:

$$\begin{aligned}2(u-\Pi u, \delta_u)_{\partial T\setminus\partial\Omega} &\leq \|u-\Pi u\|^2_{\partial T} + \|\delta_u\|^2_{\partial T}, \\ 2([u-\Pi u], [\delta_u])_{F_I} &\leq \|[u-\Pi u]\|^2_{F_I} + \|[\delta_u]\|^2_{F_I}.\end{aligned}$$

The divergence term can not be estimated by Cauchy-Schwarz and Young inequalities. The orthogonality $\delta_u \in \mathcal{P}^p_\perp$ is used to insert $\Pi^{p-1}(\operatorname{div}\sigma)$ the L^2-projection into the discontinuous polynomial space of order $p-1$, giving

$$-(\operatorname{div}\sigma,\delta_u)_T = -(\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma),\delta_u)_T \leq \|\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma)\|_T\|\delta_u\|_T.$$

Applying Lemma A.5 to δ_u leads to

$$\begin{aligned}-2(\operatorname{div}\sigma,\delta_u)_T &\leq 2Ch^{1/2}\|\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma)\|_T\|\delta_u\|_{\partial T}\\ &\leq 2C^2h\|\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma)\|_T^2+\frac{1}{2}\|\delta_u\|_{\partial T}.\end{aligned}$$

Inserting all these estimates and separating the jump into its elements gives

$$\begin{aligned}&\sum_{T\in\mathcal{T}}\|\delta_u\|^2_{\partial T\setminus\partial\Omega}+(1+\alpha\beta)\|\delta_u\|^2_{\partial T\cap\partial\Omega}+\sum_{F_I\in\mathcal{F}_I}\alpha\beta\|[\delta_u]\|^2_{F_I}\\ &\leq\sum_{T\in\mathcal{T}}2(1+\alpha\beta)\|u-\Pi u\|^2_{\partial T}+4C^2\left(\frac{1}{\alpha}+\beta\right)^2h\|\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma)\|^2_T.\end{aligned}$$

Applying Lemma A.5 again to δ_u finally gives

$$\begin{aligned}\sum_{T\in\mathcal{T}}\|\delta_u\|^2_T &\leq C^2h\sum_{T\in\mathcal{T}}\|\delta_u\|^2_{\partial T}\\ &\leq C^2h\sum_{T\in\mathcal{T}}\|\delta_u\|^2_{\partial T\setminus\partial\Omega}+(1+\alpha\beta)\|\delta_u\|^2_{\partial T\cap\partial\Omega}+\sum_{F_I\in\mathcal{F}_I}\alpha\beta\|[\delta_u]\|^2_{F_I}\\ &\leq C^2h\sum_{T\in\mathcal{T}}2(1+\alpha\beta)\|u-\Pi u\|^2_{\partial T}+4C^2\left(\frac{1}{\alpha}+\beta\right)^2h\|\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma)\|^2_T,\end{aligned}$$

concluding with

$$\begin{aligned}h\sum_{T\in\mathcal{T}}\|u-\Pi u\|^2_{\partial T}&\leq C^2h^{\min\{2s,2p+2\}}|u|_{H^{\min\{s,p+1\}}(\Omega)},\\ h^2\sum_{T\in\mathcal{T}}\|\operatorname{div}(\sigma)-\Pi^{p-1}(\operatorname{div}\sigma)\|^2_T&\leq C^2h^{\min\{2s-2,2p+2\}}|\operatorname{div}\sigma|_{H^{\min\{s-2,p\}}(\Omega)}.\end{aligned}$$

□

The following corollary gives the quasi-optimality of the interpolation P_u.

Corollary A.8. *Assuming H^s-regularity of the Helmholtz equation with $s>3/2$ then*

$$\begin{aligned}\sum_{T\in\mathcal{T}}\|u-P_u\|_T &\leq C\left(1+\sqrt{2(1+\alpha\beta)}\right)\inf_{v_h\in\mathcal{P}^p_{pw}(\mathcal{T})}\|u-v_h\|_\Omega\\ &\quad+2C\left(\frac{1}{\alpha}+\beta\right)h\inf_{\tau_h\in\mathcal{BMD}^p_{pw}(\mathcal{T})}\|\operatorname{div}(\sigma-\tau_h)\|_\Omega.\end{aligned}$$

Proof. The last lines in the proof of the previous proposition are changed by

$$h\sum_{T\in\mathcal{T}}\|u-\Pi u\|_{\partial T}^2 \le C^2 \inf_{v_h\in\mathcal{P}^p_{pw}(\mathcal{T})}\|u-v_h\|_\Omega^2,$$
$$h^2\sum_{T\in\mathcal{T}}\|\operatorname{div}\sigma-\Pi^{p-1}(\operatorname{div}\sigma)\|_T^2 \le C^2h^2 \inf_{\tau_h\in\mathcal{BMD}^p_{pw}(\mathcal{T})}\|\operatorname{div}(\sigma-\tau_h)\|_\Omega^2.$$

□

Next, the interpolation of the flux P_σ is analysed similarly as in [CGS10, A.3]. For that Lemma 46 and 47 are adapted.

Lemma A.9. *The space $\mathcal{BDM}^p_\perp(T)$ is defined by*

$$\mathcal{BDM}^p_\perp(T) := \{\sigma\in\mathcal{BMD}^(T) : (\sigma,\tau)_T = 0, \forall\tau\in[\mathcal{P}^{p-1}(T)]^d\}.$$

Assume $\sigma\in\mathcal{BDM}^p_\perp(T)$ then there holds

$$\|\sigma\|_T \le Ch^{\frac12}\|\sigma\cdot\mathbf{n}\|_{\partial T\setminus F^*},$$

with a constant $C>0$ independent of h,κ,α,β and an arbitrary facet F^* of T.

Proof. The proof is similar to [CGS10, Proposition A.3] and [Say13, Lemma 2.1]. First, it is shown that the boundary term is a norm on the space $\mathcal{BDM}^p_\perp(T)$. Assuming $\sigma\cdot\mathbf{n}=0$ on $\partial T\setminus F^*$ and by splitting σ into

$$\sigma=\sum_{i=1}^{d-1}\sigma\cdot\mathbf{n}_i$$

with $\mathbf{n}_i$ the normal vectors of facets different to F^* then there holds $\sigma\cdot\mathbf{n}_i\in\mathcal{P}^p_\perp(T)$ as well as $\sigma\cdot\mathbf{n}_i=0$ on the facet F_I. Then Lemma 45 implies that $\sigma\cdot\mathbf{n}_i$ vanishes on the whole element and therefore $\sigma=0$. The estimate is proven by a standard scaling argument. □

The main adaptation is the fact that the normal flux on one facet F^* can not be controlled, but it also is not required.

Lemma A.10. *Assuming H^1-regularity of σ and u, there exists a constant $C>0$, independent of κ,h,α,β, so that*

$$\|\sigma-P_\sigma(\sigma,u)\|_\Omega^2 \le C\left(\eta(\sigma)+\frac{1+\alpha\beta}{\alpha^{-1}+\beta}\eta(u)\right).$$

Proof. Let F^* be an arbitrary but fixed facet of T and consider the interpolant satisfying

$$(\mathcal{B}(\sigma),\tau_h)_T = (\sigma,\tau_h)_T \qquad \forall\tau_h\in[\mathcal{P}^{p-1}(T)]^d,$$
$$((\sigma-\mathcal{B}(\sigma))\cdot\mathbf{n},\mu_h)_F = 0 \qquad \forall\mu_h\in\mathcal{P}^p(F)$$

for all facets F of T except F^* and define $\delta_\sigma := \mathcal{B}(\sigma) - P_\sigma(\sigma, u)$. There holds for δ_σ

$$\begin{aligned}
(\delta_\sigma, \tau_h)_T &= 0 && \forall \tau_h \in [\mathcal{P}^{p-1}(T)]^d, \\
\left(\frac{1}{\alpha} + \beta\right)(\delta_\sigma \cdot \mathbf{n}, \mu_h)_{\partial T \cap F_I} &= \left(u - P_u + \frac{\alpha\beta}{2}[u - P_u], \mu_h\right)_{\partial T \cap F_I} && \forall \mu_h \in \mathcal{P}^p(F_I), \\
\left(\frac{1}{\alpha} + \beta\right)(\delta_\sigma \cdot \mathbf{n}, \mu_h)_{\partial T \cap F_O} &= (1 + \alpha\beta)(u - P_u, \mu_h)_{\partial T \cap F_O} && \forall \mu_h \in \mathcal{P}^p(F_O),
\end{aligned} \tag{A.4}$$

excluding the facet F^*. The idea is to choose $\mu_h = \delta_\sigma \cdot \mathbf{n}$ in the equations above and to estimate facet terms in the same way as above for the interpolation P_u or in the case of $\mathcal{RT}$-spaces.

Due to (A.4) Lemma A.9 can be applied yielding

$$\begin{aligned}
\left(\frac{1}{\alpha} + \beta\right)\|\delta_\sigma\|_T^2 &\leq \left(\frac{1}{\alpha} + \beta\right) Ch\|\delta_\sigma \cdot \mathbf{n}\|_{\partial T \setminus F^*}^2 \\
&= Ch((u - P_u, \delta_\sigma \cdot \mathbf{n})_{\partial T \setminus (\partial\Omega \cup F^*)} + \frac{\alpha\beta}{2}([u - P_u], \delta_\sigma \cdot \mathbf{n})_{\partial T \setminus (\partial\Omega \cup F^*)} \\
&\quad + (1 + \alpha\beta)(u - P_u, \delta_\sigma \cdot \mathbf{n})_{\partial T \cap \partial\Omega \cap F^*}) \\
&\leq C(1 + \alpha\beta)(\|u - v_h\|_\Omega + h|u - v_h|_{H^1(\Omega)})\|\delta_\sigma\|_T.
\end{aligned}$$

Note that only an element patch is required, and due to the shape regularity, a finite overlap is given, which can be inserted into the leading constant. This gives

$$\|\delta_\sigma\|_T \leq C\frac{1 + \alpha\beta}{\alpha^{-1} + \beta}\sqrt{\eta(u)}.$$

Note that the interpolant $\mathcal{B}$ also satisfies the following best approximation.

Lemma A.11. *For $\sigma \in [H^s(\Omega)]^d$ with $s > \frac{1}{2}$ there holds*

$$\|\sigma - \mathcal{B}\sigma\|_{L^2(\Omega)} \leq C \inf_{\tau_h \in \mathcal{BDM}_{pw}^p(\mathcal{T})} \left(\|\sigma - \tau_h\|_{L^2(\Omega)} + h^s \sum_{T \in \mathcal{T}} |\sigma - \tau_h|_{H^s(T)}\right).$$

Proof. The proof uses the continuity of the $\mathcal{B}$-interpolant in combination with the invariance of polynomials:

$$\begin{aligned}
\|\sigma - \mathcal{B}\sigma\|_{L^2(\Omega)} &\leq \inf_{\tau_h \in \mathcal{BDM}_{pw}^p(\mathcal{T})} \|\sigma - \tau_h\|_{L^2(\Omega)} + \|\tau_h - \mathcal{B}\sigma\|_{L^2(\Omega)} \\
&= \inf_{\tau_h \in \mathcal{BDM}_{pw}^p(\mathcal{T})} \|\sigma - \tau_h\|_{L^2(\Omega)} + \|\mathcal{B}(\tau_h - \sigma)\|_{L^2(\Omega)}.
\end{aligned}$$

Then, applying the continuity estimate from Chapter 16, Theorem 16.6 in [EG21] concludes the proof for the best approximation property of $\mathcal{B}$. □

This best approximation property is the last piece to finishing the proof for P_σ with

$$\|\sigma - P_\sigma(\sigma, u)\|_\Omega \leq C\left(\eta(\sigma) + \frac{1 + \alpha\beta}{\alpha^{-1} + \beta}\eta(u)\right)^{1/2}.$$

□

Acronyms

ABC absorbing boundary condition of first kind

BEM boundary element method

BVP boundary value problem

CG conjugate gradient

DG discontinuous Galerkin

DoF degrees of freedom

FDM finite difference method

FE finite element

FEM finite element method

GMRES generalised minimal residual

HDG hybrid discontinuous Galerkin

PCG preconditioned conjugate gradient

PML perfectly matched layer

List of Figures

3.1. Error of the discrete wave number κ_h of the H^1-formulation with respect to the mesh size h for the given wave number of $\kappa = 1$. Reference line of square convergence rate as a dashed line. 69

3.2. Error in the imaginary as well as the real part of the discrete wave number κ_h for the HDG-formulation with respect to the mesh size h for the given wave number of $\kappa = 1$. Reference lines of the second-order convergence rate and the first-order convergence rate as dashed lines. 71

4.1. First six meshes of the series of hierarchical 2D structured meshes for $N \in \{2, 4, 8, 16, 32, 64\}$ with $2N^2$ elements. The three additional meshes with $N \in \{128, 256, 512\}$ used in the simulations are omitted. 74

4.2. The errors for the plane wave 2D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 1$. 76

4.3. The errors for the plane wave 2D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 2$. 77

4.4. The series of hierarchical, structured 3D meshes for $N \in \{2, 4, 8\}$ with $6N^3$ elements are shown. 78

4.5. The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 1$ and the wave number $\kappa = 5$. 79

4.6. The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 2$ and the wave number $\kappa = 5$. 80

4.7. The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 1$ and the wave number $\kappa = 20$. . . . 81

4.8. The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The polynomial degree in the simulation has been chosen as $p = 2$ and the wave number $\kappa = 20$. . . . 82

4.9. The errors for the plane wave 3D simulations with respect to the ratio of the mesh size h to the wavelength $2\pi/\kappa$ are shown. The results for wave numbers $\kappa \in \{10, 20, 40\}$ can be seen. The polynomial degree in the simulation has been chosen as $p = 0$. 84

4.10. Dispersion for plane waves with a wave number $\kappa = 3$ and angles $\phi \in [0, \pi/2]$ on a structured mesh with $N = 2$. The H^1-formulation in green with square markers and the HDG-formulation in grey with circle marks. The reference line with radius one is highlighted in blue. 87
4.11. Dispersion and dissipation for plane waves with a wave number $\kappa = 3$ and angles $\phi \in [0, \pi/2]$ on a structured mesh with $N = 2$. The H^1-formulation in green with square markers and the HDG-formulation in grey with circle marks. 88
4.12. Dispersion and dissipation of the HDG-formulation for plane waves with wave numbers $\kappa \in \{0.75, 1.5, 3, 6\}$ and angles $\phi \in [0, \pi/2]$ on a structured mesh with $N = 2$ and the polynomial degree $p = 1$. 90
4.13. Pollution error of the HDG-formulation with respect to the wave number on structured meshes. The resolution is kept constant regarding $\kappa h = 2$ for spaces with varying polynomial degree p. 91
4.14. The mesh of the unit square for the preconditioner introduction example is shown. 92
4.15. Two element boundary patches of facet unknowns on a 2D skeleton mesh for the Jacobi and Gauss-Seidel preconditioners. The facet unknowns are highlighted with blue dots, and the patches with a red circles. 94
4.16. One facet patch of unknowns on a 3D mesh for the Jacobi and Gauss-Seidel preconditioners. The facet unknowns are highlighted with blue dots, and the patch with a red circle. 96
4.17. The behaviour of the GMRES method in combination with the Block-Jacobi preconditioner is shown. 96
4.18. The behaviour of the GMRES method in combination with the Gauss-Seidel preconditioner is shown. 97
4.19. The layers for the sweeping preconditioner on the domain of the introductory problem are shown. 98
4.20. The behaviour of the GMRES method in combination with the sweeping preconditioner is shown. 99
4.21. The subdomains for the domain decomposition preconditioner on the domain of the introductory problem are shown. 100
4.22. The behaviour of the GMRES method in combination with the domain decomposition preconditioner is shown. 100
4.23. The unit square domain with an off-centred circle as a scattering object can be seen. 101
4.24. The solution for the α-β test example can be seen. The view is tilted slightly to highlight the oscillating pressure. 102
4.25. The layers for the sweeping preconditioner on the domain of the α-β test problem are shown. 102
4.26. The subdomains for the domain decomposition preconditioner on the domain of the α-β test problem are shown. 102

4.27. The iteration numbers for the four considered preconditioners for different choices of the stabilisation parameters α and β can be seen. The maximum number of iterations was set to 300, which is an upper boundary in the graphs. On the top left are the iterations for Jacobi, on the top right Gauss-Seidel, on the bottom left, sweeping and on the bottom right, domain decomposition can be seen. 104
4.28. The sweeping layers for the computational costs example with respect to the chosen wave numbers are shown. 105
4.29. The subdomains for the computational costs example with respect to the chosen wave numbers are shown. 106
4.30. The pressure of the solutions for the computational costs example with respect to the chosen wave numbers are shown. 107
4.31. The wall time of the four preconditioners with respect to the number of hybrid facet unknowns n_{DoF_s} is shown. 107
4.32. The CPU time of the four preconditioners with respect to the number of hybrid facet unknowns n_{DoF_s} is shown. 108
4.33. The number of iterations of the four preconditioners with respect to the number of hybrid facet unknowns n_{DoF_s} is shown. 108
4.34. The real part $\Re(u_h)$ off the pressure is shown for simulations with the heterogeneous material coefficients c_1 and c_2. A Gaussian peak has been applied on the left boundary as an excitation. 111
4.35. A cut view along the x, z-plane of symmetry is shown for the 3D-example with spherical scatterers. 112

List of Tables

4.1. Number DoFs for the 2D simulation with heterogeneous materials 111
4.2. Number of iterations and computation times for heterogeneous materials . . 111
4.3. Number DoFs for the 3D scattering experiment 112

Bibliography

[Alt12] Hans Wilhelm Alt. Lineare Funktionale. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*, pages 171–227, 2012.

[Ber94] Jean-Pierre Berenger. A Perfectly Matched Layer for the Absorption of Electromagnetic Waves. *Journal of computational physics*, 114(2):185–200, 1994.

[BHPD23] Jacob Badger, Stefan Henneking, Socratis Petrides, and Leszek Demkowicz. Scalable DPG multigrid solver for Helmholtz problems: A study on convergence. *Computers & Mathematics with Applications*, 148:81–92, 2023.

[Bra77] Achi Brandt. Multi-Level Adaptive Solutions to Boundary-Value Problems. *Mathematics of computation*, 31(138):333–390, 1977.

[BS97] Ivo M. Babuška and Stefan A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Journal on numerical analysis*, 34(6):2392–2423, 1997.

[CCJP20] Xavier Claeys, Francis Collino, Patrick Joly, and Emile Parolin. A Discrete Domain Decomposition Method for Acoustics with Uniform Exponential Rate of Convergence Using Non-local Impedance Operators. In *Domain Decomposition Methods in Science and Engineering XXV 25*, pages 310–317. Springer, 2020.

[CCP22] Xavier Claeys, Francis Collino, and Emile Parolin. Nonlocal Optimized Schwarz Methods for time-harmonic Electromagnetics. *Advances in Computational Mathematics*, 48(6):72, 2022.

[CGL09] Bernardo Cockburn, Jayadeep Gopalakrishnan, and Raytcho Lazarov. Unified Hybridization of Discontinuous Galerkin, Mixed, and Continuous Galerkin Methods for Second Order Elliptic Problems. *SIAM J. Numer. Anal*, 47:1319–1365, 08 2009.

[CGS10] Bernardo Cockburn, Jayadeep Gopalakrishnan, and Francisco-Javier Sayas. A projection-based error analysis of HDG methods. *Mathematics of Computation*, 79(271):1351–1367, 2010.

[CKS05] Bernardo Cockburn, Guido Kanschat, and Dominik Schötzau. A locally conservative LDG method for the incompressible Navier-Stokes equations. *Mathematics of computation*, 74(251):1067–1095, 2005.

[cla23] Non-selfadjoint impedance in Generalized Optimized Schwarz Methods. *IMA Journal of Numerical Analysis*, 43(5):3026–3054, 2023.

[Clo60] Ray William Clough. The Finite Element Method in Plane Stress Analysis. 1960.

[CLX13] Huangxin Chen, Peipei Lu, and Xuejun Xu. A Hybridizable Discontinuous Galerkin Method for the Helmholtz Equation with High Wave Number. *SIAM Journal on Numerical Analysis*, 51(4):2166–2188, 2013.

[CP22] Xavier Claeys and Emile Parolin. Robust treatment of cross-points in optimized Schwarz methods. *Numerische Mathematik*, pages 1–38, 2022.

[d'A47] Jean le Rond d'Alembert. Recherches sur la courbe que forme une corde tendue mise en vibration. 1747.

[DCDBK+16] Arne De Coninck, Bernard De Baets, Drosos Kourounis, Fabio Verbosio, Olaf Schenk, Steven Maenhout, and Jan Fostier. Needles: Toward Large-Scale Genomic Prediction with Marker-by-Environment Interaction. 203(1):543–555, 2016.

[DG10] Leszek Demkowicz and Jayadeep Gopalakrishnan. A class of discontinuous Petrov–Galerkin methods. Part I: The transport equation. *Computer Methods in Applied Mechanics and Engineering*, 199(23-24):1558–1572, 2010.

[DG11] Leszek Demkowicz and Jayadeep Gopalakrishnan. A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions. *Numerical Methods for Partial Differential Equations*, 27(1):70–105, 2011.

[DGMZ12] Leszek Demkowicz, Jayadeep Gopalakrishnan, Ignacio Muga, and Jeff Zitelli. Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 213:126–138, 2012.

[DGN12] Leszek Demkowicz, Jayadeep Gopalakrishnan, and Antti H Niemi. A class of discontinuous Petrov–Galerkin methods. Part III: Adaptivity. *Applied numerical mathematics*, 62(4):396–427, 2012.

[duf83] The Multifrontal Solution of Indefinite Sparse Symmetric Linear. *ACM Transactions on Mathematical Software (TOMS)*, 9(3):302–325, 1983.

[duf17] *Direct Methods for Sparse Matrices*. Oxford University Press, 2017.

[EG21] Alexandre Ern and Jean-Luc Guermond. *Finite elements I: Approximation and interpolation*, volume 72. Springer Nature, 2021.

[EM77] Bjorn Engquist and Andrew Majda. Absorbing Boundary Conditions for the Numerical Simulation of Waves. *Proceedings of the National Academy of Sciences of the United States of America*, 74:1765–6, 06 1977.

[EM12] Sofie Esterhazy and Jens Markus Melenk. On stability of discretizations of the Helmholtz equation. In *Numerical analysis of multiscale problems*, pages 285–324. Springer, 2012.

[Eul49] Leonhard Euler. De vibratione chordarum exercitatio. *Nova Acta Eruditorum*, pages 512–527, 1749.

[EY11a] Björn Engquist and Lexing Ying. Sweeping preconditioner for the Helmholtz equation: hierarchical matrix representation. *Communications on pure and applied mathematics*, 64(5):697–735, 2011.

[EY11b] Björn Engquist and Lexing Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Modeling & Simulation*, 9(2):686–710, 2011.

[FLX16] Xiaobing Feng, Peipei Lu, and Xuejun Xu. A Hybridizable Discontinuous Galerkin Method for the Time-Harmonic Maxwell Equations with High Wave Number. *Computational Methods in Applied Mathematics*, 16(3):429–445, 2016.

[FW09] Xiaobing Feng and Haijun Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM Journal on Numerical Analysis*, 47(4):2872–2896, 2009.

[FW11] Xiaobing Feng and Haijun Wu. hp-Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Mathematics of Computation*, 80:1997–2024, 02 2011.

[FX13] Xiaobing Feng and Yulong Xing. Absolutely stable local discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Mathematics of Computation*, 82(283):1269–1296, 2013.

[GGS20] Shihua Gong, Ivan Graham, and Euan A. Spence. Domain decomposition preconditioners for high-order discretisations of the heterogeneous Helmholtz equation, 04 2020.

[GM11] Roland Griesmaier and Peter Monk. Error Analysis for a Hybridizable Discontinuous Galerkin Method for the Helmholtz Equation. *Journal of Scientific Computing*, 49:291–310, 2011.

[GMO14] Jayadeep Gopalakrishnan, Ignacio Muga, and Nicole Olivares. Dispersive and dissipative errors in the DPG method with scaled norms for Helmholtz equation. *SIAM Journal on Scientific Computing*, 36(1):A20–A39, 2014.

[Guy65] Robert J Guyan. Reduction of Stiffness and Mass Matrices. *AIAA journal*, 3(2):380–380, 1965.

[HPS13] Martin Huber, Astrid Pechstein, and Joachim Schöberl. Hybrid Domain Decomposition Solvers for the Helmholtz and the Time Harmonic Maxwell's

equation. In Randolph Bank, Michael Holst, Olof Widlund, and Jinchao Xu, editors, *Domain Decomposition Methods in Science and Engineering XX*, pages 279–287, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[HS+52] Magnus Rudolph Hestenes, Eduard Stiefel, et al. *Methods of Conjugate Gradients for Solving Linear Systems*, volume 49. NBS Washington, DC, 1952.

[HS14] Martin Huber and Joachim Schöberl. Hybrid Domain Decomposition Solvers for the Helmholtz Equation. In Jocelyne Erhel, Martin J. Gander, Laurence Halpern, Géraldine Pichot, Taoufik Sassi, and Olof Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXI*, pages 351–358, Cham, 2014. Springer International Publishing.

[Hub13] Martin Huber. *Hybrid discontinuous Galerkin methods for the wave equation.* PhD thesis, Technische Universität Wien, 2013.

[IB95] Frank Ihlenburg and Ivo M. Babuška. Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.

[IB97] Frank Ihlenburg and Ivo M. Babuška. Finite element solution of the Helmholtz equation with high wave number part II: the hp version of the FEM. *SIAM Journal on Numerical Analysis*, 34(1):315–358, 1997.

[KFS18] Drosos Kourounis, Alexander Fuchs, and Olaf Schenk. Towards the Next Generation of Multiperiod Optimal Power Flow Solvers. *IEEE Transactions on Power Systems*, PP(99):1–10, 2018.

[KK97] George Karypis and Vipin Kumar. METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices. 1997.

[KWH+22] Andreas Klöckner, Matt Wala, Nathan Hartland, AlDanial, Fritz Obermeyer, Cathy Wu, and Christoph Gohlke. PyMetis, 2022.

[LCQ17] Peipei Lu, Huangxin Chen, and Weifeng Qiu. An absolutely stable *hp*-HDG method for the time-harmonic Maxwell equations with high wave number. *Mathematics of Computation*, 86(306):pp. 1553–1577, 2017.

[LeV07] Randall J LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. SIAM, 2007.

[LHS22] Michael Leumüller, Karl Hollaus, and Joachim Schöberl. Domain decomposition and upscaling technique for metascreens. *COMPEL-The international journal for computation and mathematics in electrical and electronic engineering*, 41(3):938–953, 2022.

[LS23] Michael Leumüller and Joachim Schöberl. Error Analysis of an HDG Method with Impedance Traces for the Helmholtz Equation, 2023.

[MCF23] Axel Modave and Théophile Chaumont-Frelet. A hybridizable discontinuous Galerkin method with characteristic variables for Helmholtz problems. *Journal of Computational Physics*, 493:112459, 2023.

[Mel95] Jens Markus Melenk. *On Generalized Finite Element Methods.* PhD thesis, University of Maryland at College Park, 1995.

[Mon03] Peter Monk. *Finite element methods for Maxwell's equations.* Oxford university press, 2003.

[MPS13] Jens Markus Melenk, Asieh Parsania, and Stefan A. Sauter. General DG-Methods for Highly Indefinite Helmholtz Problems. *Journal of Scientific Computing*, 57, 12 2013.

[MS14] Andrea Moiola and Euan A. Spence. Is the Helmholtz equation really sign-indefinite? *Siam Review*, 56(2):274–312, 2014.

[MS22] Jens Markus Melenk and Stefan A. Sauter. Wavenumber-explicit hp-FEM analysis for Maxwell's equations with impedance boundary conditions, 2022.

[MSS10] Peter Monk, Joachim Schöberl, and Astrid Sinwel. Hybridizing Raviart-Thomas Elements for the Helmholtz Equation. *Electromagnetics*, 30:149 – 176, 2010.

[Nat93] Frédéric Nataf. On the Use of Open Boundary Conditions in Block Gauss-Seidel Methods for the Convection-Diffusion Equation. 1993.

[Pec22] Clemens Pechstein. A unified theory of non-overlapping Robin-Schwarz methods–continuous and discrete, including cross points. *arXiv preprint arXiv:2204.03436*, 2022.

[rei71] On the method of conjugate gradients for the solution of large sparse systems of linear equations. In *Proc. Conf. on Large Sparse Set of Linear Equations, 1971.* Academic Press, 1971.

[RH73] William H. Reed and Thomas R. Hill. Triangular Mesh Methods for the Neutron Transport Equation. Technical report, Los Alamos Scientific Lab., N. Mex.(USA), 1973.

[Say13] Francisco-Javier Sayas. From Raviart-Thomas to HDG: a personal voyage. 2013.

[Sch17] Issai Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. 1917.

[Sch74] Alfred H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Mathematics of Computation*, 28:959–962, 1974.

[Sch97] Joachim Schöberl. NETGEN An advancing front 2D/3D-mesh generator based on abstract rules. *Computing and Visualization in Science*, 1:41–52, 07 1997.

[Sch14] Joachim Schöberl. C++11 Implementation of Finite Elements in NGSolve, 09 2014.

[Som12] Arnold Sommerfeld. Die Greensche Funktion der Schwingungslgleichung. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 21:309–352, 1912.

[SS11] Stefan A. Sauter and Christoph Schwab. *Boundary Element Methods*. Springer, 2011.

[SZ05] Joachim Schöberl and Sabine Zaglmayr. High order Nédélec elements with local complete sequence properties. *COMPEL*, 24(2):374–384, 2005.

[VCKS17] Fabio Verbosio, Arne De Coninck, Drosos Kourounis, and Olaf Schenk. Enhancing the scalability of selected inversion factorization algorithms in genomic prediction. *Journal of Computational Science*, 22(Supplement C):99 – 108, 2017.

[vH82] Hermann von Helmholtz. *Die thermodynamik chemischer Vorgänge*. Die thermodynamik chemischer Vorgänge. 1882.

[Wes20] Markus Wess. *Frequency-Dependent Complex-Scaled Infinite Elements for Exterior Helmholtz Resonance Problems*. PhD thesis, 06 2020.

[Whi15] Edmund Taylor Whittaker. XVIII.—On the Functions which are represented by the Expansions of the Interpolation-Theory. *Proceedings of the Royal Society of Edinburgh*, 35:181–194, 1915.

[Wu14] Haijun Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: linear version. *IMA Journal of Numerical Analysis*, 34(3):1266–1288, 2014.

[Zum65] G. Zumpe. J. H. Argyris, Recent Advances in Matrix Methods of Structural Analysis. (Progress in Aeronautical Sciences, Volume 4) XIII + 187 S. m. Abb. u. Fig. Oxford/London/New York/Paris 1964. Pergamon Press. Preis geb. 60 s. net. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 45(5):366–367, 1965.

[ZW12] Lingxue Zhu and Haijun Wu. Pre-asymptotic Error Analysis of CIP-FEM and FEM for Helmholtz Equation with High Wave Number. Part II: *hp* version, 2012.

[ZW20] Bingxin Zhu and Haijun Wu. Hybridizable Discontinuous Galerkin Methods for Helmholtz Equation with High Wave Number. Part I: Linear case, 2020.

Curriculum Vitae

Personal Data

Name	**Michael Leumüller**
Birthday	15.03.1988
Nationality	Austrian
Email	michael.leumueller@tuwien.ac.at

Professional Experience

01/2019–09/2024	Research Associate, Institute for Analysis and Scientific Computing, Technische Universität Wien, Python Development of Simulation Tools for Electromagnetic Scattering Problems with the Finite Element Method
03/2019–07/2019	Tutor, Technische Universität Wien, Differential Equations
04/2018–09/2018	Research Associate, Institute for Analysis and Scientific Computing, Technische Universität Wien, Developing a Solver for Rational Resonance Problems, Co-Supervisor of the Bachelor Thesis ”FEAST Eigenvalue Solver”
10/2016–02/2017	Tutor, Technische Universität Wien, Numerical Analysis
07/2016–08/2016	Research Associate, Institute for Analysis and Scientific Computing, Technische Universität Wien, Implementation of Beyn's Method for Nonlinear Eigenvalue Problems
03/2015–08/2015	Research Associate, Institute for Analysis and Scientific Computing, Technische Universität Wien, Implementation and Analysis of the Spectral Lanczos Method
2008–2012	Software Engineer, ITPRO System Consulting, Linz, Development of Web- and Mobile- Applications with C# and ASP.NET
08/2006	Internship, TRUCK-CENTER L. Katzinger GmbH, Altenfelden
09/2006	Internship, Ganser Liftsysteme, St. Peter am Wimberg

Education

Since 2019	PhD., Technical Mathematics, Technische Universität Wien, Focus: Numerics, Subject: An HDG Method with Impedance Traces for the Helmholtz Equation
2016–2018	MSc., Technical Mathematics, Technische Universität Wien, Graduated 10/2018 with distinction, Thesis: Computing Resonances in Metallic Photonic Crystals
2012–2016	BSc., Technical Mathematics, Technische Universität Wien, Graduated 01/2016 with distinction, Thesis: Inexact Spectral Lanczos
2002–2007	High School Diploma, Betriebsinformatik, HTBLA Neufelden, Graduated 06/2007 with distinction, Thesis: Gemeindebuchhaltung – Gemeinde Hallwang

---

Publications

M. Leumüller and J. Schöberl: *Error Analysis of an HDG Method with Impedance Traces for the Helmholtz Equation*, arXiv, 2023
M. Leumüller, K. Hollaus and J. Schöberl: *Domain Decomposition and Upscaling Technique for Metascreens*, COMPEL, 2022
M. Leumüller and K. Hollaus: *Multiscale Finite Element Method for Ventilation Panels*, IEEE Transactions on Magnetics, 2022
M. Leumüller, B. Auinger, J. Schöberl, K. Hollaus: *Enhanced Technique for Metascreens using the Generalized Finite Element Method*, IEEE Transactions on Magnetics, 2021
M. Leumüller, B. Auinger, H. Hackl, J. Schöberl and K. Hollaus: *Imperfect EM Shielding by Thin Conducting Sheets with PEC and SIBC*, IEEE Xplore, 2019

Vienna, February 6, 2025

Michael Leumüller