8-2014

Numerical Methods and Algorithms for High Frequency Wave Scattering Problems in Homogeneous and Random Media

Cody Samuel Lorton
University of Tennessee - Knoxville, clorton@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Cody Samuel Lorton entitled "Numerical Methods and Algorithms for High Frequency Wave Scattering Problems in Homogeneous and Random Media." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Mathematics.

Xiaobing H. Feng, Major Professor

We have read this dissertation and recommend its acceptance:

Michael W. Berry, Ohannes A. Karakashian, Steven M. Wise

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

8-2014

Numerical Methods and Algorithms for High Frequency Wave Scattering Problems in Homogeneous and Random Media

Cody Samuel Lorton
clorton@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Cody Samuel Lorton entitled "Numerical Methods and Algorithms for High Frequency Wave Scattering Problems in Homogeneous and Random Media." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Mathematics.

Xiaobing H. Feng, Major Professor

We have read this dissertation and recommend its acceptance:

Michael W. Berry, Ohannes A. Karakashian, Steven M. Wise

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Numerical Methods and Algorithms for High Frequency Wave Scattering Problems in Homogeneous and Random Media

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Cody Samuel Lorton

August 2014

This dissertation is dedicated to Jesus Christ my Lord and Savior, my wife Rebekah, and my children Bella and Shepherd.

Acknowledgements

I would first like to thank my advisor, Professor Xiaobing Feng, for his excellent leadership and instruction throughout my graduate studies. I could not ask for a better advisor and I know that his insight, guidance, and hard work have provided a lasting foundation for my research career.

Secondly, I would like to thank the remaining members of my dissertation committee Professors Steven Wise, Ohannes Karakashian, and Michael Berry for their time and guidance.

I would also like to thank all those in the mathematics department at the University of Tennessee that helped me along during my graduate studies. In particular, I must thank Professors Steven Wise, Ohannes Karakashian, and Suzanne Lenhart for helping lay the foundation of my career in applied mathematics. I must also thank Pam Armentrout and Ben Walker for their support as well.

I would like to thank my wife Rebekah and my children Bella and Shepherd. I could not accomplish all that I have without their support and love.

Lastly, I would like to thank God. It is only by God's loving will that I completed this dissertation.

Abstract

This dissertation consists of four integral parts with a unified objective of developing efficient numerical methods for high frequency time-harmonic wave equations defined on both homogeneous and random media. The first part investigates the generalized weak coercivity of the acoustic Helmholtz, elastic Helmholtz, and time-harmonic Maxwell wave operators. We prove that such a weak coercivity holds for these wave operators on a class of more general domains called generalized star-shape domains. As a by-product, solution estimates for the corresponding Helmholtz-type problems are obtained.

The second part of the dissertation develops an absolutely stable (i.e. stable in all mesh regimes) interior penalty discontinuous Galerkin (IP-DG) method for the elastic Helmholtz equations. A special mesh-dependent sesquilinear form is proposed and is shown to be weakly coercive in all mesh regimes. We prove that the proposed IP-DG method converges with optimal rate with respect to the mesh size. Numerical experiments are carried out to demonstrate the theoretical results and compare this method to the standard finite element method.

The third part of the dissertation develops a Monte Carlo interior penalty discontinuous Galerkin (MCIP-DG) method for the acoustic Helmholtz equation defined on weakly random media. We prove that the solution to the random Helmholtz problem has a multi-modes expansion (i.e., a power series in a medium-related small parameter). Using this multi-modes expansion an efficient and accurate numerical method for computing moments of the solution to the random Helmholtz

problem is proposed. The proposed method is also shown to converge optimally. Numerical experiments are carried out to compare the new multi-modes MCIP-DG method to a classical Monte Carlo method.

The last part of the dissertation develops a theoretical framework for Schwarz preconditioning methods for general nonsymmetric and indefinite variational problems which are discretized by Galerkin-type discretization methods. Such a framework has been missing in the literature and is of great theoretical and practical importance for solving convection-diffusion equations and Helmholtz-type wave equations. Condition number estimates for the additive and hybrid Schwarz preconditioners are established under some structure assumptions. Numerical experiments are carried out to test the new framework.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As a fundamental mechanism for energy transmission, wave phenomena are ubiquitous in our world. Waves are determined by their sources and the media in which they propagate. Wave scattering describes the physical phenomena in which wave propagation is changed due to some non-uniformity in the medium in which the wave is traveling. Wave scattering problems have applications in many scientific fields including communications, defense, aviation, geoscience, medical science, manufacturing, etc.

The goal of this dissertation is to develop efficient numerical methods for high frequency time-harmonic wave equations defined on both homogeneous and random media. Specifically, it focuses on three basic mathematical models of wave scattering and propagation. These are the acoustic Helmholtz, elastic Helmholtz, and time-harmonic Maxwell's equations.

The first wave scattering problem we will consider is the acoustic/scalar Helmholtz problem given by

$$-\Delta u - k^2 u = f \qquad \text{in } \Omega, \tag{1.1}$$

$$\frac{\partial u}{\partial \mathbf{n}_+} + iku = g \qquad \text{on } \partial\Omega_+, \tag{1.2}$$

$$u = 0 \qquad \text{on } \partial\Omega_-. \tag{1.3}$$

Here, $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is a domain that consists of some acoustic medium and $u : \Omega \to \mathbb{C}$ is the pressure of the medium. k is the wave number, defined by $k := \frac{\omega}{c}$, where $\omega, c > 0$ are the angular frequency and speed of the wave in Ω, respectively. f is the external source. $\partial\Omega$ is decomposed into two pieces $\partial\Omega_+$ and $\partial\Omega_-$. $\mathbf{n}_+, \mathbf{n}_-$ denotes the unit outward normal vectors on $\partial\Omega_+$ and $\partial\Omega_-$, respectively. Typically, wave propagation problems are posed on large or unbounded domains complemented with a far-field radiation condition. For computational purposes, we choose to utilize a truncated domain. $\partial\Omega_+$ represents the boundary from this truncation. When $g = 0$, (1.2) is a first order absorbing boundary condition [35], which is an artificial boundary condition that absorbs incoming waves at the boundary. $\partial\Omega_-$ is the scattering portion of the domain boundary. (1.3) ensures that the scattering boundary $\partial\Omega_-$ is sound soft.

The acoustic Helmholtz problem comes from seeking time-harmonic solutions or applying Fourier transforms (in t) to the well-known acoustic wave problem

$$\begin{aligned}
\frac{1}{c^2}U_{tt} - \Delta U &= F && \text{in } \Omega \times (0, \infty),\\
\frac{1}{c}U_t + \frac{\partial U}{\partial \mathbf{n}_+} &= G && \text{on } \partial\Omega_+ \times (0, \infty),\\
U &= 0 && \text{on } \partial\Omega_- \times (0, \infty),\\
U = U_t &= 0 && \text{in } \Omega \times \{t = 0\}.
\end{aligned}$$

Here u, f, g from (1.1)–(1.3) take the form

$$\begin{aligned}
u(x) &= \int_{-\infty}^{\infty} e^{i\omega t} U(x, t) \mathrm{d}t,\\
f(x) &= \int_{-\infty}^{\infty} e^{i\omega t} F(x, t) \mathrm{d}t,\\
g(x) &= \int_{-\infty}^{\infty} e^{i\omega t} G(x, t) \mathrm{d}t.
\end{aligned}$$

Computing solutions to (1.1)–(1.3) is known as the frequency domain treatment for wave problems [29, 30]. This approach is favorable, because for a set of chosen frequencies one can compute time-harmonic solutions in parallel by solving a set of independent acoustic Helmholtz problems. Also, the use of frequency specific time-harmonic waves often arise from many applications.

The second problem that we will consider is the elastic Helmholtz problem given by

$$-\omega^2\rho\mathbf{u} - \mathbf{div}\,(\sigma(\mathbf{u})) = \mathbf{f} \qquad \text{in } \Omega, \tag{1.4}$$

$$i\omega A\mathbf{u} + \sigma(\mathbf{u})\mathbf{n} = \mathbf{g} \qquad \text{on } \partial\Omega. \tag{1.5}$$

Similar to the acoustic Helmholtz problem, (1.4)–(1.5) arise from seeking time-harmonic solutions to the well-known linear elastic wave equations. $\Omega \subset \mathbb{R}$ $(d = 1, 2, 3)$ is a domain that consists of some elastic medium and $\mathbf{u} : \Omega \to \mathbb{C}^d$ is the displacement vector of that medium. ω, ρ are the angular frequency of the elastic wave and the density of the elastic medium, respectively. For the elastic Helmholtz equation, the wave number is given by $k = \sqrt{\rho}\omega$. $\sigma(\mathbf{u})$ denotes the stress tensor defined by

$$\sigma(\mathbf{u}) := 2\mu\varepsilon(\mathbf{u}) + \lambda \mathrm{div}\,\mathbf{u} I, \qquad \varepsilon(\mathbf{u}) := \frac{1}{2}\big(\nabla\mathbf{u} + \nabla\mathbf{u}^T\big).$$

Here, $\mu, \lambda > 0$ are the Lamé constants for the elastic medium Ω and $\varepsilon(\mathbf{u})$ is called the strain tensor. We do not consider a scattering portion of the boundary in (1.4)–(1.5) for simplicity. Similar to (1.2), when $\mathbf{g} = \mathbf{0}$, (1.5) is a first order absorbing boundary condition [35]. A is a $d \times d$ symmetric positive-definite constant matrix.

Lastly, we consider the time-harmonic Maxwell's equations given by

$$\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} - k^2\mathbf{E} = \mathbf{f} \qquad \text{in } \Omega, \tag{1.6}$$

$$\mathbf{curl}\,\mathbf{E} \times \mathbf{n} - \mathbf{i}\lambda\mathbf{E}_T = \mathbf{g} \qquad \text{on } \partial\Omega. \tag{1.7}$$

(1.6)–(1.7) arise from seeking time-harmonic solutions to the well-known Maxwell's equations (c.f. [25]). $\Omega \subset \mathbb{R}^3$ and $\mathbf{E} : \Omega \to \mathbb{C}^3$ is the electrical field of Ω. $\mathbf{E}_T = (\mathbf{n} \times \mathbf{E}) \times \mathbf{n}$ is the tangential part of $\mathbf{E}$. The wave number k is defined as $k = \omega\sqrt{\mu_0 \varepsilon_0}$, where $\omega > 0$ is the angular frequency of the wave, $\varepsilon_0 > 0$ is the electrical permittivity of the medium, and $\mu_0 > 0$ is the magnetic permeability of the medium. Similar to the elastic Helmholtz problem, we will not consider a scattering portion of the boundary for simplicity. (1.7) is the standard impedance boundary condition, with $\lambda > 0$ called the impedance constant.

Because the acoustic Helmholtz, elastic Helmholtz, and time-harmonic Maxwell's problems all arise by seeking time-harmonic solutions to wave problems and thus have similar characteristics, these three problems will be referred to as Helmholtz-type problems in this dissertation.

1.1 The State of the Art

Many numerical methods have been developed for the three Helmholtz-type problems in homogeneous media, i.e. for constant wave number k. These include finite difference (FD), finite volume (FV), finite element (FE), and discontinuous Galerkin (DG) methods [1, 2, 3, 6, 7, 10, 13, 19, 20, 23, 26, 29, 30, 32, 36, 38, 48, 51, 52, 54, 53, 59, 60, 61, 64, 67, 70, 75, 76]. This section will discuss some of the challenges that arise from solving the three Helmholtz-type problems numerically.

Recall that Helmholtz-type problems are wave problems. Solutions to these problems are oscillatory with wave length $\ell = 2\pi/k$. Enough grid points must be used in the spacial domain to resolve the wave. The widely accepted rule-of-thumb is to use 6–12 mesh/grid points per wavelength. This rule-of-thumb was proved rigorously for the linear FE method for the 1-D acoustic Helmholtz problem [54, 53]. This yields a mesh constraint of $kh = O(1)$, where h is the mesh size. Meshes satisfying this mesh constraint make up the so-called pre-asymptotic mesh regime. Therefore,

in the high frequency case, discretizing the Helmholtz-type problems yields a large system of linear equations that must be solved.

In the case of linear FE method for the 1-D acoustic Helmholtz problem, the authors of [54] showed that the H^1 error bound for the FE solution contains a term of order k^3h^2. This term is called the pollution term and an increase in error as one increases the wave number k under the constraint $kh = O(1)$ is called the pollution effect. The authors of [13, 29, 54] showed that the pollution effect is inherent in Helmholtz-type problems and also leads to a loss of stability of standard discretization techniques. To eliminate the pollution effect, a more stringent mesh constraint $k^2h = O(1)$, called the asymptotic mesh constraint, is used. It is under this constraint that stability is proved for standard discretization techniques applied to Helmholtz-type problems. Shen and Wang obtained an absolutely stable (i.e. stable for all $k, h > 0$) spectral Galerkin discretization for the radially symmetric acoustic Helmholtz equation in [73]. Feng and Wu obtained absolutely stable interior penalty discontinuous Galerkin (IP-DG) discretizations for the acoustic Helmholtz and time-harmonic Maxwell's equations in [42, 43, 44]. Feng and Xing obtained an absolutely stable local discontinuous Galerkin (LDG) method for the acoustic Helmholtz equation in [45].

As noted previously, for k large one must solve a large linear system of equations in order to solve Helmholtz-type problems. From (1.1),(1.4), and (1.6) we see that for k (or ω) large the Helmholtz-type PDE operators are indefinite. Thus, any discretization method applied to Helmholtz-type PDEs yield indefinite, and ill-conditioned linear systems. It is known that standard iterative methods do not work well when applied to Helmholtz-type problems. In fact, many are not convergent (c.f. [37]). There is no framework in place to analyze multi-level solvers/preconditioners, such as multi-grid and Schwarz domain decomposition methods, for indefinite problems like the Helmholtz-type problems. Also, if one must adhere to the stringent mesh constraint $k^2h = O(1)$ in the high frequency case, practical coarse mesh spaces for multi-level solvers cannot be implemented.

1.2 Summary of this Dissertation

This dissertation contains five additional chapters. In Chapter 2, we study the three Helmholtz-type problems at the PDE level. In particular, we show that all three Helmholtz-type PDEs satisfy a generalized weak coercivity property. This generalized weak coercivity property was proved to hold for the time-harmonic Maxwell's equations in [43]. The techniques used to prove these generalized weak coercivity properties were first used in [27] and rely on Rellich identities for the Helmholtz-type operators as well as a star-shape condition on the domain Ω. Because a star-shape condition can be viewed as restrictive, the analysis in Chapter 2 is carried out on generalized star-shape domains. As a corollary of the generalized weak coercivity property, solution estimates are proved in energy norms for each Helmholtz-type problem.

Chapter 3 develops an absolutely stable interior penalty discontinuous Galerkin (IP-DG) method for the elastic Helmholtz problem. Recall that this was already done for the acoustic Helmholtz and time-harmonic Maxwell's problem [42, 43, 44]. This chapter uses new techniques, introduced in [42, 43, 44], to obtain stability and optimal (in h) error estimates in the pre-asymptotic mesh regime. Analysis in the asymptotic mesh regime is also carried out using the standard Schatz argument. Numerical experiments are provided to demonstrate the theoretical results presented in this chapter.

In Chapter 4, we develop a Monte Carlo interior penalty discontinuous Galerkin (MCIP-DG) method for the acoustic Helmholtz problem in random media. The random media is characterized by use of a random wave number in the acoustic Helmholtz problem. In this chapter, we show that when this random wave number is a random perturbation of some constant wave number, the solution takes the form of a power series expansion in the perturbation parameter. We call this series expansion the multi-modes expansion. Using this multi-modes expansion, an efficient and accurate MCIP-DG method is obtained. Numerical experiments presented to

show that the multi-mode MCIP-DG method is accurate compared to the classical MCIP-DG method and much more efficient.

There is no general framework to study Schwarz preconditioners for general non-Hermitian and indefinite variational problems. This includes Helmholtz-type problems. As a first step to meet this challenge, in Chapter 5, we develop a general framework to analyze Schwarz preconditioners for real-valued non-symmetric and indefinite variational problems. In this chapter the theoretical framework is introduced and different Schwarz preconditioners are developed and analyzed. This new framework is designed as a generalization of the existing Schwarz framework given in [77]. Extensive numerical experiments are also conducted to demonstrate some properties of Schwarz preconditioners applied to a non-symmetric problem. Though this framework does not apply directly to the three Helmholtz-type problems, it is our hope that this initial step will lead to a generalization that also applies to these Helmholtz-type problems.

Lastly, Chapter 6 discusses a number of future research directions that come from this dissertation.

1.3 Notation

This dissertation adopts many standard notation conventions. Much of the notation is explained when it is introduced, but we define some standard notations here that will be used throughout.

$H^\beta(\Omega)$ will be used to denote the standard Sobolev space $W^{\beta,2}(\Omega)$. For any $S \subset \Omega$ and $\Sigma \subset \partial\Omega$, let $(\cdot,\cdot)_S$ and $\langle\cdot,\cdot\rangle_\Sigma$ denote the standard L^2-inner products defined by

$$(u,v)_S := \int_S u \cdot \overline{v}\,\mathrm{d}x, \qquad \langle u,v\rangle_\Sigma := \int_S u \cdot \overline{v}\,\mathrm{d}S,$$

for all $u, v \in L^2(S)$ and $u, v \in L^2(\Sigma)$, respectively.

A bold-face font will be used to emphasize a vector or vector valued function, such as $\mathbf{x} \in \mathbb{R}^d$ or $\mathbf{u} : S \to \mathbb{C}^d$. With this in mind, we use the following bold-face convention for identifying vector-valued function spaces:

$$\mathbf{L}^p(S) := \left\{ \mathbf{v} : S \to \mathbb{C}^d \,\middle|\, v_i \in L^p(S) \text{ for all } i = 1, 2, \cdots, d \right\},$$
$$\mathbf{H}^\beta(S) := \left\{ \mathbf{v} : S \to \mathbb{C}^d \,\middle|\, v_i \in H^\beta(S) \text{ for } i = 1, 2, \cdots, d \right\}.$$

Chapter 2

Generalized Weak Coercivity of Reduced Wave Operators

2.1 Introduction to Generalized Weak Coercivity and Generalized Star-Shape Domains

This section introduces two new concepts; namely, generalized weak coercivity and generalized star-shape domains. As was already discussed, the Helmholtz-type operators are indefinite. Thus, one cannot expect the sesquilinear forms used to define the weak formulation of the Helmholtz-type operators to be coercive. In fact, in the case of Helmholtz-type operators one cannot even expect a weak coercivity property. Instead, for Helmholtz-type operators a generalized weak coercivity property of the form

$$\sup_{v\in V}\frac{|\operatorname{Im} a(u,v)|}{\|v\|_V}+\sup_{v\in W}\frac{|\operatorname{Re} a(u,v)|}{\|v\|_W}\geq C\|u\|_E \qquad \forall u\in E \tag{2.1}$$

takes the place of standard weak coercivity. Such a generalized weak coercivity property can be used to obtain a-priori wave-number explicit estimates for solutions of the three Helmholtz-type PDEs.

Generalized weak coercivity is also valuable in the development of novel discretization methods and linear solvers that are tailored to these Helmholtz-type problems. Specifically, the techniques employed in the proofs of the generalized weak coercivity properties can be useful in the development of absolutely stable discretization methods for Helmholtz-type problems (c.f. [42, 43, 44, 50]). That is, methods that are stable regardless of the mesh size h. Such an absolutely stable method for the elastic Helmholtz equation is developed and analyzed in Chapter 3. Absolutely stable methods are necessary to provide practical coarse mesh spaces, a key component for any multi-level method such as multi-grid or multi-level domain decomposition methods.

With multi-level methods in mind, the analysis of two-level domain decomposition for non-symmetric and indefinite linear problems in real valued Banach spaces is the focus of Chapter 5. The analysis in this chapter is based on a weak coercivity condition. It is believed that for non-symmetric and indefinite linear problems in complex valued Banach spaces the existing framework can be extended to include problems that satisfy a generalized weak coercivity condition in lieu of the standard weak coercivity condition. This is yet another motivation to study such generalized weak coercivity conditions.

The techniques used to obtain generalized weak coercivity properties of the Helmholtz-type operators are adapted from the techniques in [27, 43, 50]. The analysis found in these sources relies on a star-shape condition on the domain $\Omega \subset \mathbb{R}^d$. That is, for Ω there exists $\mathbf{x}_0 \in \Omega$ and a positive constant $c = c(\Omega)$ such that for $\boldsymbol{\alpha} = \mathbf{x} - \mathbf{x}_0$ the following condition holds:

$$\boldsymbol{\alpha} \cdot \mathbf{n} \geq c \qquad \text{on } \partial\Omega.$$

Practically, this constraint on the domain Ω is adequate when a scattering object is not present. In this case Ω is usually a truncation of a large or unbounded domain and can be chosen to meet this requirement. On the other hand, for a scattering

problem, a condition like this can be restrictive. This is due to the fact that portions of the boundary can be attributed to the scattering object.

With this in mind, this chapter is used to establish less restrictive generalized star-shape conditions on Ω for each Helmholtz-type operator. In particular, these new generalized star-shape conditions allow the existing analysis to hold while admitting more exotic geometry. These generalized star-shape domains are designed for each Helmholtz-type operator, separately. This idea does away with the "one-size fits all" nature of the standard star-shape condition and replaces it with "operator friendly" domain constraints.

This chapter is organized as follows: Sections 2.2–2.4 are used to tailor a generalized star-shape condition for each Helmholtz-type operator and prove a generalized weak coercivity condition for each operator. Section 2.5 applies the results of the previous sections to obtain stability estimates for each Helmholtz-type problem.

2.2 The Scalar Helmholtz Operator

First, a generalized star-shape domain for the scalar Helmholtz operator is defined. We consider an acoustic domain $\Omega = \Omega_+ \setminus \Omega_-$. Here, Ω_+ is the truncation of some unbounded acoustic medium and $\Omega_- \subset \Omega_+$ is some scattering object in the medium. For the existing analysis to hold using a classic star-shape condition one requires that Ω_+ and Ω_- are both star-shape domains with respect to the same point $\mathbf{x}_0 \in \Omega_-$. An example of such a domain is given in Figure 2.1.

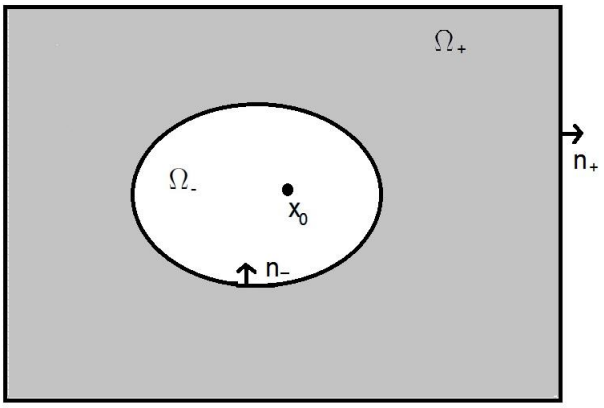


Figure 2.1: An example of a domain Ω of interest.

To generalize this idea, it is required that Ω is a domain such that there exists a vector field $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ that satisfies the following conditions:

$$\alpha_j = \alpha_j(x_j), \tag{2.2}$$

$$\boldsymbol{\alpha} \cdot \mathbf{n}_+ \geq c_+ > 0 \qquad \text{on } \partial\Omega_+, \tag{2.3}$$

$$-\boldsymbol{\alpha} \cdot \mathbf{n}_- \geq c_- > 0 \qquad \text{on } \partial\Omega_-, \tag{2.4}$$

$$|\boldsymbol{\alpha}| \leq R \qquad \text{in } \overline{\Omega}, \tag{2.5}$$

$$c_1 \leq \operatorname{div}(\boldsymbol{\alpha}) \leq c_2 \qquad \text{in } \Omega, \tag{2.6}$$

$$\min\left\{\frac{\partial \alpha_i}{\partial x_i}\right\} \geq c_3 > 0 \qquad \text{in } \Omega \text{ and } i = 1, 2 \ldots d, \tag{2.7}$$

$$c_1 - c_2 + 2c_3 \geq c_4 > 0 \qquad \text{in } \Omega, \tag{2.8}$$

where $\partial\Omega = \partial\Omega_+ \cup \partial\Omega_-$ and $\mathbf{n}_+$, $\mathbf{n}_-$ are the outward normal vectors to $\partial\Omega_+$ and $\partial\Omega_-$, respectively. A domain Ω that admits a vector field $\boldsymbol{\alpha}$ as described above will be called a ***generalized star-shape domain*** for the scalar Helmholtz equation.

Remark 2.2.1. *(a) This is a true generalization of the concept of a star-shape domain in the sense that any star-shape domain does satisfy the above properties. This includes the case discussed above (c.f. Figure 2.1).*

(b) The motivation for all of the generalized star-shape conditions introduced in this chapter comes from the use of Rellich identities for these specific Helmholtz-type operators. These identities are key in the techniques used to prove the generalized weak coercivity properties for each Helmholtz-type operator.

(c) It is conjectured that this generalization allows room for interesting computational domains that are not star-shape domains in the classical sense. At this point, no such examples are known, and this will be an item explored in future research.

For the rest of this section, Ω is assumed to be a generalized star-shape domain for the scalar Helmholtz equation. Recall that the generalized weak coercivity property is a property of the weak form of Helmholtz-type PDEs. Therefore, the weak form of (1.1)–(1.3) will need to be given. For the sake of completeness, the weak form is derived in the preceding lines. Begin by multiplying (1.1) by $v \in C^\infty(\Omega)$ and integrating over all Ω. To this, integration by parts and (1.2) are applied. These steps yield the following sequence of identities:

$$\begin{aligned}
-(\Delta u, v)_\Omega - k^2(u,v)_\Omega &= (f,v)_\Omega, \\
(\nabla u, \nabla v)_\Omega - \left\langle \frac{\partial u}{\partial \mathbf{n}}, v \right\rangle_{\partial\Omega} - k^2(u,v)_\Omega &= (f,v)_\Omega, \\
(\nabla u, \nabla v)_\Omega - \left\langle \frac{\partial u}{\partial \mathbf{n}_+}, v \right\rangle_{\partial\Omega_+} - \left\langle \frac{\partial u}{\partial \mathbf{n}_-}, v \right\rangle_{\partial\Omega_-} - k^2(u,v)_\Omega &= (f,v)_\Omega, \\
(\nabla u, \nabla v)_\Omega + \mathbf{i}k\langle u, v\rangle_{\partial\Omega_+} - \left\langle \frac{\partial u}{\partial \mathbf{n}_-}, v \right\rangle_{\partial\Omega_-} - k^2(u,v)_\Omega &= (f,v)_\Omega + \langle g, v\rangle_{\partial\Omega_+}.
\end{aligned}$$

From the above identity, we observe that an appropriate solution space for the weak formulation of (1.1)–(1.3) is given by

$$V := \left\{ u \in H^1(\Omega) \,\middle|\, u = 0 \text{ on } \partial\Omega_- \text{ and } \nabla u \in L^2(\partial\Omega) \right\}.$$

Now the weak form of (1.1)–(1.3) is defined in the following way: Find $u \in V$ such that

$$a(u,v) = (f,v)_\Omega + \langle g, v\rangle_{\partial\Omega_+} \qquad \forall v \in H^1(\Omega), \tag{2.9}$$

where $a(\cdot,\cdot)$ is a sesquilinear form defined on $V \times H^1(\Omega)$ given by

$$a(u,v) := (\nabla u, \nabla v)_\Omega - k^2(u,v)_\Omega + \mathbf{i}k\langle u, v\rangle_{\partial\Omega_+} - \left\langle \frac{\partial u}{\partial \mathbf{n}_-}, v \right\rangle_{\partial\Omega_-}. \tag{2.10}$$

The goal of this subsection is to prove a generalized weak coercivity condition (c.f. (2.1)) for the above sesquilinear form $a(\cdot,\cdot)$. To accomplish this goal we rely on the following Rellich identities quoted from [27]:

Lemma 2.2.2. *Let $u \in H^2(\Omega)$ and $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$. Then the following identity holds:*

$$-\operatorname{Re}(u, (\nabla u)\cdot\boldsymbol{\alpha})_\Omega = \frac{1}{2}\big(\operatorname{div}(\boldsymbol{\alpha}), |u|^2\big)_\Omega - \frac{1}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |u|^2\big\rangle_{\partial\Omega}.$$

Lemma 2.2.3. *Let $u \in H^2(\Omega)$ and $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$. Then the following identity holds:*

$$\begin{aligned}\operatorname{Re}(\nabla u, \nabla((\nabla u)\cdot\boldsymbol{\alpha}))_\Omega = -\frac{1}{2}\big(\operatorname{div}(\boldsymbol{\alpha}), |\nabla u|^2\big)_\Omega + \frac{1}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |\nabla u|^2\big\rangle_{\partial\Omega} \\ + \sum_{i=1}^d\sum_{j=1}^d \left(\frac{\partial u}{\partial x_i}, \frac{\partial \alpha_j}{\partial x_i}\frac{\partial u}{\partial x_j}\right)_\Omega.\end{aligned}$$

With these Rellich identities in hand, the following generalized weak coercivity property for the scalar Helmholtz operator is obtained:

Theorem 2.2.4. *Let $\Omega \subset \mathbb{R}^d$ be a generalized star-shape domain with $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.2)–(2.8). Then for any $u \in V$ the following generalized weak coercivity*

property holds for the sesquilinear form $a(\cdot,\cdot)$*:*

$$\sup_{v\in V}\frac{|\operatorname{Im} a(u,v)|}{\|v\|_E}+\sup_{v\in H^1(\Omega)}\frac{|\operatorname{Re} a(u,v)|}{|||v|||_{L^2(\Omega)}}\geq\frac{1}{\gamma}\|u\|_E,$$

where

$$\begin{aligned}
\gamma &:= \max\left\{\left[2(2c_2-4c_3+c_4)^2+16(k^2+1)R^2\right]^{\frac{1}{2}},M\right\},\\
M &:= 2\left(kR+\frac{kR^2}{c_+}+\frac{c_+}{k}\right),\\
|||u|||_{L^2(\Omega)} &:= \left(k^2c_4\|u\|^2_{L^2(\Omega)}+c_+\|u\|^2_{L^2(\partial\Omega_+)}\right)^{\frac{1}{2}},\\
\|u\|_E &:= \Big(k^2c_4\|u\|^2_{L^2(\Omega)}+c_4\|\nabla u\|^2_{L^2(\Omega)}+c_+\|u\|^2_{L^2(\partial\Omega_+)}+c_+\|\nabla u\|^2_{L^2(\partial\Omega_+)}\\
&\qquad +c_-\|\nabla u\|^2_{L^2(\partial\Omega_-)}\Big)^{\frac{1}{2}}.
\end{aligned}$$

Proof. In this proof, we assume that $u\in H^2(\Omega)\cap V$. This is possible because $u\in V$ can be approximated by a sequence of smooth functions that converge to u in $\|\cdot\|_E$. Once the result is obtained for $u\in H^2(\Omega)$ a limit can be applied to obtain the result for $u\in V$. For the sake of brevity, these details are suppressed in the steps to follow.

Begin by setting $v=u$ in (2.10) and taking the real and imaginary part separately. This yields the following identities:

$$\operatorname{Re} a(u,u)=\|\nabla u\|^2_{L^2(\Omega)}-k^2\|u\|^2_{L^2(\Omega)}, \tag{2.11}$$

$$\operatorname{Im} a(u,u)=k\|u\|^2_{L^2(\partial\Omega_+)}. \tag{2.12}$$

As will be a common theme for the analysis of all three Helmholtz-type problems, the indefiniteness of the scalar Helmholtz operator shows up here in an adverse way. That is, the signs of the terms on the right hand side of (2.11) are different. Thus the use of this one test function is not sufficient. For this reason, we employ a second test function $v=\nabla u\cdot\boldsymbol{\alpha}$, motivated by the above Rellich identities. Using this test

function in (2.10) yields

$$\operatorname{Re} a(u,v) = \operatorname{Re}(\nabla u, \nabla v)_\Omega - k^2 \operatorname{Re}(u,v)_\Omega - k \operatorname{Im}\langle u, v\rangle_{\partial\Omega_+} - \operatorname{Re}\left\langle \frac{\partial u}{\partial \mathbf{n}_-}, v \right\rangle_{\partial\Omega_-}. \tag{2.13}$$

We substitute the Rellich identities from Lemma 2.2.2 and Lemma 2.2.3 into (2.13) and rearrange the terms to get

$$\begin{aligned}
&\frac{k^2}{2}\big(\operatorname{div}(\boldsymbol{\alpha}), |u|^2\big)_\Omega - \frac{1}{2}\big(\operatorname{div}(\boldsymbol{\alpha}), |\nabla u|^2\big)_\Omega + \sum_{i=1}^{3}\left(\frac{\partial \alpha_i}{\partial x_i}, \left|\frac{\partial u_i}{\partial x_i}\right|^2\right)_\Omega \\
&\quad = \frac{k^2}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |u|^2\big\rangle_{\partial\Omega_+} - \frac{1}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |\nabla u|^2\big\rangle_{\partial\Omega_+} - \frac{1}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_-, |\nabla u|^2\big\rangle_{\partial\Omega_-} \\
&\qquad + \operatorname{Re}\left\langle \frac{\partial u}{\partial \mathbf{n}_-}, v\right\rangle_{\partial\Omega_-} + k\operatorname{Im}\langle u, v\rangle_{\partial\Omega_+} + \operatorname{Re} a(u,v) \\
&\quad = \frac{k^2}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |u|^2\big\rangle_{\partial\Omega_+} - \frac{1}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |\nabla u|^2\big\rangle_{\partial\Omega_+} + \frac{1}{2}\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}_-, |\nabla u|^2\big\rangle_{\partial\Omega_-} \\
&\qquad + k\operatorname{Im}\langle u, v\rangle_{\partial\Omega_+} + \operatorname{Re} a(u,v).
\end{aligned}$$

Notice that the first line above uses (2.2) and we get the last equality because $\nabla u = \frac{\partial u}{\partial \mathbf{n}_-}\mathbf{n}_-$ on $\partial\Omega_-$ since $u = 0$ on $\partial\Omega_-$.

Using the conditions on $\boldsymbol{\alpha}$ that are found in (2.3)–(2.6) and multiplying the previous inequality through by 2 produces the following inequality:

$$\begin{aligned}
&k^2 c_1 \|u\|^2_{L^2(\Omega)} - c_2\|\nabla u\|^2_{L^2(\Omega)} + 2c_3\|\nabla u\|^2_{L^2(\Omega)} \\
&\qquad \le k^2\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |u|^2\rangle_{\partial\Omega_+} - c_+\|\nabla u\|_{L^2(\partial\Omega_-)} - c_-\|\nabla u\|_{L^2(\partial\Omega_-)} \\
&\qquad\quad + 2k\operatorname{Im}\langle u, v\rangle_{\partial\Omega_+} + 2\operatorname{Re} a(u,v).
\end{aligned}$$

Adding $c_2 - 2c_3$ times (2.11) and $\frac{c_+}{k}$ times (2.12) to the above inequality yields

$$\begin{aligned}
&k^2(c_1 - c_2 + 2c_3)\|u\|^2_{L^2(\Omega)} + c_+\|u\|^2_{L^2(\partial\Omega_+)} \\
&\quad \le k^2\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |u|^2\rangle_{\partial\Omega_+} - c_+\|\nabla u\|_{L^2(\partial\Omega_-)} - c_-\|\nabla u\|_{L^2(\partial\Omega_-)} + 2k\,\mathrm{Im}\langle u, v\rangle_{\partial\Omega_+} \\
&\qquad + \mathrm{Re}\, a\,(u, 2v + (c_2 - 2c_3)u) + \frac{c_+}{k}\,\mathrm{Im}\, a(u,u).
\end{aligned}$$

Applying (2.8) and adding $\frac{c_4}{2}$ times (2.11) gives

$$\begin{aligned}
&\frac{k^2 c_4}{2}\|u\|^2_{L^2(\Omega)} + \frac{c_4}{2}\|\nabla u\|^2_{L^2(\Omega)} + c_+\|u\|^2_{L^2(\partial\Omega_+)} \\
&\quad \le k^2\langle \boldsymbol{\alpha}\cdot\mathbf{n}_+, |u|^2\rangle_{\partial\Omega_+} - c_+\|\nabla u\|_{L^2(\partial\Omega_-)} - c_-\|\nabla u\|_{L^2(\partial\Omega_-)} + 2k\,\mathrm{Im}\langle u, v\rangle_{\partial\Omega_+} \\
&\qquad + \mathrm{Re}\, a\left(u, 2v + \left(c_2 - 2c_3 + \frac{c_4}{2}\right)u\right) + \frac{c_+}{k}\,\mathrm{Im}\, a(u,u).
\end{aligned}$$

At this point, we apply Cauchy-Schwarz and Young's inequalities in conjunction with (2.12) to the previous inequality to obtain the following:

$$\begin{aligned}
&\frac{k^2 c_4}{2}\|u\|^2_{L^2(\Omega)} + \frac{c_4}{2}\|\nabla u\|^2_{L^2(\Omega)} + c_+\|u\|^2_{L^2(\partial\Omega_+)} \\
&\quad \le k^2 R\|u\|^2_{L^2(\partial\Omega_+)} - c_+\|\nabla u\|^2_{L^2(\partial\Omega_+)} - c_-\|\nabla u\|^2_{L^2(\partial\Omega_-)} + \frac{c_+}{k}\,\mathrm{Im}\, a(u,u) \\
&\qquad + \mathrm{Re}\, a\left(u, 2v + \left(c_2 - 2c_3 + \frac{c_4}{2}\right)u\right) + 2kR\|u\|_{L^2(\partial\Omega_+)}\|\nabla u\|_{L^2(\partial\Omega_+)} \\
&\quad \le k^2 R\|u\|^2_{L^2(\partial\Omega_+)} - c_+\|\nabla u\|^2_{L^2(\partial\Omega_+)} - c_-\|\nabla u\|^2_{L^2(\partial\Omega_-)} + \frac{c_+}{k}\,\mathrm{Im}\, a(u,u) \\
&\qquad + \mathrm{Re}\, a\left(u, 2v + \left(c_2 - 2c_3 + \frac{c_4}{2}\right)u\right) + \frac{k^2R^2}{c_+}\|u\|^2_{L^2(\partial\Omega_+)} + \frac{c_+}{2}\|\nabla u\|^2_{L^2(\partial\Omega_+)} \\
&\quad = -\frac{c_+}{2}\|\nabla u\|^2_{L^2(\partial\Omega_+)} - c_-\|\nabla u\|^2_{L^2(\partial\Omega_-)} + \left(kR + \frac{kR^2}{c_+} + \frac{c_+}{k}\right)\mathrm{Im}\, a(u,u) \\
&\qquad + \mathrm{Re}\, a\left(u, 2v + \left(c_2 - 2c_3 + \frac{c_4}{2}\right)u\right).
\end{aligned}$$

Consequently,

$$\|u\|^2_E \le M\big|\,\mathrm{Im}\, a(u,u)\big| + \big|\,\mathrm{Re}\, a\,(u, 4v + (4 + 2c_2 - 4c_3 + c_4)u)\,\big|. \tag{2.14}$$

Let $\hat{v} = 4v + (4 + 2c_2 - 4c_3 + c_4)u$. Putting this test function into $|||\cdot|||_{L^2(\Omega)}$ yields

$$\begin{aligned}
|||\hat{v}|||^2_{L^2(\Omega)} &\leq (2c_2 - 4c_3 + c_4)^2 \left[k^2 c_4 \|u\|^2_{L^2(\Omega)} + c_+ \|u\|^2_{L^2(\partial\Omega_+)}\right] \\
&\quad + 16R^2 \left[k^2 c_4 \|\nabla u\|^2_{L^2(\Omega)} + c_+ \|\nabla u\|^2_{L^2(\partial\Omega_+)}\right] \\
&\leq \left[2(2c_2 - 4c_3 + c_4)^2 + 16(k^2+1)R^2\right] \|u\|^2_E.
\end{aligned}$$

Finally, this inequality along with (2.14) implies that

$$\begin{aligned}
&\sup_{v\in V} \frac{|\operatorname{Im} a(u,v)|}{\|v\|_E} + \sup_{v\in H^1(\Omega)} \frac{|\operatorname{Re} a(u,v)|}{|||v|||_{L^2(\Omega)}} \\
&\quad \geq \frac{|\operatorname{Im} a(u,u)|}{\|u\|_E} + \frac{|\operatorname{Re} a(u,\hat{v})|}{|||\hat{v}|||_{L^2(\Omega)}} \\
&\quad \geq \frac{|\operatorname{Im} a(u,u)|}{\|u\|_E} + \frac{|\operatorname{Re} a(u,\hat{v})|}{[2(2c_2 - 4c_3 + c_4)^2 + 16(k^2+1)R^2]^{\frac{1}{2}} \|u\|_E} \\
&\quad \geq \frac{1}{\gamma} \frac{M|\operatorname{Im} a(u,u)| + |\operatorname{Re} a(u,\hat{v})|}{\|u\|_E} \\
&\quad \geq \frac{1}{\gamma} \|u\|_E.
\end{aligned}$$

Hence the generalized weak coercivity condition holds. □

2.3 The Elastic Helmholtz Operator

In this section, the focus is turned to the elastic Helmholtz operator. This operator is similar to the scalar Helmholtz operator. Due to this similarity, the analysis for the elastic Helmholtz operator should follow that of the scalar Helmholtz operator. This section is restricted to the case in which $\partial\Omega_- = \emptyset$ and thus, $\partial\Omega = \partial\Omega_+$, where Ω is the elastic medium. Such a restriction is made to compensate for the added difficulty in working with vector-valued functions. Ω is defined to be a generalized star-shape

domain for which there exists $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ such that the following properties hold:

$$\alpha_j = \alpha_j(x_j), \tag{2.15}$$

$$\boldsymbol{\alpha} \cdot \mathbf{n} \geq c_+ > 0 \qquad \text{on } \partial\Omega, \tag{2.16}$$

$$|\boldsymbol{\alpha}| \leq R \qquad \text{in } \overline{\Omega}, \tag{2.17}$$

$$\frac{\partial \alpha_i}{\partial x_i} = c_1 > 0 \qquad \text{in } \Omega \text{ and } i = 1, 2 \ldots d. \tag{2.18}$$

As was the case in Section 2.2, the analysis of this section relies on Rellich identities for the elastic Helmholtz operator. These Rellich identities are the reason behind the constraints placed on the domain. Unfortunately, the Rellich identities for the elastic Helmholtz operator do not yield as much as those for the scalar Helmholtz operator. This is mainly a result of the increase in complexity when moving from scalar-valued functions to vector-valued functions. For this reason the generalized star-shape domain criterion for the elastic Helmholtz operator is more restrictive than that of the scalar Helmholtz operator. A less restrictive domain might be possible, but different techniques will be needed to attain a generalized weak coercivity condition.

Now with a generalized star-shape domain defined for the elastic Helmholtz operator, a weak formulation of (1.4)–(1.5) will be derived. To begin, multiply (1.4) with a smooth function $\mathbf{v} \in \mathbf{C}^\infty(\Omega)$ and integrate over Ω to obtain

$$-\omega^2 \rho(\mathbf{u}, \mathbf{v})_\Omega - (\mathbf{div}\,(\sigma(\mathbf{u})), \mathbf{v})_\Omega = (\mathbf{f}, \mathbf{v})_\Omega. \tag{2.19}$$

Since $\sigma(\mathbf{u})$ is symmetric the following product rule for the divergence holds:

$$\mathbf{div}\,(\sigma(\mathbf{u})\overline{\mathbf{v}}) = \mathbf{div}\,(\sigma(\mathbf{u}))\overline{\mathbf{v}} + \sigma(\mathbf{u}) : \boldsymbol{\nabla}\overline{\mathbf{v}}.$$

This identity together with the divergence theorem yields

$$\begin{aligned} -(\mathbf{div}\,(\sigma(\mathbf{u})),\mathbf{v})_\Omega &= -\int_\Omega \mathbf{div}\,(\sigma(\mathbf{u})\overline{\mathbf{v}})\mathrm{d}\mathbf{x} + (\sigma(\mathbf{u}),\boldsymbol{\nabla}\mathbf{v})_\Omega \\ &= -\langle\sigma(\mathbf{u})\mathbf{n},\mathbf{v}\rangle_{\partial\Omega} + (\sigma(\mathbf{u}),\boldsymbol{\nabla}\mathbf{v})_\Omega. \end{aligned} \tag{2.20}$$

From Lemma 3 of [27] one obtains the following useful identity:

$$\sigma(\mathbf{u}) : \boldsymbol{\nabla}\overline{\mathbf{v}} = \lambda \mathrm{div}\,\mathbf{u}\mathrm{div}\,\overline{\mathbf{v}} + 2\mu\varepsilon(\mathbf{u}) : \varepsilon(\overline{\mathbf{v}}).$$

With this identity (2.20) becomes

$$-(\mathbf{div}\,(\sigma(\mathbf{u})),\mathbf{v})_\Omega = \lambda(\mathrm{div}\,\mathbf{u},\mathrm{div}\,\mathbf{v})_\Omega + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_\Omega - \langle\sigma(\mathbf{u})\mathbf{n},\mathbf{v}\rangle_{\partial\Omega}\,. \tag{2.21}$$

Applying this integration by parts formula along with (1.5) to (2.19) gives

$$\lambda(\mathrm{div}\,\mathbf{u},\mathrm{div}\,\mathbf{v})_\Omega + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_\Omega - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega + i\omega\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega} = (\mathbf{f},\mathbf{v})_\Omega + \langle\mathbf{g},\mathbf{v}\rangle_{\partial\Omega}.$$

Thus, a weak formulation of the elastic Helmholtz equations (1.4) - (1.5) is given by: find $\mathbf{u} \in \mathbf{H}^1(\Omega)$ such that

$$a(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v})_\Omega + \langle\mathbf{g},\mathbf{v}\rangle_{\partial\Omega} \qquad \forall\mathbf{v} \in \mathbf{H}^1(\Omega), \tag{2.22}$$

where $a(\cdot,\cdot)$ is defined on $\mathbf{H}^1(\Omega) \times \mathbf{H}^1(\Omega)$ by

$$a(\mathbf{u},\mathbf{v}) := \lambda(\mathrm{div}\,\mathbf{u},\mathrm{div}\,\mathbf{v})_\Omega + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_\Omega - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega + i\omega\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega}. \tag{2.23}$$

Now that the weak formulation is established on a generalized star-shape domain, the focus of this section shifts to obtaining a generalized weak coercivity property for $a(\cdot,\cdot)$. As stated previously, this will require the use of some Rellich identities for the

elastic Helmholtz operator. These Rellich identities were established in [27] and are quoted below as the following two lemmas.

Lemma 2.3.1. *For* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *and* $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$*, the following identity holds:*

$$\begin{aligned}
&\lambda\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\operatorname{div}\mathbf{u}|^2\rangle_{\partial\Omega} + 2\mu\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\varepsilon(\mathbf{u})|^2\rangle_{\partial\Omega} \\
&\quad = \lambda(\operatorname{div}\boldsymbol{\alpha}, |\operatorname{div}\mathbf{u}|^2)_\Omega + 2\mu(\operatorname{div}\boldsymbol{\alpha}, |\varepsilon(\mathbf{u})|^2)_\Omega \\
&\qquad + 2\lambda \operatorname{Re}\left(\operatorname{div}\mathbf{u}, \operatorname{div}\left((\boldsymbol{\nabla}\mathbf{v})\boldsymbol{\alpha}\right)\right)_\Omega + 4\mu \operatorname{Re}\left(\varepsilon(\mathbf{u}), \varepsilon((\boldsymbol{\nabla}\mathbf{v})\boldsymbol{\alpha})\right)_\Omega \\
&\qquad - 2\lambda \operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\operatorname{div}\mathbf{u}, \frac{\partial\alpha_j}{\partial x_i}\frac{\partial u_i}{\partial x_j}\right)_\Omega \\
&\qquad - 2\mu \operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial\alpha_k}{\partial x_j}\frac{\partial u_i}{\partial x_k}\right)_\Omega .
\end{aligned}$$

Lemma 2.3.2. *For* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *and* $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$*, the following identity holds:*

$$(\operatorname{div}\boldsymbol{\alpha}, |\mathbf{u}|^2)_\Omega = \langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}|^2\rangle_{\partial\Omega} - 2\operatorname{Re}(\mathbf{u}, (\boldsymbol{\nabla}\mathbf{u})\boldsymbol{\alpha})_\Omega.$$

Similar to other analysis involving the stress tensor $\sigma(\cdot)$, it is necessary to use the well-known Korn's inequality to obtain the desired generalized weak coercivity property. It is stated here as a lemma. For a proof, see [63].

Lemma 2.3.3. *There exists a positive constant* K *such that for any* $\mathbf{v} \in H^1(\Omega)$ *the following inequality holds:*

$$\|\mathbf{v}\|_{H^1(\Omega)} \leq K\Big[\|\varepsilon(\mathbf{u})\|_{L^2(\Omega)} + \|v\|_{L^2(\Omega)}\Big].$$

As was the case in Section 2.2, the analysis used in this section will follow closely to that in [27]. In [27] the authors found it necessary to use a Korn-type inequality on the boundary $\partial\Omega$ to obtain estimates that are optimal in terms of the frequency ω. This Korn-type inequality still remains a conjecture. As stated in [27], this conjecture

is believed to hold for the solution of the elastic Helmholtz problem since a similar result is shown in [28] for the solution of the Lamé systems of elastostatics.

Conjecture 2.3.4. *There exists a positive constant $\tilde{K}$ such that for any $\mathbf{u} \in \mathbf{H}^2(\Omega)$ the following Korn-type inequality holds:*

$$\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} \leq \tilde{K}\left[\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + \|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}\right]. \tag{2.24}$$

With these technical lemmas in hand, we have all the tools necessary to prove a generalized weak coercivity property on $a(\cdot,\cdot)$. This will be done in two steps. First, we prove a preliminary result that does not make use of Conjecture 2.3.4. Next, we prove a generalized weak coercivity property for $\mathbf{u}$ in a more restrictive function space (i.e. the space on which Conjecture 2.3.4 holds).

Lemma 2.3.5. *Let Ω be a generalized star-shape domain such that there exists $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.15)–(2.18). Then for all $\mathbf{u} \in H^2(\Omega)$ and $\epsilon > 0$ there holds*

$$\begin{aligned}\|\mathbf{u}\|^2_E \leq \epsilon\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} - \left(c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}\right)\\ + \operatorname{Re} a(\mathbf{u}, 2(\boldsymbol{\nabla}\mathbf{u})\boldsymbol{\alpha} + (1-d)c_1\mathbf{u}) + \frac{1}{c_A}\left(R\omega\rho + \frac{R^2\omega C_A}{\epsilon} + 2\right)\operatorname{Im} a(u,u),\end{aligned}$$

where $\|\cdot\|_E$ is defined by

$$\begin{aligned}\|\mathbf{u}\|^2_E := c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)} + 2c_1\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)}\\ + c_+\mu\|\mathbf{u}\|_{L^2(\partial\Omega)} + c_+\lambda\big\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}.\end{aligned}$$

Proof. In a manner similar to the proof of Theorem 2.2.4, setting $\mathbf{v} = \mathbf{u}$ in (2.23) and taking both real and imaginary parts separately we get

$$\operatorname{Re} a(\mathbf{u},\mathbf{u}) = \lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)} + 2\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)} - \omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)}, \tag{2.25}$$

$$\operatorname{Im} a(\mathbf{u},\mathbf{u}) = \omega\langle A\mathbf{u},\mathbf{u}\rangle_{\partial\Omega}. \tag{2.26}$$

Again, it is clear that this first test function alone cannot yield the desired result because of the sign difference in (2.25). For this reason, a second test function, motivated by our Rellich identities, will be selected. Let $\mathbf{v} = (\boldsymbol{\nabla}\mathbf{u})\boldsymbol{\alpha}$ for the rest of this proof. Substituting this test function into (2.23) and multiplying through by 2 gives

$$\begin{aligned} 2\operatorname{Re} a(\mathbf{u},\mathbf{v}) = {} & 2\lambda(\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v})_\Omega + 4\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_\Omega - 2\operatorname{Re}\omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\ & - 2\omega\operatorname{Im}\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega}. \end{aligned}$$

By the Rellich identities for the elastic Helmholtz operator (i.e. Lemmas 2.3.1 and 2.3.2), we get

$$\begin{aligned} 2\operatorname{Re} a(\mathbf{u},\mathbf{v}) = {} & \lambda\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\operatorname{div}\mathbf{u}|^2\big\rangle_{\partial\Omega} + 2\mu\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\varepsilon(\mathbf{u})|^2\big\rangle_{\partial\Omega} - \lambda\big(\operatorname{div}\boldsymbol{\alpha}, |\operatorname{div}\mathbf{u}|^2\big)_\Omega \\ & - 2\mu\big(\operatorname{div}\boldsymbol{\alpha}, |\varepsilon(\mathbf{u})|^2\big)_\Omega + 2\lambda\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\operatorname{div}\mathbf{u}, \frac{\partial\alpha_j}{\partial x_i}\frac{\partial u_i}{\partial x_j}\right)_\Omega \\ & + 2\mu\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial\alpha_k}{\partial x_j}\frac{\partial u_i}{\partial x_k}\right)_\Omega \\ & - \omega^2\rho\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}|^2\big\rangle_\Omega + \omega^2\rho\big(\operatorname{div}\boldsymbol{\alpha}, |\mathbf{u}|^2\big)_\Omega - 2\omega\operatorname{Im}\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega}. \end{aligned}$$

Equivalently,

$$\begin{aligned} & \omega^2\rho\big(\operatorname{div}\boldsymbol{\alpha}, |\mathbf{u}|^2\big)_\Omega - \lambda\big(\operatorname{div}\boldsymbol{\alpha}, |\operatorname{div}\mathbf{u}|^2\big)_\Omega - 2\mu\big(\operatorname{div}\boldsymbol{\alpha}, |\varepsilon(\mathbf{u})|^2\big)_\Omega \\ & \qquad + 2\lambda\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\operatorname{div}\mathbf{u}, \frac{\partial\alpha_j}{\partial x_i}\frac{\partial u_i}{\partial x_j}\right)_\Omega \\ & \qquad + 2\mu\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial\alpha_k}{\partial x_j}\frac{\partial u_i}{\partial x_k}\right)_\Omega \\ & = \omega^2\rho\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}|^2\big\rangle_{\partial\Omega} - \lambda\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\operatorname{div}\mathbf{u}|^2\big\rangle_{\partial\Omega} - 2\mu\big\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\varepsilon(\mathbf{u})|^2\big\rangle_{\partial\Omega} \\ & \qquad + 2\omega\operatorname{Im}\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega} + 2\operatorname{Re} a(\mathbf{u},\mathbf{v}). \end{aligned}$$

Applying the properties of $\boldsymbol{\alpha}$ from (2.15)–(2.18) to the above identity produces

$$\begin{aligned}
&dc_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} - dc_1\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)} - 2dc_1\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)} \\
&\qquad + 2c_1\lambda\operatorname{Re}\sum_{i=1}^{d}\left(\operatorname{div}\mathbf{u}, \frac{\partial u_i}{\partial x_i}\right)_\Omega + 2c_1\mu\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial u_i}{\partial x_j}\right)_\Omega \\
&\le R\omega^2\rho\big\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\lambda\big\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} - 2c_+\mu\big\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\
&\qquad + 2\omega\operatorname{Im}\langle A\mathbf{u}, \mathbf{v}\rangle_{\partial\Omega} + 2\operatorname{Re} a(\mathbf{u}, \mathbf{v}).
\end{aligned} \tag{2.27}$$

At this stage, we appeal to the following simplifications:

$$\begin{aligned}
2c_1\lambda\operatorname{Re}\sum_{i=1}^{d}\left(\operatorname{div}\mathbf{u}, \frac{\partial u_i}{\partial x_i}\right)_\Omega &= 2c_1\lambda\operatorname{Re}\left(\operatorname{div}\mathbf{u}, \sum_{i=1}^{d}\frac{\partial u_i}{\partial x_i}\right)_\Omega \\
&= 2c_1\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)},
\end{aligned}$$

and

$$\begin{aligned}
&2c_1\mu\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial u_i}{\partial x_j}\right)_\Omega \\
&= c_1\mu\operatorname{Re}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial u_i}{\partial x_j}\right)_\Omega + \sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial u_j}{\partial x_i}\right)_\Omega\right] \\
&= c_1\mu\operatorname{Re}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}, \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)_\Omega \\
&= 4c_1\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)}.
\end{aligned}$$

We apply these simplifications to (2.27) to get

$$\begin{aligned}
&dc_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} + (2-d)c_1\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)} + 2(2-d)c_1\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)} \\
&\qquad \le R\omega^2\rho\big\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\lambda\big\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} - 2c_+\mu\big\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\
&\qquad\qquad + 2\omega\operatorname{Im}\langle A\mathbf{u}, \mathbf{v}\rangle_{\partial\Omega} + 2\operatorname{Re} a(\mathbf{u}, \mathbf{v}).
\end{aligned}$$

Note, for $d > 2$, the terms with coefficient $(2-d)$ are negative. To eliminate these terms from the left hand side (LHS), add $(d-2)c_1$ times (2.25) to the above identity to obtain the following:

$$\begin{aligned} 2c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} \leq R\omega^2\rho\big\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\lambda\big\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} - 2c_+\mu\big\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\ + 2\omega\operatorname{Im}\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega} + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (2-d)c_1\mathbf{u}). \end{aligned}$$

We notice that (2.26) allows us control over the terms on the right hand side (RHS) involving $\|\mathbf{u}\|_{L^2(\partial\Omega)}$. With this in mind, we apply the Cauchy-Schwarz inequality along with Young's inequality to $2\omega\operatorname{Im}\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega}$ to obtain

$$\begin{aligned} &2c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} \\ &\quad\leq R\omega^2\rho\big\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\lambda\big\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} - 2c_+\mu\big\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\ &\qquad + 2\omega\|A\mathbf{u}\|_{L^2(\partial\Omega)}\|\mathbf{v}\|_{L^2(\partial\Omega)} + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (2-d)c_1\mathbf{u}) \\ &\quad\leq R\omega^2\rho\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} - 2c_+\mu\big\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\ &\qquad + 2R\omega\|A\mathbf{u}\|_{L^2(\partial\Omega)}\|\boldsymbol{\nabla}\mathbf{u}\|_{L^2(\partial\Omega)} + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (2-d)c_1\mathbf{u}) \\ &\quad\leq R\omega^2\rho\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} - 2c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\ &\qquad + \frac{R^2\omega^2 C_A}{\epsilon}\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + \epsilon\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (2-d)c_1\mathbf{u}). \end{aligned}$$

The term $\epsilon\|\boldsymbol{\nabla}\mathbf{u}\|_{L^2(\partial\Omega)}$ will be controlled later using the boundary Korn-type inequality (c.f. Conjecture 2.3.4). With this in mind, we add and subtract

$2c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)}$ and apply (2.26) to yield

$$\begin{aligned}
2c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} + c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)} &\le -c_+\lambda\big\|\operatorname{div}\mathbf{u}\big\|^2_{L^2(\partial\Omega)} - c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\
&\quad + \epsilon\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} - \big(c_+\mu\|\mathbf{u}\|^2_{\partial\Omega} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{\partial\Omega}\big) \\
&\quad + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (2-d)c_1\mathbf{u}) + \left(R\omega^2\rho + \frac{R^2\omega^2 C_A}{\epsilon} + 2c_+\mu\right)\|\mathbf{u}\|^2_{L^2(\partial\Omega)} \\
&\le -c_+\lambda\big\|\operatorname{div}\mathbf{u}\big\|^2_{L^2(\partial\Omega)} - c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} \\
&\quad + \epsilon\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} - \big(c_+\mu\|\mathbf{u}\|^2_{\partial\Omega} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{\partial\Omega}\big) \\
&\quad + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (2-d)c_1\mathbf{u}) + \frac{1}{c_A}\left(R\omega\rho + \frac{R^2\omega C_A}{\epsilon} + \frac{2c_+\mu}{\omega}\right)\operatorname{Im} a(\mathbf{u},\mathbf{u}).
\end{aligned}$$

To obtain a norm on $H^1(\Omega)$ on the LHS, we subtract (2.25) from the above inequality, and move some terms to the LHS to get

$$\begin{aligned}
c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)} + 2c_1\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)} &\\
\quad + c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\lambda\big\|\operatorname{div}\mathbf{u}\big\|^2_{L^2(\partial\Omega)} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)} &\\
\le \epsilon\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} - \big(c_+\mu\|\mathbf{u}\|^2_{\partial\Omega} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{\partial\Omega}\big) &\\
\quad + \operatorname{Re} a(\mathbf{u}, 2\mathbf{v} + (1-d)c_1\mathbf{u}) + \frac{1}{c_A}\left(R\omega\rho + \frac{R^2\omega C_A}{\epsilon} + \frac{2c_+\mu}{\omega}\right)\operatorname{Im} a(\mathbf{u},\mathbf{u}).&
\end{aligned}$$

Therefore, the assertion holds. □

In order to prove a generalized weak coercivity property on $a(\cdot,\cdot)$, an estimate to control the term $\epsilon\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)}$ on the RHS of the inequality in Lemma 2.3.5 needs to be established. This is where a Korn-type inequality on the boundary would be helpful. With this in mind, we introduce the special function spaces

$$\begin{aligned}
\mathbf{V} &:= \Big\{\mathbf{v} \in \mathbf{H}^1(\Omega)\Big|\ \varepsilon(\mathbf{v}) \in \mathbf{L}^2(\partial\Omega)\Big\}, \\
\mathbf{V}_{\tilde{K}} &:= \Big\{\mathbf{u} \in \mathbf{V}\Big|\ \|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} \le \tilde{K}\left[\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + \|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}\right]\Big\},
\end{aligned}$$

where $\tilde{K}$ is a positive constant. With the help of these spaces, the following generalized weak coercivity property holds.

Theorem 2.3.6. *Let Ω be a domain for which there exists $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.15)–(2.18). Then for any $\tilde{K} > 0$ and $\mathbf{u} \in \mathbf{V}_{\tilde{K}}$ the following inequality holds:*

$$\sup_{\mathbf{v}\in\mathbf{V}} \frac{|\operatorname{Im} a(\mathbf{u},\mathbf{v})|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\mathbf{H}^1(\Omega)} \frac{|\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{|||\mathbf{v}|||_{L^2(\Omega)}} \geq \frac{1}{\gamma}\|\mathbf{u}\|_E,$$

where

$$\begin{aligned}
\gamma &:= \max\left\{ \left[4R^2K\left(1+\frac{\omega^2\rho}{2\mu}\right) + 4R^2\tilde{K} + (1-d)^2c_1^2\right]^{\frac{1}{2}}, M\right\},\\
M &:= \frac{1}{c_A}\left(R\omega\rho + \frac{R^2\omega C_A\tilde{K}}{c_+\mu} + \frac{2c_+\mu}{\omega}\right),\\
\|\mathbf{u}\|_E^2 &:= c_1\omega^2\rho\|\mathbf{u}\|_{L^2(\Omega)}^2 + c_1\lambda\|\operatorname{div}\mathbf{u}\|_{L^2(\Omega)}^2 + 2c_1\mu\|\varepsilon(\mathbf{u})\|_{L^2(\Omega)}^2\\
&\qquad + c_+\mu\|\mathbf{u}\|_{L^2(\partial\Omega)} + c_+\lambda\big\|\operatorname{div}\mathbf{u}\|_{L^2(\partial\Omega)}^2 + c_+\mu\|\varepsilon(\mathbf{u})\|_{L^2(\partial\Omega)}^2\\
|||\mathbf{u}|||_{L^2(\Omega)}^2 &:= c_1\omega^2\rho\|\mathbf{u}\|_{L^2(\Omega)}^2 + c_+\mu\|\mathbf{u}\|_{L^2(\partial\Omega)}^2.
\end{aligned}$$

Proof. As was the case in the proof of Theorem 2.2.4, we only give a proof for $\mathbf{u} \in \mathbf{V}_{\tilde{K}} \cap \mathbf{H}^2(\Omega)$. After we prove the result for this more restrictive case, we can use a limiting process to yield the result for $\mathbf{u} \in \mathbf{V}_{\tilde{K}}$.

By Lemma 2.3.5 with $\epsilon = \frac{c_+\mu}{\tilde{K}}$ we obtain

$$\begin{aligned}
\|\mathbf{u}\|_E^2 \leq \frac{c_+\mu}{\tilde{K}}\|\boldsymbol{\nabla}\mathbf{u}\|_{L^2(\partial\Omega)}^2 - \left(c_+\mu\|\mathbf{u}\|_{L^2(\partial\Omega)}^2 + c_+\mu\|\varepsilon(\mathbf{u})\|_{L^2(\partial\Omega)}^2\right)\\
+ \operatorname{Re} a(\mathbf{u},\hat{\mathbf{v}}) + \frac{1}{c_A}\left(R\omega\rho + \frac{R^2\omega C_A\tilde{K}}{c_+\mu} + \frac{2c_+\mu}{\omega}\right)\operatorname{Im} a(u,u),
\end{aligned}$$

where $\hat{\mathbf{v}} := (2(\boldsymbol{\nabla}\mathbf{u})\boldsymbol{\alpha} + (1-d)c_1\mathbf{u})$. With this choice of ϵ, there holds

$$\begin{aligned}
&\frac{c_+\mu}{\tilde{K}}\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)} - \big(c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}\big)\\
&\qquad \le c_+\mu\big(\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + \|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}\big) - \big(c_+\mu\|\mathbf{u}\|^2_{\partial\Omega} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}\big)\\
&\qquad \le 0.
\end{aligned}$$

Thus,

$$\|\mathbf{u}\|^2_E \le \big|\operatorname{Re} a(\mathbf{u},\hat{\mathbf{v}})\big| + M\big|\operatorname{Im} a(\mathbf{u},\mathbf{u})\big|. \tag{2.28}$$

Next, by the definitions of $\|\cdot\|_E$ and $|||\cdot|||_{L^2(\Omega)}$ we get

$$\begin{aligned}
|||\hat{\mathbf{v}}|||^2_{L^2(\Omega)} &= c_1\omega^2\rho\|\hat{\mathbf{v}}\|^2_{L^2(\Omega)} + c_+\mu\|\hat{\mathbf{v}}\|^2_{L^2(\partial\Omega)} \\
&\le 4c_1R^2\omega^2\rho\|\boldsymbol{\nabla}\mathbf{u}\|_{L^2(\Omega)} + 4R^2c_+\mu\|\boldsymbol{\nabla}\mathbf{u}\|^2_{L^2(\partial\Omega)}\\
&\qquad + (1-d)^2c_1^3\omega^2\rho\|\mathbf{u}\|_{L^2(\Omega)} + (1-d)^2c_1^2c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)}\\
&\le \left[4R^2K\left(1+\frac{\omega^2\rho}{2\mu}\right) + 4R^2\tilde{K} + (1-d)^2c_1^2\right]\|\mathbf{u}\|^2_E\\
&\le \gamma^2\|\mathbf{u}\|^2_E.
\end{aligned} \tag{2.29}$$

It follows from (2.28) and (2.29) that

$$\begin{aligned}
\sup_{\mathbf{v}\in\mathbf{V}}\frac{|\operatorname{Im} a(\mathbf{u},\mathbf{v})|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\mathbf{H}^1(\Omega)}\frac{|\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{|||\mathbf{v}|||_{L^2(\Omega)}} &\ge \frac{|\operatorname{Im} a(\mathbf{u},\mathbf{u})|}{\|\mathbf{u}\|_E} + \frac{|\operatorname{Re} a(\mathbf{u},\hat{\mathbf{v}})|}{|||\hat{\mathbf{v}}|||_{L^2(\Omega)}}\\
&\ge \frac{M|\operatorname{Im} a(\mathbf{u},\mathbf{u})|}{M\|\mathbf{u}\|_E} + \frac{|\operatorname{Re} a(\mathbf{u},\hat{\mathbf{v}})|}{\gamma\|\mathbf{u}\|_E}\\
&\ge \frac{M|\operatorname{Im} a(\mathbf{u},\mathbf{u})| + |\operatorname{Re} a(\mathbf{u},\hat{\mathbf{v}})|}{\gamma\|\mathbf{u}\|_E}\\
&\ge \frac{1}{\gamma}\|\mathbf{u}\|_E.
\end{aligned}$$

Thus, the desired generalized weak coercivity property holds. □

2.4 The Time-Harmonic Maxwell Operator

As was the case in Sections 2.2 and 2.3, we begin this section by defining a generalized star-shape domain that is specially suited to the time-harmonic Maxwell operator. In this section, the restriction that $d = 3$ (i.e. $\Omega \subset \mathbb{R}^3$) is assumed. Similar to section 2.3, the domain Ω is also restricted to the case where a scattering portion of the boundary is not present. That is, $\partial\Omega_- = \emptyset$ and thus $\partial\Omega_+ = \partial\Omega$. Lastly, in this section, Ω is defined to be a generalized star-shape domain such that there exists $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying the following properties:

$$\alpha_i = \alpha_i(x_i) \qquad \text{in } \Omega \text{ and for } \ i = 1, 2, 3, \tag{2.30}$$

$$|\boldsymbol{\alpha}| \leq R \qquad \text{in } \Omega, \tag{2.31}$$

$$\boldsymbol{\alpha} \cdot \mathbf{n} \geq c_+ > 0 \qquad \text{on } \partial\Omega, \tag{2.32}$$

$$\operatorname{div} \boldsymbol{\alpha} - 2 \max_{i=1,2,3} \left\{ \frac{\partial \alpha_i}{\partial x_i} \right\} \geq c_1 > 0 \qquad \text{in } \Omega. \tag{2.33}$$

Maxwell's equations are defined using the **curl** operator. For this reason, some special function spaces need to be defined on Ω before a weak formulation can be defined.

$$\begin{aligned}
\mathbf{H}(\mathbf{curl}\,, \Omega) &:= \Big\{ \mathbf{v} \in \mathbf{L}^2(\Omega) \Big| \ \mathbf{curl}\, \mathbf{v} \in \mathbf{L}^2(\Omega) \Big\}, \\
\mathbf{H}(\operatorname{div}\,, \Omega) &:= \Big\{ \mathbf{v} \in \mathbf{L}^2(\Omega) \Big| \ \operatorname{div} \mathbf{v} \in L^2(\Omega) \Big\}, \\
\mathbf{H}(\operatorname{div}_0, \Omega) &:= \Big\{ \mathbf{v} \in \mathbf{L}^2(\Omega) \Big| \ \operatorname{div} \mathbf{v} = 0 \Big\}, \\
\boldsymbol{\mathcal{V}} &:= \Big\{ \mathbf{v} \in \mathbf{H}(\mathbf{curl}\,, \Omega) \Big| \ \mathbf{v} \in \mathbf{L}^2(\Omega) \Big\}, \\
\hat{\boldsymbol{\mathcal{V}}} &:= \Big\{ \mathbf{v} \in \mathbf{H}(\mathbf{curl}\,, \Omega) \Big| \ \mathbf{curl}\, \mathbf{v} \in \mathbf{H}(\mathbf{curl}\,, \Omega) \text{ and } \mathbf{v} \in \mathbf{H}(\mathbf{curl}\,, \partial\Omega) \Big\}.
\end{aligned}$$

Following the example set forth in Sections 2.2 and 2.3, the weak formulation of (1.6)–(1.7) is derived below. Multiplying (1.6) with a smooth test function $\mathbf{v} \in \mathbf{C}^\infty(\Omega)$

and integrating over the domain Ω gives

$$(\mathbf{curl}\,\mathbf{curl}\,\mathbf{E},\mathbf{v})_\Omega - k^2(\mathbf{E},\mathbf{v})_\Omega = (\mathbf{f},\mathbf{v})_\Omega. \tag{2.34}$$

In order to derive the appropriate integration by parts formula, we start with the following identity:

$$\mathrm{div}\,(\mathbf{curl}\,\mathbf{E}\times\overline{\mathbf{v}}) = \mathbf{curl}\,\mathbf{curl}\,\mathbf{E}\cdot\overline{\mathbf{v}} - \mathbf{curl}\,\mathbf{E}\cdot\mathbf{curl}\,\overline{\mathbf{v}}. \tag{2.35}$$

This identity is easily obtained from the well-known vector calculus identity:

$$\mathrm{div}\,(\mathbf{a}\times\mathbf{b}) = \mathbf{b}\cdot\mathbf{curl}\,\mathbf{a} - \mathbf{a}\cdot(\mathbf{curl}\,\mathbf{b}). \tag{2.36}$$

(2.35) along with the divergence theorem yields the following integration by parts identity:

$$\begin{aligned}(\mathbf{curl}\,\mathbf{curl}\,\mathbf{E},\mathbf{v})_\Omega &= (\mathbf{curl}\,\mathbf{E},\mathbf{curl}\,\mathbf{v})_\Omega + \int_\Omega \mathrm{div}\,(\mathbf{curl}\,\mathbf{E}\times\overline{\mathbf{v}})\mathrm{d}\mathbf{x} \\ &= (\mathbf{curl}\,\mathbf{E},\mathbf{curl}\,\mathbf{v})_\Omega + \big\langle \mathbf{curl}\,\mathbf{E}\times\overline{\mathbf{v}},\mathbf{n}\big\rangle_{\partial\Omega} \\ &= (\mathbf{curl}\,\mathbf{E},\mathbf{curl}\,\mathbf{v})_\Omega - \big\langle \mathbf{curl}\,\mathbf{E}\times\mathbf{n},\mathbf{v}\big\rangle_{\partial\Omega} \\ &= (\mathbf{curl}\,\mathbf{E},\mathbf{curl}\,\mathbf{v})_\Omega - \big\langle \mathbf{curl}\,\mathbf{E}\times\mathbf{n},\mathbf{v}_T\big\rangle_{\partial\Omega}.\end{aligned} \tag{2.37}$$

Here, the identities $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \mathbf{c}\cdot(\mathbf{a}\times\mathbf{b})$ and $\mathbf{a}\times\mathbf{b} = -\mathbf{b}\times\mathbf{a}$ along with the decomposition $\mathbf{v} = \mathbf{v}_T + (\mathbf{v}\cdot\mathbf{n})\mathbf{n}$ and the fact $(\mathbf{a}\times\mathbf{n})\cdot\mathbf{n} = 0$ have been used.

By the boundary conditions (1.7) we get

$$(\mathbf{curl}\,\mathbf{curl}\,\mathbf{E},\mathbf{v})_\Omega = (\mathbf{curl}\,\mathbf{E},\mathbf{curl}\,\mathbf{v})_\Omega - \mathbf{i}\lambda\big\langle\mathbf{E}_T,\mathbf{v}_T\big\rangle_{\partial\Omega} - \big\langle\mathbf{g},\mathbf{v}_T\big\rangle_{\partial\Omega}.$$

Applying the above identity to (2.34) yields the following weak formulation of the time-harmonic Maxwell's equations: find $\mathbf{E} \in \boldsymbol{\mathcal{V}}$ such that

$$a(\mathbf{E}, \mathbf{v}) = (\mathbf{f}, \mathbf{v})_\Omega + \langle \mathbf{g}, \mathbf{v}_T \rangle_{\partial\Omega} \qquad \forall \mathbf{v} \in \boldsymbol{\mathcal{V}}, \tag{2.38}$$

where the sesquilinear form $a(\cdot, \cdot)$ on $\boldsymbol{\mathcal{V}} \times \boldsymbol{\mathcal{V}}$ is defined by

$$a(\mathbf{u}, \mathbf{v}) := (\mathbf{curl}\, \mathbf{u}, \mathbf{curl}\, \mathbf{v})_\Omega - k^2 (\mathbf{u}, \mathbf{v})_\Omega - \mathbf{i}\lambda \langle \mathbf{u}_T, \mathbf{v}_T \rangle_{\partial\Omega}. \tag{2.39}$$

After having derived the above weak formulation for the time-harmonic Maxwell's equations, the focus of this section shifts to the goal of obtaining a generalized weak coercivity property for the sesquilinear form $a(\cdot, \cdot)$. Again, Rellich identities will be used to achieve this goal. These Rellich identities are generalizations of those that can be found in [39]. Similar identities are derived and used to achieve stability estimates for the time-harmonic Maxwell's equations when Ω is a star-shape domain (c.f. [50]). Since the general case of $\boldsymbol{\alpha}$ being a $\mathbf{C}^1(\overline{\Omega})$ function was not considered, detailed proofs for these Rellich identities are given below. The following notation will be used in the Rellich identities:

$$\nabla_{\mathbf{a}} \mathbf{b} := (\mathbf{a} \cdot \nabla)\mathbf{b}.$$

Lemma 2.4.1. *Suppose* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *and* $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$. *Then the following identity holds:*

$$\begin{aligned} &\left(\operatorname{div} \boldsymbol{\alpha}, |\mathbf{curl}\, \mathbf{u}|^2\right)_\Omega + \left\langle \boldsymbol{\alpha} \cdot \mathbf{n}, |\mathbf{curl}\, \mathbf{u}|^2 \right\rangle_{\partial\Omega} \\ &\qquad = 2 \operatorname{Re} \left(\mathbf{curl}\, \mathbf{u}, \mathbf{curl}\, (\mathbf{curl}\, \mathbf{u} \times \boldsymbol{\alpha})\right)_\Omega + 2 \operatorname{Re} \left(\mathbf{curl}\, \mathbf{u}, \nabla_{\mathbf{curl}\, \mathbf{u}} \boldsymbol{\alpha}\right)_\Omega . \end{aligned}$$

Proof. To prove this lemma two additional differential identities along with the divergence theorem are used. Set $\mathbf{v} = \mathbf{curl}\, \mathbf{u} \times \boldsymbol{\alpha}$. To derive the first identity,

we recall the following well-known identity involving the **curl** operator:

$$\begin{aligned}\mathbf{curl}\,\mathbf{v} &= \mathbf{curl}\,(\mathbf{curl}\,\mathbf{u}\times\boldsymbol{\alpha})\\ &= \mathrm{div}\,(\boldsymbol{\alpha})\mathbf{curl}\,\mathbf{u} - \boldsymbol{\alpha}\mathrm{div}\,(\mathbf{curl}\,\mathbf{u}) + \nabla_{\boldsymbol{\alpha}}\mathbf{curl}\,\mathbf{u} - \nabla_{\mathbf{curl}\,\mathbf{u}}\boldsymbol{\alpha}.\end{aligned}$$

Using the fact that $\mathrm{div}\,(\mathbf{curl}\,\mathbf{u}) = 0$, the above identity gives the first sought-after identity:

$$\mathbf{curl}\,\mathbf{v} = \mathrm{div}\,(\boldsymbol{\alpha})\mathbf{curl}\,\mathbf{u} + \nabla_{\boldsymbol{\alpha}}\mathbf{curl}\,\mathbf{u} - \nabla_{\mathbf{curl}\,\mathbf{u}}\boldsymbol{\alpha}. \tag{2.40}$$

To derive the second sought-after identity, expanding $\mathrm{div}\,(\mathbf{a}|\mathbf{b}|^2)$ using the product rule for the divergence and gradient yields

$$\begin{aligned}\mathrm{div}\,\left(\mathbf{a}|\mathbf{b}|^2\right) &= \mathrm{div}\,(\mathbf{a})|\mathbf{b}|^2 + \mathbf{a}\cdot\nabla(\mathbf{b}\cdot\overline{\mathbf{b}})\\ &= \mathrm{div}\,(\mathbf{a})|\mathbf{b}|^2 + \mathbf{b}\cdot\nabla_{\mathbf{a}}\overline{\mathbf{b}} + \overline{\mathbf{b}}\cdot\nabla_{\mathbf{a}}\mathbf{b}\\ &= \mathrm{div}\,(\mathbf{a})|\mathbf{b}|^2 + 2\,\mathrm{Re}\,\overline{\mathbf{b}}\cdot\nabla_{\mathbf{a}}\mathbf{b}.\end{aligned} \tag{2.41}$$

(2.41) immediately gives the second sought-after identity

$$\mathrm{div}\,\left(\boldsymbol{\alpha}|\mathbf{curl}\,\mathbf{u}|^2\right) = \mathrm{div}\,\boldsymbol{\alpha}|\mathbf{curl}\,\mathbf{u}|^2 + 2\,\mathrm{Re}\,\mathbf{curl}\,\overline{\mathbf{u}}\cdot\nabla_{\boldsymbol{\alpha}}\mathbf{curl}\,\mathbf{u}. \tag{2.42}$$

Taking the complex conjugate of (2.40), applying the dot product with $2\mathbf{curl}\,\mathbf{u}$, and taking the real part gives

$$\begin{aligned}2\,\mathrm{Re}\,\mathbf{curl}\,\mathbf{u}\cdot\mathbf{curl}\,\overline{\mathbf{v}} = 2\mathrm{div}\,(\boldsymbol{\alpha})|\mathbf{curl}\,\mathbf{u}|^2 &+ 2\,\mathrm{Re}\,\mathbf{curl}\,\mathbf{u}\cdot\nabla_{\boldsymbol{\alpha}}\mathbf{curl}\,\overline{\mathbf{u}}\\ &- 2\,\mathrm{Re}\,\mathbf{curl}\,\mathbf{u}\cdot\nabla_{\mathbf{curl}\,\overline{\mathbf{u}}}\boldsymbol{\alpha},\end{aligned}$$

which together with (2.42) gives

$$\begin{aligned} 2\,\mathrm{Re}\,\mathbf{curl}\,\mathbf{u}\cdot\mathbf{curl}\,\overline{\mathbf{v}} &= 2\mathrm{div}\,(\boldsymbol{\alpha})|\mathbf{curl}\,\mathbf{u}|^2 - \mathrm{div}\,(\boldsymbol{\alpha})|\mathbf{curl}\,\mathbf{u}|^2 + \mathrm{div}\,\left(\boldsymbol{\alpha}|\mathbf{curl}\,\mathbf{u}|^2\right) \\ &\quad - 2\,\mathrm{Re}\,\mathbf{curl}\,\mathbf{u}\cdot\nabla_{\mathbf{curl}\,\overline{\mathbf{u}}}\boldsymbol{\alpha} \\ &= \mathrm{div}\,(\boldsymbol{\alpha})|\mathbf{curl}\,\mathbf{u}|^2 + \mathrm{div}\,\left(\boldsymbol{\alpha}|\mathbf{curl}\,\mathbf{u}|^2\right) - 2\,\mathrm{Re}\,\mathbf{curl}\,\mathbf{u}\cdot\nabla_{\mathbf{curl}\,\overline{\mathbf{u}}}\boldsymbol{\alpha}. \end{aligned}$$

Integrating the above identity over Ω and using the divergence theorem yields

$$\begin{aligned} 2\,\mathrm{Re}(\mathbf{curl}\,\mathbf{u},\mathbf{curl}\,\mathbf{v})_\Omega &= \left(\mathrm{div}\,(\boldsymbol{\alpha}),|\mathbf{curl}\,\mathbf{u}|^2\right)_\Omega + \int_\Omega \mathrm{div}\,\left(\boldsymbol{\alpha}|\mathbf{curl}\,\mathbf{u}|^2\right)\mathrm{d}\mathbf{x} \\ &\quad - 2\,\mathrm{Re}\left(\mathbf{curl}\,\mathbf{u},\nabla_{\mathbf{curl}\,\overline{\mathbf{u}}}\boldsymbol{\alpha}\right) \\ &= \left(\mathrm{div}\,(\boldsymbol{\alpha}),|\mathbf{curl}\,\mathbf{u}|^2\right)_\Omega + \left\langle\boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{curl}\,\mathbf{u}|^2\right\rangle_{\partial\Omega} \\ &\quad - 2\,\mathrm{Re}\left(\mathbf{curl}\,\mathbf{u},\nabla_{\mathbf{curl}\,\overline{\mathbf{u}}}\boldsymbol{\alpha}\right)_\Omega. \end{aligned}$$

By rearranging terms and recalling $\mathbf{v} = \mathbf{curl}\,\mathbf{u}\times\boldsymbol{\alpha}$, the desired Rellich identity is obtained. □

Lemma 2.4.2. *Suppose* $\mathbf{u}\in\mathbf{H}^1(\Omega)$ *and* $\boldsymbol{\alpha}\in\mathbf{C}^1(\overline{\Omega})$. *Then the following identity holds:*

$$\begin{aligned} &\mathrm{Re}(\mathbf{u},\mathbf{curl}\,\mathbf{u}\times\boldsymbol{\alpha})_\Omega + \frac{1}{2}\left(\mathrm{div}\,\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega + \frac{1}{2}\left\langle\boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{u}|^2\right\rangle_{\partial\Omega} \\ &\quad = \mathrm{Re}(\boldsymbol{\alpha}\mathrm{div}\,\mathbf{u},\mathbf{u})_\Omega + \mathrm{Re}\left(\mathbf{u},\nabla_{\mathbf{u}}\boldsymbol{\alpha}\right)_\Omega - \mathrm{Re}\langle\boldsymbol{\alpha}\times\mathbf{u},\mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega}. \end{aligned}$$

Proof. Like the proof of Lemma 2.4.1, this proof relies on two differential identities along with integration by parts. The first differential identity is the following identity involving the curl operator:

$$\mathbf{curl}\,(\boldsymbol{\alpha}\times\mathbf{u}) = \boldsymbol{\alpha}\mathrm{div}\,\mathbf{u} - \mathbf{u}\mathrm{div}\,\boldsymbol{\alpha} + \nabla_{\mathbf{u}}\boldsymbol{\alpha} - \nabla_{\boldsymbol{\alpha}}\mathbf{u}. \tag{2.43}$$

The next identity follows from (2.41) and reads as

$$\operatorname{div}\left(\boldsymbol{\alpha}|\mathbf{u}|^2\right) = \operatorname{div}(\boldsymbol{\alpha})|\mathbf{u}|^2 + 2\operatorname{Re}\overline{\mathbf{u}}\nabla_{\boldsymbol{\alpha}}\mathbf{u}. \tag{2.44}$$

By (2.36) and the divergence theorem we get

$$\begin{aligned}(\mathbf{curl}\,\mathbf{u},\mathbf{v})_\Omega &= (\mathbf{u},\mathbf{curl}\,\mathbf{v})_\Omega + \int_\Omega \operatorname{div}(\mathbf{u}\times\overline{\mathbf{v}})\mathrm{d}\mathbf{x} \\ &= (\mathbf{u},\mathbf{curl}\,\mathbf{v})_\Omega + \langle \mathbf{u}\times\overline{\mathbf{v}},\mathbf{n}\rangle_{\partial\Omega}.\end{aligned} \tag{2.45}$$

It follows from the identities $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \mathbf{b}\cdot(\mathbf{c}\times\mathbf{a})$, (2.45), (2.43), and (2.44) that

$$\begin{aligned}\operatorname{Re}(\mathbf{u},\mathbf{curl}\,\mathbf{u}\times\boldsymbol{\alpha})_\Omega &= \operatorname{Re}(\mathbf{curl}\,\mathbf{u},\boldsymbol{\alpha}\times\mathbf{u}) \\ &= \operatorname{Re}\langle \mathbf{u}\times\boldsymbol{\alpha}\times\overline{\mathbf{u}},\mathbf{n}\rangle_{\partial\Omega} + \operatorname{Re}(\mathbf{u},\mathbf{curl}\,(\boldsymbol{\alpha}\times\mathbf{u}))_\Omega \\ &= -\operatorname{Re}\langle \boldsymbol{\alpha}\times\mathbf{u},\mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} + \operatorname{Re}(\mathbf{u},\boldsymbol{\alpha}\operatorname{div}\mathbf{u})_\Omega - \left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega \\ &\quad + \operatorname{Re}(\mathbf{u},\nabla_{\mathbf{u}}\boldsymbol{\alpha})_\Omega - \operatorname{Re}(\mathbf{u},\nabla_{\boldsymbol{\alpha}}\mathbf{u})_\Omega \\ &= -\operatorname{Re}\langle \boldsymbol{\alpha}\times\mathbf{u},\mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} + \operatorname{Re}(\mathbf{u},\boldsymbol{\alpha}\operatorname{div}\mathbf{u})_\Omega - \left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega \\ &\quad + \operatorname{Re}(\mathbf{u},\nabla_{\mathbf{u}}\boldsymbol{\alpha})_\Omega - \frac{1}{2}\left\langle \boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{u}|^2\right\rangle_{\partial\Omega} + \frac{1}{2}\left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega \\ &= -\operatorname{Re}\langle \boldsymbol{\alpha}\times\mathbf{u},\mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} + \operatorname{Re}(\mathbf{u},\boldsymbol{\alpha}\operatorname{div}\mathbf{u})_\Omega - \frac{1}{2}\left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega \\ &\quad + \operatorname{Re}(\mathbf{u},\nabla_{\mathbf{u}}\boldsymbol{\alpha})_\Omega - \frac{1}{2}\left\langle \boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{u}|^2\right\rangle_{\partial\Omega}.\end{aligned}$$

Thus, the desired Rellich identity is obtained by rearranging the terms above. □

The above Rellich identities can be used to establish a generalized weak coercivity property for $a(\cdot,\cdot)$. Note that a generalized weak coercivity property for $a(\cdot,\cdot)$ was already established in [43] for a star-shape domain and what is below is an extension of that result to a generalized star-shape domain.

Theorem 2.4.3. *Let Ω be a generalized star-shape domain such that there exists $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.30)–(2.33). Then for any $\mathbf{u} \in \boldsymbol{\mathcal{V}} \cap \mathbf{H}(\operatorname{div}_0, \Omega)$ the following generalized weak coercivity property holds on $a(\cdot,\cdot)$:*

$$\sup_{\mathbf{v}\in\hat{\boldsymbol{\mathcal{V}}}} \frac{|\operatorname{Im} a(\mathbf{u},\mathbf{v})|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\boldsymbol{\mathcal{V}}} \frac{|\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{|||\mathbf{v}|||_{L^2(\Omega)}} \geq \frac{1}{\gamma}\|\mathbf{u}\|_E,$$

where

$$\gamma := 4kR + M, \qquad M := \frac{2Rc_+ + 2R^2k^2 + 2R^2\lambda^2}{\lambda c_+},$$
$$|||\mathbf{u}|||_{L^2(\Omega)} := \left(c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}},$$
$$\|\mathbf{u}\|_E := \left(c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} + c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}}.$$

Proof. Similar to the proof of Theorems 2.2.4 and 2.3.6, this proof makes use of two specific test functions. The first is $\mathbf{v} = \mathbf{u}$. Using this test function in (2.39) and taking the real and imaginary parts separately yield

$$\operatorname{Re} a(\mathbf{u},\mathbf{u}) = \|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} - k^2\|\mathbf{u}\|^2_{L^2(\Omega)}, \tag{2.46}$$
$$\operatorname{Im} a(\mathbf{u},\mathbf{u}) = -\lambda\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)}. \tag{2.47}$$

The second test function is $\mathbf{v} = \mathbf{curl}\,\mathbf{u} \times \boldsymbol{\alpha}$ motivated by the Rellich identities. Recall that $\boldsymbol{\alpha}$ is the vector field defined by the generalized star-shape condition on Ω. Using Lemmas 2.4.1 and 2.4.2 gives

$$\begin{aligned} 2\operatorname{Re} a(\mathbf{u},\mathbf{v}) &= 2\operatorname{Re}(\mathbf{curl}\,\mathbf{u},\mathbf{curl}\,\mathbf{v})_\Omega - 2k^2\operatorname{Re}(\mathbf{u},\mathbf{v})_\Omega + 2\lambda\operatorname{Im}\langle\mathbf{u}_T,\mathbf{v}_T\rangle_{\partial\Omega} \\ &= \left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{curl}\,\mathbf{u}|^2\right)_\Omega - 2\operatorname{Re}(\mathbf{curl}\,\mathbf{u},\nabla_{\mathbf{curl}\,\mathbf{u}}\boldsymbol{\alpha})_\Omega + \left\langle\boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{curl}\,\mathbf{u}|^2\right\rangle_{\partial\Omega} \\ &\quad + k^2\left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega - 2k^2\operatorname{Re}(\mathbf{u},\nabla_{\mathbf{u}}\boldsymbol{\alpha})_\Omega + k^2\left\langle\boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{u}|^2\right\rangle_{\partial\Omega} \\ &\quad + 2k^2\operatorname{Re}\langle\boldsymbol{\alpha}\times\mathbf{u},\mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} + 2\lambda\operatorname{Re}\langle\mathbf{u}_T,\mathbf{v}_T\rangle_{\partial\Omega}. \end{aligned} \tag{2.48}$$

Here the fact that $\operatorname{div}\mathbf{u} = 0$ has been used.

By using (2.30) we get

$$\mathbf{u}\cdot\nabla_{\overline{\mathbf{u}}}\boldsymbol{\alpha}=\sum_{i=1}^{3}u_i(\overline{\mathbf{u}}\cdot\nabla)\alpha_i=\sum_{i=1}^{3}\sum_{j=1}^{3}u_i\Big(\overline{u}_j\frac{\partial}{\partial x_j}\Big)\alpha_i=\sum_{i=1}^{3}\sum_{j=1}^{3}|u_i|^2\frac{\partial\alpha_i}{\partial x_i} \tag{2.49}$$
$$\leq|\mathbf{u}|^2\max_{i=1,2,3}\left\{\frac{\partial\alpha_i}{\partial x_i}\right\}.$$

Similarly,

$$\mathbf{curl}\,\mathbf{u}\cdot\nabla_{\mathbf{curl}\,\overline{\mathbf{u}}}\boldsymbol{\alpha}\leq|\mathbf{curl}\,\mathbf{u}|^2\max_{i=1,2,3}\left\{\frac{\partial\alpha_i}{\partial x_i}\right\}. \tag{2.50}$$

Combining (2.49) and (2.50) with (2.33) gives

$$k^2\left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{u}|^2\right)_\Omega-2k^2\operatorname{Re}(\mathbf{u},\nabla_{\mathbf{u}}\boldsymbol{\alpha})_\Omega\geq k^2\left(\operatorname{div}\boldsymbol{\alpha}-\max_{i=1,2,3}\left\{\frac{\partial\alpha_i}{\partial x_i}\right\},|\mathbf{u}|^2\right)_\Omega$$
$$\geq c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)}$$

and

$$\left(\operatorname{div}\boldsymbol{\alpha},|\mathbf{curl}\,\mathbf{u}|^2\right)_\Omega-2\operatorname{Re}(\mathbf{curl}\,\mathbf{u},\nabla_{\mathbf{curl}\,\mathbf{u}}\boldsymbol{\alpha})_\Omega\geq c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)}.$$

Rearranging the terms in (2.48) and substituting the above inequalities yield

$$c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)}+c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)}$$
$$\leq-k^2\left\langle\boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{u}|^2\right\rangle_{\partial\Omega}-\left\langle\boldsymbol{\alpha}\cdot\mathbf{n},|\mathbf{curl}\,\mathbf{u}|^2\right\rangle_{\partial\Omega}$$
$$+2k^2\langle\boldsymbol{\alpha}\times\mathbf{u},\mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega}-2\lambda\langle\mathbf{u}_T,\mathbf{v}_T\rangle_{\partial\Omega}+2\operatorname{Re}a(\mathbf{u},\mathbf{v}). \tag{2.51}$$

To bound the term $2k^2\langle \boldsymbol{\alpha}\times\mathbf{u}, \mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega}$, by the decomposition $\boldsymbol{\alpha} = \boldsymbol{\alpha}_T + (\boldsymbol{\alpha}\cdot\mathbf{n})\boldsymbol{\alpha}$ and the identity $(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{c}\times\mathbf{d}) = (\mathbf{a}\cdot\mathbf{c})(\mathbf{b}\cdot\mathbf{d}) - (\mathbf{a}\cdot\mathbf{d})(\mathbf{b}\cdot\mathbf{c})$ we get

$$
\begin{aligned}
&-2k^2\langle \boldsymbol{\alpha}\times\mathbf{u}, \mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} \\
&\quad = -2k^2\langle \boldsymbol{\alpha}_T\times\mathbf{u}, \mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} - 2k^2\langle (\boldsymbol{\alpha}\cdot\mathbf{n})\mathbf{n}\times\mathbf{u}, \mathbf{u}\times\mathbf{n}\rangle_{\partial\Omega} \\
&\quad = -2k^2\langle \boldsymbol{\alpha}_T\cdot\mathbf{u}, \mathbf{u}\cdot\mathbf{n}\rangle_{\partial\Omega} + 2k^2\langle \boldsymbol{\alpha}_T\cdot\mathbf{n}, |\mathbf{u}|^2\rangle_{\partial\Omega} + 2k^2\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}\times\mathbf{n}|^2\rangle_{\partial\Omega} \\
&\quad = -2k^2\langle \boldsymbol{\alpha}_T\cdot\mathbf{u}_T, \mathbf{u}\cdot\mathbf{n}\rangle_{\partial\Omega} + 2k^2\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}\times\mathbf{n}|^2\rangle_{\partial\Omega}.
\end{aligned}
$$

This identity allows us to rewrite (2.51) as

$$
\begin{aligned}
&c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} \qquad (2.52)\\
&\quad \le -k^2\left\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}|^2\right\rangle_{\partial\Omega} - \left\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{curl}\,\mathbf{u}|^2\right\rangle_{\partial\Omega} + 2k^2\,\mathrm{Re}\left\langle \boldsymbol{\alpha}\cdot\mathbf{n}, |\mathbf{u}\times\mathbf{n}|^2\right\rangle_{\partial\Omega} \\
&\qquad - 2k^2\,\mathrm{Re}\langle \mathbf{u}\cdot\mathbf{n}, \boldsymbol{\alpha}_T\cdot\mathbf{u}_T\rangle_{\partial\Omega} - 2\lambda\,\mathrm{Re}\langle \mathbf{u}_T, \mathbf{v}_T\rangle_{\partial\Omega} + 2\,\mathrm{Re}\,a(\mathbf{u},\mathbf{v}).
\end{aligned}
$$

Noting that

$$
|\mathbf{u}_T|^2 = |(\mathbf{n}\times\mathbf{u})\times\mathbf{n}|^2 = |\mathbf{u}\times\mathbf{n}|^2 - \big((\mathbf{u}\times\mathbf{n})\cdot\mathbf{n}\big)^2 = |\mathbf{u}\times\mathbf{n}|^2 \qquad \text{on } \partial\Omega.
$$

We use this identity along with (2.31), (2.32), Cauchy Schwarz and Young's inequality in (2.52) to get

$$\begin{aligned}
&c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} \\
&\quad\le -c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\|\mathbf{curl}\,\mathbf{u}\|_{L^2(\partial\Omega)} + 2Rk^2\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)} \\
&\qquad + 2Rk^2\|\mathbf{u}\|_{L^2(\partial\Omega)}\|\mathbf{u}_T\|_{L^2(\partial\Omega)} + 2R\lambda\|\mathbf{u}_T\|_{L^2(\partial\Omega)}\|\mathbf{curl}\,\mathbf{u}\|_{L^2(\partial\Omega)} \\
&\qquad + 2\,\mathrm{Re}\,a(\mathbf{u},\mathbf{v}) \\
&\quad\le -c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - c_+\|\mathbf{curl}\,\mathbf{u}\|_{L^2(\partial\Omega)} + 2Rk^2\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)} \\
&\qquad + \frac{c_+k^2}{2}\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + \frac{2R^2k^2}{c_+}\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)} \\
&\qquad + \frac{2R^2\lambda^2}{c_+}\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)} + \frac{c_+}{2}\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\partial\Omega)} \\
&\qquad + 2\,\mathrm{Re}\,a(\mathbf{u},\mathbf{v}) \\
&\quad\le -\frac{c_+}{2}k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)} - \frac{c_+}{2}\|\mathbf{curl}\,\mathbf{u}\|_{L^2(\partial\Omega)} + 2\,\mathrm{Re}\,a(\mathbf{u},\mathbf{v}) \\
&\qquad + \frac{2Rc_+ + 2R^2k^2 + 2R^2\lambda^2}{c_+}\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)}.
\end{aligned}$$

Multiplying the above inequality by 2 and making use of (2.47) yield

$$\begin{aligned}
&2c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + 2c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} + c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)} \\
&\qquad + c_+\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\partial\Omega)} + 4c_+k^2\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)} \\
&\quad\le \frac{2Rc_+ + 2R^2k^2 + 2R^2\lambda^2}{\lambda c_+}\lambda\|\mathbf{u}_T\|^2_{L^2(\partial\Omega)} + \mathrm{Re}\,a(\mathbf{u},4\mathbf{v}). \\
&\quad\le M|\,\mathrm{Im}\,a(\mathbf{u},\mathbf{u})| + |\,\mathrm{Re}\,a(\mathbf{u},4\mathbf{v})|.
\end{aligned}$$

Thus,

$$\|\mathbf{u}\|^2_E \le M\big|\,\mathrm{Im}\,a(\mathbf{u},\mathbf{u})\big| + \big|\,\mathrm{Re}\,a(\mathbf{u},4\mathbf{v})\big|. \tag{2.53}$$

By the definitions of $\mathbf{v}$, $\|\cdot\|_E$, and $|||\cdot|||_{L^2(\Omega)}$ we have

$$\begin{aligned}
|||\mathbf{v}|||_{L^2(\Omega)} &= \Big(c_1 k^2 \|\mathbf{curl}\,\mathbf{u} \times \boldsymbol{\alpha}\|^2_{L^2(\Omega)} + c_+ k^2 \|\mathbf{curl}\,\mathbf{u} \times \boldsymbol{\alpha}\|^2_{L^2(\partial\Omega)}\Big)^{\frac{1}{2}} \\
&\leq \Big(c_1 R^2 k^2 \|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} + c_+ R^2 k^2 \|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\partial\Omega)}\Big)^{\frac{1}{2}} \\
&\leq kR\|\mathbf{u}\|_E. \qquad (2.54)
\end{aligned}$$

It follows from (2.53) and (2.54) that

$$\begin{aligned}
\sup_{\mathbf{v}\in\mathbf{H}^2(\Omega)} \frac{|\operatorname{Im} a(\mathbf{u},\mathbf{v})|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\mathbf{H}^1(\Omega)} \frac{|\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{|||\mathbf{v}|||_{L^2(\Omega)}} &\geq \frac{|\operatorname{Im} a(\mathbf{u},\mathbf{u})|}{\|\mathbf{u}\|_E} + \frac{|\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{|||4(\mathbf{curl}\,\mathbf{u} \times \boldsymbol{\alpha})|||_{L^2(\Omega)}} \\
&\geq \frac{M|\operatorname{Im} a(\mathbf{u},\mathbf{u})|}{M\|\mathbf{u}\|_E} + \frac{|\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{\gamma\|\mathbf{u}\|_E} \\
&\geq \frac{1}{\gamma} \cdot \frac{M|\operatorname{Im} a(\mathbf{u},\mathbf{u})| + |\operatorname{Re} a(\mathbf{u},\mathbf{v})|}{\|\mathbf{u}\|_E} \\
&\geq \frac{1}{\gamma}\|\mathbf{u}\|_E,
\end{aligned}$$

which yields the desired generalized weak coercivity property. □

2.5 Applications to Stability Estimates

In Sections 2.2, 2.3, and 2.4, it was demonstrated that each Helmholtz-type problem satisfies a generalized weak coercivity property. The goal of this section is to give one application of the generalized weak coercivity properties. Namely, we apply the generalized weak coercivity properties to derive wave-number explicit solution estimates for each Helmholtz-type problem. These solution estimates are stated in the following three theorems.

Theorem 2.5.1. *Let $\Omega \subset \mathbb{R}^d$ be a generalized star-shape domain with $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.2)–(2.8). Suppose $u \in V$ solves (2.9) with $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega_+)$.*

Then the following estimate holds:

$$\|u\|_E \leq 2\gamma\beta\Big(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_+)}\Big),$$

where

$$\begin{aligned}
&\gamma := \max\left\{\left[2(2c_2 - 4c_3 + c_4)^2 + 16(k^2+1)R^2\right]^{\frac{1}{2}}, M\right\},\\
&M := 2\left(kR + \frac{kR^2}{c_+} + \frac{c_+}{k}\right),\\
&\beta := \frac{1}{\sqrt{c_4}k} + \frac{1}{\sqrt{c_+}},\\
&|||u|||_{L^2(\Omega)} := \left(k^2 c_4\|u\|^2_{L^2(\Omega)} + c_+\|u\|^2_{L^2(\partial\Omega_+)}\right)^{\frac{1}{2}},\\
&\|u\|_E := \Big(k^2 c_4\|u\|^2_{L^2(\Omega)} + c_4\|\nabla u\|^2_{L^2(\Omega)} + c_+\|u\|^2_{L^2(\partial\Omega_+)} + c_+\|\nabla u\|^2_{L^2(\partial\Omega_+)}\\
&\qquad\qquad + c_-\|\nabla u\|^2_{L^2(\partial\Omega_-)}\Big)^{\frac{1}{2}}.
\end{aligned}$$

Proof. Let $v \in H^1(\Omega)$. The Cauchy-Schwarz inequality yields the following series of inequalities:

$$\begin{aligned}
|a(u,v)| &= |(f,v)_\Omega + \langle g, v\rangle_{\partial\Omega_+}|\\
&\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_+)}\|v\|_{L^2(\partial\Omega_+)}\\
&\leq \left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\partial\Omega_+)}\right)^{\frac{1}{2}}\left(\|v\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\partial\Omega_+)}\right)^{\frac{1}{2}}\\
&\leq \left(\frac{1}{\sqrt{c_4}k} + \frac{1}{\sqrt{c_+}}\right)\left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\partial\Omega_+)}\right)^{\frac{1}{2}}|||v|||_{L^2(\Omega)}.
\end{aligned}$$

Similarly, for $v \in V$ we have

$$\begin{aligned}
|a(u,v)| &= |(f,v)_\Omega + \langle g, v\rangle_{\partial\Omega_+}|\\
&\leq \left(\frac{1}{\sqrt{c_4}k} + \frac{1}{\sqrt{c_+}}\right)\left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}}\|v\|_E.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{1}{\gamma}\|u\|_E &\leq \sup_{v\in H^2(\Omega)} \frac{|\operatorname{Im} a(u,v)|}{\|v\|_E} + \sup_{v\in H^1(\Omega)} \frac{|\operatorname{Re} a(u,v)|}{|||v|||_{L^2(\Omega)}} \\
&\leq \sup_{v\in H^2(\Omega)} \frac{|a(u,v)|}{\|v\|_E} + \sup_{v\in H^1(\Omega)} \frac{|a(u,v)|}{|||v|||_{L^2(\Omega)}} \\
&\leq \left(\frac{1}{\sqrt{c_4}k} + \frac{1}{\sqrt{c_+}}\right)\left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}} \\
&\quad + \left(\frac{1}{\sqrt{c_4}k} + \frac{1}{\sqrt{c_+}}\right)\left(\|f\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\partial\Omega_+)}\right)^{\frac{1}{2}} \\
&\leq 2\beta\Big(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_+)}\Big).
\end{aligned}$$

The proof is complete. □

Theorem 2.5.2. *Let $\Omega \subset \mathbb{R}^d$ be a generalized star-shape domain such that there exists $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.15)–(2.18). Suppose that there exists some positive constant $\tilde{K}$ such that $\mathbf{u} \in \mathbf{V}_{\tilde{K}}$ solves (2.22) for $\mathbf{f} \in \mathbf{L}^2(\Omega)$ and $\mathbf{g} \in \mathbf{L}^2(\partial\Omega)$. Then the following stability estimate holds:*

$$\|\mathbf{u}\|_E \leq 2\gamma\left(\frac{1}{\omega\sqrt{c_1}} + \frac{1}{\sqrt{c_+\mu}}\right)\Big(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\Big),$$

where

$$\begin{aligned}
\gamma &:= \max\left\{\left[4R^2K\left(1+\frac{\omega^2\rho}{2\mu}\right) + 4R^2\tilde{K} + (1-d)^2c_1^2\right]^{\frac{1}{2}}, M\right\}, \\
M &:= \frac{1}{c_A}\left(R\omega\rho + \frac{R^2\omega C_A\tilde{K}}{c_+\mu} + \frac{2c_+\mu}{\omega}\right), \\
\|\mathbf{u}\|^2_E &:= c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\lambda\|\operatorname{div}\mathbf{u}\|^2_{L^2(\Omega)} + 2c_1\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\Omega)} \\
&\qquad + c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\lambda\big\|\operatorname{div}\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\mu\|\varepsilon(\mathbf{u})\|^2_{L^2(\partial\Omega)}, \\
|||\mathbf{u}|||^2_{L^2(\Omega)} &:= c_1\omega^2\rho\|\mathbf{u}\|^2_{L^2(\Omega)} + c_+\mu\|\mathbf{u}\|^2_{L^2(\partial\Omega)}.
\end{aligned}$$

Proof. This proof is very similar to that of Theorem 2.5.1. We begin with finding an upper bound on $|a(\mathbf{u},\mathbf{v})|$ for some $\mathbf{v} \in \mathbf{H}^1(\Omega)$. By the Cauchy-Schwarz inequality,

we get

$$
\begin{aligned}
|a(\mathbf{u},\mathbf{v})| &= |(\mathbf{f},\mathbf{v})_\Omega + \langle \mathbf{g},\mathbf{v}\rangle_{\partial\Omega}| \\
&\leq \|\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\|\mathbf{v}\|_{L^2(\partial\Omega)} \\
&\leq \left(\|\mathbf{f}\|^2_{L^2(\Omega)} + \|\mathbf{g}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}} \left(\|\mathbf{v}\|^2_{L^2(\Omega)} + \|\mathbf{v}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{c_1\omega^2} + \frac{1}{c_+\mu}\right)^{\frac{1}{2}} \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) \left(c_1\omega^2\|\mathbf{v}\|^2_{L^2(\Omega)} + c_+\mu\|\mathbf{v}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}}.
\end{aligned}
$$

Therefore, for $\mathbf{v} \in \mathbf{H}^1(\Omega)$,

$$
|a(\mathbf{u},\mathbf{v})| \leq \left(\frac{1}{c_1\omega^2} + \frac{1}{c_+\mu}\right)^{\frac{1}{2}} \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) |||\mathbf{v}|||_{L^2(\Omega)}. \tag{2.55}
$$

By noting that, $|||\mathbf{v}|||_{L^2(\Omega)} \leq \|\mathbf{v}\|_E$, we find

$$
|a(\mathbf{u},\mathbf{v})| \leq \left(\frac{1}{c_1\omega^2} + \frac{1}{c_+\mu}\right)^{\frac{1}{2}} \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) \|\mathbf{v}\|_E. \tag{2.56}
$$

These bounds on $|a(\cdot,\cdot)|$ in conjunction with Theorem 2.3.6 imply that

$$
\begin{aligned}
\frac{1}{\gamma}\|\mathbf{u}\|_E &\leq \sup_{\mathbf{v}\in\mathbf{H}^2(\Omega)} \frac{\left|\operatorname{Im} a(\mathbf{u},\mathbf{v})\right|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\mathbf{H}^1(\Omega)} \frac{\left|\operatorname{Re} a(\mathbf{u},\mathbf{v})\right|}{|||\mathbf{v}|||_E} \\
&\leq \sup_{\mathbf{v}\in\mathbf{H}^2(\Omega)} \frac{|a(\mathbf{u},\mathbf{v})|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\mathbf{H}^1(\Omega)} \frac{|a(\mathbf{u},\mathbf{v})|}{|||\mathbf{v}|||_E} \\
&\leq 2\left(\frac{1}{c_1\omega^2} + \frac{1}{c_+\mu}\right)^{\frac{1}{2}} \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right).
\end{aligned}
$$

Hence the stability estimate holds. □

Theorem 2.5.3. *Let $\Omega \subset \mathbb{R}^3$ be a generalized star-shape domain with $\boldsymbol{\alpha} \in \mathbf{C}^1(\overline{\Omega})$ satisfying (2.30)–(2.33). Suppose $\mathbf{E} \in \hat{\boldsymbol{\mathcal{V}}}$ solves (2.38) for $\mathbf{f} \in \mathbf{H}(\operatorname{div},\Omega)$, $\mathbf{g} \in \mathbf{L}^2(\partial\Omega)$.*

Then the following solution estimate on $\mathbf{E}$ *holds:*

$$\|\mathbf{E}\|_E \leq \frac{4\gamma}{k}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\right) + \frac{(c_1+R)^{\frac{1}{2}}}{k}\|\mathbf{f}\|_\Omega + \frac{R^{\frac{1}{2}}}{k}\|\operatorname{div}\mathbf{f}\|_{L^2(\Omega)}$$

where

$$\gamma := 4kR + M, \qquad M := \frac{2Rc_+ + 2R^2k^2 + 2R^2\lambda^2}{\lambda c_+},$$
$$|||\mathbf{u}|||_{L^2(\Omega)} := \left(c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}},$$
$$\|\mathbf{u}\|_E := \left(c_1k^2\|\mathbf{u}\|^2_{L^2(\Omega)} + c_1\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} + c_+k^2\|\mathbf{u}\|^2_{L^2(\partial\Omega)} + c_+\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}}.$$

Proof. This proof follows the proof of Theorem 2.3 from [43] with changes in some details dealing with the generalized star-shape condition imposed on the domain Ω. Now since $\mathbf{E} \in \hat{\boldsymbol{\mathcal{V}}}$ solves (2.38) then $\mathbf{E}$ satisfies (1.6) a.e. in Ω and we find

$$-k^2\operatorname{div}\mathbf{E} = \operatorname{div}\left(\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} - k^2\mathbf{E}\right) = \operatorname{div}\mathbf{f} \qquad \text{a.e. in } \Omega.$$

This implies $\operatorname{div}\mathbf{E} = -k^{-2}\operatorname{div}\mathbf{f}$ a.e. in Ω.

To apply Theorem 2.4.3 we need a vector field $\mathbf{u} \in \mathbf{H}^2(\Omega)\cap\mathbf{H}(\operatorname{div}_0,\Omega)$. Therefore, unlike the proofs of Theorems 2.5.1 and 2.5.2, we cannot directly apply Theorem 2.4.3 to $\mathbf{E}$. To overcome this difficulty, consider an auxiliary vector field $\mathbf{F} = \nabla\phi$, where $\phi \in H^1_0(\Omega)$ solves the following Poisson equation:

$$\Delta\phi = k^{-2}\operatorname{div}\mathbf{f} \qquad \text{a.e. in } \Omega. \tag{2.57}$$

By definition, $\operatorname{div}\mathbf{F} = k^{-2}\operatorname{div}\mathbf{f}$ a.e. in Ω. Also, the definition of $\mathbf{F}$ ensures $\mathbf{curl}\,\mathbf{F} = 0$ so $\mathbf{F} \in \hat{\boldsymbol{\mathcal{V}}}$. Thus for $\mathbf{u} := \mathbf{E} + \mathbf{F}$, $\mathbf{u} \in \hat{\boldsymbol{\mathcal{V}}} \cap \mathbf{H}(\operatorname{div}_0,\Omega)$. With this in mind, the estimate on $\mathbf{E}$ will be obtained by estimating $\mathbf{F}$ and $\mathbf{u}$ separately. Estimates on $\mathbf{F}$

can be obtained based on its definition and estimates on $\mathbf{u}$ can be obtained from the generalized weak coercivity property for the time-harmonic Maxwell operator.

To derive estimates for $\mathbf{F}$, we test (2.57) with ϕ and integrate by parts to obtain

$$\|\nabla\phi\|^2_{L^2(\Omega)} = k^{-2}(\mathbf{f}, \nabla\phi)_\Omega \leq k^{-2}\|\mathbf{f}\|_{L^2(\Omega)}\|\nabla\phi\|_{L^2(\Omega)}.$$

Hence,

$$\|\mathbf{F}\|_{L^2(\Omega)} = \|\nabla\phi\|_{L^2(\Omega)} \leq k^{-2}\|\mathbf{f}\|_{L^2(\Omega)}. \tag{2.58}$$

Next, testing (2.57) by the test function $\nabla\phi\cdot\boldsymbol{\alpha} = \mathbf{F}\cdot\boldsymbol{\alpha}$ and applying integration by parts and Lemma 2.2.3 we obtain

$$\begin{aligned}
2\,\mathrm{Re}(\Delta\phi, \nabla\phi\cdot\boldsymbol{\alpha})_\Omega &= 2\,\mathrm{Re}(\Delta\phi, \nabla\phi\cdot\boldsymbol{\alpha})_\Omega \\
&= -2\,\mathrm{Re}\left(\nabla\phi, \nabla(\nabla\phi\cdot\boldsymbol{\alpha})\right)_\Omega + 2\,\mathrm{Re}\left\langle\frac{\partial\phi}{\partial\mathbf{n}}, \nabla\phi\cdot\boldsymbol{\alpha}\right\rangle_{\partial\Omega} \\
&= \left(\mathrm{div}\,\boldsymbol{\alpha}, |\phi|^2\right)_\Omega - 2\,\mathrm{Re}\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\frac{\partial\phi}{\partial x_i}, \frac{\partial\alpha_j}{\partial x_i}\frac{\partial\phi}{\partial x_j}\right)_\Omega \\
&\quad - \left\langle\boldsymbol{\alpha}\cdot\mathbf{n}, |\nabla\phi|^2\right\rangle_{\partial\Omega} + 2\,\mathrm{Re}\left\langle\frac{\partial\phi}{\partial\mathbf{n}}, \nabla\phi\cdot\boldsymbol{\alpha}\right\rangle_{\partial\Omega}.
\end{aligned}$$

Now $\mathbf{F}_T = (\nabla\phi)_T = 0$ on $\partial\Omega$ since $\phi \in H^1_0(\Omega)$. Using this fact along with (2.30), (2.32), and (2.33) in the above inequality gives

$$\begin{aligned}
2\,\mathrm{Re}(\Delta\phi, \nabla\phi\cdot\boldsymbol{\alpha})_\Omega &= \left(\mathrm{div}\,\boldsymbol{\alpha}, |\phi|^2\right)_\Omega - 2\,\mathrm{Re}\sum_{i=1}^{3}\left(\frac{\partial\alpha_j}{\partial x_i}, \left|\frac{\partial\phi}{\partial x_j}\right|^2\right)_\Omega \\
&\quad - \left\langle\boldsymbol{\alpha}\cdot\mathbf{n}, |\nabla\phi|^2\right\rangle_{\partial\Omega} + 2\,\mathrm{Re}\left\langle\frac{\partial\phi}{\partial\mathbf{n}}, \left((\nabla\phi)_T + \frac{\partial\phi}{\partial\mathbf{n}}\mathbf{n}\right)\cdot\boldsymbol{\alpha}\right\rangle_{\partial\Omega} \\
&\geq \left(\mathrm{div}\,\boldsymbol{\alpha} - 2\max_{i=1,2,3}\left\{\frac{\partial\alpha_i}{\partial x_i}\right\}, |\nabla\phi|^2\right)_\Omega + \left\langle\boldsymbol{\alpha}\cdot\mathbf{n}, |\nabla\phi|^2\right\rangle_{\partial\Omega} \\
&\geq c_1\|\mathbf{F}\|^2_{L^2(\Omega)} + c_+\|\mathbf{F}\|^2_{\partial\Omega}.
\end{aligned}$$

Therefore, it follows from (2.57), (2.58), (2.31), and the Cauchy-Schwarz inequality that

$$\begin{aligned} c_+\|\mathbf{F}\|^2_{L^2(\partial\Omega)} = c_+\|\nabla\phi\|^2_{L^2(\partial\Omega)} &\le 2k^{-2}\,\mathrm{Re}(\mathrm{div}\,\mathbf{f}, \mathbf{F}\cdot\boldsymbol{\alpha})_\Omega \\ &\le 2Rk^{-2}\|\mathrm{div}\,\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{F}\|_{L^2(\Omega)} \\ &\le 2Rk^{-4}\|\mathrm{div}\,\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{f}\|_{L^2(\Omega)} \\ &\le Rk^{-4}\|\mathrm{div}\,\mathbf{f}\|^2_{L^2(\Omega)} + Rk^{-4}\|\mathbf{f}\|^2_{L^2(\Omega)}. \end{aligned} \tag{2.59}$$

Since $\mathbf{curl}\,\mathbf{F} = 0$ in $\overline{\Omega}$, (2.58) and (2.59) yield the following estimate for $\|\mathbf{F}\|_E$:

$$\begin{aligned} \|\mathbf{F}\|^2_E &= c_1k^2\|\mathbf{F}\|_{L^2(\Omega)} + c_+k^2\|\mathbf{F}\|^2_{L^2(\partial\Omega)} \\ &\le c_1k^{-2}\|\mathbf{f}\|^2_\Omega + Rk^{-2}\|\mathrm{div}\,\mathbf{f}\|^2_{L^2(\Omega)} + Rk^{-2}\|\mathbf{f}\|^2_{L^2(\Omega)} \\ &= \frac{c_1+R}{k^2}\|\mathbf{f}\|^2_\Omega + \frac{R}{k^2}\|\mathrm{div}\,\mathbf{f}\|^2_{L^2(\Omega)}. \end{aligned} \tag{2.60}$$

Next, we derive estimates for $\mathbf{u}$. Note that since $\phi \in H^1_0(\Omega)$ and $\mathbf{F} = \nabla\phi$, $\mathbf{F}_T = 0$ on $\partial\Omega$ and $\mathbf{curl}\,\mathbf{F} = 0$ in Ω. These two facts imply that $\mathbf{u}$ satisfies the following weak form of the time-harmonic Maxwell's equations:

$$a(\mathbf{u},\mathbf{v}) = \big(\mathbf{f} - k^{-2}\mathbf{F}, \mathbf{v}\big)_\Omega + \langle \mathbf{g}, \mathbf{v}_T\rangle_{L^2(\partial\Omega)} \qquad \forall \mathbf{v} \in \boldsymbol{\mathcal{V}}. \tag{2.61}$$

where $a(\cdot,\cdot)$ is the sesquilinear form defined in (2.39). Thus, by the Cauchy-Schwarz inequality, (2.58), and (2.61) we get

$$\begin{aligned}
|a(\mathbf{u},\mathbf{v})| &= |(\mathbf{f}-k^2\mathbf{F},\mathbf{v})_\Omega + \langle \mathbf{g},\mathbf{v}_T\rangle_{\partial\Omega}| \\
&\le |(\mathbf{f},\mathbf{v})_\Omega| + k^2|(\mathbf{F},\mathbf{v})_\Omega| + |\langle \mathbf{g},\mathbf{v}_T\rangle_{\partial\Omega}| \\
&\le \|\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} + k^2\|\mathbf{F}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\|\mathbf{v}_T\|_{L^2(\partial\Omega)} \\
&= 2\|\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\|\mathbf{v}_T\|_{L^2(\partial\Omega)} \\
&\le \left(4\|\mathbf{f}\|^2_{L^2(\Omega)} + \|\mathbf{g}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \left(\|\mathbf{v}\|^2_{L^2(\Omega)} + \|\mathbf{v}_T\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}} \\
&\le 2k^{-1}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right)\left(c_1k^2\|\mathbf{v}\|^2_{L^2(\Omega)} + c_+k^2\|\mathbf{v}\|^2_{L^2(\partial\Omega)}\right)^{\frac{1}{2}} \\
&= 2k^{-1}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right)|||\mathbf{v}|||_{L^2(\Omega)}.
\end{aligned}$$

Thus for $\mathbf{v}\in\mathbf{H}^1(\Omega)$,

$$|a(\mathbf{u},\mathbf{v})| \le 2k^{-1}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right)|||\mathbf{v}|||_{L^2(\Omega)}. \tag{2.62}$$

Note for $\mathbf{v}\in\hat{\boldsymbol{\mathcal{V}}}$, $|||\mathbf{v}|||_{L^2(\Omega)} \le \|\mathbf{v}\|_E$. Hence,

$$|a(\mathbf{u},\mathbf{v})| \le 2k^{-1}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right)\|\mathbf{v}\|_E. \tag{2.63}$$

Substituting (2.62) and (2.63) into the generalized weak coercivity condition given in Theorem 2.4.3 gives

$$\begin{aligned}
\frac{1}{\gamma}\|\mathbf{u}\|_E &\le \sup_{\mathbf{v}\in\mathbf{H}^2(\Omega)} \frac{\left|\operatorname{Im} a(\mathbf{u},\mathbf{v})\right|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\mathbf{H}^1(\Omega)} \frac{\left|\operatorname{Re} a(\mathbf{u},\mathbf{v})\right|}{|||\mathbf{v}|||_E} \\
&\le \sup_{\mathbf{v}\in\mathbf{H}^2(\Omega)} \frac{|a(\mathbf{u},\mathbf{v})|}{\|\mathbf{v}\|_E} + \sup_{\mathbf{v}\in\mathbf{H}^1(\Omega)} \frac{|a(\mathbf{u},\mathbf{v})|}{|||\mathbf{v}|||_E} \\
&\le 4k^{-1}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right).
\end{aligned}$$

Thus,

$$\|\mathbf{u}\|_E \leq \frac{4\gamma}{k}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\Big(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\Big). \tag{2.64}$$

Recall $\mathbf{E}=\mathbf{u}-\mathbf{F}$. Thus (2.60) and (2.64) yield

$$\begin{aligned}\|\mathbf{E}\|_E \leq \frac{4\gamma}{k}\left(\frac{1}{c_1}+\frac{1}{c_+}\right)\Big(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\Big)\\ +\frac{(c_1+R)^{\frac{1}{2}}}{k}\|\mathbf{f}\|_{\Omega}+\frac{R^{\frac{1}{2}}}{k}\|\operatorname{div}\mathbf{f}\|_{L^2(\Omega)}.\end{aligned}$$

□

Remark 2.5.4. *(a) The above solution estimates ensure uniqueness of the solution to each Helmholtz-type problem in their respective solution spaces.*

(b)The adjoint problem for each Helmholtz-type problem differs only in the sign of the boundary integral terms. For this reason, all of the results of this chapter also hold for these adjoint problems. In particular, the uniqueness results. By the Fredholm Alternative Principle this ensures existence of the solutions to the Helmholtz-type problems in their respective solution spaces.

Chapter 3

Absolutely Stable Discontinuous Galerkin Methods for the Elastic Helmholtz Equations

As is the case for the scalar Helmholtz equation, the angular frequency ω plays a key role in the analysis and implementation of any numerical method used to solve the elastic Helmholtz equations. It is a well-known fact that in order to resolve the wave numerically one must use some minimum number of grid points in each wave length $\ell = 2\pi/\omega$ in every coordinate direction. This yields the minimum mesh constraint $\omega h = O(1)$, where h is the mesh size parameter. In fact, the widely held "rule of thumb" is to use 6–12 grid points per wave length. In [54], Babŭska *et al* proved the necessity of this "rule of thumb" in the 1-dimensional case for the scalar Helmholtz equation. In [54], it was also shown that the H^1 error bound for the finite element solution contains a pollution term that contributes to the loss of stability for this method in the case of a large ω. The pollution term also causes the error to increase as ω increases under the mesh constraint $\omega h = O(1)$. This forces one to adopt a more stringent mesh condition to guarantee an accurate numerical solution for high frequency waves.

The loss of stability of the standard finite element method applied to Helmholtz-type problems is an important issue to address and a fundamental limitation to overcome. Specifically, in [6, 29, 30], a strict mesh condition of $\omega^2 h = O(1)$ (called the asymptotic mesh constraint) was required to obtain optimal and quasi-optimal error estimates for finite element approximations applied to the scalar Helmholtz equation. In [26], this same mesh condition was used to obtain error estimates for the elastic Helmholtz equations. Requiring such a stringent mesh constraint makes the use of a practical coarse mesh space impossible in the case that ω is large. This is a hurdle that must be overcome if one wishes to use multi-level algebraic solvers, such as the multi-grid method or multi-level domain decomposition method.

Thus, it is the goal of this chapter to develop and analyze an interior penalty discontinuous Galerkin (IP-DG) method that will be absolutely stable for the elastic Helmholtz equations. In other words, a method in which a-priori solution estimates can be obtained for any $\omega, h > 0$. This chapter follows the example of [42, 79, 43] which give similar methods for the scalar Helmholtz equation and the time-harmonic Maxwell's equations.

Section 3.1 introduces standard notation required to formulate a discontinuous Galerkin method and presents the IP-DG method. Also, in this section, some key properties of the proposed method are demonstrated. In Section 3.2, error estimates are obtained for the asymptotic mesh regime (i.e. when $\omega^2 h \leq C$). To accomplish this, we define and analyze a specific elliptic projection operator for the elastic Helmholtz equations. With this projection operator, Schatz argument is carried out to obtain the optimal error estimates. Section 3.3 is devoted to establishing stability and error estimates for the pre-asymptotic mesh regime (i.e. when $\omega^2 h > C$). This is an important feature of the IP-DG method proposed in this chapter since it has not been shown that previous discretization techniques yield stability in this mesh regime. Section 3.4 is devoted to numerical experiments that validate properties of the proposed IP-DG method and compare it to the standard finite element method.

3.1 Formulation of the IP-DG Method

In this section, an interior penalty discontinuous Galerkin (IP-DG) method for the elastic Helmholtz equations is formulated. This formulation will follow those in [42, 44]. The methods referenced in the previous papers are absolutely stable, (i.e. stable for all $\omega, h > 0$) a trait sought in the discretization methods for Helmholtz-type problems.

First, some standard IP-DG notation needs to be introduced. Let $\mathcal{T}_h$ be a shape regular partition of the domain Ω, such that for each cell $K \in \mathcal{T}_h$, $h_K := \text{diam}(K)$. Also, for each edge/face e of a cell K, define $h_e := \text{diam}(e)$. $\mathcal{T}_h$ is called shape regular if there exist positive constants m_1, m_2 such that for any $K \in \mathcal{T}_h$ and e an edge/face of K the following inequality holds:

$$m_1 h_e \leq h_K \leq m_2 h_e.$$

The partition $\mathcal{T}_h$ is parameterized by h, which denotes the maximum spatial cell size, i.e. $h := \max_{K \in \mathcal{T}_h}\{h_K\}$. We note that the discontinuous Galerkin (DG) methodology allows greater flexibility in terms of meshing the domain Ω. In particular, one can use any polyhedral elements in the partition. In some cases, the partition is made of elements with curved boundaries.

Let

$$\begin{aligned} \mathcal{E}_h^I &:= \text{ set of all interior edges/faces of } \mathcal{T}_h, \\ \mathcal{E}_h^B &:= \text{ set of all boundary edges/faces of } \mathcal{T}_h \text{ on } \partial\Omega. \end{aligned}$$

For any edge $e \in \mathcal{E}_h^I$, let $K_e, K_e' \in \mathcal{T}_h$ such that $e = \partial K_e \cap \partial K_e'$. For such an edge e, we define the following jump and average operators:

$$[\mathbf{v}]|_e := \begin{cases} \mathbf{v}|_{K_e} - \mathbf{v}|_{K_e'}, \text{ if the global labeling number of } K_e \text{ is greater than that of } K_e', \\ \mathbf{v}|_{K_e'} - \mathbf{v}|_{K_e}, \text{ if the global labeling number of } K_e' \text{ is greater than that of } K_e, \end{cases}$$

and

$$\{\mathbf{v}\}|_e := \frac{1}{2}\left(\mathbf{v}|_{K_e} + \mathbf{v}|_{K_e'}\right).$$

For $e \in \mathcal{E}_h^B$, we use the convention $[\mathbf{v}]|_e = \{\mathbf{v}\}|_e := \mathbf{v}|_e$. Also keeping in mind the idea of cell by cell integration by parts, the outward normal vector $\mathbf{n}_e$ to $e \in \mathcal{E}_h^I$ will need to be defined. Let $\mathbf{n}_e$ be the unit outward normal vector to K_e on e, where $e = \partial K_e \cap \partial K_e'$ and K_e has a bigger global labeling number than that of K_e'.

Define the DG energy space $\mathbf{E}$ as

$$\mathbf{E} := \prod_{K \in \mathcal{T}_h} \mathbf{H}^2(K).$$

Note that unlike a conforming finite element method, this energy space can be discontinuous across cell boundaries. Multiplying (1.4) by some $\mathbf{v} \in \mathbf{E}$ and integrating

by parts piecewisely we get

$$\begin{aligned}
(\mathbf{f},\mathbf{v})_\Omega &= -(\mathbf{div}\,(\sigma(\mathbf{u})),\mathbf{v})_\Omega - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\
&= -\sum_{K\in\mathcal{T}_h}(\mathbf{div}\,(\sigma(\mathbf{u})),\mathbf{v})_K - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\
&= \sum_{K\in\mathcal{T}_h}\big(\lambda(\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v})_K + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_K\big) - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\
&\quad - \sum_{K\in\mathcal{T}_h}\langle\sigma(\mathbf{u})\mathbf{n}_K,\mathbf{v}\rangle_{\partial K} \\
&= \sum_{K\in\mathcal{T}_h}\big(\lambda(\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v})_K + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_K\big) - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\
&\quad - \sum_{e\in\mathcal{E}_h^I}\int_e[\sigma(\mathbf{u})\mathbf{n}_e\cdot\overline{\mathbf{v}}]\mathrm{d}S - \sum_{e\in\mathcal{E}_h^B}\langle\sigma(\mathbf{u})\mathbf{n}_e,\mathbf{v}\rangle_e\,.
\end{aligned}$$

To the above identity we apply (1.5) along with a well-known identity concerning the jump of a product, i.e. $[\mathbf{a}\cdot\mathbf{b}] = \{\mathbf{a}\}\cdot[\mathbf{b}] + [\mathbf{a}]\cdot\{\mathbf{b}\}$, and get

$$\begin{aligned}
(\mathbf{f},\mathbf{v})_\Omega + \langle\mathbf{g},\mathbf{v}\rangle_{\partial\Omega} = &\sum_{K\in\mathcal{T}_h}\big(\lambda(\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v})_K + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_K\big) - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\
&- \sum_{e\in\mathcal{E}_h^I}\big(\langle\{\sigma(\mathbf{u})\mathbf{n}_e\},[\mathbf{v}]\rangle_e + \langle[\sigma(\mathbf{u})\mathbf{n}_e],\{\mathbf{v}\}\rangle_e\big) + \mathbf{i}\omega\langle A\mathbf{u},\mathbf{v}\rangle_e\,.
\end{aligned}$$

Now assuming that the solution $\mathbf{u}$ of (1.4)–(1.5) is smooth enough, we find $[\sigma(\mathbf{u})\mathbf{n}_e] = [\mathbf{u}] = 0$ on all $e\in\mathcal{E}_h^I$. Thus the above identity is equivalent to

$$\begin{aligned}
(\mathbf{f},\mathbf{v})_\Omega + \langle\mathbf{g},\mathbf{v}\rangle_{\partial\Omega} = &\sum_{K\in\mathcal{T}_h}\big(\lambda(\operatorname{div}\mathbf{u},\operatorname{div}\mathbf{v})_K + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_K\big) - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \qquad (3.1)\\
&- \sum_{e\in\mathcal{E}_h^I}\big(\langle\{\sigma(\mathbf{u})\mathbf{n}_e\},[\mathbf{v}]\rangle_e + \eta\langle[\mathbf{u}],\{\sigma(\mathbf{v})\mathbf{n}_e\}\rangle_e\big) + \mathbf{i}\omega\langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega}\,. \qquad (3.2)
\end{aligned}$$

The term $\eta\langle[\mathbf{u}],\{\sigma(\mathbf{v})\mathbf{n}_e\}\rangle_e$ is introduced as a possible avenue toward symmetrizing the RHS. Thus, η is called a symmetrization parameter. It is standard for one

to consider three possible values for η. These values are $\eta = 1$ for a symmetric formulation, $\eta = 0$, and $\eta = -1$ for an anti-symmetric formulation. In this chapter, we only focus on the symmetric case with $\eta = 1$. Here it is left as a variable parameter to offer different formulations of this method.

An important aspect of the IP-DG methods is the use of penalty terms to ensure the coercivity of the sesquilinear forms involved in the formulation. To this end, we introduce two penalty sesquilinear forms $J_0(\cdot,\cdot)$ and $J_1(\cdot,\cdot)$. They are defined for $\mathbf{w}, \mathbf{v} \in \mathbf{E}$ as

$$
\begin{aligned}
J_0(\mathbf{w},\mathbf{v}) &:= \sum_{e\in\mathcal{E}_h^I} \frac{\gamma_{0,e}}{h_e} \big\langle [\mathbf{w}], [\mathbf{v}] \big\rangle_e, \\
J_1(\mathbf{w},\mathbf{v}) &:= \sum_{e\in\mathcal{E}_h^I} \gamma_{1,e} h_e \big\langle \, [\sigma(\mathbf{w})\mathbf{n}_e] \, , [\sigma(\mathbf{v})\mathbf{n}_e] \, \big\rangle_e,
\end{aligned}
$$

where $\gamma_{0,e}, \gamma_{1,e} > 0$ are called the penalty parameters for $e \in \mathcal{E}_h^I$. Note that by the smoothness of the solution $\mathbf{u}$ we have $J_0(\mathbf{u},\mathbf{v}) = J_1(\mathbf{u},\mathbf{v}) = 0$. Thus, (3.2) can be rewritten as

$$
\begin{aligned}
(\mathbf{f},\mathbf{v})_\Omega + \langle \mathbf{g},\mathbf{v}\rangle_{\partial\Omega} = \sum_{K\in\mathcal{T}_h} \big(\lambda(\operatorname{div}\mathbf{u}, \operatorname{div}\mathbf{v})_K + 2\mu(\varepsilon(\mathbf{u}),\varepsilon(\mathbf{v}))_K\big) - \omega^2\rho(\mathbf{u},\mathbf{v})_\Omega \\
- \sum_{e\in\mathcal{E}_h^I} \big(\langle\{\sigma(\mathbf{u})\mathbf{n}_e\}, [\mathbf{v}]\rangle_e + \eta \langle [\mathbf{u}], \{\sigma(\mathbf{v})\mathbf{n}_e\}\rangle_e \big) \\
+ \mathbf{i}\big(J_0(\mathbf{u},\mathbf{v}) + J_1(\mathbf{u},\mathbf{v})\big) + \mathbf{i}\omega \langle A\mathbf{u},\mathbf{v}\rangle_{\partial\Omega}.
\end{aligned}
$$

Therefore, $\mathbf{u}$ satisfies

$$
A_h(\mathbf{u},\mathbf{v}) = (\mathbf{f},\mathbf{v})_\Omega + \langle \mathbf{g},\mathbf{v}\rangle_{\partial\Omega} \qquad \forall \mathbf{v} \in \mathbf{E}, \tag{3.3}
$$

where $A_h(\cdot,\cdot)$ is defined on $\mathbf{E}\times\mathbf{E}$ as

$$A_h(\mathbf{w},\mathbf{v}) := a_h(\mathbf{w},\mathbf{v}) - \omega^2\rho(\mathbf{w},\mathbf{v})_\Omega + i\omega\langle A\mathbf{w},\mathbf{v}\rangle_{\partial\Omega}, \tag{3.4}$$

$$\begin{aligned} a_h(\mathbf{w},\mathbf{v}) := & \sum_{K\in\mathcal{T}_h}\Big(\lambda\big(\operatorname{div}\mathbf{w},\operatorname{div}\mathbf{v}\big)_K + 2\mu\big(\varepsilon(\mathbf{w}),\varepsilon(\mathbf{v})\big)_K\Big) \\ & - \sum_{e\in\mathcal{E}_h^I}\Big(\big\langle\,\{\sigma(\mathbf{w})\mathbf{n}_e\}\,,[\mathbf{v}]\big\rangle_e + \eta\big\langle[\mathbf{w}],\{\sigma(\mathbf{v})\mathbf{n}_e\}\,\big\rangle_e\Big) \\ & + \mathbf{i}\big(J_0(\mathbf{w},\mathbf{v}) + J_1(\mathbf{w},\mathbf{v})\big). \end{aligned} \tag{3.5}$$

With the DG sesquilinear form $A_h(\cdot,\cdot)$ in hand, a discrete function space is needed to formulate the IP-DG method. For this chapter, only piecewise linear polynomial functions over the partition $\mathcal{T}_h$ will be considered. Thus the IP-DG approximation space $\mathbf{V}_h$ is defined as

$$\mathbf{V}_h := \prod_{K\in\mathcal{T}_h}\mathbf{P}_1(K).$$

With all the building blocks in place, our IP-DG method is defined by seeking $\mathbf{u}_h\in\mathbf{V}_h$ such that

$$A_h(\mathbf{u}_h,\mathbf{v}_h) = \big(\mathbf{f},\mathbf{v}_h\big)_\Omega + \big\langle\mathbf{g},\mathbf{v}_h\big\rangle_{\partial\Omega} \qquad \forall\mathbf{v}_h\in\mathbf{V}_h. \tag{3.6}$$

3.1.1 Some Properties of the IP-DG Method

In this subsection, some useful properties of the above IP-DG method (3.6) are established. From this point on, we only consider the case $\eta = 1$ for simplicity. Also, we assume there exists a positive constant C such that

$$\begin{aligned} h_K \le h \le Ch_K \qquad & \forall K\in\mathcal{T}_h, \\ \gamma_{0,e}\le\gamma_0\le C\gamma_{0,e} \qquad & \forall e\in\mathcal{E}_h^I, \\ \gamma_{1,e}\le\gamma_1\le C\gamma_{1,e} \qquad & \forall e\in\mathcal{E}_h^I. \end{aligned}$$

The above constraints are not necessary for the analysis of this chapter, but rather are in place to make the constants in the inequalities more tractable.

In order to analyze the proposed IP-DG method, we introduce the following semi-norms:

$$|\mathbf{v}|_{1,h} := \left(\sum_{K \in \mathcal{T}_h} \lambda \|\operatorname{div} \mathbf{v}\|^2_{L^2(K)} + 2\mu \|\varepsilon(\mathbf{v})\|^2_{L^2(K)} \right)^{\frac{1}{2}},$$

$$\|\mathbf{v}\|_{1,h} := \left(|\mathbf{v}|^2_{1,h} + J_0(\mathbf{v},\mathbf{v}) + J_1(\mathbf{v},\mathbf{v}) \right)^{\frac{1}{2}},$$

$$|||\mathbf{v}|||_{1,h} := \left(\|\mathbf{v}\|^2_{1,h} + \sum_{e \in \mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \|\{\sigma(\mathbf{v})\mathbf{n}_e\}\|^2_{L^2(e)} \right)^{\frac{1}{2}}.$$

Note on $\mathbf{V}_h$, the semi-norms $\|\cdot\|_{1,h}$ and $|||\cdot|||_{1,h}$ are equivalent. This is trivial since $\mathbf{V}_h$ is a finite dimensional vector space. On the other hand, it can be shown that this equivalence is independent of the dimension of $\mathbf{V}_h$. This result is non-trivial, and thus it is proved in a lemma below.

To prove the equivalence, we need two inequalities which hold for polynomial functions. Namely, the inverse and trace inequalities as given below:

$$\|\mathbf{v}_h\|_{L^2(e)} \leq C h_e^{-\frac{1}{2}} \|\mathbf{v}_h\|_{L^2(K)} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{3.7}$$

$$\|\sigma(\mathbf{v}_h)\mathbf{n}_e\|_{L^2(e)} \leq C h_e^{-\frac{1}{2}} \|\sigma(\mathbf{v}_h)\|_{L^2(K)} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{3.8}$$

where $K \in \mathcal{T}_h$ and e is an edge/face of K. These inequalities will be used throughout the rest of this chapter.

Lemma 3.1.1. *For any $\mathbf{v}_h \in \mathbf{V}_h$, there holds*

$$\|\mathbf{v}_h\|_{1,h} \leq |||\mathbf{v}_h|||_{1,h} \leq C\xi^{\frac{1}{2}} \|\mathbf{v}_h\|_{1,h},$$

where $\xi := \left(1 + \frac{1}{\gamma_0}\right)$ and C is a positive constant independent of $\omega, h, \gamma_0, \gamma_1$.

Proof. Note that the first inequality is trivial. To show the second inequality we use (3.8) as follows:

$$\begin{aligned}
|||\mathbf{v}_h|||_{1,h}^2 &= \|\mathbf{v}\|_{1,h}^2 + \sum_{e\in\mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \|\{\sigma(\mathbf{v})\mathbf{n}_e\}\|_{L^2(e)}^2 \\
&\leq \|\mathbf{v}_h\|_{1,h}^2 + C\sum_{e\in\mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \cdot \frac{1}{h_e} \left(\|\sigma(\mathbf{v})\|_{L^2(K_e)}^2 + \|\sigma(\mathbf{v})\|_{L^2(K_e')}^2 \right) \\
&\leq \|\mathbf{v}_h\|_{1,h}^2 + \frac{C(\lambda+2\mu)}{\gamma_0} \sum_{K\in\mathcal{T}_h} \left(\lambda \|\operatorname{div} \mathbf{v}_h\|_K^2 + 2\mu \|\varepsilon(\mathbf{v}_h)\|_{L^2(K)}^2 \right) \\
&\leq C\left(1 + \frac{\lambda+2\mu}{\gamma_0}\right) \|\mathbf{v}_h\|_{1,h}^2.
\end{aligned}$$

□

In order to show that (3.6) is well posed, we need to verify that $A_h(\cdot,\cdot)$ is both continuous and weakly coercive on $\mathbf{V}_h$. Weak coercivity in this case relies on the fact that $\mathbf{V}_h$ is made up of piecewise linear polynomials. This fact infers the following lemma.

Lemma 3.1.2. *For any* $0 < \delta < 1$ *and* $\mathbf{v}_h \in \mathbf{V}_h$,

$$\begin{aligned}
|\mathbf{v}_h|_{1,h}^2 \leq \delta\omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 &+ \frac{C_\delta(\lambda+2\mu)}{\omega h}\omega\|\mathbf{v}_h\|_{L^2(\partial\Omega)}^2 \\
&+ \frac{C_\delta(\lambda+2\mu)}{\gamma_0} J_0(\mathbf{v}_h,\mathbf{v}_h) + \frac{C_\delta}{\omega^2\rho h^2\gamma_1} J_1(\mathbf{v}_h,\mathbf{v}_h).
\end{aligned} \tag{3.9}$$

Proof. Note that $\mathbf{v}_h|_K \in \mathbf{P}_1(K)$ for all $K \in \mathcal{T}_h$ and thus $\mathbf{div}\left(\sigma(\mathbf{v}_h)\right)|_K = \mathbf{0}$. For any $\mathbf{w}_h, \mathbf{v}_h \in \mathbf{V}_h$ and $K \in \mathcal{T}_h$, integrating by parts yields

$$\begin{aligned}
0 &= \left(\mathbf{div}\,\sigma(\mathbf{v}_h), \mathbf{w}_h\right)_K \\
&= -\left\langle \sigma(\mathbf{v}_h)\mathbf{n}_K, \mathbf{w}_h \right\rangle_{\partial K} + \lambda\left(\operatorname{div}\mathbf{v}_h, \operatorname{div}\mathbf{w}_h\right)_K + 2\mu\left(\varepsilon(\mathbf{v}_h), \varepsilon(\mathbf{w}_h)\right)_K.
\end{aligned}$$

Thus,

$$\lambda\big(\operatorname{div}\mathbf{v}_h, \operatorname{div}\mathbf{w}_h\big)_K + 2\mu\big(\varepsilon(\mathbf{v}_h), \varepsilon(\mathbf{w}_h)\big)_K = \big\langle \sigma(\mathbf{v}_h)\mathbf{n}_K, \mathbf{w}_h\big\rangle_{\partial K}.$$

Setting $\mathbf{w}_h = \mathbf{v}_h$ and summing over all $K \in \mathcal{T}_h$ gives

$$\begin{aligned}
|\mathbf{v}_h|_{1,h}^2 &= \sum_{K\in\mathcal{T}_h} \big\langle \sigma(\mathbf{v}_h)\mathbf{n}_K, \mathbf{v}_h\big\rangle_{\partial K} \\
&= \sum_{e\in\mathcal{E}_h^I} \Big(\big\langle \{\sigma(\mathbf{v}_h)\mathbf{n}_e\}, [\mathbf{v}_h]\big\rangle_e + \big\langle [\sigma(\mathbf{v}_h)\mathbf{n}_e], \{\mathbf{v}_h\}\big\rangle_e \Big) + \sum_{e\in\mathcal{E}_h^B} \big\langle \sigma(\mathbf{v}_h)\mathbf{n}_e, \mathbf{v}_h\big\rangle_e \\
&\le \sum_{e\in\mathcal{E}_h^I} \big(\|\{\sigma(\mathbf{v}_h)\mathbf{n}_e\}\|_{L^2(e)} \|[\mathbf{v}_h]\|_{L^2(e)} + \|[\sigma(\mathbf{v}_h)\mathbf{n}_e]\|_{L^2(e)} \|\{\mathbf{v}_h\}\|_{L^2(e)} \big) \\
&\quad + \sum_{e\in\mathcal{E}_h^B} \|\sigma(\mathbf{v}_h)\mathbf{n}_e\|_{L^2(e)} \|\mathbf{v}_h\|_{L^2(e)}.
\end{aligned}$$

By the trace and inverse inequalities (3.7) and (3.8) along with the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
|\mathbf{v}_h|_{1,h}^2 \le C &\sum_{e\in\mathcal{E}_h^B} h_e^{-\frac{1}{2}} \|\sigma(\mathbf{v}_h)\|_{L^2(K_e)} \|\mathbf{v}_h\|_{L^2(e)} \\
&+ C \sum_{e\in\mathcal{E}_h^I} h_e^{-\frac{1}{2}} \|[\mathbf{v}_h]\|_{L^2(e)} \left(\|\sigma(\mathbf{v}_h)\|_{L^2(K_e)} + \|\sigma(\mathbf{v}_h)\|_{L^2(K_e')} \right) \\
&+ C \sum_{e\in\mathcal{E}_h^I} h_e^{-\frac{1}{2}} \|[\sigma(\mathbf{v}_h)\mathbf{n}_e]\|_{L^2(e)} \left(\|\mathbf{v}_h\|_{L^2(K_e)} + \|\mathbf{v}_h\|_{L^2(K_e')} \right).
\end{aligned}$$

Finally, it follows from the discrete Cauchy-Schwarz inequality along with Young's inequality that

$$
\begin{aligned}
|\mathbf{v}_h|_{1,h}^2 &\leq Ch^{-\frac{1}{2}} \left(\sum_{K\in\mathcal{T}_h} \|\sigma(\mathbf{v}_h)\|_{L^2(K)} \right)^{\frac{1}{2}} \|\mathbf{v}_h\|_{L^2(\partial\Omega)} \\
&\quad + C\gamma_0^{-\frac{1}{2}} \Big(\sum_{e\in\mathcal{E}_h^I} \frac{\gamma_{0,e}}{h_e} \|[\mathbf{v}_h]\|_{L^2(e)}^2 \Big)^{\frac{1}{2}} \Big(\sum_{K\in\mathcal{T}_h} \|\sigma(\mathbf{v}_h)\|_{L^2(K)}^2 \Big)^{\frac{1}{2}} \\
&\quad + C\gamma_1^{-\frac{1}{2}} h^{-1} \Big(\sum_{e\in\mathcal{E}_h^I} \gamma_{1,e} h_e \|[\sigma(\mathbf{v}_h)\mathbf{n}_e]\|_{L^2(e)}^2 \Big)^{\frac{1}{2}} \|\mathbf{v}_h\|_{L^2(\Omega)} \\
&\leq \delta |\mathbf{v}_h|_{1,h}^2 + \frac{C(\lambda+2\mu)}{\delta\omega h} \omega \|\mathbf{v}_h\|_{L^2(\partial\Omega)}^2 + \frac{\delta}{2} |\mathbf{v}_h|_{1,h}^2 + \frac{C(\lambda+2\mu)}{\delta\gamma_0} J_0(\mathbf{v}_h, \mathbf{v}_h) \\
&\quad + \delta(1-\delta)\omega^2\rho \|\mathbf{v}_h\|_{L^2(\Omega)}^2 + \frac{C}{\delta(1-\delta)\omega^2\rho h^2\gamma_1} J_1(\mathbf{v}_h, \mathbf{v}_h).
\end{aligned}
$$

Thus (3.9) holds. □

With this lemma in place, the following theorem establishes the continuity and weak coercivity of $A_h(\cdot,\cdot)$. We note that since A is a constant symmetric positive definite (SPD) matrix that there exists positive constants c_A, C_A such that

$$
c_A \|\mathbf{v}\|_{L^2(\partial\Omega)}^2 \leq \langle A\mathbf{v}, \mathbf{v}\rangle_{\partial\Omega} \leq C_A \|\mathbf{v}\|_{L^2(\partial\Omega)}^2 \qquad \forall \mathbf{v} \in \mathbf{E}. \tag{3.10}
$$

Theorem 3.1.3. *The sesquilinear form $A_h(\cdot,\cdot)$ is continuous on the space $\mathbf{E}$ and weakly coercive on the space $\mathbf{V}_h$. That is, there exist positive constants M, C independent ω,h,γ_0,γ_1 such that*

$$
|A_h(\mathbf{w}, \mathbf{v})| \leq M \left(|||\mathbf{w}|||_{1,h}^2 + \omega^2\rho \|\mathbf{w}\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}} \left(|||\mathbf{v}|||_{1,h}^2 + \omega^2\rho \|\mathbf{v}\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}} \tag{3.11}
$$

for all $\mathbf{w}, \mathbf{v} \in \mathbf{E}$ and

$$
|A_h(\mathbf{v}_h, \mathbf{v}_h)| \geq C\Big(\xi + \frac{1}{\omega h} + \frac{1}{\omega^2 h^2 \gamma_1} \Big)^{-1} \Big(|\mathbf{v}_h|_{1,h}^2 + \omega^2\rho \|\mathbf{v}_h\|_{L^2(\Omega)}^2 \Big), \tag{3.12}
$$

$$
|A_h(\mathbf{v}_h, \mathbf{v}_h)| \geq J_0(\mathbf{v}_h, \mathbf{v}_h) + J_1(\mathbf{v}_h, \mathbf{v}_h) + \omega \big\langle A\mathbf{v}_h, \mathbf{v}_h \big\rangle_{\partial\Omega}. \tag{3.13}
$$

for all $\mathbf{v}_h \in \mathbf{V}_h$. *Here* $\xi = 1 + \frac{1}{\gamma_0}$.

Proof. To show (3.11), we appeal to the Cauchy-Schwarz and triangle inequalities. Thus, for any $\mathbf{w}, \mathbf{v} \in \mathbf{E}$, we find

$$
\begin{aligned}
|A_h(\mathbf{w},\mathbf{v})| &\le |a_h(\mathbf{w},\mathbf{v})| + \omega^2\rho\|\mathbf{w}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} \\
&\le \|\mathbf{w}\|_{1,h}\|\mathbf{v}\|_{1,h} + \sum_{e\in\mathcal{E}_h^I} \|\{\sigma(\mathbf{w})\mathbf{n}_e\}\|_{L^2(e)}\|[\mathbf{v}]\|_{L^2(e)} \\
&\quad + \sum_{e\in\mathcal{E}_h^I} \|[\mathbf{w}]\|_{L^2(e)}\|\{\sigma(\mathbf{v})\mathbf{n}_e\}\|_{L^2(e)} + \omega^2\rho\|\mathbf{w}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} \\
&\le \|\mathbf{w}\|_{1,h}\|\mathbf{v}\|_{1,h} + \omega^2\rho\|\mathbf{w}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} \\
&\quad + \sum_{e\in\mathcal{E}_h^I} \left(\left(\frac{h_e}{\gamma_{0,e}}\right)^{\frac{1}{2}} \|\{\sigma(\mathbf{w})\mathbf{n}_e\}\|_{L^2(e)} \left(\frac{\gamma_{0,e}}{h_e}\right)^{\frac{1}{2}} \|[\mathbf{v}]\|_{L^2(e)} \right) \\
&\quad + \sum_{e\in\mathcal{E}_h^I} \left(\left(\frac{\gamma_{0,e}}{h_e}\right)^{\frac{1}{2}} \|[\mathbf{w}]\|_{L^2(e)} \left(\frac{h_e}{\gamma_{0,e}}\right)^{\frac{1}{2}} \|\{\sigma(\mathbf{v})\mathbf{n}_e\}\|_{L^2(e)} \right) \\
&\le \|\mathbf{w}\|_{1,h}\|\mathbf{v}\|_{1,h} + +\omega^2\rho\|\mathbf{w}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^2(\Omega)} \\
&\quad + \left(\sum_{e\in\mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \|\{\sigma(\mathbf{w})\mathbf{n}_e\}\|^2_{L^2(e)} \right)^{\frac{1}{2}} J_0(\mathbf{v},\mathbf{v})^{\frac{1}{2}} \\
&\quad + J_0(\mathbf{w},\mathbf{w}) \left(\sum_{e\in\mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \|\{\sigma(\mathbf{v})\mathbf{n}_e\}\|^2_{L^2(e)} \right)^{\frac{1}{2}} \\
&\le \left(\|\mathbf{w}\|^2_{1,h} + J_0(\mathbf{w},\mathbf{w}) + \sum_{e\in\mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \|\{\sigma(\mathbf{w})\mathbf{n}_e\}\|^2_{L^2(e)} + \omega^2\rho\|\mathbf{w}\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}} \\
&\quad \cdot \left(\|\mathbf{v}\|^2_{1,h} + J_0(\mathbf{v},\mathbf{v}) + \sum_{e\in\mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e}} \|\{\sigma(\mathbf{v})\mathbf{n}_e\}\|^2_{L^2(e)} + \omega^2\rho\|\mathbf{v}\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}} \\
&\le M \left(|||\mathbf{w}|||^2_{1,h} + \omega^2\rho\|\mathbf{w}\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}} \left(|||\mathbf{v}|||^2_{1,h} + \omega^2\rho\|\mathbf{v}\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}.
\end{aligned}
$$

To verify weak coercivity, for any $\mathbf{v}_h \in \mathbf{V}_h$, taking the real and imaginary parts of $A_h(\mathbf{v}_h, \mathbf{v}_h)$ yields

$$\operatorname{Re} A_h(\mathbf{v}_h, \mathbf{v}_h) = |\mathbf{v}_h|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 + 2\operatorname{Re}\sum_{e\in\mathcal{E}_h^I}\big\langle\{\sigma(\mathbf{v}_h)\mathbf{n}_e\}, [\mathbf{v}_h]\big\rangle_e, \tag{3.14}$$

$$\operatorname{Im} A_h(\mathbf{v}_h, \mathbf{v}_h) = J_0(\mathbf{v}_h, \mathbf{v}_h) + J_1(\mathbf{v}_h, \mathbf{v}_h) + \omega\big\langle A\mathbf{v}_h, \mathbf{v}_h\big\rangle_{\partial\Omega}. \tag{3.15}$$

Then (3.13) follows directly from (3.15).

To verify (3.12), we need to bound the term $\sum_{e\in\mathcal{E}_h^I}\big\langle\{\sigma(\mathbf{v}_h)\mathbf{n}_e\}, [\mathbf{v}_h]\big\rangle_e$. This step involves using the trace and inverse inequality and was already carried out in previous calculations (c.f. Lemma 3.1.2). Thus,

$$\begin{aligned}\operatorname{Re} A_h(\mathbf{v}_h, \mathbf{v}_h) &\le |\mathbf{v}_h|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 + C\gamma_0^{-\frac{1}{2}}\Big(\sum_{e\in\mathcal{E}_h^I}\frac{\gamma_0}{h_e}\|[\mathbf{v}_h]\|_{L^2(\Omega)}^2\Big)^{\frac{1}{2}}|\mathbf{v}_h|_{1,h}\\ &\le \frac{3}{2}|\mathbf{v}_h|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 + \frac{C}{\gamma_0}J_0(\mathbf{v}_h, \mathbf{v}_h).\end{aligned}$$

Combining the above inequality with (3.9) and using $\delta = \frac{1}{4}$ we get

$$\begin{aligned}\frac{1}{2}|\mathbf{v}_h|_{1,h}^2 + \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 &\le -\operatorname{Re} A_h(\mathbf{v}_h, \mathbf{v}_h) + 2|\mathbf{v}_h|_{1,h}^2 + \frac{C}{\gamma_0}J_0(\mathbf{v}_h, \mathbf{v}_h)\\ &\le -\operatorname{Re} A_h(\mathbf{v}_h, \mathbf{v}_h) + \frac{1}{2}\omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 + \frac{C}{\omega h c_A}\omega c_A\|\mathbf{v}_h\|_{L^2(\partial\Omega)}^2\\ &\quad + \frac{C}{\gamma_0}J_0(\mathbf{v}_h, \mathbf{v}_h) + \frac{C}{\omega^2\rho h^2\gamma_1}J_1(\mathbf{v}_h, \mathbf{v}_h).\end{aligned}$$

Thus, subtracting both sides of the above inequality by $\frac{1}{2}\omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2$ and using both (3.10) and (3.15) yield

$$\begin{aligned}|\mathbf{v}_h|_{1,h}^2 + \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 &\le -2\operatorname{Re} A_h(\mathbf{v}_h, \mathbf{v}_h) + C\Big(\frac{1}{\gamma_0} + \frac{1}{\omega h c_A} + \frac{1}{\omega^2\rho h^2\gamma_1}\Big)\operatorname{Im} A_h(\mathbf{v}_h, \mathbf{v}_h)\\ &\le C\Big(1 + \frac{1}{\gamma_0} + \frac{1}{\omega h c_A} + \frac{1}{\omega^2\rho h^2\gamma_1}\Big)|A_h(\mathbf{v}_h, \mathbf{v}_h)|.\end{aligned}$$

Hence, (3.12) is verified. □

Remark 3.1.4. (3.12)–(3.13) *is called weak coercivity because of the complex magnitude used in the left-hand side of these inequalities.*

Theorem 3.1.5. *For every choice of* $\omega, h, \gamma_0, \gamma_1 > 0$, $\mathbf{f} \in \mathbf{L}^2(\Omega)$, *and* $\mathbf{g} \in \mathbf{L}^2(\partial\Omega)$ *there exists a unique solution* $\mathbf{u}_h$ *of* (3.6).

Proof. This is an immediate consequence of Theorem 3.1.3 and the well-known Lax-Milgram-Babuška theorem [8, 9]. □

We note that the weak coercivity of $A_h(\cdot,\cdot)$ in (3.12) depends in an adverse way on the mesh parameter h. For this reason, this weak coercivity cannot be used to obtain optimal error estimates in the case that h is allowed to be arbitrarily small. In the case of small h, which belongs to the asymptotic mesh regime, we instead rely on a Gärding's inequality for $A_h(\cdot,\cdot)$ to derive more robust estimates. This Gärding's inequality is proved in the following theorem:

Theorem 3.1.6. *For* $\mathbf{v}_h \in \mathbf{V}_h$ *the following Gärding's inequality for* $A_h(\cdot,\cdot)$ *holds:*

$$\frac{1}{2}\|\mathbf{v}_h\|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 \leq C\xi|A_h(\mathbf{v}_h,\mathbf{v}_h)|, \tag{3.16}$$

where $\xi := \left(1 + \frac{1}{\gamma_0}\right)$ *and* C *is independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. Similar to the proof of the weak coercivity inequalities in Theorem 3.1.3, we begin by taking the real and imaginary part of $A_h(\cdot,\cdot)$ separately (c.f. (3.14)–(3.15)). Also, in a similar fashion as was done in the proofs of Theorem 3.1.2 and 3.1.3, the Cauchy-Schwarz and Young's inequality along with (3.7) and (3.8) are used to bound the term $\sum_{e\in\mathcal{E}_h^I}\left\langle\{\sigma(\mathbf{v}_h)\mathbf{n}_e\},[\mathbf{v}_h]\right\rangle_e$. Thus the following sequence of inequalities are

obtained:

$$\begin{aligned}\operatorname{Re} A_h(\mathbf{v}_h,\mathbf{v}_h) &= |\mathbf{v}_h|_{1,h}^2 - \omega^2\rho\|\mathbf{v}\|_{L^2(\Omega)}^2 + 2\operatorname{Re}\sum_{e\in\mathcal{E}_h^I}\left\langle\{\sigma(\mathbf{v}_h)\mathbf{n}_e\},[\mathbf{v}_h]\right\rangle_e \\ &\geq |\mathbf{v}_h|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 - \frac{1}{2}|\mathbf{v}_h|_{1,h}^2 - \frac{C(\lambda+2\mu)}{\gamma_0}J_0(\mathbf{v}_h,\mathbf{v}_h) \\ &\geq \frac{1}{2}|\mathbf{v}_h|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 - \frac{C(\lambda+2\mu)}{\gamma_0}\operatorname{Im} A_h(\mathbf{v}_h,\mathbf{v}_h).\end{aligned}$$

Rearranging the above inequality and adding $\frac{1}{2}$ times (3.15) yield

$$\begin{aligned}\frac{1}{2}\|\mathbf{v}_h\|_{1,h}^2 - \omega^2\rho\|\mathbf{v}_h\|_{L^2(\Omega)}^2 &\leq C\operatorname{Re} A_h(\mathbf{v}_h,\mathbf{v}_h) + C\left(1+\frac{\lambda+2\mu}{\gamma_0}\right)\operatorname{Im} A_h(\mathbf{v}_h,\mathbf{v}_h) \\ &\leq C\left(1+\frac{\lambda+2\mu}{\gamma_0}\right)|A_h(\mathbf{v}_h,\mathbf{v}_h)|.\end{aligned}$$

Hence, (3.16) holds. □

3.2 Asymptotic Error Estimates

Recall that Theorem 3.1.3 guarantees both continuity and weak coercivity of $A_h(\cdot,\cdot)$ for any positive values of $\omega, h, \gamma_0, \gamma_1$. Unfortunately, as is observed in Theorem 3.1.3, the weak coercivity inequality degrades as h becomes small. For this reason, weak coercivity of $A_h(\cdot,\cdot)$ cannot be used to obtain optimal order error estimates in the asymptotic mesh regime, i.e. $\omega^2 h = O(1)$.

Instead, a standard argument called Schatz argument (c.f. [71]) is often used to obtain error estimates in the asymptotic mesh regime. Schatz argument is useful for deriving error estimates for consistent discretizations of indefinite problems that are characterized by sesquilinear forms that satisfy a Gärding's inequality. This method has been used in the past to prove optimal order error estimates for finite element approximations of Helmholtz-type problems. For references of Schatz argument applied to finite element formulations of the scalar Helmholtz equation, see [6, 29, 30]. For an example of Schatz argument applied to a finite element approximation of the

elastic Helmholtz equation, see [26]. For a general reference of Schatz argument applied to the finite element approximations of general second order PDEs satisfying Gärding's inequality, see [16].

The consistency of (3.6) immediately infers Galerkin orthogonality on $\mathbf{V}_h$. Namely,

$$A_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{3.17}$$

where $\mathbf{u}$ solves (1.4)–(1.5) and $\mathbf{u}_h$ solves (3.6).

We also quote frequency-explicit a priori estimates for the PDE solution $\mathbf{u}$. These are taken from [27]. Note that the estimates in [27] were only carried out in the case in which $\mathbf{g} = \mathbf{0}$ but these estimates should hold as well for any $\mathbf{g} \in \mathbf{L}^2(\partial\Omega)$. This extension can be expected because the analysis in Theorem 2.5.2 holds when $\mathbf{g} \in \mathbf{L}^2(\partial\Omega)$ and this analysis is based on that of [27]. The estimate that will be used in the analysis of the proposed IP-DG method is quoted below.

Theorem 3.2.1. *Suppose that Ω is a convex polygonal domain or Ω is a smooth domain. Further suppose that $\mathbf{u} \in \mathbf{H}^2(\Omega)$ solves (1.4)–(1.5). Then $\mathbf{u}$ satisfies the following regularity estimate:*

$$\|\mathbf{u}\|_{H^2(\Omega)} \leq C \left(\omega^\alpha + \frac{1}{\omega^2}\right) \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right), \tag{3.18}$$

where $\alpha = 1$ if $\mathbf{u} \in \mathbf{V}_{\tilde{K}}$ for some positive constant $\tilde{K}$ as defined in Chapter 2 and otherwise $\alpha = 2$.

3.2.1 Elliptic Projection and its Error Estimates

The primary goal of this section is to estimate the error $\mathbf{u} - \mathbf{u}_h$ in the asymptotic mesh regime. This will be done based on the error decomposition $\mathbf{u} - \mathbf{u}_h = (\mathbf{u} - \tilde{\mathbf{u}}_h) + (\tilde{\mathbf{u}}_h - \mathbf{u}_h)$ with some $\tilde{\mathbf{u}}_h \in \mathbf{V}_h$ which is sufficiently close to $\mathbf{u}$, such as a projection of $\mathbf{u}$. To this end, for any $\mathbf{w} \in \mathbf{E}$, we define its elliptic projection $\tilde{\mathbf{w}}_h \in \mathbf{V}_h$ as the

solution to the following problem:

$$a_h(\tilde{\mathbf{w}}_h, \mathbf{v}_h) + \mathbf{i}\omega\big\langle A\tilde{\mathbf{w}}_h, \mathbf{v}_h\big\rangle_{\partial\Omega} = a_h(\mathbf{w}, \mathbf{v}_h) + \mathbf{i}\omega\big\langle A\mathbf{w}, \mathbf{v}_h\big\rangle_{\partial\Omega} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h. \tag{3.19}$$

Since the elliptic projection is defined using the bilinear form $a_h(\cdot,\cdot)$, it would be prudent to prove some properties of $a_h(\cdot,\cdot)$. The following theorem is given with this in mind.

Theorem 3.2.2. *For any $\mathbf{v}, \mathbf{w} \in \mathbf{E}$ there exists a positive constant C independent of $\omega, h, \gamma_0, \gamma_1$ such that*

$$|a_h(\mathbf{v}, \mathbf{w})| \leq C|||\mathbf{v}|||_{1,h}|||\mathbf{w}|||_{1,h}. \tag{3.20}$$

Also for any $0 < \delta < 1$ and $\mathbf{v}_h \in \mathbf{V}_h$ there holds

$$\operatorname{Re} a_h(\mathbf{v}_h, \mathbf{v}_h) + \left(1 - \delta + \frac{C_\delta}{\gamma_0}\right) \operatorname{Im} a_h(\mathbf{v}_h, \mathbf{v}_h) \geq (1-\delta)\|\mathbf{v}_h\|_{1,h}^2. \tag{3.21}$$

Proof. Note that (3.20) is easy to prove with the techniques used to prove continuity of $A_h(\cdot,\cdot)$ and thus we omit it. By using Cauchy-Schwarz, trace, inverse, and Young's inequalities we get

$$\begin{aligned}
\operatorname{Re} a_h(\mathbf{v}_h, \mathbf{v}_h) &\geq |\mathbf{v}_h|_{1,h}^2 + 2\operatorname{Re}\sum_{e\in\mathcal{E}_h^I} \big\langle [\mathbf{v}_h], \{\sigma(\mathbf{v}_h)\mathbf{n}_e\}\big\rangle_e \\
&\geq |\mathbf{v}_h|_{1,h}^2 - 2\sum_{e\in\mathcal{E}_h^I} \|[\mathbf{v}_h]\|_{L^2(e)}\|\{\sigma(\mathbf{v}_h)\mathbf{n}_e\}\|_{L^2(e)} \\
&\geq |\mathbf{v}_h|_{1,h}^2 - \sum_{e\in\mathcal{E}_h^I} \frac{C}{\gamma_0^{\frac{1}{2}}} \frac{\gamma_0^{\frac{1}{2}}}{h_e^{\frac{1}{2}}} \|[\mathbf{v}_h]\|_{L^2(e)} \left(\|\sigma(\mathbf{v}_h)\|_{L^2(K_e)} + \|\sigma(\mathbf{v}_h)\|_{L^2(K_e')}\right) \\
&\geq (1-\delta)|\mathbf{v}_h|_{1,h}^2 - \frac{C_\delta}{\gamma_0} J_0(\mathbf{v}_h, \mathbf{v}_h).
\end{aligned}$$

Also,

$$\operatorname{Im} a_h(\mathbf{v}_h, \mathbf{v}_h) = J_0(\mathbf{v}_h, \mathbf{v}_h) + J_1(\mathbf{v}_h, \mathbf{v}_h).$$

Combining these two results yields (3.21). □

Using this theorem, it is easy to check that the bilinear form $a_h(\cdot,\cdot) + \mathbf{i}\omega\langle A\cdot,\cdot\rangle_{\partial\Omega}$ is both continuous and coercive when γ_0 is large enough. Hence, by the Lax-Milgram theorem, the above elliptic projection is well-defined.

Trivially, the following Galerkin orthogonality holds.

Lemma 3.2.3. *Suppose that* $\mathbf{w} \in \mathbf{E}$ *and* $\tilde{\mathbf{w}}_h \in \mathbf{V}_h$ *is its elliptic projection then*

$$a_h(\mathbf{w} - \tilde{\mathbf{w}}_h, \mathbf{v}_h) + \mathbf{i}\omega\big\langle A(\mathbf{w} - \tilde{\mathbf{w}}_h), \mathbf{v}_h\big\rangle = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}_h. \tag{3.22}$$

Theorem 3.2.4. *Let* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *solve* (1.4)–(1.5) *and let* $\tilde{\mathbf{u}}_h \in \mathbf{V}_h$ *be its elliptic projection defined in* (3.19). *Then the following estimates hold:*

$$|||\mathbf{u} - \tilde{\mathbf{u}}_h|||_{1,h} + \omega^{\frac{1}{2}}\xi\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\partial\Omega)} \tag{3.23}$$

$$\leq C\xi^2 h\big(\xi + \gamma_1 + \omega h\big)^{\frac{1}{2}} \left(\omega^\alpha + \frac{1}{\omega^2}\right) \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right), \tag{3.24}$$

and

$$\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\Omega)} \leq C\xi^2 h^2\big(\xi + \gamma_1 + \omega h\big) \left(\omega^\alpha + \frac{1}{\omega^2}\right) \left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right), \tag{3.25}$$

where $\xi = 1 + \gamma_0^{-1}$, α *is defined in Theorem 3.2.1, and* C *is a positive constant independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. Let $\hat{\mathbf{u}}_h \in \mathbf{V_h}$ denote the P_1 conforming finite element interpolant of $\mathbf{u}$ on $\mathcal{T}_h$. Then the following estimates are well-known (c.f. [16, 24]):

$$\|\mathbf{u} - \hat{\mathbf{u}}_h\|_{L^2(\Omega)} \leq Ch^2|\mathbf{u}|_{H^2(\Omega)} \qquad \text{and} \qquad \|\boldsymbol{\nabla}(\mathbf{u} - \hat{\mathbf{u}}_h)\|_{L^2(\Omega)} \leq Ch|\mathbf{u}|_{H^2(\Omega)}. \tag{3.26}$$

Applying the trace and inverse inequalities to these estimates yields

$$|||\mathbf{u} - \hat{\mathbf{u}}_h|||_{1,h} \leq C(\xi + \gamma_1)^{\frac{1}{2}} h |\mathbf{u}|_{H^2(\Omega)}, \tag{3.27}$$

and

$$\|\mathbf{u} - \hat{\mathbf{u}}_h\|_{L^2(\partial\Omega)} \leq C h^{\frac{3}{2}} |\mathbf{u}|_{H^2(\Omega)}. \tag{3.28}$$

Set $\boldsymbol{\psi}_h := \tilde{\mathbf{u}}_h - \hat{\mathbf{u}}_h$. By Galerkin orthogonality along with the fact that $\boldsymbol{\psi}_h + \mathbf{u} - \tilde{\mathbf{u}}_h = \mathbf{u} - \hat{\mathbf{u}}_h$ we get

$$a_h(\boldsymbol{\psi}_h, \boldsymbol{\psi}_h) + \mathbf{i}\omega \big\langle A\boldsymbol{\psi}_h, \boldsymbol{\psi}_h \big\rangle_{\partial\Omega} = a_h(\mathbf{u} - \hat{\mathbf{u}}_h, \boldsymbol{\psi}_h) + \mathbf{i}\omega \big\langle A(\mathbf{u} - \hat{\mathbf{u}}_h), \boldsymbol{\psi}_h \big\rangle_{\partial\Omega}.$$

Next, it follows from Theorem 3.2.2 with $\delta = \frac{1}{2}$ and Lemma 3.1.1 that

$$
\begin{aligned}
\frac{1}{2}|||\boldsymbol{\psi}_h|||_{1,h}^2 &\le C\xi\|\boldsymbol{\psi}_h\|_{1,h}^2 \\
&\le C\xi\,\mathrm{Re}\,a_h(\boldsymbol{\psi}_h,\boldsymbol{\psi}_h) + C\xi\left(\frac{1}{2}+\frac{C_{\frac{1}{2}}}{\gamma_0}\right)\mathrm{Im}\,a_h(\boldsymbol{\psi}_h,\boldsymbol{\psi}_h) \\
&= C\xi\,\mathrm{Re}\left(a_h(\boldsymbol{\psi}_h,\boldsymbol{\psi}_h)+\mathbf{i}\omega\big\langle A\boldsymbol{\psi}_h,\boldsymbol{\psi}_h\big\rangle_{\partial\Omega}\right) - C\xi\omega\left(\frac{1}{2}+\frac{C_{\frac{1}{2}}}{\gamma_0}\right)\big\langle A\boldsymbol{\psi}_h,\boldsymbol{\psi}_h\big\rangle_{\partial\Omega} \\
&\quad + C\xi\left(\frac{1}{2}+\frac{C_{\frac{1}{2}}}{\gamma_0}\right)\mathrm{Im}\left(a_h(\boldsymbol{\psi}_h,\boldsymbol{\psi}_h)+\mathbf{i}\omega\big\langle A\boldsymbol{\psi}_h,\boldsymbol{\psi}_h\big\rangle_{\partial\Omega}\right) \\
&= C\xi\,\mathrm{Re}\left(a_h(\mathbf{u}-\hat{\mathbf{u}}_h,\boldsymbol{\psi}_h)+\mathbf{i}\omega\big\langle A(\mathbf{u}-\hat{\mathbf{u}}_h),\boldsymbol{\psi}_h\big\rangle_{\partial\Omega}\right) \\
&\quad + C\xi\left(\frac{1}{2}+\frac{C_{\frac{1}{2}}}{\gamma_0}\right)\mathrm{Im}\left(a_h(\mathbf{u}-\hat{\mathbf{u}}_h,\boldsymbol{\psi}_h)+\mathbf{i}\omega\big\langle A(\mathbf{u}-\hat{\mathbf{u}}_h),\boldsymbol{\psi}_h\big\rangle_{\partial\Omega}\right) \\
&\quad - C\xi\omega\left(\frac{1}{2}+\frac{C_{\frac{1}{2}}}{\gamma_0}\right)\big\langle A\boldsymbol{\psi}_h,\boldsymbol{\psi}_h\big\rangle_{\partial\Omega} \\
&\le C\xi|||\boldsymbol{\psi}_h|||_{1,h}|||\mathbf{u}-\hat{\mathbf{u}}|||_{1,h} + C\omega\xi\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)}\|\mathbf{u}-\hat{\mathbf{u}}\|_{L^2(\partial\Omega)} - C\omega\xi^2\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)}^2 \\
&\quad + C\xi^2\Big(|||\boldsymbol{\psi}_h|||_{1,h}|||\mathbf{u}-\hat{\mathbf{u}}|||_{1,h} + \omega\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)}\|\mathbf{u}-\hat{\mathbf{u}}\|_{L^2(\partial\Omega)}\Big) \\
&\le C\xi^2|||\boldsymbol{\psi}_h|||_{1,h}|||\mathbf{u}-\hat{\mathbf{u}}|||_{1,h} + 2C\omega\xi^2\|\mathbf{u}-\hat{\mathbf{u}}_h\|_{L^2(\partial\Omega)}^2 - \frac{C}{4}\omega\xi^2\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)} \\
&\le C\xi^4|||\mathbf{u}-\hat{\mathbf{u}}_h|||_{1,h}^2 + \frac{1}{4}|||\boldsymbol{\psi}_h|||_{1,h}^2 - \frac{C}{4}\omega\xi^2\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)}^2 + 2C\omega\xi^2\|\mathbf{u}-\hat{\mathbf{u}}_h\|_{L^2(\partial\Omega)}^2.
\end{aligned}
$$

Substituting (3.27) and (3.28) into the above estimate gives

$$
\begin{aligned}
|||\boldsymbol{\psi}_h|||_{1,h}^2 + \omega\xi^2\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)}^2 &\le C\Big(\xi^4|||\mathbf{u}-\hat{\mathbf{u}}_h|||_{1,h}^2 + \omega\xi^2\|\mathbf{u}-\hat{\mathbf{u}}_h\|_{L^2(\partial\Omega)}^2\Big) \\
&\le C\Big(\xi^4(\xi+\gamma_1)h^2|\mathbf{u}|_{H^2(\Omega)}^2 + \omega\xi^2h^3|\mathbf{u}|_{H^2(\Omega)}^2\Big) \\
&= C\xi^4h^2\big(\xi+\gamma_1+\omega h\big)|\mathbf{u}|_{H^2(\Omega)}^2 \\
&\le C\xi^4h^2\big(\xi+\gamma_1+\omega h\big)\left(\omega^\alpha+\frac{1}{\omega^2}\right)^2\Big(\|\mathbf{f}\|_{L^2(\Omega)}^2 + \|\mathbf{g}\|_{L^2(\partial\Omega)}\Big).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&|||\boldsymbol{\psi}_h|||_{1,h} + \omega^{\frac{1}{2}}\xi\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)} \\
&\qquad \leq C\xi^2 h\big(\xi+\gamma_1+\omega h\big)^{\frac{1}{2}}\left(\omega^{\alpha}+\frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\right).
\end{aligned}
$$

Recall that $\mathbf{u}-\tilde{\mathbf{u}}_h = \mathbf{u}-\hat{\mathbf{u}}_h - \boldsymbol{\psi}_h$. By the triangle inequality we get

$$
\begin{aligned}
&|||\mathbf{u}-\tilde{\mathbf{u}}_h|||_{1,h} + \omega^{\frac{1}{2}}\xi\|\mathbf{u}-\tilde{\mathbf{u}}_h\|_{L^2(\partial\Omega)} \\
&\qquad \leq |||\mathbf{u}-\hat{\mathbf{u}}_h|||_{1,h} + \omega^{\frac{1}{2}}\xi\|\mathbf{u}-\hat{\mathbf{u}}_h\|_{L^2(\partial\Omega)} + |||\boldsymbol{\psi}_h|||_{1,h} + \omega^{\frac{1}{2}}\xi\|\boldsymbol{\psi}_h\|_{L^2(\partial\Omega)} \\
&\qquad \leq C\xi^2 h\big(\xi+\gamma_1+\omega h\big)^{\frac{1}{2}}\left(\omega^{\alpha}+\frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\right).
\end{aligned}
$$

Therefore, (3.24) holds.

To obtain (3.25), we appeal to the duality argument by considering the following auxiliary PDE problem:

$$
\begin{aligned}
-\mathbf{div}\left(\sigma(\mathbf{w})\right) &= \mathbf{u}-\tilde{\mathbf{u}}_h \qquad && \text{in } \Omega, \\
\sigma(\mathbf{w})\mathbf{n} - \mathbf{i}\omega A\mathbf{w} &= \mathbf{0} && \text{on } \partial\Omega.
\end{aligned}
$$

It can be shown that there exists a unique solution $\mathbf{w}\in\mathbf{H}^2(\Omega)$ such that

$$
\|\mathbf{w}\|_{H^2(\Omega)} \leq C\|\mathbf{u}-\tilde{\mathbf{u}}_h\|_{L^2(\Omega)}. \tag{3.29}
$$

Let $\hat{\mathbf{w}}_h \in \mathbf{V}_h$ be the P_1 conforming finite element interpolant of $\mathbf{w}$. Testing the above auxiliary problem with $\mathbf{u} - \tilde{\mathbf{u}}_h$ yields

$$\begin{aligned}
\|\mathbf{u} - \tilde{\mathbf{u}}_h\|^2_{L^2(\Omega)} &= -\big(\mathbf{u} - \tilde{\mathbf{u}}_h, \mathbf{div}\,\sigma(\mathbf{w})\big)_\Omega \\
&= a_h(\mathbf{u} - \tilde{\mathbf{u}}_h, \mathbf{w}) + \mathbf{i}\omega\big\langle A(\mathbf{u} - \tilde{\mathbf{u}}_h), \mathbf{w}\big\rangle_{\partial\Omega} \\
&= a_h(\mathbf{u} - \tilde{\mathbf{u}}_h, \mathbf{w} - \hat{\mathbf{w}}_h) + \mathbf{i}\omega\big\langle A(\mathbf{u} - \tilde{\mathbf{u}}_h), \mathbf{w} - \hat{\mathbf{w}}_h\big\rangle_{\partial\Omega} \\
&\leq C\Big(|||\mathbf{u} - \tilde{\mathbf{u}}_h|||_{1,h}|||\mathbf{w} - \hat{\mathbf{w}}_h|||_{1,h} + \omega\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\partial\Omega)}\|\mathbf{w} - \hat{\mathbf{w}}_h\|_{L^2(\Omega)}\Big) \\
&\leq C\Big((\xi + \gamma_1)^{\frac{1}{2}} h|\mathbf{w}|_{H^2(\Omega)}|||\mathbf{u} - \tilde{\mathbf{u}}_h|||_{1,h} + \omega h^{\frac{3}{2}}|\mathbf{w}|_{H^2(\Omega)}\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\partial\Omega)}\Big) \\
&\leq C\Big((\xi + \gamma_1)^{\frac{1}{2}} h\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\Omega)}|||\mathbf{u} - \tilde{\mathbf{u}}_h|||_{1,h} \\
&\qquad + \omega h^{\frac{3}{2}}\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\Omega)}\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\partial\Omega)}\Big).
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\Omega)} &\leq Ch(\xi + \gamma_1 + \omega h)^{\frac{1}{2}}\Big(|||\mathbf{u} - \tilde{\mathbf{u}}_h|||_{1,h} + \omega^{\frac{1}{2}}\xi\|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{L^2(\Omega)}\Big) \\
&\leq C\xi^2 h^2(\xi + \gamma_1 + \omega h)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right).
\end{aligned}$$

Thus, (3.25) holds. □

With estimates for the elliptic projection in hand, all of the components are in place to complete Schatz argument and obtain asymptotic error estimates. The next subsection is devoted to carrying this out.

3.2.2 Asymptotic Error Estimates Via Schatz Argument

In this subsection, error estimates for the proposed IP-DG method are proved in the asymptotic mesh regime (i.e. when $\omega^2 h = O(1)$). This will be carried out using the well-known Schatz argument (c.f. [6, 29, 30, 26, 16]).

To carry out Schatz argument, we introduce the error decomposition

$$\mathbf{e}_h := \mathbf{u} - \mathbf{u}_h = \boldsymbol{\psi} + \boldsymbol{\phi}_h,$$

where $\boldsymbol{\psi} := \mathbf{u} - \tilde{\mathbf{u}}_h$, $\boldsymbol{\phi}_h := \tilde{\mathbf{u}}_h - \mathbf{u}_h$, and $\tilde{\mathbf{u}}_h$ is the elliptic projection of $\mathbf{u}$ as defined in (3.19). Recall that estimates on $\boldsymbol{\psi}$ have already been established (c.f. Theorem 3.2.4). Thus it is left to bound $\boldsymbol{\phi}_h$.

The next step relates the semi-norm $\|\boldsymbol{\phi}_h\|_{1,h}$ to norms on $\boldsymbol{\psi}$ and a lower order norm $\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}$. This step relies on the Gärding's inequality for $A_h(\cdot,\cdot)$ and the continuity of $A_h(\cdot,\cdot)$. After this is complete, the next step is to relate the norm $\|\mathbf{e}_h\|_{L^2(\Omega)}$ to the semi-norm $h|||\mathbf{e}_h|||_{1,h}$ using a duality argument. The final step is to put all of the previous steps together with h chosen to be small enough to obtain the desired error estimates.

The following lemmas carry out the intermediate steps leading to error estimates in the asymptotic mesh regime.

Lemma 3.2.5. *Suppose that* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *solves* (1.4)–(1.5), $\mathbf{u}_h$ *is its IP-DG approximation, and* $\tilde{\mathbf{u}}_h$ *is its elliptic projection. Then for* $\boldsymbol{\phi}_h = \tilde{\mathbf{u}}_h - \mathbf{u}_h$

$$\|\boldsymbol{\phi}_h\|_{1,h}^2 \leq C\left(\xi^3|||\boldsymbol{\psi}|||_{1,h}^2 + \xi^3\omega^2\rho\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2 + \omega^2\rho\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}^2\right), \tag{3.30}$$

where $\boldsymbol{\psi} = \mathbf{u} - \tilde{\mathbf{u}}_h$, $\xi = 1 + \frac{1}{\gamma_0}$ *and* C *is a positive constant independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. By Gärding's inequality (c.f. Theorem 3.1.6), the continuity of $A_h(\cdot,\cdot)$, Galerkin orthogonality, the Cauchy-Schwarz and Young's inequalities we have

$$\begin{aligned}
\frac{1}{2}\|\boldsymbol{\phi}_h\|_{1,h}^2 - \omega^2\rho\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}^2 &\leq C\xi|A_h(\boldsymbol{\phi}_h,\boldsymbol{\phi}_h)| \\
&= C\xi|A_h(\boldsymbol{\psi},\boldsymbol{\phi}_h)| \\
&\leq C\xi\left(|||\boldsymbol{\psi}|||_{1,h}^2 + \omega^2\rho\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}}\left(\xi\|\boldsymbol{\phi}_h\|_{1,h}^2 + \omega^2\rho\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}} \\
&\leq C\left(\xi^3|||\boldsymbol{\psi}|||_{1,h}^2 + \xi^3\omega^2\rho\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2 + \frac{\omega^2\rho}{\xi}\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}^2\right) + \frac{1}{4}\|\boldsymbol{\phi}_h\|_{1,h}^2.
\end{aligned}$$

Thus, rearranging the terms and using the fact that $1 + \frac{1}{\xi} \leq 2$ yields the desired inequality (3.30). □

Lemma 3.2.6. *Suppose that* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *solves* (1.4)–(1.5) *and* $\mathbf{u}_h$ *is its IP-DG approximation. There exists a positive constant* C_1 *independent of* $\omega, h, \gamma_0, \gamma_1$ *such that for*

$$h \leq \min\left\{\frac{1}{\omega}, \left(2C_1\xi^2\omega(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\right)^{-1}\right\},$$

there holds

$$\|\mathbf{u}-\mathbf{u}_h\|^2_{L^2(\Omega)} \leq C\xi^2 h\,(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)|||\mathbf{u}-\mathbf{u}_h|||_{1,h}, \tag{3.31}$$

where $\xi = 1 + \frac{1}{\gamma_0}$ *and* C *is a positive constant independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. Let $\mathbf{w}$ be the solution to the following problem:

$$A_h(\mathbf{v}, \mathbf{w}) = (\mathbf{v}, \mathbf{e}_h)_\Omega \qquad \forall \mathbf{v} \in \mathbf{E}.$$

Let $\tilde{\mathbf{w}}_h$ be the elliptic projection of $\mathbf{w}$ defined by (3.19). Using the continuity of $A_h(\cdot,\cdot)$ and Galerkin orthogonality we get

$$\begin{aligned}
\|\mathbf{e}_h\|^2_{L^2(\Omega)} &= (\mathbf{e}_h, \mathbf{e}_h)_{L^2(\Omega)} \\
&= A_h(\mathbf{e}_h, \mathbf{w}) \\
&= A_h(\mathbf{e}_h, \mathbf{w} - \tilde{\mathbf{w}}_h) \\
&\leq C\left(|||\mathbf{e}_h|||_{1,h} + \omega^2\|\mathbf{e}_h\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}\left(|||\mathbf{w}-\tilde{\mathbf{w}}_h|||_{1,h} + \omega^2\|\mathbf{w}-\tilde{\mathbf{w}}_h\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}.
\end{aligned}$$

Applying Theorem 3.2.4 and using the fact that $h \leq \frac{1}{\omega}$ yield

$$\begin{aligned}\|\mathbf{e}_h\|^2_{L^2(\Omega)} &\leq C\left(|||\mathbf{e}_h|||_{1,h} + \omega^2\|\mathbf{e}_h\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \\ &\quad \cdot \xi^2(\xi+\gamma_1+\omega h)h(1+\omega h)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\|\mathbf{e}_h\|_{L^2(\Omega)} \\ &\leq C_1\left(|||\mathbf{e}_h|||_{1,h} + \omega^2\|\mathbf{e}_h\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \\ &\quad \cdot \xi^2 h(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\|\mathbf{e}_h\|_{L^2(\Omega)}.\end{aligned}$$

Dividing both sides of this inequality by $\|\mathbf{e}_h\|_{L^2(\Omega)}$ and using the fact that $h \leq \left(2C_1\xi^2\omega(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\right)^{-1}$ gives

$$\begin{aligned}\|\mathbf{e}_h\|_{L^2(\Omega)} &\leq C_1\xi^2 h(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)|||\mathbf{e}_h|||_{1,h} \\ &\quad + C_1\xi^2\omega h(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\|\mathbf{e}_h\|_{L^2(\Omega)} \\ &\leq C_1\xi^2 h(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)|||\mathbf{e}_h|||_{1,h} + \frac{1}{2}\|\mathbf{e}_h\|_{L^2(\Omega)}.\end{aligned}$$

Subtracting $\frac{1}{2}\|\mathbf{e}_h\|_{L^2(\Omega)}$ from both sides of the above inequality yields the desired result. □

Lemma 3.2.5 and lemma 3.2.6 are now put together to derive optimal order error estimates for h small.

Theorem 3.2.7. *Let $\mathbf{e}_h = \mathbf{u} - \mathbf{u}_h$ where $\mathbf{u} \in \mathbf{H}^2(\Omega)$ solves (1.4)–(1.5) and $\mathbf{u}_h$ is its IP-DG approximation. There exists a positive constant $\tilde{C}$, independent of $\omega, h, \gamma_0, \gamma_1$, such that for $h \leq h_0 := \min\left\{\frac{1}{\omega}, \left(2\tilde{C}\xi^{\frac{5}{2}}\omega(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\right)^{-1}\right\}$ the following error*

estimates hold:

$$|||\mathbf{e}_h|||_{1,h} \leq C\xi^4(\xi+\gamma_1)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\left(h+\omega h^2\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right), \tag{3.32}$$

$$\|\mathbf{e}_h\|_{L^2(\Omega)} \leq C\xi^6(\xi+\gamma_1)^2\left(\omega^\alpha + \frac{1}{\omega^2}\right)^2\left(h^2+\omega h^3\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right), \tag{3.33}$$

where $\xi = 1 + \frac{1}{\gamma_0}$, α *is defined as in Theorem 3.2.1, and* C *is a positive constant independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. Let $\tilde{\mathbf{u}}_h$ be the elliptic projection of $\mathbf{u}$. Then $\mathbf{e}_h = \boldsymbol{\psi} + \boldsymbol{\phi}_h$, where $\boldsymbol{\psi} = \mathbf{u} - \tilde{\mathbf{u}}_h$ and $\boldsymbol{\phi}_h = \tilde{\mathbf{u}}_h - \mathbf{u}_h$. By Lemma 3.2.5 we get

$$\begin{aligned}
|||\mathbf{e}_h|||_{1,h}^2 &\leq C\left(|||\boldsymbol{\psi}|||_{1,h}^2 + |||\boldsymbol{\phi}_h|||_{1,h}^2\right) \\
&\leq C\left(|||\boldsymbol{\psi}|||_{1,h}^2 + \xi\|\boldsymbol{\phi}_h\|_{1,h}^2\right) \\
&\leq C\left(\xi^4|||\boldsymbol{\psi}|||_{1,h}^2 + \xi^4\omega^2\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2 + \xi\omega^2\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}^2\right).
\end{aligned} \tag{3.34}$$

Note that $\boldsymbol{\phi}_h = -\boldsymbol{\psi} + \mathbf{e}_h$, and the triangle inequality implies that

$$\|\boldsymbol{\phi}_h\|_{L^2(\Omega)}^2 \leq C\left(\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2 + \|\mathbf{e}_h\|_{L^2(\Omega)}^2\right).$$

Substituting the above inequality into (3.34) and using Lemma 3.2.6 (at this point we assume that $\tilde{C} \geq C_1$ and thus h is small enough to satisfy the condition of Lemma 3.2.6), we have

$$\begin{aligned}
|||\mathbf{e}_h|||_{1,h}^2 &\leq C\left(\xi^4|||\boldsymbol{\psi}|||_{1,h}^2 + \xi^4\omega^2\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2 + \xi\omega^2\|\mathbf{e}_h\|_{L^2(\Omega)}^2\right) \\
&\leq C\left(\xi^4|||\boldsymbol{\psi}|||_{1,h}^2 + \xi^4\omega^2\|\boldsymbol{\psi}\|_{L^2(\Omega)}^2\right) \\
&\quad + \tilde{C}\xi^5\omega^2h^2(\xi+\gamma_1)^2\left(\omega^\alpha + \frac{1}{\omega^2}\right)^2|||\mathbf{e}_h|||_{1,h}^2.
\end{aligned}$$

Now the fact that $h \leq h_0$ implies

$$|||\mathbf{e}_h|||_{1,h}^2 \leq C\left(\xi^4|||\boldsymbol{\psi}|||_{1,h}^2 + \xi^4\omega^2\|\boldsymbol{\psi}\|_{1,h}^2\right) + \frac{1}{4}|||\mathbf{e}_h|||_{1,h}^2.$$

Thus,

$$|||\mathbf{e}_h|||_{1,h} \leq C\left(\xi^2|||\boldsymbol{\psi}|||_{1,h} + \xi^2\omega\|\boldsymbol{\psi}\|_{L^2(\Omega)}\right),$$

which together with Theorem 3.2.4 infers (3.32).

(3.33) follows from applying Lemma 3.2.6 to (3.32). □

We note that a mesh constraint of the form $\omega^{\alpha+1}h = O(1)$ must be used to ensure an optimal order error estimate when using Schatz argument. If a Korn's inequality holds on the boundary (c.f. Conjecture 2.3.4) then $\alpha = 1$ and the constraint becomes $\omega^2 h = O(1)$. This is consistent with the mesh constraint used to characterize the asymptotic mesh regime for discretization methods applied to the other Helmholtz-type problems. Stability in the asymptotic mesh regime can be derived as a consequence of the error estimates above. For standard discretization methods applied to the elastic Helmholtz problem, stability has only been proved in the asymptotic mesh regime. In the next section, we will carry out a stability and convergence analysis for our IP-DG method in the pre-asymptotic mesh regime, i.e. for $\omega^{\alpha+1}h > C$.

3.3 Pre-asymptotic Error Estimates

In the previous section, optimal error estimates were obtained in the asymptotic mesh regime, i.e. when $\omega^{\alpha+1}h \leq C$, where α is defined in Theorem 3.2.1 and C is some positive constant independent of ω. In this section, stability estimates for our IP-DG approximation will be established in the pre-asymptotic mesh regime, i.e. when $\omega^{\alpha+1}h > C$. These stability estimates will then be used to establish optimal order

error estimates in the pre-asymptotic regime. Thus, Section 3.2 together with this section ensure that the proposed IP-DG method is absolutely stable. This property makes the proposed IP-DG method especially well suited to approximate the elastic Helmholtz equations.

3.3.1 Stability Estimates in the Pre-Asymptotic Mesh Regime

In this subsection, stability estimates for the proposed IP-DG method are established. Specifically, these estimates are established in the pre-asymptotic mesh regime, i.e. when $\omega^{\alpha+1}h > C$. It turns out that stability in this case is just a consequence of Theorem 3.1.3 which proves a weak coercivity property of $A_h(\cdot,\cdot)$. As stated previously, the coercivity constant from this theorem is adversely dependent on h for small values of h. In the case that h is bounded away from zero, this constant can be replaced with one that is not dependent on h. Thus, stability estimates for the pre-asymptotic mesh regime are obtained as a consequence of the weak coercivity of $A_h(\cdot,\cdot)$. The stability of the IP-DG method in the pre-asymptotic mesh regime is given by the next theorem.

Theorem 3.3.1. *Suppose that $\mathbf{u}_h \in \mathbf{V}_h$ solves the IP-DG method given by (3.6). Then the following inequalities hold:*

$$|\mathbf{u}_h|_{1,h}^2 + \omega^2\rho\|\mathbf{u}_h\|_{L^2(\Omega)}^2 \leq C\left(C_{\text{sta}}^2\|\mathbf{f}\|_{L^2(\Omega)}^2 + C_{\text{sta}}\|\mathbf{g}\|_{L^2(\partial\Omega)}^2\right), \quad (3.35)$$

$$J_0(\mathbf{u}_h,\mathbf{u}_h) + J_1(\mathbf{u}_h,\mathbf{u}_h) + \left\langle A\mathbf{u}_h,\mathbf{u}_h\right\rangle_{\partial\Omega} \leq \frac{C}{\omega}\left(C_{\text{sta}}\|\mathbf{f}\|_{L^2(\Omega)}^2 + \|\mathbf{g}\|_{L^2(\partial\Omega)}^2\right), \quad (3.36)$$

where $C_{\text{sta}} = \left(\frac{\xi}{\omega} + \frac{1}{\omega^2 h} + \frac{1}{\omega^3 h^2\gamma_1}\right)$, $\xi = 1 + \frac{1}{\gamma_0}$, and C is a positive constant independent of $\omega, h, \gamma_0, \gamma_1$.

Proof. By (3.12) and (3.13) we get

$$
\begin{aligned}
&|\mathbf{u}_h|_{1,h}^2+\omega^2\rho\|\mathbf{u}_h\|_{L^2(\Omega)}^2+\omega\Big(1+\frac{1}{\gamma_0}+\frac{1}{\omega h}+\frac{1}{\omega^2\rho h^2\gamma_1}\Big)\big\langle A\mathbf{u}_h,\mathbf{u}_h\big\rangle_{\partial\Omega}\\
&\quad\le C\Big(1+\frac{1}{\gamma_0}+\frac{1}{\omega h}+\frac{1}{\omega^2\rho h^2\gamma_1}\Big)|A_h(\mathbf{u}_h,\mathbf{u}_h)|\\
&\quad\le C\Big(1+\frac{1}{\gamma_0}+\frac{1}{\omega h}+\frac{1}{\omega^2\rho h^2\gamma_1}\Big)\Big(\|\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{u}_h\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\|\mathbf{u}_h\|_{L^2(\partial\Omega)}\Big)\\
&\quad\le \frac{C}{\omega^2\rho}\Big(1+\frac{1}{\gamma_0}+\frac{1}{\omega h}+\frac{1}{\omega^2\rho h^2\gamma_1}\Big)^2\|\mathbf{f}\|_{L^2(\Omega)}^2+\frac{\omega^2\rho}{2}\|\mathbf{u}_h\|_{L^2(\Omega)}^2\\
&\qquad\quad+\frac{C}{\omega c_A}\Big(1+\frac{1}{\gamma_0}+\frac{1}{\omega h}+\frac{1}{\omega^2\rho h^2\gamma_1}\Big)\|\mathbf{g}\|_{L^2(\partial\Omega)}^2\\
&\qquad\quad+\omega c_A\Big(1+\frac{1}{\gamma_0}+\frac{1}{\omega h}+\frac{1}{\omega^2\rho h^2\gamma_1}\Big)\|\mathbf{u}_h\|_{L^2(\partial\Omega)}^2.
\end{aligned}
$$

Substituting (3.10) into the above inequality infers (3.35).

Now, combining (3.13) with (3.35) yields

$$
\begin{aligned}
&J_0(\mathbf{u}_h,\mathbf{u}_h)+J_1(\mathbf{u}_h,\mathbf{u}_h)+\omega\big\langle A\mathbf{u}_h,\mathbf{u}_h\big\rangle_{\partial\Omega}\\
&\quad\le |A_h(\mathbf{u}_h,\mathbf{u}_h)|\\
&\quad\le \|\mathbf{f}\|_{L^2(\Omega)}\|\mathbf{u}_h\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\|\mathbf{u}_h\|_{L^2(\partial\Omega)}\\
&\quad\le \frac{C}{\omega\rho^{\frac12}}\|\mathbf{f}\|_{L^2(\Omega)}\Big(\frac{1}{\rho}C_{\mathrm{sta}}^2\|\mathbf{f}\|_{L^2(\Omega)}^2+\frac{1}{c_A}C_{\mathrm{sta}}\|\mathbf{g}\|_{L^2(\partial\Omega)}^2\Big)^{\frac12}\\
&\qquad\quad+\frac{C}{\omega c_A}\|\mathbf{g}\|_{L^2(\partial\Omega)}^2+\frac{\omega c_A}{2}\|\mathbf{u}_h\|_{L^2(\partial\Omega)}^2\\
&\quad\le \frac{C}{\omega\rho}C_{\mathrm{sta}}\|\mathbf{f}\|_{L^2(\Omega)}^2+\frac{C}{\omega c_A}\|\mathbf{g}\|_{L^2(\partial\Omega)}^2+\frac{\omega c_A}{2}\|\mathbf{u}_h\|_{L^2(\partial\Omega)}^2.
\end{aligned}
$$

Using (3.10) in the above inequality infers (3.36). □

Remark 3.3.2. *When* $\omega^{\alpha+1}h>C$

$$
C_{\mathrm{sta}}<C\Big(\frac{\xi}{\omega}+\omega^{\alpha-1}+\frac{\omega^{2\alpha-1}}{\gamma_1}\Big). \tag{3.37}
$$

Thus, the constant in the above stability estimate is independent of h *in the pre-asymptotic mesh regime.*

3.3.2 Error Estimates for the IP-DG Method

In this subsection, the stability estimates established in Subsection 3.3.1 are utilized to obtain optimal order error estimates for the proposed IP-DG method in the pre-asymptotic mesh regime (i.e. $\omega^{\alpha+1}h > C$). These estimates are obtained with the help of the elliptic projection defined in Subsection 3.2.1, the stability estimates established in Subsection 3.3.1, and Galerkin orthogonality for the sesquilinear form $A_h(\cdot,\cdot)$.

Let $\mathbf{u} \in \mathbf{H}^2(\Omega)$ solve (1.4)–(1.5) and $\mathbf{u}_h \in \mathbf{V}_h$ be its IP-DG approximation defined in (3.6). As before, the error $\mathbf{e}_h$ is defined by $\mathbf{e}_h := \mathbf{u} - \mathbf{u}_h$. Subtracting (3.6) from (3.3) immediately yields the following Galerkin orthogonality property for $A_h(\cdot,\cdot)$:

$$a_h(\mathbf{e}_h, \mathbf{v}_h) - \omega^2 \rho\big(\mathbf{e}_h, \mathbf{v}_h\big)_\Omega + \mathbf{i}\omega\big\langle A\mathbf{e}_h, \mathbf{v}_h\big\rangle_{\partial\Omega} = \mathbf{0} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h. \tag{3.38}$$

Let $\tilde{\mathbf{u}}_h \in \mathbf{V}_h$ denote the elliptic projection of $\mathbf{u}$ as defined in (3.19). In the same fashion as was done in subsection 3.2.2, the error $\mathbf{e}_h$ can be decomposed as $\mathbf{e}_h = \boldsymbol{\psi} + \boldsymbol{\phi}_h$ where $\boldsymbol{\psi} = \mathbf{u} - \tilde{\mathbf{u}}_h$ and $\boldsymbol{\phi}_h = \tilde{\mathbf{u}}_h - \mathbf{u}_h$. Again, estimates on $\mathbf{e}_h$ will be established from estimates on $\boldsymbol{\psi}$ and $\boldsymbol{\phi}_h$ that are obtained separately. By Galerkin orthogonality given in (3.17) and (3.38) we have the following identity:

$$\begin{aligned} a_h(\boldsymbol{\phi}_h, \mathbf{v}_h) - \omega^2 \rho\big(\boldsymbol{\phi}_h, \mathbf{v}_h\big)_\Omega + \mathbf{i}\omega\big\langle A\boldsymbol{\phi}_h, \mathbf{v}_h\big\rangle_{\partial\Omega} \\ = -a_h(\boldsymbol{\psi}, \mathbf{v}_h) + \omega^2 \rho\big(\boldsymbol{\psi}, \mathbf{v}_h\big)_\Omega - \mathbf{i}\omega\big\langle A\boldsymbol{\psi}, \mathbf{v}_h\big\rangle_{\partial\Omega} \\ = \omega^2 \rho\big(\boldsymbol{\psi}, \mathbf{v}_h\big)_\Omega \qquad \forall \mathbf{v}_h \in \mathbf{V}_h. \end{aligned} \tag{3.39}$$

In other words, $\boldsymbol{\phi}_h \in \mathbf{V}_h$ solves (3.6) with $\mathbf{f} = \omega^2\rho\boldsymbol{\psi}$ and $\mathbf{g} \equiv \mathbf{0}$. This allows us to establish estimates on $\boldsymbol{\phi}_h$ by using the stability estimates from Theorem 3.3.1. Specifically, we have the next lemma.

Lemma 3.3.3. *Let* $\mathbf{u} \in H^2(\Omega)$ *solve* (1.4)–(1.5), $\mathbf{u}_h$ *be its IP-DG approximation, and* $\tilde{\mathbf{u}}_h$ *be its elliptic projection. Then* $\boldsymbol{\phi}_h = \tilde{\mathbf{u}}_h - \mathbf{u}_h$ *satisfies the following inequality:*

$$\begin{aligned}
&\|\boldsymbol{\phi}_h\|_{1,h} + \omega\rho^{\frac{1}{2}}\|\boldsymbol{\phi}_h\|_{L^2(\Omega)} \\
&\qquad \leq C\xi^2\omega^2h^2C_{\mathrm{sta}}\big(\xi+\gamma_1+\omega h\big)\Big(\omega^\alpha+\frac{1}{\omega^2}\Big)\Big(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\Big), \qquad (3.40)
\end{aligned}$$

where $C_{\mathrm{sta}} = \Big(\frac{\xi}{\omega}+\frac{1}{\omega^2h}+\frac{1}{\omega^3h^2\gamma_1}\Big)$, $\xi = 1+\frac{1}{\gamma_0}$, *and* C *is a positive constant independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. (3.39) implies that $\boldsymbol{\phi}_h$ solves (3.6) with $\mathbf{f} = \omega^2\rho\boldsymbol{\psi}$ and $\mathbf{g} \equiv \mathbf{0}$. Thus, an application of Theorem 3.3.1 yields

$$\begin{aligned}
\|\boldsymbol{\phi}_h\|_{1,h} + \omega\rho^{\frac{1}{2}}\|\boldsymbol{\phi}_h\|_{L^2(\Omega)} &\leq \frac{C}{\rho^{\frac{1}{2}}}C_{\mathrm{sta}}\|\omega^2\rho\boldsymbol{\psi}\|_{L^2(\Omega)} \\
&\leq C\omega^2\rho^{\frac{1}{2}}C_{\mathrm{sta}}\|\boldsymbol{\psi}\|_{L^2(\Omega)},
\end{aligned}$$

which along with Theorem 3.2.4 infers (3.40). □

We are now ready to derive error estimates for our IP-DG method in the pre-asymptotic mesh regime. The next theorem is a consequence of combining Theorem 3.2.4 and Lemma 3.3.3.

Theorem 3.3.4. *Let* $\mathbf{u} \in \mathbf{H}^2(\Omega)$ *solve* (1.4)–(1.5) *and* $\mathbf{u}_h$ *be its IP-DG approximation. Then* $\mathbf{e}_h = \mathbf{u} - \mathbf{u}_h$ *satisfies the following inequality:*

$$\begin{aligned}
&\|\mathbf{e}_h\|_{1,h} + \omega\rho^{\frac{1}{2}}\|\mathbf{e}_h\|_{L^2(\Omega)} \qquad (3.41)\\
&\quad \leq C\xi^2h\big(\xi+\gamma_1+\omega h\big)\Big(\omega^\alpha+\frac{1}{\omega^2}\Big)\Big(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\Big) \\
&\qquad + C\xi^2\omega h^2\big(1+\omega C_{\mathrm{sta}}\big)\big(\xi+\gamma_1+\omega h\big)\Big(\omega^\alpha+\frac{1}{\omega^2}\Big)\Big(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\Big), \\
&\|\mathbf{e}_h\|_{L^2(\Omega)} \qquad (3.42)\\
&\quad \leq C\xi^2h^2\big(1+\omega C_{\mathrm{sta}}\big)\big(\xi+\gamma_1+\omega h\big)\Big(\omega^\alpha+\frac{1}{\omega^2}\Big)\Big(\|\mathbf{f}\|_{L^2(\Omega)}+\|\mathbf{g}\|_{L^2(\partial\Omega)}\Big),
\end{aligned}$$

where $C_{\text{sta}} = \left(\frac{\xi}{\omega} + \frac{1}{\omega^2 h} + \frac{1}{\omega^3 h^2 \gamma_1}\right)$, $\xi = 1 + \frac{1}{\gamma_0}$, *and* C *is a positive constant independent of* $\omega, h, \gamma_0, \gamma_1$.

Proof. Recall that estimates for $\boldsymbol{\psi}$ and $\boldsymbol{\phi}_h$ have already been established in Theorem 3.2.4 and Lemma 3.3.3, respectively. These estimates are combined in the following steps to obtain (3.41):

$$
\begin{aligned}
&\|\mathbf{e}_h\|_{1,h} + \omega\rho^{\frac{1}{2}}\|\mathbf{e}_h\|_{L^2(\Omega)} \\
&\qquad \le |||\boldsymbol{\psi}|||_{1,h} + \omega\rho^{\frac{1}{2}}\|\boldsymbol{\psi}\|_{L^2(\Omega)} + \|\boldsymbol{\phi}_h\|_{1,h} + \omega\rho^{\frac{1}{2}}\|\boldsymbol{\phi}_h\|_{L^2(\Omega)} \\
&\qquad \le C\xi^2 h\big(\xi + \gamma_1 + \omega h\big)^{\frac{1}{2}}\left(\omega^\alpha + \frac{1}{\omega^2}\right)^{\frac{1}{2}}\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) \\
&\qquad\quad + C\xi^2\omega h^2\big(\xi + \gamma_1 + \omega h\big)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) \\
&\qquad\quad + C\xi^2\omega^2 h^2 C_{\text{sta}}\big(\xi + \gamma_1 + \omega h\big)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) \\
&\qquad \le C\xi^2 h\big(\xi + \gamma_1 + \omega h\big)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right) \\
&\qquad\quad + C\xi^2\omega h^2\big(1 + \omega C_{\text{sta}}\big)\big(\xi + \gamma_1 + \omega h\big)\left(\omega^\alpha + \frac{1}{\omega^2}\right)\left(\|\mathbf{f}\|_{L^2(\Omega)} + \|\mathbf{g}\|_{L^2(\partial\Omega)}\right).
\end{aligned}
$$

Similarly, (3.42) is obtained by combining Theorem 3.2.4 and Lemma 3.3.3. □

Remark 3.3.5. *When* $\omega^{\alpha+1}h > C$

$$
C_{\text{sta}} < C\left(\frac{\xi}{\omega} + \omega^{\alpha-1} + \frac{\omega^{2\alpha-1}}{\gamma_1}\right).
$$

Therefore, the above error estimates are optimal in h *in the pre-asymptotic mesh regime.*

3.4 Numerical Experiments

In this section, numerical tests are carried out in order to demonstrate key features of the proposed IP-DG method. We choose $\Omega = (-0.5, 0.5)^2 \subset \mathbb{R}^2$ (i.e. the unit square in $\mathbb{R}^2$ centered at the origin), along with the material constants $\rho = \mu = \lambda = 1$,

and penalty constants $\gamma_0 = 10$ and $\gamma_1 = 0.1$. For the sake of testing the exact error, $\mathbf{f}$ and $\mathbf{g}$ are chosen so that the exact solution to the elastic Helmholtz equations is $\mathbf{u} = \frac{1}{\omega^2 r}[e^{\mathbf{i}\omega r} - 1, e^{-\mathbf{i}\omega r} - 1]^T$, where $r = \|\mathbf{x}\|_2$. This simple problem along with the subsequent numerical tests are chosen to mirror those for the IP-DG method proposed in [42] for the scalar Helmholtz problem. Some example plots are given in Figures 3.2 and 3.3. These plots demonstrate how well the proposed IP-DG method can capture an example with large wave frequency when using a relatively coarse mesh.

To partition the domain Ω, a uniform triangulation $\mathcal{T}_h$ is used. For a positive integer n, define $\mathcal{T}_{1/n}$ to be a triangulation of $2n^2$ congruent isosceles triangles with side lengths $1/n, 1/n$, and $\sqrt{2}/n$. Figure 4.1 shows a sample triangulation $\mathcal{T}_{1/10}$.

The numerical tests in this section intend to demonstrate the following:

- absolute stability of our IP-DG method,
- error of our IP-DG solution,
- pollution effect on the error when $\omega h = O(1)$,
- absence of the pollution effect when $\omega^3 h^2 = O(1)$,
- comparisons between standard FE and our IP-DG method for this problem.

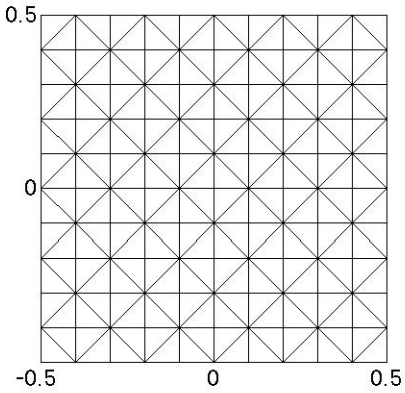


Figure 3.1: Example of the triangulation $\mathcal{T}_{1/10}$.

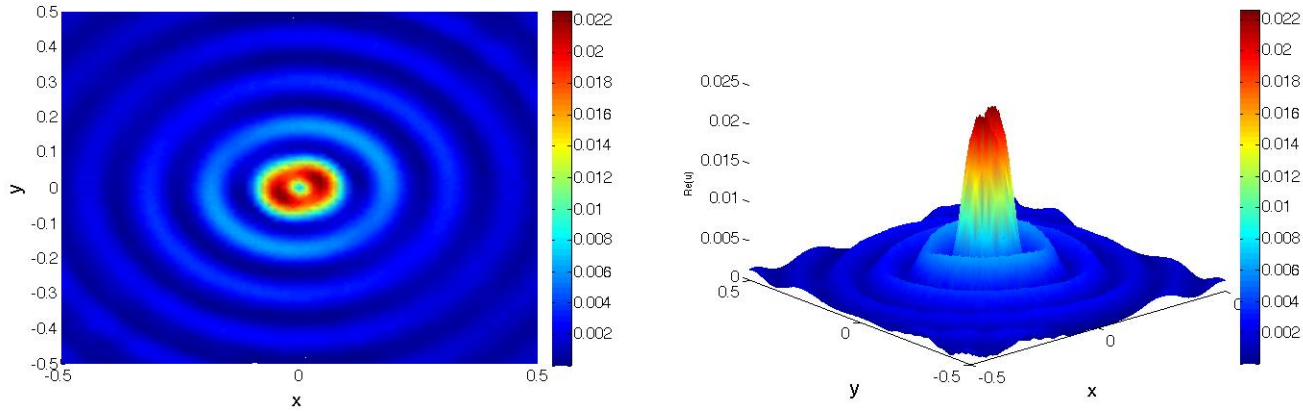


Figure 3.2: Plot of $\| \operatorname{Re}(\mathbf{u}_h) \|_2$ for $\omega = 50$ and $h = 1/70$. Both a top down view (left) and a side view (right) are shown.

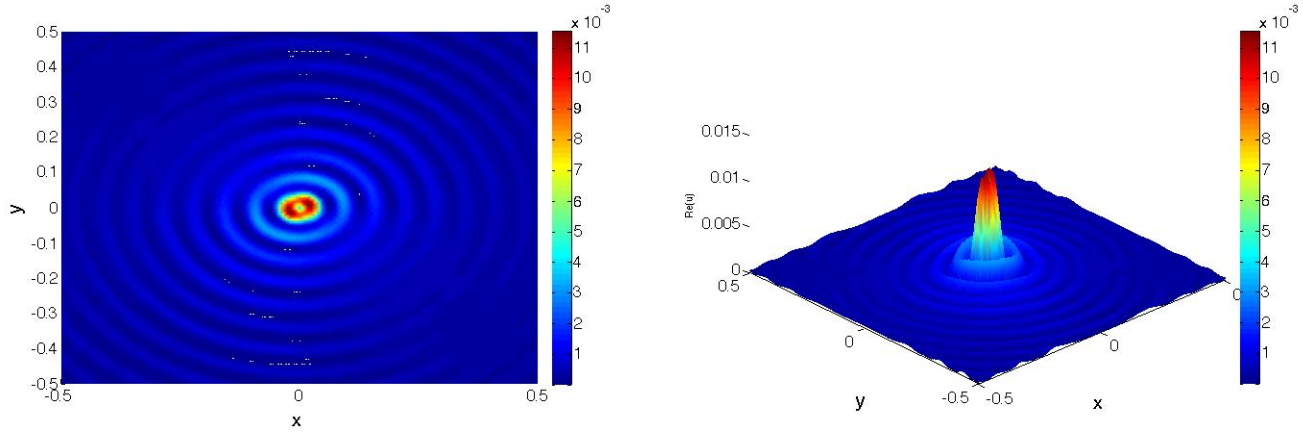


Figure 3.3: Plot of $\| \operatorname{Re}(\mathbf{u}_h) \|_2$ for $\omega = 100$ and $h = 1/120$. Both a top down view (left) and a side view (right) are shown.

3.4.1 Stability

In this subsection, the stability of both the proposed IP-DG method and the P_1-conforming finite element method will be discussed. Let $\mathbf{u}_h^{FEM}$ denote the P_1-conforming finite element approximation of $\mathbf{u}$. Recall that the proposed IP-DG approximation is absolutely stable, i.e. it is stable for all $\omega, h, \gamma_0, \gamma_1 > 0$. This has not been established for the P_1-conforming finite element approximation. In fact, the stability of the P_1-conforming finite element approximation is known to hold only when h satisfies $\omega^2 h \leq C$.

Figure 3.4 plots both $\|\mathbf{u}_h\|_{1,h}$ and $\|\mathbf{u}^{FEM}\|_{1,h}$ for $h = 0.05, 0.01$ and $\omega = 1, 2, ..., 200$. We observe that $\|\mathbf{u}_h\|_{1,h}$ decreases in a smooth fashion as ω increases. This smooth behavior of $\|\mathbf{u}_h\|_{1,h}$ is indicative of the absolute stability of the IP-DG approximation. On the other hand, we observe oscillations in $\|\mathbf{u}_h^{FEM}\|_{1,h}$ that occur when we vary ω. This oscillation is indicative of the instability of the P_1-conforming finite element method when h is too large.

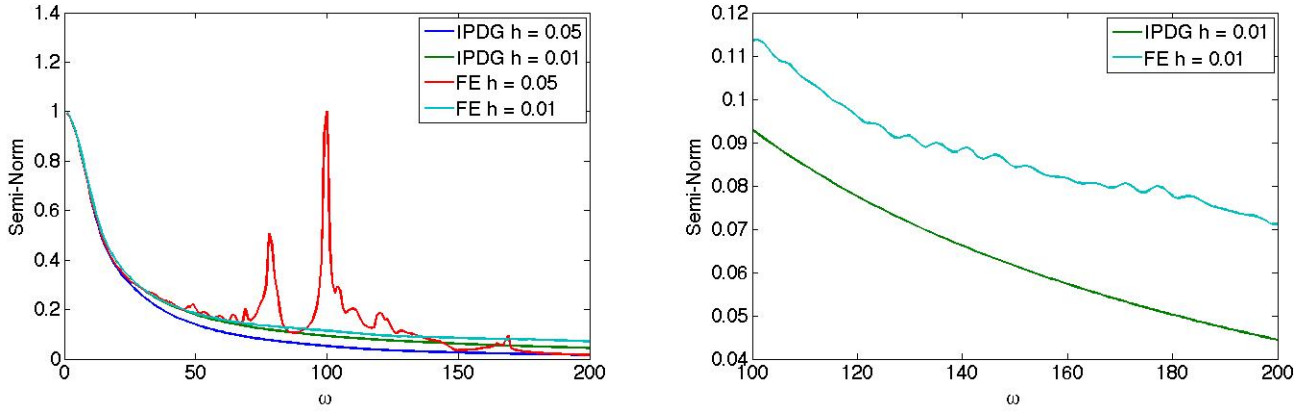


Figure 3.4: Plots of $\|\mathbf{u}_h\|_{1,h}$ and $\|\mathbf{u}_h^{FEM}\|_{1,h}$.

3.4.2 Error

In this subsection, the optimal order of convergence for the proposed IP-DG method will be demonstrated. The pollution effect will also be demonstrated. From Theorems 3.3.4 and 3.2.7 we expect the error in $\|\cdot\|_{1,h}$ to decrease at an optimal order in both the pre-asymptotic and asymptotic mesh regimes. In other words, $\|\mathbf{u}-\mathbf{u}_h\|_{1,h} = O(h)$ is expected. Figure 3.5 is a log-log plot of the relative error $\|\mathbf{u}-\mathbf{u}_h\|_{1,h}/\|\mathbf{u}\|_{1,h}$ against the value $1/h$ for frequencies $\omega = 5, 10, 20, 30$. From this plot, it is observed that the relative error decreases at the same rate as h, thus displaying the optimal order of convergence in the relative semi-norm. Also displayed in Figure 3.5 is the error when ω varies according to the constraint $\omega h = 0.25$. From this figure it is observed that the error increases as ω increases under this constraint. This is due to the pollution effect on the error for the elastic Helmholtz equations.

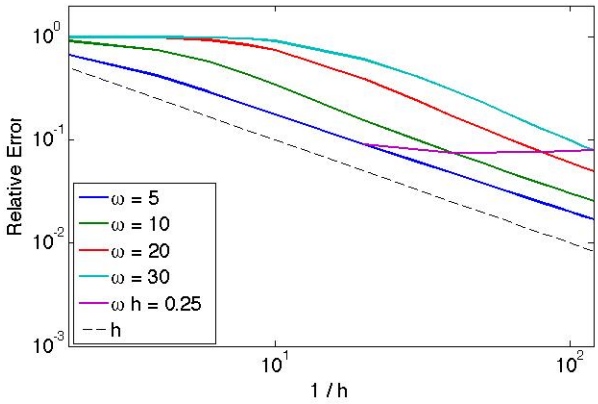


Figure 3.5: Log-log plot of the relative error for the IP-DG approximation measured in the H^1-seminorm for different values of ω.

The pollution effect for Helmholtz-type problems is characterized by the increase in error as ω is increased under the constraint $\omega h = O(1)$. This effect is intrinsic to Helmholtz-type problems (c.f. [54]). It is well-known that the pollution effect can be eliminated if h is chosen to fulfill the more stringent constraint $\omega^3 h^2 = O(1)$. In Figure 3.6 the relative error is plotted against ω as h is chosen to satisfy different constraints. Under the constraints $\omega h = 1$ and $\omega h = 0.5$, the pollution effect is present and the relative error increases as ω is increased. On the other hand, when $\omega^3 h^2 = 1$ is used to choose the the mesh size h, the pollution effect is eliminated.

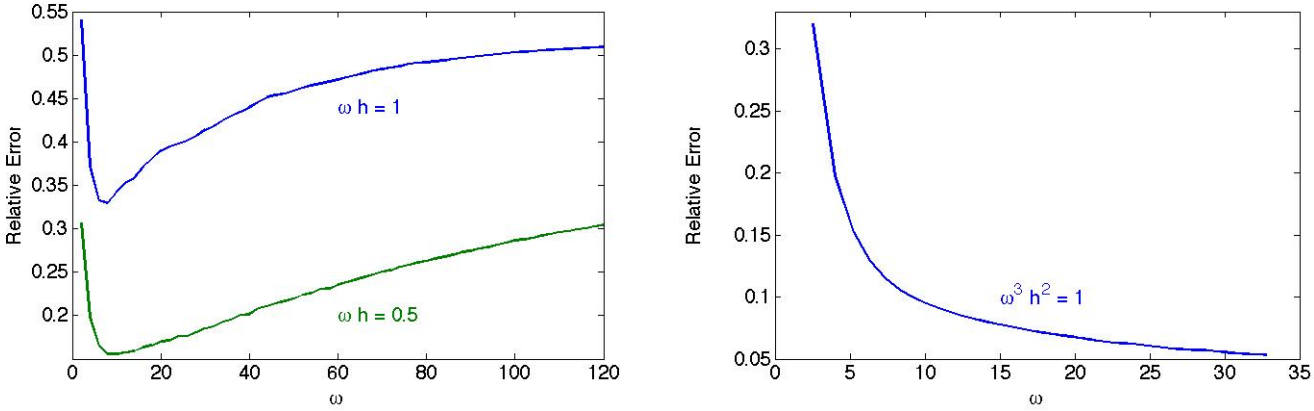


Figure 3.6: Relative error of the IP-DG approximation measured in the H^1 seminorm computed for different values of ω and h is chosen to satisfy the given constraints.

3.4.3 IP-DG vs. FEM

In this subsection, the proposed IP-DG solution is compared to the P_1-conforming finite element solution. As stated previously, the proposed IP-DG method is absolutely stable while the P_1-conforming finite element method is only shown to be stable when h satisfies $\omega^2 h = O(1)$. With this in mind, one can anticipate that in the case that the frequency ω is allowed to be large, the IP-DG method becomes a better method.

In Figures 3.7–3.9, $\|\operatorname{Re}(\mathbf{u}_h)\|_2$ and $\|\operatorname{Re}(\mathbf{u}_h^{FEM})\|_2$ are plotted for $\omega = 100$ and $h = 1/50, 1/120, 1/200$ on a cross-section over the line $y = x$. In addition, $\|\operatorname{Re}(\mathbf{u})\|_2$ is plotted to measure how well the respective approximations capture the true solution. In Figure 3.7, it is observed that $\mathbf{u}_h$ already captures the phase of $\mathbf{u}$ with $h = 1/50$ while not fully capturing the large changes in magnitude. On the other hand, for $h = 1/50$, $\mathbf{u}_h^{FEM}$ has spurious oscillations. In this case, $\mathbf{u}_h^{FEM}$ also fails to capture the changes in the magnitude of the wave. In Figure 3.8, we see that for $h = 1/120$, $\mathbf{u}_h$ captures the phase and changes in magnitude of the wave very well while $\mathbf{u}_h^{FEM}$ still displays spurious oscillations. In Figure 3.9, we see for $h = 1/200$, both methods capture the wave well. However, the IP-DG method captures the wave slightly better. These examples demonstrate that the IP-DG method approximates high frequency waves better than the standard finite element when a coarse mesh is employed. This is of great importance when memory is limited or one wishes to employ a multi-level solver such as multigrid or multi-level Schwarz space/domain decomposition methods.

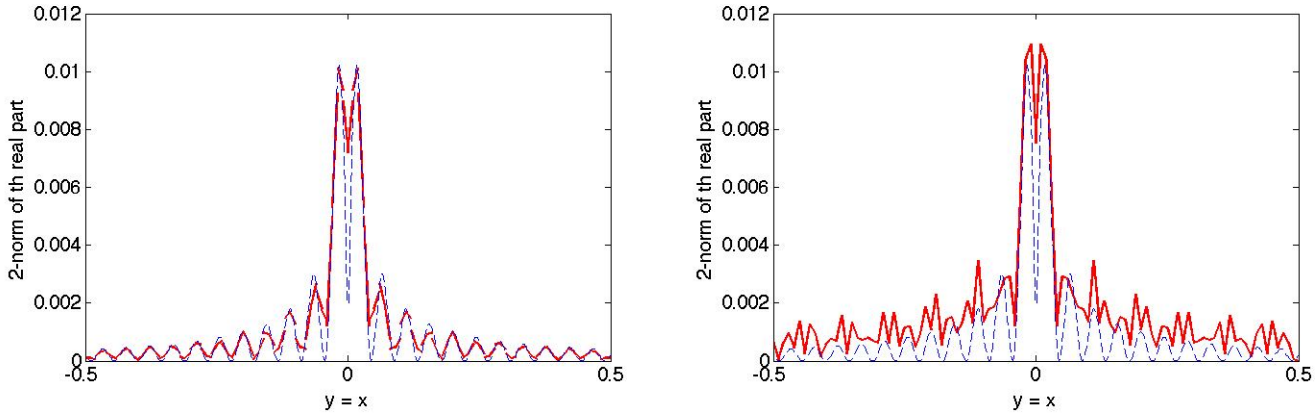


Figure 3.7: The left plot is of $\| \operatorname{Re}(\mathbf{u}_h) \|_2$ (solid red line) vs. $\| \operatorname{Re}(\mathbf{u}) \|_2$ (dashed blue line) for $h = 1/50$. The right plot is of $\| \operatorname{Re}(\mathbf{u}_h^{FEM}) \|_2$ (solid red line) vs. $\| \operatorname{Re}(\mathbf{u}) \|_2$ (dashed blue line) for $h = 1/50$.

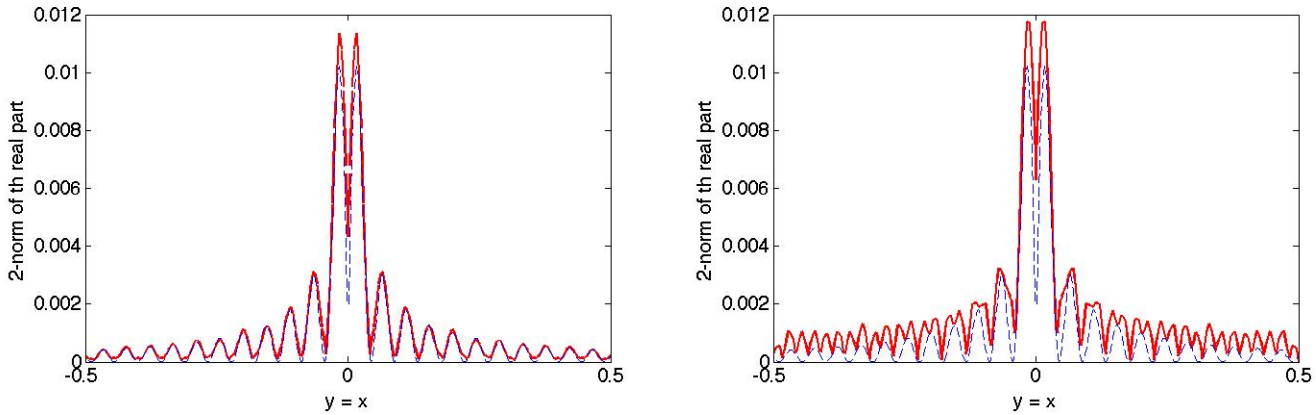


Figure 3.8: The left plot is of $\| \operatorname{Re}(\mathbf{u}_h) \|_2$ (solid red line) vs. $\| \operatorname{Re}(\mathbf{u}) \|_2$ (dashed blue line) for $h = 1/120$. The right plot is of $\| \operatorname{Re}(\mathbf{u}_h^{FEM}) \|_2$ (solid red line) vs. $\| \operatorname{Re}(\mathbf{u}) \|_2$ (dashed blue line) for $h = 1/120$.

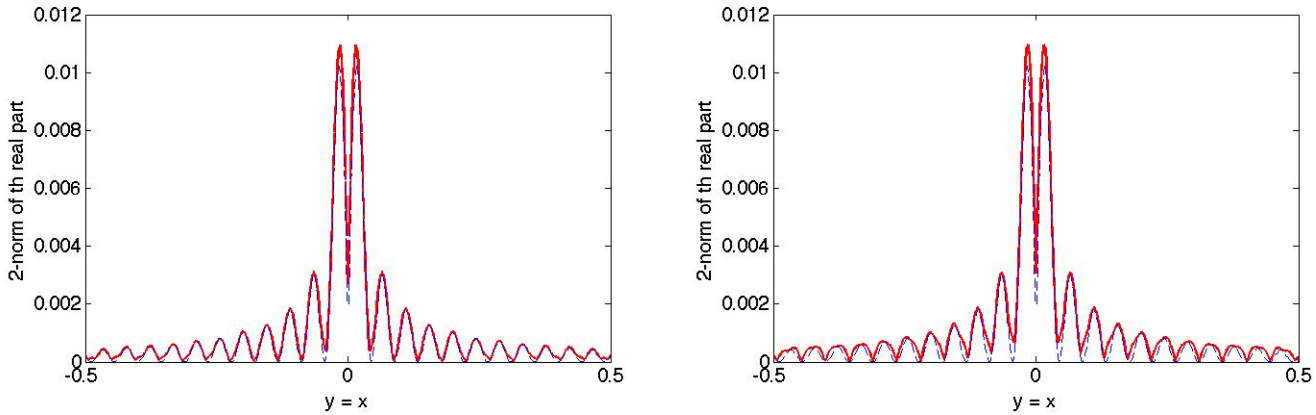


Figure 3.9: The left plot is of $\| \operatorname{Re}(\mathbf{u}_h)\|_2$ (solid red line) vs. $\| \operatorname{Re}(\mathbf{u})\|_2$ (dashed blue line) for $h = 1/200$. The right plot is of $\| \operatorname{Re}(\mathbf{u}_h^{FEM})\|_2$ (solid red line) vs. $\| \operatorname{Re}(\mathbf{u})\|_2$ (dashed blue line) for $h = 1/200$.

Chapter 4

A Multi-modes Monte Carlo Interior Penalty Discontinuous Galerkin Method for Acoustic Wave Scattering in Random Media

4.1 Introduction

Partial differential equations with random coefficients arise naturally in the modeling of many physical phenomena. This is due to the fact that some level of uncertainty is usually involved if the knowledge of the physical behavior or when noise is present in the experimental measurements. In recent years, substantial progress has been made in the numerical approximation of such PDEs due to the significant development in computational resources. We refer to [11, 12, 21, 68, 80] and the references therein for more details.

In this chapter, we consider the propagation of acoustic waves in a medium where the wave velocity is characterized by a random process. More precisely, we study the

approximation of the solution to the following Helmholtz problem:

$$-\Delta u(\omega,\cdot) - k^2\alpha(\omega,\cdot)^2 u(\omega,\cdot) = f(\omega,\cdot) \qquad \text{in } D, \tag{4.1}$$

$$\partial_\nu u(\omega,\cdot) + \mathbf{i}k\alpha(\omega,\cdot)u(\omega,\cdot) = 0 \qquad \text{on } \partial D, \tag{4.2}$$

where k is the wave number, and $D \subset \mathbb{R}^d\,(d = 1, 2, 3)$ is a convex bounded polygonal domain with boundary ∂D. Let $(\Omega, \mathcal{F}, P)$ be a probability space with sample space Ω, $\sigma-$algebra $\mathcal{F}$ and probability measure P. For each fixed $x \in D$, the refractive index $\alpha(\cdot, x)$ is a real-valued random variable defined over Ω. We assume that the medium is a small random perturbation of a uniform background medium in the sense that

$$\alpha(\omega,\cdot) := 1 + \varepsilon\eta(\omega,\cdot). \tag{4.3}$$

Here ε represents the magnitude of the random fluctuation, and $\eta \in L^2(\Omega, L^\infty(D))$ is some random process satisfying

$$P\left\{\omega \in \Omega;\ \|\eta(\omega,\cdot)\|_{L^\infty(D)} \le 1\right\} = 1.$$

For notational brevity, we only consider the case that η is real-valued. However, we note that the results of this chapter are also valid for complex-valued η. On the boundary ∂D, an absorbing boundary condition is imposed to absorb incoming waves [35]. Here ν denotes the unit outward normal to ∂D, and $\partial_\nu u$ stands for the normal derivative of u. The boundary value problem (4.1)–(4.2) arises in the modeling of wave propagation in complex environments, such as composite materials, oil reservoirs and geological basins [46, 55]. In such instances, it is of practical interest to characterize the uncertainty of the wave energy transport when the medium contains some randomness. In particular, we are interested in the computation of some statistics of the wave field, e.g. the expected value of the solution u.

To solve stochastic (or random) partial differential equations (SPDEs) numerically, the simplest and most natural approach is to use the Monte Carlo method, where

a set of independent identically distributed (i.i.d.) solutions are obtained by sampling the PDE coefficients, and the expected value of the solution is calculated via a statistical average over all the sampling in the probability space [21]. An alternative is the stochastic Galerkin method, where the SPDE is reduced to a high dimensional deterministic equation by expanding the random field in the equation using the Karhunen-Loève or Wiener Chaos expansions. We refer the reader to [11, 12, 33, 68, 80] for detailed discussions. However, it is known that a brute-force Monte Carlo or stochastic Galerkin method applied directly to the Helmholtz equation with random coefficients is computationally prohibitive even for a moderate wave number k, since a large number of degrees of freedom is involved in the spatial discretization. It is apparent that in such cases, the Monte Carlo method requires solving a PDE with many sampled coefficients, while the high dimensional deterministic equation associated with the stochastic Galerkin method will be too expensive to be solved.

In this chapter, we propose an efficient numerical method for solving the Helmholtz problem (4.1)–(4.2) when the medium is weakly random defined by (4.3). A multi-modes representation of the solution is derived, where each mode is governed by a Helmholtz equation with deterministic coefficients and a random source. We develop a Monte Carlo interior penalty discontinuous Galerkin (MCIP-DG) method for approximating the mode functions. In particular, we take advantage of the fact that the coefficients of the Helmholtz equation for all the modes are identical; hence, the associated discretized equations share the same constant coefficient matrix. Using this crucial fact, it is observed that an LU direct solver for the discretized equations leads to a tremendous saving in the computational costs, since the LU decomposition matrices can be used repeatedly. Thus, all of the solutions for all modes and all samples can be obtained in an efficient way by performing simple forward and backward substitutions. Indeed, it turns out that the computational complexity of the proposed algorithm is comparable to that of solving a few deterministic Helmholtz problems using the LU direct solver.

The rest of the chapter is organized as follows. A wave-number explicit estimate for the solution of the random Helmholtz equation is established in Section 4.2. In Section 4.3, we introduce the multi-modes expansion of the solution as a power series of ε and analyze the error estimation for its finite-modes approximation. The Monte Carlo interior penalty discontinuous Galerkin method is presented in Section 4.4, where the error estimates for the approximation of each mode function is also obtained. In Section 4.5, a numerical procedure for solving (4.1)–(4.2) is described and its computational complexity is analyzed in detail. In addition, we derive optimal order error estimates for the proposed procedure. Several numerical experiments are provided in Section 6 to demonstrate the efficiency of the method and to validate the theoretical results.

4.2 PDE Analysis

The focus of this section will be to derive a priori solution estimates for the random Helmholtz problem introduced in (4.1)–(4.2). These a priori estimates will be used to prove existence and uniqueness of the solution to the random Helmholtz problem. The techniques in this chapter will mirror those carried out for the deterministic scalar Helmholtz problem (1.1)–(1.2) (c.f. Chapter 2 and [27]).

4.2.1 Preliminaries

Similar to Chapter 2, analysis in this section will be carried out using the following special function spaces:

$$H^1_+(D) := \left\{v \in H^1(D);\, |\nabla v|\big|_\Gamma \in L^2(\partial D)\right\}, \tag{4.4}$$

$$V := \left\{v \in H^1(D);\, \Delta v \in L^2(D)\right\}. \tag{4.5}$$

Without loss of generality, we assume that the domain $D \subset B_R(0)$. Throughout this chapter we also assume that D is a star-shaped domain with respect to the origin

in the sense that there exists a positive constant c_0 such that

$$x \cdot \nu \geq c_0 \quad \text{on } \partial D,$$

Recall that the analysis in Chapter 2 also relied on a star-shape condition on the domain.

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which all the random variables of this chapter are defined. $\mathbb{E}(\cdot)$ denotes the expectation operator on this probability space. The abbreviation *a.s.* stands for *almost surely*.

As it will be needed in the late sections of this chapter, in this section we analyze the boundary value problem for the Helmholtz equation (4.1) with the following slightly more general nonhomogeneous boundary condition:

$$\partial_\nu u(\omega, \cdot) + \mathbf{i}k\alpha(\omega, \cdot)u(\omega, \cdot) = g(\omega, \cdot). \tag{4.6}$$

As in Chapter 2, analysis of the random Helmholtz problem (4.1),(4.6) will be carried out on its weak formulation. With this in mind, we introduce the following definition.

Definition 4.2.1. *Let $f \in L^2(\Omega, L^2(D))$ and $g \in L^2(\Omega, L^2(\partial D))$. A function $u \in L^2(\Omega, H^1(D))$ is called a weak solution to problem (4.1),(4.6) if it satisfies the following identity:*

$$\int_\Omega a(u, v)\, dP = \int_\Omega \big((f, v)_D + \langle g, v\rangle_{\partial D}\big)\, dP \qquad \forall v \in L^2(\Omega, H^1(D)), \tag{4.7}$$

where

$$a(w, v) := \big(\nabla w, \nabla v\big)_D - k^2\big(\alpha^2 w, v\big)_D + \mathbf{i}k\, \langle \alpha w, v\rangle_{\partial D}\,. \tag{4.8}$$

Remark 4.2.2. *Using (4.10) below, it is easy to show that any solution u of (4.1),(4.6) satisfies $u \in L^2(\Omega, H^1_+(D) \cap V)$.*

4.2.2 Wave-number Explicit Solution Estimates

In this subsection, we shall derive stability estimates for the solution of problem (4.1),(4.6) which is defined in Definition 4.2.1. Similar to Chapter 2, our focus is to obtain explicit dependence of the stability constants on the wave number k. Such wave-number explicit stability estimates will play a vital role in our convergence analysis in the later sections. We note that wave-number explicit stability estimates also play a pivotal role in the development of numerical methods, such as finite element and discontinuous Galerkin methods, for deterministic reduced wave equations (cf. Chapter 3 and [42, 44]). As a byproduct of the stability estimates, the existence and uniqueness of solutions to problem (4.1),(4.6) will be conveniently established.

Lemma 4.2.3. *Let $u \in L^2(\Omega, H^1(D))$ be a solution of (4.1),(4.6), then for any $\delta_1, \delta_2 > 0$ and $\varepsilon < 1$ there hold*

$$\begin{aligned}\mathbb{E}(\|\nabla u\|^2_{L^2(D)}) \leq &\left(k^2(1+\varepsilon)^2 + \delta_1\right)\mathbb{E}(\|u\|^2_{L^2(D)}) \\ &+ \left(\frac{\delta_1}{2k^2(1-\varepsilon)^2} + \frac{1}{2\delta_1}\right)\left(\mathbb{E}(\|f\|^2_{L^2(D)}) + \mathbb{E}(\|g\|^2_{L^2(\partial D)})\right), \end{aligned} \tag{4.9}$$

$$\begin{aligned}\mathbb{E}(\|u\|^2_{L^2(\partial D)}) \leq &\frac{\delta_2}{k(1-\varepsilon)}\mathbb{E}(\|u\|^2_{L^2(D)}) + \frac{1}{\delta_2 k(1-\varepsilon)}\mathbb{E}(\|f\|^2_{L^2(D)}) \\ &+ \frac{1}{k^2(1-\varepsilon)^2}\mathbb{E}(\|g\|^2_{L^2(\partial D)}). \end{aligned} \tag{4.10}$$

Proof. Setting $v = u$ in (4.7) yields

$$\int_\Omega a(u,u)\, dP = \int_\Omega \left((f,u)_D + \langle g, v\rangle_{\partial D}\right) dP.$$

Taking the real and imaginary parts and using the definition of $a(\cdot,\cdot)$, we get

$$\int_\Omega \left(\|\nabla u\|^2_{L^2(D)} - k^2\|(1+\varepsilon\eta)u\|^2_{L^2(D)}\right) dP = \mathrm{Re}\int_\Omega \left((f,u)_D + \langle g, v\rangle_{\partial D}\right) dP, \tag{4.11}$$

$$k\int_\Omega \langle 1+\varepsilon\eta, |u|^2\rangle_{\partial D}\, dP = \mathrm{Im}\int_\Omega \left((f,u)_D + \langle g, v\rangle_{\partial D}\right) dP. \tag{4.12}$$

Applying the Cauchy-Schwarz inequality to (4.12) produces

$$k(1-\varepsilon)\mathbb{E}(\|u\|^2_{L^2(\partial D)}) \leq \frac{\delta_2}{2}\mathbb{E}(\|u\|^2_{L^2(D)}) + \frac{1}{2\delta_2}\mathbb{E}(\|f\|^2_{L^2(D)}) + \frac{k(1-\varepsilon)}{2}\mathbb{E}(\|u\|^2_{L^2(\partial D)}) + \frac{1}{2k(1-\varepsilon)}\mathbb{E}(\|g\|^2_{L^2(\partial D)}).$$

Thus, (4.10) holds. Applying Cauchy-Schwarz to (4.11) yields

$$\mathbb{E}(\|\nabla u\|^2_{L^2(D)}) \leq \left(k^2(1+\varepsilon)^2 + \frac{\delta_1}{2}\right)\mathbb{E}(\|u\|^2_{L^2(D)}) + \frac{1}{2\delta_1}\mathbb{E}(\|f\|^2_{L^2(D)}) + \frac{\delta_1}{2}\mathbb{E}(\|u\|^2_{L^2(\partial D)}) + \frac{1}{2\delta_1}\mathbb{E}(\|g\|^2_{L^2(\partial D)}),$$

which together with (4.10) (using $\delta_2 = k(1-\varepsilon)$) infers (4.9). The proof is complete. □

From Lemma 4.2.3, We observe that the test function $v = u$ is not enough to obtain solution estimates for the weak solution of (4.1), (4.6). Recall that this was also the case for the deterministic scalar Helmholtz equation. To overcome this difficulty, in Chapter 2 and [27] we made use of Rellich identities for the Laplacian operator. With this in mind, we prove the following lemma.

Lemma 4.2.4. *Let $u \in L^2(\Omega, H^2(D))$, then*

$$\mathrm{Re}\int_\Omega \big(u, x\cdot\nabla u\big)_D \, dP = -\frac{d}{2}\int_\Omega \|u\|^2_{L^2(D)}\, dP + \frac{1}{2}\int_\Omega \langle x\cdot\nu, |u|^2\rangle_{\partial D}\, dP, \tag{4.13}$$

$$\mathrm{Re}\int_\Omega \big(\nabla u, \nabla(x\cdot\nabla u)\big)_D \, dP = \frac{2-d}{2}\int_\Omega \|\nabla u\|^2_{L^2(D)}\, dP + \frac{1}{2}\int_\Omega \big\langle x\cdot\nu, |\nabla u|^2\big\rangle_{\partial D}\, dP. \tag{4.14}$$

Proof. To obtain the above result, we apply Lemmas 2.2.2 and 2.2.3 with $\boldsymbol{\alpha} = x$ and integrate the subsequent identities over the probability space $(\Omega, \mathcal{F}, P)$. □

Remark 4.2.5. (4.14) *could be called a stochastic Rellich identity for the Laplacian.*

We are now ready to state and prove our wave-number explicit estimate for solutions of problem (4.1), (4.6) defined in Definition 4.2.1.

Theorem 4.2.6. *Let $u \in L^2(\Omega, H^1_+(D))$ be a solution of (4.1)–(4.6) and R be the smallest number such that $B_R(0)$ contains the domain D. Then there hold the following estimates:*

$$\mathbb{E}\Big(\|u\|^2_{L^2(D)} + \|u\|^2_{L^2(\partial D)} + c_0\,\|\nabla u\|^2_{L^2(\partial D)}\Big) \leq C_0\Big(\frac{1}{k} + \frac{1}{k^2}\Big)^2 M(f,g), \tag{4.15}$$

$$\mathbb{E}(\|u\|^2_{H^1(D)}) \leq C_0\Big(1 + \frac{1}{k^2}\Big)^2 M(f,g), \tag{4.16}$$

provided that $\varepsilon(2+\varepsilon) < \gamma_0 := \min\big\{1, \frac{13-2d}{2(4d-7)+25kR}\big\}$. Where C_0 is some positive constant independent of k and u, and

$$M(f,g) := \mathbb{E}\Big(\|f\|^2_{L^2(D)} + \|g\|^2_{L^2(\partial D)}\Big). \tag{4.17}$$

Moreover, if $g \in L^2(\Omega, H^{\frac{1}{2}}(D))$ and $u \in L^2(\Omega, H^2(D))$, there also holds

$$\mathbb{E}(\|u\|^2_{H^2(D)}) \leq C\Big(k + \frac{1}{k^2}\Big)^2 \mathbb{E}\Big(\|f\|^2_{L^2(D)} + \|g\|^2_{H^{\frac{1}{2}}(\partial D)}\Big). \tag{4.18}$$

Proof. To avoid some technicalities, below we only give a proof for the case $u \in L^2(\Omega, H^2(D))$. For the general case, u needs be replaced by its mollification u_ρ at the beginning of the proof and followed by taking the limit $\rho \to 0$ after the integration by parts is done. A similar strategy was adopted in the proofs of the generalized weak coercivity properties in Chapter 2.

Setting $v = x \cdot \nabla u$ in (4.7) yields

$$\int_\Omega \Big((\nabla u, \nabla v)_D - k^2(\alpha^2 u, v)_D + \mathbf{i}k\langle \alpha u, v\rangle_{\partial D}\Big)\, dP = \int_\Omega \left((f,v)_D + \langle g, v\rangle_{\partial D}\right) dP. \tag{4.19}$$

Using (4.13) and (4.14) after taking the real part of (4.19) and regrouping we get

$$\begin{aligned}\frac{dk^2}{2}\int_\Omega \|u\|^2_{L^2(D)}\,dP &= \int_\Omega \Big(\frac{d-2}{2}\|\nabla u\|^2_{L^2(D)} + k^2\varepsilon\,\mathrm{Re}\big(\eta(2+\varepsilon\eta),v\big)_D\Big)\,dP\\ &\quad - \int_\Omega \Big(k\mathrm{Im}\langle(1+\varepsilon\eta)u,v\rangle_{\partial D} + \frac{1}{2}\left\langle x\cdot\nu,|\nabla u|^2\right\rangle_{\partial D} - \frac{k^2}{2}\left\langle x\cdot\nu,|u|^2\right\rangle_{\partial D}\Big)\,dP\\ &\quad + \int_\Omega \left(\mathrm{Re}(f,v)_D + \mathrm{Re}\langle g,v\rangle_{\partial D}\right)\,dP.\end{aligned}$$

It then follows from the Cauchy-Schwarz inequality, the star-shape condition, and the facts that $|x|\le R$ for $x\in D$ and $\|\eta\|_{L^\infty(D)}\le 1$ a.s. that

$$\begin{aligned}\frac{dk^2}{2}\mathbb{E}(\|u\|^2_{L^2(D)}) &\le k^2\varepsilon R(2+\varepsilon)\Big(\frac{1}{2\delta_1}\mathbb{E}(\|u\|^2_{L^2(D)}) + \frac{\delta_1}{2}\mathbb{E}(\|\nabla u\|^2_{L^2(D)})\Big)\\ &\quad + \frac{d-2}{2}\mathbb{E}(\|\nabla u\|^2_{L^2(D)}) + \frac{R}{2\delta_2}\mathbb{E}(\|f\|^2_{L^2(D)}) + \frac{R\delta_2}{2}\mathbb{E}(\|\nabla u\|^2_{L^2(D)})\\ &\quad + \frac{R}{2\delta_3}\mathbb{E}(\|g\|^2_{L^2(\partial D)}) + \frac{R\delta_3}{2}\mathbb{E}(\|\nabla u\|^2_{L^2(\partial D)})\\ &\quad + \frac{kR}{\delta_4}\mathbb{E}(\|u\|^2_{L^2(\partial D)}) + kR\delta_4\mathbb{E}(\|\nabla u\|^2_{L^2(\partial D)})\\ &\quad - \frac{c_0}{2}\mathbb{E}\big(\|\nabla u\|^2_{L^2(\partial D)}\big) + \frac{k^2R}{2}\mathbb{E}\big(\|u\|^2_{L^2(\partial D)}\big).\end{aligned}$$

At this point, we note that $\varepsilon(2+\varepsilon)\le 1$ implies $\varepsilon\le\frac{1}{2}$. Setting $\delta_3=\frac{c_0}{4R}$, $\delta_4=\frac{c_0}{8kR}$, denoting $\gamma=\varepsilon(2+\varepsilon)$, and using Lemma 4.2.3 we can bound right-hand side as

follows:

$$\begin{aligned}
\frac{dk^2}{2}\mathbb{E}(\|u\|^2_{L^2(D)}) &\le \Big(\frac{d-2}{2}+\frac{k^2R\gamma\delta_1}{2}+\frac{R\delta_2}{2}\Big)\mathbb{E}(\|\nabla u\|^2_{L^2(D)})+\frac{k^2R\gamma}{2\delta_1}\mathbb{E}(\|u\|^2_{L^2(D)})\\
&\quad+\Big(\frac{8k^2R^2}{c_0}+\frac{k^2R}{2}\Big)\mathbb{E}(\|u\|^2_{L^2(\partial D)})-\frac{c_0}{4}\mathbb{E}\big(\|\nabla u\|^2_{L^2(\partial D)}\big)\\
&\quad+\frac{R}{2\delta_2}\mathbb{E}(\|f\|^2_{L^2(D)})+\frac{2R^2}{c_0}\mathbb{E}(\|g\|^2_{L^2(\partial D)})\\
&\le \Big(\frac{d-2}{2}+\frac{k^2R\gamma\delta_1}{2}+\frac{R\delta_2}{2}\Big)\big(k^2(1+\gamma)+\delta_5\big)\,\mathbb{E}(\|u\|^2_{L^2(D)})\\
&\quad+\Big(\frac{d-2}{2}+\frac{k^2R\gamma\delta_1}{2}+\frac{R\delta_2}{2}\Big)\Big(\frac{2\delta_5}{k^2}+\frac{1}{2\delta_5}\Big)\Big(\mathbb{E}(\|f\|^2_{L^2(D)})+\mathbb{E}(\|g\|^2_{L^2(\partial D)})\Big)\\
&\quad+\Big(\frac{8k^2R^2}{c_0}+\frac{k^2R}{2}\Big)\Big(2\delta_6\mathbb{E}(\|u\|^2_{L^2(D)})+\frac{2}{k^2\delta_6}\mathbb{E}(\|f\|^2_{L^2(D)})+\frac{4}{k^2}\mathbb{E}(\|g\|^2_{L^2(\partial D)})\Big)\\
&\quad+\frac{k^2R\gamma}{2\delta_1}\mathbb{E}(\|u\|^2_{L^2(D)})+\frac{R}{2\delta_2}\mathbb{E}(\|f\|^2_{L^2(D)})+\frac{2R^2}{c_0}\mathbb{E}(\|g\|^2_{L^2(\partial D)})\\
&\quad-\frac{c_0}{4}\mathbb{E}\big(\|\nabla u\|^2_{L^2(\partial D)}\big),
\end{aligned}$$

which is equivalent to

$$c_1\mathbb{E}(\|u\|^2_{L^2(D)})+\frac{c_0}{4}\mathbb{E}(\|\nabla u\|^2_{L^2(\partial D)})\le c_2\mathbb{E}(\|f\|^2_{L^2(D)}), \tag{4.20}$$

where

$$\begin{aligned}
c_1 &:= k^2-\frac{d-2}{2}\Big(k^2\gamma+\delta_5\Big)-\Big(\frac{k^2R\gamma\delta_1}{2}+\frac{R\delta_2}{2}\Big)\Big(k^2(1+\gamma)+\delta_5\Big)\\
&\qquad-\Big(\frac{16k^2R^2}{c_0}+k^2R\Big)\delta_6-\frac{k^2R\gamma}{2\delta_1},\\
c_2 &:= \Big(\frac{d-2}{2}+\frac{k^2R\gamma\delta_1}{2}+\frac{R\delta_2}{2}\Big)\Big(\frac{2\delta_5}{k^2}+\frac{1}{2\delta_5}\Big)\\
&\qquad+\Big(\frac{32k^2R^2}{c_0}+2k^2R\Big)\Big(\frac{2}{k^2\delta_6}+\frac{4}{k^2}\Big)+\frac{R}{2\delta_2}+\frac{2R^2}{c_0}.
\end{aligned}$$

Let $\delta_1 = \frac{1}{2k}$, $\delta_2 = \frac{1}{4R}$, $\delta_5 = \frac{k^2}{4}$, and $\delta_6 = \frac{1}{4\left(\frac{16R^2}{c_0}+R\right)}$, then

$$
\begin{aligned}
c_1 &= k^2\Big[\frac{27-4d}{32} - \Big(\frac{4d-7}{8} + \frac{(21+4\gamma)Rk}{16}\Big)\gamma\Big],\\
c_2 &= \Big(\frac{d-2}{2} + \frac{kR\gamma}{4} + \frac{1}{8}\Big)\Big(\frac{1}{2} + \frac{1}{8k^2}\Big) + 2R^2(1+c_0)\\
&\quad + 8\Big(\frac{16R^2}{c_0} + R\Big)\Big(\frac{16R^2}{c_0} + R + 1\Big).
\end{aligned}
$$

If $\gamma < \gamma_0$, it is easy to check that $c_1 \geq \frac{k^2}{32}$. Thus, (4.20) infers that

$$
\mathbb{E}(\|u\|^2_{0,D}) + c_0\mathbb{E}(\|\nabla u\|^2_{L^2(\partial D)}) \leq \frac{C}{k^2}\Big(1+\frac{1}{k^2}\Big)\Big(\mathbb{E}(\|f\|^2_{L^2(D)}) + \mathbb{E}(\|g\|^2_{L^2(\partial D)})\Big) \tag{4.21}
$$

for some constant $C > 0$ independent of k and u. We combine this result with (4.10) (using $\delta_2 = k$) to obtain (4.15).

By (4.9) (using $\delta_1 = k^2$) and (4.21) we get

$$
\begin{aligned}
\mathbb{E}(\|u\|^2_{H^1(D)}) &= \mathbb{E}(\|u\|^2_{L^2(D)}) + \mathbb{E}(\|\nabla u\|^2_{L^2(D)})\\
&\leq \frac{C}{k^2}\Big(1+\frac{1}{k^2}\Big)\Big(\mathbb{E}(\|f\|^2_{L^2(D)}) + \mathbb{E}(\|g\|^2_{L^2(\partial D)})\Big)\\
&\qquad + \big(k^2(1+\varepsilon)^4 + k^2\big)\,\mathbb{E}(\|u\|^2_{L^2(D)})\\
&\qquad + \Big(1+\frac{1}{2k^2}\Big)\Big(\mathbb{E}(\|f\|^2_{L^2(D)}) + \mathbb{E}(\|g\|^2_{L^2(\partial D)})\Big)\\
&\leq C\Big(1+\frac{1}{k^2}\Big)^2\Big(\mathbb{E}(\|f\|^2_{L^2(D)}) + \mathbb{E}(\|g\|^2_{L^2(\partial D)})\Big).
\end{aligned}
$$

Hence, (4.16) holds.

Finally, it follows from the standard elliptic regularity theory for the Poisson equation and the trace inequality (cf. [47]) that

$$\begin{aligned}\mathbb{E}(\|u\|^2_{H^2(D)}) &\le C\Big(\mathbb{E}(\big\|k^2u\big\|^2_{L^2(D)}) + \mathbb{E}(\|f\|^2_{L^2(D)}) + \mathbb{E}(\|g\|^2_{H^{\frac{1}{2}}(\partial D)}) \\ &\quad + \mathbb{E}(\|ku\|^2_{H^{\frac{1}{2}}(\partial D)} + \mathbb{E}(\|u\|^2_{L^2(D)})\Big) \\ &\le C\mathbb{E}\Big(k^4\|u\|^2_{L^2(D)} + \|f\|^2_{L^2(D)} + \|g\|^2_{H^{\frac{1}{2}}(\partial D)}\Big) \\ &\quad + C\mathbb{E}\Big(k^2\|\nabla u\|^2_{L^2(D)} + \|u\|^2_{L^2(D)}\Big) \\ &\le C\Big(k + \frac{1}{k^2}\Big)^2\mathbb{E}\Big(\|f\|^2_{L^2(D)} + \|g\|^2_{H^{\frac{1}{2}}(\partial D)}\Big).\end{aligned}$$

Hence (4.18) holds. The proof is complete. □

Remark 4.2.7. *By the definition of γ_0, we see that $\gamma_0 = O\big(\frac{1}{kR}\big)$. In practice, this is not a restrictive condition because R is often taken to be proportional to the wave length. Hence, $kR = O(1)$.*

As a non-trivial byproduct, the above stability estimates can be used conveniently to establish the existence and uniqueness of solutions to problem (4.1),(4.6) as defined in Definition 4.2.1. This strategy was mentioned at the end of Chapter 2 and is carried out explicitly in the following theorem.

Theorem 4.2.8. *Let $f \in L^2(\Omega, L^2(D))$ and $g \in L^2(\Omega, L^2(\partial D)$. For each fixed pair of positive numbers k and ε such that $\varepsilon(2+\varepsilon) < \gamma_0$, there exists a unique solution $u \in L^2(\Omega, H^1_+(D) \cap V)$ to problem (4.1),(4.6).*

Proof. The proof is based on the well-known Fredholm Alternative Principle (cf. [47]). First, it is easy to check that the sesquilinear form on the right-hand side of (4.7) satisfies a Gärding's inequality on the space $L^2(\Omega, H^1(D))$. Second, to apply the Fredholm Alternative Principle we need to prove that solutions to the adjoint problem of (4.7)–(4.8) is unique. It is easy to verify that the adjoint problem is associated

with the sesquilinear form

$$\widehat{a}(w,v) := \big(\nabla w, \nabla v\big)_D - k^2\big(\alpha^2 w, v\big)_D - \mathbf{i}k\,\langle \alpha w, v\rangle_{\partial D}\,,$$

which differs from $a(\cdot,\cdot)$ only in the sign of the last term. As a result, all the stability estimates for problem (4.7)–(4.8) still hold for its adjoint problem. Since the adjoint problem is a linear problem (so is problem (4.7)–(4.8)), the stability estimates immediately infer uniqueness. Finally, the Fredholm Alternative Principle then implies that problem (4.7)–(4.8) has a unique solution $u \in L^2(\Omega, H^1(D))$. The proof is complete. □

Remark 4.2.9. *The uniqueness of the adjoint problem can also be shown using the classical unique continuation argument (cf. [57]).*

4.3 Multi-modes Representation of the Solution and its Finite Modes Approximations

The first goal of this section is to develop a multi-modes representation for the solution to problem (4.1)–(4.2) in terms of powers of the parameter ε. We first postulate such a representation and then prove its validity by establishing some energy estimates for all the mode functions. The second goal of this section is to establish an error estimate for finite modes approximations of the solution. Both the multi-modes representation and its finite modes approximations play a pivotal role in our overall solution procedure for solving problem (4.1)–(4.2) as they provide the theoretical foundation for the solution procedure. Throughout this section, we use u^ε to denote the solution to problem (4.1)–(4.2) which is proved in Theorem 4.2.8.

We start by postulating that the solution u^ε has the following multi-modes expansion:

$$u^\varepsilon = \sum_{n=0}^{\infty} \varepsilon^n u_n, \tag{4.22}$$

whose validity will be justified later. Without loss of generality, we assume that $k \geq 1$ and $D \subset B_1(0)$. Otherwise, the problem can be rescaled to this regime by a suitable change of variable. We note that the normalization $D \subset B_1(0)$ implies that $R = 1$.

Substituting the above expansion into the Helmholtz equation (4.1) we get

$$\begin{aligned} f &= -\Delta u^\varepsilon - k^2\alpha^2 u^\varepsilon \\ &= \sum_{n=0}^{\infty} \varepsilon^n \Big(-\Delta u_n - k^2\big(1 + 2\varepsilon\eta + \varepsilon^2\eta^2\big) u_n \Big) \\ &= \sum_{n=0}^{\infty} \Big(\varepsilon^n \big(-\Delta u_n - k^2 u_n \big) - 2\varepsilon^{n+1}\eta k^2 u_n - \varepsilon^{n+2} k^2 \eta^2 u_n \Big) \\ &= -\Delta u_0 - k^2 u_0 - \varepsilon\big(-\Delta u_1 - k^2 u_1 - 2k^2\eta u_0 \big) \\ &\quad + \sum_{n=2}^{\infty} \varepsilon^n \big(-\Delta u_n - k^2 u_n - 2k^2\eta u_{n-1} - k^2\eta^2 u_{n-2} \big). \end{aligned}$$

Matching the coefficients of ε^n order terms for $n = 0, 1, 2, \cdots$, we obtain

$$u_{-1} :\equiv 0, \tag{4.23}$$

$$-\Delta u_0 - k^2 u_0 = f, \tag{4.24}$$

$$-\Delta u_n - k^2 u_n = 2k^2\eta u_{n-1} + k^2\eta^2 u_{n-2} \qquad \text{for } n \geq 1. \tag{4.25}$$

Similarly, the boundary condition (4.2) gives

$$\begin{aligned} 0 &= \frac{\partial u^\varepsilon}{\partial \nu} + \mathbf{i}k(1 + \varepsilon\eta) u^\varepsilon \\ &= \sum_{n=0}^{\infty} \Big(\varepsilon^n \frac{\partial u_n}{\partial \nu} + \varepsilon^n \mathbf{i}k u_n + \varepsilon^{n+1} \mathbf{i}k\eta u_n \Big) \\ &= \frac{\partial u_0}{\partial \nu} + \mathbf{i}k u_0 + \sum_{n=1}^{\infty} \varepsilon^n \Big(\frac{\partial u_n}{\partial \nu} + \mathbf{i}k u_n + \mathbf{i}k\eta u_{n-1} \Big). \end{aligned}$$

This translates to each mode function u_n as follows:

$$\partial_\nu u_n + \mathbf{i}k u_n = -\mathbf{i}k\eta u_{n-1} \qquad \text{for } n \geq 0. \tag{4.26}$$

We note that the non-zero right hand side term in (4.26) was the motivation to study the random Helmholtz equation with non-homogeneous boundary data given by (4.1), (4.6) in Section 4.2.

A remarkable feature of the above multi-modes expansion is that all the mode functions satisfy the same type (nearly deterministic) Helmholtz equation and the same boundary condition. The only difference is that the Helmholtz equations have different right-hand side source terms, and each pair of consecutive mode functions supply the source term for the Helmholtz equation satisfied by the next mode function. This remarkable feature will be crucially utilized in Section 4.5 to construct our overall numerical methodology for solving problem (4.1)–(4.2).

Next, we address the existence and uniqueness of each mode function u_n.

Theorem 4.3.1. *Let $f \in L^2(\Omega, L^2(D))$. Then for each $n \geq 0$, there exists a unique solution $u_n \in L^2(\Omega, H^1(D))$ (understood in the sense of Definition 4.2.1) to problem (4.24),(4.26) for $n = 0$ and problem (4.25),(4.26) for $n \geq 1$. Moreover, for $n \geq 0$, u_n satisfies*

$$\begin{aligned} &\mathbb{E}\Big(\|u_n\|^2_{L^2(D)} + \|u_n\|^2_{L^2(\partial D)} + c_0\,\|\nabla u_n\|^2_{L^2(\partial D)}\Big) \\ &\qquad \leq \Big(\frac{1}{k} + \frac{1}{k^2}\Big)^2 C(n,k)\mathbb{E}(\|f\|^2_{L^2(D)}), \end{aligned} \tag{4.27}$$

$$\mathbb{E}\big(\|u_n\|^2_{H^1(D)}\big) \leq \Big(1 + \frac{1}{k^2}\Big)^2 C(n,k)\mathbb{E}(\|f\|^2_{L^2(D)}), \tag{4.28}$$

where

$$C(0,k) := C_0, \qquad C(n,k) := 4^{2n-1}C_0^{n+1}(1+k)^{2n} \quad \text{for } n \geq 1. \tag{4.29}$$

Furthermore, if $u_n \in L^2(\Omega, H^2(D))$, there also holds

$$\mathbb{E}(\|u_n\|^2_{H^2(D)}) \leq \frac{1}{\bar{c}_0}\Big(k + \frac{1}{k^2}\Big)^2 C(n,k)\mathbb{E}(\|f\|^2_{L^2(D)}), \tag{4.30}$$

where $\bar{c}_0 := \min\{1, kc_0\}$.

Proof. For each $n \geq 0$, the PDE problem associated with u_n is the same type Helmholtz problem as the original problem (4.1)–(4.2) (with $\varepsilon = 0$ in the left-hand side of the PDE). Hence, all a priori estimates of Theorem 4.2.6 hold for each u_n (with its respective right-hand source side function). First, we have

$$\begin{aligned}\mathbb{E}\Big(\|u_0\|^2_{L^2(D)} + \|u_0\|^2_{L^2(\partial D)} + c_0\,\|\nabla u_0\|^2_{L^2(\partial D)}\Big) \\ \leq C_0\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2\mathbb{E}(\|f\|^2_{L^2(D)}),\end{aligned} \tag{4.31}$$

$$\mathbb{E}(\|u_0\|^2_{H^1(D)}) \leq C_0\Big(1+\frac{1}{k^2}\Big)^2\mathbb{E}(\|f\|^2_{L^2(D)}). \tag{4.32}$$

Thus, (4.27) and (4.28) hold for $n = 0$. Without loss of generality we assume that $C_0 \geq 1$.

Next, we use induction to prove that (4.27) and (4.28) hold for all $n > 0$. Assume that (4.27) and (4.28) hold for all $0 \leq n \leq \ell - 1$, then

$$\begin{aligned}&\mathbb{E}\Big(\|u_\ell\|^2_{L^2(D)} + \|u_\ell\|^2_{L^2(\partial D)} + c_0\,\|\nabla u_\ell\|^2_{L^2(\partial D)}\Big)\\
&\quad\leq 2C_0\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2\mathbb{E}\Big(\big\|2k^2\eta u_{\ell-1}\big\|^2_{L^2(D)} + \overline{\delta}_{1\ell}\,\big\|k^2\eta^2 u_{\ell-2}\big\|^2_{L^2(D)} + \|k\eta u_{\ell-1}\|^2_{L^2(\partial D)}\Big)\\
&\quad\leq 2C_0\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2(1+k)^2\Big(4C(\ell-1,k)+C(\ell-2,k)\Big)\mathbb{E}(\|f\|^2_{L^2(D)})\\
&\quad\leq \Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2 8C_0(1+k)^2C(\ell-1,k)\left(1+\frac{C(\ell-2,k)}{C(\ell-1,k)}\right)\mathbb{E}(\|f\|^2_{L^2(D)})\\
&\quad\leq \Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2 C(\ell,k)\mathbb{E}(\|f\|^2_{L^2(D)}),\end{aligned}$$

where $\overline{\delta}_{1\ell} = 1 - \delta_{1\ell}$ and $\delta_{1\ell}$ denotes the Kronecker delta. To obtain the above result, we note that $k, C_0 \geq 1$ yield the following inequality:

$$\begin{aligned}
&8C_0(1+k)^2 C(\ell-1,k)\left(1+\frac{C(\ell-2,k)}{C(\ell-1,k)}\right) \\
&\quad = 8C_0(1+k)^2 C(\ell-1,k)\left(1+\frac{4^{2(\ell-2)-1}C_0^{\ell-1}(1+k)^{2(\ell-2)}}{4^{2(\ell-1)-1}C_0^{\ell}(1+k)^{2(\ell-1)}}\right) \\
&\quad = 8C_0(1+k)^2 C(\ell-1,k)\left(1+\frac{1}{4^2 C_0(1+k)^2}\right) \\
&\quad \leq 4^2 C_0(1+k)^2 C(\ell-1,k) \\
&\quad = C(\ell,k)
\end{aligned}$$

for $\ell \geq 2$.

Similarly,

$$\begin{aligned}
&\mathbb{E}\big(\|u_\ell\|^2_{H^1(D)}\big) \\
&\quad \leq 2C_0\Big(1+\frac{1}{k^2}\Big)^2 \mathbb{E}\Big(\big\|2k^2\eta u_{\ell-1}\big\|^2_{L^2(D)} + \overline{\delta}_{1\ell}\big\|k^2\eta^2 u_{\ell-2}\big\|^2_{L^2(D)} + \|k\eta u_{\ell-1}\|^2_{L^2(\partial D)}\Big) \\
&\quad \leq 2C_0\Big(1+\frac{1}{k^2}\Big)^2 (1+k)^2\Big(4C(\ell-1,k) + C(\ell-2,k)\Big)\mathbb{E}\big(\|f\|^2_{L^2(D)}\big) \\
&\quad \leq \Big(1+\frac{1}{k^2}\Big)^2 C(\ell,k)\mathbb{E}\big(\|f\|^2_{L^2(D)}\big).
\end{aligned}$$

Hence, (4.27) and (4.28) hold for $n = \ell$. So the induction argument is complete.

We now use (4.27) and the elliptic theory for Poisson problems directly to verify estimate (4.30).

$$
\begin{aligned}
&\mathbb{E}\big(\|u_n\|^2_{H^2(\partial D)}\big)\\
&\quad\le 2C_0\Big(k+\frac{1}{k^2}\Big)^2\mathbb{E}\Big(\big\|2k^2\eta u_{n-1}\big\|^2_{L^2(D)}+\big\|k^2\eta^2 u_{n-2}\big\|^2_{L^2(D)}+\|k\eta u_{n-1}\|^2_{H^{\frac{1}{2}}(\partial D)}\Big)\\
&\quad\le 2C_0\Big(k+\frac{1}{k^2}\Big)^2k^2\mathbb{E}\Big(4\,\|u_{n-1}\|^2_{L^2(D)}+\|u_{n-2}\|^2_{L^2(D)}+\frac{c_0}{kc_0}\,\|\nabla u_{n-1}\|^2_{L^2(\partial D)}\Big)\\
&\quad\le \frac{2}{\bar{c}_0}C_0\Big(k+\frac{1}{k^2}\Big)^2(1+k)^2\Big(4C(n-1,k)+C(n-2,k)\Big)\mathbb{E}\big(\|f\|^2_{L^2(D)}\big)\\
&\quad\le \frac{1}{\bar{c}_0}\Big(k+\frac{1}{k^2}\Big)^2C(n,k)\mathbb{E}\big(\|f\|^2_{L^2(D)}\big).
\end{aligned}
$$

Hence, (4.30) holds for all $n\ge 0$.

With a priori estimates (4.27) and (4.28) in hand, the proof of existence and uniqueness of each u_n follows verbatim the proof of Theorem 4.2.8. The proof is complete. □

Now we are ready to justify the multi-modes representation (4.22) for the solution u^ε of problem (4.1)–(4.2). To carry this step out we define the partial multi-modes expansion U^ε_N as

$$U^\varepsilon_N:=\sum_{n=0}^{N}\varepsilon^n u_n,$$

where N is some positive integer. U^ε_N will also play a key role in our overall numerical approximation method. Since the full series u^ε cannot be computed, we approximate u^ε by its truncation U^ε_N.

Theorem 4.3.2. *Let $\{u_n\}$ be the same as in Theorem 4.3.1. Then (4.22) is valid in $L^2(\Omega,H^1(D))$ provided that $\sigma:=4\varepsilon C_0^{\frac{1}{2}}(1+k)<1$.*

Proof. The proof consists of two parts: (i) the infinite series on the right-hand side of (4.22) converges in $L^2(\Omega,H^1(D))$; (ii) the limit coincides with the solution u^ε. To

prove (i), note for any fixed positive integer p we have

$$U^{\varepsilon}_{N+p} - U^{\varepsilon}_{N} = \sum_{n=N}^{N+p-1} \varepsilon^n u_n.$$

It follows from the Cauchy-Schwarz inequality and (4.27) that for $j = 0, 1$

$$\begin{aligned}
\mathbb{E}\big(\|U^{\varepsilon}_{N+p} - U^{\varepsilon}_{N}\|^2_{H^j(D)}\big) &\le p \sum_{n=N}^{N+p-1} \varepsilon^{2n} \mathbb{E}(\|u_n\|^2_{H^j(D)}) \\
&\le p\Big(k^{j-1} + \frac{1}{k^2}\Big)^2 \mathbb{E}(\|f\|^2_{L^2(D)}) \sum_{n=N}^{N+p-1} \varepsilon^{2n} C(n,k) \\
&\le C_0 p\Big(k^{j-1} + \frac{1}{k^2}\Big)^2 \mathbb{E}(\|f\|^2_{L^2(D)}) \sum_{n=N}^{N+p-1} \sigma^{2n} \\
&\le C_0 p\Big(k^{j-1} + \frac{1}{k^2}\Big)^2 \mathbb{E}(\|f\|^2_{L^2(D)}) \cdot \frac{\sigma^{2N}\big(1-\sigma^{2p}\big)}{1-\sigma^2}.
\end{aligned}$$

Thus, if $\sigma < 1$ we have

$$\lim_{N\to\infty} \mathbb{E}\big(\|U^{\varepsilon}_{N+p} - U^{\varepsilon}_{N}\|^2_{H^1(D)}\big) = 0.$$

Therefore, $\{U^{\varepsilon}_N\}$ is a Cauchy sequence in $L^2(\Omega, H^1(D))$. Since $L^2(\Omega, H^1(D))$ is a Banach space, then there exists a function $U^{\varepsilon} \in L^2(\Omega, H^1(D))$ such that

$$\lim_{N\to\infty} U^{\varepsilon}_N = U^{\varepsilon} \qquad \text{in } L^2(\Omega, H^1(D)).$$

To show (ii), we first notice that by the definitions of u_n and U_N^ε,

$$
\begin{aligned}
&a(U_N^\varepsilon, v)\\
&\quad= \sum_{n=0}^{N-1} \int_\Omega \varepsilon^n \Big((\nabla u_n, \nabla v)_D - k^2(\alpha^2 u_n, v)_D + \mathbf{i}k\langle u_n, v\rangle_{\partial D} \Big)\, dP\\
&\quad= \sum_{n=0}^{N-1} \int_\Omega \varepsilon^n \Big((\nabla u_n, \nabla v)_D - k^2(u_n, v)_D + \mathbf{i}k\langle u_n, v\rangle_{\partial D} \Big)\, dP\\
&\qquad\quad - \sum_{n=0}^{N-1} \int_\Omega \varepsilon^n \Big(k^2\big((2\varepsilon\eta + \varepsilon^2\eta^2)u_n, v\big)_D - \mathbf{i}k\langle \varepsilon\eta u_n, v\rangle_{\partial D} \Big)\, dP\\
&\quad= \int_\Omega (f, v)_D\, dP + \sum_{n=1}^{N-1} \int_\Omega \varepsilon^n \Big(\big(2k^2\eta u_{n-1} + k^2\eta^2 u_{n-2}, v\big)_D - \mathbf{i}k\langle \eta u_{n-1}, v\rangle_{\partial D} \Big)\, dP\\
&\qquad\quad - \sum_{n=0}^{N-1} \int_\Omega \varepsilon^n \Big(k^2\big((2\varepsilon\eta + \varepsilon^2\eta^2)u_n, v\big)_D - \mathbf{i}k\langle \varepsilon\eta u_n, v\rangle_{\partial D} \Big)\, dP\\
&\quad= \int_\Omega (f, v)_D\, dP - k^2\varepsilon^N \int_\Omega \big(\eta(2+\varepsilon\eta)u_{N-1} + \eta^2 u_{N-2},\, v\big)_D\, dP\\
&\qquad\qquad + \mathbf{i}k\varepsilon^N \int_\Omega \langle \eta u_{N-1}, v\rangle_{\partial D}\, dP,
\end{aligned}
$$

Thus, U_N^ε satisfies

$$
\begin{aligned}
&\int_\Omega \Big(\big(\nabla U_N^\varepsilon, \nabla v\big)_D - k^2\big(\alpha^2 U_N^\varepsilon, v\big)_D + \mathbf{i}k\,\langle \alpha U_N^\varepsilon, v\rangle_{\partial D} \Big)\, dP \qquad (4.33)\\
&\qquad = \int_\Omega (f, v)_D\, dP - k^2\varepsilon^N \int_\Omega \big(\eta(2+\varepsilon\eta)u_{N-1} + \eta^2 u_{N-2},\, v\big)_D\, dP\\
&\qquad\qquad + \mathbf{i}k\varepsilon^N \int_\Omega \langle \eta u_{N-1}, v\rangle_{\partial D}\, dP
\end{aligned}
$$

for all $v \in L^2(\Omega, H^1(D))$. Where $\alpha = 1 + \varepsilon\eta$. In other words, U_N^ε solves the following Helmholtz problem:

$$
\begin{aligned}
-\Delta U_N^\varepsilon - k^2\alpha^2 U_N^\varepsilon &= f - k^2\varepsilon^N\big(\eta(2+\varepsilon\eta)u_{N-1} + \eta^2 u_{N-2}\big) && \text{in } D,\\
\partial_\nu U_N^\varepsilon + \mathbf{i}k\alpha U_N^\varepsilon &= -\mathbf{i}k\varepsilon^N \eta u_{N-1} && \text{on } \partial D.
\end{aligned}
$$

By (4.27) and the Cauchy-Schwarz inequality we have

$$
\begin{aligned}
&k^2\varepsilon^N \left| \int_\Omega \big(\eta(2+\varepsilon\eta)u_{N-1} + \eta^2 u_{N-2},\, v\big)_D \, dP \right| \\
&\quad \le 3k^2\varepsilon^N \Big(\big(\mathbb{E}(\|u_{N-1}\|^2_{L^2(D)})\big)^{\frac{1}{2}} + \big(\mathbb{E}(\|u_{N-2}\|^2_{L^2(D)})\big)^{\frac{1}{2}} \Big) \big(\mathbb{E}(\|v\|^2_{L^2(D)})\big)^{\frac{1}{2}} \\
&\quad \le 6k^2\varepsilon^N \Big(\frac{1}{k} + \frac{1}{k^2}\Big) C(N-1,k)^{\frac{1}{2}} \big(\mathbb{E}(\|f\|^2_{L^2(D)})\big)^{\frac{1}{2}} \big(\mathbb{E}(\|v\|^2_{L^2(D)})\big)^{\frac{1}{2}} \\
&\quad \le 3\varepsilon(k+1) C_0^{\frac{1}{2}} \sigma^{N-1} \big(\mathbb{E}(\|f\|^2_{L^2(D)})\big)^{\frac{1}{2}} \big(\mathbb{E}(\|v\|^2_{L^2(D)})\big)^{\frac{1}{2}} \\
&\quad \longrightarrow 0 \quad \text{as } N \to \infty \quad \text{provided that } \sigma < 1.
\end{aligned}
$$

Similarly, we get

$$
\begin{aligned}
&k\varepsilon^N \left| \int_\Omega \langle \eta u_{N-1}, v\rangle_{\partial D} \, dP \right| \\
&\quad \le k\varepsilon^N \big(\mathbb{E}(\|u_{N-1}\|^2_{L^2(\partial D)})\big)^{\frac{1}{2}} \big(\mathbb{E}(\|v\|^2_{L^2(\partial D)})\big)^{\frac{1}{2}} \\
&\quad \le k\varepsilon^N \Big(\frac{1}{k} + \frac{1}{k^2}\Big) C(N-1,k) \big(\mathbb{E}(\|f\|^2_{L^2(D)})\big)^{\frac{1}{2}} \big(\mathbb{E}(\|v\|^2_{L^2(\partial D)})\big)^{\frac{1}{2}} \\
&\quad \le \frac{\varepsilon}{2} \Big(1 + \frac{1}{k}\Big) C_0^{\frac{1}{2}} \sigma^{N-1} \big(\mathbb{E}(\|f\|^2_{L^2(D)})\big)^{\frac{1}{2}} \big(\mathbb{E}(\|v\|^2_{L^2(\partial D)})\big)^{\frac{1}{2}} \\
&\quad \longrightarrow 0 \quad \text{as } N \to \infty \quad \text{provided that } \sigma < 1.
\end{aligned}
$$

Setting $N \to \infty$ in (4.33) immediately yields

$$
\int_\Omega \Big(\big(\nabla U^\varepsilon, \nabla v\big)_D - k^2 \big(\alpha^2 U^\varepsilon, v\big)_D + \mathbf{i}k \, \langle \alpha U^\varepsilon, v\rangle_{\partial D} \Big) \, dP = \int_\Omega (f, v)_D \, dP, \tag{4.34}
$$

for all $v \in L^2(\Omega, H^1(D))$. Thus, U^ε is a solution to problem (4.1)–(4.2). By the uniqueness of the solution, we conclude that $U^\varepsilon = u^\varepsilon$. Therefore, (4.22) holds in $L^2(\Omega, H^1(D))$. The proof is complete. □

The above proof also infers an upper bound for the error $u^\varepsilon - U^\varepsilon_N$ as stated in the next theorem.

Theorem 4.3.3. *Let U_N^ε be the same as above and u^ε denote the solution to problem (4.1)–(4.2) and $\sigma := 4\varepsilon C_0^{\frac{1}{2}}(1+k)$. Then there holds for $\varepsilon(2\varepsilon+1) < \gamma_0$*

$$\mathbb{E}(\|u^\varepsilon - U_N^\varepsilon\|^2_{H^j(D)}) \leq \frac{9C_0\sigma^{2N}}{32(1+k)^2}\Big(k^j + \frac{1}{k}\Big)^4 \mathbb{E}(\|f\|^2_{L^2(D)}), \quad j = 0, 1, \tag{4.35}$$

provided that $\sigma < 1$. Where C_0 is a positive constant independent of k and ε.

Proof. Let $E_N^\varepsilon := u^\varepsilon - U_N^\varepsilon$, subtracting (4.33) from (4.34) we get

$$\begin{aligned}\int_\Omega \Big(\big(\nabla E_N^\varepsilon, \nabla v\big)_D - k^2\big(\alpha^2 E_N^\varepsilon, v\big)_D + \mathbf{i}k\,\langle \alpha E_N^\varepsilon, v\rangle_{\partial D}\Big)\, dP \\ = k^2\varepsilon^N \int_\Omega \big(\eta(2+\varepsilon\eta)u_{N-1} + \eta^2 u_{N-2},\, v\big)_D\, dP - \mathbf{i}k\varepsilon^N \int_\Omega \langle \eta u_{N-1}, v\rangle_{\partial D}\, dP,\end{aligned} \tag{4.36}$$

for all $v \in L^2(\Omega, H^1(D))$. In other words, E_N^ε solves the following Helmholtz problem:

$$\begin{aligned} -\Delta E_N^\varepsilon - k^2\alpha^2 E_N^\varepsilon &= k^2\varepsilon^N\big(\eta(2+\varepsilon\eta)u_{N-1} + \eta^2 u_{N-2}\big) && \text{in } D, \\ \partial_\nu E_N^\varepsilon + \mathbf{i}k\alpha E_N^\varepsilon &= -\mathbf{i}k\varepsilon^N \eta u_{N-1} && \text{on } \partial D. \end{aligned}$$

By Theorem 4.2.6 and (4.27) we obtain for $j = 0, 1$

$$\begin{aligned}\mathbb{E}(\|E_N^\varepsilon\|^2_{H^j(D)}) &\leq 18C_0\Big(k^{j-1} + \frac{1}{k^2}\Big)^2 \Big[k^4\varepsilon^{2N}\Big(\mathbb{E}(\|u_{N-1}\|^2_{L^2(D)}) + \mathbb{E}(\|u_{N-2}\|^2_{L^2(D)})\Big) \\ &\qquad + k^2\varepsilon^{2N}\mathbb{E}(\|u_{N-1}\|^2_{L^2(\partial D)})\Big] \\ &\leq 18C_0k^4\varepsilon^{2N}\Big(k^{j-1} + \frac{1}{k^2}\Big)^4 C(N-1,k)\,\mathbb{E}(\|f\|^2_{L^2(D)}) \\ &\leq \frac{18C_0\sigma^{2N}}{64(1+k)^2}\Big(k^j + \frac{1}{k}\Big)^4 \mathbb{E}(\|f\|^2_{L^2(D)}).\end{aligned}$$

The proof is complete. □

Remark 4.3.4. *Theorem 4.3.3 shows that the error introduced by truncating the multi-modes expansion is on the order of ε^N where N is the number of modes in the truncated multi-modes expansion. Since ε is small, U_N^ε can be used to approximate u^ε using only a few mode functions, i.e. N is relatively small.*

4.4 Monte Carlo Discontinuous Galerkin Approximation of the Truncated Multi-modes Expansion U_N^ε

In the previous section, we present a multi-modes representation of the solution u^ε and a convergence rate estimate for its truncated multi-modes approximation. These results will serve as the theoretical foundation for our overall numerical methodology for approximating the solution u^ε of problem (4.1)–(4.2).

As stated previously, we start by approximating u^ε through its truncated multi-modes expansion U_N^ε. Note that the linear nature of the expectation operator, along with the definition of U_N^ε yields the following expansion:

$$\mathbb{E}(U_N^\varepsilon) = \sum_{n=0}^{N-1} \varepsilon^n \mathbb{E}(u_n).$$

Hence, to gain an accurate approximation of $\mathbb{E}(U_N^\varepsilon)$ one only needs to seek an accurate approximation of $\mathbb{E}(u_n)$ for each mode function u_n. Observe that we can apply the expectation operator to (4.24) and (4.26) to find

$$-\Delta \mathbb{E}(u_0) - k^2 \mathbb{E}(u_0) = \mathbb{E}(f), \qquad \text{in } \Omega, \tag{4.37}$$

$$\frac{\partial}{\partial \nu}\mathbb{E}(u_0) + \mathbf{i}k\mathbb{E}(u_0) = 0, \qquad \text{on } \partial\Omega. \tag{4.38}$$

Therefore, for $\mathbb{E}(f)$ known, $\mathbb{E}(u_0)$ can be computed directly by solving a deterministic Helmholtz equation. On the other hand, for $n \geq 1$, we apply the same reasoning to (4.25) and (4.26) to find the following:

$$-\Delta \mathbb{E}(u_n) - k^2 \mathbb{E}(u_n) = 2k^2 \mathbb{E}(\eta u_{n-1}) + k^2 \mathbb{E}(\eta^2 u_{n-2}), \qquad \text{in } \Omega,$$

$$\frac{\partial}{\partial \nu}\mathbb{E}(u_n) + \mathbf{i}k\mathbb{E}(u_n) = -\mathbf{i}k\mathbb{E}(\eta u_{n-1}), \qquad \text{on } \partial\Omega.$$

We note the terms $\mathbb{E}(\eta u_{n-1})$ and $\mathbb{E}(\eta^2 u_{n-2})$ cannot be further broken apart due to the multiplicative nature of these terms and the fact that η and u_n are not independent. Thus for $n \geq 1$, $\mathbb{E}(u_n)$ cannot be computed directly in the same manner as $E(u_0)$.

The goal of this section is to develop a *Monte Carlo interior penalty discontinuous Galerkin* (MCIP-DG) method for the above mentioned Helmholtz problem. Our MCIP-DG method is the direct generalization of the deterministic IP-DG method proposed in [42, 44] for the related deterministic Helmholtz problem. It should be noted that although various numerical methods (such as finite difference, finite element and spectral methods) can be used for the job, the IP-DG method presented below is the only general purpose discretization method which is absolutely stable (i.e., stable without mesh constraint) and optimally convergent. This is indeed the primary reason why we choose this IP-DG method as our spatial discretization method.

4.4.1 DG Notations

To define the IP-DG method used in this chapter, we must introduce some standard DG notation. This notation was first introduced in Chapter 3. Let $\mathcal{T}_h$ be a quasi-uniform partition of D such that $\overline{D} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$. Let h_K denote the diameter of $K \in \mathcal{T}_h$ and $h := \max\{h_K; K \in \mathcal{T}_h\}$. $H^s(\mathcal{T}_h)$ denotes the standard broken Sobolev space and V_r^h denotes the DG finite element space which are defined as

$$H^s(\mathcal{T}_h) := \prod_{K\in\mathcal{T}_h} H^s(K), \qquad V_r^h := \prod_{K\in\mathcal{T}_h} P_r(K),$$

where $P_r(K)$ is the set of all polynomials whose degrees do not exceed a given positive integer r. Let $\mathcal{E}^I$ denote the set of all interior faces/edges of $\mathcal{T}_h$, $\mathcal{E}^B$ denote the set of all boundary faces/edges of $\mathcal{T}_h$, and $\mathcal{E} := \mathcal{E}^I \cup \mathcal{E}^B$. The L^2-inner product for piecewise functions over the mesh $\mathcal{T}_h$ is naturally defined by

$$(v, w)_{\mathcal{T}_h} := \sum_{K\in\mathcal{T}_h} \int_K v \cdot \overline{w}\, dx,$$

and for any set $\mathcal{S}_h \subset \mathcal{E}$, the L^2-inner product over $\mathcal{S}_h$ is defined by

$$\langle v, w \rangle_{\mathcal{S}_h} := \sum_{e \in \mathcal{S}_h} \int_e v \cdot \overline{w}\, dS.$$

Let $K, K' \in \mathcal{T}_h$ and $e = \partial K \cap \partial K'$ and assume global labeling number of K is bigger than that of K'. We choose $n_e := n_K|_e = -n_{K'}|_e$ as the unit normal on e outward to K and define the following standard jump and average notations across the face/edge e:

$$\begin{aligned} [v] &:= v|_K - v|_{K'} \qquad \text{on } e \in \mathcal{E}^I, \qquad [v] := v \quad \text{on } e \in \mathcal{E}^B, \\ \{v\} &:= \frac{1}{2}\big(v|_K + v|_{K'}\big) \quad \text{on } e \in \mathcal{E}^I, \qquad \{v\} := v \quad \text{on } e \in \mathcal{E}^B \end{aligned}$$

for $v \in V_r^h$. We also define the following semi-norms on $H^s(\mathcal{T}_h)$:

$$\begin{aligned} |v|_{1,h,D} &:= \|\nabla v\|_{L^2(\mathcal{T}_h)}, \\ \|v\|_{1,h,D} &:= \left(|v|_{1,h,D}^2 + \sum_{e \in \mathcal{E}_h^I} \left(\frac{\gamma_{0,e}\, r}{h_e} \|[v]\|_{L^2(e)}^2 + \sum_{\ell=1}^{d-1} \frac{\beta_{1,e} r}{h_e} \left\| [\partial_{\tau_e^\ell} v] \right\|_{L^2(e)}^2 \right) \right. \\ &\qquad \left. + \sum_{j=1}^{r} \sum_{e \in \mathcal{E}_h^I} \gamma_{j,e} \Big(\frac{h_e}{r} \Big)^{2j-1} \left\| [\partial_{n_e}^j v] \right\|_{L^2(e)}^2 \right)^{\frac{1}{2}}, \\ |||v|||_{1,h,D} &:= \left(\|v\|_{1,h,D}^2 + \sum_{e \in \mathcal{E}_h^I} \frac{h_e}{\gamma_{0,e} r} \|\{\partial_{n_e} v\}\|_{L^2(e)}^2 \right)^{\frac{1}{2}}. \end{aligned}$$

4.4.2 IP-DG Method for Deterministic Helmholtz Problem

In this subsection, we consider following deterministic Helmholtz problem and its IP-DG approximations proposed in [42, 44].

$$-\Delta \Phi_0 - k^2 \Phi_0 = F_0 \qquad \text{in } D, \tag{4.39}$$

$$\partial_\nu \Phi_0 + \mathbf{i} k \Phi_0 = G_0 \qquad \text{on } \partial D. \tag{4.40}$$

Recall that $\Phi_0 = \mathbb{E}(u_0)$ satisfies the above equations with $F_0 = \mathbb{E}(f)$ and $G_0 = 0$. As an interesting byproduct, all the results to be presented in this subsection apply to $\mathbb{E}(u_0)$.

The IP-DG weak formulation for (4.39)–(4.40) is defined by (cf. [42, 44]) seeking $\Phi_0 \in H^1(D) \cap H^{r+1}_{\text{loc}}(D)$ such that

$$a_h(\Phi_0, \psi) = (F_0, \psi)_D + \langle G_0, \psi \rangle_{\partial D} \qquad \forall \psi \in H^1(D) \cap H^{r+1}(\mathcal{T}_h), \tag{4.41}$$

where

$$\begin{aligned}
a_h(\phi, \psi) &:= b_h(\phi, \psi) - k^2(\phi, \psi)_{\mathcal{T}_h} + \mathbf{i}k\langle \phi, \psi \rangle_{\mathcal{E}_h^B} + \mathbf{i}\Big(L_1(\phi, \psi) + \sum_{j=0}^{r} J_j(\phi, \psi)\Big), \qquad (4.42)\\
b_h(\phi, \psi) &:= (\nabla\phi, \nabla\psi)_{\mathcal{T}_h} - \Big(\langle \{\partial_n \phi\}, [\psi] \rangle_{\mathcal{E}_h^I} + \langle [\phi], \{\partial_n \psi\} \rangle_{\mathcal{E}_h^I}\Big),\\
L_1(\phi, \psi) &:= \sum_{e \in \mathcal{E}_h^I} \sum_{\ell=1}^{d-1} \beta_{1,e} h_e^{-1} \left\langle [\partial_{\tau^\ell} \phi], [\partial_{\tau^\ell} \psi] \right\rangle_e,\\
J_j(\phi, \psi) &:= \sum_{e \in \mathcal{E}_h^I} \gamma_{j,e} h_e^{2j-1} \left\langle [\partial_n^j \phi], [\partial_n^j \psi] \right\rangle_e, \qquad j = 0, 1, \cdots, r.
\end{aligned}$$

$\{\beta_{1,e}\}$ and $\{\gamma_{j,e}\}$ are piecewise constant nonnegative functions defined on $\mathcal{E}_h^I$. $\{\tau^\ell\}_{\ell=1}^{d-1}$ denotes an orthonormal basis of the edge and ∂_{τ^ℓ} denotes the tangential derivative in the direction of τ^ℓ.

Remark 4.4.1. *L_1 and $\{J_j\}$ terms are called interior penalty terms, $\{\beta_{1,e}\}$ and $\{\gamma_{j,e}\}$ are called penalty parameters. The two distinct features of the DG sesquilinear form $a_h(\cdot, \cdot)$ are: (i) it penalizes not only the jumps of the function values but also penalizes the jumps of the tangential derivatives as well the jumps of all normal derivatives up to rth order; (ii) the penalty parameters are purely imaginary numbers with nonnegative imaginary parts.*

Following [42, 44] and based on the DG weak formulation (4.41), our IP-DG method for problem (4.39)–(4.40) is defined by seeking $\Phi_0^h \in V_r^h$ such that

$$a_h(\Phi_0^h, \psi^h) = (F_0, \psi^h)_D + \langle G_0, \psi^h \rangle_{\partial D} \qquad \forall \psi^h \in V_r^h. \tag{4.43}$$

For the above IP-DG method, it was proved in [42, 44] that the method is absolutely stable and its solutions satisfy some wave-number explicit stability estimates. Its solutions also satisfy optimal order (in h) error estimates, which are described below.

Theorem 4.4.2. *Let $\Phi_0^h \in V_r^h$ be a solution to scheme (4.43), then there hold*

(i) For all $h, k > 0$, there exists a positive constant $\hat{C}_0$ independent of k and h such that

$$\|\Phi_0^h\|_{L^2(D)} + \frac{1}{k}\left\|\Phi_0^h\right\|_{1,h,D} + \|\Phi_0^h\|_{L^2(\partial D)} \le \hat{C}_0 C_s \hat{M}(F_0, G_0), \tag{4.44}$$

where

$$C_s := \frac{d-2}{k} + \frac{1}{k^2} + \frac{1}{k^2}\max_{e\in\mathcal{E}_h^I}\Big(\frac{r\,k^2 h_e^2 + r^5}{\gamma_{0,e}\, h_e^2} + \frac{r}{h_e}\max_{0\le j\le r-1}\sqrt{\frac{\gamma_{j,e}}{\gamma_{j+1,e}}} + \frac{r^2}{h_e} + \frac{r^3}{h_e^2}\sqrt{\frac{\beta_{1,e}}{\gamma_{1,e}}}\Big), \tag{4.45}$$

$$\hat{M}(F_0, G_0) := \|F_0\|_{L^2(D)} + \|G_0\|_{L^2(\partial D)}. \tag{4.46}$$

(ii) If $k^3 h^2 r^{-2} = O(1)$, then there exists a positive constant $\hat{C}_0$ independent of k and h such that

$$\|\Phi_0^h\|_{L^2(D)} + \|\Phi_0^h\|_{L^2(\partial D)} + \frac{1}{k}\|\Phi_0^h\|_{1,h,D} \le \hat{C}_0\Big(\frac{1}{k} + \frac{1}{k^2}\Big)\hat{M}(F_0, G_0). \tag{4.47}$$

An immediate consequence of (4.44) is the following unconditional solvability and uniqueness result.

Corollary 4.4.3. *There exists a unique solution to scheme (4.43) for all $k, h > 0$.*

Theorem 4.4.4. *Let $\Phi_0^h \in V^h$ solve (4.43), $\Phi_0 \in H^s(\Omega)$ be the solution of (4.39)–(4.40), and $\mu = \min\{r+1, s\}$. Suppose $\gamma_{j,e}, \beta_{1,e} > 0$. Let $\gamma_j = \max_{e\in\mathcal{E}^I} \gamma_{j,e}$ and $\lambda = 1 + \frac{1}{\gamma_0}$.*

(i) For all $h, k > 0$, there exists a positive constant $\tilde{C}_0$ independent of k and h such that

$$\|\Phi_0 - \Phi_0^h\|_{1,h,D} \le \tilde{C}_0\Big(C_r + \frac{k^3 h}{r}\, C_s \hat{C}_r\Big)\, \frac{h^{\mu-1}}{r^{s-1}} \|\Phi_0\|_{H^s(D)}, \tag{4.48}$$

$$\|\Phi_0 - \Phi_0^h\|_{L^2(D)} + \|\Phi_0 - \Phi_0^h\|_{L^2(\partial D)} \le \tilde{C}_0 \hat{C}_r \left(1 + k^2 C_s\right) \frac{h^\mu}{r^s} \|\Phi_0\|_{H^s(D)}, \tag{4.49}$$

where

$$C_r := \lambda\Big(1 + \frac{r}{\gamma_0} + \sum_{j=1}^{r} r^{2j-1}\gamma_j + \frac{kh}{\lambda r}\Big)^{\frac{1}{2}},$$

$$\hat{C}_r := \Big(1 + \frac{r}{\gamma_0} + r\,\gamma_1 + \sum_{j=2}^{r} r^{2j-2}\gamma_j + \frac{kh}{\lambda r}\Big)^{\frac{1}{2}} C_r.$$

(ii) If $k^3 h^2 r^{-2} = O(1)$, then there exists a positive constant $\tilde{C}_0$ independent of k and h such that

$$\big\|\Phi_0 - \Phi_0^h\big\|_{1,h,D} \le \frac{\tilde{C}_0 (r + k^2 h) h^{\mu-1}}{r^s}\, \|\Phi_0\|_{H^s(D)}\,, \tag{4.50}$$

$$\big\|\Phi_0 - \Phi_0^h\big\|_{L^2(D)} + \big\|\Phi_0 - \Phi_0^h\big\|_{L^2(\partial D)} \le \frac{\tilde{C}_0 k h^\mu}{r^s}\, \|\Phi_0\|_{H^s(D)}\,. \tag{4.51}$$

Remark 4.4.5. *It was proved in [27] (also by Theorem 4.2.6 with $\varepsilon = 0$) that*

$$\|\Phi_0\|_{H^s(D)} \le \tilde{C}_0\Big(k^{s-1} + \frac{1}{k}\Big)\hat{M}(F_0, G_0), \qquad s = 0, 1, 2.$$

It is expected that the following higher order norm estimates also hold (cf. [42] for an explanation):

$$\|\Phi_0\|_{H^s(D)} \leq \tilde{C}_0\Big(k^{s-1}+\frac{1}{k}\Big)\Big(\|F_0\|_{H^{s-2}(D)}+\|G_0\|_{H^{s-\frac{5}{2}}(\partial D))}\Big), \qquad s\geq 3 \tag{4.52}$$

provided that F_0, G_0 and D are sufficiently smooth. In such a case, $\|\Phi_0\|_{H^s(D)}$ in (4.48)–(4.51) can be replaced by the above bound so explicit constants can be obtained in these estimates.

Theorems 4.4.2 and 4.4.4 give stability and optimal error estimates (in h) for the IP-DG method in both the pre-asymptotic and asymptotic regime. Here the asymptotic regime is characterized as k, h, r chosen to satisfy the constraint $k^3h^2r^{-2} = O(1)$. In the remainder of this chapter we only consider the asymptotic regime of $k^3h^2r^{-2} = O(1)$. This choice was made because the constants for the asymptotic regime in both Theorem 4.4.2 and 4.4.4 are more tractable.

4.4.3 MCIP-DG Method for Approximating $\mathbb{E}(\mathbf{U}_\mathbf{n}^\varepsilon)$

We recall that each mode function u_n satisfies the following Helmholtz problem:

$$-\Delta u_n - k^2 u_n = S_n \qquad \text{in } D, \tag{4.53}$$

$$\partial_\nu u_n + \mathbf{i}k u_n = Q_n \qquad \text{on } \partial D, \tag{4.54}$$

where $u_{-1} := 0$ and

$$S_0 := f, \qquad Q_0 := 0, \tag{4.55}$$

$$S_n := 2k^2\eta u_{n-1} + k^2\eta^2 u_{n-2}, \qquad Q_n := -\mathbf{i}k\eta u_{n-1}, \tag{4.56}$$

for $n \geq 1$. Clearly, $S_n(\cdot, x)$ and $Q_n(\cdot, x)$ are random variables for *a.e.* $x \in D$, $S_n \in L^2(\Omega, L^2(D))$ and $Q_n \in L^2(\Omega, L^2(\partial D))$. We remark again that due to its multiplicative structure $\mathbb{E}(S_n)$ and $\mathbb{E}(Q_n)$ can not be computed directly for $n \geq 1$.

Otherwise, (4.53) and (4.54) would be easily converted into deterministic equations for $\mathbb{E}(u_n)$, as we did early for $\mathbb{E}(u_0)$. In other words, (4.53)–(4.54) is a genuine random PDE problem. On the other hand, since all the coefficients of the equations are constants, then the problem is nearly deterministic. Such a remarkable property will be fully exploited in our overall numerical methodology which will be described in the next section.

Several numerical methodologies are well known in the literature for discretizing random PDEs, Monte Carlo Galerkin and stochastic Galerkin (or polynomial chaos) methods and stochastic collocation methods are three of well-known methods (cf. [12, 11] and the references therein). Due to the nearly deterministic structure of (4.53)–(4.54), we propose to discretize it using the Monte Carlo IP-DG approach which combines the classical Monte Carlo method for stochastic variable and the IP-DG method, which is presented in the proceeding subsection, for the spatial variable.

Following the standard formulation of the Monte Carlo method (cf. [12]), let M be a (large) positive integer which will be used to denote the number of realizations and V_r^h be the DG space defined in Section 4.4.1. For each $j = 1, 2, \cdots, M$, we sample i.i.d. realizations of the source term $f(\omega_j, \cdot)$ and random medium coefficient $\eta(\omega_j, \cdot)$, and recursively find corresponding approximation $u_n^h(\omega_j, \cdot) \in V_r^h$ such that

$$a_h\big(u_n^h(\omega_j,\cdot),\psi^h\big) = \big(S_n^h(\omega_j,\cdot),\psi^h\big)_D + \langle Q_n^h(\omega_j,\cdot),\psi^h\rangle_{\partial D} \qquad \forall \psi^h \in V_r^h \tag{4.57}$$

for $n = 0, 1, 2, \cdots, N-1$. Where

$$S_0^h(\omega_j,\cdot) := f(\omega_j,\cdot), \quad Q_0^h := 0, \tag{4.58}$$

$$u_{-1}^h(\omega_j,\cdot) := 0, \tag{4.59}$$

$$S_n^h(\omega_j,\cdot) := 2k^2\eta u_{n-1}^h(\omega_j,\cdot) + k^2\eta^2 u_{n-2}^h(\omega_j,\cdot), \qquad n = 1, 2, \cdots, N-1, \tag{4.60}$$

$$Q_n^h(\omega_j,\cdot) := -\mathbf{i}k\eta u_{n-1}^h(\omega_j,\cdot), \qquad n = 1, 2, \cdots, N-1. \tag{4.61}$$

We point out that in order for u_n^h to be computable, S_n^h and Q_n^h, not S_n and Q_n, are used on the right-hand side of (4.57). This (small) perturbation on the right-hand side will result in an additional discretization error which must be accounted for later.

We approximate $\mathbb{E}(u_n)$ by the following sample average

$$\Phi_n^h := \frac{1}{M}\sum_{j=1}^{M} u_n^h(\omega_j, \cdot). \tag{4.62}$$

Thus, $\mathbb{E}(U_N^\varepsilon)$ can be approximated by

$$\Psi_N^h = \sum_{n=0}^{N-1} \varepsilon^n \Phi_n^h. \tag{4.63}$$

The rest of this section is used to analyze the error generated by this MCIP-DG method. This will be carried out in the following two steps: (i) We estimate the error from the IP-DG approximation, i.e. $U_N^\varepsilon - U_N^h$, where $U_N^h = \sum_{n=0}^{N-1} \varepsilon^n u_n^h$. (ii) We estimate the error from the Monte Carlo method, i.e. $\mathbb{E}(U_N^h) - \Psi_N^h$.

We begin by obtaining stability estimates for each u_n^h.

Lemma 4.4.6. *Assume $k^3h^2r^{-2} = O(1)$. Then for each $n \geq 0$ the following stability estimate holds:*

$$\mathbb{E}\big(\|u_n^h\|_{L^2(D)}^2 + \|u_n^h\|_{L^2(\partial D)}^2\big) \leq \Big(\frac{1}{k} + \frac{1}{k^2}\Big)^2 \hat{C}(n,k)\, \mathbb{E}(\|f\|_{L^2(D)}^2), \tag{4.64}$$

$$\mathbb{E}(\|u_n^h\|_{1,h,D}^2) \leq \Big(1 + \frac{1}{k}\Big)^2 \hat{C}(n,k)\, \mathbb{E}(\|f\|_{L^2(D)}^2), \tag{4.65}$$

where

$$\hat{C}(0,k) := \hat{C}_0^2, \quad \hat{C}(n,k) := 4^{2n-1}\hat{C}_0^{2n+2}(1+k)^{2n} \quad \textit{for } n \geq 1. \tag{4.66}$$

Proof. Using estimate (4.47) one immediately obtains

$$\begin{aligned}
&\mathbb{E}\big(\|u_0^h\|^2_{L^2(D)} + \|u_0^h\|^2_{L^2(\partial D)}\big) \\
&\qquad \leq \hat{C}_0^2\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2\mathbb{E}(\|S_0^h\|^2_{L^2(D)}) \leq \hat{C}_0^2\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2\mathbb{E}(\|f\|^2_{L^2(D)}), \\
&\mathbb{E}\big(\|u_0^h\|^2_{1,h,D}\big) \leq \hat{C}_0^2\Big(1+\frac{1}{k}\Big)^2\mathbb{E}(\|S_0^h\|^2_{L^2(D)}) \leq \hat{C}_0^2\Big(1+\frac{1}{k}\Big)^2\mathbb{E}(\|f\|^2_{L^2(D)}),
\end{aligned}$$

which verifies (4.64) and (4.65) for $n = 0$. Suppose (4.64) and (4.65) hold for all $n = 0, 1, 2, \cdots, \ell-1$. It remains to show (4.64) and (4.65) hold for $n = \ell$.

Using (4.64) with $n = \ell - 1$ and steps that were used previously in the proof of Theorem 4.3.1 one obtains the following:

$$\begin{aligned}
&\mathbb{E}\big(\|u_\ell^h\|^2_{L^2(D)} + \|u_\ell^h\|^2_{L^2(\partial D)}\big) \\
&\quad \leq \hat{C}_0^2\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2\mathbb{E}\Big(\|S_\ell^h\|^2_{L^2(D)} + \|Q_n^h\|^2_{L^2(\partial D)}\Big) \\
&\quad \leq 2\hat{C}_0^2\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2 k^4 E\Big(4\|u_{\ell-1}^h\|^2_{L^2(D)} + \|u_{\ell-2}^h\|^2_{L^2(D)} + \frac{1}{k^2}\|u_{\ell-1}^h\|^2_{L^2(\partial D)}\Big) \\
&\quad \leq 2\hat{C}_0^2\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2 (1+k)^2\Big(4\hat{C}(\ell-1,k) + \hat{C}(\ell-2,k)\Big)\mathbb{E}(\|f\|^2_{L^2(D)}) \\
&\quad \leq 8\hat{C}_0^2\Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2 (1+k)^2\hat{C}(\ell-1,k)\left(1+\frac{\hat{C}(\ell-2,k)}{4\hat{C}(\ell-1,k)}\right)\mathbb{E}(\|f\|^2_{L^2(D)}) \\
&\quad \leq \Big(\frac{1}{k}+\frac{1}{k^2}\Big)^2\hat{C}(\ell,k)\mathbb{E}(\|f\|^2_{L^2(D)}).
\end{aligned}$$

Here we have used the fact that

$$8\hat{C}_0^2(1+k)^2\hat{C}(\ell-1,k)\left(1+\frac{\hat{C}(\ell-2,k)}{4\hat{C}(\ell-1,k)}\right) \leq \hat{C}(\ell,k).$$

Thus (4.64) is proved for $n = \ell$.

Using (4.65) along with a similar argument the following is found:

$$\begin{aligned}
\mathbb{E}(\|u_\ell^h\|_{1,h,D}^2) &\le \hat{C}_0^2\Big(1+\frac{1}{k}\Big)^2\mathbb{E}\Big(\|S_\ell^h\|_{L^2(D)}^2+\|Q_n^h\|_{L^2(\partial D)}^2\Big)\\
&\le 2\hat{C}_0^2\Big(1+\frac{1}{k}\Big)^2k^4E\Big(4\|u_{\ell-1}^h\|_{L^2(D)}^2+\mathbb{E}(\|u_{\ell-2}^h\|_{L^2(D)}^2+\frac{1}{k^2}\|u_{\ell-1}^h\|_{L^2(\partial D)}^2\Big)\\
&\le 2\hat{C}_0^2\Big(1+\frac{1}{k}\Big)^2(1+k)^2\Big(4\hat{C}(\ell-1,k)+\hat{C}(\ell-2,k)\Big)\mathbb{E}(\|f\|_{L^2(D)}^2)\\
&\le \Big(1+\frac{1}{k}\Big)^2\hat{C}(\ell,k)\mathbb{E}(\|f\|_{L^2(D)}^2).
\end{aligned}$$

Hence (4.65) holds for $n=\ell$. Therefore, the induction argument is complete. □

To complete step (i), we need a set of auxiliary mode functions $\{\tilde{u}_n^h\}_{n\ge0}$. For fixed realizations $\eta(\omega_j,\cdot)$ and $f(\omega_j,\cdot)$, we define $\tilde{u}_n^h(\omega_j,\cdot)\in V_r^h$ as the solution to the following problem:

$$a_h\big(\tilde{u}(\omega_j,\cdot)_n^h,\psi^h\big)=\big(S_n(\omega_j,\cdot),\psi^h\big)_D+\langle Q_n(\omega_j,\cdot),\psi^h\rangle_{\partial D}\qquad\forall\psi^h\in V_r^h. \tag{4.67}$$

$S_n(\omega_j,\cdot)$ and $Q_n(\omega_j,\cdot)$ were defined in (4.55) and (4.56) and are different from S_n^h and Q_n^h that are used in the definition of u_n^h.

We also need the following lemma.

Lemma 4.4.7. *Let $\gamma,\beta>0$ be two real numbers, $\{c_n\}_{n\ge0}$ and $\{\alpha_n\}_{n\ge0}$ be two sequences of nonnegative numbers such that*

$$c_0\le\gamma\alpha_0,\quad c_n\le\beta c_{n-1}+\gamma\alpha_n\quad \textit{for } n\ge1. \tag{4.68}$$

Then there holds

$$c_n\le\gamma\sum_{j=0}^{n}\beta^{n-j}\alpha_j\qquad \textit{for } n\ge1. \tag{4.69}$$

The proof to this lemma is trivial and thus is omitted.

Now, we are ready to estimate the error $u_n-u_n^h$.

Lemma 4.4.8. *Suppose $k^3h^2r^{-2} = O(1)$. Then the following error estimates hold:*

$$\mathbb{E}\big(\|u_n - u_n^h\|_{L^2(D)} + \|u_n - u_n^h\|_{L^2(\partial D)}\big) \leq \frac{\tilde{C}_0 k h^\mu}{r^s} \sum_{j=0}^{n} \big[\tilde{C}_0(2k+3)\big]^{n-j} \mathbb{E}\big(\|u_j\|_{H^s(D)}\big), \tag{4.70}$$

$$\mathbb{E}\big(\|u_n - u_n^h\|_{1,h,D}\big) \leq \frac{C\tilde{C}_0^2 k(1+k)h^{\mu-1}}{r^s} \sum_{j=0}^{n} \big[\tilde{C}_0(2k+3)\big]^{n-j} \mathbb{E}\big(\|u_j\|_{H^s(D)}\big), \tag{4.71}$$

where $\mu = \min\{r+1, s\}$.

Proof. To begin, we introduce the following error decomposition:

$$u_n - u_n^h = \big(u_n - \tilde{u}_n^h\big) + \big(\tilde{u}_n^h - u_n^h\big).$$

Thus, we get estimates on the error $u_n - u_n^h$ by first estimating $u_n - \tilde{u}_n^h$ and then estimating $\tilde{u}_n^h - u_n^h$. As an immediate consequence of Theorem 4.4.4 part (ii), for $k^3h^2r^{-2} = O(1)$ the following estimates hold:

$$\mathbb{E}\big(\big\|u_n - \tilde{u}_n^h\big\|_{1,h,D}\big) \leq \frac{\tilde{C}_0(r + k^2 h)h^{\mu-1}}{r^s} \mathbb{E}\big(\|u_n\|_{H^s(D)}\big), \tag{4.72}$$

$$\mathbb{E}\big(\big\|u_n - \tilde{u}_n^h\big\|_{L^2(D)} + \big\|u_n - \tilde{u}_n^h\big\|_{L^2(\partial D)}\big) \leq \frac{\tilde{C}_0 k h^\mu}{r^s} \mathbb{E}\big(\|u_n\|_{H^s(D)}\big). \tag{4.73}$$

To bound $\tilde{u}_n^h - u_n^h$, we observe that subtracting (4.57) from (4.67) yields

$$a_h\big(\tilde{u}_n^h - u_n^h, \psi^h\big) = \big(S_n - S_n^h, \psi^h\big)_D + \langle Q_n - Q_n^h, \psi^h\rangle_{\partial D} \qquad \forall \psi^h \in V_r^h, \text{ a.s.}$$

We apply Theorem 4.4.2 (ii) to obtain the following estimate:

$$\begin{aligned}
&\mathbb{E}\Big(k\|\tilde{u}_n^h - u_n^h\|_{L^2(D)} + k\|\tilde{u}_n^h - u_n^h\|_{L^2(\partial D)} + \|\tilde{u}_n^h - u_n^h\|_{1,h,D}\Big) \qquad (4.74)\\
&\quad \le \hat{C}_0\Big(1+\frac{1}{k}\Big)\mathbb{E}\Big(\|S_n - S_n^h\|_{L^2(D)} + \|Q_n - Q_n^h\|_{L^2(\partial D)}\Big)\\
&\quad \le 2\tilde{C}_0 k(k+1)E\Big(\|u_{n-1} - u_{n-1}^h\|_{L^2(D)} + \|u_{n-2} - u_{n-2}^h\|_{L^2(D)}\\
&\qquad + \|u_{n-1} - u_{n-1}^h\|_{L^2(\partial D)}\Big).
\end{aligned}$$

Combining (4.73) and (4.74) and applying the triangle inequality, we get

$$\begin{aligned}
&\mathbb{E}\big(\|u_n - u_n^h\|_{L^2(D)} + \|u_n - u_n^h\|_{L^2(\partial D)}\big) \qquad (4.75)\\
&\quad \le \mathbb{E}\Big(\|\tilde{u}_n^h - u_n^h\|_{L^2(D)} + \|\tilde{u}_n^h - u_n^h\|_{L^2(\partial D)} + \|u_n - \tilde{u}_n^h\|_{L^2(D)}\\
&\qquad + \|u_n - \tilde{u}_n^h\|_{L^2(\partial D)}\Big)\\
&\quad \le 2\tilde{C}_0(k+1)E\Big(\|u_{n-1} - u_{n-1}^h\|_{L^2(D)} + \|u_{n-2} - u_{n-2}^h\|_{L^2(D)}\\
&\qquad + \|u_{n-1} - u_{n-1}^h\|_{L^2(\partial D)}\Big) + \frac{\tilde{C}_0 k h^\mu}{r^s}\mathbb{E}\big(\|u_n\|_{H^s(D)}\big).
\end{aligned}$$

To estimate the error in $\|\cdot\|_{1,h,D}$, we apply an inverse inequality along with (4.73), (4.74), and the triangle inequality to get

$$\begin{aligned}
&\mathbb{E}\big(\|u_n - u_n^h\|_{1,h,D}\big) \le \mathbb{E}\big(\|\tilde{u}_n^h - u_n^h\|_{1,h,D} + \|u_n - \tilde{u}_n^h\|_{1,h,D}\big) \qquad (4.76)\\
&\quad \le Ch^{-1}\mathbb{E}\big(\|\tilde{u}_n^h - u_n^h\|_{L^2(D)}\big) + \mathbb{E}\big(\|u_n - \tilde{u}_n^h\|_{1,h,D}\big)\\
&\quad \le C\tilde{C}_0 h^{-1}(k+1)E\Big(2\|u_{n-1} - u_{n-1}^h\|_{L^2(D)} + \|u_{n-2} - u_{n-2}^h\|_{L^2(D)}\\
&\qquad + \frac{1}{k}\|u_{n-1} - u_{n-1}^h\|_{L^2(\partial D)}\Big) + \frac{\tilde{C}_0(r + k^2 h)h^{\mu-1}}{r^s}\,\mathbb{E}\big(\|u_n\|_{H^s(D)}\big).
\end{aligned}$$

Thus, (4.75) and (4.76) give recursive estimates for the error $u_n - u_n^h$. Next, we note the following estimates:

$$\mathbb{E}\big(\|u_{-1} - u_{-1}^h\|_{L^2(D)}\big) = \mathbb{E}\big(\|u_{-1} - u_{-1}^h\|_{1,h,D}\big) = 0, \tag{4.77}$$

$$\mathbb{E}\big(\|u_0 - u_0^h\|_{L^2(D)} + \|u_0 - u_0^h\|_{L^2(\partial D)}\big) \le \frac{\tilde{C}_0 k h^\mu}{r^s}\mathbb{E}\big(\|u_0\|_{H^s(D)}\big), \tag{4.78}$$

$$\mathbb{E}\big(\|u_0 - u_0^h\|_{1,h,D}\big) \le \frac{\tilde{C}_0(r + k^2 h)h^{\mu-1}}{r^s}\,\mathbb{E}\big(\|u_0\|_{H^s(D)}\big), \tag{4.79}$$

and define

$$\begin{aligned}
&u_{-2} = u_{-1} = u_{-2}^h = u_{-1}^h = 0,\\
&c_n := \mathbb{E}\big(\|u_n - u_n^h\|_{L^2(D)} + \|u_{n-1} - u_{n-1}^h\|_{L^2(D)}\big)\\
&\qquad\quad + \mathbb{E}\big(\|u_n - u_n^h\|_{L^2(\partial D)} + \|u_{n-1} - u_{n-1}^h\|_{L^2(\partial D)}\big),\\
&\beta := \tilde{C}_0(2k+3), \quad \gamma := \frac{\tilde{C}_0 k h^\mu}{r^s}, \quad \alpha_n := \mathbb{E}\big(\|u_n\|_{H^s(D)}\big).
\end{aligned}$$

Then by (4.76) these defined quantities meet the assumptions in Lemma 4.4.7. Applying Lemma 4.4.7 yields (4.70). Now (4.76) and (4.70) can be combined to produce (4.71). □

Now Lemma 4.4.8 can be used to bound the error due to IP-DG discretization, i.e. $U_N^\varepsilon - U_N^h$.

Theorem 4.4.9. *Assume that $u_n \in L^2(\Omega, H^s(D))$ for $n \ge 0$. Then the spatial error $U_N^\varepsilon - U_N^h$ satisfies the following estimates:*

$$\mathbb{E}\big(\|U_N^\varepsilon - U_N^h\|_{L^2(D)}\big) \le \frac{\tilde{C}_0 k h^\mu}{r^s}\sum_{n=0}^{N-1}\sum_{j=0}^{n}\varepsilon^n\big[\tilde{C}_0(2k+3)\big]^{n-j}\mathbb{E}\big(\|u_j\|_{H^s(D)}\big). \tag{4.80}$$

$$\begin{aligned}
&\mathbb{E}\big(\|U_N^\varepsilon - U_N^h\|_{1,h,D}\big) \\
&\qquad \le \frac{C\tilde{C}_0^2 k(1+k)h^{\mu-1}}{r^s}\sum_{n=0}^{N-1}\sum_{j=0}^{n}\varepsilon^n\big[\tilde{C}_0(2k+3)\big]^{n-j}\mathbb{E}\big(\|u_j\|_{H^s(D)}\big).
\end{aligned} \tag{4.81}$$

To simplify the above spatial error estimates, bounds for $\mathbb{E}(\|u_n\|_{H^s(D)})$ in terms of higher order norms of f are necessary. Only the case $s = 2$ is considered below for simplicity. When $s = 2$, the required estimates have been obtained in (4.30). These estimates in conjunction with Theorem 4.4.9 yield the following results:

Theorem 4.4.10. *Assume that $u_n \in L^2(\Omega, H^2(D))$ for $n \geq 0$. Then the following estimates hold:*

$$\mathbb{E}\big(\|U_N^\varepsilon - U_N^h\|_{L^2(D)}\big) \leq C_3(N,k,\varepsilon)\, h^2 \|f\|_{L^2(\Omega,L^2(D))}, \tag{4.82}$$

$$\mathbb{E}\big(\|U_N^\varepsilon - U_N^h\|_{1,h,D}\big) \leq C_4(N,k,\varepsilon)\, h \|f\|_{L^2(\Omega,L^2(D))}, \tag{4.83}$$

where

$$C_3(N,k,\varepsilon) := \frac{\tilde{C}_0 k}{r^2} \cdot \frac{C_0(k^3+1)}{k^2(2\sqrt{C_0}-1)} \cdot \frac{1-\big(2\tilde{C}_0\sqrt{C_0}(2k+3)\varepsilon\big)^N}{1-2\tilde{C}_0\sqrt{C_0}(2k+3)\varepsilon}, \tag{4.84}$$

$$C_4(N,k,\varepsilon) := \frac{C\tilde{C}_0^2 k(1+k)}{r^2} \cdot \frac{C_0(k^3+1)}{k^2(2\sqrt{C_0}-1)} \cdot \frac{1-\big(2\tilde{C}_0\sqrt{C_0}(2k+3)\varepsilon\big)^N}{1-2\tilde{C}_0\sqrt{C_0}(2k+3)\varepsilon}. \tag{4.85}$$

Proof. To obtain the above result, we need to estimate the double sum

$$\sum_{n=0}^{N-1}\sum_{j=0}^{n} \varepsilon^n \big[\tilde{C}_0(2k+3)\big]^{n-j} \mathbb{E}\big(\|u_j\|_{H^2(D)}\big),$$

which is in Theorem 4.4.9. Applying (4.30) to estimate $\|u_j\|_{H^2(D)}$ and exploiting some geometric series properties of the sum we get

$$\begin{aligned}
&\sum_{n=0}^{N-1}\sum_{j=0}^{n}\varepsilon^n\big[\tilde{C}_0(2k+3)\big]^{n-j}\mathbb{E}\big(\|u_j\|_{H^s(D)}\big)\\
&\quad\le\Big(k+\frac{1}{k^2}\Big)\|f\|_{L^2(\Omega,L^2(D))}\sum_{n=0}^{N-1}\sum_{j=0}^{n}\varepsilon^n\big[\tilde{C}_0(2k+3)\big]^{n-j}C(j,k)^{\frac{1}{2}}\\
&\quad=\frac{C_0^{\frac{1}{2}}(k^3+1)}{2k^2}\|f\|_{L^2(\Omega,L^2(D))}\sum_{n=0}^{N-1}\sum_{j=0}^{n}\varepsilon^n 4^j C_0^{\frac{j}{2}}(1+k)^j\big[\tilde{C}_0(2k+3)\big]^{n-j}\\
&\quad\le\frac{C_0^{\frac{1}{2}}(k^3+1)}{2k^2}\|f\|_{L^2(\Omega,L^2(D))}\sum_{n=0}^{N-1}\varepsilon^n\big[\tilde{C}_0(2k+3)\big]^{n}\sum_{j=0}^{n}2^j C_0^{\frac{j}{2}}\\
&\quad\le\frac{C_0(k^3+1)}{k^2(2\sqrt{C_0}-1)}\|f\|_{L^2(\Omega,L^2(D))}\sum_{n=0}^{N-1}\big[2\varepsilon\tilde{C}_0\sqrt{C_0}(2k+3)\big]^{n}\\
&\quad\le\frac{C_0(k^3+1)}{k^2(2\sqrt{C_0}-1)}\cdot\frac{1-\big(2\tilde{C}_0\sqrt{C_0}(2k+3)\varepsilon\big)^N}{1-2\tilde{C}_0\sqrt{C_0}(2k+3)\varepsilon}\|f\|_{L^2(\Omega,L^2(D))}.
\end{aligned}$$

We get (4.82) and (4.83) by applying the above inequality to (4.80) and (4.81) respectively. □

Remark 4.4.11. *Theorem 4.4.10, shows that the error generated by the proposed IP-DG method is optimal in the mesh size h.*

With (i) complete, we turn our attention to (ii), i.e. estimating the error generated by using the Monte Carlo method. The following lemma is well known (cf. [12, 58]).

Lemma 4.4.12. *For $n\ge 0$ the following estimates hold:*

$$\mathbb{E}\big(\|\mathbb{E}(u_n^h)-\Phi_n^h\|_{L^2(D)}^2\big)\le\frac{1}{M}\mathbb{E}(\|u_n^h\|_{L^2(D)}^2),\tag{4.86}$$

$$\mathbb{E}\big(\|\mathbb{E}(u_n^h)-\Phi_n^h\|_{1,h,D}^2\big)\le\frac{1}{M}\mathbb{E}(\|u_n^h\|_{1,h,D}^2).\tag{4.87}$$

Combining Lemmas 4.4.12 and 4.4.6, we get the following error estimate theorem.

Theorem 4.4.13. *Under the constraint $k^3h^2r^{-2} = O(1)$ the following estimates hold:*

$$\mathbb{E}\big(\|\mathbb{E}(u_n^h) - \Phi_n^h\|_{L^2(D)}^2\big) \leq \frac{1}{M}\Big(\frac{1}{k} + \frac{1}{k^2}\Big)^2 \hat{C}(n,k)\,\mathbb{E}(\|f\|_{L^2(D)}^2), \tag{4.88}$$

$$\mathbb{E}\big(\|\mathbb{E}(u_n^h) - \Phi_n^h\|_{1,h,D}^2\big) \leq \frac{1}{M}\Big(1 + \frac{1}{k}\Big)^2 \hat{C}(n,k)\,\mathbb{E}(\|f\|_{L^2(D)}^2), \tag{4.89}$$

Remark 4.4.14. *Estimates (4.88) and (4.89) show that for each fixed $n \geq 0$ the statistical error due to sampling is controlled by the number of realizations of u_n^h. Indeed, it can be easily proved by using Markov's inequality and Borel-Cantelli lemma that the statistical error converges to zero as M tends to infinity, see [12, Proposition 4.1] and [58, Theorem 3.2].*

The above estimates on $\mathbb{E}(u_N^h) - \Phi_n^h$ are now used to obtain the following theorem.

Theorem 4.4.15. *Suppose $k^3h^2r^{-2} = O(1)$ and $\hat{\sigma} := 4\varepsilon\hat{C}_0(1+k) < 1$. Then the following estimates hold:*

$$\mathbb{E}\big(\big\|E(U_N^h) - \Psi_N^h\big\|_{L^2(D)}\big) \leq \frac{\hat{C}_0}{2\sqrt{M}}\Big(\frac{1}{k} + \frac{1}{k^2}\Big)\|f\|_{L^2(\Omega,L^2(D)} \cdot \frac{1}{1-\hat{\sigma}}, \tag{4.90}$$

$$\mathbb{E}\big(\big\|E(U_N^h) - \Psi_N^h\big\|_{1,h,D}\big) \leq \frac{\hat{C}_0}{2\sqrt{M}}\Big(1 + \frac{1}{k}\Big)\|f\|_{L^2(\Omega,L^2(D)} \cdot \frac{1}{1-\hat{\sigma}}. \tag{4.91}$$

Proof. First, note that

$$U_N^h - \Psi_N^h = \sum_{n=0}^{N-1} \varepsilon^n \big(u_n^h - \Phi_n^h\big).$$

By (4.88) we get

$$\begin{aligned}\mathbb{E}\big(\big\|E(U_N^h)-\Psi_N^h\big\|_{L^2(D)}\big) &\le \sum_{n=0}^{N-1}\varepsilon^n\mathbb{E}\big(\|E(u_n^h)-\Phi_n^h\|_{L^2(D)}\big)\\ &\le \frac{1}{\sqrt{M}}\Big(\frac{1}{k}+\frac{1}{k^2}\Big)\|f\|_{L^2(\Omega,L^2(D)}\sum_{n=0}^{N-1}\varepsilon^n\hat{C}(n,k)^{\frac{1}{2}}\\ &\le \frac{\hat{C}_0}{2\sqrt{M}}\Big(\frac{1}{k}+\frac{1}{k^2}\Big)\|f\|_{L^2(\Omega,L^2(D)}\sum_{n=0}^{N-1}4^n\varepsilon^n\hat{C}_0^n(1+k)^n\\ &\le \frac{\hat{C}_0}{2\sqrt{M}}\Big(\frac{1}{k}+\frac{1}{k^2}\Big)\|f\|_{L^2(\Omega,L^2(D)}\cdot\frac{1}{1-\hat{\sigma}},\end{aligned}$$

where $\hat{\sigma}:=4\varepsilon\hat{C}_0(1+k)<1$.

Similarly, by (4.89)

$$\begin{aligned}\mathbb{E}\big(\big\|E(U_N^h)-\Psi_N^h\big\|_{1,h,D}\big) &\le \sum_{n=0}^{N-1}\varepsilon^n\mathbb{E}\big(\|E(u_n^h)-\Phi_n^h\|_{1,h,D}\big)\\ &\le \frac{\hat{C}_0}{2\sqrt{M}}\Big(1+\frac{1}{k}\Big)\|f\|_{L^2(\Omega,L^2(D)}\cdot\frac{1}{1-\hat{\sigma}}.\end{aligned}$$

The proof is complete. □

Remark 4.4.16. *Theorem 4.4.15 shows that the error generated by using the Monte Carlo method is on the order of $O(M^{-\frac{1}{2}})$. Thus, for an accurate approximation a large number of realizations M must be taken.*

4.5 The Overall Numerical Procedure

This section is devoted to defining and analyzing the overall efficient MCIP-DG algorithm for computing u^ε. The key to efficiency is the exploitation of the special structure inherent in the multi-modes expansion. In Subsection 4.5.1, the efficient MCIP-DG algorithm is defined and its computational complexity is analyzed. Subsection 4.5.2 summarizes all of the error estimates given in the previous sections

to obtain an estimate of the total error produced by using the multi-modes MCIP-DG method.

4.5.1 The Numerical Algorithm, Linear Solver and Computational Complexity

The goal of this subsection is to introduce an efficient MCIP-DG method to approximate the expectation of the solution to the random Helmholtz problem (4.1)–(4.2). The key to this method is the exploitation of the special structure of the multi-modes expansion of the solution described in Section 4.3. In order to judge the efficiency of this method, we must establish a reliable standard upon which to compare and contrast our method. For this standard, we use the classical MCIP-DG method that does not utilize the multi-modes expansion of the solution. In order to define such a method, we need to introduce a new IP-DG formulation. Given a sample realization of the coefficient $\eta(\omega_j, \cdot)$ and source data $f(\omega_j, \cdot)$ define $\hat{u}^h(\omega_j, \cdot) \in V_h^r$ as the solution to

$$\hat{a}_j^h(\hat{u}^h(\omega_j, \cdot), v^h) = (f(\omega_j, \cdot), v^h)_D, \qquad \forall v^h \in V_h^r, \tag{4.92}$$

where

$$\begin{aligned}\hat{a}_j^h(\phi, \psi) := b_h(\phi, \psi) - k^2\big((1 + \varepsilon\eta(\omega_j, \cdot))^2\phi, \psi\big)_{\mathcal{T}_h} + \mathbf{i}k\langle(1 + \varepsilon\eta(\omega_j, \cdot))\phi, \psi\rangle_{\mathcal{E}_h^B} \\ + \mathbf{i}\Big(L_1(\phi, \psi) + \sum_{m=0}^{r} J_m(\phi, \psi)\Big).\end{aligned}$$

Here $b_h(\cdot, \cdot)$, $L_1(\cdot, \cdot)$, and $J_m(\cdot, \cdot)$ are defined previously in Subsection 4.4.2. Notice that the main difference between (4.92) and (4.57) which was used to define u_n^h is that the sesquilinear form $\hat{a}_j^h(\cdot, \cdot)$ depends on the realization $\eta(\omega_j, \cdot)$. This is the key observation that makes the use of the multi-modes expansion worth-while when seeking an efficient MCIP-DG method.

Based on (4.92) the classical MCIP-DG method for solving the random Helmholtz problem is defined by the following algorithm:

Algorithm 1 (Classical MCIP-DG)

Inputs: $f, \eta, \varepsilon, k, h, M$.

Set $\tilde{\Psi}^h(\cdot) = 0$ (initializing).

For $j = 1, 2, \cdots, M$

Obtain realizations $\eta(\omega_j, \cdot)$ and $f(\omega_j, \cdot)$.

Solve for $\hat{u}^h(\omega_j, \cdot) \in V_r^h$ such that

$$\hat{a}_j^h\big(\hat{u}^h(\omega_j, \cdot), v_h\big) = \big(f(\omega_j, \cdot), v_h\big)_D \qquad \forall v_h \in V_r^h.$$

Set $\tilde{\Psi}^h(\cdot) \leftarrow \tilde{\Psi}^h(\cdot) + \frac{1}{M}\hat{u}^h(\omega_j, \cdot)$.

Endfor

Output $\tilde{\Psi}^h(\cdot)$.

This algorithm is very expensive for M large, because at each step of the loop a deterministic Helmholtz equation must be solved. This requires one to solve a large (especially for k large), ill-conditioned, indefinite linear system. It is well-known that no standard iterative method works well for such a system [37]. For this reason, Gaussian elimination is considered for each solve in the loop. Since the Gaussian elimination step is the most costly portion of the loop, the computational complexity is estimated in terms of Gaussian elimination steps.

Let h be the mesh size of a quasi-uniform partition $\mathcal{T}_h$ of the domain D. Then each coefficient matrix that appeared in the for-loop of Algorithm 1 has approximate size $O(L^d \times L^d)$, where $L = \frac{1}{h}$. Each Gaussian elimination solve will have computational complexity $O\left(\frac{3L^{3d}}{2}\right)$. Thus, the overall computational complexity of Algorithm 1 is $O\left(\frac{3L^{3d}M}{2}\right)$. Recall that the Monte Carlo method converges at a rate of $O(M^{-\frac{1}{2}})$.

Thus, M must be chosen sufficiently large in order to gain sufficient error reduction. Therefore, a computational complexity of $O\left(\frac{3L^{3d}M}{2}\right)$ is quite costly, and this makes Algorithm 1 not practical.

For this reason, the following algorithm is introduced.

Algorithm 2 (Multi-Modes MCIP-DG)

Inputs: $f, \eta, \varepsilon, k, h, M, N$

Set $\Psi_N^h(\cdot) = 0$ (initializing).

Generate the coefficient matrix A associated to the sesquilinear form $a_h(\cdot,\cdot)$ over $V_r^h \times V_r^h$.

Compute and store the LU decomposition of A.

For $j = 1, 2, \cdots, M$

Obtain realizations $\eta(\omega_j, \cdot)$ and $f(\omega_j, \cdot)$.

Set $S_0^h(\omega_j, \cdot) = f(\omega_j, \cdot)$.

Set $Q_0^h(\omega_j, \cdot) = 0$.

Set $u_{-1}^h(\omega_j, \cdot) = 0$.

Set $U_N^h(\omega_j, \cdot) = 0$ (initializing).

For $n = 0, 1, \cdots, N-1$

Solve for $u_n^h(\omega_j, \cdot) \in V_r^h$ such that

$$a_h\big(u_n^h(\omega_j, \cdot), v_h\big) = \big(S_n^h(\omega_j, \cdot), v_h\big)_D + \big\langle Q_n^h(\omega_j, \cdot), v_h\big\rangle_{\partial D} \qquad \forall v_h \in V_r^h,$$

using the LU decomposition and forward and backward substitution.

Set $U_N^h(\omega_j, \cdot) \leftarrow U_N^h(\omega_j, \cdot) + \varepsilon^n u_n^h(\omega_j, \cdot)$.

Set $S_{n+1}^h(\omega_j, \cdot) = 2k^2\eta(\omega_j, \cdot)u_n^h(\omega_j, \cdot) + k^2\eta(\omega_j, \cdot)^2 u_{n-1}^h(\omega_j, \cdot)$.

Set $Q_{n+1}^h(\omega_j, \cdot) = -\mathbf{i}k\eta(\omega_j, \cdot)u_n^h(\omega_j, \cdot)$.

Endfor

Set $\Psi_N^h(\cdot) \leftarrow \Psi_N^h(\cdot) + \frac{1}{M} U_N^h(\omega_j, \cdot)$.

Endfor

Output $\Psi_N^h(\cdot)$.

The key difference between Algorithm 1 and Algorithm 2 is the fact that the bilinear form $a_h(\cdot, \cdot)$ used in the "nearly deterministic" Helmholtz equation in the inner for loop of Algorithm 2 does not depend on the current mode number n or the current realization number j. Thus, only one stiffness matrix A must be computed and its LU decomposition can be reused when seeking a solution to the equation in the inner loop. This results in a great savings in terms of computational time required by the algorithm.

To analyze Algorithm 2, again the coefficient matrix A will have approximate size $O(L^d \times L^d)$. Thus, Gaussian elimination used to produce the LU decomposition has order $O(\frac{3L^{3d}}{2})$. After this LU decomposition is computed, solving the system using forward and backward substitution has complexity order $O(L^{2d})$. Thus, the computational complexity for Algorithm 2 is on the order $O\left(\frac{3L^{3d}}{2} + MNL^{2d}\right)$. M will be chosen large (c.f. Remark 4.4.16). On the other hand, N will be chosen to be a small positive integer (c.f. Remark 4.3.4). With the intent of choosing M large, using $M = L^d$ for Algorithm 2 yields a computational complexity $O\left(\frac{3L^{3d}}{2} + NL^{3d}\right)$. This is on the same order as a few Gaussian elimination solves.

The Monte Carlo method is naturally parallelizable and it is in this setting that Algorithm 1 should be implemented. The structure of Algorithm 2 also allows parallel implementation. This being said, unless one uses computational resources in which all Gaussian elimination solves in Algorithm 1 can be carried out at the same time, Algorithm 2 should be more efficient in terms of computation time.

4.5.2 Convergence Analysis

The goal of this subsection is to analyze the error of the multi-modes MCIP-DG approximation produced by Algorithm 2. Recall that Algorithm 2 uses the following three sequential approximations:

- Approximation of u^ε with a partial multi-modes expansion U_N^ε
- Approximation of U_N^ε with its IP-DG approximation U_N^h
- Approximation of $\mathbb{E}(U_N^h)$ with its Monte Carlo approximation Ψ_N^h

Thus, the error $\mathbb{E}(u^\varepsilon) - \Psi_N^h$ associated with Algorithm 2 can be decomposed in the following manner:

$$\mathbb{E}(u^\varepsilon) - \Psi_N^h = \big(\mathbb{E}(u^\varepsilon) - \mathbb{E}(U_N^\varepsilon)\big) + \big(\mathbb{E}(U_N^\varepsilon) - \mathbb{E}(U_N^h)\big) + \big(\mathbb{E}(U_N^h) - \Psi_N^h\big).$$

Each piece of this error decomposition has already been estimated in the previous sections of this chapter. The following theorem puts these results together to obtain estimates for the total error of Algorithm 2:

Theorem 4.5.1. *Under the assumptions that $u_n \in L^2(\Omega, H^2(D))$ for $n \geq 0$, $k^3h^2r^{-2} = O(1)$ and $\sigma, \hat{\sigma} < 1$ (i.e. $\varepsilon = O(k^{-1})$), the following error estimates hold:*

$$\mathbb{E}\big(\|\mathbb{E}(u^\varepsilon) - \Psi_N^h\|_{L^2(D)}\big) \leq C_1\varepsilon^N + C_2h^2 + C_3M^{-\frac{1}{2}}, \tag{4.93}$$

$$\mathbb{E}\big(\|\mathbb{E}(u^\varepsilon) - \Psi_N^h\|_{H^1(D)}\big) \leq C_4\varepsilon^N + C_5h + C_6M^{-\frac{1}{2}}, \tag{4.94}$$

where $C_j = C_j(C_0, \hat{C}_0, k, \varepsilon)$ are positive constants for $j = 1, 2, \cdots, 6$.

Proof. To begin, we apply the triangle inequality to the error decomposition given above. Then each term can be estimated separately using Theorems 4.3.3, 4.4.9, and 4.4.15. Note that Theorem 4.3.3 cannot be used directly; instead, the Cauchy-Schwarz

inequality must be used in the following manner:

$$\mathbb{E}\left(\|u^{\varepsilon}-U_N^{\varepsilon}\|_{H^j(D)}\right)^2 \leq \mathbb{E}\left(1^2\right)\mathbb{E}\left(\|u^{\varepsilon}-U_N^{\varepsilon}\|^2_{H^j(D)}\right)$$
$$\leq \frac{9C_0\sigma^{2N}}{32(1+k)^2}\left(k^j+\frac{1}{k}\right)^4\mathbb{E}\left(\|f\|^2_{L^2(D)}\right).$$

Thus, taking the square root on both sides and applying the definition of σ yields the first term in (4.93) and (4.94). With this result and those listed above, the desired inequalities follow. □

4.6 Numerical Experiments

In this section we present a series of numerical experiments in order to accomplish the following:

- compare our MCIP-DG method using the multi-modes expansion to a classical MCIP-DG method
- illustrate examples using our MCIP-DG method in which the perturbation parameter ε satisfies the constraint required by the convergence theory
- illustrate examples using our MCIP-DG method in which the perturbation parameter constraint is violated

In all our numerical experiments we use the spatial domain $D=(-0.5,0.5)^2$. To partition D we use a uniform triangulation $\mathcal{T}_h$. For a positive integer n, $\mathcal{T}_{1/n}$ denotes the triangulation of D consisting of $2n^2$ congruent isosceles triangles with side lengths $1/n, 1/n$, and $\sqrt{2}/n$. Figure 4.1 gives the sample triangulation $\mathcal{T}_{1/10}$.

To implement the random noise η, we note that η only appears in the integration component of our computations. Therefore, we made the choice to implement η only at quadrature points of the triangulation. To simulate the random media, we let $\eta(\cdot,\hat{x})$ be an independent random number chosen from a uniform distribution on

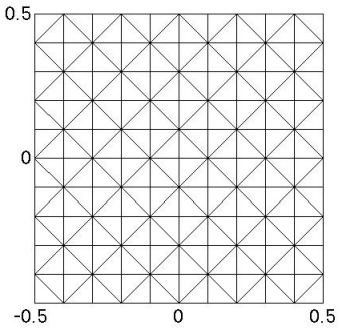


Figure 4.1: Triangulation $\mathcal{T}_{1/10}$.

some closed interval at each quadrature point $\hat{x}$. Figure 4.2 shows an example of such random media.

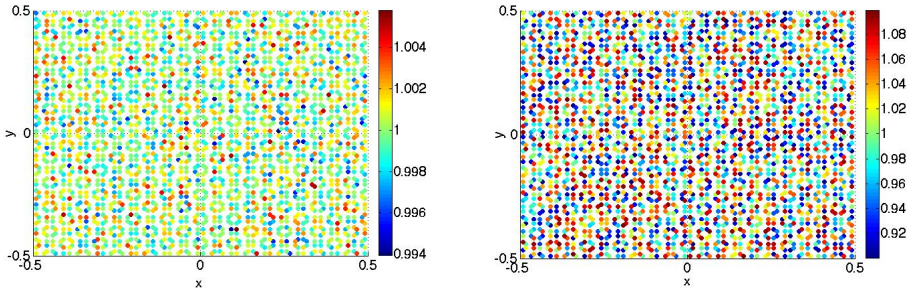


Figure 4.2: Discrete average media $\frac{1}{M}\sum_{j=1}^{M}\alpha(\omega_j,\cdot)$ (left) and a sample media $\alpha(\omega,\cdot)$ (right) computed for $h = 1/20$, $\varepsilon = 0.1$, $\eta(\cdot,x) \sim \mathcal{U}[-1,1]$, and $M = 1000$.

4.6.1 MCIP-DG with Multi-modes Expansion Compared to Classical MCIP-DG

The goal of this subsection is to verify the accuracy and efficiency of the proposed multi-modes MCIP-DG method. As a benchmark we compare this method to the classical MCIP-DG (i.e. produced using Algorithm 1). Throughout this section $\tilde{\Psi}^h$ is used to denote the computed approximation to $\mathbb{E}(u)$ using the classical MCIP-DG.

In this subsection, we set $f = 1$, $k = 5$, $1/h = 50$, $M = 1000$, and $\varepsilon = 1/(k+1)$. Here ε is chosen with the intent of satisfying the constraint set by the convergence

theory in the preceding section. η is sampled as described above from a uniform distribution on the interval $[0, 1]$. Ψ_N^h is computed for $N = 1, 2, 3, 4, 5$.

In our first test, we compute $\|\Psi_N^h - \tilde{\Psi}^h\|_{L^2(D)}$. The results are displayed in Figure 4.3. As expected, we find that the difference between Ψ_N^h and $\tilde{\Psi}^h$ is very small. We also observe that we are benefited more by the first couple modes while the help from the later modes is relatively small. From this experiment, we see the error decrease is similar to ε^N. This is expected from Theorem 4.4.15.

To test the efficiency of our MCIP-DG method with multi-modes expansion, we compare the CPU time for computing Ψ_N^h and $\tilde{\Psi}^h$. Both methods are implemented on the same computer using Matlab. Matlab's built-in LU factorization is called to solve the linear systems. The results of this test are shown in Table 4.1. As expected, we find that the use of the multi-modes expansion improves the CPU time for the computation considerably. In fact, the table shows that this improvement is an order of magnitude. Also, as expected, as the number of modes used is increased the CPU time increases in a linear fashion.

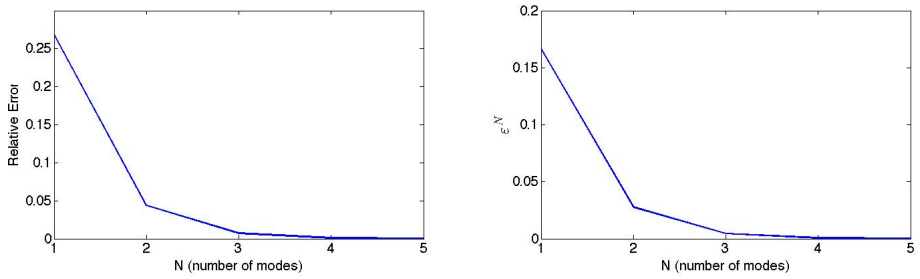


Figure 4.3: (left) Relative error in the L^2-norm between Ψ_N^h computed using the multi-modes MCIP-DG method and $\tilde{\Psi}^h$ computed using the classical MCIP-DG method. (right) ε^N vs. N for $N = 1, 2, \cdots, 5$.

4.6.2 More Numerical Tests

The goal of this subsection is to demonstrate the approximations obtained by our multi-modes MCIP-DG method using different magnitudes of parameter ε. We only consider the case $0 < \varepsilon < 1$ in order to legitimize the series expansion u^ε. Our

Table 4.1: CPU times required to compute the multi-modes MCIP-DG approximation Ψ_N^h and the classical MCIP-DG approximation $\tilde{\Psi}^h$.

Approximation	CPU Time (s)
$\tilde{\Psi}^h$	3.4954×10^5
Ψ_1^h	1.0198×10^4
Ψ_2^h	2.0307×10^4
Ψ_3^h	3.0037×10^4
Ψ_4^h	3.9589×10^4
Ψ_5^h	4.9011×10^4

hope is that $\varepsilon = O(k^{-1})$ required by the convergence theory (c.f. Theorem 4.4.15) is not sharp in practice, and thus our multi-modes MCIP-DG method produces good approximations for larger values of ε. Similar to the numerical experiments from [42], we choose the function $f = \sin\big(k\alpha(\omega,\cdot)r\big)/r$, where r is the radial distance from the origin and $\alpha(\omega,\cdot)$ is implemented as described in the beginning of this section. Since our intention is to observe what happens as we vary ε, we fix $k = 50$, $h = 1/100$, and $M = 1000$.

In Figures 4.4 and 4.5, we set $\varepsilon = 0.02$ and $|\eta| \leq 1$. In Figure 4.4 we present plots of the magnitude of the computed mean $\mathrm{Re}\big(\Psi_3^h\big)$ and a computed sample $\mathrm{Re}\big(U_3^h\big)$, respectively, over the whole domain D. Figure 4.5 gives the plots of a cross section of the computed mean $\mathrm{Re}\big(\Psi_3^h\big)$ and a computed sample $\mathrm{Re}\big(U_3^h\big)$, respectively, over the line $y = x$. In this first example, we observe that the computed sample does not differ greatly from the computed mean because ε is very small.

In Figures 4.6–4.11, we fix $|\eta| \leq 1$ and increase ε past the constraint established in the preceding convergence theory. As expected, we see that as ε increases the computed sample differs more from the computed mean. We also observe that as ε increases the phase of the wave remains relatively intact but the magnitude of the wave becomes more uniform.

In Table 4.2, the relative error (measured in the L^2-norm) between the multi-modes approximation Ψ_N^h and the classical Monte Carlo approximation Ψ^h is given for $\varepsilon = 0.02, 0.1, 0.5, 0.8$. In this table only three modes (i.e., $N = 3$) are used. Recall

that the convergence theory in this case only holds for ε on the order of the first value 0.02. That being said, we observe that the approximations corresponding to $\varepsilon = 0.1$ and $\varepsilon = 0.5$ are relatively close to those obtained using the classical Monte Carlo method. Another observation that can be made from Table 4.2 is that as ε increases the relative error increases. This is expected from the convergence theory.

Recall that the error predicted in the convergence theory can be bounded by a term with the factor ε^N. Thus for ε relatively large, one must use more modes to decrease the error. Keeping this in mind, Table 4.3 records the relative error (measured in the L^2-norm) between the multi-modes approximation Ψ_N^h and the classical Monte Carlo approximation $\tilde{\Psi}^h$ for $\varepsilon = 0.5, 0.8$ and $N = 4, 5, 6, 7$. We observe that the relative error decreases as N increases when $\varepsilon = 0.5$. On the other hand, the relative error increases as N increases when $\varepsilon = 0.8$. From Tables 4.2 and 4.3, we observe that multi-modes approximation Ψ_N^h is relatively accurate (measured against an approximation from the classical Monte Carlo method) even in cases when ε does not satisfy the constraint set forth in the convergence theory. We also observe that when ε becomes too large, the multi-modes approximation no longer agrees with the classical Monte Carlo method.

Table 4.2: Relative error in the L^2-norm between the multi-modes MCIP-DG approximation Ψ_3^h and the classical MCIP-DG approximation $\tilde{\Psi}^h$.

ε	0.02	0.1	0.5	0.8
Relative L^2 Error	3.0125×10^{-4}	6.0073×10^{-4}	0.2865	1.6979

Table 4.3: Relative error in the L^2-norm between the multi-modes MCIP-DG approximation Ψ_N^h and the classical MCIP-DG approximation $\tilde{\Psi}^h$.

ε	$N = 4$	$N = 5$	$N = 6$	$N = 7$
0.5	0.2866	0.1125	0.1137	0.0554
0.8	1.7036	1.6713	1.6839	1.7887

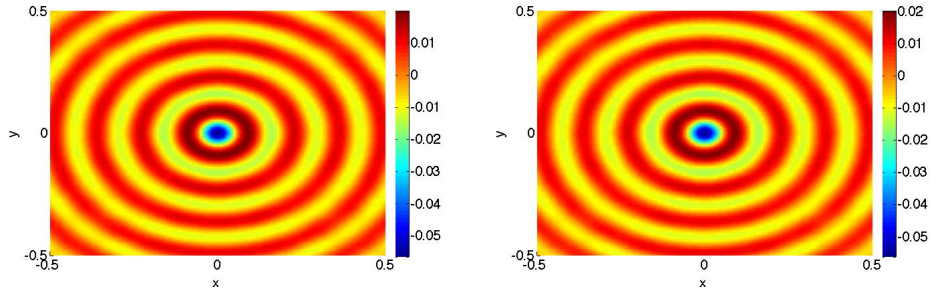


Figure 4.4: $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$(right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.02$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

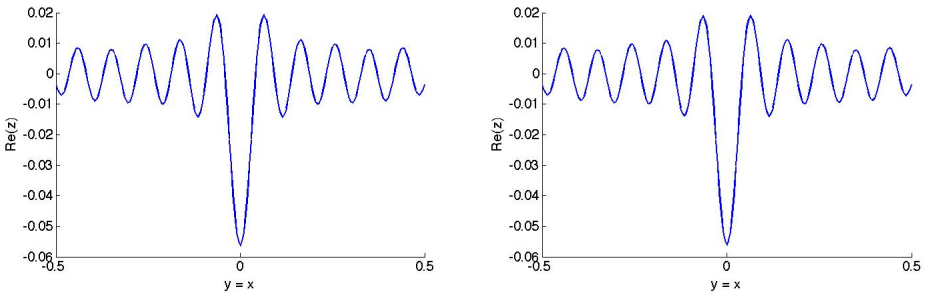


Figure 4.5: Cross sections of $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.02$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$, over the line $y = x$.

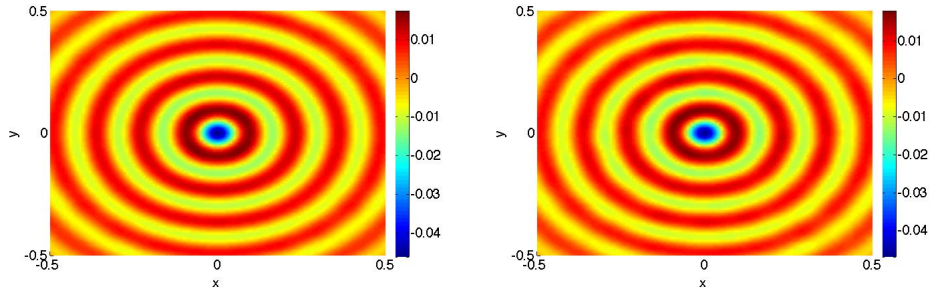


Figure 4.6: $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.1$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

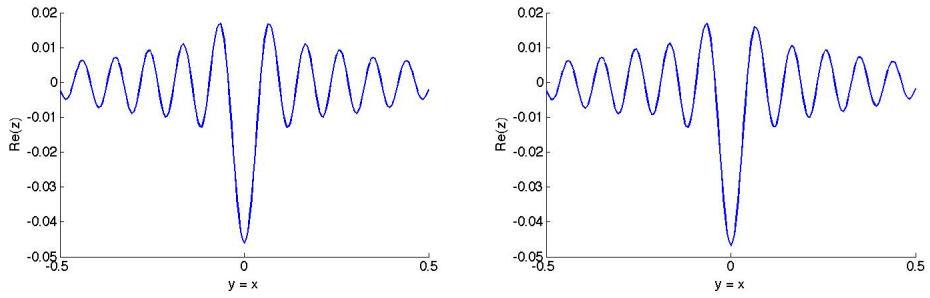


Figure 4.7: Cross sections of $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.1$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

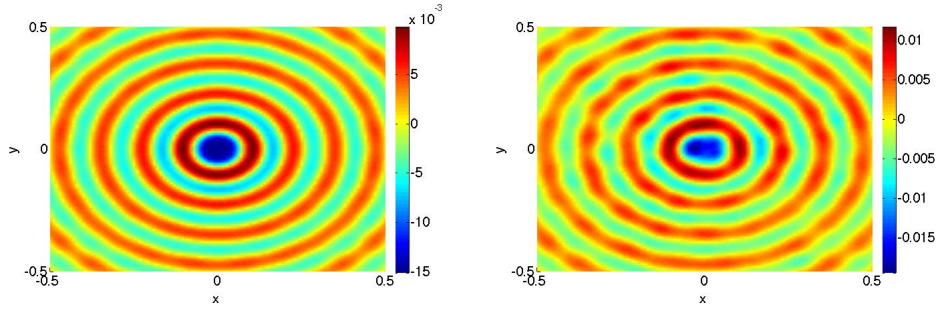


Figure 4.8: $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.5$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

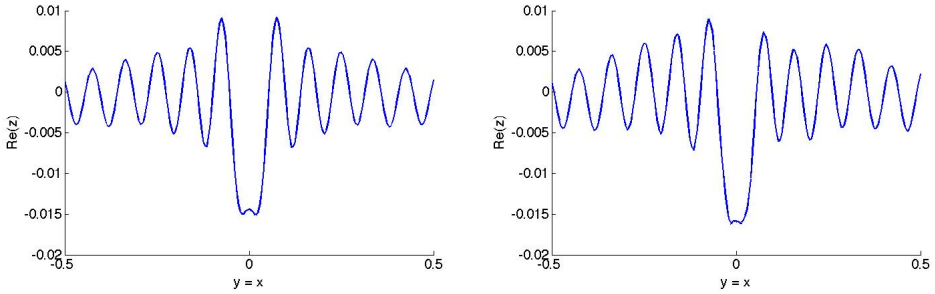


Figure 4.9: Cross sections of $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.5$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

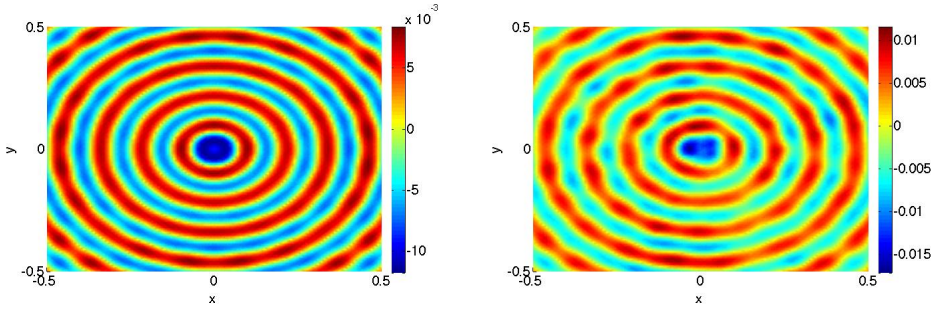


Figure 4.10: $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.8$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

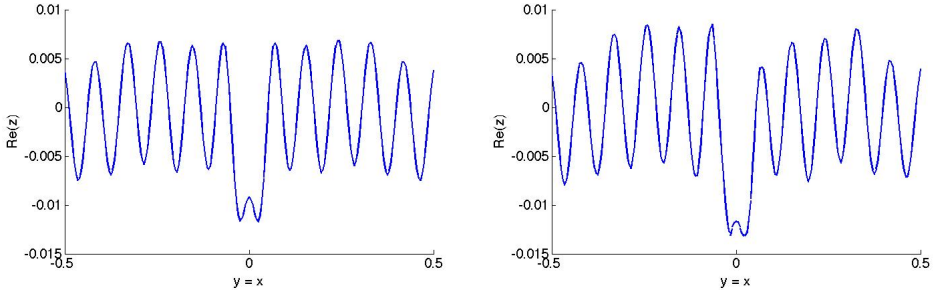


Figure 4.11: Cross sections of $\mathrm{Re}\big(\Psi_3^h\big)$ (left) and $\mathrm{Re}\big(U_3^h\big)$ (right) computed for $k = 50$, $h = 1/100$, $\varepsilon = 0.8$, $\eta(\cdot, x) \sim \mathcal{U}[-1, 1]$, and $M = 1000$.

Chapter 5

Schwarz Space Decomposition Methods for Nonsymmetric and Indefinite Problems

5.1 Introduction

The original Schwarz method, proposed and analyzed by Hermann Schwarz in 1870 [72], is an iterative method to find the solution of a partial differential equation (PDE) on a complicated domain which is the union of two overlapping simpler subdomains. The method solves the equation on each of the two subdomains by using the latest values of the approximate solution as the boundary conditions on the parts of the subdomain boundaries which are inside of the given domain. The idea of splitting a given problem posed on a large (and possibly complicated) domain into several subproblems posed on smaller subdomains and then solving the subdomain problems either sequentially or in parallel is a very appealing idea. Such a "divide-and-conquer" idea is at the heart of every domain decomposition or Schwarz method.

It is well-known that [77] the domain decomposition strategy can be introduced at the following three different levels: the continuous level for PDE analysis as proposed

and analyzed by Hermann Schwarz in 1870, the discretization level for constructing (hybrid and composite) discretization methods, and the algebraic level for solving algebraic systems arising from the numerical approximations of PDE problems. These three levels are often interconnected, and each of them has its own merit to be studied. Most of the recent efforts and attentions have been focused on the algebraic level. The field of domain decomposition methods has blossomed and undergone intensive and phenomenal development during the last thirty years (cf. [74, 65, 77] and the references therein). The phenomenal development has largely been driven by the ever-increasing demands for fast solvers for solving important and complicated scientific, engineering, and industrial application problems which are often governed mathematically by a PDE or a system of PDEs. It has also been infused and facilitated by the rapid advances in computer hardware and the emergence of parallel computing technologies.

At the algebraic level, domain decomposition methods or Schwarz methods have been well developed and studied for various numerical approximations (discretizations) of many types of PDE problems including finite element methods (cf. [31, 81]), mixed finite element methods and spectral methods (cf. [77]), and discontinuous Galerkin methods (cf. [40, 56, 41, 4]). A general abstract framework, backed by an elegant convergence theory, was well established many years ago for symmetric and positive definite (SPD) PDE problems and their numerical approximations (cf. [31, 81, 74, 65, 77, 83] and the references therein).

Despite the tremendous advances in domain decomposition (Schwarz) methods over the past thirty years, the current framework and convergence theory are mainly confined to SPD problems in Hilbert spaces. Because the framework and especially the convergence theory indispensably rely on the SPD properties of the underlying problem and the Hilbert space structures, they do not apply to genuinely nonsymmetric and/or indefinite problems. As a result, the SPD framework and theory leave many important and interesting problems uncovered as pointed out in [77, page 311].

This chapter attempts to address this important issue in Schwarz methods. The goal of this chapter is to introduce a new Schwarz framework and theory, based on the well-known idea of space decomposition as in the SPD case, for nonsymmetric and indefinite linear systems arising from continuous and discontinuous Galerkin approximations of general nonsymmetric and indefinite elliptic partial differential equations under some "minimum" structure assumptions. Unlike the SPD framework and theory, our new framework and theory are presented in a variational setting in Banach spaces instead of Hilbert spaces. Such a general framework allows broader applications of Schwarz methods. Additive, multiplicative, and hybrid Schwarz methods are developed. A comprehensive Schwarz preconditioner theory is provided which includes condition number estimates for the additive Schwarz preconditioners and hybrid Schwarz preconditioners. The main idea of our nonsymmetric and indefinite Schwarz framework and theory is to use weak coercivity (satisfied by the nonsymmetric and indefinite bilinear form) induced norms to replace the standard bilinear form induced norm in the SPD Schwarz framework and theory (see Sections 5.2–5.4 for a detailed exposition). As expected, working with such weak coercivity induced norms and nonsymmetric and indefinite bilinear forms is quite delicate. It requires new and different technical tools in order to establish our preconditioner theory.

The remainder of this chapter is organized in the following way. In Section 5.2, we introduce notation, the functional setting, and the variational problems which we aim to solve. Section 5.2 also contains some further discussions on the main idea of the chapter. Section 5.3 is devoted to establishing an abstract additive Schwarz, multiplicative Schwarz, and hybrid Schwarz framework for general nonsymmetric and indefinite algebraic problems in a variational setting in general Banach spaces. In Section 5.4, we present an abstract preconditioner theory for the additive and hybrid Schwarz methods proposed in Section 5.3. In Section 5.5, we present some applications of the proposed nonsymmetric and indefinite Schwarz framework to discontinuous Galerkin approximations of convection-diffusion

(in particular, convection-dominated) problems. We also provide extensive 1-D numerical experiments to gauge the performance of the proposed nonsymmetric and indefinite Schwarz methods.

5.2 Functional Setting and Statement of Problems

5.2.1 Variational Problem

Let X be a real Hilbert space with the inner product $(\cdot,\cdot)_X$ and the induced norm $\|\cdot\|_X$. Let $V, W \subset X$ be two reflexive Banach spaces endowed with the norms $\|\cdot\|_V$ and $\|\cdot\|_W$ respectively. Let $\mathcal{A}(\cdot,\cdot)$ be a real bilinear form defined on the product space $V \times W$ and $\mathcal{F}$ be a real linear functional defined on W. We consider the following variational problem: Find $u \in V$ such that

$$\mathcal{A}(u,w) = \mathcal{F}(w) \qquad \forall w \in W. \tag{5.1}$$

The well-posedness of the above variational problem has been extensively studied. One such result is summarized in the following theorem:

Theorem 5.2.1. (cf. [9]) *Suppose that $\mathcal{F}$ is a bounded linear functional on W. Assume that $\mathcal{A}(\cdot,\cdot)$ is continuous and* weakly coercive *in the sense that there exist constants $C_{\mathcal{A}}, \gamma_{\mathcal{A}} > 0$ such that*

$$|\mathcal{A}(v,w)| \le C_{\mathcal{A}}\|v\|_V\|w\|_W \qquad \forall v \in V,\ w \in W, \tag{5.2}$$

$$\sup_{w\in W} \frac{\mathcal{A}(v,w)}{\|w\|_W} \ge \gamma_{\mathcal{A}}\|v\|_V \qquad \forall v \in V, \tag{5.3}$$

$$\sup_{v\in V} \mathcal{A}(v,w) > 0 \qquad \forall\, 0 \ne w \in W. \tag{5.4}$$

Then problem (5.1) has a unique solution $u \in V$. Moreover,

$$\|u\|_V \le \frac{\|\mathcal{F}\|}{\gamma_{\mathcal{A}}}. \tag{5.5}$$

Remark 5.2.2. *(a) Theorem 5.2.1 is called Lax-Milgram-Babuška theorem in the literature (cf. [69]). It was first introduced to the finite element context in [8] (also see [9]). An earlier version of the theorem can also be found in [62].*

(b) As pointed out in [9, page 117], condition (5.4) can be replaced by the following more restrictive condition: There exists a constant $\beta_{\mathcal{A}} > 0$ such that

$$\sup_{v \in V} \frac{\mathcal{A}(v, w)}{\|v\|_V} \geq \beta_{\mathcal{A}} \|w\|_W \qquad \forall w \in W. \tag{5.6}$$

The above condition can be viewed as a weak coercivity *condition for the adjoint bilinear form $\mathcal{A}^*(\cdot, \cdot)$ of $\mathcal{A}(\cdot, \cdot)$.*

(c) Weak coercivity condition (5.3) is often called the inf-sup *or* Babuška–Brezzi *condition in the finite element literature [16, 24] for a different reason. It appears and plays a vital role for saddle point problems and their (mixed) finite element approximations (cf. [17, 18]).*

(d) Theorem 5.2.1 is certainly valid when $V = W$. Since condition (5.3) is weaker than strong coercivity, Theorem 5.2.1 is a stronger result than the classical Lax-Milgram Theorem for the case $V = W$. Indeed, for most convection-dominated convection-diffusion problems, $V = W$. However, there are situations where condition (5.3) holds but strong coercivity fails (c.f. Section 5.5).

(e) There are also situations where one prefers to use different norms for the trial space V and the test space W even if $V = W$ (c.f. the generalize weak coercivity properties in Chapter 2). Theorem 5.2.1 also provides a convenient framework to handle such a situation.

5.2.2 Discrete Problem

As problem (5.1) is posed on infinite dimensional spaces V and W, to solve it numerically, one must approximate V and W by some finite dimensional spaces $V_n, W_n \subset X$. Here $n = \dim(V_n) = \dim(W_n)$ is a positive integer which denotes the dimension of V_n and W_n. If one of (or both) V_n and W_n is not a subspace

of its corresponding infinite dimensional space, then one also needs to provide an approximate bilinear form $a(\cdot,\cdot)$ for $\mathcal{A}(\cdot,\cdot)$ so that $a(\cdot,\cdot)$ is well defined on $V_n \times W_n$. In addition, if W_n is not a subspace of W one also needs to provide an approximate linear functional f for $\mathcal{F}$ so that f is well defined on W_n.

Once V_n, W_n, a and f are constructed, the Galerkin method for problem (5.1) is defined as seeking $u_n \in V_n$ such that

$$a(u_n, w_n) = f(w_n) \qquad \forall w_n \in W_n. \tag{5.7}$$

Pick a basis $\{\phi^{(j)}\}_{j=1}^n$ of V_n and a basis $\{\psi^{(j)}\}_{j=1}^n$ of W_n. It is trivial to check that the discrete variational problem (5.7) can be rewritten as the following linear system of equations:

$$A\mathbf{u} = \mathbf{f}, \tag{5.8}$$

where $\mathbf{u} = [u^{(j)}]_{j=1}^n$ is the coefficient vector of the representation of u_n in terms of the basis $\{\phi^{(j)}\}_{j=1}^n$ and

$$A = \left[a_{ij}\right]_{i,j=1}^n, \qquad a_{ij} = a(\phi^{(j)}, \psi^{(i)}), \tag{5.9}$$

$$\mathbf{f} = \left[f^{(i)}\right]_{i=1}^n, \qquad f^{(i)} = f(\psi^{(i)}). \tag{5.10}$$

The properties of matrix A (called a stiffness matrix) are obviously determined by the properties of the discrete bilinear form $a(\cdot,\cdot)$ and the approximate spaces V_n and W_n. When $V_n = W_n$ it is well known that [49] A is symmetric if and only if $a(\cdot,\cdot)$ is symmetric and A is positive definite provided that $a(\cdot,\cdot)$ is strongly coercive on $V_n \times V_n$. In general, A is just an $n \times n$ nonsymmetric real matrix if $a(\cdot,\cdot)$ is not symmetric. A also can be indefinite (i.e., A has at least one negative and one positive eigenvalue) if $a(\cdot,\cdot)$ fails to be coercive.

As (5.8) is a square linear system, by a well-known algebraic fact we know that (5.8) has a unique solution $\mathbf{u}$ provided that the stiffness matrix A is nonsingular.

This nonsingular condition on A becomes necessary if one wants (5.8) to be uniquely solvable for *arbitrary* vector $\mathbf{f}$. For most application problems (such as boundary value problems for elliptic PDEs), one needs to consider various choices of the "load" functional $\mathcal{F}$, so the vector $\mathbf{f}$ is practically "arbitrary" in (5.8). Hence, besides some deeper mathematical and algorithmic considerations, asking for the stiffness matrix A to be *nonsingular* is a "minimum" requirement for the discretization method (5.7) to be practically useful.

Sufficient conditions on the discrete bilinear form $a(\cdot,\cdot)$ and the approximate spaces V_n and W_n which infer the unique solvability of the linear system (5.8) have been well studied and understood in the past thirty years. In particular, for the SPD type (algebraic) problems arising from various discretizations of boundary value problems for elliptic PDEs [8, 9, 24, 16, 14, 66]. In the following, we shall quote some of these well-known results in a theorem which is a counterpart of Theorem 5.2.1.

Theorem 5.2.3. (cf. [8, 9]) *Suppose that f is a bounded linear functional on W_n. Assume that $a(\cdot,\cdot)$ is continuous and weakly coercive in the sense that there exist constants $C_a, \gamma_a, \beta_a > 0$ such that*

$$|a(v,w)| \le C_a \|v\|_{V_n} \|w\|_{W_n} \qquad \forall v \in V_n,\ w \in W_n, \tag{5.11}$$

$$\sup_{w \in W_n} \frac{a(v,w)}{\|w\|_{W_n}} \ge \gamma_a \|v\|_{V_n} \qquad \forall v \in V_n, \tag{5.12}$$

$$\sup_{v \in V_n} \frac{a(v,w)}{\|v\|_{V_n}} \ge \beta_a \|w\|_{W_n} \qquad \forall w \in W_n. \tag{5.13}$$

Then problem (5.7) has a unique solution $u_n \in V_n$. Moreover,

$$\|u_n\|_{V_n} \le \frac{\|f\|}{\gamma_a}. \tag{5.14}$$

A few remarks are in order about the above well-posedness theorem.

Remark 5.2.4. *(a) The constants C_a, γ_a, and β_a do not need to be independent of n for Theorem 5.2.3 to hold. From a practical perspective, if these constants are*

dependent on n then the numerical discretization method characterized by (5.7) *may not be convergent. Since the convergence of the numerical discretization method is not the focus of this chapter, this independence is not necessary.*

(b) Condition (5.13) *is equivalent to requiring that the adjoint $a^*(\cdot,\cdot)$ of $a(\cdot,\cdot)$ is weakly coercive.*

(c) Conditions (5.11)–(5.13) *are analogies of their continuous counterparts* (5.2)–(5.4). *The discrete weak coercivity condition* (5.12) *is often called the* inf-sup *or* Babuška–Brezzi *condition in the finite element literature [16, 24] for a different reason. It is the most important one in a set of sufficient conditions for a mixed finite element to be stable (cf. [17, 18]).*

(d) A numerical method which fulfills conditions (5.11)–(5.13) *is guaranteed to be uniquely solvable and stable. Hence, these conditions can be used as a test stone to determine whether a numerical method is a "good" method. For this reason, we shall call the numerical method* (5.7) *an* inf-sup preserving method *or a* weak coercivity preserving method *if it satisfies* (5.11)–(5.13).

(e) Theorem 5.2.3 focuses on the unique solvability and the stability of the numerical method (5.7) *not on the accuracy of the method. We like to note that method* (5.7) *indeed is an accurate numerical method provided that approximate spaces V_n and W_n are accurate approximations of V and W (cf. [9]).*

5.2.3 Main Objective

As we briefly explained above, approximating the variational problem (5.1) by a Galerkin method certainly results in solving the linear system (5.8). It is well known that the common dimension n of the approximation spaces V_n and W_n has to be sufficiently large in order for the Galerkin method to be accurate. As a result, the size of the linear system (i.e., the size of the matrix A) is expected to be very large in applications. Moreover, if (5.1) is a variational formulation of some elliptic boundary value problem, then the stiffness matrix A is certainly ill-conditioned in the sense

that the condition number $\kappa(A) := \|A\|\|A^{-1}\|$ is very large. Here $\|A\|$ denotes a matrix norm of A. For example, in the case of second and fourth order elliptic boundary value problems, $\kappa(A) = O(n^{\frac{2}{d}})$ and $\kappa(A) = O(n^{\frac{4}{d}})$, respectively, where d is the spatial dimension of the domain (cf. [16, 77]). Consequently, it is not efficient to solve linear system (5.8) directly using classical iterative methods even if they converge. Furthermore, unlike in the SPD case, classical iterative methods often do not converge for general nonsymmetric and indefinite linear system (5.8) (cf. [49, 77]).

As a first step toward developing better iterative solvers for nonsymmetric and indefinite linear system (5.8), it is natural to design a "good" preconditioner (i.e., an $n \times n$ real matrix B) such that BA is well-conditioned (i.e., $\kappa(BA)$ is relatively small, say, significantly smaller than $\kappa(A)$). Then one can try classical iterative methods. In particular, the Generalized Minimal Residual (GMRES) method can be used on the preconditioned system

$$BA\mathbf{x} = B\mathbf{b}. \tag{5.15}$$

One can also develop some new (and hopefully better) iterative methods if classical iterative methods still do not work as well on (5.15) as one had hoped.

As was already mentioned in Section 5.1, the focus of this chapter is exactly what is described above. Our goal is to develop a new Schwarz framework and theory, based on the well-known idea of space decomposition, for solving nonsymmetric and indefinite linear system (5.8) which arises from the Galerkin method (5.7) as an approximation of the variational problem (5.1). As expected, our nonsymmetric and indefinite Schwarz framework and theory are natural extensions of the well-known SPD Schwarz framework and theory which were nicely described in [31, 81, 74, 65, 77].

5.3 An Abstract Schwarz Framework for Nonsymmetric and Indefinite Problems

For the sake of notational brevity, throughout the remainder of this chapter we shall suppress the sub-index n in the discrete spaces V_n and W_n and in discrete functions u_n, v_n and w_n. In other words, V and W are used to denote V_n and W_n, and u, v and w are used to denote u_n, v_n and w_n. In addition, we shall make an effort below to use the same or similar terminologies, as well as space and norm notation as those in [77] for the symmetric and positive definite (SPD) Schwarz framework and theory. We shall also make comments about notation and terminologies which have no SPD counterparts and try to make links between the well known SPD Schwarz framework and theory and our nonsymmetric and indefinite Schwarz framework and theory.

To motivate, we recall that in the SPD Schwarz framework and theory [31, 81, 74, 65, 77], since $V = W$ and the discrete bilinear form $a(\cdot, \cdot)$ is symmetric and strongly coercive, $\sqrt{a(v, v)}$ defines a convenient norm (which is also equivalent to the $\|\cdot\|_V$-norm) on the space V (as well as on its subspaces). This bilinear form induced norm plays a vital role in the SPD Schwarz framework and theory.

Unfortunately, without the symmetry and strong coercivity assumptions on $a(\cdot, \cdot)$, $\sqrt{a(v, v)}$ is not a norm anymore when $V = W$. It is not even well defined if $V \neq W$. To overcome this difficulty, the existing nonsymmetric and indefinite Schwarz framework and theory (cf. [22, 77, 82]), which only deal with the case $V = W$, assume that $a(\cdot, \cdot)$ has a decomposition $a(\cdot, \cdot) = a_0(\cdot, \cdot) + a_1(\cdot, \cdot)$, where $a_0(\cdot, \cdot)$ is assumed to be symmetric and strongly coercive (i.e., it is SPD) and $a_1(\cdot, \cdot)$ is a perturbation of $a_0(\cdot, \cdot)$. In this setting, $a_0(\cdot, \cdot)$ then induces an equivalent (to $\|\cdot\|_V$) norm $\sqrt{a_0(v, v)}$ and one then works with this norm as in the SPD case. Unfortunately, such a setting requires that $a_1(\cdot, \cdot)$ is a *small* perturbation of $a_0(\cdot, \cdot)$, which is why the existing nonsymmetric and indefinite Schwarz framework and theory only apply to "nearly" SPD problems. Hence, it leaves more interesting and more difficult nonsymmetric and indefinite problems unresolved.

5.3.1 Main Assumptions and Main Idea

To develop a new Schwarz framework and theory for general nonsymmetric and indefinite problems, our only assumptions on the discrete problem (5.7) are those stated in the well-posedness Theorem 5.2.3. We now restate those assumptions on the discrete bilinear form $a(\cdot,\cdot)$ and its adjoint $a^*(\cdot,\cdot)$ using the new function and space notation (i.e., after suppressing the sub-index n) as follows:

Main Assumptions

(MA_1) *Continuity* There exists a positive constant C_a such that

$$|a(v,w)| \leq C_a \|v\|_V \|w\|_W \qquad \forall v \in V,\, w \in W. \tag{5.16}$$

(MA_2) *Weak coercivity* There exists positive constants γ_a, β_a such that

$$\sup_{w\in W} \frac{a(v,w)}{\|w\|_W} \geq \gamma_a \|v\|_V \qquad \forall v \in V, \tag{5.17}$$

$$\sup_{v\in V} \frac{a(v,w)}{\|v\|_V} \geq \beta_a \|w\|_W \qquad \forall w \in W. \tag{5.18}$$

Remark 5.3.1. *(a) Since $a^*(w,v) = a(v,w)$, then the continuity condition (5.16) is equivalent to*

$$|a^*(w,v)| \leq C_a \|w\|_W \|v\|_V \qquad \forall w \in W,\, v \in V, \tag{5.19}$$

and (5.18) is equivalent to

$$\sup_{v\in V} \frac{a^*(w,v)}{\|v\|_V} \geq \beta_a \|w\|_W \qquad \forall w \in W. \tag{5.20}$$

(b) Assumptions (MA_1) and (MA_2) impose some restrictions on the underlying Galerkin method (5.7). But as we noted in Remark 5.2.4, these are some "minimum"

conditions for the Galerkin method to be practically useful. From that point of view, (MA_1) and (MA_2) are not restrictions at all.

As it was pointed out in the previous subsection, for a general nonsymmetric and indefinite problem, since the discrete bilinear form $a(\cdot,\cdot)$ is not strongly coercive, then $a(v,v)$ is not a norm anymore. In fact, $a(v,v)$ may not even be defined if $V \neq W$. So a crucial question is what norms (if any) would $a(\cdot,\cdot)$ induce on V and W which are equivalent to $\|\cdot\|_V$ and $\|\cdot\|_W$. It turns out that $a(\cdot,\cdot)$ does induce equivalent norms on both V and W, and these norms are hidden in the weak coercivity conditions (5.17) and (5.18). This key observation leads to the main idea of this chapter; that is, we define the following weak coercivity induced norms:

$$\|v\|_a := \sup_{w\in W} \frac{a(v,w)}{\|w\|_{\mathrm{W}}} \qquad \forall v \in V, \tag{5.21}$$

$$\|w\|_{a^*} := \sup_{v\in V} \frac{a^*(w,v)}{\|v\|_{\mathrm{V}}} \qquad \forall w \in W. \tag{5.22}$$

Assumptions (MA_1) and (MA_2) immediately infer the following norm equivalence result. Since its proof is trivial, we omit it.

Lemma 5.3.2. *The following inequalities hold:*

$$\gamma_a \|v\|_V \leq \|v\|_a \leq C_a \|v\|_V \qquad \forall v \in V, \tag{5.23}$$

$$\beta_a \|w\|_W \leq \|w\|_{a^*} \leq C_a \|w\|_W \qquad \forall w \in W. \tag{5.24}$$

We conclude this subsection by noting that the variational setting laid out so far is a Banach space setting. No Hilbert space structure is required for the spaces V and W. This is not only mathematically interesting but also practically valuable because for some PDE application problems it is imperative to work in a Banach space setting. We also note that if $V = W$ and $a(\cdot,\cdot)$ is SPD (i.e., it is symmetric and strongly coercive), then $\|v\|_a = \|v\|_{a^*} = \sqrt{a(v,v)}$. Hence, we recover the standard bilinear form induced norm.

5.3.2 Space Decomposition and Local Solvers

It is well known [31, 81, 74, 83, 77] that Schwarz domain decomposition methods can be presented abstractly in the framework of the space decomposition method. In particular, the physical domain decomposition provides a practical and effective way to construct the required space decomposition and local solvers in the method. To some extent, the space decomposition method to the Schwarz domain decomposition method is what the LU factorization is to the classical Gaussian elimination method.

Like in the SPD Schwarz framework (cf. [77]), there are two essential ingredients in our nonsymmetric and indefinite Schwarz framework, namely, (i) *construction of a pair of "compatible" space decompositions for V and W* and (ii) *construction of a local solver (or local discrete bilinear form) on each pair of local spaces.* However, there is an obvious and crucial difference between the SPD Schwarz framework and our nonsymmetric and indefinite Schwarz framework. When $V \neq W$, our framework requires space decompositions for both spaces V and W, and these two space decompositions must be chosen compatibly in the sense to be described below.

Let

$$V_j \subset X, \qquad W_j \subset X \qquad \text{for } j = 0, 1, 2, \cdots, J,$$

be two sets of reflexive Banach spaces with norms $\|\cdot\|_{\mathrm{V}_j}$ and $\|\cdot\|_{\mathrm{W}_j}$ respectively. We note that V_0 and W_0 are used to denote the so-called *coarse spaces* in the domain decomposition context. For $j = 0, 1, 2, \cdots, J$, let

$$\mathcal{R}_j^{\dagger} : V_j \to V, \qquad \mathcal{S}_j^{\dagger} : W_j \to W$$

denote some *prolongation operators*.

Remark 5.3.3. *In the Schwarz method literature (cf. [77, 74, 81]), R_j^T is often used to denote both the prolongation operator from V_j to V and its matrix representation. Such a choice of notation is due to the fact that the matrix representation of the* not-explicitly-defined *restriction operator R_j from V to V_j is always chosen to be the*

transpose of the matrix representation of the prolongation operator. As expected, such a dual role notation may be confusing to some readers. To avoid such a potential confusion we use different notations for operators and their matrix representations throughout this chapter.

We also like to note that in the construction of all Schwarz methods the restriction operators/matrices are not "primary" operators/matrices but "derivative" operators/matrices in the sense that they are not chosen independently. Instead, they are determined by the prolongation operators/matrices. One often first defines the matrix representation of the (desired) restriction operator as the transpose of the the matrix representation of the prolongation operator and then defines the restriction operator to be the unique linear operator which has the chosen matrix representation (under the same bases in which the prolongation matrix is obtained). This will also be the approach adopted in this chapter for defining our restriction operators (see Definition 5.3.6). Clearly, such a definition of the restriction operators is not only abstract but also depends on the choices of the bases of the underlying function spaces. However, its simplicity and convenience at the matrix level make the definition appealing.

Suppose that the following relations hold:

$$\mathcal{R}_j^\dagger V_j \subsetneq V, \qquad \mathcal{S}_j^\dagger W_j \subsetneq W \quad \text{for } j = 0, 1, 2, \cdots, J, \tag{5.25}$$

$$V = \sum_{j=0}^{J} \mathcal{R}_j^\dagger V_j, \qquad W = \sum_{j=0}^{J} \mathcal{S}_j^\dagger W_j, \tag{5.26}$$

where $\mathcal{R}_j^\dagger V_j$ and $\mathcal{S}_j^\dagger W_j$ stand for the ranges of the linear operators $\mathcal{R}_j^\dagger$ and $\mathcal{S}_j^\dagger$ respectively.

Associated with each pair of local spaces (V_j, W_j) for $j = 0, 1, 2 \cdots, J$, we introduce a local discrete bilinear form $a_j(\cdot, \cdot)$ defined on $V_j \times W_j$, which can be taken either as the restriction of global discrete bilinear form $a(\cdot, \cdot)$ on $V_j \times W_j$ or as some approximation of the restriction of $a(\cdot, \cdot)$ on $V_j \times W_j$. We call these two choices

of local discrete bilinear form $a_j(\cdot,\cdot)$ an *exact local solver* and an *inexact local solver*, respectively. After the local discrete bilinear forms are chosen, we can define what constitutes as a *compatible space decomposition*.

Definition 5.3.4. *(i) A pair of spaces V_j and W_j are said to be compatible with respect to $a_j(\cdot,\cdot)$ if they satisfy the following conditions:*

(LA_1) Local continuity. *There exists a positive constant C_{a_j} such that*

$$|a_j(v,w)| \leq C_{a_j}\|v\|_{V_j}\|w\|_{W_j} \qquad \forall v \in V_j,\ w \in W_j. \tag{5.27}$$

(LA_2) Local weak coercivity. *There exist positive constants γ_{a_j} and β_{a_j} such that*

$$\sup_{w\in W_j} \frac{a_j(v,w)}{\|w\|_{W_j}} \geq \gamma_{a_j}\|v\|_{V_j} \qquad \forall v \in V_j, \tag{5.28}$$

$$\sup_{v\in V_j} \frac{a_j(v,w)}{\|v\|_{V_j}} \geq \beta_{a_j}\|w\|_{W_j} \qquad \forall w \in W_j. \tag{5.29}$$

(ii) A pair of space decompositions $\{V_j\}_{j=0}^J$ and $\{W_j\}_{j=0}^J$ of V and W satisfying (5.25)–(5.26) are said to be compatible if each pair of V_j and W_j is compatible with respect to $a_j(\cdot,\cdot)$ for $j = 0, 1, 2, \cdots, J$.

Obviously conditions (LA_1) and (LA_2) on $a_j(\cdot,\cdot)$ are the analogies of (MA_1) and (MA_2) on $a(\cdot,\cdot)$. By Theorem 5.2.3, these conditions guarantee that the local problem of seeking $u_j \in V_j$ such that

$$a_j(u_j, w_j) = f_j(w_j) \qquad \forall\ w_j \in W_j, \tag{5.30}$$

is uniquely solvable for any given bounded linear functional f_j on W_j. Moreover, (LA_1) and (LA_2) are "minimum" conditions for achieving such a guaranteed unique solvability (cf. Remark 5.2.4). Furthermore, like its global counterpart, the local weak coercivity condition (LA_2) induces the following two equivalent norms on V_j

and W_j:

$$\|v\|_{a_j} := \sup_{w \in W_j} \frac{a_j(v,w)}{\|w\|_{\mathrm{W_j}}} \qquad \forall v \in V_j, \tag{5.31}$$

$$\|w\|_{a_j^*} := \sup_{v \in V_j} \frac{a_j^*(w,v)}{\|v\|_{\mathrm{V_j}}} \qquad \forall w \in W_j, \tag{5.32}$$

where $a_j^*(w,v) := a_j(v,w)$ for any $(v,w) \in V_j \times W_j$.

Trivially, we have

Lemma 5.3.5. *Suppose that V_j and W_j are compatible with respect to $a_j(\cdot,\cdot)$. Then the following inequalities hold:*

$$\gamma_{a_j}\|v\|_{V_j} \leq \|v\|_{a_j} \leq C_{a_j}\|v\|_{V_j} \qquad \forall v \in V_j, \tag{5.33}$$

$$\beta_{a_j}\|w\|_{W_j} \leq \|w\|_{a_j^*} \leq C_{a_j}\|w\|_{W_j} \qquad \forall w \in W_j. \tag{5.34}$$

5.3.3 Additive Schwarz Method

Throughout this section, we assume that we are given a global discrete problem (5.7), and the global discrete bilinear form $a(\cdot,\cdot)$ fulfills the main assumptions (MA_1) and (MA_2) so that problem (5.7) has a unique solution $u \in V$. In addition, we assume we are given a pair of space decompositions $\{V_j\}_{j=0}^J$ and $\{W_j\}_{j=0}^J$ of V and W, the prolongation operators $\{\mathcal{R}_j^\dagger\}_{j=0}^J$ and $\{\mathcal{S}_j^\dagger\}_{j=0}^J$, and the local discrete bilinear forms $\{a_j(\cdot,\cdot)\}_{j=0}^J$ such that the given space decompositions are compatible with respect to the given local discrete bilinear forms in the sense of Definition 5.3.4. Our goal in this subsection is to construct the additive Schwarz method for problem (5.7) using the given information.

To continue, we now introduce two sets of *projection-like* operators $\widetilde{\mathcal{P}}_j : V \to V_j$ and $\widetilde{\mathcal{Q}}_j : W \to W_j$ for $j = 0, 1, 2, \cdots, J$. These projection-like-operators will serve as the building blocks for the constructions of both our additive and multiplicative Schwarz methods. For any fixed $v \in V$ and $w \in W$, define $\widetilde{\mathcal{P}}_j v \in V_j$ and $\widetilde{\mathcal{Q}}_j w \in W_j$

by

$$a_j\big(\widetilde{\mathcal{P}}_j v, w_j\big) := a\big(v, \mathcal{S}_j^\dagger w_j\big) \qquad \forall w_j \in W_j, \tag{5.35}$$

$$a_j^*\big(\widetilde{\mathcal{Q}}_j w, v_j\big) := a^*\big(w, \mathcal{R}_j^\dagger v_j\big) \qquad \forall v_j \in V_j. \tag{5.36}$$

We recall that $a_j^*(w_j, v_j) = a_j(v_j, w_j)$ for all $v_j \in V_j$ and $w_j \in W_j$. We also note that since V_j and W_j are assumed to be compatible, Theorem 5.2.3 then ensures both $\widetilde{\mathcal{P}}_j$ and $\widetilde{\mathcal{Q}}_j$ are well defined for $j = 0, 1, \cdots, J$.

Since V_j and W_j may not be subspaces of V and W, $\widetilde{\mathcal{P}}_j v$ and $\widetilde{\mathcal{Q}}_j w$ may not belong to V and W. To pull them back to the global discrete spaces V and W, we appeal to the *prolongation operators* $\mathcal{R}_j^\dagger$ and $\mathcal{S}_j^\dagger$ for help. Define the composite operators

$$\mathcal{P}_j := \mathcal{R}_j^\dagger \circ \widetilde{\mathcal{P}}_j, \qquad \mathcal{Q}_j := \mathcal{S}_j^\dagger \circ \widetilde{\mathcal{Q}}_j \qquad \text{for } j = 0, 1, 2, \cdots, J. \tag{5.37}$$

Trivially, we have $\mathcal{P}_j : V \to V$ and $\mathcal{Q}_j : W \to W$ for $j = 0, 1, 2, \cdots, J$.

We now are ready to define the following additive Schwarz operators. Following [31, 81, 74, 77] we define

$$\mathcal{P}_{\text{ad}} := \mathcal{P}_0 + \mathcal{P}_1 + \mathcal{P}_2 + \cdots + \mathcal{P}_J, \tag{5.38}$$

$$\mathcal{Q}_{\text{ad}} := \mathcal{Q}_0 + \mathcal{Q}_1 + \mathcal{Q}_2 + \cdots + \mathcal{Q}_J. \tag{5.39}$$

The matrix interpretation of the additive operator $\mathcal{P}_{\text{ad}}$ is similar to but slightly more complicated than the one in the SPD Schwarz framework. In particular, the additive operator $\mathcal{Q}_{\text{ad}}$ does not exist in the the SPD framework. For the reader's convenience, we give below a brief matrix interpretation for both $\mathcal{P}_{\text{ad}}$ and $\mathcal{Q}_{\text{ad}}$.

Fixing a basis for each of V, W, V_j and W_j, let A and A_j denote respectively the global and local stiffness matrices of the bilinear forms $a(\cdot,\cdot)$ and $a_j(\cdot,\cdot)$ with respect to the given bases. Let $R_j^\dagger, S_j^\dagger, \widetilde{P}_j, \widetilde{Q}_j, P_j, Q_j, P_{\text{ad}}$ and Q_{ad} denote the matrix representations of the linear operators $\mathcal{R}_j^\dagger, \mathcal{S}_j^\dagger, \widetilde{\mathcal{P}}_j, \widetilde{\mathcal{Q}}_j, \mathcal{P}_j, \mathcal{Q}_j, \mathcal{P}_{\text{ad}}$ and $\mathcal{Q}_{\text{ad}}$ with respect

to the given bases. Lastly, let $A^T, A_j^T, R_j^{\dagger T}$ and $S_j^{\dagger T}$ denote the matrix transposes of $A, A_j, R_j^\dagger$ and $S_j^\dagger$.

Using the above notation and the well-known fact that composite linear operators are represented by matrix multiplications, we obtain from (5.35) and (5.36) that

$$A_j \widetilde{P}_j \mathbf{v} := S_j^{\dagger T} A \mathbf{v} \qquad \forall \mathbf{v} \in \mathbf{R}^n, \tag{5.40}$$

$$A_j^T \widetilde{Q}_j \mathbf{w} := R_j^{\dagger T} A^T \mathbf{w} \qquad \forall \mathbf{w} \in \mathbf{R}^n. \tag{5.41}$$

Thus,

$$\widetilde{P}_j = A_j^{-1} S_j^{\dagger T} A, \qquad P_j = R_j^\dagger A_j^{-1} S_j^{\dagger T} A, \tag{5.42}$$

$$\widetilde{Q}_j = A_j^{-T} R_j^{\dagger T} A^T, \qquad Q_j = S_j^\dagger A_j^{-T} R_j^{\dagger T} A^T, \tag{5.43}$$

where A_j^{-1} and A_j^{-T} denote the inverse matrices of A_j and A_j^T, respectively. We also note that the compatibility assumptions (LA_1) and (LA_2) imply that A_j^{-1} and A_j^{-T} do exist.

Finally, it follows from (5.38), (5.39), (5.42) and (5.43) that

$$P_{\mathrm{ad}} = R_0^\dagger A_0^{-1} S_0^{\dagger T} A + \sum_{j=1}^{J} R_j^\dagger A_j^{-1} S_j^{\dagger T} A, \tag{5.44}$$

$$Q_{\mathrm{ad}} = S_0^\dagger A_0^{-T} R_0^{\dagger T} A^T + \sum_{j=1}^{J} S_j^\dagger A_j^{-T} R_j^{\dagger T} A^T. \tag{5.45}$$

From the above expressions, we obtain the following two additive Schwarz preconditioners for both A and its transpose A^T:

$$B := R_0^\dagger A_0^{-1} S_0^{\dagger T} + \sum_{j=1}^{J} R_j^\dagger A_j^{-1} S_j^{\dagger T}, \tag{5.46}$$

$$B^\dagger := S_0^\dagger A_0^{-T} R_0^{\dagger T} + \sum_{j=1}^{J} S_j^\dagger A_j^{-T} R_j^{\dagger T}. \tag{5.47}$$

It is interesting to note that $B^{\dagger} = B^T$ which means that the nonsymmetric Schwarz preconditioner B can be used to precondition both the linear system (5.7) and its adjoint system without any additional cost.

As it was already alluded to in Remark 5.3.3, we now formally define our restriction operators $\{\mathcal{R}_j\}$ and $\{\mathcal{S}_j\}$.

Definition 5.3.6. *For $j = 0, 1, 2, \cdots, J$, let $\mathcal{R}_j : V \to V_j$ (resp. $\mathcal{S}_j : W \to W_j$) be the unique linear operator whose matrix representation is given by $S_j^{\dagger T}$ (resp. $R_j^{\dagger T}$) under the same bases of V, W, V_j and W_j in which $R_j^{\dagger T}$ and $S_j^{\dagger T}$ are obtained.*

By the design, the matrix representations R_j and S_j of $\mathcal{R}_j$ and $\mathcal{S}_j$ satisfy $R_j = S_j^{\dagger T}$ and $S_j = R_j^{\dagger T}$.

5.3.4 Multiplicative Schwarz Method

The multiplicative Schwarz methods for solving problem (5.7) refer to various generalizations of the original Schwarz alternating iterative method (cf. [15, 81]). However, they also can be formulated as linear iterations on some preconditioned systems (cf. [77]). In this chapter we, adopt the latter point of view to present our nonsymmetric and indefinite multiplicative Schwarz methods. We shall use the same notation as in Subsection 5.3.3.

We first introduce the following two so-called *error propagation operators*:

$$\mathcal{E}_{\text{mu}} := (\mathcal{I} - \mathcal{P}_J) \circ (\mathcal{I} - \mathcal{P}_{J-1}) \circ \cdots \circ (\mathcal{I} - \mathcal{P}_0), \tag{5.48}$$

$$\mathcal{E}_{\text{sy}} := (\mathcal{I} - \mathcal{P}_0) \circ (\mathcal{I} - \mathcal{P}_1) \circ \cdots \circ (\mathcal{I} - \mathcal{P}_J) \circ (\mathcal{I} - \mathcal{P}_J) \circ \cdots \circ (\mathcal{I} - \mathcal{P}_0). \tag{5.49}$$

where $\mathcal{I}$ denotes the identity operator on V or on W. We then define the following two "preconditioned" operators:

$$\mathcal{P}_{\text{mu}} := \mathcal{I} - \mathcal{E}_{\text{mu}}, \qquad \mathcal{P}_{\text{sy}} := \mathcal{I} - \mathcal{E}_{\text{sy}}. \tag{5.50}$$

It is easy to check that the algebraic matrix representations of the above operators are, respectively,

$$E_{\text{mu}} := (I - P_J)(I - P_{J-1}) \cdots (I - P_0), \tag{5.51}$$

$$E_{\text{sy}} := (I - P_0)(I - P_1) \cdots (I - P_J)(I - P_J) \cdots (I - P_1)(I - P_0), \tag{5.52}$$

$$P_{\text{mu}} := I - E_{\text{mu}}, \tag{5.53}$$

$$P_{\text{sy}} := I - E_{\text{sy}}. \tag{5.54}$$

Then our multiplicative Schwarz iterative methods are defined as

$$\mathbf{u}^{(k+1)} = (I - C)\mathbf{u}^{(k)} + \mathbf{g} = E\mathbf{u}^{(k)} + \mathbf{g}, \qquad k \geq 0 \tag{5.55}$$

where (C, E) are either $(P_{\text{mu}}, E_{\text{mu}})$ or $(P_{\text{sy}}, E_{\text{sy}})$, and $\mathbf{g}$ takes either $\mathbf{g}_{\text{mu}} \in \mathbf{R}^n$ or $\mathbf{g}_{\text{sy}} \in \mathbf{R}^n$ which are easily computable from $\mathbf{f}$ in (5.8).

Remark 5.3.7. *(a) Clearly, the case with the triple $(P_{mu}, E_{mu}, \mathbf{g}_{mu})$ corresponds to the classical multiplicative Schwarz method for (5.8) (cf. [15]).*

(b) The case with the triple $(P_{sy}, E_{sy}, \mathbf{g}_{sy})$ can be regarded as a "symmetrized" multiplicative Schwarz method for nonsymmetric and indefinite problems. However, we note that the operator $\mathcal{E}_{sy}$ and matrix E_{sy} are not symmetric in general because $\{\mathcal{P}_j\}$ and $\{P_j\}$ may not be symmetric.

(c) Unlike in the SPD case, the norm $\|\mathcal{E}_{mu}\|_a$ could be larger than 1 for convection-dominant problems as shown by the numerical tests given in Section 5.5, although the multiplicative Schwarz method appears to be convergent in all those tests. Consequently, the convergent behavior of the multiplicative Schwarz method presented above is more complicated than its SPD counterpart.

5.3.5 A Hybrid Schwarz Method

In this subsection, we consider a hybrid Schwarz method which combines the additive Schwarz idea (between subdomains) and the multiplicative Schwarz idea (between levels). The hybrid method is expected to take advantage of both additive and multiplicative Schwarz methods.

The iteration operator of our hybrid Schwarz method is given by

$$\mathcal{E}_{\text{hy}} := (\mathcal{I} - \alpha\mathcal{P}_0)(\mathcal{I} - \widehat{\mathcal{P}}), \qquad \text{where} \quad \widehat{\mathcal{P}} := \textstyle\sum_{j=1}^{J} \mathcal{P}_j, \tag{5.56}$$

$$\mathcal{G}_{\text{hy}} := (\mathcal{I} - \alpha\mathcal{Q}_0)(\mathcal{I} - \widehat{\mathcal{Q}}), \qquad \text{where} \quad \widehat{\mathcal{Q}} := \textstyle\sum_{j=1}^{J} \mathcal{Q}_j. \tag{5.57}$$

Thus, the "preconditioned" hybrid Schwarz operator has the following form:

$$\mathcal{P}_{\text{hy}} := \mathcal{I} - \mathcal{E}_{\text{hy}} = \alpha\mathcal{P}_0 + (\mathcal{I} - \alpha\mathcal{P}_0)\widehat{\mathcal{P}}, \tag{5.58}$$

$$\mathcal{Q}_{\text{hy}} := \mathcal{I} - \mathcal{G}_{\text{hy}} = \alpha\mathcal{Q}_0 + (\mathcal{I} - \alpha\mathcal{Q}_0)\widehat{\mathcal{Q}}, \tag{5.59}$$

where α, called a relaxation parameter, is an undetermined positive constant.

Since the corresponding matrix representations of $\mathcal{E}_{\text{hy}}, \mathcal{P}_{\text{hy}}, \mathcal{G}_{\text{hy}}$, and $\mathcal{Q}_{\text{hy}}$ are easy to write down, we omit them to save space.

5.4 An Abstract Schwarz Preconditioner Theory for Nonsymmetric and Indefinite Problems

In this section, we shall first establish condition number estimates for additive Schwarz operator $\mathcal{P}_{\text{ad}}$ and for its matrix representation P_{ad}. We then present a condition number estimate for the hybrid operator $\mathcal{P}_{\text{hy}}$.

5.4.1 Structure Assumptions

Our preconditioner theory rests on the following structure assumptions. The validity of these structure assumptions is dependent on the numerical discretization method that is being implemented along with the choice of space decomposition, local solvers, and prolongation operators that are made. These choices must be made carefully in order to ensure a good Schwarz preconditioner is obtained.

(SA_0) *Compatibility assumption.* Assume that $\{(V_j, W_j)\}_{j=0}^J$ is a compatible decomposition of (V, W) in the sense of Definition 5.3.4.

(SA_1) *Energy stable decomposition assumption.* There exist positive constants C_{v} and C_{w} such that every pair $(v, w) \in V \times W$ admits a decomposition

$$v = \sum_{j=0}^{J} \mathcal{R}_j^\dagger v_j, \qquad w = \sum_{j=0}^{J} \mathcal{S}_j^\dagger w_j,$$

with $v_j \in V_j$ and $w_j \in W_j$ such that

$$\sum_{j=0}^{J} \|v_j\|_{a_j} \leq C_{\mathrm{v}} \|v\|_a, \tag{5.60}$$

$$\sum_{j=0}^{J} \|w_j\|_{\mathrm{w}_j} \leq C_{\mathrm{w}} \|w\|_{\mathrm{w}}. \tag{5.61}$$

(SA_2) *Strengthened generalized Cauchy-Schwarz inequality assumption.* There exist constants $\theta_{ij} \in [0, 1]$ for $i, j = 0, 1, 2, \cdots, J$ such that

$$a(\mathcal{R}_i^\dagger v_i, \mathcal{S}_j^\dagger w_j) \leq \theta_{ij} \|\mathcal{R}_i^\dagger v_i\|_a \|\mathcal{S}_j^\dagger w_j\|_{\mathrm{w}} \qquad \forall v_i \in V_i,\, w_j \in W_j. \tag{5.62}$$

(SA$_3$) *Local stability assumption.* There exist positive constants $\omega_{\rm V}$ and $\omega_{\rm W}$ such that for $j = 0, 1, 2, \cdots, J$

$$\|\mathcal{R}_j^\dagger v_j\|_a \le \omega_{\rm V} \|v_j\|_{a_j} \qquad \forall v_j \in V_j, \tag{5.63}$$

$$\|\mathcal{S}_j^\dagger w_j\|_{\rm W} \le \omega_{\rm W} \|w_j\|_{{\rm W}_j} \qquad \forall w_j \in W_j. \tag{5.64}$$

(SA$_4$) *Approximability assumption.* There exist (small) positive constants $\delta_{\rm V}, \widehat{\delta}_{\rm V}, \delta_{\rm W}$ and $\widehat{\delta}_{\rm W}$ such that for $i = 0, 1, 2, \cdots, J$ and $j = 1, 2, \cdots, J$

$$\|v - \mathcal{P}_0 v\|_a \le \delta_{\rm V} \|v\|_a \qquad \forall v \in V, \tag{5.65}$$

$$\|v - \widehat{\mathcal{P}} v\|_a \le \widehat{\delta}_{\rm V} \|v\|_a \qquad \forall v \in V, \tag{5.66}$$

$$\|w - \mathcal{Q}_0 w\|_{{\rm W}_0} \le \delta_{\rm W} \|w\|_{\rm W} \qquad \forall w \in W, \tag{5.67}$$

$$\|w - \widehat{\mathcal{Q}} w\|_{{\rm W}_0} \le \widehat{\delta}_{\rm W} \|w\|_{\rm W} \qquad \forall w \in W, \tag{5.68}$$

where $\widehat{\mathcal{P}} := \sum_{i=1}^J \mathcal{P}_i$ and $\widehat{\mathcal{Q}} := \sum_{i=1}^J \mathcal{Q}_i$.

We now explain the rationale and motivation of each assumption listed above.

Remark 5.4.1. *(a) We note that $\|\cdot\|_a$ and $\|\cdot\|_{a^*}$ are defined in (5.21) and (5.22), and $\|\cdot\|_{a_j}$ and $\|\cdot\|_{a_j^*}$ are defined in (5.31) and (5.32).*

(b) For a given compatible pair of space decompositions $\{(V_j, W_j)\}_{j=0}^J$, decompositions of each function $v \in V$ and $w \in W$ may not be unique. Assumption (SA$_1$) assumes that there exists at least one decomposition which is energy stable for every function in V and W. It imposes a constraint on both the choice of the space decompositions $\{(V_j, W_j)\}_{j=0}^J$ and on the choice of the local bilinear forms $\{a_j(\cdot,\cdot)\}_{j=0}^J$.

(c) We note that different norms are used for two functions on the right-hand side of (5.62), and θ_{ij} is defined for $i, j = 0, 1, 2, \cdots, J$. We set $\Theta = [\theta_{ij}]_{i,j=0}^J$ and note that Θ is a $(J+1) \times (J+1)$ matrix. We shall also use the submatrix $\widehat{\Theta} := [\theta_{ij}]_{i,j=1}^J$ in our analysis to be given in Section 5.4. Since the bilinear form $a(\cdot,\cdot)$ is not an inner product, the standard Cauchy-Schwarz inequality does not hold in general. But it does

hold in this generalized sense with $\theta_{ij} = 1$, *see Lemma 5.4.2. Moreover, we expect that each pair* (V_j, W_j) *for* $1 \le i, j \le J$ *only interacts with very few remaining pairs in the space decomposition* $\{(V_j, W_j)\}_{j=1}^{J}$. *Hence, the matrix* $\widehat{\Theta}$, *which is symmetric, is expected to be sparse and nearly diagonal in most applications. On the other hand, we expect that* $\theta_{0j} = \theta_{i0} = 1$ *for* $i, j = 1, 2, \cdots, J$.

(d) Local stability assumption (SA_3) *imposes a condition on the choice of the prolongation operators* $\mathcal{R}_j^\dagger$ *and* $\mathcal{S}_j^\dagger$. *It requires that these operators are bounded operators.*

(e) Assumption (SA_4), *which does not appear in the SPD theory, imposes a local approximation condition on the projection-like operators* $\{\widetilde{\mathcal{P}}_j\}$ *and* $\{\widetilde{\mathcal{Q}}_j\}$. *Consequently it imposes conditions on the prolongation operators* $\{\mathcal{R}_j^\dagger\}, \{\mathcal{S}_j^\dagger\}$ *and the local solvers* $a_j(\cdot, \cdot)$.

(f) Because of the norm equivalence properties (5.23), (5.24), (5.33) and (5.34), one can easily replace the weak coercivity induced norms by their equivalent underlying space norms or vice versa in all assumptions (SA_1)*–*(SA_4). *However, one must track all the constants resulting from the changes. The main reason for using the current forms of the assumptions is that they allow us to give a cleaner presentation of our nonsymmetric and indefinite Schwarz preconditioner theory to be described below.*

5.4.2 Condition Number Estimate for $\mathcal{P}_{\mathbf{ad}}$

First, we state the following simple lemma.

Lemma 5.4.2. *The following generalized Cauchy-Schwarz inequalities hold:*

$$a(v, w) \le \|v\|_a \|w\|_W \qquad \forall v \in V,\ w \in W, \tag{5.69}$$

$$a(v, w) \le \|v\|_V \|w\|_{a^*} \qquad \forall v \in V,\ w \in W, \tag{5.70}$$

$$a_j(v_j, w_j) \le \|v_j\|_{a_j} \|w\|_{W_j} \qquad \forall v_j \in V_j,\ w_j \in W_j,\ j = 0, 1, \cdots, J, \tag{5.71}$$

$$a_j(v_j, w_j) \le \|v\|_{V_j} \|w_j\|_{a_j^*} \qquad \forall v_j \in V_j,\ w_j \in W_j,\ j = 0, 1, \cdots, J. \tag{5.72}$$

Proof. (5.69)–(5.72) are immediate consequences of the definitions of the norms $\|\cdot\|_a, \|\cdot\|_{a^*}$, $\|\cdot\|_{a_j}$ and $\|\cdot\|_{a_j^*}$. □

Lemma 5.4.3. *Under assumptions (SA_0) and (SA_3), the following estimates hold:*

$$\|\widetilde{\mathcal{P}}_j v\|_{a_j} \le \omega_W \|v\|_a \qquad \forall v \in V, \quad j = 0, 1, \cdots, J, \tag{5.73}$$

$$\|\mathcal{P}_j v\|_a \le \omega_V \omega_W \|v\|_a \qquad \forall v \in V, \quad j = 0, 1, \cdots, J, \tag{5.74}$$

$$\|\widetilde{\mathcal{Q}}_j w\|_{a_j^*} \le \omega_V C_{a_j} \beta_a^{-1} \|w\|_{a^*} \qquad \forall w \in W, \quad j = 0, 1, \cdots, J, \tag{5.75}$$

$$\|\mathcal{Q}_j w\|_{a^*} \le \omega_V \omega_W C_a C_{a_j} \beta_a^{-1} \beta_{a_j}^{-1} \|w\|_{a^*} \qquad \forall w \in W, \quad j = 0, 1, \cdots, J, \tag{5.76}$$

$$\|\mathcal{P}_j v\|_V \le \omega_V \omega_W C_a \gamma_a^{-1} \|v\|_V \qquad \forall v \in V, \quad j = 0, 1, \cdots, J, \tag{5.77}$$

$$\|\mathcal{Q}_j w\|_W \le \omega_V \omega_W C_{a_j} \beta_{a_j}^{-1} \|w\|_W \qquad \forall w \in W, \quad j = 0, 1, \cdots, J. \tag{5.78}$$

Proof. For any $v \in V$, by assumption (SA_3) and Lemma 5.4.2 we get for $j = 0, 1, \cdots, J$,

$$\begin{aligned}
\|\widetilde{\mathcal{P}}_j v\|_{a_j} &= \sup_{w_j \in W_j} \frac{a_j(\widetilde{\mathcal{P}}_j v, w_j)}{\|w_j\|_{\mathrm{W}_j}} \\
&= \sup_{w_j \in W_j} \frac{a(v, \mathcal{S}_j^\dagger w_j)}{\|w_j\|_{\mathrm{W}_j}} && \text{(by (5.35))} \\
&\le \sup_{w_j \in W_j} \frac{\|v\|_a \, \|\mathcal{S}_j^\dagger w_j\|_{\mathrm{W}}}{\|w_j\|_{\mathrm{W}_j}} && \text{(by (5.69))} \\
&\le \omega_{\mathrm{W}} \|v\|_a. && \text{(by (5.64))}
\end{aligned} \tag{5.79}$$

Hence, (5.73) holds. (5.74) follows immediately from (5.73) and (5.63). By assumption (SA_3) and Lemma 5.4.2 we obtain

$$\begin{aligned}
\|\widetilde{\mathcal{Q}}w\|_{a_j^*} &= \sup_{v_j\in V_j}\frac{a_j(v_j,\widetilde{\mathcal{Q}}w)}{\|v_j\|_{\mathrm{V}_j}} \\
&= \sup_{v_j\in V_j}\frac{a(\mathcal{R}_j^{\dagger}v_j,w)}{\|v_j\|_{\mathrm{V}_j}} && \text{(by (5.36))} \\
&\le \sup_{v_j\in V_j}\frac{\|\mathcal{R}_j^{\dagger}v_j\|_a\|w\|_{\mathrm{W}}}{\|v_j\|_{\mathrm{V}_j}} && \text{(by (5.69))} \\
&\le \sup_{v_j\in V_j}\frac{\omega_{\mathrm{V}}\|v_j\|_{a_j}\|w\|_{\mathrm{W}}}{\|v_j\|_{\mathrm{V}_j}} && \text{(by (5.63))} \\
&\le \omega_{\mathrm{V}}C_{a_j}\|w\|_{\mathrm{W}} && \text{(by (5.33))} \\
&\le \omega_{\mathrm{V}}C_{a_j}\beta_a^{-1}\|w\|_{a^*}. && \text{(by (5.24))}
\end{aligned}$$

Hence, (5.75) holds. (5.76) follows from (5.75), (5.23), (5.64), and (5.34). From the proof for (5.75) we can obtain $\|\widetilde{\mathcal{Q}}_j w\|_{a_j^*}\le\omega_{\mathrm{V}}C_{a_j}\|w\|_W$. This result along with (5.64) and (5.34) yields (5.78). The proof is complete. □

We now are ready to give an upper bound estimate for the additive Schwarz operator $\mathcal{P}_{\mathrm{ad}}$.

Proposition 5.4.4. *Under assumptions (SA_0)–(SA_3) the following estimate holds:*

$$\|\mathcal{P}_{ad}v\|_a\le\omega_V\omega_W\big[1+\omega_W C_W N(\Theta)\big]\|v\|_a\qquad\forall v\in V, \tag{5.80}$$

where $\Theta=[\theta_{ij}]_{i,j=0}^{J}$, $N(\Theta)=\max\{N_j(\Theta);\,0\le j\le J\}$ and $N_j(\Theta)$ denotes the number of nonzero entries in the vector $\Theta_j:=[\theta_{ij}]_{i=0}^{J}$, i.e., the number of nonzero entries of the jth column of the matrix Θ.

Proof. For any $w\in W$, let $\{w_j\}$ be an energy stable decomposition of w as defined in (SA_1). By the definition of $\mathcal{P}_{\mathrm{ad}}$, (5.69), (5.62), (5.74), (5.64), and (5.61) we get for

any $v \in V$

$$
\begin{aligned}
a(\mathcal{P}_{\mathrm{ad}} v, w) &= a(\mathcal{P}_0 v, w) + \sum_{i=1}^{J} a(\mathcal{P}_i v, w) \qquad (5.81)\\
&= a(\mathcal{P}_0 v, w) + \sum_{i=1}^{J}\sum_{j=0}^{J} a(\mathcal{R}_i^{\dagger}\widetilde{\mathcal{P}}_i v, \mathcal{S}_j^{\dagger} w_j)\\
&\le \|\mathcal{P}_0 v\|_a \|w\|_{\mathrm{W}} + \sum_{i=1}^{J}\sum_{j=0}^{J} \theta_{ij} \|\mathcal{P}_i v\|_a \|\mathcal{S}_j^{\dagger} w_j\|_{\mathrm{W}}\\
&\le \omega_{\mathrm{V}}\omega_{\mathrm{W}} \|v\|_a \Big\{ \|w\|_{\mathrm{W}} + \sum_{j=0}^{J} N_j(\Theta) \|\mathcal{S}_j^{\dagger} w_j\|_{\mathrm{W}} \Big\}\\
&\le \omega_{\mathrm{V}}\omega_{\mathrm{W}} \|v\|_a \Big\{ \|w\|_{\mathrm{W}} + \omega_{\mathrm{W}} N(\Theta) \sum_{j=0}^{J} \|w_j\|_{\mathrm{W}_j} \Big\}\\
&\le \omega_{\mathrm{V}}\omega_{\mathrm{W}} \|v\|_a \Big\{ \|w\|_{\mathrm{W}} + \omega_{\mathrm{W}} N(\Theta) C_{\mathrm{W}} \|w\|_{\mathrm{W}} \Big\}\\
&= \omega_{\mathrm{V}}\omega_{\mathrm{W}} \big[1 + \omega_{\mathrm{W}} C_{\mathrm{W}} N(\Theta)\big] \|v\|_a \|w\|_{\mathrm{W}}.
\end{aligned}
$$

Hence, (5.80) holds. The proof is complete. □

As expected, it is harder to get a lower bound estimate for the additive Schwarz operator $\mathcal{P}_{\mathrm{ad}}$. Such a bound then readily provides an upper bound for $\mathcal{P}_{\mathrm{ad}}^{-1}$. To this end, we first establish the following key lemma.

Lemma 5.4.5. *(i) Suppose that for every $v \in V$, $\{\widetilde{\mathcal{P}}_j v; j = 0, 1, 2, \cdots, J\}$ forms a stable decomposition of $\mathcal{P}_{ad} v$. Then under assumptions (SA_0) and (SA_1) the following inequality holds:*

$$
\sum_{j=0}^{J} \|\widetilde{\mathcal{P}}_j v\|_{a_j} \le C_V \|\mathcal{P}_{ad} v\|_a \qquad \forall v \in V. \qquad (5.82)
$$

(ii) If the condition of (i) does not hold, then under assumptions (SA_0)–(SA_4) we have

$$\sum_{j=0}^{J} \|\widetilde{\mathcal{P}}_j v\|_{a_j} \leq \frac{\omega_W}{2}(J+1)\Big[\|\mathcal{P}_{ad} v\|_a + \big(\delta_V + \widehat{\delta}_V\big)\|v\|_a\Big]. \tag{5.83}$$

Proof. (i) For any $v \in V$, let $u = \mathcal{P}_{\rm ad} v$, $u_j = \widetilde{\mathcal{P}}_j v$ for $j = 0, 1, 2, \cdots, J$. Since

$$u = \mathcal{P}_{\rm ad} v = \sum_{j=0}^{J} \mathcal{P}_j v = \sum_{j=0}^{J} \mathcal{R}_j^{\dagger} \circ \widetilde{\mathcal{P}}_j v = \sum_{j=0}^{J} \mathcal{R}_j^{\dagger} u_j,$$

$\{u_j\}$ is indeed a decomposition of u which is assumed to be stable. By assumption (SA_1) we conclude that (5.60) holds for u, which gives (5.82).

(ii) Let u be same as in part (i). Recall that $\widehat{\mathcal{P}} = \sum_{j=1}^{J} \mathcal{P}_j$. Using the identity

$$\widetilde{\mathcal{P}}_j v = \frac{1}{2}\Big[\widetilde{\mathcal{P}}_j u + \widetilde{\mathcal{P}}_j(v - \mathcal{P}_0 v) + \widetilde{\mathcal{P}}_j(v - \widehat{\mathcal{P}} v)\Big] \qquad \text{for } j = 0, 1, \cdots, J,$$

the triangle inequality, (SA_4) and (5.73) we get

$$\begin{aligned} \|\widetilde{\mathcal{P}}_j v\|_{a_j} &\leq \frac{1}{2}\Big[\|\widetilde{\mathcal{P}}_j u\|_{a_j} + \|\widetilde{\mathcal{P}}_j\big(v - \mathcal{P}_0 v\big)\|_{a_j} + \|\widetilde{\mathcal{P}}_j\big(v - \widehat{\mathcal{P}} v\big)\|_{a_j}\Big] \\ &\leq \frac{\omega_{\rm W}}{2}\Big[\|u\|_a + \big(\delta_{\rm V} + \widehat{\delta}_{\rm V}\big)\|v\|_a\Big] \qquad \text{for } j = 0, 1, \cdots, J. \end{aligned}$$

Then summing the above inequality we obtain

$$\sum_{j=0}^{J} \|\widetilde{\mathcal{P}}_j v\|_{a_j} \leq \frac{\omega_{\rm W}}{2}(J+1)\Big[\|u\|_a + \big(\delta_{\rm V} + \widehat{\delta}_{\rm V}\big)\|v\|_a\Big].$$

Hence, (5.83) holds. The proof is complete. □

We now are ready to establish a lower bound estimate for the additive Schwarz operator $\mathcal{P}_{\rm ad}$.

Proposition 5.4.6. *(i) Under the assumptions of (i) of Lemma 5.4.5, the following estimate holds:*

$$\|\mathcal{P}_{ad}v\|_a \geq (C_V C_W)^{-1}\|v\|_a \qquad \forall v \in V. \tag{5.84}$$

(ii) Under the assumptions of (ii) of Lemma 5.4.5, the following estimate holds:

$$\|\mathcal{P}_{ad}v\|_a \geq K_0^{-1}\|v\|_a \qquad \forall v \in V, \tag{5.85}$$

provided that $C_W\big(J+1\big)\big(\delta_V + \widehat{\delta}_V\big) < 1$ *where*

$$K_0 := \frac{\omega_W C_W}{2 - 2C_W\big(J+1\big)\big(\delta_V + \widehat{\delta}_V\big)}. \tag{5.86}$$

Consequently, operator $\mathcal{P}_{ad}$ *is invertible.*

Proof. For any $w \in W$, let $\{w_j\}$ be an energy stable decomposition of w, that is,

$$w = \sum_{j=0}^{J} \mathcal{S}_j^\dagger w_j,$$

and (5.61) holds. Then we have

$$\begin{aligned}
a(v,w) &= \sum_{j=0}^{J} a(v, \mathcal{S}_j^\dagger w_j) && (5.87)\\
&= \sum_{j=0}^{J} a_j(\widetilde{\mathcal{P}}_j v, w_j) && \text{(by (5.35))}\\
&\leq \sum_{j=0}^{J} \|\widetilde{\mathcal{P}}_j v\|_{a_j}\|w_j\|_{\mathrm{w}_j} && \text{(by (5.71))}\\
&\leq \sum_{j=0}^{J} \|\widetilde{\mathcal{P}}_j v\|_{a_j} \sum_{j=0}^{J} \|w_j\|_{\mathrm{w}_j} && \text{(by discrete Schwarz inequality)}\\
&\leq C_{\mathrm{w}}\|w\|_{\mathrm{w}} \sum_{j=0}^{J} \|\widetilde{\mathcal{P}}_j v\|_{a_j}. && \text{(by (5.61))}
\end{aligned}$$

The desired estimates (5.84) and (5.85) follow from substituting (5.82) and (5.83) into (5.87), respectively. The proof is complete. □

Remark 5.4.7. *We note that the argument used in the proof of lower bound estimate* (5.85) *is in the spirit of the so-called Schatz argument (cf. [16]) which is often used to derive finite element error estimates for nonsymmetric and indefinite problems. It is interesting to see that a similar argument also plays an important role in our Schwarz preconditioner theory.*

Combining Propositions 5.4.4 and 5.4.6 we obtain our first main theorem of this chapter.

Theorem 5.4.8. *(i) If for every* $v \in V$, $\{\widetilde{\mathcal{P}}_j v; j = 0, 1, 2, \cdots, J\}$ *forms a stable decomposition of* $\mathcal{P}_{ad} v$, *then under assumptions* (SA_0)*–*(SA_3) *the following condition number estimate holds:*

$$\kappa_a(\mathcal{P}_{ad}) \leq \omega_V \omega_W C_V C_W \big[1 + \omega_W C_W N(\Theta)\big]. \tag{5.88}$$

(ii) If the condition of (i) does not hold, then under assumptions (SA_0)*–*(SA_4) *the following condition number estimate holds:*

$$\kappa_a(\mathcal{P}_{ad}) \leq \omega_V \omega_W \big[1 + \omega_W C_W N(\Theta)\big] K_0. \tag{5.89}$$

Where

$$\kappa_a(\mathcal{P}_{ad}) := \|\mathcal{P}_{ad}\|_a \|\mathcal{P}_{ad}^{-1}\|_a, \tag{5.90}$$

$$\|\mathcal{P}_{ad}\|_a := \sup_{0 \neq v \in V} \frac{\|\mathcal{P}_{ad} v\|_a}{\|v\|_a}. \tag{5.91}$$

The above condition number estimates for the operator $\mathcal{P}_{\text{ad}}$ also translates to its matrix representation.

Theorem 5.4.9. *(i) Under assumptions of (i) of Theorem 5.4.8 the following condition number estimate holds:*

$$\kappa_A(P_{ad}) \le \omega_V \omega_W C_V C_W \big[1 + \omega_W C_W N(\Theta)\big]. \tag{5.92}$$

(ii) Under assumptions of (ii) of Theorem 5.4.8 the following condition number estimate holds:

$$\kappa_A(P_{ad}) \le \omega_V \omega_W \big[1 + \omega_W C_W N(\Theta)\big] K_0, \tag{5.93}$$

where

$$\kappa_A(P_{ad}) := \|P_{ad}\|_A \|P_{ad}^{-1}\|_A, \tag{5.94}$$

$$\|P_{ad}\|_A := \sup_{0 \neq \mathbf{v} \in \mathbf{R}^d} \frac{\|P_{ad}\mathbf{v}\|_A}{\|\mathbf{v}\|_A}, \tag{5.95}$$

$$\|\mathbf{v}\|_A := \sqrt{A\mathbf{v} \cdot A\mathbf{v}} = \sqrt{A^T A\mathbf{v} \cdot \mathbf{v}}. \tag{5.96}$$

Proof. Given bases for the spaces V and W, we can write $v \in V$ and $w \in W$ with vector representations $\mathbf{v} \in \mathbf{R}^n$ and $\mathbf{w} \in \mathbf{R}^n$, respectively. Also there exists $A \in \mathbf{R}^{n \times n}$ such that $a(v, w) = \mathbf{w}^T A\mathbf{v}$. If $\|w\|_W = \|\mathbf{w}\|_2$ then we get

$$\|v\|_a = \sup_{\mathbf{w} \in \mathbf{R}^n} \frac{\mathbf{w}^T A\mathbf{v}}{\|\mathbf{w}\|_2} \le \sup_{\mathbf{w} \in \mathbf{R}^n} \frac{\|\mathbf{w}\|_2 \|A\mathbf{v}\|_2}{\|\mathbf{w}\|_2} = \|A\mathbf{v}\|_2 = \|\mathbf{v}\|_A,$$

and

$$\|v\|_a = \sup_{\mathbf{w} \in \mathbf{R}^n} \frac{\mathbf{w}^T A\mathbf{v}}{\|\mathbf{w}\|_2} \ge \frac{(A\mathbf{v})^T A\mathbf{v}}{\|A\mathbf{v}\|_2} = \|\mathbf{v}\|_A$$

for $v \neq 0$. Thus, for $\|w\|_W = \|\mathbf{w}\|_2$, (5.92) and (5.93) are immediate consequences of (5.88) and (5.89), respectively. □

5.4.3 Condition Number Estimate for $\mathcal{P}_{\text{hy}}$

As in the case of SPD problems [77, section 2.5.2], we replace the structure assumption (SA_1) by the following one:

$(\widetilde{\text{SA}_1})$ *Energy stable decomposition assumption.* There exist positive constants $\widetilde{C}_{\text{v}}$ and $\widetilde{C}_{\text{w}}$ such that every pair $(\varphi, \psi) \in \text{range}(\mathcal{I} - \alpha\mathcal{P}_0) \times \text{range}(\mathcal{I} - \alpha\mathcal{Q}_0)$ admits a decomposition

$$\varphi = \sum_{j=1}^{J} \mathcal{R}_j^{\dagger}\varphi_j, \qquad \psi = \sum_{j=1}^{J} \mathcal{S}_j^{\dagger}\psi_j,$$

with $\varphi_j \in V_j$ and $\psi_j \in W_j$ such that

$$\sum_{j=1}^{J} \|\varphi_j\|_{a_j} \leq \widetilde{C}_{\text{v}} \|\varphi\|_a, \tag{5.97}$$

$$\sum_{j=1}^{J} \|\psi_j\|_{\text{w}_j} \leq \widetilde{C}_{\text{w}} \|\psi\|_{\text{w}}. \tag{5.98}$$

We remark that the new *energy stable decomposition assumption* $(\widetilde{\text{SA}_1})$ implies that any pair $(v, w) \in V \times W$ has a stable decomposition (in the sense of (SA_1)) of the following form:

$$v = \alpha\mathcal{P}_0 v + \sum_{j=1}^{J} \mathcal{R}_j^{\dagger}\varphi_j, \qquad w = \alpha\mathcal{Q}_0 w + \sum_{j=1}^{J} \mathcal{S}_j^{\dagger}\psi_j,$$

where $\{(\varphi_j, \psi_j)\}_{j=1}^{J}$ is a stable decomposition (in the sense of $(\widetilde{\text{SA}_1})$) for $\big((\mathcal{I} - \alpha\mathcal{P}_0)v, (\mathcal{I} - \alpha\mathcal{Q}_0)w\big)$.

Next lemma shows that $\mathcal{P}_j$ (resp. $\mathcal{P}_{\text{ad}}$) and $\mathcal{Q}_j$ (resp. $\mathcal{Q}_{\text{ad}}$) are mutually conjugate with respect to the bilinear form $a(\cdot, \cdot)$.

Lemma 5.4.10. *The following identities hold:*

$$a(\mathcal{P}_j v, w) = a(v, \mathcal{Q}_j w) \qquad \forall (v, w) \in V \times W,\ j = 0, 1, 2, \cdots, J, \tag{5.99}$$

$$a(\mathcal{P}_{ad} v, w) = a(v, \mathcal{Q}_{ad} w) \qquad \forall (v, w) \in V \times W. \tag{5.100}$$

Since the proof is trivial, we omit it to save space.

The following proposition is the analogue to Proposition 5.4.4 for the hybrid operator $\mathcal{P}_{\text{hy}}$.

Proposition 5.4.11. *Under assumptions (SA_0), ($\widetilde{SA}_1$), (SA_2) and (SA_3) the following estimate holds:*

$$\|\mathcal{P}_{hy} v\|_a \leq \omega_V \omega_W \big[\alpha + \omega_W \widetilde{C}_W N(\widehat{\Theta}) \big(1 + \alpha \omega_V \omega_W C_{a_j} \beta_{a_j}^{-1}\big)\big] \|v\|_a, \tag{5.101}$$

for all $v \in V$. Where $\widehat{\Theta} = [\theta_{ij}]_{i,j=1}^{J}$.

Proof. Let $\widehat{\mathcal{P}} := \sum_{j=1}^{J} \mathcal{P}_j$ and $\widehat{\mathcal{Q}} := \sum_{j=1}^{J} \mathcal{Q}_j$. For any $v \in V$ and $w \in W$, let $\varphi := (\mathcal{I} - \alpha \mathcal{P}_0) v$ and $\psi := (\mathcal{I} - \alpha \mathcal{Q}_0) w$. Obviously, $\varphi \in \text{range}(\mathcal{I} - \alpha \mathcal{P}_0)$ and $\psi \in \text{range}(\mathcal{I} - \mathcal{Q}_0)$. By assumption ($\widetilde{\text{SA}}_1$), (φ, ψ) admits an energy stable decomposition

$\{(\varphi_j, \psi_j)\}_{j=1}^J$. Thus,

$$
\begin{aligned}
a((\mathcal{I} - \alpha\mathcal{P}_0)\widehat{\mathcal{P}}v, w) &= a(\widehat{\mathcal{P}}v, (\mathcal{I} - \alpha\mathcal{Q}_0)w) \\
&= a(\widehat{\mathcal{P}}v, \psi) \\
&= \sum_{i=1}^{J}\sum_{j=1}^{J} a(\mathcal{R}_i^\dagger \widetilde{\mathcal{P}}_i v, \mathcal{S}_j^\dagger \psi_j) \\
&\leq \sum_{i=1}^{J}\sum_{j=1}^{J} \theta_{ij} \|\mathcal{P}_i v\|_a \|\mathcal{S}_j^\dagger \psi_j\|_{\mathrm{W}} \\
&\leq \omega_{\mathrm{V}}\omega_{\mathrm{W}} \|v\|_a \sum_{j=1}^{J} N_j(\widehat{\Theta}) \|\mathcal{S}_j^\dagger \psi_j\|_{\mathrm{W}} \\
&\leq \omega_{\mathrm{V}}\omega_{\mathrm{W}}^2 N(\widehat{\Theta}) \|v\|_a \sum_{j=1}^{J} \|\psi_j\|_{\mathrm{W}_j} \\
&\leq \omega_{\mathrm{V}}\omega_{\mathrm{W}}^2 \widetilde{C}_{\mathrm{W}} N(\widehat{\Theta}) \|v\|_a \|\psi\|_{\mathrm{W}} \\
&\leq \omega_{\mathrm{V}}\omega_{\mathrm{W}}^2 \widetilde{C}_{\mathrm{W}} N(\widehat{\Theta}) \big(1 + \alpha\omega_{\mathrm{V}}\omega_{\mathrm{W}} C_{a_j} \beta_{a_j}^{-1}\big) \|v\|_a \|w\|_{\mathrm{W}},
\end{aligned}
\tag{5.102}
$$

where we have used (5.78) to obtain the last inequality. The above inequality in turn implies that

$$
\|(\mathcal{I} - \alpha\mathcal{P}_0)\widehat{\mathcal{P}}v\|_a \leq \omega_{\mathrm{V}}\omega_{\mathrm{W}}^2 \widetilde{C}_{\mathrm{W}} N(\widehat{\Theta}) \big(1 + \alpha\omega_{\mathrm{V}}\omega_{\mathrm{W}} C_{a_j} \beta_{a_j}^{-1}\big) \|v\|_a,
$$

and

$$
\begin{aligned}
\|\mathcal{P}_{\mathrm{hy}} v\|_a &\leq \alpha \|\mathcal{P}_0 v\|_a + \|(\mathcal{I} - \alpha\mathcal{P}_0)\widehat{\mathcal{P}}v\|_a \\
&\leq \alpha\omega_{\mathrm{V}}\omega_{\mathrm{W}} + \omega_{\mathrm{V}}\omega_{\mathrm{W}}^2 \widetilde{C}_{\mathrm{W}} N(\widehat{\Theta}) \big(1 + \alpha\omega_{\mathrm{V}}\omega_{\mathrm{W}} C_{a_j} \beta_{a_j}^{-1}\big) \|v\|_a.
\end{aligned}
$$

Hence, (5.101) holds and the proof is complete. □

Next, we derive a lower bound estimate for $\|\mathcal{P}_{\mathrm{hy}}\|_a$. The following proposition is an analogue of Proposition 5.4.6.

Proposition 5.4.12. *Under assumptions (SA_0), ($\widetilde{SA}_1$), (SA_2)–(SA_4), along with the assumption range$\big(I - \alpha\mathcal{Q}_0\big) = W$ the following estimate holds:*

$$\|\mathcal{P}_{hy}v\|_a \geq K_1^{-1}\|v\|_a \qquad \forall v \in V, \tag{5.103}$$

provided that $\widehat{\delta}_{V}\big(\widetilde{C}_{W}\omega_{W} + \alpha\omega_{V}\omega_{W}\big) < 1$. Where

$$K_1 := \frac{1}{1 - \widehat{\delta}_{V}\big(\widetilde{C}_{W}\omega_{W} + \alpha\omega_{V}\omega_{W}\big)}. \tag{5.104}$$

Consequently, operator $\mathcal{P}_{hy}$ is invertible.

Proof. For any $v \in V$ and $w \in W$. Let $\psi := (\mathcal{I} - \alpha\mathcal{Q}_0)$, $w \in \text{range}(\mathcal{I} - \alpha\mathcal{Q}_0)$ and $u := \mathcal{P}_{\text{hy}}v$. Assumption ($\widetilde{\text{SA}}_1$) ensures that ψ has an energy stable decompositions $\{\psi_j\}_{j=1}^{J}$ with $\psi_j \in W_j$, that is,

$$\psi = \sum_{j=1}^{J} \mathcal{S}_j^{\dagger}\psi_j \qquad \text{and} \qquad \sum_{j=1}^{J} \|\psi_j\|_{\text{w}_j} \leq \widetilde{C}_{\text{w}}\|\psi\|_{\text{w}}. \tag{5.105}$$

Using the following identity

$$v = u + \big(v - \widehat{\mathcal{P}}v\ \big) + \alpha\mathcal{P}_0\big(\widehat{\mathcal{P}}v - v\big),$$

(SA$_4$), (5.63), (5.73) and (5.105) we get

$$\begin{aligned}
a(v,\psi) &= a(u,\psi) + a\big(v-\widehat{\mathcal{P}}v,\psi\big) + \alpha a\big(\mathcal{P}_0\big(\widehat{\mathcal{P}}v-v\big),\psi\big)\\
&\le \|u\|_a\|\psi\|_{\mathrm{w}} + \sum_{j=1}^{J} a\big(v-\widehat{\mathcal{P}}v,\mathcal{S}_j^{\dagger}\psi_j\big) + \alpha\big\|\mathcal{P}_0\big(\widehat{\mathcal{P}}v-v\big)\big\|_a\|\psi\|_{\mathrm{w}}\\
&\le \|u\|_a\|\psi\|_{\mathrm{w}} + \sum_{j=1}^{J} a_j\big(\widetilde{\mathcal{P}}_j\big(v-\widehat{\mathcal{P}}v\big),\psi_j\big) + \alpha\omega_{\mathrm{v}}\big\|\widetilde{\mathcal{P}}_0\big(\widehat{\mathcal{P}}v-v\big)\big\|_{a_0}\|\psi\|_{\mathrm{w}}\\
&\le \|u\|_a\|\psi\|_{\mathrm{w}} + \sum_{j=1}^{J} \omega_{\mathrm{w}}\widehat{\delta}_{\mathrm{v}}\|v\|_a\|\psi_j\|_{\mathrm{w}_j} + \alpha\omega_{\mathrm{v}}\omega_{\mathrm{w}}\widehat{\delta}_{\mathrm{v}}\|v\|_a\|\psi\|_{\mathrm{w}}\\
&\le \|u\|_a\|\psi\|_{\mathrm{w}} + \widehat{\delta}_{\mathrm{v}}\big(\widetilde{C}_{\mathrm{w}}\omega_{\mathrm{w}} + \alpha\omega_{\mathrm{v}}\omega_{\mathrm{w}}\big)\|v\|_a\|\psi\|_{\mathrm{w}}
\end{aligned}$$

The desired estimate follows from the assumption range$(I-\alpha\mathcal{Q}_0)=W$. □

Remark 5.4.13. *We note that the assumption* range $(\mathcal{I}-\alpha\mathcal{Q}_0)=W$ *is equivalent to asking* $\mathcal{I}-\alpha\mathcal{Q}_0$ *to be invertible, which holds for sufficiently small or large relaxation parameter* α*.*

Combining Propositions 5.4.11 and 5.4.12 we obtain our third main theorem of this chapter.

Theorem 5.4.14. *Under assumptions (SA$_0$), ($\widetilde{\text{SA}}_1$), (SA$_2$)–(SA$_4$) and range*$\big(I-\alpha\mathcal{Q}_0\big)=W$ *the following condition number estimate holds:*

$$\kappa_a(\mathcal{P}_{hy}) \le \omega_V\omega_W\big[\alpha+\omega_W\widetilde{C}_W N(\widehat{\Theta})\big(1+\alpha\omega_V\omega_W C_{a_j}\beta_{a_j}^{-1}\big)\big]\,K_1. \tag{5.106}$$

5.5 Application to DG Discretizations for Convection-diffusion Problems

In this section, we shall use our abstract framework and the abstract preconditioner theory developed in Sections 5.3 and 5.4 to construct three types of Schwarz

methods for discontinuous Galerkin approximations of the following general diffusion-convection problem:

$$\mathcal{L}u := -\mathrm{div}(\sigma(u)) + \gamma(x)u = f \qquad \text{in } \Omega, \tag{5.107}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{5.108}$$

where $\Omega \subset \mathbf{R}^d\,(d = 1, 2, 3)$ is a bounded domain with Lipschitz continuous boundary $\partial\Omega$ and $\sigma(u) := -D(x)\nabla u + \mathbf{b}(x)u$. $D(x) \in \mathbf{R}^{d\times d}$ satisfies $\lambda|\xi|^2 \le D(x)\xi\cdot\xi \le \Lambda|\xi|^2$ $\forall\xi \in \mathbf{R}^d$ for some positive constants λ and Λ. So (5.107) is uniformly elliptic in Ω [47, Chapter 8]. Assume that $\mathbf{b} \in H(\mathrm{div}, \Omega)$ or $\mathbf{b} \in [C^0(\overline{\Omega})]^d$, $\gamma \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$. Let $V = W = H_0^1(\Omega)$, then the variational formulation of (5.107)–(5.108) is defined as [9, 47]

$$\mathcal{A}(u, w) = \mathcal{F}(w) \qquad \forall w \in W, \tag{5.109}$$

where

$$\mathcal{A}(u, w) := \int_\Omega \Big(D(x)\nabla u\cdot\nabla w + \mathbf{b}(x)u\cdot\nabla w + \gamma(x)uw \Big)\, dx, \tag{5.110}$$

$$\mathcal{F}(w) := \int_\Omega fw\, dx. \tag{5.111}$$

Clearly, when $\mathbf{b}(x) \not\equiv 0$, the bilinear form $\mathcal{A}(\cdot,\cdot)$ is nonsymmetric. The problem can be further classified as follows:

(i) *Positive definite case:* If $\mathbf{b}$ and γ satisfies

$$\gamma(x) + \frac{1}{2}\mathrm{div}\mathbf{b}(x) \ge 0 \qquad \text{in } \Omega. \tag{5.112}$$

(ii) *Indefinite case:* If $\mathbf{b}$ and c satisfies

$$\gamma(x) + \frac{1}{2}\mathrm{div}\mathbf{b}(x) < 0 \qquad \text{in } \Omega. \tag{5.113}$$

It is easy to check that all the conditions of the classical Lax-Milgram Theorem hold in the *positive definite case*. It also can be shown [9] that in the *indefinite case* all the conditions of Theorem 5.2.1 are satisfied provided that problem (5.107)–(5.108) and its adjoint problem are uniquely solvable for arbitrary source terms. It is also well known [9, 47] that in *indefinite case* the bilinear form $\mathcal{A}(\cdot,\cdot)$ satisfies a Gärding-type inequality instead of strong coercivity.

5.5.1 Discontinuous Galerkin Approximations

Consider a special case of (5.107) where $D(x) = \epsilon > 0$ and $\mathbf{b} \in [W^{1,\infty}(\Omega)]^d$. To discretize this problem, we shall use an interior penalty discontinuous Galerkin (IPDG) scheme developed in [5]. For this scheme we require a shape-regular triangulation $\mathcal{T}_h$ of the domain Ω. The scheme can then be written in the form (5.7) where

$$V = W := \left\{ v \in L^2(\Omega) \text{ such that } v|_K \in \mathbb{P}_r(K)\ \forall K \in \mathcal{T}_h \right\}, \tag{5.114}$$

$$\begin{aligned} a(u,w) := &\sum_{K\in\mathcal{T}_h} \int_K (\gamma uw + (\epsilon\nabla u - \mathbf{b}u)\cdot\nabla w)\,dx + \sum_{e\notin\Gamma^+} c_e \frac{\epsilon}{|e|}\int_e [u]\cdot[w]\,ds \\ &+ \sum_{e\in\mathcal{E}_h^\circ}\int_e \{\mathbf{b}u\}_{upw}\cdot[w]\,ds - \sum_{e\notin\Gamma^+}\int_e \{\epsilon\nabla_h u\}\cdot[w]\,ds + \sum_{e\in\Gamma^+}\int_e \mathbf{b}\cdot\mathbf{n}uw\,ds, \end{aligned} \tag{5.115}$$

$$f(w) := \sum_{K\in\mathcal{T}_h}\int_K fw\,dx. \tag{5.116}$$

Where $r \geq 1, \Gamma = \partial\Omega$, $\mathbf{n}$ is the unit outward normal vector to Γ, and Γ^+ indicates the outflow portion of Γ defined as

$$\Gamma^+ = \{x \in \Gamma \text{ such that } \mathbf{b}(x)\cdot\mathbf{n}(x) \geq 0\}.$$

$\mathcal{E}_h^\circ$ is the set of interior edges associated to the partition $\mathcal{T}_h$. $[\cdot]$ and $\{\cdot\}$ are the standard jump and average operators, respectively, and $\{\cdot\}_{upw}$ is the upwind flux. To define this flux, we consider a vector valued function $\boldsymbol{\tau}$ defined on two neighboring

elements K_1 and K_2 of $\mathcal{T}_h$ with common edge e. Suppose that $\boldsymbol{\tau}^i = \boldsymbol{\tau}|_{K_i}$ for $i = 1, 2$. Then $\{\boldsymbol{\tau}\}_{upw}$ is defined on the edge e as follows:

$$\{\boldsymbol{\tau}\}_{upw} = \frac{1}{2}(\text{sign}(\mathbf{b} \cdot \mathbf{n}^1) + 1)\boldsymbol{\tau}^1 + \frac{1}{2}(\text{sign}(\mathbf{b} \cdot \mathbf{n}^2) + 1)\boldsymbol{\tau}^2,$$

where $\mathbf{n}^i$ is the unit outward normal vector of K_i on e for $i = 1, 2$. The choice of this scheme was made because it was shown [5] that in the positive definite case (i.e. when (5.112) holds) this scheme satisfies (MA1) and (MA2) (cf. Section 5.3.1).

Once a discretization scheme is chosen we can begin to develop our space decomposition and local solvers. In this example, we will obtain the space decomposition by using a nonoverlapping domain decomposition. Let $\mathcal{T}_H$ be a coarse mesh of Ω and $\mathcal{T}_s$ a nonoverlapping partition $\{\Omega_j\}_{j=1}^J$ of Ω such that $\mathcal{T}_s \subseteq \mathcal{T}_H \subseteq \mathcal{T}_h$. Then we define

$$V_0 = W_0 := \left\{v \in L^2(\Omega) \text{ such that } v|_K \in \mathbb{P}_r \ \forall K \in \mathcal{T}_H\right\}, \tag{5.117}$$

$$V_j = W_j := \left\{v \in L^2(\Omega_j) \text{ such that } v|_K \in \mathbb{P}_r \ \forall K \in \mathcal{T}_h \text{ with } K \subseteq \Omega_j\right\} \tag{5.118}$$

for $j = 1, 2, \ldots, J$ and $r \geq 1$. For the prolongation operator $\mathcal{R}_0^\dagger = \mathcal{S}_0^\dagger$ we use the polynomial interpolation on each element $K \in \mathcal{T}_h$.

$$\mathcal{R}_0^\dagger u_0|_K = \text{ the interpolant of } u_0 \text{ in } \mathbb{P}_r(K) \tag{5.119}$$

for each $u_0 \in V_0$ and $K \in \mathcal{T}_h$. For the prolongation operators $\mathcal{R}_j^\dagger = \mathcal{S}_j^\dagger$, when $j = 1, 2, \cdots, J$, we use the following natural injection into V:

$$\mathcal{R}_j^\dagger u_j = \begin{cases} u_j & \text{in } \overline{\Omega}_j \\ 0 & \text{in } \Omega \setminus \overline{\Omega}_j. \end{cases} \tag{5.120}$$

For the local bilinear forms $a_j(\cdot,\cdot)$ we use the exact local solvers defined by

$$a_j(u_j, w_j) := a(\mathcal{R}_j^\dagger u_j, \mathcal{R}_j^\dagger w_j) \qquad \forall\ u_j, w_j \in V_j, \tag{5.121}$$

and $j = 0, 1, \ldots, J$. Note that in this example we only have one set of subspaces $\{V_j\}_{j=0}^J$ and one set of prolongation operators $\{\mathcal{R}_j^\dagger\}_{j=0}^J$ so we shall only have one set of projection-like operators $\{\mathcal{P}_j\}_{j=0}^J$ defined in (5.35) and (5.37). Using these projection-like operators we can then build the Schwarz operators $\mathcal{P}_{ad}$, $\mathcal{P}_{mu}$, and $\mathcal{P}_{hy}$ defined in (5.38), (5.50), and (5.58), respectively.

5.5.2 Partial Analysis of the 1-D Convection Diffusion Problem

In this subsection, we only consider a special 1-D case of (5.107)–(5.108) where $\Omega = (0,1)$, $D(x) \equiv 1$, $\gamma(x) \equiv 1$, and $b(x)$ is a positive constant. Here, the goal is to demonstrate techniques used to prove some of the necessary structure assumptions presented in Subsection 5.4.1, namely assumptions (SA_0) and (SA_1). Structure assumptions (SA_2) and (SA_3) should be easy to verify and we leave these to the reader. (SA_4) will be more challenging to prove and requires the correct choices of prolongation operators and local solvers to be made. It is our intention to explore (SA_4) in more depth in subsequent works.

Let $\{x_\ell\}_{\ell=0}^n$ be a uniform partition of $[0,1]$ with step size h. Then define $\mathcal{T}_h := \{K_\ell\}_{\ell=1}^n$ where $K_\ell = (x_\ell, x_{\ell-1})$. Let

$$V = W := \left\{ v \in L^2(\Omega) \,\middle|\, v|_{K_\ell} \in \mathbb{P}_r(K_\ell), \qquad \forall\, K_\ell \in \mathcal{T}_h \right\}.$$

We will take a uniform penalty parameter $c_e = c_0$. In this special case, $a(u,v)$ in (5.114) is given by

$$\begin{aligned} a(u,v) = &\sum_{\ell=1}^{n} \Big((u',v')_{K_\ell} - (bu,v')_{K_\ell} + (u,v)_{K_\ell} \Big) \\ &+ \sum_{\ell=0}^{n} \Big(\frac{c_0}{h}[u(x_\ell)][v(x_\ell)] - \{u'(x_\ell)\}[v(x_\ell)] \Big) \\ &+ \sum_{\ell=1}^{n} bu^-(x_\ell)[v(x_\ell)], \end{aligned}$$

for any $u, v \in V$. Here, $[u(x_\ell)]$ and $\{u(x_\ell)\}$ are the jump and average operators defined as

$$\begin{aligned} &[u(x_\ell)] := u^-(x_\ell) - u^+(x_\ell) && \{u(x_\ell)\} := \frac{1}{2}\Big(u^-(x_\ell) + u^+(x_\ell) \Big), \\ &[u(x_0)] = -\{u(x_0)\} = -u^+(x_0) && [u(x_n)] = \{u(x_n)\} := u^-(x_n), \end{aligned}$$

where $\ell = 1, 2, \cdots, n-1$ and

$$u^-(y) := \lim_{x \to y^-} u(x) \qquad \text{and} \qquad u^+(y) := \lim_{x \to y^+} u(x).$$

Let $\{x_i^{(0)}\}_{i=0}^{n_0} \subseteq \{x_\ell\}_{\ell=0}^{n}$ be a coarse partition of $[0,1]$ with uniform step size $H > h$ and $\mathcal{T}_H := \{K_i^{(0)}\}_{i=1}^{n_0}$, where $K_i^{(0)} = (x_i^{(0)}, x_{i-1}^{(0)})$ for $i = 1, 2, \cdots, n_0$. Define an even coarser partition $\{x_0^{(j)}\}_{j=1}^{J} \cup \{x_{n_J}^{(J)}\} \subseteq \{x_i^{(0)}\}_{i=0}^{n_0}$ and subdomain decomposition $\mathcal{T}_S := \{\Omega_j\}_{j=1}^{J}$, where $\Omega_j = (x_0^{(j)}, x_0^{(j-1)})$ for $j = 1, \cdots, J-1$ and $\Omega_J = (x_0^{(J)}, x_{n_J}^{(J)})$. These choices ensure that

$$\mathcal{T}_h \supseteq \mathcal{T}_H \supseteq \mathcal{T}_S.$$

Also, there exists $n_1, n_2, \cdots, n_J$ and subsequences $\{x_i^{(j)}\}_{i=0}^{n_j} \subseteq \{x_\ell\}_{\ell=0}^{n}$ for $j = 1, 2, \cdots, J$ such that

$$\sum_{j=1}^{J} n_j = n \qquad \text{and} \qquad \{x_\ell\}_{\ell=0}^{n} = \bigcup_{j=1}^{J} \{x_i^{(j)}\}_{i=0}^{n_j}.$$

Thus,

$$\Omega_j = \bigcup_{i=1}^{n_j} K_i^{(j)},$$

for $j = 1, 2, \cdots, J$, and $K_i^{(j)}$ is defined as $K_i^{(j)} = (x_i^{(j)}, x_{i-1}^{(j)})$ for $i = 1, 2, \cdots, n_j$.

Define the subspaces V_j and local solvers $a_j(\cdot, \cdot)$ by

$$V_0 = W_0 := \left\{ v \in L^2(\Omega) \,\middle|\, v|_{K_i^{(0)}} \in P_r\big(K_i^{(0)}\big) \text{ for } i = 1, 2, \cdots, n_0 \right\},$$
$$V_j = W_j := \left\{ v \in L^2(\Omega_j) \,\middle|\, v|_{K_i^{(j)}} \in P_r\big(K_i^{(j)}\big) \text{ for } i = 1, 2, \cdots, n_j \right\},$$

for $j = 1, 2, \cdots, J$ and

$$a_j(u, v) := a(\mathcal{R}_j^\dagger u, \mathcal{R}_j^\dagger v),$$

for $u, v \in V_j$ and $j = 0, 1, \cdots, J$. Here $\mathcal{R}_j^\dagger$ is taken to be the prolongation operators described in Subsection 5.5.1.

Define the norm $|||\cdot|||$ on V in the following way:

$$|||v|||^2 := |||v|||^2 + |||v|||_{rc}^2$$
$$|||v|||_d^2 := |v|_{1,h}^2 + \sum_{\ell=0}^{n} \frac{1}{h}[v(x_\ell)]^2,$$
$$|||v|||_{rc}^2 := (1 + b)\|v\|_{L^2(\Omega)}^2 + \sum_{\ell=0}^{n} b[v(x_\ell)]^2.$$

Using this norm, Ayuso and Marini proved that there exists $h_0 = h_0(b)$ and $\gamma_a = \gamma_a(b, \Omega)$ such that

$$\sup_{v \in V^h} \frac{a(u,v)}{|||v|||} \geq \gamma_a |||u|||,$$

for all $u \in V$ when $h \leq h_0$ (c.f. [5]).

From the above definitions, we see that the local solvers take the following form

$$\begin{aligned}
a_j(u,v) = &\sum_{i=1}^{n_j} \Big((u', v')_{K_i^{(j)}} - (bu, v')_{K_i^{(j)}} + (u, v)_{K_i^{(j)}} \Big) \\
&+ \sum_{i=1}^{n_j - 1} \Big(\frac{c_0}{h} [u(x_i^{(j)})][v(x_i^{(j)})] - \{u'(x_i^{(j)})\}[v(x_i^{(j)})] + bu^-(x_i^{(j)})[v(x_i^{(j)})] \Big) \\
&+ \frac{c_0}{h} \Big(u^-(x_{n_j}^{(j)}) v^-(x_{n_j}^{(j)}) + u^+(x_0^{(j)}) v^+(x_0^{(j)}) \Big) + bu^-(x_{n_j}^{(j)}) v^-(x_{n_j}^{(j)}) \\
&+ \frac{1}{2}(\delta_{j,0} + \delta_{j,1} + 1) u'^+(x_0^{(j)}) v^+(x_0^{(j)}) - \frac{1}{2}(\delta_{j,0} + \delta_{j,J} + 1) u'^-(x_{n_j}^{(j)}) v^-(x_{n_j}^{(j)}),
\end{aligned}$$

for all $u, v \in V_j$ and $j = 0, 1, 2, \cdots, J$. Here $\delta_{\ell,m}$ denotes the Kronecker delta symbol. We note that the local solvers take the a similar form as the DG bilinear form $a(\cdot, \cdot)$, noting that the penalty parameter for the coarse local solver $a_0(\cdot, \cdot)$ should be thought of as $\frac{c_0 H}{h}$ to gain the correct scaling. This immediately implies that there exists $\gamma_{a_j} = \gamma_{a_j}(b, \Omega_j)$ for $j = 1, 2, \cdots, J$ such that

$$\sup_{v_j \in V_j} \frac{a_j(u_j, v_j)}{|||v_j|||} \geq \gamma_{a_j} |||u_j|||,$$

for all $u_j \in V_j$ and $h < h_0$. This also holds for $j = 0$ when the coarse mesh size satisfies $H < h_0$. Therefore, (SA_0) is satisfied.

Let $I(\cdot,\cdot)$ denote an interface bilinear form on $V \times V$ defined as

$$I(u,v) := \sum_{j=1}^{J-1} \Big(-\frac{c_0}{h}\big(u^-(x_{n_j}^{(j)})v^+(x_{n_j}^{(j)}) + u^+(x_{n_j}^{(j)})v^-(x_{n_j}^{(j)})\big) - bu^-(x_{n_j}^{(j)})v^+(x_{n_j}^{(j)})\Big)$$
$$+ \sum_{j=1}^{J-1} \frac{1}{2}\Big(u'^-(x_{n_j}^{(j)})v^+(x_{n_j}^{(j)}) - u'^+(x_{n_j}^{(j)})v^-(x_{n_j}^{(j)})\Big).$$

Similarly, define the interface functional $\langle\cdot\rangle_I$ on V by

$$\langle u\rangle_I := -2\Big(\frac{1}{h}+b\Big)\sum_{j=1}^{J-1} u^+(x_{n_j}^{(j)})u^-(x_{n_j}^{(j)}).$$

For the rest of this subsection we aim to prove (SA_1), i.e. prove the existence of an energy stable decomposition of every $v \in V$. To do this, we need to establish a series of technical lemmas. Using the definitions of $I(\cdot,\cdot)$ and $\langle\cdot\rangle_I$, we immediately obtain the following lemma.

Lemma 5.5.1. *For all $u, v \in V$ there exist unique decompositions $u = \sum_{j=1}^{J} \mathcal{R}_j^\dagger u_j$ and $v = \sum_{j=1}^{J} \mathcal{R}_j^\dagger v_j$, where $u_j, v_j \in V_j$ for $j = 1, 2, \cdots, J$. Moreover,*

$$a(u,v) = \sum_{j=1}^{J} a_j(u_j, v_j) + I(u,v),$$
$$|||u|||^2 = \sum_{j=1}^{J} |||u_j|||^2 + \langle u\rangle_I.$$

From [40] we obtain the following two technical lemmas.

Lemma 5.5.2. *For any $u \in V$, there holds the trace inequality*

$$|u|^2_{\partial K_i^{(0)}} \le c\Big(H^{-1}\|u\|^2_{L^2(K_i^{(0)})} + H|u|^2_{1,h,K_i^{(0)}}\Big) \qquad \text{for } i = 1, 2, \cdots, n_0,$$

where

$$|u|^2_{1,h,K_i^{(0)}} := \sum_{K_\ell \in K_i^{(0)}} \|u'\|_{L^2(K_\ell)} + \sum_{x_\ell \in K_i^{(0)}} \frac{1}{h}[u(x_\ell)]^2.$$

Lemma 5.5.3. *For any $u \in V$, let $u_0 \in V_0$ be defined by*

$$u_0|_{K_i^{(0)}} = \frac{1}{\text{meas}(K_i^{(0)})} \int_{K_i^{(0)}} u \mathrm{d}x \qquad \text{for } i = 1, 2, \cdots, n_0,$$

then

$$\|u - u_0\|_{L^2(K_i^{(0)})} \leq cH|u|_{1,h,K_i^{(0)}}.$$

The following lemma verifies (SA_1).

Lemma 5.5.4 (Energy Stable Decomposition)**.** *For every $u \in V$, there exists $u_j \in V_j$ for $j = 0, 1, \cdots, J$ such that*

$$\sum_{j=0}^{J} \|u\|_{a_j} \leq C_V \|u\|_a,$$

where

$$C_V = C\Big(JC_a\gamma_a\Big(\frac{1}{h} + b\Big)\big(H^2 + H\big)\Big)^{\frac{1}{2}}.$$

Proof. Let $u \in V$ and define $u_0 \in V_0$ by

$$u_0|_{K_i^{(0)}} = \frac{1}{\text{meas}(K_i^{(0)})} \int_{K_i^{(0)}} u \mathrm{d}x \qquad \text{for } i = 1, 2, \cdots, n_0.$$

Then let $u_j \in V_j$, for $j = 1, 2, \cdots, J$, be defined uniquely by $u - u_0 = u_1 + u_2 + \cdots + u_J$. From Lemma 5.5.1 we get

$$|||u - u_0|||^2 = \sum_{j=1}^{J} |||u_j|||^2 + \langle u - u_0 \rangle_I.$$

Thus,

$$\begin{aligned} \sum_{j=0}^{J} |||u_j|||^2 &= |||u - u_0|||^2 + |||u_0|||^2 - \langle u - u_0 \rangle_I \\ &\leq 2|||u|||^2 + 3|||u_0|||^2 + \Big|\langle u - u_0 \rangle_I\Big|. \end{aligned} \tag{5.122}$$

We will estimate $|||u_0|||$ and $\langle u - u_0 \rangle_I$ separately. Using Lemmas 5.5.2 and 5.5.3 we find

$$\begin{aligned} |||u_0|||^2 &= (1+b)\|u_0\|_{L^2(\Omega)} + \Big(\frac{1}{h} + b\Big) \sum_{i=0}^{n_0} [u_0(x_i^{(0)})]^2 \\ &\leq C(1+b)\Big(\sum_{i=1}^{n_0} \|u - u_0\|^2_{L^2(K_i^{(0)})} + \|u\|^2_{L^2(\Omega)} \Big) \\ &\quad + C\Big(\frac{1}{h} + b\Big)\Big(\sum_{i=0}^{n_0} [u(x_i^{(0)}) - u_0(x_i^{(0)})]^2 + \sum_{\ell=0} n[u(x_\ell)]^2 \Big) \\ &\leq C|||u|||^2 + C\Big(\frac{1}{h} + b\Big) \sum_{i=1}^{n_0} H^2 |u|^2_{1,h,K_i^{(0)}} \\ &\quad + C\Big(\frac{1}{h} + b\Big) \sum_{i=0}^{n_0 - 1} \Big(\big(u^-(x_i^{(0)}) - u_0^-(x_i^{(0)})\big)^2 + \big(u^+(x_{i+1}^{(0)}) - u_0^+(x_{i+1}^{(0)})\big)^2 \Big) \\ &\leq C|||u|||^2 + C\Big(\frac{1}{h} + b\Big) \sum_{i=1}^{n_0} \Big(H^2 |u|^2_{1,h,K_i^{(0)}} + H|u|^2_{1,h,K_i^{(0)}} \Big) \\ &\leq C\Big(\frac{1}{h} + b\Big)(H^2 + H)|||u|||^2. \end{aligned}$$

Above, we used the facts $|u-u_0|_{1,h,K_i^{(0)}} = |u|_{1,h,K_i^{(0)}}$ for $i = 1, 2, \cdots, n_0$ and $Hh^{-1} \geq 1$. Using Lemmas 5.5.2 and 5.5.3, we find

$$\begin{aligned}
\Big|\langle u-u_0\rangle_I\Big| &= 2\Big(\frac{1}{h}+b\Big)\Big|\sum_{j=1}^{J-1}\big(u^+(x_{n_j}^{(j)})-u_0^+(x_{n_j}^{(j)})\big)\big(u^-(x_{n_j}^{(j)})-u_0^-(x_{n_j}^{(j)})\big)\Big| \\
&\leq 2\Big(\frac{1}{h}+b\Big)\Big(\sum_{j=1}^{J-1}\big(u^+(x_{n_j}^{(j)})-u_0^+(x_{n_j}^{(j)})\big)^2\Big)^{\frac{1}{2}}\Big(\sum_{j=1}^{J-1}\big(u^-(x_{n_j}^{(j)})-u_0^-(x_{n_j}^{(j)})\big)^2\Big)^{\frac{1}{2}} \\
&\leq 2\Big(\frac{1}{h}+b\Big)\sum_{i=0}^{n_0-1}\Big(\big(u^-(x_i^{(0)})-u_0^-(x_i^{(0)})\big)^2+\big(u^+(x_{i+1}^{(0)})-u_0^+(x_{i+1}^{(0)})\big)^2\Big) \\
&\leq C\Big(\frac{1}{h}+b\Big)\sum_{i=1}^{n_0}\Big(H^{-1}\|u-u_0\|^2_{L^2(K_i^{(0)})}+H|u-u_0|^2_{L^2(K_i^{(0)})}\Big) \\
&\leq C\Big(\frac{1}{h}+b\Big)\sum_{i=1}^{n_0}\Big(H|u|^2_{L^2(K_i^{(0)})}\Big) \\
&\leq C\Big(\frac{1}{h}+b\Big)H|||u|||^2.
\end{aligned}$$

We apply these two estimates to (5.122) and get

$$\sum_{j=0}^{J}|||u_j|||^2 \leq C\Big(\frac{1}{h}+b\Big)\big(H^2+H\big)|||u|||^2.$$

Now using this result along with the norm equivalence results, we find

$$\begin{aligned}
\sum_{j=0}^{J}\|u_j\|_{a_j}^2 &\leq C_a\sum_{j=1}^{J}|||u_j|||^2 \\
&\leq CC_a(\frac{1}{h}+b)(H^2+H)|||u|||^2 \\
&\leq CC_a\gamma_a\Big(\frac{1}{h}+b\Big)\big(H^2+H\big)\|u\|_a^2.
\end{aligned}$$

We obtain the desired result by applying the equivalence of $\|\cdot\|_1$ and $\|\cdot\|_2$ to the above inequality. □

Remark 5.5.5. *We note that in [40], $C_V = C\,(Hh^{-1})$ but here for $H \leq 1$ we find $C_V = CH\,(b+h^{-1})$.*

5.5.3 Numerical Experiments

In this section, we present several 1-D numerical experiments to gauge the theoretical results proved in the previous section. For these experiments we concentrated on equation (5.107) in the domain $\Omega = (0, 1)$ with the following choices of constant coefficient:

Test 1. $D(x) = 1$, $b(x) = 1,000$, and $\gamma(x) = 1$.

Test 2. $D(x) = 1$, $b(x) = 2,000$, and $\gamma(x) = 1$.

Test 3. $D(x) = 1$, $b(x) = 10,000$, and $\gamma(x) = 1$.

Test 4. $D(x) = 1$, $b(x) = 100,000$, and $\gamma(x) = 1$.

Note that these choices of coefficients put us in the convection dominated regime and fit the criteria of the positive definite case characterized by (5.112). For this reason we are able to use the discretization scheme and domain decomposition techniques described in Section 5.5.1. In these experiments, we use a uniform fine mesh size $h = 1/256$ and a uniform coarse mesh size $H = 1/64$. The equations are solved numerically using standard GMRES, GMRES after using $\mathcal{P}_{ad}$ preconditioning, the multiplicative Schwarz iterative method (5.55), and GMRES after using $\mathcal{P}_{hy}$ preconditioning, all with a stopping tolerance of 10^{-6}. To verify the dependence of $\kappa_a(\mathcal{P}_{ad})$ and $\kappa_a(\mathcal{P}_{hy})$, we use a varying number of subdomains $J = 4, 8, 16, 32, 64$.

Our first goal in these experiments is to compare the performance of the Schwarz methods to that of standard GMRES in order to verify the usefulness of such methods. We would like to verify numerically that the estimates given in previous sections are tight. In particular, we would like to find an example that shows that $\kappa_A(P_{ad})$ does in fact depend linearly on the number of subdomains J as predicted in Theorem 5.4.9. For multiplicative Schwarz iteration we would like to estimate $\|E_{mu}\|_A$, noting that if this norm is less than 1 it guarantees convergence of the method. If not, we shall need to rely on the spectral radius $\rho(E_{mu})$ to guarantee this convergence.

Tables 5.1–5.4 collect the test results on the additive, multiplicative, and hybrid Schwarz methods proposed in Section 5.3. Where $J = \text{NA}$ represents the original

system with no preconditioning. From these numerical results we can make the following observations:

(a) Any of these methods offers an improvement in terms of the CPU time needed to solve the system when compared to solving the system using standard GMRES.

(b) GMRES after using $\mathcal{P}_{ad}$ or $\mathcal{P}_{hy}$ for preconditioning performs better when the number of subdomains J is not too large.

(c) In all of these tests, $\kappa_A(P_{ad})$ and $\kappa_A(P_{hy})$ depend on the number of subdomains J. Particularly in **Test 2**, we see an example that exhibits approximate linear dependence. See figure 5.1.

(d) For $\|E_{mu}\|_A$ we do not observe such a strong dependence on the number of subdomains J.

(e) In these tests $\|E_{mu}\|_A$ is greater than 1; thus, we cannot rely upon this as an indicator for convergence of the multiplicative Schwarz iterative method.

(f) κ_A is not a unique metric in judging the convergence of GMRES after preconditioning with P_{ad} and P_{hy}. For instance, in **Test 4** $\kappa_A(P_{ad})$ decreases while the number of iterations necessary for GMRES increases as J increases. This is opposite of the behavior that is observed in the previous tests.

Our numerical experiments verify that κ_A is not a unique metric for the convergence of GMRES. Therefore, we must rely on other metrics to predict the convergence behavior of GMRES. A result that could be of help in this area is the following theorem (cf. [78]).

Theorem 5.5.6. *Consider the linear system* $A\mathbf{x} = \mathbf{b}$ *where* $A \in \mathbb{R}^{d\times d}$ *and* $\mathbf{x}, \mathbf{b} \in \mathbb{R}^d$. *Further suppose that* A *is diagonalizable. Then after* k *steps of GMRES, the residual*

$\mathbf{r}_k := \mathbf{b} - A\mathbf{x}^{(k)}$ *satisfies*

$$\frac{\|\mathbf{r}_k\|_2}{\|\mathbf{b}\|_2} \leq \kappa_2(V) \inf_{\substack{p \in \mathbb{P}_k \\ p(0)=1}} \sup_{\lambda \in \sigma(A)} |p(\lambda)|,$$

where V is a nonsingular matrix of eigenvectors of A and $\sigma(A)$ denotes the spectrum of A.

The above theorem says that the spread of the spectrum is a metric to judge the performance of GMRES with GMRES performing better when the spectrum of A is clustered. With this theorem in mind, let us examine the spectrum of the matrix A and P_{ad} for $J = 4, 8, 16, 32, 64$ obtained in **Test 2** and **Test 4**.

Note that in Figure 5.2(a) and Figure 5.3(a) the spectrum has a large spread which is consistent with the fact that GMRES performed poorly on the original system without preconditioning. We also see that after preconditioning, the spectrum of P_{ad} is clustered which corresponds to improved performance of GMRES after preconditioning with P_{ad}. Lastly, we note that as the number of subdomains J increases, the spread of the spectrum of P_{ad} increases. This corresponds to a decreased performance in GMRES after preconditioning with P_{ad} as J increases.

This result leads us to believe that to accurately judge the behavior of GMRES after preconditioning one needs to analyze the spectrum of the preconditioned system. Similarly, we find that to accurately predict the performance of the multiplicative Schwarz iterative method one needs to analyze the spectral radius of E_{mu}.

Table 5.1: Performance of three Schwarz methods on **Test 1**

(a) GMRES after preconditioning with P_{ad} and P_{hy}

J	Iteration # of GMRES		CPU Time		κ_A	
NA	552		14.3760		3.3893×10^4	
	P_{ad}	P_{hy}	P_{ad}	P_{hy}	P_{ad}	P_{hy}
4	7	2	1.3638	1.1922	460.5713	397.3567
8	7	3	1.3343	1.2367	436.7967	398.2544
16	11	5	1.6873	1.4040	438.2207	412.1700
32	17	8	2.6431	1.9066	521.3530	478.9537
64	30	15	6.2315	3.7889	774.7091	619.3973

(b) Multiplicative Schwarz Iteration

J	Iterations # of Mult. Schwartz	CPU Time	$\|E_{mu}\|_A$	$\rho(E_{mu})$
4	2	1.1060	19.8830	4.4793×10^{-6}
8	2	1.1016	19.8889	0.0029
16	3	1.1352	19.8469	0.0725
32	5	1.2768	19.7658	0.3179
64	8	1.7129	19.7176	0.5926

Table 5.2: Performance of three Schwarz methods on **Test 2**

(a) GMRES after preconditioning with P_{ad} and P_{hy}

J	Iteration # of GMRES		CPU Time		κ_A	
NA	550		14.4971		1.7388×10^4	
	P_{ad}	P_{hy}	P_{ad}	P_{hy}	P_{ad}	P_{hy}
4	8	3	1.3249	1.2069	741.9511	699.5729
8	10	5	1.4463	1.2835	749.0976	713.3674
16	17	8	1.9924	1.5557	847.4815	800.9121
32	27	14	5.5602	2.4255	1.1221×10^3	1.0029×10^3
64	44	24	8.7063	5.3089	1.6247×10^3	1.2918×10^3

(b) Multiplicative Schwarz Iteration

J	Iterations # of Mult. Schwartz	CPU Time	$\|E_{mu}\|_A$	$\rho(E_{mu})$
4	2	1.1010	26.4005	0.0011
8	3	1.1131	26.3187	0.0451
16	4	1.1679	26.1222	0.2529
32	6	1.3214	25.9832	0.5277
64	10	1.8713	25.9270	0.7167

Table 5.3: Performance of three Schwarz methods on **Test 3**

(a) GMRES after preconditioning with P_{ad} and P_{hy}

J	Iteration # of GMRES		CPU Time		κ_A	
NA	554		14.3919		4.0782×10^3	
	P_{ad}	P_{hy}	P_{ad}	P_{hy}	P_{ad}	P_{hy}
4	8	3	1.3422	1.1772	647.6787	615.1005
8	12	5	1.4953	1.2517	658.7462	627.0064
16	18	9	2.0216	1.5588	726.3005	690.1682
32	27	15	3.5402	2.4623	854.1450	788.5277
64	35	23	7.1266	4.9327	939.5190	816.1892

(b) Multiplicative Schwarz Iteration

J	Iteration # of Mult. Schwartz	CPU Time	$\|E_{mu}\|_A$	$\rho(E_{mu})$
4	2	1.1067	24.7399	0.0021
8	2	1.0982	24.6200	0.0526
16	3	1.1394	24.4247	0.2350
32	5	1.2778	24.2986	0.4369
64	7	1.6321	24.2524	0.5302

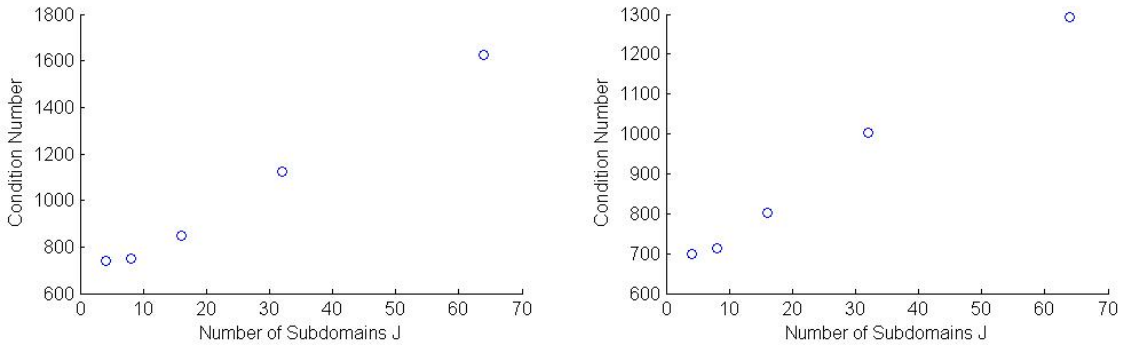
Table 5.4: Performance of three Schwarz methods on **Test 4**

(a) GMRES after preconditioning with P_{ad} and P_{hy}

J	Iteration # of GMRES		CPU Time		κ_A	
NA	468		9.4276		1.0769×10^3	
	P_{ad}	P_{hy}	P_{ad}	P_{hy}	P_{ad}	P_{hy}
4	8	2	1.3305	1.2039	103.5739	31.7538
8	11	2	1.4551	1.2558	75.7527	31.6954
16	14	3	1.8217	1.3019	56.6486	31.6803
32	13	5	2.2940	1.6227	46.4141	31.8710
64	15	8	3.7025	2.5950	44.1292	32.2846

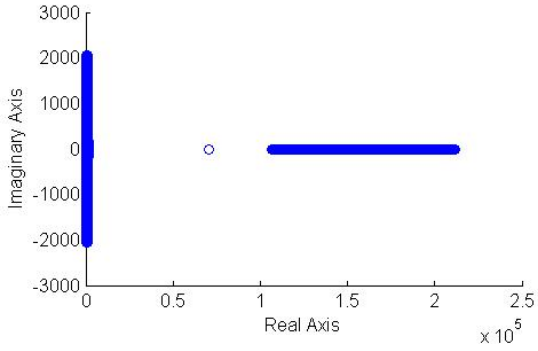
(b) Multiplicative Schwarz Iteration

J	Iteration # of Mult. Schwartz	CPU Time	$\|E_{mu}\|_A$	$\rho(E_{mu})$
4	2	1.0996	5.4574	85394×10^{-9}
8	2	1.0984	5.4575	1.0873×10^{-6}
16	2	1.1157	5.4575	6.6472×10^{-4}
32	2	1.1566	5.4560	0.0158
64	2	1.2399	5.4540	0.0678

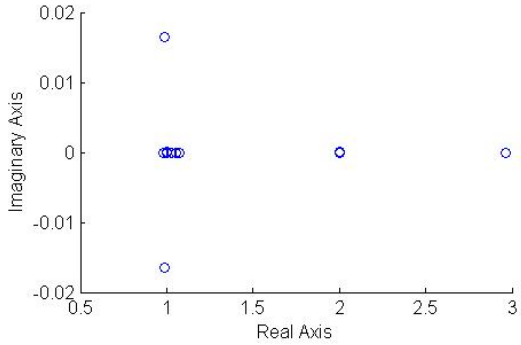


(a) Plot of J vs. $\kappa_A(P_{ad})$ **(b)** Plot of J vs. $\kappa_A(P_{hy})$

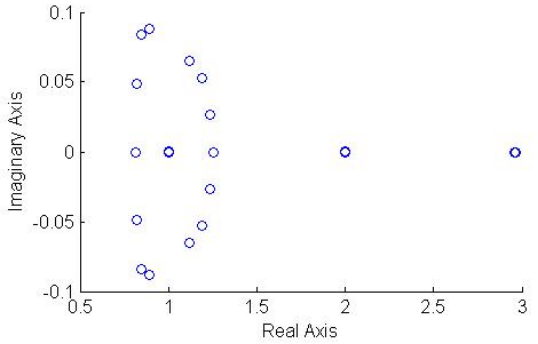
Figure 5.1: Dependence of $\kappa_A(P_{ad})$ and $\kappa_A(P_{hy})$ on J in **Test 2**



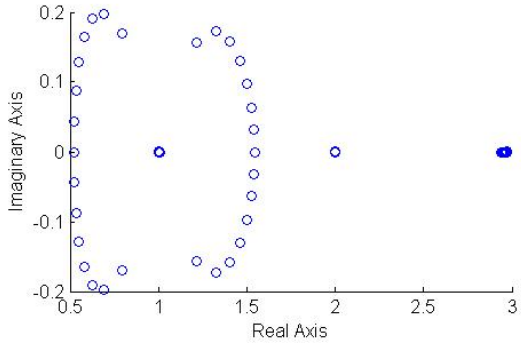
(a) Plot of $\sigma(A)$



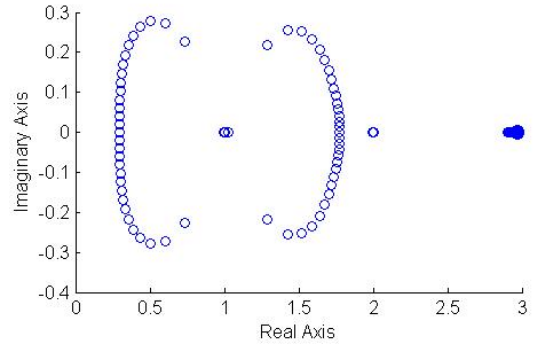
(b) Plot of $\sigma(P_{ad})$ with J = 4



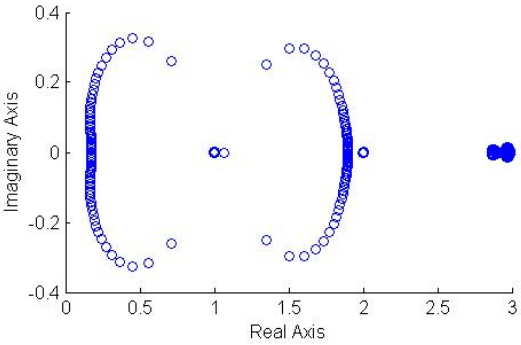
(c) Plot of $\sigma(P_{ad})$ with J = 8



(d) Plot of $\sigma(P_{ad})$ with J = 16

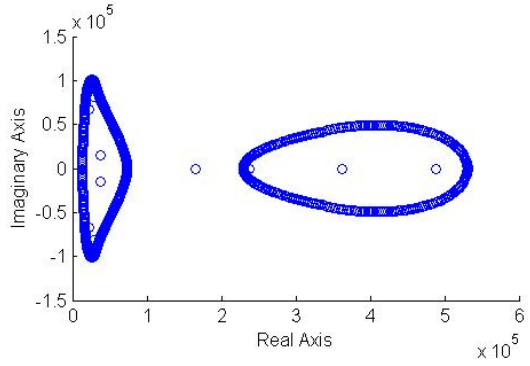


(e) Plot of $\sigma(P_{ad})$ with J = 32

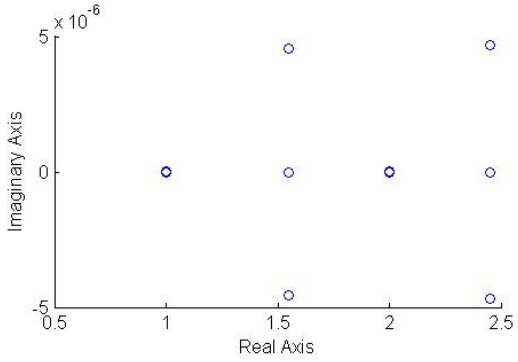


(f) Plot of $\sigma(P_{ad})$ with J = 64

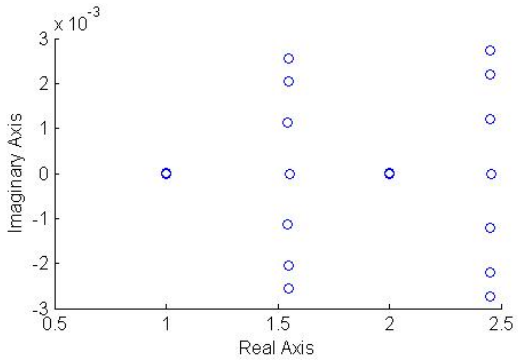
Figure 5.2: Spectrum plots from **Test 2**



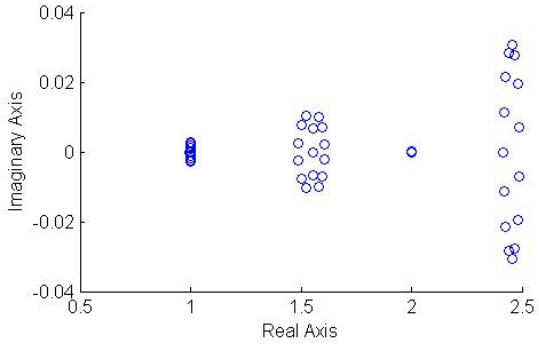
(a) Plot of $\sigma(A)$



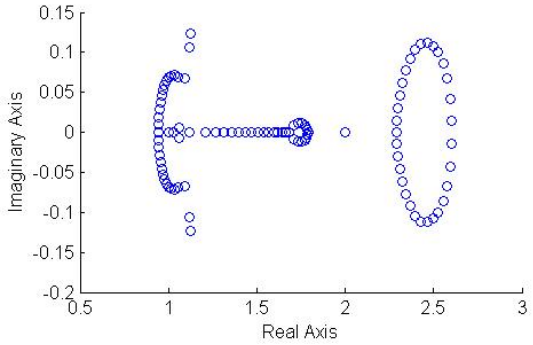
(b) Plot of $\sigma(P_{ad})$ with J = 4



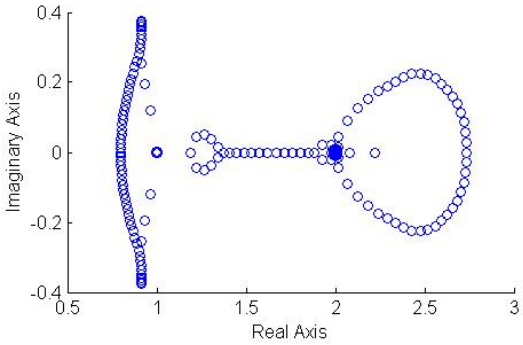
(c) Plot of $\sigma(P_{ad})$ with J = 8



(d) Plot of $\sigma(P_{ad})$ with J = 16



(e) Plot of $\sigma(P_{ad})$ with J = 32



(f) Plot of $\sigma(P_{ad})$ with J = 64

Figure 5.3: Spectrum plots from **Test 4**

Chapter 6

Future Directions

The goal of this chapter is to list a few future research directions that come directly from the work in the previous chapters of this dissertation.

- *Give examples of domains that satisfy a generalized star-shape condition but not a classical star-shape condition*

  Generalized star-shape domain conditions were introduced in Chapter 2 to replace the more restrictive star-shape condition. Novel examples that satisfy the generalized star-shape condition but not the classical star-shape condition need to be obtained to justify this generalization.

- *Prove the Korn-type inequality on the boundary of a domain for solutions of the elastic Helmholtz equations, i.e. prove Conjecture 2.3.4*

  This conjecture is important to obtain the results for the elastic Helmholtz equations in Chapter 2 as well as to obtain optimal stability estimates in the wave frequency ω in [27].

- *Continue to develop new absolutely stable discretization methods for the elastic Helmholtz problem*

In particular, the IP-DG method given in Chapter 3 should be extended to an *hp* method using higher order polynomials. Other discretization methods can be considered as well, such as the local discontinuous Galerkin (LDG) method [45].

- *Develop multi-modes MCIP-DG methods for the other Helmholtz-type problems*

  The method and analysis demonstrated in Chapter 4 can be extended to the other Helmholtz-type problems in random media. This is worthwhile since these problems have the same numerical challenges as the acoustic Helmholtz problem.

- *Develop multi-modes MCIP-DG methods for other PDEs*

  In particular, this approach can be extended to general elliptic PDEs with random coefficients. For a problem like the Poisson problem where fast solvers are available, is this approach worthwhile? This is a question that should be explored.

- *Continue to develop the Schwarz framework to include Helmholtz-type problems*

  The new Schwarz framework should extend easily to the case of complex non-Hermitian and indefinite problems that satisfy a weak coercivity property. To generalize it to the Helmholtz-type problems, we need to extend the framework to include problems that only satisfy a generalized weak coercivity property.

- *Apply the nonsymmetric and indefinite Schwarz framework to various PDE problems*

  Many problems do not fit the classical SPD Schwarz framework. The new Schwarz framework given in Chapter 5 should be applied to these problems. An example would be the Navier-Stokes equations.

Bibliography

[1] M. Ainsworth, P. Monk, and W. Muniz. Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. *J. Sci. Comput.*, 27:5 – 40, 2006. 4

[2] M. Ainsworth and H.A. Wajid. Optimally blended spectral-finite element scheme for wave propagation and nonstandard reduced integration. *SIAM J. Numer. Anal.*, 48(1):346 – 371, 2010. 4

[3] M. Amara, R. Djellouli, and C. Farhat. Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems. *SIAM J. Numer. Anal.*, 47(2):1038 – 1066, 2009. 4

[4] P.F. Anonietti and B. Ayuso. Schwarz domain decomposition preconditioners for discontinuous Galerkin approximations of elliptic problems:non-overlapping case. *M2AN Math. Model. Numer. Anal.*, 41:21 – 54, 2007. 141

[5] B. Ayuso and L.D. Marini. Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 47:1391 – 1420, 2009. 177, 178, 182

[6] A.K. Aziz and R. B. Kellogg. A scattering problem for the Helmholtz equation. In *Advances in Computer Methods for Partial Differential Equations - III; Proceedings of the Third International Symposium, Bethlem, PA, June 20 - 22, 1979*, pages 93 – 95, New Brunswick, NJ, 1979. IMACS. 4, 49, 62, 69

[7] A.K. Aziz and A. Werschulz. On the numerical solutions of Helmholtz's equation by the finite element method. *SIAM J. Numer. Anal.*, 17(5):681 – 686, 1980. 4

[8] I. Babuška. Error bound for the finite element method. *Numer. Math.*, 16:322 – 333, 1971. 61, 144, 146

[9] I. Babuška and A.K. Aziz. Survey lectures on the mathematical foundations of finite element method. In A.K. Aziz, editor, *The Mathematical Foundations of*

the FEM with Applications to PDE, pages 5 – 359. Academic Press, 1972. 61, 143, 144, 146, 147, 176, 177

[10] I. Babuška, F. Ihlenburg, E.T. Paik, and S.A. Sauter. A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution. *Comput. Methods Appl. Mech. Engrg.*, 128(3 - 4):325 – 359, 1995. 4

[11] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Rev.*, 52:317 – 355, 2010. 87, 89, 116

[12] I. Babuška, R. Tempone, and G.E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42:800 – 825, 2004. 87, 89, 116, 124, 125

[13] I.M. Babuška and S.A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM J. Numer. Anal.*, 34(6):2392 – 2423, 1997. 4, 5

[14] C. Bernardi and Y. Maday. Spectral methods. In *Handbook of Numerical Analysis*, volume V, pages 209 – 485. North-Holland, Amsterdam, 1997. 146

[15] J.H. Bramble, J.E. Pasciak, J. Wang, and J Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comp.*, 57:1 – 21, 1991. 158, 159

[16] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York, 3rd edition, 2008. 63, 65, 69, 144, 146, 147, 148, 169

[17] F. Brezzi. On the existence, uniqueness, and approximation of saddle-point problems arising from Lagrange multipliers. *RAIRO*, 2:129 – 151, 1974. 144, 147

[18] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Method*. Springer, New York, 1991. 144, 147

[19] A. Buffa and P. Monk. Error estimates for the ultra weak variational formulation of the Helmholtz equation. *M2AN Math. Model. Numer. Anal.*, 42(6):925 – 940, 2008. 4

[20] A. Buffa and I. Perugia. Discontinuous Galerkin approximation of heterogeneous Maxwell source problem. *Bol. Sov. Esp. Mat. Apl. SMA*, 32(34):33 – 44, 2006. 4

[21] R. Caflisch. Monte carlo and quasi-monte carlo methods. *Acta Numerica*, 7:1–49, 1998. 87, 89

[22] X.C. Cai and O.B. Widlund. Domain decomposition algorithms for indefinite elliptic problems. *SIAM J. Sci. Statist. Comput.*, 13:243 – 258, 1992. 149

[23] C.L. Chang. A least-squares finite element method for the Helmholtz equation. *Comput. Methods. Appl. Mech. Engrg*, 83(1):1 – 7, 1990. 4

[24] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978. 65, 144, 146, 147

[25] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer-Verlag, Berlin, 1992. 4

[26] P. Cummings. *Analysis of Finite Element Based Numerical Methods for Acoustic Waves, Elastic Waves and Fluid-Solid Interactions in the Frequency Domain*. PhD thesis, The University of Tennessee, 2001. 4, 49, 63, 69

[27] P. Cummings and X. Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Mathematical Models and Methods in Applied Sciences*, 16:139 – 160, 2006. 6, 10, 14, 20, 21, 63, 90, 93, 114, 197

[28] B.E.J Dahlberg, C.E. Kenig, and G.C. Verchota. Boundary value problems for the systems of elastostatics in Lipschitz domains. *Duke Math. J*, 57:795 – 818, 1988. 22

[29] J. Douglas Jr., J.E. Santos, D. Sheen, and L.S. Bennethum. Frequency domain treatment of one-dimensional scalar waves. M^3AS, 3(2):171 – 194, 1993. 3, 4, 5, 49, 62, 69

[30] J. Douglas Jr., D. Sheen, and J.E. Santos. Approximations of scalar waves in the space-frequency domain. M^3AS, 4(4):509 – 531, 1994. 3, 4, 49, 62, 69

[31] M. Dryja and O.B. Widlund. Towards a unified theory of domain decomposition algorithms for elliptic problems. In T. Chan, R. Glowinski, J. Periaux, and O.B. Widlund, editors, *Proceedings of Third International Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 3 – 21, Philadelphia, 1990. SIAM. 141, 148, 149, 152, 156

[32] E.G. Dutra do Carmo, G.B. Alvarez, A.F.D Loula, and F.A. Rochinha. A nearly optimal Galerkin projected residual finite element method for Helmholtz problem. *Comput. Methods Appl. Mech. Engrg*, 197(13 - 16):1362 – 1375, 2008. 4

[33] M. Eiermann, O. Ernst, and E. Ullmann. Computational aspects of the stochastic finite element method. *Proceedings of ALGORITMY*, pages 1–10, 2005. 89

[34] S.C. Eisenstat, H.C. Elman, and M.H. Schultz. Variational iterative methods for nonsymetric systems of linear equations. *SIAM J. Numer. Anal.*, 20(2):345 – 357, 1983.

[35] B. Engquist and A. Majda. Radiation bounday conditions for acoustic and elastics wave calculations. *Comm. Pure Appl. Math.*, 32(3):314 – 358, 1979. 2, 3, 88

[36] B. Engquist and O. Runborg. Computational high frequency wave propagation. *Acta Numer.*, 12:181 – 266, 2003. 4

[37] O. Ernst and M. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods? In I. Graham, T. Hou, O. Lakkis, and R. Scheichl, editors, *Numerical Analysis of Multiscale Problems*, Lecture Notes in Computational Science and Engineering 83, pages 325 – 363. Springer Verlag, 2012. 5, 128

[38] C. Farhat, I. Harari, and U. Hetmaniuk. A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime. *Comput. Methods Appl. Mech. Engrg.*, 192(11 - 12):1389 – 1419, 03. 4

[39] X. Feng. Wave number-explicit a priori estimates for the time-harmonic Maxwell equations. preprint, July 2010. 31

[40] X. Feng and O. Karakashian. Two-level additive Schwarz methods for discontinuous Galerkin approximations of second order elliptic problems. *SIAM J. Numer. Anal.*, 39(39):1343 – 1365, 2001. 141, 183, 186

[41] X. Feng and O. Karakashian. Two-level non-overlapping Schwarz preconditioners for a discontinuous Galerkin approximation of the biharmonic equation. *J. Sci. Comput.*, 22/23:289 – 314, 2005. 141

[42] X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave numbers. *SIAM J. Numer. Anal.*, 47:2872 – 2896, 2009. 5, 6, 10, 49, 50, 80, 92, 110, 111, 112, 113, 115, 135

[43] X. Feng and H. Wu. Absolutely stable interior penalty discontinuous Galerkin methods for the time-harmonic Maxwell equations with large wave number. http://arxiv.org/pdf/1210.5837v2.pdf, December 2010. 5, 6, 10, 34, 43, 49

[44] X. Feng and H. Wu. hp-discontinuous Galerkin methods for the Helmholtz equation with large wave numbers. *Math. Comp.*, 80:1997 – 2024, 2011. 5, 6, 10, 50, 92, 110, 111, 112, 113

[45] X. Feng and Y. Xing. Absolutely stable local discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Math. Comp.*, 82:1269 – 1296, 2012. 5, 198

[46] J. Fouque, J. Garnier, G. Papanicolaou, and K. Solna. *Wave Propogation and Time Reversal in Randomly Layered Media*, volume 56 of *Stochastic Modeling and Applied Probability*. Springer, 2007. 88

[47] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer Verlag, Berlin, 2001. reprint of the 1998 edition. 98, 176, 177

[48] C.I. Goldstein. The finite element method with nonuniform mesh sizes applied to the exterior Helmholtz problem. *Numer. Math.*, 38(1):61 – 82, 1981. 4

[49] G.H. Golub and C.F. Van Loan. *Matrix Computation*. Johns Hopkins University Press, 3rd edition, 1996. 145, 148

[50] R. Hiptmair, A. Moiola, and I. Perugia. Stability results for the time-harmonic Maxwell equations with impedance boundary conditions. *M^3AS*, 21(11):2263 – 2287, 2011. 10, 31

[51] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Interior penalty method for the indefinite time-harmonic Maxwell equations. *Numer. Math.*, 100:485 – 518, 2005. 4

[52] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*. Springer-Verlag, New York, 1998. 4

[53] F. Ihlenburg and I Babuška. Finite element solution of the Helmholtz equation with high wave number. ii. the h-p version of the FEM. *SIAM J. Numer. Anal.*, 34(1):315 – 358, 1997. 4

[54] F. Ihlenburg and I. Babŭska. Finite element solution of the Helmholtz equation with high wave number. the h-version of FEM. *Comp. Math. Appl.*, 30(9):9 – 37, 1995. 4, 5, 48, 83

[55] A. Ishimaru. *Wave Propagation and Scattering in Random Media.* IEEE Press, New York, 1997. 88

[56] C. Lasser and A. Toselli. An overlapping domain decomposition preconditioner for a class of discontinuous Galerkin approximations of advection-diffusion problems. *Math. Comp.*, 72:1215 – 1238, 2003. 141

[57] R. Leis. *Initial-Boundary Value Problems in Mathematical Physics.* Tübner, 1986. 99

[58] K. Liu and B. Rivière. Discontinuous Galerkin methods for elliptic partial differential equations with random coefficients. *Int. J. Computer Math.*, 90(11):2477 – 2490. 124, 125

[59] J.M. Melenk and S. Sauter. Wavenumber explicit convergence analysis for Galerkin discretization of the Helmholtz equation. *SIAM J. Numer. Anal.*, 49(3):1210 – 1243, 2011. 4

[60] P. Monk. *Finite Element Methods for Maxwell's Equations.* Oxford University Press, New York, 2003. 4

[61] N.C. Nguyena, J. Perairea, and B. Cockburn. Hybridizable discontinuous Galerkin methods for the time-harmonic Maxwell's equations. *J. Comput. Phys.*, 230:7151 – 7175, 2011. 4

[62] L. Nirenberg. Remarks on strongly elliptic partial differential equations. *Comm. Pure Appl. Math.*, 8:648 – 674, 1955. 144

[63] A. J. Nitsche. On Korn's second inequality. *R.A.I.R.O. Anal. Numér.*, 15:237–248, 1998. 21

[64] A.A. Oberai and P.M. Pinsky. A multiscale finite element method for the Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 154(3-4):281 – 297, 1998. 4

[65] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, New York, 1999. 141, 148, 149

[66] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation*. SIAM, Philadelphia, PA, 2008. 146

[67] F.A. Rochinha, G.B. Alvarez, E.F.D do Carmo, and A.F.D. Loula. A locally discontinuous enriched finite element formulation for acoustics. *Comm. Numer. Methods Engrg*, 23(6):623 – 637, 2007. 4

[68] L. Roman and M. Sarkis. Stochastic Galerkin method for elliptic SPDEs: A white noise approach. *Discret. Contin. Dyn. S.*, 6:941 – 955, 2006. 87, 89

[69] I. Rosca. On the Babuška Lax Milgram theorem. In *An. Univ. Bucuresti, XXXVIII*, volume 3, pages 61 – 65, 1989. 144

[70] O. Runborg. Mathematical models and numerical methods for high frequency waves. *Commun. Comput. Phys.*, 2(5):827 – 880, 2007. 4

[71] A.H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959 – 962, 1974. 62

[72] H.A. Schwarz. Über einen grenzbergang durch alternierendes verfahren. *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, 15:272 – 286, 1870. 140

[73] J. Shen and L. Wang. Analysis of a spectral-Galerkin approximation to the Helmholtz equation in exterior domains. *SIAM J. Numer. Anal.*, 45, 2007. 5

[74] B.E. Smith, P.E. Bjørstad, and W.D. Gropp. *Domain Decomposition, Parellel Multilevel Methods for Elliptic Partial Differential Equations.* Cambridge University Press, New York, 1996. 141, 148, 149, 152, 156

[75] R. Tezaur and C. Farhat. Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems. *Internat. J. Numer. Methods Engrg.*, 66(5):796 – 815, 2006. 4

[76] L.L. Thompson and P.M. Pinsky. A Galerkin least-squares finite element method for the two-dimensional Helmholtz equation. *Internat. J. Numer. Methods Engrg.*, 38(3):371 – 397, 1995. 4

[77] A. Toselli and O. Widlund. *Domain Decomposition Methods - Algorithms and Theory.* Springer, 2005. 7, 140, 141, 148, 149, 152, 156, 158, 171

[78] L. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, 1997. 188

[79] H. Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. part i: linear version. *IMA J. Numer. Anal.*, 2013. 49

[80] D. Xiu and G. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24:619 – 644, 2002. 87, 89

[81] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581 – 613, 1992. 141, 148, 149, 152, 156, 158

[82] J. Xu and X.C. Cai. A preconditioned GMRES method for nonsymmetric or indefinite problems. *Math. Comp.*, 59(200):311 – 319, 1992. 149

[83] J. Xu and L. Zikatanov. The method of alternating projections and the method of subspace corrections in hilbert space. *J. Amer. Math. Soc.*, 15:573 – 597, 2002. 141, 152

Vita

Cody Samuel Lorton was born in Pomona, California to George and Tamara Lorton in 1985. He graduated from Barren County High School in Glasgow, KY in 2003. In the fall of 2003, he began attending Western Kentucky University in Bowling Green, KY. He graduated with a Bachelor of Arts degree in Mathematics with a minor in computer science in the Spring of 2007. He continued at Western Kentucky as a graduate student and graduated with a Master of Science degree in Mathematics in the Spring of 2009. In the fall of 2009, Cody began attending the University of Tennessee in Knoxville, TN as a graduate student in Mathematics. Cody graduated with a Doctor of Philosophy degree in Mathematics in the Summer of 2014.