Asymptotic Bounds to Outputs of the Exact Weak Solution of the Three-Dimensional Helmholtz Equation

Shahin Ghomeshi

A Thesis

in

The Department

of

Mechanical and Industrial Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Mechanical Engineering) at
Concordia University
Montreal, Quebec, Canada

March 2010

Your file Votre référence
ISBN: 978-0-494-80163-5
Our file Notre référence
ISBN: 978-0-494-80163-5

ABSTRACT

Asymptotic Bounds to Outputs of the Exact Weak Solution of the Three-Dimensional Helmholtz Equation

Shahin Ghomeshi, Ph.D.

Concordia University, 2010

In engineering practice, the design is based on certain design quantities or "outputs of interest" which are functionals of field variables such as displacement, velocity field, or pressure. In order to gain confidence in the numerical approximation of "outputs," a method of obtaining sharp, rigorous upper and lower bounds to outputs of the exact solution have been developed for symmetric and coercive problems (the Poisson equation and the elasticity equation), for non-symmetric coercive problems (advection-diffusion-reaction equation), and more recently for certain constrained problems (Stokes equation). In this thesis we develop the method for the Helmholtz equation.

The common approach relies on decomposing the global problem into independent local elemental sub-problems by relaxing the continuity along the edges of a partitioning of the entire domain, using approximate Lagrange multipliers. The method exploits the Lagrangian saddle point property by recasting the output problem as a constrained minimization problem. The upper and lower computed bounds then hold for all levels of refinement and are shown to approach the exact solution at the same rate as its underlying finite element approach. The certificate of precision can then determine the best as well as the worst case scenario in an engineering design problem. This thesis addresses bounds to outputs of interest for the complex Helmholtz equation. The Helmholtz equation is in gen-

eral non-coercive for high wave numbers and therefore, the previous approaches that relied on duality theory of convex minimization do not directly apply. Only in the asymptotic regime does the Helmholtz equation become coercive, and reliable (guaranteed) bounds can thus be obtained. Therefore, in order to achieve good bounds, several new ingredients have been introduced. The bounds procedure is firstly formulated with appropriate extension to complex-valued equations. Secondly, in the computation of the inter-subdomain continuity multipliers we follow the FETI-H approach and regularize the system matrix with a complex term to make the system non-singular. Finally, in order to obtain sharper output bounds in the presence of pollution errors, higher order nodal spectral element method is employed which has several computational advantages over the traditional finite element approach. We performed verification of our results and demonstrate the bounding properties for the Helmholtz problem.

Acknowledgements

First and foremost, I would like to thank my advisor, Professor Marius Paraschivoiu, for his continual supply of ideas, as well as for his encouragement and patience during the course of this research. Secondly, I am grateful to Professor Frédéric Magoulès from Ecole Centrale Paris for his feedback and insight into my work, which was not only very helpful, but also very encouraging. I would also like thank my committee members at Concordia university for their useful feedback and advice on the thesis. These include Professors Georgios Vatistas, and Ramin Sedaghati from the mechanical engineering department and Professor Tien Bui from the computer science department. I am also extremely thankful for the warm atmosphere of the CFD lab where many of my colleagues there provided me with very fruitful discussions. Finally, I could not have overcome the many frustrations and disappointments experienced during the course of my studies without the support of my family and friends. I would especially like to thank my parents Eshrat and Behrang, whose constant love and encouragement gave me the strength and encouragement when times seemed most difficult.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Acoustics refers to the transmission of sound through solid and fluid media. Sound can be described by pressure oscillations in an elastic medium resulting from the vibrations imparted to that medium. These oscillations are traveling waves, which act as disturbances generated by space-time evolution of mechanical perturbations in a fluid (which produce sound waves) or solid (which produce elastic waves such as in the plucking of a violin string). As a result, the mathematical models describing acoustic phenomena are derived from the linearization of the equations of continuum mechanics, where only the first order terms are retained after making small perturbations to some ambient values for the velocity, density and pressure. There are three phenomena encountered in acoustics, the most important of which is the wave propagation, and is also the only one which occurs in an infinite homogeneous medium. The second, is scattering, which can occur due to various obstacles encountered by the wave. Finally, the third phenomenon deals with dissipation processes that can cause the absorption and dispersion of waves.

Wave propagation, in general has been studied in numerous scientific fields. For example, the simulation of vibro-acoustic problems for mufflers and silencers can be approximated by linear time-harmonic wave propagation, which is governed by the Helmholtz's equation. The search for efficient computational approaches for solving such problems more accu-

rately have been going on for decades. These efforts have contributed to the developments of many advanced noise control technology. The work herein is focused on the development of a numerical approach in solving the linear time-harmonic wave propagation problems, where the acoustic phenomena occurs in enclosures. The numerical procedure used, enables more accurate simulations of the phenomena. However, every numerical result contains discretization errors and we often do not know quantitatively, how large is this error. Error estimation procedures are therefore used in assessing the accuracy of the numerical simulation result which can then quantify the accuracy of the solution. Ultimately, the aim of the error estimation is to deliver certainty information, which will bolster trust in the numerical simulation.

1.1 Acoustic Wave Propagation Problem

In this section we present the basic assumptions leading to the model equation for time harmonic wave equation, in particular the Helmholtz equation. Moreover, we briefly discuss some typical boundary value problems which frequently occur in practical applications. For more details on the subject of acoustics, the reader can consult [30, 54]. We begin with the following assumptions:

- The unperturbed values of pressure, density, temperature, and velocity are assumed to be time independent and are given by (p, ρ, T, u);
- An acoustic signal passing through a fluid is small perturbation of the pressure, density, temperature, and velocity and are expressed as $p + \widetilde{p}$, $\rho + \widetilde{\rho}$, $T + \widetilde{T}$, and $\widetilde{u}$. The unperturbed velocity u does not undergo macroscopic motion and is set to zero. The assumption on these perturbed values are $\widetilde{p} \ll p$, $\widetilde{\rho} \ll \rho$, and $\widetilde{T} \ll T$.
- Sound transmission through the fluid is sufficiently transient that there is no time

for heat transfer to occur resulting in an adiabatic process. Furthermore, no energy losses occur due to friction or dissipative effects which indicate that the flow through the system is also reversible.

- No external forces act on the flow and that the fluid is inviscid and that the pressure depends on the density, eg. $p = g(\rho)$.

Under these assumptions, expanding the continuity and momentum equations in terms of the perturbed values and applying the fact that we have an isentropic (adiabatic and reversible) process with an ideal gas law (compressible flow) we arrive at the linear wave equation

$$\frac{\partial^2 \widetilde{p}}{\partial t^2} - c^2 \nabla^2 \widetilde{p} = 0, \tag{1.1}$$

where c is the speed of sound in an acoustic medium and is $c^2 = \frac{\partial p}{\partial \rho}\Big|_{isentropic}$ evaluated at the background density. Equation (1.1) is a hyperbolic equation and thus it represents a non-dispersive wave. For a thorough discussion of hyperbolic and dispersive waves, the reader is referred to [63]. Much of real life acoustic phenomena have a periodic (or quasi periodic) time dependency and are composed of a linear combination of harmonic components. Therefore, for time harmonic (steady state) waves, we write

$$\widetilde{p}(\mathbf{x}, t) = \Phi(\mathbf{x}) e^{-i\omega t}, \quad i = \sqrt{-1}$$

in which the wave equation reduces to the Helmholtz equation:

$$\nabla^2 \Phi + k^2 \Phi = 0, \tag{1.2}$$

where $k = \omega/c$ is the wave number, and ω is the angular frequency.

1.1.1 Model Problems in Acoustics

Here we present a review of certain kinds of applications which exist in acoustic wave propagation problems in fluids. The applications can be either for interior problems or

exterior problems where the model equation is the Helmholtz equation, however the different situations lead to different boundary conditions. In the following, we let Ω be a bounded open domain of $\mathbb{R}^3$, with boundary $\partial\Omega$. Moreover, the normal derivatives at the boundaries are given by $\frac{\partial\Phi}{\partial\mathbf{n}} = \mathbf{n}\cdot\nabla\Phi$ in the direction of the outward normal vector $\mathbf{n}$ to $\partial\Omega$.

Interior problems

Interior problems deal with acoustic phenomena in enclosed regions of space, examples may include a cavity, room acoustic, or mufflers and silencers. In vibro-acoustics, the boundary value problems consist of finding the spatial components of the acoustic pressure field $\Phi : \overline{\Omega} \mapsto \mathbb{C}$ such that:

$$\begin{cases} \nabla^2\Phi + k^2\Phi = 0, & \text{in } \Omega \quad (\mathbf{a}) \\ \Phi = \Phi_0, & \text{on } \Gamma_D \quad (\mathbf{b}) \\ \frac{\partial\Phi}{\partial\mathbf{n}} = i\rho ckv_n, & \text{on } \Gamma_N \quad (\mathbf{c}) \\ \frac{\partial\Phi}{\partial\mathbf{n}} - i\rho ckG\Phi = g, & \text{on } \Gamma_R \quad (\mathbf{d}) \end{cases} \tag{1.3}$$

where $\overline{\Gamma_D \cup \Gamma_N \cup \Gamma_R} = \partial\Omega$, and $\Gamma_D \cap \Gamma_N = \Gamma_N \cap \Gamma_R = \emptyset$. The boundary conditions (1.3(**b**)) is the part of the boundary $\Gamma_D \subset \partial\Omega$ which contains the Dirichlet boundary conditions and consists of a simple pressure release condition $\Phi = \Phi_o$. The part of the boundary $\Gamma_N \subset \partial\Omega$ is the Neumann boundary condition (1.3(**c**)) where the wall is rigid and vibrates with normal velocity v_n. Finally, for vibro-acoustic problems, the vibrational parts of the wall introduce a forcing term(see Figure 1.1), and so the complex-valued function g is prescribed in the general Robin boundary condition (1.3(**d**)). Here G is the field admittance in the normal direction and is related to the impedance $Z = 1/G$. In an acoustic medium, the force is generated by a change in the pressure; and the ratio of this change in pressure by the velocity in the normal direction is defined as the impedance Z. G depends on the nature of the enclosure. For example, in the case of homogenous Robin boundary conditions (i.e.

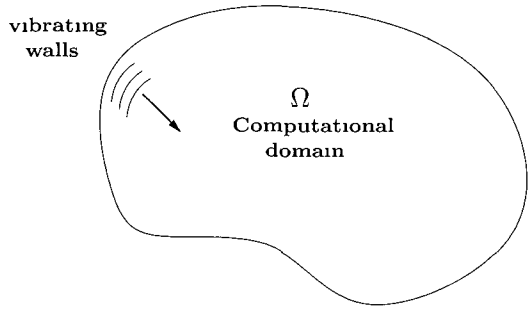


Figure 1 1 Computational domain Ω for a vibro-acoustic problem (picture adapted from [46])

$g = 0$), having $G = 0$, will imply that the walls are rigid with only Neumann boundary conditions $\frac{\partial \Phi}{\partial \mathbf{n}} = 0$ prescribed If $G \to \infty$, then the wall is considered to be acoustically soft and one then obtains the homogenous Dirichlet boundary conditions $\Phi = 0$ For the case $0 < G < \infty$, the wall acts as an absorbing surface and so we have an absorbing boundary condition We note that the boundary value problem Equation (1 3) is very general

Exterior problems

Exterior problems may include radiation, scattering or transmission problems which involve the characterization of the acoustic field surrounding a given structure For these type of problems, the computational domain is unbounded in space As an example, we consider a radiation problem to find the radiated acoustic pressure field $\Phi \quad \overline{\Omega} \mapsto \mathbb{C}$ such that

$$\begin{cases} \nabla^2 \Phi + k^2 \Phi = 0, & \text{in } \Omega \qquad (\mathbf{a}) \\ \frac{\partial \Phi}{\partial \mathbf{n}} = \imath \rho c k v_n, & \text{on } \partial D \qquad (\mathbf{b}) \\ \frac{\partial \Phi}{\partial \mathbf{n}} - \imath \rho c k \Phi = \mathsf{q}, & \text{on } \partial B \qquad (\mathbf{c}) \end{cases} \tag{1 4}$$

It is assumed that the region $D \subset \mathbb{R}^3$ occupied by the body is embedded in a homogeneous isotropic medium at rest Here D is a bounded simply connected domain with boundary ∂D If the walls of the body vibrate with a normal velocity v_n and that radiated waves propagate into free space then the physical requirement is that the radiated waves cannot

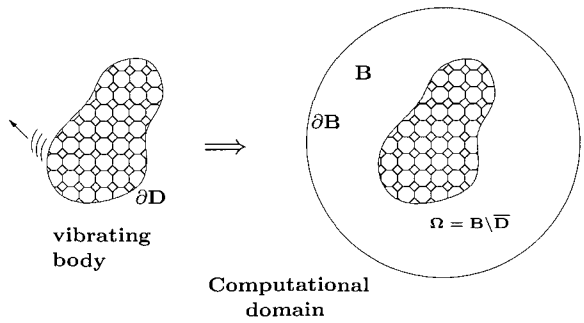


Figure 1.2: Radiation problem for a vibrating body D and computational domain Ω (picture adapted from [46])

reflect at infinity. Such a requirement leads to the Sommerfield radiation condition given by

$$\lim_{r\to\infty} r\left(\frac{\partial\Phi}{\partial r} - \imath k\Phi\right) = 0. \tag{1.5}$$

where $r = |\mathbf{x}|$ and $\partial/\partial r$ is the derivative in the radial direction. The Sommerfield radiation condition requires that the Helmholtz equation be solved in an infinite domain. Since computational techniques such as the finite element method require a bounded domain, a sufficiently large ball $B \subset \mathbb{R}^3$ with boundary ∂B is introduced that contains D (see Figure 1.2). One can approximate the Sommerfield condition at infinity by Robin (non-reflecting) boundary condition on ∂B using the Dirichlet-to-Neumann map (DtN) technique in Chapter 3 of [34] by Equation (1.4(**c**)), namely,

$$\frac{\partial\Phi}{\partial\mathbf{n}} - \imath k\Phi = \mathsf{q} \quad \text{on} \quad \partial B,$$

where $\mathbf{n}$ is the unit normal to ∂B. With this boundary condition, the computational domain is given by $\Omega = B\backslash\overline{D}$ with boundary $\partial\Omega = \partial B \cup \partial D$. In addition to the given boundary conditions, the pressure release condition $\Phi = \Phi_\circ$, and the absorbing condition $\frac{\partial\Phi}{\partial n} - \imath\rho ckG\Phi = g$ may also be prescribed on parts of the boundary.

1.1.2 Numerical Challenges

Solving the Helmholtz problem by the finite element method results in stability problems associated with the Helmholtz operator, in particular the loss of ellipticity for high wave numbers. Such stability issues are a consequence of the pollution effect for high wave numbers which is associated with the phase error. Regardless of the stability properties of the finite element method used in the discretization, a certain minimal number of elements per wave length is always required to correctly represent the physical phenomena and to obtain a fully resolved numerical solution. However, as discussed in [34], due to the pollution effect, it is difficult to meet the appropriate resolution requirements for higher frequencies. This is mainly because in order to obtain a solution with high precision, very fine meshes are often necessary which will in turn lead to a significantly large-scale system of equations that must be solved. Higher-order methods however, have been shown [36] to significantly reduce the pollution error. A comparison between the traditional finite element method and the higher-order nodal spectral element methods for applications to mufflers and silencers is given in [41]. The approach that we will undertake is on employing the higher-order methods with a domain decomposition procedure. Therefore in addition to higher-order accuracy, the domain decomposition procedure is not only required in the bounds approach, but will also make the method amenable to parallel computation.

1.2 Error Estimation

A major drawback of all computational results obtained from pde models of an event, is the presence of numerical errors which leads to uncertainty. Since much of the present day technology depends upon simulation based engineering design, modeling and simulation predictions can have a significant impact. Consequently, for the engineering designers and practitioners, the credibility of the computational result becomes of great concern. Knowl-

edge of such errors, however, can provide a means for assessing the accuracy and reliability of the computation.

The primary means by which to assess the accuracy and reliability in computational simulations are through the process of verification and validation (V&V). An extensive review of verification and validation in computational fluid dynamics is discussed in [42], but here we briefly highlight the differences between the two. In verification studies, the accuracy of a computational solution is measured relative to either an exact solution (if available or by construction), or a very highly accurate solution of the mathematical model. The strategy then becomes a purely mathematical and computer science issue where the objective is to identify, quantify, and then reduce the error between the computed solution and the true solution of the model. In the validation process, the goal is to assess how accurately the computed results measure in relation to the experimental data (if available) or the real world, which is both a mathematical and physical science issue. Therefore, verification provides evidence that the continuum model is solved correctly by the discretized formulations, while validation provides evidence of how the computed results simulate reality, and in particular, it addresses the question of the fidelity of the model to the specified real world problem. However, the terms "evidence" and "fidelity" in a computed solution, bring about the important concept of error estimation, which can be based either on *a priori* error estimates, or *a posteriori* error estimates.

A priori information relies on the part of the numerical algorithm which is associated with the partial differential operators and its initial and boundary data; and does not rely on direct knowledge of the solution of the PDE. *A priori* estimates generally involve normed estimates of the error which demonstrate stability and convergence of the approximation, and that are also used in facilitating existence theorems. *A priori* estimates for the Helmholtz equation can be readily found in [60, 61]. Although such estimates reveal

the correct asymptotic rate of convergence, they do involve norms of the unknown solution and thus are of limited use if one requires numerical estimate of the accuracy. *A posteriori* information on the other hand, does require knowledge of the solution of the relevant PDE. Hence, *a posteriori* error estimates are expressed in terms of the computed solution, using the output of the Finite element computation to assess the accuracy, which enables the quantification of the error. Examples of *a posteriori* error estimators for the Helmholtz equation are available in [5, 35] .

In order to predict the accuracy of the numerical solution using *a priori* estimation in the verification process, will generally require spatial convergence studies which can become computationally very expensive especially for three dimensional problems. The application of *a posteriori* estimators forms the basis for effective control of adaptive grid strategies which can be used to reduce the computational expense. However, adaptive methods have the shortcoming that a desired error reduction target is not guaranteed, and in order to gain reliability in the numerical solution, efficient error estimators are necessary, but the questions "what error is relevant" and "will an adaptive mesh refinement lead to a desired solution accuracy" must be answered.

Much of the present interest in *a posteriori* error estimation began with the work of Babuska and Rheinboldt [7] in 1978 on the development of rigorous global error bounds for finite element approximations of the linear elliptic two-point boundary value problems. Here, *a posteriori* error estimation techniques are then used in approximating the error in energy or an energy norm on each finite element K. These estimates formed the basis for adaptive mesh techniques used in the control and minimization of the error. Much of the recent work on error estimation is based on the idea of using complementary energy techniques to obtain bounds. Prior to the work of Babuska and Rheinboldt, de Veubeke [62] introduced the idea of obtaining upper and lower bounds for the energy norm of the error minimizing

the complementary energy of a global dual approximation. However, the method failed in popularity because the error estimates were based on global calculations. Ladevèze and Lequillon [38] advanced the idea of using complementary energy formulations by proposing the construction of solving local dual problem in conjunction with the equilibrated fluxes as boundary conditions, in order to avoid the global computation of the dual approximation. This approach is widely referred to as the *equilibrated residual* method, but it has also been given the name *hybrid-flux* residual method. The notion of obtaining equilibrated fluxes or tractions through a post-processing of the finite element approximation have also been pursued by Bank and Weiser [9]. They obtained estimators for affine approximation on linear triangular elements by solving a primal problem. From their numerical results they then conjectured that the resulting estimators always lead to an upper bound on the error. This conjecture, was then proved later by Ainsworth and Oden [2] for general *hp*-finite element approximations. For a more detailed review of *a posteriori* estimates, the reader is encouraged to consult [3] and [64] for more recent review. However, we briefly mention that the works described above belong to a large family of methods within the general class of *a posteriori* estimation called implicit residual methods, which are based on a series of local problems with appropriate boundary conditions. Other methods within this class also exists such as the constitutive relation error developed by Ladevèze and co-workers [39]; and on recovery-based methods developed by Zienkiewicz and Zhu [65] for a certain types of problems and finite element approximations.

These original implicit methods [2, 9, 38] were developed for linear self-adjoint operators, where the aim was to provide bounds for the energy norm of the error. In practice however, one is seldom interested in the error in the energy norm. A latter study on the investigation of possible extensions of error norms [6], lead to the idea of error estimators associated with engineering outputs that are often referred to in the literature as "goal ori-

ented" error estimation and which was further extended in the late 1990's [44, 48, 50, 53] These quantities of interest which may include the flux across a boundary, the normal force on a surface, heat transfer, transmission loss etc are from a functional analysis context, manifested as functionals of field variables such as the velocity field or displacement, temperature, and pressure While the earlier work was based on estimating the error $e = u_h - u_{ex}$ (where u_h is an approximate solution and u_{ex} is the exact solution) measured in the energy norm, the work of Oden and Prudhomme [44, 53] exploited this idea to obtain quantitative estimates for the error in quantities of interest The various work of Paraschivoiu et al [48, 50, 49], employed a different technique towards goal-oriented error estimation, where they obtained computable upper and lower bounds consisting of solving the problem of interest using two discretization schemes of different accuracy and using the difference in the approximations as an estimate for the error This is referred to as the two-level residual based approach where the bounds are obtained on a fine mesh (or "*truth mesh*") based on coarse mesh global solves Initially bounds on quantities of interest was applied to symmetric elliptic problems (i e the Poisson equation) [50], non-symmetric coercive problems (i e convection-diffusion equation) [48], certain constrained problems (i e the Stokes problem) [49, 43, 45], and non-coercive problems (i e 2D Helmholtz equation) [57] More recently though, goal oriented estimation techniques have been advanced in the study of transient parabolic problems [21], and also acoustic wave propagation problems [46]

It was not until the early to mid 2000's that the focus on employing *a posteriori* error estimation shifted from obtaining estimates using local computations for better adaptive strategies to focusing more on employing the error estimation techniques in order to obtain certainty information The work on guaranteed bounds on exact outputs called *"exact bounds"* were first proposed in [58, 59, 51] and for time dependent outputs [13] However very recently, the work was developed for the Stokes problem [15], and through a different

approach, strict error bounds have been obtained with improved effectivities for the Poisson equation [52], and for applications to linear solid mechanics problems [19]. The upper and lower computed bounds holds for all levels of refinement and are shown to approach the exact quantity of interest at the same rate as its underlying finite element approach. This certificate of precision can determine the best as well as the worst case scenario in an engineering design problem. Moreover, the method can be termed "cost effective" as it can be used to determine the size of the mesh required to achieve a desired level of accuracy. This approach also answers the initially stated question where error in a design quantity is important.

The strategy involved in the computing of bounds on exact outputs of interest is similar to the former hierarchical method [48] in that it involves decomposing the global mesh into several elemental subdomains and relaxing the continuity requirements along the edges of each subdomain. A Lagrangian is first constructed so that the output problem is recasted as a constrained minimization problem where the constraints are the continuity requirements along the edges of the subdomains and the equilibrium equation. The gradient condition of the Lagrangian will then lead to the primal-adjoint pair and the equilibration equation that will determine the candidate inter-element continuity multipliers. The bounds are finally obtained through local sub-problem calculations. At this stage, the method differs from the former two-level residual method because by exploiting the Lagrangian saddle point property, existence of such bounds on the exact solution output is guaranteed, however the bounds are practically un-computable. The key ingredient relies on constructing a complementary energy functional chosen from a suitable finite dimensional set that can be used to bound the infinite dimensional problem [58, 59, 15]. Finally, we point out that amongst the implicit residual methods, there exists two approaches that have been employed in the computation of the exact bounds. One is the approach undertaken in [58, 59, 15] for

obtaining exact bounds, which is based on hybrid-flux methods where the local problems are element based (subdomains are non-intersecting) and the other approach used in [19, 52] is based on flux-free methods where the subdomains are patches of elements. A comparison between the two methods has been done in [17].

1.3 Objectives and Scope of Thesis

Our goal in the thesis is to develop the formulation of the exact bounds method for the three-dimensional Helmholtz equation based on the work developed in [59] via the hybrid-flux approach. The work here also includes several new ingredients as compared to previous work in this field. Firstly, the exact bounds procedure is formulated with particular emphasis on appropriate extension to complex-valued equations. A key ingredient to the bounds method is based on decomposing the global mesh into several elemental subdomains and relaxing the continuity requirements along the edges of each subdomain. A Lagrangian is first constructed in such a way that as the error in the approximation is minimized (goes to zero in the limit), the output of interest is obtained. Such a Lagrangian contains the continuity requirements at the edges of each subdomain and the equilibrium equation as constraints. For the Helmholtz equation, the Lagrangian is modified by adding a complex lumped interface mass matrix as the interface problem associated with solving the equilibrated fluxes for the Helmholtz equation can become singular. Such an approach follows the work done in [26] where the additional complex regularizing term will result in a non-singular system of equations. The current approach in calculating the equilibrated flux has an additional advantage that it avoids the need for global calculations as in previous approaches, moreover, the computations are calculated locally and consequently, are intrinsically parallel. Lastly, in order to obtain more accurate solutions for higher wave numbers, high-order finite element method (nodal spectral element method) has been incorporated

with the exact bounds approach.

In Chapter 2 we will describe some of the technical aspects associated with the computational method for the Helmholtz equation where we again discuss the numerical challenges involved. In chapter 3 we will present the domain decomposition procedure that we will use in the bounds method. Chapter 4 will go over the theory of the exact bounds method and its application to outputs of the Poisson equation. In Chapter 5 we will present the formulation for the method as applied to the complex Helmholtz equation. Finally Chapter 6 will contain the discussion of the results, and Chapter 7 will conclude with possible ways to improve the method and potential extensions to practical engineering applications.

Chapter 2

Numerical Methods for the Helmholtz Equation

In pursuit of our objectives, we proceed with the mathematical model describing acoustic wave propagation in fluids, in particular, vibro-acoustic problems in three-space dimensions. It was previously mentioned that upon considering time-harmonic acoustics (steady state) waves, then the model of interest is the Helmholtz equation. Although there exists an abundant of numerical techniques to solve the Helmholtz equation, we will confine ourselves to polynomial based methods such as the Galerkin Finite Element method (FEM) and higher-order nodal spectral element method (SEM). Moreover, a review of certain technical issues pertaining to the numerical simulation of acoustic phenomena will be discussed, as this will also in part, give cause to our choice of a higher-order polynomial based numerical method.

2.1 Model Problem

2.1.1 Strong Formulation

We are concerned with interior problems where the acoustic phenomena occurs in enclosed regions of space, for potential applications to mufflers and silencers. Vibro-acoustic problems, deal with forced acoustic fields due to the vibrational parts of the enclosed walls.

Therefore, with an external acoustic source term denoted by function $f(x, y, z)$ we write the strong statement for the general complex Helmholtz equation as: finding the spatial component of the acoustic pressure $\Phi : \overline{\Omega} \mapsto \mathbb{C}$ such that

$$-\nabla^2\Phi - k^2\Phi = f \quad \text{in domain } \Omega \tag{2.1}$$

$$\Phi = g_D \quad \text{on boundary } \Gamma = \Gamma_D. \tag{2.2}$$

where we have assumed for simplicity of presentation that we have only Dirichlet boundary conditions and g_D is the boundary data on Γ. For notational simplicity, we will throughout our presentation designate the acoustic pressure Φ with u.

2.1.2 Preliminaries from Functional Analysis

The discretization process associated with the finite element method, involves reformulating the given differential equation as an equivalent variational problem. For elliptic problems this is transformed into a minimization statement of the form

$$\text{Find } u \in V \text{ such that } \quad F(u) \leq F(v) \quad \text{for all } v \in V,$$

where V is a given set of admissible functions and $F : V \mapsto \mathbb{R}$ is a functional representing the total energy associated with the functions $v \in V$. The gradient condition of such functional leads to the variational weak form of the model problem, which is discretized by the finite element method. The functions v often are continuously varying quantities, however when working with variational formulations of boundary value problems for partial differential equations it is natural to work with spaces which are larger (contain more functions) then the spaces of bounded continuous functions. Therefore, we introduce a special category of Sobolev spaces namely, the Hilbert Spaces and allow V to be a Hilbert space, where the most common of these are the spaces $H^0(\Omega) \equiv L^2(\Omega), H^1(\Omega), H_0^1(\Omega), H^2(\Omega)$, and $H_0^2(\Omega)$. For the convenience of the reader we give a brief description of these spaces, however for

details the reader is referred to the many references [1, 55]. Throughout we shall assume that the problems are posed in a domain Ω of $\mathbb{R}^3$, with a sufficiently smooth boundary $\partial\Omega = \Gamma$. Now we define $L^2(\Omega)$ to be the space of square integrable functions on Ω:

$$L^2(\Omega) = \left\{ v \Big| \int_\Omega |v|^2 \, dx = \|v\|^2_{L^2(\Omega)} \leq +\infty \right\}$$

and in general, for integer $m > 0$ the spaces $H^m(\Omega)$ are:

$$H^m(\Omega) = \left\{ v \Big| D^\alpha v \in L^2(\Omega), \forall |\alpha| \leq m \right\},$$

where

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}, \ |\alpha| = \alpha_1 + \cdots + \alpha_n,$$

these derivatives being taken in the sense of distributions. On these spaces we shall use the semi-norm

$$|v|^2_{H^m} = \sum_{|\alpha|=m} \int_\Omega |D^\alpha u|^2 \, dx$$

and the norm

$$\|v\|^2_{H^m} = \sum_{|\alpha|\leq m} \int_\Omega (D^\alpha u)^2 \, dx = \sum_{s\leq m} |v|^2_{H^s}.$$

Since we deal with boundary value problems, we are not only concerned with the value of functions on open domains Ω, but also with the value of the functions on the boundary Γ. According to the Trace theorem (see [1, 55]), for a sufficiently smooth boundary $\partial\Omega$, there exists a unique bounded linear operator $\gamma : H^1(\Omega) \mapsto L^2(\Gamma)$ for every v smooth (say for $v \in C^1(\overline{\Omega})$). $\gamma(v)$ is then called the trace of v on Γ and written as $v|_\Gamma$, even if v is a general function in $H^1(\Omega)$. A more involved analysis indicates that the range of γ is not all of $L^2(\Gamma)$, only a portion of it (i.e. a subspace of $L^2(\Gamma)$); moreover, such a subspace contains $H^1(\Gamma)$ as a proper subset. Hence

$$H^1(\Gamma) \subset \gamma(H^1(\Omega)) \subset L^2(\Gamma) \equiv H^0(\Gamma),$$

and so it can be seen that $\gamma(H^1(\Omega))$ belongs to the space $H^{1/2}(\Gamma)$ for functions $v \in H^1(\Omega)$. More precisely we have

$$H^{1/2}(\Gamma) = \gamma(H^1(\Omega))$$

and the norm for functions g_b defined on part of the boundary is given as

$$\|g_b\|_{H^{1/2}(\Gamma)} = \inf_{\substack{v \in H^1(\Omega) \\ \gamma v = g_b}} \|v\|_{H^1(\Omega)}.$$

More generally though, for $m \geq 1$ and $u \in H^m(\Omega)$, the trace will have the mapping $\gamma : H^m(\Omega) \mapsto H^{m-1/2}(\Gamma)$. We now define the spaces

$$H_0^1(\Omega) = \left\{ v | v \in H^1(\Omega), v|_\Gamma = 0 \right\},$$

and

$$H_0^2(\Omega) = \left\{ v | v \in H^2(\Omega), v|_\Gamma = 0, \frac{\partial v}{\partial n}\Big|_\Gamma = 0 \right\}$$

We will see in the next subsection that the variational weak form of the model problem is written as

$$b(u, v) = \ell(v)$$

where the form $b : V_1 \times V_2 \mapsto \mathbb{K}$, and $\ell : V_2 \mapsto \mathbb{K}$ for $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Depending on whether or not we are dealing with the complex Helmholtz equation , the mapping $b(u, v)$ will be bilinear, or sesquilinear. It is bilinear (occurs in the case $\mathbb{K} = \mathbb{R}$) if it is linear in both arguments, but sesquilinear (occurs in the case $\mathbb{K} = \mathbb{C}$)if it is linear in the first argument and antilinear in the second, namely if

$$b(\alpha(u_1 + u_2), v) = \alpha(b(u_1, v) + b(u_2, v))$$

$$b(u, \alpha(v_1 + v_2)) = \overline{\alpha}(b(u, v_1) + b(u, v_2))$$

for complex number α, and its complex conjugate $\overline{\alpha}$. Although our demonstration of the bounds method has been applied to a more theoretical problem where the solution u is real, our formulation and the C code used allow for general complex valued functions, for this reason we often refer to the forms used as sesquilinear or antilinear.

The spaces V_1 and V_2 for the weak formulation, are typically $H^1(\Omega)$ and $H_0^1(\Omega)$ respectively, and the operator $\ell(v)$ is a bounded antilinear operator on the vector space $H_0^1(\Omega)$ belonging to the space of bounded antilinear functionals with the name dual space denoted by a prime with the convention:

$$[H_0^1(\Omega)]' \equiv H^{-1}(\Omega).$$

Remark 2.1 *For bounded linear functionals the action of a functional ℓ on an element v is more correctly denoted by $\langle \ell, v \rangle$ rather than $\ell(v)$, however in certain areas where it is convenient the later notation may be used.*

2.1.3 Notations and Weak Formulation

For any complex number $z \in \mathbb{C}$, let $\overline{z}$ be the complex conjugate of z, and $|z|$ its modulus. Now introduce the space

$$\widetilde{Z}(\Omega) = \left\{ v = v^R + \imath v^I : v^R \in H^1(\Omega), v^I \in H^1(\Omega) \right\}, \tag{2.3}$$

Superscript R and I are the real and imaginary parts, respectively, that is, $v^R = \Re(v)$ and $v^I = \Im(v)$. We then introduce the sets

$$Z = \left\{ v \in \tilde{Z}(\Omega) : v|_{\Gamma_D} = 0 \right\} \tag{2.4}$$

$$Z_D = \left\{ v \in \tilde{Z}(\Omega) : v|_{\Gamma_D} = g_D \right\} \tag{2.5}$$

which reflect the essential boundary conditions. We proceed by introducing the weak formulation of problem (2.1)–(2.2). For a sufficiently smooth function $v : \overline{\Omega} \mapsto \mathbb{C}$ belonging

to the space Z, we integrate the Helmholtz equation against a test function $\overline{v}$ (where we recall the over line designates the complex conjugate of v). Upon integrating by parts and imposing the boundary conditions, we obtain the weak formulation of equation (2.1), as follows:

find $u \in Z_D$, such that

$$\mathcal{A}(u, v) = 0, \quad \forall v \in Z, \tag{2.6}$$

where the form $\mathcal{A} : \tilde{Z}(\Omega) \times \tilde{Z}(\Omega) \mapsto \mathbb{C}$ is defined as

$$\mathcal{A}(u, v) = a(u, v) - m(u, v) - \langle f, v \rangle, \tag{2.7}$$

and the sesquilinear forms are given by:

$$a(w, v) = \int_\Omega \nabla w \cdot \nabla \overline{v} \, d\Omega, \tag{2.8}$$

$$m(w, v) = k^2 \int_\Omega w \overline{v} \, d\Omega, \tag{2.9}$$

with the duality pairing

$$\langle f, v \rangle = \int_\Omega f \overline{v} \, d\Omega. \tag{2.10}$$

Remark 2.2 *We note that in order to be consistent with the definition of the inner product on complex vector spaces (where we have sesquilinear forms), the complex conjugate of v, namely, $\overline{v}$ is used as the test function.*

2.1.4 Linear Functional Outputs

Engineering design is based on the prediction of certain quantities of interest that are generally expressed as functionals of the field variables. In the present work we are interested in real outputs s, that measure the acoustic performance of mufflers and silencers, however we will simplify the matter by expressing the outputs as linear functionals of the solution $u = u^R + \imath u^I$. More generally though, we set $s = \Re\{S(u)\}$, where $S(u) : \tilde{Z}(\Omega) \mapsto \mathbb{C}$. We will give details of the specific outputs used, in the following chapters.

2.2 Discretization Methods

In this section we briefly describe the finite element method and the nodal spectral element method which are used in the approximation to the Helmholtz equation. The first step consists of making a partitioning of the domain Ω into a finite number $\mathcal{N}_{e\ell}$ of tetrahedrons $\Omega^1, \Omega^2, \cdots \Omega^{\mathcal{N}^{e\ell}}$ with the property that the elements are nonoverlapping and cover Ω so that:

1.
$$\overline{\Omega} = \bigcup_{k_e=1}^{\mathcal{N}^{e\ell}} \overline{\Omega}^{k_e}$$

2.
$$\Omega^{k_e} \cap \Omega^{k_f} = \emptyset \text{ for } k_e \neq k_f.$$

Instead of defining local basis functions for each element Ω^{k_e} with respect to its physical coordinates $\mathbf{x} = (x_1, x_2, x_3)$, the problem is simplified by setting up a standard (reference) tetrahedron $\widehat{\Omega}$ given in terms of a barycentric coordinates system $(\xi_1, \xi_2, \xi_3, \xi_4)$, and having vertices

$$\mathbf{V}_I = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \ \mathbf{V}_{II} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \ \mathbf{V}_{III} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{V}_{IV} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

which is isolated from the actual finite element mesh. We note that actually as $0 \leq \xi_n \leq 1$, $n = 1, 2, 3, 4$, for the reference tetrahedron $\xi_1 + \xi_2 + \xi_3 + \xi_4 = 1$ and only three of these local coordinates are independent. The reference element has the same system of nodal points as the elements Ω^{k_e} and can be used to generate a smooth and invertible map $\Psi^{k_e} : \widehat{\Omega} \mapsto \Omega^{k_e}$ as seen in Figure (2.1). This type of mapping is called an affine mapping and is a function of some elemental local basis functions $h_i(\xi)$ for $i = 1 \cdots \mathcal{N}^{\ell}$. For further explanation on the mappings we refer the reader to [16]. The relation between the physical coordinates

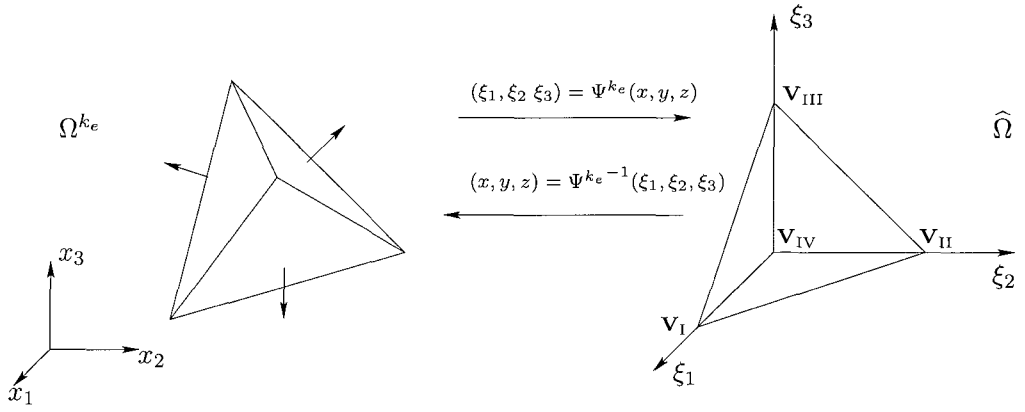


Figure 2.1: Mapping between the tetrahedral element Ω^{k_e} in the partitioned mesh and the reference element $\widehat{\Omega}$ given in terms of barycentric coordinates.

and the local coordinates of the reference element are given by

$$\mathbf{x}(\xi_1.\xi_2, \xi_3) = \sum_{i=1}^{N_\ell} \widehat{\mathbf{x}}_i^{k_e} h_i(\xi_1, \xi_2, \xi_3) \tag{2.11}$$

where N_ℓ is the number of local nodes in each element, and $\widehat{\mathbf{x}}_i^{k_e}$ represents the coordinate of the local node i of element Ω^{k_e}. The local basis functions h_i defined on the reference element are polynomials and are associated with a nodal set $\widehat{\xi}_i$ in $\widehat{\Omega}$. Consequently, the h_i's have the requisite property which allows one to approximate any given function w by polynomial interpolation w_h,

$$w(\xi_1, \xi_2, \xi_3) \approx w_h(\xi_1, \xi_2, \xi_3) = \sum_{i=1}^{N_\ell} w_i h_i(\xi_1, \xi_2, \xi_3). \tag{2.12}$$

with

$$h_i(\widehat{\xi}_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{2.13}$$

Under the approximation (2.12), the problem (2.6) is recast as: find $u_h \in Z_D^h$ such that

$$\mathcal{A}(u_h, v) = 0 \quad \forall v \in Z^h. \tag{2.14}$$

Having partitioned the domain into $\mathcal{N}^{e\ell}$ elements and defined the suitable local basis functions $h_i(\widehat{\xi_j})$ over each element k_e, the integrals over the domain Ω in Equation (2.6) can be rewritten as a sum of integrals over the reference element $\widehat{\Omega}$ where we have for the volume integrations

$$\int_\Omega \nabla u \cdot \nabla \overline{v}\, d\Omega - k^2 \int_\Omega u\overline{v}\, d\Omega = \sum_{k_e=1}^{\mathcal{N}^{e\ell}} \left(\int_{\Omega^{k_e}} \nabla u \cdot \nabla \overline{v}\, d\Omega - k^2 \int_{\Omega^{k_e}} u\overline{v}\, d\Omega \right).$$

Doing a change of variables and using the approximation (2.12), we can transform the integrations one over the reference element $\widehat{\Omega}$, thus obtaining

$$\begin{aligned}
&\int_{\Omega^{k_e}} \nabla u \cdot \nabla \overline{v}\, d\Omega - k^2 \int_{\Omega^{k_e}} u\overline{v}\, d\Omega \\
&= \int_{\widehat{\Omega}} \left[\sum_{m=1}^{3} \left(\sum_{n=1}^{3} \frac{\partial u}{\partial \xi_n} \frac{\partial \xi_n}{\partial x_m} J \right) \left(\sum_{n'=1}^{3} \frac{\partial \overline{v}}{\partial \xi'_n} \frac{\partial \xi'_n}{\partial x_m} J \right) \right] \frac{1}{|J|}\, d\Omega - k^2 \int_{\widehat{\Omega}} u\overline{v}|J|\, d\Omega \\
&\approx \sum_{i=1}^{N_\ell} \sum_{j=1}^{N_\ell} u_i v_j \left(\int_{\widehat{\Omega}} \sum_{m=1}^{3} \left[\left(\sum_{n=1}^{3} \frac{\partial h_i}{\partial \xi_n} \frac{\partial \xi_n}{\partial x_m} J \right) \left(\sum_{n'=1}^{3} \frac{\partial h_j}{\partial \xi'_n} \frac{\partial \xi'_n}{\partial x_m} J \right) \right] \frac{1}{|J|}\, d\Omega - k^2 \int_{\widehat{\Omega}} h_i h_j |J|\, d\Omega \right)
\end{aligned}$$

where J is the Jacobian of the transformation which for straight edged tetrahedrons, is given by $6 \times vol^{k_e}$ where vol^{k_e} is the volume of each element. However more generally it is a function of the reference coordinates ξ_1, ξ_2, ξ_3, and is given by

$$\begin{aligned}
J(\xi_1,\xi_2,\xi_3) \;=\; & \frac{\partial x_1}{\partial \xi_1}\left(\frac{\partial x_2}{\partial \xi_2}\frac{\partial x_3}{\partial \xi_3} - \frac{\partial x_2}{\partial \xi_3}\frac{\partial x_3}{\partial \xi_2}\right) - \frac{\partial x_1}{\partial \xi_2}\left(\frac{\partial x_2}{\partial \xi_1}\frac{\partial x_3}{\partial \xi_3} - \frac{\partial x_2}{\partial \xi_3}\frac{\partial x_3}{\partial \xi_1}\right) \\
& + \frac{\partial x_1}{\partial \xi_3}\left(\frac{\partial x_2}{\partial \xi_1}\frac{\partial x_3}{\partial \xi_2} - \frac{\partial x_2}{\partial \xi_2}\frac{\partial x_3}{\partial \xi_1}\right)
\end{aligned} \tag{2.15}$$

The element stiffness matrix in terms of the local coordinates (ξ_1, ξ_2, ξ_3) is now given by

$$K_{ij}^{k_e} = \int_{\widehat{\Omega}} \sum_{m=1}^{3} \left[\left(\sum_{n=1}^{3} \frac{\partial h_i}{\partial \xi_n} \frac{\partial \xi_n}{\partial x_m} J \right) \left(\sum_{n'=1}^{3} \frac{\partial h_j}{\partial \xi_{n'}} \frac{\partial \xi_{n'}}{\partial x_m} J \right) \right] \frac{1}{|J|}\, d\Omega \tag{2.16}$$

and the elemental mass matrix is given by

$$M_{ij}^{k_e} = \int_{\widehat{\Omega}} h_i h_j |J|\, d\Omega. \tag{2.17}$$

The problem matrix is now written as

$$\widetilde{K}_{ij}^{k_e} = K_{ij}^{k_e} - k^2 M_{ij}^{k_e} + S_{ij}^{k_e} \tag{2.18}$$

where the coefficients $S_{ij}^{k_e}$ represent the contributions from the surface integrals of each element Ω^{k_e} needed in invoking boundary conditions. The analogous mass matrix needed in two-dimensions is represented as

$$[M_f^{k_e}]_{ij} = \int_{\widehat{\gamma}} [h_f]_i [h_f]_j |J_f| \, d\Gamma \quad 1 \leq i, j \geq N_{\ell_f}, \tag{2.19}$$

where the subscript f implies that the terms are evaluated on the faces of the elements. Moreover, the two-dimensional Jacobian J_f is constant for the straight faced elements and is equal to $J_f = 2Area_{\widehat{\Gamma}}$, and can be evaluated from the three-dimensional Jacobian by

$$J_f(\xi_1, \xi_2) = J(\xi_1, \xi_2, 0).$$

These mappings depends on the basis functions $h(\xi_1, \xi_2, \xi_3)$ and must be defined appropriately for the tetrahedral elements. The work here uses basis functions of varying orders based on higher-order Jacobi polynomials, in which we will now briefly review.

2.2.1 Nodal Points

The subdivided tetrahedral domains contains nodes or nodal points which plays a key role in the finite element method. These nodes are allocated at least at the vertices of the elements as seen in Figure(2.1), but in order to improve the approximation, further nodes are introduced. For example in the finite element method to go from a linear approximation (nodes only on vertices) to a quadratic approximation, additional nodes at the midpoints of the sides of the element are added. Basis functions of higher than quadratic are also available, but typically when using higher-order spectral elements the nodal sets consist of non-equispaced points. For nodal sets of varying polynomial orders we refer the reader to [41].

The focus of the numerical procedure used, is on the *hp*-version of the finite element method (referred to a spectral element method), where we simultaneously increase the number of elements and increase the interpolation order within the element in order to improve the approximation. This is referred to as an h refinement and a p enrichment. The parameter $h \in (0, 1)$, and its magnitude gives some indication of how close the approximation space V^h is to the infinite dimensional space V; and the number N represents the number of basis functions or nodes in the approximation. In the limits $N \to \infty$, $(h \to 0)$, the basis functions are chosen in such a way such that V^h approaches V. Increasing the polynomial order p (corresponds to more nodes on each element) can provide fast convergence, small diffusion and dispersion errors, and consequently, as will be discussed further, can be an advantageous numerical approach for the Helmholtz equation.

2.2.2 Basis functions

The expansion basis associated with the nodal points for the finite element method are the Lagrange polynomials h_i associated with the nodal sets $\widehat{\xi}_i$. For the nodal spectral element methods, higher-order Lagrange polynomials over tetrahedrons are described in [33], and are extended in the work [41] for the local coordinates (ξ_1, ξ_2, ξ_3) covering $[0, 1]^3$. Here we briefly review the higher-order basis functions described in [41], as we incorporate these into our Code and adapt it for the exact bounds method. Considering the complete three-dimensional polynomial basis of order at most n, where $\mathsf{p}_i(\boldsymbol{\xi}) \in \mathbb{P}^n$, any function can be interpolated as

$$\mathsf{f}(\widehat{\xi}_i) = \sum_{j=1}^{N_\ell} \hat{\mathsf{f}}_j \mathsf{p}_j(\widehat{\xi}_i), \quad \forall i \tag{2.20}$$

where $\hat{\mathsf{f}} = [\hat{f}_1, \cdots \hat{f}_{N_\ell}]$ is the vector of expansion coefficients and $\mathsf{f} = [f(\widehat{\xi}_1), \cdots f(\widehat{\xi}_{N_\ell})]$ is the vector of nodal values at the grid points. Equation(2.20) can be re-expressed as the matrix equation $\mathbf{V}\hat{\mathsf{f}} = \mathsf{f}$ where $\mathbf{V}_{ij} = \mathsf{p}_j(\widehat{\xi}_i)$ is the Vandermonde matrix. Through the Lagrange

basis functions $h_i(\xi_1, \xi_2, \xi_3)$ satisfying (2.13), one can also write the relationship

$$\mathsf{f}(\widehat{\xi}_i) = \sum_{j=1}^{N_\ell} \mathsf{f}(\widehat{\xi}_i) h_j(\widehat{\xi}_i). \tag{2.21}$$

Upon combining Equations (2.20) and (2.21), by invoking $\hat{\mathsf{f}} = \mathbf{V}^{-1}\mathsf{f}$ in (2.20) we obtain

$$h_i(\xi_1, \xi_2, \xi_3) = \sum_{j=1}^{N_\ell} \mathbf{V}_{ji}^{-1} \mathsf{p}_j(\xi_1, \xi_2, \xi_3) \tag{2.22}$$

Therefore, the Lagrange polynomials h_i are related to another set of polynomials (which can be chosen to be any set) through the Vandermonde matrix. The inverse of the Vandermonde matrix must exist in order for the interpolation to exist. Moreover, the polynomial basis p_i must be orthonormal with respect to some inner product in order to maximize the degree of linear independence of the basis and consequently avoid severe problems in computing $\mathbf{V^{-1}}$ where the condition number of the matrix $\mathbf{V}$ can grow exponentially with increasing order n. Henceforth, the basis that has been applied in [41] is given in terms of the Jacobi polynomials $P^{(\alpha,\beta)}$ to be:

$$\begin{aligned} \psi_{ijk}(\xi_1, \xi_2, \xi_3) &= \left(\frac{(1-s)(1-t)}{4}\right)^i P_i^{(0,0)}(r) \left(\frac{1-t}{2}\right)^j P_j^{(2i+1,0)}(s) \\ &\times P_k^{(2i+2j+2,0)}(t), \quad i, j, k \geq 0, i+j+k \leq n, \end{aligned} \tag{2.23}$$

where the mappings (r, s, t) are the collapsed coordinate system described in [37] and are given by

$$r = \frac{2\xi_1}{1-\xi_2-\xi_3} - 1, \quad s = \frac{2\xi_2}{1-\xi_3} - 1, \quad t = 2\xi_3 - 1. \tag{2.24}$$

The $\psi_{i,j,k}$'s satisfy the following orthogonality condition

$$\int_{\widehat{\Omega}} \psi_{i,j,k} \psi_{p\,q\,r}\, d\Omega = \gamma_{i,j,k} \delta_{i,j,k,p,q,r}, \tag{2.25}$$

where $\delta_{i,j,k,p,q,r}$ is the three-dimensional Dirac delta function, and $\gamma_{i,j,k}$ is a normalization coefficient given by

$$\gamma_{i,j,k} = \frac{1}{2i+1} \frac{1}{2(i+j)+2} \frac{1}{2(i+j+k)+3}.$$

The polynomial set $\psi_{\imath,\jmath,k}$ is now expressed in terms of modified Jacobi polynomials

$$\psi_{\imath,\jmath,k}(\xi_1,\xi_2,\xi_3) = [P_r]_{\imath}^{(0,0)}(\xi_1,\xi_2,\xi_3) \times [P_s]_{\jmath}^{(2\imath+1,0)}(\xi_2,\xi_3) \times [P_t]_{k}^{(2\imath+2\jmath+2,0)}(\xi_3), \tag{2.26}$$

where these modified polynomials are defined to be

$$\begin{aligned} [P_r]_n^{(\alpha,\beta)}(\xi_1,\xi_2,\xi_3) &= (1-\xi_2-\xi_3)^n P_n^{(\alpha,\beta)}\left(\frac{2\xi_1}{1-\xi_2-\xi_3}-1\right) \\ [P_s]_n^{(\alpha,\beta)}(\xi_2,\xi_3) &= (1-\xi_3)^n P_n^{(\alpha,\beta)}\left(\frac{2\xi_2}{1-\xi_3}-1\right), \\ [P_t]_n^{(\alpha,\beta)}(\xi_3) &= P_n^{(\alpha,\beta)}(2\xi_3-1), \end{aligned} \tag{2.27}$$

which can be expressed alternatively, as

$$[P_r]_n^{(\alpha,\beta)} = \frac{1}{2^n}\sum_{m=0}^{n}\binom{n+\alpha}{m}\binom{n+\beta}{n-m}(2\xi_1+2\xi_2+2\xi_3-2)^{n-m}(2\xi_1)^m, \tag{2.28}$$

$$[P_s]_n^{(\alpha,\beta)} = \frac{1}{2^n}\sum_{m=0}^{n}\binom{n+\alpha}{m}\binom{n+\beta}{n-m}(2\xi_2+2\xi_3-2)^{n-m}(2\xi_2)^m, \tag{2.29}$$

$$[P_t]_n^{(\alpha,\beta)} = \frac{1}{2^n}\sum_{m=0}^{n}\binom{n+\alpha}{m}\binom{n+\beta}{n-m}(2\xi_3-2)^{n-m}(2\xi_3)^m, \tag{2.30}$$

where

$$\binom{m}{n} = \frac{m!}{n!(m-n)!}.$$

Now the polynomials $\frac{\psi_\jmath(\xi_1,\xi_2,\xi_3)}{\sqrt{\gamma_\jmath}}$ are orthonormal (i.e. satisfying Kronecker delta property) and are modal (Hierarchical set of polynomials). For simplicity of notation, we designate the combination of $i,\jmath,k$ in $\psi_{\imath,\jmath,k}$ with δ, where $1 \leq \delta \leq N_\ell$ and write ψ_δ. Moreover, the normalized polynomials $\frac{\psi_\delta}{\sqrt{\gamma_\delta}}$ are a suitable choice for the p_δ's presented above. Therefore from Equation (2.22) we can express the Lagrange polynomials which are a nodal (non-hierarchical basis) set of basis in terms of the normalized modal basis functions as

$$h_\imath(\xi_1,\xi_2,\xi_3) = \sum_{\jmath=1}^{N_\ell}\mathbf{V}^{-1}_{\jmath\imath}\frac{\psi_\jmath(\xi_1,\xi_2,\xi_3)}{\sqrt{\gamma_\jmath}}, \quad \forall i \tag{2.31}$$

where the components of the Vandermonde matrix are given by

$$\mathbf{V}_{ij} = \frac{\psi_j(\widehat{\xi}_i)}{\sqrt{\gamma_j}}. \tag{2.32}$$

In order to calculate the surface integrals for the boundary conditions, we require the two-dimensional version of these basis functions. These can be retrieved from the three-dimensional modal basis functions

$$[\psi_f]_{ij}(\xi_1, \xi_2) = \psi_{ij0}(\xi_1, \xi_2, 0), \quad \text{for} \quad i, j \geq 0, i + j \leq n \tag{2.33}$$

The corresponding orthogonality condition for the face reference element is now

$$\int_{\widehat{\Gamma}} [\psi_f]_{ij}[\psi_f]_{pq}\, d\Gamma = [\gamma_f]_{ij}[\delta_f]_{ijpq}, \tag{2.34}$$

where $[\delta_f]_{ijpq}$ is the two-dimensional Kronecker delta, and $[\gamma_f]_{ij}$ is the normalization factor

$$[\gamma_f]_{ij} = \frac{1}{2i+1}\,\frac{1}{2(i+j)+2}. \tag{2.35}$$

We again combine the indices i and j and designate the combination by δ', where $1 \leq \delta' \leq N_{\ell_f}$ and write $[\psi_f]_{ij}$ as $[\psi_f]_{\delta'}$. Analogous definitions for the two-dimensional higher-order Lagrange polynomials in terms of these normalized modal basis functions is written as

$$[h_f]_i(\xi_1, \xi_2) = \sum_{j=1}^{N_{\ell_f}} [\mathbf{V}_f]^{-1}_{ji} \frac{[\psi_f]_j(\xi_1, \xi_2)}{\sqrt{[\gamma_f]_j}}. \tag{2.36}$$

These higher-order basis functions h_i are then used in the calculation of the elemental matrices described earlier which include the stiffness and mass matrices, (2.16), (2.17), and (2.19).

2.3 Some Key Concepts

Variational forms which arise from the Helmholtz equation are in general, not positive definite and since we are concerned with solving the Helmholtz problem via FEM or SEM, we

must address the issue of whether or not the boundary value problem is weakly solvable and if the solutions are unique, as these generalizations to indefinite forms do not immediately follow. Moreover, for a boundary value problem to be well-posed for a given set of data, a solution must not only exist, and be unique but must depend continuously on the initial data. Therefore, stability of the numerical solution also becomes of paramount importance, especially when addressing convergence issues, as small errors in the data may cause a large error in the solution. For the Helmholtz problem, existence and uniqueness results are generalized from the Lax-Milgram theorem which holds only for positive definite variational forms. There are two such generalizations, one where the variational form satisfies an *inf-sup* condition, and another which satisfies a Gårding inequality. As these concepts are important for the understanding of the numerical difficulties associated with the Helmholtz equation, and in the computation of the bounds, a brief discussion of them is warranted.

In order to explain some of the ideas related to solvability conditions for the Helmholtz problem, let us first consider the variational (weak) formulation of a boundary value problem of the form

$$\begin{cases} \text{Find } u \in V_1 : \\ b(u,v) = (f,v) \ \ \forall v \in V_2, \end{cases} \tag{2.37}$$

where V_1 and V_2 are normed linear spaces for the trial and test spaces respectively. The form b can be either biliear or sesquilinear, and f is a bounded linear (or antilinear) functional defined on V_2.

2.3.1 Positive definite forms

The Lax-Milgram theorem states that for sesquilinear forms $a : V \times V \mapsto \mathbb{C}$ defined on a Hilbert space V satisfying

1. Continuity(boundedness):

$$\exists M > 0 : \quad |a(u,v)| \leq M\|u\|_V \|v\|_V, \quad \forall u, v \in V, \tag{2.38}$$

2. V-Ellipticity(positive definiteness):

$$\exists \alpha > 0 : \quad |a(u,v)| \geq \alpha \|u\|_V^2, \quad \forall u \in V, \tag{2.39}$$

and for a bounded linear functional f defined on V, we can find a unique element $u = u_* \in V$ such that $a(u_*, v) = (f, v)$. From the ellipticity condition, it can be shown that the solution u_* is bounded by the data f, requiring that

$$\|u_*\|_V \leq \frac{1}{\alpha} \|f\|_{V'},$$

thus showing stability and regularity. As an example, for the Poisson equation with homogeneous Dirichlet boundary conditions, the sesquilinear form $a(u,u)$ is given by Equation (2.8) defined on the Hilbert space $V = H_\circ^1(\Omega) \subset H^1(\Omega)$ containing all H^1 functions that vanish on Γ can be shown to be V-elliptic, i.e.

$$a(u,u) \geq \frac{1}{1+C^2} \|\nabla u\|_1^2$$

where C is a positive constant and $\|.\|_1$ denotes the H^1-norm. This follows from an application of the *Poincare* inequality:

$$\|u\|_{L^2} \leq C \|\nabla u\|_{L^2}.$$

2.3.2 inf − sup condition

For the Helmholtz problem, the sesquilinear form is

$$b(u,v) = \int_\Omega (\nabla u \cdot \nabla \overline{v} - k^2 u \overline{v})\, d\Omega \tag{2.40}$$

which becomes indefinite for large k, thus the Lax-Milgram theorem cannot be applied as the condition of V-ellipticity no longer holds. Under slightly weaker conditions (where V-ellipticity is not assumed) however, a generalization of the theorem to indefinite forms was shown by Babuška and guarantees the existence of a unique solution $u_* \in V_1$ such that $b(u_*, v) = (f, v)$ for a sesquilinear form $b : V_1 \times V_2 \mapsto \mathbb{C}$ on *Hilbert spaces* V_1, V_2 satisfying

1. Continuity

$$\exists M > 0: \quad |b(u,v)| \leq M\|u\|_{V_1}\|v\|_{V_2}, \quad \forall u \in V_1, v \in V_2, \tag{2.41}$$

2. $\inf - \sup$ Condition:

$$\exists \beta > 0: \quad \beta \leq \sup_{0 \neq v \in v_2} \frac{|b(u,v)|}{\|u\|_{V_1}\|v\|_{V_2}} \quad \forall 0 \neq u \in V_1, \tag{2.42}$$

3. Transposed $\inf - \sup$ Condition:

$$\sup_{0 \neq u \in V_1} |b(u,v)| > 0, \quad \forall 0 \neq v \in V_2, \tag{2.43}$$

and an antilinear bounded functional $f : V_2 \mapsto \mathbb{C}$ defined on V_2. The solution u_* then can be shown to satisfy the stability estimate

$$\|u_*\|_{V_1} \leq \frac{1}{\beta}\|f\|_{V_2'}.$$

As already mentioned, in the FEM (SEM) we are concerned which approximations which belong to finite dimensional subspaces $V^N \subset V$. Therefore, the above conditions for existence of a unique solution must also extend to the approximation spaces. For definite forms (Lax-Milgram theorem) this has an immediate extension, however for indefinite forms, the theorem of Babuška does not extend to subspaces V^N because by restricting to a subspace of V, the supremum can decrease, where as the infimum may not and so the $\inf - \sup$ condition will not be satisfied. This motivates the definition of the discrete $\inf - \sup$ condition which guarantees a unique solution to problem (2.37) with $u \in {V_1}^N$, and $v \in {V_2}^N$ and gives stability estimates for these approximation. The question of when the numerical solution for the Helmholtz equation satisfies the discrete $\inf - \sup$ condition, will rely on the notion of V-coercive forms.

2.3.3 V-Coercive forms

The Helmholtz equation is an elliptic partial differential equation, however we know that the sesquilinear (bilinear) form corresponding to Helmholtz problem is not *V-elliptic* Therefore, elliptic boundary value problems do not generally correspond to V-elliptic variational forms, rather V-coercive forms [1] will be associated with elliptic boundary value problems Henceforth, a sesquilinear form $b \quad V \times V \mapsto \mathbb{C}$ defined on a Hilbert space $V = H^1(\Omega)$ (Ω is a bounded domain) is called V-coercive if for $u \in V$ it satisfies a Gårding inequality

$$\left| b(u,u) + C\|u\|^2_{L^2(\Omega)} \right| \geq \alpha \|u\|^2_{H^1(\Omega)} \tag{2 44}$$

with positive constants C, α We point out that this definition holds for a Gelfand triple (see [32], Section 6 5 13) which for the special case $V = H^1(\Omega)$ is $H^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega)$ This definition can be interpreted as the V-ellipticity property of

$$a(u,v) \ = b(u,v) + C(u,v)_{L^2(\Omega)}$$

It is seen that by setting $C = k^2$, the Helmholtz variational form will satisfy the Gårding inequality

Remark 2.3 *Although we are concerned here with the well-posedness of the Helmholtz problem, and that the theorem of Babuška gives conditions for existence and uniqueness of indefinite variational forms, there are those frequencies where the interior Helmholtz problem does not have a unique solution Let us consider the homogeneous Helmholtz equation in a unit cube* Ω,

$$\begin{aligned} \nabla^2\psi + k^2\psi &= 0 \ \textit{in} \ \Omega, \\ \psi &= 0 \ \textit{on} \ \Gamma, \end{aligned}$$

[1] Often many textbooks on the subject, use the notion of coercive bilinear forms i e $a(u,u) \geq \alpha\|u\|^2$ for constant $\alpha > 0$ and speak of the Helmholtz problem being noncoercive However, we follow the definition used in Hackbusch [32], and speak of the Helmholtz variational form as being V-coercive and satisfying a Gårding inequality

where the problem has nontrivial solutions (the eigenmodes) for $\lambda_{n,m,l} = k^2 = \pi^2(n^2+m^2+l^2)$, $l, m, n \in \mathbb{N}$ *given by*

$$\psi_{n,m,l}(x, y, z) = \sin(n\pi x)\sin(m\pi y)\sin(l\pi z).$$

As an example we see that $\lambda_{1,7,2} = \lambda_{2,5,5}$, *but this value corresponds to different eigenfunctions. At these frequencies, the inhomogeneous problem is also no longer uniquely solvable. However, if the vibrations are damped, which occurs by introducing a nonzero imaginary boundary term, then the interior problem will generally have a unique solution.*

We will see in the next chapter that in the discrete model, having no damping will result in the system matrix becoming singular, whenever k^2 *becomes numerically close to an eigenvalue of the problem. Therefore a damping term is introduced in order to make the system matrix non-singular.*

The above remark hints at the fact that having a nonvanishing imaginary boundary condition is essential for the proof of uniqueness. However, once a uniqueness is established, the existence of a solution to the boundary value problem (2.37) is shown in (see [32], Theorem 6.5.15) to follow from the Fredholm alternative for a sufficiently regular domain Ω, in which the embedding $H^1(\Omega) \subset L^2(\Omega)$ is compact. The Fredholm alternative states: that either the problem (2.37) has a solution $u \in H^1(\Omega)$ for all f, or there exists a non-trivial solution to the homogeneous problem (with $f \equiv 0$). Thus, for the above remark, if for the homogeneous case one obtains non-unique solutions, then it can be expected that the non-homogenous problem, if a solution exists, will also yield non-unique solutions. Moreover, the number of solutions will be sum of the solutions to the homogeneous and non-homogeneous problems.

2.3.4 Variational methods

The aim of this subsection is to demonstrate some properties of the convergence behavior related to the numerical approximation of Helmholtz boundary value problem as compared

to boundary value problems having V-elliptic variational forms, where in particular, we will use the Poisson equation as an example. The variational weak form of the Poisson problem is given by find $u \in V = H^1(\Omega)$ such that

$$a(u, v) = (f, v) \quad \forall v \in V \tag{2.45}$$

with $a : V \times V \mapsto \mathbb{R}$ given by

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, d\Omega. \tag{2.46}$$

In the FEM(SEM), the solution to the variational weak form, is also the minimizer of an energy functional $J(u)$ given by

$$J(u) = \frac{1}{2} a(u, u) - (f, u).$$

Thus, the same problem of solving Equation (2.45) can be expressed as a minimization problem, which is the basis for the Ritz method. By considering approximation spaces $V^N \subset V$ the Ritz method seeks an approximation $u^N \in V^N \subset V$ that minimizes an energy functional in V^N. The consequence of such a minimization leads to the Galerkin weak form

$$a(u^N, v) = (f, v) \quad \forall v \in V^N. \tag{2.47}$$

Subtracting Equation (2.47) from Equation (2.45) gives the finite element error equation

$$a(u - u^N, v) = 0 \quad \forall v \in V^N, \tag{2.48}$$

which indicates the error $u - u^N$ is orthogonal to V^N with respect to the energy norm $\|.\|_a$ given by

$$\|u\|_a = (|a(u, u)|)^{1/2}. \tag{2.49}$$

Convergence properties

Using Equation (2.48), and applying the Cauchy-Schwartz inequality, it is apparent that

$$\begin{aligned} \|u-u^N\|_a^2 &= a(u-u^N, u-u^N) = a(u-u^N, u-v-u^N+v) \\ &= a(u-u^N, u-v) - a(e, u^N-v) \\ &= a(u-u^N, u-v) \leq \|u-u^N\|_a \|u-v\|_a \end{aligned} \tag{2.50}$$

for $e = u - u^N$ and all $v \in V^N$. Therefore, by canceling out the error norm on both sides, we see that the error is asymptotic when we estimate the error of best approximation as

$$\inf_{v \in V^N} \|u-v\|_a \to 0 \quad \text{for} \quad N \to \infty, \quad \forall u \in V.$$

For V-elliptic variational forms $a(.,.)$, the solution $u^N \in V^N$ is the best approximation (in the energy norm) of the exact solution $u \in V$. V-ellipticity of $a(.,.)$ implies that the energy functional $J(u)$ is convex, and thus has a unique minimizer. An important observation for the Ritz method of minimizing an energy functional is that the solution u^N begins to converge from $N = 1$ and asymptotically approaches the exact solution as the number of degrees of freedom in the discrete model increases. Moreover, by letting $v = u$ in the continuity and ellipticity equation (2.38)-(2.39) respectively, one obtains

$$\alpha \|u\|_V^2 \leq |a(u,u)| \leq M \|u\|_V^2 \quad \forall u \in V,$$

which suggests that the energy norm is equivalent to the V-norm and thus from (2.50), one can prove Cea's Lemma which states that for a continuous,V-elliptic, bilinear form $a(.,.)$ and a bounded linear functional f on V, there exists a constant C independent of N (or h), such that

$$\|u-u^N\|_V \leq C \inf_{v \in V^h} \|u-v\|_V,$$

where C is a stability constant and for positive definite forms it is $\frac{M}{\alpha}$ where M and α come from Equations (2.38) and (2.39) respectively. Cea's Lemma can be viewed as an estimation of how far off u is from the subspace V^N, which can indicate the quality of the approximation u^N. Since for some particular $\tilde{v}^N \in V^N$ we have that

$$\inf_{v^N \in V^N} \|u - v^N\| \leq \|u - \tilde{v}^N\|,$$

one can get a better understanding of the convergence of approximation, if $\tilde{v}^N$ is chosen to be some interpolate of u. This is a function $\tilde{u}^N \in V^N$ which has values corresponding to that of u at the N points $\mathbf{x_1}, \mathbf{x_2}, \cdots \mathbf{x_N}$ in Ω [2] such that $u(\mathbf{x}_\ell) = \tilde{u}^N(\mathbf{x}_\ell)$ for $\ell = 1 \cdots N$. This indicates that the problem of convergence can be understood as whether $\tilde{u}^N \mapsto u$ as $N \mapsto \infty$.

In the case of indefinite variational forms, the convergence behavior is quite different. Generally an indefinite variational form, satisfying Equation (2.37) is solvable if the continuous form $b(.,.)$ satisfies the discrete $\inf - \sup$ condition

$$\exists \beta_N > 0 : \quad \beta_N \leq \sup_{0 \neq v \in V_2^N} \frac{|b(u,v)|}{\|u\|_{V_1} \|v\|_{V_2}} \quad \forall 0 \neq u \in V_1^N, \tag{2.51}$$

and the discrete transposed condition

$$\sup_{u \in V_1^N} |b(u,v)| > 0, \quad \forall 0 \neq u \in V_2^N. \tag{2.52}$$

However, a sufficient condition for the existence of a unique solution requires that the sesquilinear form $b(.,\ )$ be V-coercive on $V = H^1(\Omega)$, satisfying the Gårding inequality (2.44). In this case there exists a number N_o such that the variational form in (2.37) has a unique solution $u^N \in V^N$ for all $N \geq N_o$ and $\|u^N - u\|_{H^1(\Omega)} \to 0$ as $N \to \infty$. The proof of this statement is given in [32], and it indicates that the convergence for indefinite but V-coercive forms behaves erratically when N is small, but when N passes a critical number, we

[2] The interpolate can also be viewed locally on $\widehat{\Omega}$, but for preciseness we need to define those approximation spaces corresponding to the individual elements appropriately.

begin to observe a regular convergence pattern. Consequently, for the Helmholtz problem, increasing the wave number will cause the variational form to become indefinite, however, as it satisfies a Gårding inequality, for subspaces approximating well enough, the problem can be made to be V-elliptic. Thus the energy functional for the Helmholtz variational weak form will become convex, and the method of convex minimization for obtaining bounds will apply. Moreover, for $N > N_\circ$ when the discrete approximation has a unique solution u^N the error $u - u^N$ will satisfy

$$\|u - u^N\|_V \leq C_1 \inf_{v \in V^N} \|u - v\|_V \quad \forall N > N_\circ$$

for some constant C_1 not depending on N. This is the same in form as Cea's Lemma, but it must be understood that the stability constant is different in this case, and typically larger for the indefinite forms.

2.3.5 An observation

We let $e := u - u_h$, where u_h is the finite element approximation to u and e is the error. Letting $v = u - u_h \in V_h$ we recast the Helmholtz equation as the error equation

$$b(e, e) =< f, e > \quad \forall e \in V^h \tag{2.53}$$

where we recall the sesquilinear operator $b(e, e)$ to be

$$b(e, e) = \int_\Omega \nabla e \cdot \nabla \overline{e}\, d\Omega - k^2 \int_\Omega e \cdot \overline{e}\, d\Omega. \tag{2.54}$$

If we assume a particular polynomial approximation, that u_h is approximated by piecewise linear polynomials, then $e = \mathrm{O}(h^2)$. Since the second term is quadratic in e, then the integral of the second term will be $\mathrm{O}(h^4)$. The first term is quadratic in the derivatives of e and thus the integral in the first term will reduce as $\mathrm{O}(h^2)$. Therefore, integral in the second term goes to zero much faster than the first term, and although the second term is

$O(h^4k^2)$, one can expect that at some h, the first term will dominate and the sesquilinear form will retain a positive definite structure.

2.4 Pollution Effect

2.4.1 Piecewise linear case

The above observations hints at a restriction which must be applied in the design of a mesh for a given k. As a simple example, consider the three-dimensional Helmholtz equation as in problem (2.1)-(2.2) with homogeneous Dirichlet boundaries ($g_D = 0$) and with a forcing function: $f = 3(\pi^2 - 1)k^2 \sin(k\pi x)\sin(k\pi y)\sin(k\pi z)$. Then the exact solution is given by $u = \sin(k\pi x)\sin(k\pi y)\sin(k\pi z)$. Thus the solution is periodic and for the three-dimensional wavelength λ, with components $\lambda = \lambda_x = \lambda_y = \lambda_z = \frac{2\pi}{k\pi}$ will satisfy

$$n_{res} = \frac{\lambda}{h} \approx \text{constant},$$

where the number n_{res} is the resolution of the mesh in each direction. Extensive study of the one dimensional wave equation [34, 35] have referred to this as the "rule of thumb", where in the particular case when approximating an oscillatory function with piecewise linear interpolants, have recommended the choice $n_{res} = 10$ in practice, which gives reliable results. As was mentioned in the previous section, the problem of convergence of a Galerkin approximation can be transformed into one of interpolants. Letting $\mathbb{I}_h(u)$ to be the projection operator which maps u to its interpolate $\tilde{u}_h$, we can write some typical estimates on the interpolation error $u - \mathbb{I}_h(u)$. These estimates hold for finite elements that are not allowed to be arbitrarily thin, where the finite element is said to be *regular*. We refer to [18] for details. For piecewise polynomials on *regular* finite elements of degree $p \geq 1$ of dimension

$d \leq 3$ which form a partitioning of the mesh, these interpolation estimates satisfy

$$\|u - \mathbb{I}_h(u)\|_{L^2(\Omega)} \leq Ch^{p+1}|u|_{H^{p+1}(\Omega)} \tag{2.55}$$

$$\|u - \mathbb{I}_h(u)\|_{H^1(\Omega)} \leq Ch^{p}|u|_{H^{p+1}(\Omega)}. \tag{2.56}$$

For piecewise linear polynomials $p = 1$, it can be shown for the solution u given above that

$$\frac{|u|_{H^2}}{\|u\|_{L^2}} \leq C_1 k^2, \quad \text{and} \quad \frac{|u|_{H^2}}{|u|_{H^1}} \leq C_2 k$$

for positive constants C_1, C_2. Consequently, bounds on the relative errors for interpolation of an oscillatory solution u yield

$$\frac{\|u - \mathbb{I}_h(u)\|_{L^2}}{\|u\|_{L^2}} \leq C_1 h^2 k^2 \quad \text{and} \quad \frac{\|u - \mathbb{I}_h(u)\|_{H^1}}{|u|_{L^1}} \leq Chk.$$

Under the h-version of the finite element method, we see that since the resolution rule is determined upon by the estimate on hk, it thus controls the interpolation error.

Despite imposing a "rule of thumb" strategy, we know from computations that the finite element error still grows with the wave number. In fact from a proof in [4], which generalizes Cea's Lemma, we can estimate the error $u - u_h$ as:

$$\|e\|_V \leq \|u - v\|_V + \|u_h - v\|_V \leq \|u - v\|_V + \frac{M}{\beta_h}\|u - v\|_V$$

where v is an arbitrary function of V^h, M is the continuity constant from (2.38) and β_h is the constant in the discrete $\inf - \sup$ condition. For smooth functions u, v is chosen to be the interpolant $\mathbb{I}_h u$ of u in V^h. The error thus is characterized by two terms, where the first term is due to a discretization error $u - \mathbb{I}_h u$ and the second, called a pollution error which is related to the stability properties of the approximation of the discrete sesquilinear form, where large values of $\frac{M}{\beta_h}$ indicate a loss of stability. In the case of piecewise linear approximations, estimates which hold under the assumption $hk < 1$ are referred to as preasymptotic estimates. It has been shown in [34] that for the one dimensional Helmholtz

equation, the interpolant $\mathbb{I}_h u$ is the best approximation and that under the assumption $hk < 1$, the finite element error can be shown to be

$$|u - u_h|_{H^1} \leq (1 + Ck^2 h) \inf_{v \in V^h} |u - v|_{H^1}. \tag{2.57}$$

Assuming oscillatory behavior of the solution (i.e. $\frac{|u|_{H^2}}{|u|_{H^1}} = \mathrm{O}(k)$), it is shown that the relative error (i.e. $\tilde{e} = \frac{|u-u_h|_{H^1}}{|u|_{H^1}}$) is estimated as

$$\tilde{e}_1 \leq C_1 hk + C_2 k^3 h^2, \tag{2.58}$$

where C is a constant independent of k and h. The second term is the pollution term and is $\mathrm{O}(k^3 h^2)$.

Although the model equation representing time harmonic acoustic waves is non-dispersive wave equation, it should be noted that the numerical solution to the Helmholtz equation do not preserve the non-dispersive character. Just as in the mathematical model, the wave number is given as $k = \omega/c$, where ω is the frequency of the wave, and c is the speed of sound, the numerical wave will have a wave number $k^h = \omega/c^h$ which differs from k. This dispersive effect is a numerical artifact which is consequence of the discretization scheme. Moreover, it has been shown in the one dimensional setting that in fact the pollution error has the same order as the phase lag, namely

$$k^h = k - \frac{k^3 h^2}{24} + \mathrm{O}(k^5 h^2). \tag{2.59}$$

Therefore, due to the pollution error, the optimal order of convergence is not achieved through the estimate $hk < 1$, and numerical experiments demonstrate in the case of piecewise linear approximations, that the condition $k^2 h \ll 1$ is sufficient for quasi-optimality of the finite element error. In this case, as seen in (2.57), the finite element error on fine meshes is similar to that of the interpolation error. In practical computations, however, in

order to keep the error below some tolerance, one considers a mesh design under the estimate $k^2h \leq$ constant, but a mesh size $h = \mathrm{O}(k^{-2})$ is considered impractical for engineering applications when k is large.

2.4.2 Higher Order Elements - *hp* FEM

It is seen from the interpolation estimates of order p given in Equations (2.56)-(2.56), that addressing the effect of pollution error for higher order polynomials that where the error approximation in the H^1 norm is bounded by $h^p|u|_{H^{p+1}}$ then the derivatives of higher order must exist and so higher regularity of the functions are required. A detailed study of this is given in [34].

After establishing the regularity of the solution, just as in Equation (2.58), it is shown in [34, 36] that the relative error for general polynomial orders of $p \geq 2$, assuming oscillatory behavior where $|u|_{H^{\ell+1}}/|u|_{H^1} = \mathrm{O}(k^\ell)$ is given by

$$\tilde{e}_1 \leq C_1 \left(\frac{hk}{2p}\right)^p + C_2 k \left(\frac{hk}{2p}\right)^{2p} \tag{2.60}$$

where the first term is the approximation error and the second term represents the pollution term. It is apparent that when $p \geq 2$, the pollution effect is reduced significantly when the mesh size is restricted to satisfy $\frac{kh}{2p} < 1$, since this expression is taken to the power of $2p$ in the pollution term.

Although going to higher order significantly improves the approximation by reducing the pollution error, it should not be considered a panacea as this type of error is inherent in the polynomial based methods despite its improvement for high-orders. Moreover, according to Babuška et. al. [8], only in one-dimensions can the pollution error be eliminated completely. Other approaches such as the discontinuous enrichment method of Farhat [24] have been applied to improve approximations for the Helmholtz equation which involve using analytical functions such as plane wave equation which satisfy the homogeneous Helmholtz

equation as part of the elemental basis functions. However, this can be computationally more expensive, especially in the local integrations. Regardless, we are restricted to using polynomial based approximations as the bounds method as will be presented, is based upon approximations from varying polynomial fields. To obtain sharper bounds we apply the method with higher order polynomials in order to reduce the effect of pollution for higher k.

We will see in the next chapter that the method is based on recasting the problem into a constrained minimization problem where the constraints involved are relaxing the continuity along elemental subdomains and then enforcing the continuity through the use of Lagrange multipliers. The subdomains which we deal with are either individual elements that were used in the partitioning of the domain in this case tetrahedrons, or a cube comprised of six tetrahedrons. Therefore we will now review a domain decomposition method that will be used herein in the approximation of the Lagrange multipliers.

Chapter 3

Domain Decomposition Methods

In this chapter we will present two domain decomposition (DD) procedures which are based on a family of FETI (Finite Element Tearing and Interconnecting) methods developed for the parallel finite element solution of equilibrium equations [23, 25, 26, 28]. The method relies on partitioning the domain into totally disconnected subdomains, where the continuity along the interfaces of the subdomains is relaxed (torn) by the introduction of Lagrange multipliers called hybrid fluxes. The decomposed subdomains communicate through the use of Lagrange multipliers which act as forces (traction) along the interfaces. The problem is then reformulated as a hybrid variational principle, which can be thought of as the sum of an interior functional which is subdomain localized, and an interface potential which is the inter-partition connection. The minimization of such an energy functional leads to a set of equations for the hybrid fluxes. These equations include an elliptic problem for each subdomain with Neumann boundary conditions on the interfaces, and an inter-subdomain field continuity enforced via Lagrange multipliers. The family of FETI methods are exploited, in a fast, iterative, computationally efficient solver for the numerical solution of the vector of Lagrange multipliers used to enforce continuity constraints (inter-connect) along the interfaces of the subdomains.

The FETI method is well established in the literature [28, 23] and is based on an

iterative procedure for solving systems of equations where the system matrix is symmetric positive definite (SPD) such as those arising from the discretization of the Poisson or Stokes' problems. This will then be the launching point in presenting the FETI-H method [25, 26] which is an extension of the regularized FETI method suited for solving indefinite problems arising from the Helmholtz equation.

One of the advantages of the family of FETI methods is their numerical scalability with respect to both the mesh size h and the subdomain size. In the particular case of the FETI-H method, for Helmholtz equation, it has also demonstrated scalability with respect to the wave number. This means that for this class of DD-based iterative solver, increasing the mesh size of the problem or the number of subdomains, only causes the convergence rate to deteriorate weakly. However, for both FETI and FETI-H methods, numerical scalability with respect to the number of subdomains requires a coarse space preconditioner which augments the DD iterative solver with a coarse grid problem that is large enough to propagate significant information globally, and thus accelerate convergence, while at the same time keep the computations affordable. In the case of the original FETI method the coarse space is naturally derived from the solvability conditions. However, it is not the case for the FETI-H, where any suitable preconditioner, both local (subdomain localized) and global (coarse grid) are determined based on the spectral analysis of the Laplace operator(see [11]) which give a physical interpretation of the wave nature of the solution as it oscillates on the interface and propagates in the subdomain. Such preconditioners then filter out any low or high frequency waves for better convergence of the algorithm.

As our aim in this work is to demonstrate the effectiveness of the bounds method, and not necessarily on computational efficiency, we have not implemented the coarse space global preconditioner in the interface problem for the Helmholtz equation. The methods we have tried, exploits the complex symmetry of the problem, demonstrates good convergence, and

is scalable with respect to mesh size. However, the coarse space preconditioner is required to make the regularized FETI method scalable with respect to the number of subdomains and wave number, although without this preconditioner the number of iterations for convergence is shown [23] to increase only sub-linearly with respect to the wave number. We will briefly discuss some of the iterative methods we have applied and furthermore, for the sake of completeness we give a brief overview of the coarse space preconditioner in the FETI-H method.

3.1 The FETI Procedure - Poisson Example

Non-overlapping domain decomposition methods have been used extensively for solving elliptic problems; and often are the method of choice. Here, one splits the global domain Ω into a finite set of subdomains $\Omega^{(s)}$ satisfying

$$\overline{\Omega} = \bigcup_{s=1}^{N_s} \overline{\Omega}^{(s)}, \quad \Omega^{(s)} \cap \Omega^{(q)} = \emptyset \quad \forall s \neq q$$

where we recall that the overline denotes the closure; for example here $\overline{\Omega}^{(s)} = \partial\Omega^{(s)} \cup \Omega^{(s)}$ for subdomain s.

Domain decomposition methods usually involve solving for the restrictions to each subdomain of the global problem where appropriate boundary conditions on the subdomain interfaces are given. The original FETI method [28, 29], which is often referred to as the dual Schur complement method, invokes the Neumann boundary conditions

$$\frac{\partial u^s}{\partial \mathbf{n}^s} = \lambda$$

at each interface $\overline{\Gamma}^{(s,q)} = \partial\Omega^s \cap \partial\Omega^q$, where $\lambda's$ are the Lagrange multipliers which serve as the inter-connectivity between the subdomains and $\mathbf{n}^s$ is the normalized outward normal vector to $\overline{\Gamma}^{(s,q)}$. Henceforth, for $f \in L^2(\Omega)$, each subdomain problem is given as: Find

$u^s \in H^1(\Omega^s)$ satisfying

$$-\nabla^2 u^{(s)} = f^{(s)} \quad \text{in domain} \quad \Omega^{(s)} \tag{3.1}$$

$$\frac{\partial u^s}{\partial \mathbf{n}^s} = \lambda \quad \text{on} \quad \overline{\Gamma}^{(s,q)} \tag{3.2}$$

$$\text{boundary conditions on } \Gamma \cap \partial\Omega^{(s)} \tag{3.3}$$

with the continuity restrictions

$$u^{(s)} = u^{(q)} \quad \text{on} \quad \overline{\Gamma}^{(s,q)} \tag{3.4}$$

$$\frac{\partial u^s}{\partial \mathbf{n}^s} = -\frac{\partial u^q}{\partial \mathbf{n}^q} \quad \text{on} \quad \overline{\Gamma}^{(s,q)} \tag{3.5}$$

where we recall that the superscripts s are the restrictions of the global quantity to the subdomain, i.e. $f^{(s)} = f|_{\Omega^{(s)}}$, and that the continuity conditions ensure that u which is equal to $u^{(s)}$ on each subdomain is the solution of the global problem, and belongs to $H^1(\Omega)$.

As the FEM(SEM) seeks out a weak solution, we proceed by recasting problem (3.1)-(3.3) as a hybrid formulation. We first recall that the variational weak form for a Poisson problem with Dirichlet boundary condition is written for $u \in H^1(\Omega)$ and $f \in H^{-1}(\Omega)$:

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \quad \forall v \in H^1(\Omega).$$

Under a domain decomposition procedure just described this can be equivalently expressed as the argument minimum of the Lagrangian

$$\mathcal{L}(v^{(s)}, \tilde{\lambda}) = \sum_{s=1}^{N_s} \left[\frac{1}{2} \int_{\Omega^{(s)}} \nabla v^{(s)} \cdot \nabla v^{(s)} \, d\Omega - \int_{\Omega^{(s)}} f^{(s)} v^{(s)} \, d\Omega \right] + \sum_{s=1}^{N_s} \int_{\partial\Omega^{(s)}} \sigma_{\Omega^{(s)}} \tilde{\lambda} v^{(s)} \, d\Gamma^{(s)}.$$

where for some arbitrary ordering of subdomains, $\Omega^{(s)} < \Omega^{(q)}$,

$$\sigma_{\Omega^{(s)}}(x) = \begin{cases} -1 & x \in \overline{\Omega}^{(s)} \cap \overline{\Omega}^{(q)}, \ \Omega^{(s)} < \Omega^{(q)} \\ +1 & \text{otherwise.} \end{cases}$$

The optimality conditions are expressed as

$$\frac{\partial}{\partial v^{(s)}}\mathcal{L}(v^{(s)},\tilde{\lambda})|_{\tilde{\lambda}=\lambda}=0, \qquad \frac{\partial}{\partial \tilde{\lambda}}\mathcal{L}(v^{(s)},\tilde{\lambda})|_{v^{(s)}=u^{(s)}}=0 \tag{3.6}$$

where the first of Equation (3.6) leads to the equilibration equation

$$\int_{\Omega^{(s)}} \nabla u^{(s)}\,\nabla v^{(s)}\,d\Omega = \int_{\Omega^{(s)}} f^{(s)}v^{(s)}\,d\Omega - \int_{\partial\Omega^{(s)}} \sigma_{\Omega^{(s)}}\lambda v^{(s)}\,d\overline{\Gamma} \tag{3.7}$$

and the second of Equation (3.6) enforces the continuity of the decoupled solution between subdomains,

$$\sum_{s=1}^{N_s}\int_{\partial\Omega^{(s)}} \sigma_{\Omega^{(s)}}u^{(s)}\,d\Omega = 0. \tag{3.8}$$

3.1.1 Discretization

In this thesis the FETI method is reviewed only as a means to give the necessary background and preliminaries in understanding the domain decomposition approach we undertake for the Helmholtz problem. The FETI method is an iterative method based on a preconditioned conjugate projected gradient method which targets the system

$$\mathbf{Ku} = \mathbf{f} \tag{3.9}$$

where $\mathbf{K}$ represents an $n \times n$ symmetric positive semi-definite sparse matrix, which is typically the stiffness matrix. $\mathbf{u}$ is the n-long vector representing the discrete field solution and $\mathbf{f}$ is the n-long vector representing a generalized forcing term. Here we illustrate the FETI method for the Poisson equation, not only because of its symmetric positive definite (SPD) structure, but also because it closely resembles the Helmholtz equation, and demonstrating the bounds method for the Poisson equation is a stepping stone towards developing the method for the Helmholtz equation as will be seen in Chapter 5. We have seen that after a partitioning of the computational domain Ω into a set of N_s subdomains $\mathcal{T}_h$, the Equation

(3.9) can be replaced by the equivalent system

$$\mathbf{K}^{(s)}\mathbf{u}^{(s)} = \mathbf{f}^{(s)} - \mathbf{B}^{(s)^T}\boldsymbol{\lambda}, \quad s = 1, \cdots, N_s \tag{3.10}$$

$$\sum_{s=1}^{N_s} \mathbf{B}^{(s)}\mathbf{u}^{(s)} = 0, \tag{3.11}$$

where the restrictions of $\mathbf{K}$, $\mathbf{f}$ and the solution $\mathbf{u}$ to each disconnected subdomain, $\Omega^{(s)}$ is denoted by $\mathbf{K}^{(s)}$, $\mathbf{f}^{(s)}$ and $\mathbf{u}^{(s)}$ respectively. Equation (3.10)-(3.11) is just the discrete form of the gradient conditions (3.7)-(3.8). Here $(.)^T$ denotes the transpose of a matrix, and $\mathbf{B}^{sT}\boldsymbol{\lambda}$ represent the array of nodal points on the faces of each elemental subdomain corresponding to the discretization of the subdomain surface integral $\int_{\partial\Omega^{(s)}} \sigma_{\Omega^{(s)}} \lambda v^{(s)}\, d\Omega$. $\mathbf{B}^{sT}\boldsymbol{\lambda}$ should be thought of as an extraction process rather than a matrix vector multiplication, as $\mathbf{B}^{\mathbf{s}}$ is a Boolean matrix corresponding to $\sigma_{\Omega^{(s)}}$ which extracts the signed ($\pm$) restriction of a subdomain solution $u^{(s)}$ to the interface boundary. We note that $\boldsymbol{\lambda}$ here designates the nodal values corresponding to the product of a surface mass matrix and the hybrid fluxes.

The subdomains N_s used in the partitioning are floating subdomains which means that they are not attached to any boundaries and as a result do not have any essential boundary conditions to prevent the matrices $\mathbf{K}^{(s)}$ from becoming singular. Therefore, the subdomain problems in Equation (3.10) are ill-posed and do not yield a unique solution. To ensure that Equation (3.10) is solvable, it is required that the right-hand side data be in the column space of $\mathbf{K}^{(s)}$, which is orthogonal to the left null space i.e. $\ker\left(\mathbf{K}^{(s)^T}\right)$. In the case of a symmetric real matrix $\mathbf{K}^{(s)}$ this amounts to saying,

$$(\mathbf{f}^{(s)} - \mathbf{B}^{(s)^T}\boldsymbol{\lambda}) \perp \ker(\mathbf{K}^{(s)}). \tag{3.12}$$

The general solution to (3.10) is written as

$$\mathbf{u}^{(s)} = \mathbf{K}^{(s)^+}(\mathbf{f}^{(s)} - \mathbf{B}^{(s)^T}\boldsymbol{\lambda}) + \mathbf{R}^{(s)}\boldsymbol{\alpha}^{(s)} \tag{3.13}$$

where $\mathbf{K}^{(s)^+}$ is the generalized inverse of $\mathbf{K}^{(s)}$, and $\mathbf{R}^{(s)} = \ker\left(\mathbf{A}^{(s)}\right)$, with the additional set of unknowns $\boldsymbol{\alpha}^{(s)}$ representing a set of amplitudes that specify the contribution of the null space $\mathbf{R}^{(s)}$ to the solution $\mathbf{u}^{(s)}$. In order to solve for these unknowns, an additional set of constraints is required which must come from the solvability condition (3.12), that is,

$$\mathbf{R}^{(s)^T}\left(\mathbf{f}^{(s)} - \mathbf{B}^{(s)^T}\boldsymbol{\lambda}\right) = 0 \quad \text{for } s = 1, \cdots, N_s. \tag{3.14}$$

Upon substituting the expression for the general solution (3.13) into the Equation (3.11) and using (3.12) the interface problem solving for $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ is the revealed as

$$\begin{bmatrix} \mathbf{F}_I & -\mathbf{G}_I \\ -\mathbf{G}_I^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ -\mathbf{e} \end{bmatrix} \tag{3.15}$$

where

$$\mathbf{F}_I = \sum_{s=1}^{N_s} \mathbf{B}^{(s)}\mathbf{K}^{(s)^+}\mathbf{B}^{(s)^T}, \quad \mathbf{G}_I = \left[\mathbf{B}^{(1)}\mathbf{R}^{(1)} \cdots \mathbf{B}^{(N_s)}\mathbf{R}^{(N_s)}\right] \tag{3.16}$$

$$\mathbf{d} = \sum_{s=1}^{N_s} \mathbf{B}^{(s)}\mathbf{K}^{(s)^+}\mathbf{f}^{(s)}, \quad \boldsymbol{\alpha}^{(s)} = \left[\boldsymbol{\alpha}^{(1)} \cdots \boldsymbol{\alpha}^{(N_s)}\right] \tag{3.17}$$

$$\mathbf{e} = \left[\mathbf{f}^{(1)}\mathbf{R}^{(1)} \cdots \mathbf{f}^{(N_s)}\mathbf{R}^{(N_s)}\right]^T. \tag{3.18}$$

Remark 3.1 *We point out here that the FETI method solves Equation (3.10) for* $\boldsymbol{\lambda}$ *which has non-unique solution. However after calculating the* $\boldsymbol{\alpha}^{(s)}$*'s, the decoupled solutions* $\mathbf{u}^{(s)}$ *are uniquely obtained for each subdomain. In the implementation of our code, we take the vector of all* $\boldsymbol{\alpha}^{(s)}$*'s as* 1 *which specifies* $\mathbf{u}^{(s)}$.

Now we see that the FETI iterative procedure is based on a preconditioned conjugate projected gradient (PCPG) algorithm which augments the basic DD-based iterative method (here the conjugate gradient method is used) with a coarse grid problem that is large enough so that it can disseminate important information globally and accelerate convergence; yet small enough to be computationally affordable.

3.1.2 The FETI PCPG Iterative Procedure

The FETI iterative method consists of premultiplying the first of Equation (3.15) by an operator P, and transform the DD interface problem into

$$\begin{aligned} \mathrm{P}\mathbf{F}_I\boldsymbol{\lambda} &= \mathrm{P}\mathbf{d} \\ \mathbf{G}_I^T\boldsymbol{\lambda} &= \mathbf{e}. \end{aligned} \tag{3.19}$$

Here P is defined as

$$\mathrm{P} = \mathbf{I} - \mathbf{G}_I(\mathbf{G}_I^T\mathbf{G}_I)^{-1}\mathbf{G}_I^T \tag{3.20}$$

which is an orthogonal projector onto $\ker(\mathbf{G}_I^T)$. In the FETI procedure, given an initial value $\boldsymbol{\lambda}^0$ that satisfies

$$\mathbf{G}^T\boldsymbol{\lambda}^0 = \mathbf{e}, \tag{3.21}$$

one can obtain iteratively, a solution to (3.19) by solving the homogenous problem

$$\begin{aligned} \mathrm{P}\mathbf{F}_I\boldsymbol{\lambda} &= \mathrm{P}\mathbf{d} \\ \mathbf{G}_I^T\boldsymbol{\lambda} &= 0. \end{aligned} \tag{3.22}$$

The iterates $\boldsymbol{\lambda}^n$ are then generated by a preconditioned conjugate gradient (PCG) algorithm applied to $\mathrm{P}\mathbf{F}_I\boldsymbol{\lambda} = \mathrm{P}\mathbf{d} \Rightarrow \mathbf{G}_I^T\boldsymbol{\lambda}^n = 0$. Let $\{\mathbf{p}_1, \mathbf{p}_2, \cdots \mathbf{p}_n\}$ denote the set of search directions generated by the first n iterations applied to the interface problem (3.15). It is well known that the CG method is a Kryolov subspace method implied by

$$\boldsymbol{\lambda}^n \in \mathrm{span}\{\mathbf{p}_1, \cdots, \mathbf{p}_n\} = \mathrm{span}\{\mathbf{w}_0, \mathrm{P}\mathbf{F}_I\mathbf{w}_0, \cdots, (\mathrm{P}\mathbf{F}_I)^{n-1}\mathbf{w}_0\}$$

where $\mathbf{w}_0$ is the initial residual associated with $\boldsymbol{\lambda}^0$ and is given by

$$\mathbf{w}_0 = \mathrm{P}\mathbf{d} - \mathrm{P}\mathbf{F}_I\boldsymbol{\lambda}^0 = \mathrm{P}(\mathbf{d} - \mathbf{F}_I\boldsymbol{\lambda}^0).$$

The initial vector $\boldsymbol{\lambda}^0$ is given as $\boldsymbol{\lambda}^0 = (\mathbf{G}_I^T)^{-1}\mathbf{e} = \mathbf{G}_I(\mathbf{G}_I^T\mathbf{G}_I)^{-1}\mathbf{e}$. Therefore, the FETI algorithm can be viewed as a two step preconditioned conjugate gradient method to solve the interface problem and can be found in the literature [22, 23, 28]. The algorithm can be summarized as in [27]:

1. Initialize

$$\begin{aligned} \boldsymbol{\lambda}^0 &= \mathbf{G}_I(\mathbf{G}_I^T\mathbf{G}_I)^{-1}\mathbf{e} \\ \mathbf{w}^0 &= \mathrm{P}^T(\mathbf{d} - \mathbf{F}_I\boldsymbol{\lambda}^0) \end{aligned}$$

2. Iterate $n = 1, 2, \cdots$

$$\begin{aligned} \mathbf{y}^n &= \mathrm{P}\widetilde{F}^{-1}\mathbf{w}^n, \\ \mathbf{p}^n &= \mathbf{y}^n - \sum_{i=0}^{n-1} \frac{\mathbf{y}^n\mathbf{F}_I\mathbf{p}^i}{\mathbf{p}^{iT}\mathbf{F}_I\mathbf{p}^i}\mathbf{p}^i, \\ \boldsymbol{\nu}^n &= \frac{\mathbf{y}^{nT}\mathbf{w}^n}{\mathbf{p}^{nT}\mathbf{F}_I\mathbf{p}^n}, \\ \boldsymbol{\lambda}^{n+1} &= \lambda^n + \boldsymbol{\nu}^n\mathbf{p}^n, \\ \mathbf{w}^{n+1} &= \mathbf{w}^n - \boldsymbol{\nu}^n\mathrm{P}^T\mathbf{F}_I\mathbf{p}^n, \end{aligned}$$

where $\widetilde{F}^{-1}$ denotes a chosen preconditioner. The FETI iteration residual satisfies the following stopping criterion

$$\frac{\|\widetilde{F}^{-1}\mathbf{w}^n\|_{H^2}}{\|\mathbf{f}^n\|_{H^2}} \leq \epsilon_g \tag{3.23}$$

where ϵ_g is the global FETI tolerance.

3.2 Domain Decomposition for Helmholtz

If the dual Schur complement method described above, is applied to the Helmholtz problem, then the local problems may become ill-posed whenever the wave number k of the global problem corresponds to a resonant frequency. As was briefly mentioned in the previous chapter, one can have a unique solution and make the problem well-posed by imposing imaginary boundary conditions, which is due to the rationale that this moves the spectrum of the operator associated with the Helmholtz problem in each subdomain into the complex

plane. However, as it will be shown, there must be special care taken to assure that the problem retains its desired solution.

Although there can be several different ways of including an imaginary boundary term, we follow Farhat [25, 26] and replace the Neumann boundary conditions described in the previous section with Robin boundary conditions where

$$\frac{\partial u^s}{\partial \mathbf{n}^s} + \imath k u^{(s)} = \lambda, \qquad \text{where} \quad \imath = \sqrt{-1}$$

at each interface $\overline{\Gamma}^{(s,q)} = \partial\Omega^s \cap \partial\Omega^q$. Thus the local problems that are to be solved can be stated as: Find $u^{(s)} \in H^1(\Omega^{(s)})$ that satisfies

$$-\nabla^2 u^{(s)} - k^2 u^{(s)} = f^{(s)} \quad \text{in domain} \quad \Omega^{(s)} \tag{3.24}$$

$$\frac{\partial u^s}{\partial \mathbf{n}^s} + \imath k u^{(s)} = \lambda \quad \text{on} \quad \overline{\Gamma}^{(s,q)} \tag{3.25}$$

$$u^{(s)} = g_D^{(s)} \quad \text{on} \quad \Gamma \cap \partial\Omega^{(s)} \tag{3.26}$$

where we recall g_D to be the Dirichlet data from Chapter 2. Moreover, at interfaces the following continuity constraints must be respected

$$u^{(s)} = u^{(q)} \quad \text{on} \quad \overline{\Gamma}^{(s,q)} \tag{3.27}$$

$$\frac{\partial u^{(s)}}{\partial \mathbf{n}^s} + \imath k u^{(s)} = -\frac{\partial u^{(q)}}{\partial \mathbf{n}^q} + \imath k u^{(q)} = \lambda \quad \text{on} \quad \overline{\Gamma}^{(s,q)}. \tag{3.28}$$

The above reformulation (3.24)-(3.26) is valid only for a checkerboard like decomposition of Ω where the subdomains have a special $\pm$ signing assigned to them which ensures well-posed local problems while at the same time guarantee that the accumulation of the contributions of the local decoupled solutions $u^{(s)}$ will yield the global solution u of the original Helmholtz problem. Problem (3.24)-(3.26) then can be written equivalently as the minimization of a modified Lagrangian with continuity constraints between subdomains assembled in a checkerboard fashion satisfying (3.27)-(3.28). For N_s subdomains we write the modified

Lagrangian

$$\begin{aligned}\mathcal{L}(v^{(s)},\tilde{\lambda}) &= \sum_{s=1}^{N_s}\left[\frac{1}{2}\int_{\Omega^{(s)}}\nabla v^{(s)}\cdot\nabla\overline{v}^{(s)}\,d\Omega - k^2\int_{\Omega^{(s)}}v^{(s)}\overline{v}^{(s)}\,d\Omega^{(s)} - \int_{\Omega^{(s)}}f^{(s)}\overline{v}^{(s)}\,d\Omega\right] \\ &+ \sum_{s=1}^{N_s}\int_{\partial\Omega^{(s)}}\sigma_{\Omega^{(s)}}\tilde{\lambda}\overline{v}^{(s)}\,d\Gamma^{(s)} + \sum_{s=1}^{N_s}\sum_{\Omega^{(s)}\cap\Omega^{(q)}\neq\{\emptyset\}}\frac{1}{2}\left[\int_{\overline{\Gamma}^{(s)}}(-1)^{\delta_{\mathbf{n}_{s,q}}}\,ikv^{(s)}\overline{v}^{(s)}\,d\overline{\Gamma}^{(s)}\right. \\ &- \left.\int_{\overline{\Gamma}^{(q)}}(-1)^{\delta_{\mathbf{n}_{q,s}}}\,ikv^{(q)}\overline{v}^{(q)}d\overline{\Gamma}^{(q)}\right]\end{aligned}$$

where $\delta_{\mathbf{n}_{s,q}}$ is equal to 1 if $\mathbf{n}$ is the outgoing normal unit vector to $\Omega^{(s)}$, and is equal to 0 otherwise. The last term in the Lagrangian then is the modified term and is only nonzero at an interface between two subdomains. As in the case for Poisson problem, the gradient conditions of the Lagrangian are then sought, which leads to the equilibration of the hybrid fluxes

$$\begin{aligned}\int_{\Omega^{(s)}}\nabla u^{(s)}\cdot\nabla\overline{v}^{(s)}\,d\Omega - k^2\int_{\Omega^{(s)}}u^{(s)}\overline{v}^{(s)}\,d\Omega + ik\sum_{\Omega^{(s)}\cap\Omega^{(q)}\neq\{\emptyset\}}(-1)^{\delta_{\mathbf{n}_{s,q}}}\int_{\overline{\Gamma}^{s,q}}u^{(s)}\overline{v}^{(s)}\,d\overline{\Gamma}^{(s)} \\ = \int_{\Omega^{(s)}}f^{(s)}\overline{v}^{(s)}\,d\Omega - \int_{\partial\Omega^{(s)}}\sigma_{\Omega^{(s)}}\lambda\overline{v}^{(s)}\,d\Gamma^{(s)},\end{aligned} \tag{3.29}$$

and the continuity enforcement of the decoupled solution between the subdomains, namely,

$$\sum_{s=1}^{N_s}\int_{\partial\Omega^{(s)}}\sigma_{\Omega^{(s)}}u^{(s)}\,d\Omega = 0. \tag{3.30}$$

3.2.1 Discretization

The global Helmholtz problem in discrete form to be solved is

$$\widetilde{\mathbf{K}}\,\mathbf{u} = \mathbf{f} \quad \text{where} \quad \widetilde{\mathbf{K}} = \mathbf{K} - k^2\mathbf{M}$$

where we recall $\mathbf{K}$ and $\mathbf{M}$ are the stiffness and mass matrices respectively, arising from the finite element discretization and k is the wave number which is positive. From Equations (3.29)-(3.30) we know that in the FETI-H method we replace this equation with the

equivalent system of subdomain equations

$$\begin{aligned}\left(\widetilde{\mathbf{K}}^s + ik\mathbf{M}_I^s\right)\mathbf{u}^s &= \left(\mathbf{K}^s - k^2\mathbf{M}^s + ik\mathbf{M}_I^s\right)\mathbf{u}^s \\ &= \mathbf{f}^s - \mathbf{B}^{sT}\boldsymbol{\lambda} \qquad (3.31)\\ \sum_{s=1}^{N_s}\mathbf{B}^s\mathbf{u}^s &= 0 \qquad (3.32)\end{aligned}$$

where the $\mathbf{K}^s$ and $\mathbf{M}^s$ are the local subdomain stiffness and mass matrices, and $\mathbf{M}_I^s$ the matrix corresponding to the discretization of the modified part of the Lagrangian in Equation (3.29) which is nonzero only at the nodal values shared at the interfaces between two subdomains. It regularizes the subdomain matrix, making it non-singular, which in turn leads to (3.31)-(3.32) having a unique solution. Moreover, the solution of nodal values of the subdomain is given by $\mathbf{u}^s$ which includes interior values designated by $\mathbf{u}_{\mathrm{i}}$ and boundary values given by $\mathbf{u}_{\mathrm{b}}$. These are explicitly, written as

$$\mathbf{M}_I^s = \begin{bmatrix} 0 & 0 \\ 0 & \sum_{\Omega^q\cap\Omega^q\neq\{\emptyset\}}\epsilon^{s,q}\mathbf{M}_{\mathrm{bb}}^s \end{bmatrix}, \quad \mathbf{u}^s = \begin{bmatrix} \mathbf{u}_{\mathrm{i}} \\ \mathbf{u}_{\mathrm{b}} \end{bmatrix}$$

where $\mathbf{M}_{\mathrm{bb}}$ is the interface mass matrix given by

$$\{\mathbf{M}_{\mathrm{bb}}^s\}_{ij} = \int_{\Omega^s\cap\Omega^q} h_i\, h_j\, d\xi$$

where we recall that h_i and h_j are the finite element shape functions associated with nodes i and j on the interface between the subdomains. $\epsilon^{s,q}$ is due to the checkerboard structure of the subdomains shown in Figure 3.1, which allows for the well-posed structure of the local subdomain problems and has the special signing: $\epsilon^{s,q} = -\epsilon^{q,s} = \pm 1$. We point out that for computational efficiency, the matrix $\mathbf{M}_I^s$ is constructed as a lumped matrix where all contributions of the matrix is on the diagonal. Our numerical experiments have shown that when we increase the number of subdomains, and more iterations is required for convergence, mass lumping improves our iterative method by reducing the number of iterations. We also

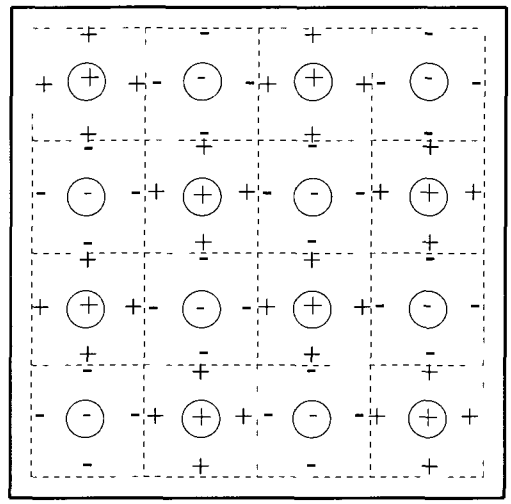


Figure 3.1: Representation of the checkerboard partitioning of the mesh (depicted in 2D), with the special $\pm$ signing used in the regularization of the local subdomain matrices in three dimensions In this partitioning, each subdomain is given arbitrary sign in such a way that at least one of its neighbors has an opposite sign. Furthermore, the faces of each subdomain must also have the same signing as the subdomain.

note that $\mathbf{f^s}$ is the array of nodal values corresponding to the discretization of the duality pairing $\langle f^{(s)}\overline{v}^{(s)}\rangle$, namely,

$$f^s_\alpha = \sum_{\beta=1}^{N_\ell} M^s_{\alpha\beta} f^s(\hat{x}_\beta)$$

where here N_ℓ indicates number of nodes in each subdomain, and $\hat{x}_\beta$ are the the corresponding coordinates.

Finally, substituting Equation (3.31) into (3.32) yields the desired interface problem associated with the regularized subdomain equations in (3.31) given explicitly as

$$\mathbf{F}_I\boldsymbol{\lambda} = \mathbf{d} \tag{3.33}$$

for

$$\mathbf{F}_I = \sum_{s=1}^{N_s} \mathbf{B^s}(\widetilde{\mathbf{K}}^s + ik\mathbf{M}_I^s)^{-1}\mathbf{B}^{sT} \tag{3.34}$$

$$\mathbf{d} = \sum_{s=1}^{N_s} \mathbf{B^s}(\widetilde{\mathbf{K}}^s + ik\mathbf{M}_I^s)^{-1}\mathbf{f}^s. \tag{3.35}$$

We make the observation that $\mathbf{F}_I$ is not Hermitian ($\mathbf{F}_I \neq \overline{\mathbf{F}}_I^T$, where over line indicates complex conjugate of the matrix), it is however, symmetric ($\mathbf{F}_I = \mathbf{F}^T$). This fact is important in choosing the right iterative algorithm to solve the interface problem which exploit the complex symmetry of $\mathbf{F}_I$.

3.2.2 Iterative method used

Several iterative methods have been applied to system (3.33), but as $\mathbf{F}_I$ is non-Hermitian, the classical conjugate gradient method cannot be directly applied. The iterative methods in which we have applied are best suited for solving Helmholtz problems and in particular for complex symmetric linear systems. The iterative procedure we follow is that used in [26], which is a modification to the conjugate residual (CR), and the generalized conjugate residual (GCR) methods which have been derived from the GMRES method for the special case that the system matrix is Hermitian. We briefly summarize the CR and GCR methods as a means to adding clarity in the description of the algorithm used in [26], however, for a detailed description, on the CR and GCR methods the reader should consult [56]. We have also tried the symmetric quasi-minimal residual method (QMR), and the CSYM method which are also good methods in solving large sparse systems with complex symmetric coefficient matrices. With the exception of the CSYM method, all of the other iterative schemes are Krylov subspace methods. For the QMR method we refer the reader to [31, 56], and for the CSYM method the main source is [12]. We will not go into the description of of the other methods as they are discussed fully in the references. However we will comment on

the reasoning behind the choice of iterative method used and will briefly mention from our experience which methods perform best in solving the interface problem (3.33).

The CR method very closely resembles the CG method which becomes eminent if we make the observation, that the nth residual vector $\mathbf{r}^n := \mathbf{b} - \mathbf{A}\mathbf{x}^n$ (in the iterative solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$), where $\mathbf{x}^n$ is the n approximate solution, and search directions $\mathbf{p}^n$ are related in both methods through the relations

$$\mathbf{r}^0 = \mathbf{b} - \mathbf{A}\mathbf{r}^0, \quad \mathbf{p}^0 = \mathbf{r}^0, \tag{3.36}$$

$$\mathbf{r}^n = \mathbf{r}^{n-1} - \eta^{n-1}\mathbf{A}\mathbf{p}^{n-1}, \tag{3.37}$$

$$\mathbf{p}^n = \mathbf{r}^n + \zeta^{n-1}\mathbf{p}^{n-1}, \quad \text{for iterates } n = 1, 2, \cdots, \tag{3.38}$$

where the only difference in the algorithms are in the formulas η^{n-1} and ζ^{n-1}. The algorithms are based on seeking out approximations in the affine space $\mathbf{x}^0 + \mathcal{W}$ given an initial guess x^0, using a succession of orthogonal projections onto the subspace $\mathcal{W}$ which leads to the conditions

$$\mathbf{r}^0 \perp \mathcal{W} \quad \text{and} \quad \mathbf{A}\mathbf{p}^n \perp \mathcal{W} \tag{3.39}$$

that determine the parameters η^{n-1} and ζ^{n-1}. When $\mathbf{A}$ is Hermitian positive definite,

- $\mathcal{W} = \mathcal{K}^n(\mathbf{A}, \mathbf{r}^0)$ leads to the CG method
- $\mathcal{W} = \mathbf{A}\mathcal{K}^n(\mathbf{A}, \mathbf{r}^0)$ leads to the CR method,

where $\mathcal{K}^n(\mathbf{A}, \mathbf{r}^0)$ is the Krylov subspace. Thus, in the CR method residual vectors are $\mathbf{A}$-orthogonal, and the $\mathbf{A}\mathbf{p}^n$'s are orthogonal, or equivalently stated, $\mathbf{p}^n$'s are $\mathbf{A}^T\mathbf{A}$-orthogonal, i.e. the inner product $(\mathbf{A}\mathbf{p}^n, \mathbf{A}\mathbf{p}^n) = (\mathbf{A}\mathbf{p}^n)^T(\mathbf{A}\mathbf{p}^n) = \mathbf{p}^{nT}\mathbf{A}^T\mathbf{A}\mathbf{p}^n = (\mathbf{A}^T\mathbf{A}\mathbf{p}^n, \mathbf{p}^n) = 0$. In the GCR method, rather than storing only the previous iteration of search directions, one considers a sequence of these search directions $\mathbf{p}^0, \mathbf{p}^1, \cdots, \mathbf{p}^{n-1}$, where each set

$\{\mathbf{p}^0, \mathbf{p}^1, \cdots, \mathbf{p}^{j-1}\}$ for $j \leq n$ forms a basis for the Krylov subspace $\mathcal{K}^j(\mathbf{A}, \mathbf{r}^0)$, where

$$(\mathbf{A}\mathbf{p}^i, \mathbf{A}\mathbf{p}^k) = 0, \quad \text{for} \quad i \neq k. \tag{3.40}$$

The associated residual and search direction vector analogues to (3.36)-(3.38) in addition to the approximate solution vector are

$$\mathbf{x}^n = \mathbf{x}^{n-1} + \eta^{n-1}\mathbf{A}\mathbf{p}^n, \quad \mathbf{r}^n = \mathbf{r}^{n-1} - \eta^{n-1}\mathbf{A}\mathbf{p}^n \tag{3.41}$$

$$\mathbf{p}^n = \mathbf{r}^n + \sum_{i=0}^{n-1} \zeta^{in}\mathbf{A}\mathbf{p}^i \tag{3.42}$$

where the parameters are then determined via orthogonality conditions

$$(\mathbf{r}^n, \mathbf{A}\mathbf{p}^i) = 0 \quad i = 0, \cdots, n-1 \tag{3.43}$$

and (3.40). This will in turn yield the approximate solution $\mathbf{x}^n$, with the smallest residual norm in the affine space $\mathbf{x}^0 + \mathcal{K}^n$.

As the interface problem $\mathbf{F}_I$ here is complex symmetric and not Hermitian as was the case for the CR and GCR methods, the main difference becomes in the complex conjugate. Therefore, in order to determine the parameters η^{n-1} and ζ^{in}, needed in the computation of $\mathbf{r}^n$ and the sequence $\mathbf{p}^0, \mathbf{p}^1, \cdots, \mathbf{p}^{n-1}$ for a complex symmetric matrix $\mathbf{A}$, one considers the following choice of subspace

- $\mathcal{W} = \overline{\mathbf{A}}\mathcal{K}^n(\overline{\mathbf{A}}, \overline{\mathbf{r}}^0)$,

and thus we have that for $\mathbf{A} = \mathbf{F}_I$, the search directions must satisfy the orthogonality conditions

$$(\mathbf{F}_I\mathbf{p}^i, \overline{\mathbf{F}_I\mathbf{p}^k}) = (\overline{\mathbf{F}}_I^T\mathbf{F}_I\mathbf{p}^i, \overline{\mathbf{p}^k}) = 0 \text{ for } i \neq k, \tag{3.44}$$

$$(\mathbf{r}^n, \overline{\mathbf{F}_I\mathbf{p}^i}) = 0 \text{ for } i = 0, \cdots, n-1. \tag{3.45}$$

Any approximate solution iterate, $\boldsymbol{\lambda}^n$, can be computed from $\boldsymbol{\lambda}^{n-1}$, where by exploiting orthogonality with the vectors $\overline{\mathbf{F}_I\mathbf{p}^j}$ for $j = 1, \cdots, n-1$, gives

$$\eta^{n-1} = \frac{(\mathbf{r}^{n-1}, \overline{\mathbf{F}_I\mathbf{p}^{n-1}})}{(\mathbf{F}_I\mathbf{p}^{n-1}, \overline{\mathbf{F}_I\mathbf{p}^{n-1}})}, \tag{3.46}$$

so that $\boldsymbol{\lambda}^n = \boldsymbol{\lambda}^{n-1} + \eta^{n-1}\mathbf{p}^{n-1}$. Multiplying this equation with $\overline{\mathbf{F}}_I^T\mathbf{F}$ and taking the inner product with any $\overline{\mathbf{p}^n} \in \overline{\mathbf{A}}\mathcal{K}^n$, and using (3.44) we have the orthogonality condition

$$(\overline{\mathbf{F}}_I^T\mathbf{F}(\boldsymbol{\lambda}^n - \boldsymbol{\lambda}^{n-1}), \overline{\mathbf{p}^n})) = 0,$$

which implies that the algorithm used to solve (3.33), minimizes the $\overline{\mathbf{F}}_I^T\mathbf{F}$ norm of the error, i.e. *minimizes* $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^n\|_{\overline{\mathbf{F}}_I^T\mathbf{F}_I}$. Below we summarize the algorithm we used in calculating the interface problem, where here we designate the complex conjugate transpose with $*$.

1. Initialize vectors

$$\begin{aligned} \lambda^0 &= 0, \quad \mathbf{r}^0 = \mathbf{d} - \mathbf{F}_I\lambda^0 \\ \mathbf{y}^0 &= \mathbf{r}^0 \\ \mathbf{p}^0 &= \mathbf{y}^0 \end{aligned}$$

2. iterate until convergence $n = 1, 2, \cdots$

$$\begin{aligned} \eta^n &= (\mathbf{F}_I\mathbf{p})^{n-1*}\mathbf{r}^{n-1}/(\mathbf{F}_I\mathbf{p})^{n-1*}(\mathbf{F}_I\mathbf{p})^{n-1} \\ \boldsymbol{\lambda}^n &= \boldsymbol{\lambda}^{n-1} + \eta^n\mathbf{p}^{n-1}, \\ \mathbf{r}^n &= \mathbf{r}^{n-1} - \eta^n(\mathbf{F}_I\mathbf{p})^{n-1}, \\ \mathbf{y}^n &= \mathbf{r}^n, \\ \text{Compute } \zeta^{in} &= -(\mathbf{F}_I\mathbf{p})^{i*}(\mathbf{F}_I\mathbf{p})^n/(\mathbf{F}_I\mathbf{p})^{i*}(\mathbf{F}_I\mathbf{p})^i \text{ for } i = 0, 1, \cdots n-1 \\ \mathbf{p}^n &= \mathbf{y}^n + \sum_{i=1}^{n-1}\zeta^{in}\mathbf{p}^i \end{aligned}$$

Some Comments on Preconditioning

The above algorithm is referred to as the unpreconditioned regularized FETI method for complex problems. While the unpreconditioned method converges, and is numerically scalable with respect to the mesh size, it is however not scalable with respect to the number of subdomains. Moreover, a study done in [25] also shows that in the absence of preconditioners the convergence of the regularized FETI method increases sublinearly with the wave number. The methodology for preconditioning the regularized FETI method is described in [25, 26] where at each iteration, the interface residual generated by the GCR algorithm is preconditioned by solving an auxiliary second-level problem obtained by projecting the interface problem onto a suitable coarse space. Such a coarse space preconditioner numerically scales with respect to the number of subdomains and the wave number and thus accelerates the convergence.

The current work does not implement the coarse space preconditioner as the main purpose of the work is to develop the exact bound method and to demonstrate the bounding properties. However, as the method is applied to three-dimensional wave equation in simple geometries using structured meshes, the computational cost in solving the interface problem is still affordable even in the absence of preconditioner.

Chapter 4

Exact Bounds Method for the Poisson Equation

The strategy involved in the computation of bounds to exact outputs of interest is similar to the former hierarchical method [48, 50], in that it involves decomposing the global mesh into several elemental subdomains and relaxing the continuity requirements along the edges of each subdomain. A Lagrangian is first constructed so that the output problem is recast as a constrained minimization problem where the constraints are the continuity requirements along the edges of the subdomains and the equilibrium equation. The gradient condition of the Lagrangian will then lead to the primal-adjoint pair and the equilibration equation that will determine the candidate inter-element continuity multipliers. The bounds are finally obtained through a local sub-problem calculation. At this stage, the method differs from the former two-level residual method because by exploiting the Lagrangian saddle point property, existence of such bounds to the exact solution is guaranteed, however the bounds are practically un-computable. The key ingredient relies on constructing a complementary energy functional chosen from a suitable finite dimensional set that can be used to bound the infinite dimensional problem [58, 59].

In this chapter, we extend the exact bounds method proposed in [58] for the Poisson's equation to three space dimensions. While in [58], Ladevéze's procedure [38] is used in the

calculation of the hybrid fluxes, we invoke the FETI method which has been shown [47] in the context of the two-level residual method of obtaining bounds to reduce computational time and memory.

4.1 Bounds on Energy

Preceding our objective of obtaining bounds on selected quantities of interest for the Helmholtz equation, we highlight some of the main theoretical aspects of the method, by applying the ideas to obtain energy bounds for the Poisson equation. Considering a simple boundary value problem with homogenous Dirichlet boundary conditions, we can define both trial and test space as: $\mathcal{U}(\Omega) \equiv \{u \in H^1(\Omega) \Big| \ u|_{\Gamma \cap \Omega} = 0\}$ and consider the problem of finding $u \in \mathcal{U}$, such that

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega, \ \ \forall v \in \mathcal{U}. \tag{4.1}$$

The total energy of the system which we would want to bound is

$$\varepsilon(u) = \frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \, d\Omega - \int_\Omega f \, u \, d\Omega = -\frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \, d\Omega \tag{4.2}$$

using Equation (4.1). The physical principle involved here is that the solution u minimizes the energy with respect to all other candidates in $\mathcal{U}$, which can be stated mathematically as

$$u = \arg \inf_{\omega \in \mathcal{U}} \frac{1}{2}\int_\Omega \nabla \omega \cdot \nabla \omega \, d\Omega - \int_\Omega f \omega \, d\Omega.$$

This statement is a result owing to the convexity of the energy functional which guarantees a unique minimizer. If we search for a discrete approximation of u from the finite set of functions $\mathcal{U}_h \subset \mathcal{U}$, then the convexity implies that any approximation of the energy term, will approach the minimum from above as depicted in Figure 4.1. Therefore, in the case of bounding the energy, the upper bound is just a consequence of the finite element

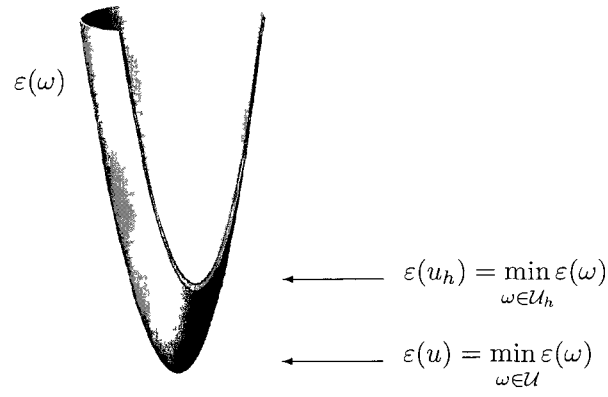


Figure 4.1: Conforming nature of the upper bound. Here we see that any set of conforming functions will approach the exact minimum from above; thus $\varepsilon(u_h) \geq \varepsilon(u)$.

approximation, however the upper bound alone is not enough to certify the error in the solution.

In order to obtain a lower bound, we commence with the domain decomposition strategy discussed in the previous chapter, and relax the continuity between subdomains in the partitioning of the domain Ω, through the use of Lagrange multipliers. Such a strategy leads to the introduction of the broken spaces (hat spaces)

$$\widehat{\mathcal{U}} \equiv \{v | v \in L^2(\Omega), v|_T \in H^1(T), \forall T \in \mathcal{T}_h\} \tag{4.3}$$

where here we use the notation given in [58] and denote $\mathcal{T}_h$ as the mesh consisting of these non-overlapping subdomains $T := \Omega^{(s)} \in \mathcal{T}_h$ (as was earlier presented in the previous chapter where $\Omega^{(s)}$ has been defined). As the global mesh is the accumulation of these subdomains we identify the integral over the broken domains as sums of integrals over the subdomains, i.e.

$$\int_\Omega \nabla\hat{\omega} \cdot \nabla\hat{v}\, d\Omega = \sum_{T \in \mathcal{T}_h} \int_T \nabla\hat{\omega}|_T \cdot \nabla\hat{v}|_T\, d\Omega, \quad \text{where recall } \hat{\omega}|_T := \omega^{(s)}.$$

4.1.1 Lagrangian Formulation

By enforcing continuity, we can re-formulate the above minimization statement where again λ is used to indicate the Lagrange multipliers which have ($\pm$) signing at the interfaces between two subdomains tracked by σ_T. Let $\gamma \in \partial\mathcal{T}_h$ be a face of an elemental subdomain, then we express the continuity constraint as the strong statement given by: $\hat{\omega}|_{T,\gamma} - \hat{\omega}|_{T_N,\gamma} = 0$ on γ. Since $\hat{\omega} \in H^1(T)$, the trace of $\hat{\omega}$ on γ, will satisfy $\hat{\omega}|_\gamma \in H^{1/2}(\partial T)$, and $\lambda|_\gamma$ will belong to the dual of the trace space: $\Lambda(\partial T) = H^{-1/2}(\partial T)$. The weak form of such a statement then can be expressed for all $\lambda_\gamma \in \Lambda(\gamma)$ as $\int_\gamma (\hat{\omega}|_{T,\gamma} - \hat{\omega}_{T_N,\gamma})\lambda_\gamma \, d\Gamma = 0$. Therefore, such a re-formulation is now written as

$$\inf_{\hat{\omega}\in\widehat{\mathcal{U}}} \frac{1}{2}\int_\Omega \nabla\hat{\omega}\cdot\nabla\hat{\omega}\, d\Omega - \int_\Omega f\hat{\omega}\, d\Omega \tag{4.4}$$

$$\text{s.t.} \sum_{T\in\mathcal{T}_h} \int_{\partial T} \sigma_T \lambda\hat{\omega}\, d\Gamma = 0 \quad \forall\lambda\in\Lambda. \tag{4.5}$$

with $\Lambda = \Pi_{T\in\mathcal{T}_h} H^{-1/2}(\partial T)$. We now express this as the Lagrangian

$$\mathcal{L}(\hat{\omega},\lambda) \equiv \frac{1}{2}\int_\Omega \nabla\hat{\omega}\cdot\nabla\hat{\omega}\, d\Omega - \int_\Omega f\hat{\omega}\, d\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T}\sigma_T\lambda\hat{\omega}\, d\Gamma \tag{4.6}$$

where optimality is arrived at by imposing the conditions $\nabla_{\hat{\omega}}\mathcal{L} = 0$ and $\nabla_\lambda\mathcal{L} = 0$ which correspond to a saddle point. From the saddle point property and strong duality of convex minimization leads to the inequality

$$\varepsilon^- \le \inf_{\hat{\omega}\in\widehat{\mathcal{U}}} \mathcal{L}(\hat{\omega},\tilde{\lambda}) \le \sup_{\lambda\in\Lambda}\inf_{\hat{\omega}\in\widehat{\mathcal{U}}} \mathcal{L}(\hat{\omega},\lambda) = \inf_{\hat{\omega}\in\widehat{\mathcal{U}}}\sup_{\lambda\in\Lambda}\mathcal{L}(\hat{\omega},\lambda) = \varepsilon$$

which is the basis for the bounds method and for the energy term, obtaining a lower bound. Here $\tilde{\lambda}$ is some candidate Lagrange multiplier which cannot be chosen arbitrarily. The choice for such candidates emerges from the finite element or spectral element approximation of the equilibration equation (as explained in Chapter 2) which is an equation balancing the contribution of the forces acting on the subdomains and consequently, guarantees that the $\inf_{\hat{\omega}\in\widehat{\mathcal{U}}}\mathcal{L}(\hat{\omega},\tilde{\lambda})$ is bounded from below for any choice of $\tilde{\lambda}$, as shown in [58].

4.1.2 Spaces and Multiplier Approximation

The approximation spaces used for the finite dimensional problem are

$$\mathcal{U}_h \equiv \{v \in \mathcal{U} \Big| \hat{v}|_T \in \mathbb{P}^p(T), \forall T \in \mathcal{T}_h\} \tag{4.7}$$

$$\Lambda_h \equiv \{\lambda \in \Lambda \Big| \lambda|_\gamma \in \mathbb{P}^p, \forall \gamma \in \partial\mathcal{T}_h\} \tag{4.8}$$

for $\mathbb{P}^p(T)$ being the space of polynomials on element T (in three-space dimensions) with degree less than or equal to p, and $\mathbb{P}^p(\gamma)$ being the space of polynomials on element faces γ (in two dimensions) with degree less than or equal to p. Moreover, the global representation of the broken space is given by

$$\hat{\mathcal{U}}_h \equiv \left\{ v \in \widehat{\mathcal{U}} \Big| \, v|_{T_h} \in \mathbb{P}^p(T_h), \; \forall T_h \in \mathcal{T}_h \right\}. \tag{4.9}$$

Thus an approximation to the equilibration equation resulting from the gradient condition of the Lagrangian of the last section is given by

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} \sigma_T \lambda_h \hat{v} \, d\Gamma = \int_\Omega \nabla u_h \cdot \nabla \hat{v} \, d\Omega - \int_\Omega f \hat{v} \, d\Omega, \quad \hat{v} \in \widehat{\mathcal{U}}_h \tag{4.10}$$

which solves for λ_h, and thus from the discussion of Chapter 2 is used to calculate $u_h^{(s)}$ in which the accumulation of such contributions from every subdomain leads to the global approximation u_h. As mentioned previously, the particular choice for such a candidate Lagrange multiplier $\tilde{\lambda}$ is taken to be λ_h. Thus it is apparent that we have a lower bound to the exact energy, that is

$$\inf_{\hat{\omega} \in \widehat{\mathcal{U}}_h} \mathcal{L}(\hat{\omega}, \lambda_h) \leq \varepsilon. \tag{4.11}$$

However, as the infimum of the Lagrangian is taken over all $\hat{\omega} \in \widehat{\mathcal{U}}$, it involves the exact solution, and is thus un-computable. The strategy then, involves the use of a complementary energy functional, in order to bound the infinite dimensional problem from below. We will see that any approximation to this new functional will involve a computable lower bound to the lower bound in (4.11).

4.1.3 Subproblem Calculations

Computable lower bounds are obtained through local independent subproblem calculations where from the lower bounding property given in (4.11) we write

$$\inf_{\hat{\omega}\in\widehat{\mathcal{U}}} \mathcal{L}(\hat{\omega};\lambda_h) = \sum_{T\in\mathcal{T}_h} \inf_{\omega\in\mathcal{U}(T)} J_T(\omega)$$

where $J_T(\omega)$ is the localized Lagrangian (subdomain localized) given by

$$J_T(\omega) = \frac{1}{2}\int_T \nabla\omega\cdot\nabla\omega\, d\Omega - \int_T f\omega\, d\Omega - \int_{\partial T} \sigma_T\lambda_h\omega\, d\Gamma. \tag{4.12}$$

Now by defining the complimentary functional by

$$J_T(\mathbf{q}) \equiv \frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\, d\Omega \tag{4.13}$$

for functions $\mathbf{q}\in H(\mathrm{div};T)$ where

$$H(\mathrm{div};T) \equiv \{\mathbf{q}|\mathbf{q}\in(L^2(T))^3, \nabla\cdot\mathbf{q}\in L^2(T)\},$$

we consider the inequality

$$\frac{1}{2}\int_T(\mathbf{q}-\nabla\omega)^2\, d\Omega = \frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\, d\Omega + \frac{1}{2}\int_T \nabla\omega\cdot\nabla\omega\, d\Omega - \int_T \mathbf{q}\cdot\nabla\omega\, d\Omega \geq 0.$$

Thus from Green's identity: $-\int_T \mathbf{q}\cdot\nabla\omega\, d\Omega = \int_T \nabla\cdot\mathbf{q}\,\omega\, d\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,\omega\, d\Gamma$ we have

$$\frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\, d\Omega + \frac{1}{2}\int_T \nabla\omega\cdot\nabla\omega\, d\Omega + \int_T \nabla\cdot\mathbf{q}\,\omega\, d\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,\omega\, d\Gamma \geq 0$$

Choosing $\mathbf{q}$ from a space $\mathcal{Q}$ defined over each subdomain T given by

$$\mathcal{Q}(T) \equiv \Big\{\mathbf{q}\in H(\mathrm{div};T)\Big|\ \nabla\cdot\mathbf{q} = f \quad \text{in } T,$$
$$\mathbf{q}\cdot\mathbf{n} = \sigma_T\lambda_h \quad \text{on } \partial T\Big\}$$

where upon substitution of this choice of $\mathbf{q}$ in the above inequality leads to

$$\frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\, d\Omega + \frac{1}{2}\int_T \nabla\omega\cdot\nabla\omega\, d\Omega + \int_T f\,\omega\, d\Omega - \int_{\partial T}\sigma_T\lambda_h\,\omega\, d\Gamma \geq 0.$$

Using (4.12) and (4.13), we arrive at the lower bound, but in order to find the best possible we consider the supremum over all $\mathbf{q} \in \mathcal{Q}$, in order to obtain

$$\sup_{\mathbf{q}\in\mathcal{Q}(T)} -J_T^c(\mathbf{q}) \leq \inf_{\omega\in\mathcal{U}(T)} J_T(\omega). \tag{4.14}$$

4.1.4 Bounds Procedure

Here we will not go into the detail of the discrete equation used to approximate the bound components $\mathbf{q}$ as we discuss this in the next section. However, we point out that the global lower bound consists of an aggregate of these local quantities, where

$$J(\hat{\omega}) = \sum_{T\in\mathcal{T}_h} J_T(\omega|_T), \quad J^c(\hat{\mathbf{q}}) = \sum_{T\in\mathcal{T}_h} J_T^c(\mathbf{q}|_T)$$

and the space $\hat{\mathcal{Q}} = \Pi_{T\in\mathcal{T}_h}\mathcal{Q}(T)$. To elucidate the idea behind the method further, the functional $-J^c(\hat{\mathbf{q}})$ is negative definite and is a lower bound to the infinite dimensional problem, where we see from Figure 4.2 that any approximation $-J^c(\mathbf{p}_h)$ for $\mathbf{p}_h \in \mathcal{Q}_h(T)$, will approach the exact solution $-J^c(\hat{\mathbf{q}})$ from below, and thus will procure a computable lower bound to the exact energy $J(\omega)$. Therefore, in summary in order to achieve the bounds $\varepsilon_h^- \leq \varepsilon \leq \varepsilon_h^+$, the bounding procedure contains of the following steps:

1. Calculate the global equilibration using FETI: Find $\lambda_h \in \Lambda_h$ such that

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} \sigma_T \lambda_h \hat{v}\, d\Gamma = \int_\Omega \nabla u_h \cdot \nabla \hat{v}\, d\Omega - \int_\Omega f\hat{v}\, d\Omega \quad \forall \hat{v} \in \widehat{\mathcal{U}}_h.$$

2. Use λ_h to calculate the decoupled solutions $u_h^{(s)}$.

3. Calculate the upper bound, which is just a result of a finite element approximation to Equation (4.2) given by

$$\varepsilon_h^+ = -\frac{1}{2}\int_\Omega \nabla u_h \cdot \nabla u_h\, d\Omega = -\frac{1}{2}\sum_{T\in\mathcal{T}_h}\int_T \nabla u_h|_T \cdot \nabla u_h|_T\, d\Omega.$$

4. Calculate the dual Approximations: Find ε_h^- such that

$$\varepsilon_h^- = \sup_{\hat{\mathbf{q}}_h \in \widehat{\mathcal{Q}}_h} -J^c(\hat{\mathbf{q}}_h).$$

where the approximations $\mathbf{q}_h$ require knowledge of already computed values λ_h,and $u^{(s)}$.

The last step consists of several sub-problem computations where we discuss in more detail when we introduce the output bounds procedure in the next section. We would only point out from the properties of the energy bounds [58] that the lower bound ε_h^- must hold for all levels of refinement and converge asymptotically to the exact solution at the same rate as the finite element approximation converges to the exact energy. In the case of the energy, the finite element approximation of the energy term coincides with the upper bound.

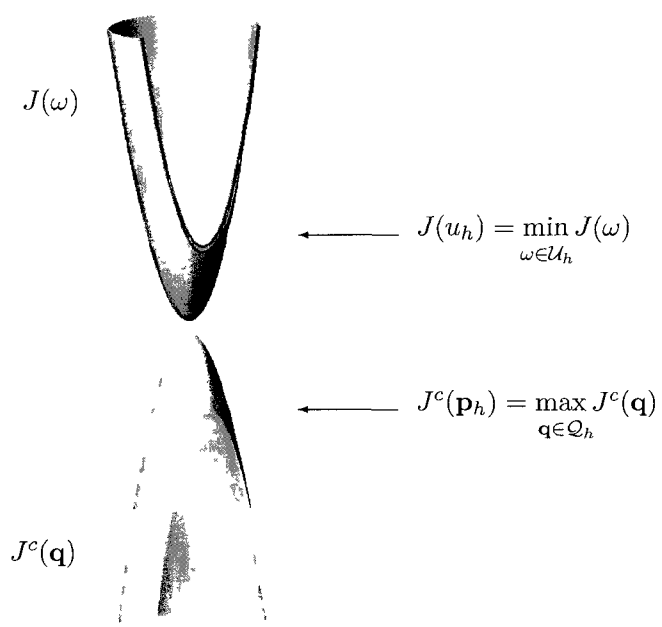


Figure 4.2: Lower bounding property. Here we see that any approximation of the complimentary energy functional will bound the exact energy functional such that $J^c(\mathbf{q}_h) \leq J(\omega)$ for the choice of a $\mathbf{q}_h$ chosen appropriately, namely from the finite dimensional set $\mathcal{Q}_h$.

4.2 Bounds on Quantitative Outputs

In this section we formulate the bounds for quantities of interest s to the Poisson equation (4.1). Such quantities are expressed as linear functionals

$$s = \int_{\Omega^\circ} f^\circ u \, d\Omega \tag{4.15}$$

for $f^\circ \in H^{-1}(\Omega^\circ)$, where Ω° is the region which the output component is specified. The objective here is to find rigorous upper and lower bounds $s^\pm$ to the exact output s such that $s^- \leq s \leq s^+$ holds for all levels of refinement of the mesh.

In the energy bound procedure, the minimization of the Lagrangian also corresponded to the solution of the original problem. In contrast, for the case of output bounds, the variational procedure must be set up such that a minimization of the Lagrangian, will yield the exact output. However, in order to guarantee that the minimizer also corresponds to the solution of the original problem, we must impose it as a constraint. Now we define the following linear and bi-linear forms

$$\begin{aligned} a(w,v) &= \int_\Omega \nabla w \cdot \nabla v \, d\Omega, \quad \ell(v) = \int_\Omega f v \, d\Omega, \\ \ell^\circ(v) &= \int_{\Omega^\circ} f^\circ v \, d\Omega, \quad \mathsf{h}(w,\lambda) = \sum_{T\in\mathcal{T}_h} \int_{\partial T} \sigma_T \lambda w \, d\Gamma \end{aligned}$$

and express the exact output with energy functional $\mathcal{E}(\hat{\omega}^\pm) : \widehat{U} \mapsto \mathbb{R}$ as:

$$\mp s = \inf_{\hat{\omega}^\pm \widehat{\mathcal{U}}} \mp \ell^\circ(\hat{\omega}^\pm) + \frac{\kappa}{2} \overbrace{\left\{ a(\hat{\omega}^\pm, \hat{\omega}^\pm) - \ell(\hat{\omega}^\pm) + \ell(\tilde{u}) - a(\hat{\omega}^\pm, \tilde{u}) \right\}}^{\mathcal{E}(\hat{\omega}^\pm)} \tag{4.16}$$

$$\text{s.t} \quad a(\hat{\omega}^\pm, \psi) = \ell(\psi) \tag{4.17}$$

$$\mathsf{h}(\hat{\omega}^\pm, \lambda) = 0 \tag{4.18}$$

where by varying the sign of the original output we can obtain an upper and lower bound. Here $\tilde{u}$ is chosen to be an element from the set $\mathcal{U}$, thus as $\hat{\omega} \in \widehat{\mathcal{U}}$ and $\mathcal{U} \subset \widehat{\mathcal{U}}$, then $\hat{e} = \hat{\omega} - \tilde{u} \in$

$\widehat{\mathcal{U}}$ can be thought of as the error in the energy functional $\mathcal{E}(\hat{\omega}^{\pm}) = a(\hat{\omega}^{\pm}, \hat{\omega}^{\pm} - \tilde{u}) - \ell(\hat{\omega}^{\pm} - \tilde{u})$, with the property that $\mathcal{E}(\hat{\omega}) = 0$ when $\hat{\omega} = u$ for an element $\tilde{u}$ chosen appropriately from the set $\mathcal{U}$. Therefore, the objective functional: $\mp\ell^{\circ}(\hat{\omega}^{\pm}) + \frac{\kappa}{2}\mathcal{E}(\hat{\omega}^{\pm})$ will be minimized and the minimum will be the desired output of interest. However to ensure that the original Poisson equation is satisfied, we have included it as a constraint with the Lagrange multipliers ψ in addition to enforcing continuity along the faces with the hybrid fluxes λ. The use of a parameter κ is made, which is standard in the literature, and its purpose is made later in optimizing the bounds.

4.2.1 Lagrangian Formulation

The Lagrangian $\mathcal{L} : \widehat{\mathcal{U}} \times \mathcal{U} \times \Lambda \mapsto \mathbb{R}$, now having constraints where weak continuity is imposed along the faces of the subdomains and the equilibrium equation is re-enforced is expressed as

$$\begin{aligned} \mathcal{L}^{\pm}(\hat{\omega}^{\pm}; \psi^{\pm}, \lambda^{\pm}) \equiv \mp\ell^{\circ}(\hat{\omega}^{\pm}) \; &+ \; \frac{\kappa}{2}\left\{a(\hat{\omega}^{\pm}, \hat{\omega}^{\pm} - \tilde{u}) - \ell(\hat{\omega}^{\pm} - \tilde{u})\right\} \\ &+ \; \ell(\psi^{\pm}) - a(\hat{\omega}^{\pm}, \psi^{\pm}) - \mathsf{h}(\hat{\omega}^{\pm}, \lambda^{\pm}), \end{aligned} \tag{4.19}$$

where in order to achieve the output bounds the saddle point property of Lagrange multipliers and strong duality of convex minimization must be employed leading to the inequality:

$$\inf_{\hat{\omega}^{\pm} \in \hat{\mathcal{U}}} \mathcal{L}^{\pm}(\hat{\omega}^{\pm}, \tilde{\psi}^{\pm}, \tilde{\lambda}^{\pm}) \leq \sup_{\substack{\psi^{\pm} \in \mathcal{U} \\ \lambda^{\pm} \in \Lambda}} \inf_{\hat{\omega}^{\pm} \in \hat{\mathcal{U}}} \mathcal{L}^{\pm}(\hat{\omega}^{\pm}, \psi^{\pm}, \lambda^{\pm}) = \mp s \tag{4.20}$$

for some candidate Lagrange multipliers $(\tilde{\psi}^{\pm}, \tilde{\lambda}^{\pm}) \in \mathcal{U} \times \Lambda$. The bound (4.20) is uncomputable in general since it requires knowledge of the exact solution. We follow the work presented in [58] in order to procure computable upper and lower bounds to the bound in (4.20). This task first requires the approximate solutions to the Lagrange multi-

pliers $\lambda_h^{(s)}$, and the decoupled solutions, $u_h^{(s)}$, $\psi_h^{(s)}$. These approximations are then used in several subdomain calculations.

4.2.2 Lagrange Multiplier Approximation

The argument infimum of the Lagrangian for candidate Lagrange multipliers is derived at through its first variational form. Making the decompositions $\lambda^{\pm} = \frac{\kappa}{2}\lambda_h^u \pm \lambda_h^{\psi}$, and $\psi^{\pm} = \pm\psi_h$ the gradient condition of the Lagrangian leads to two equilibration equations, which upon using the FETI method yields the hybrid flux approximations and the approximation to the decoupled solutions $u^{(s)}$, and $\psi^{(s)}$. The resulting set of equilibrium equations are independent of κ and are written as:

1. Find $\lambda_h^u \in \Lambda_h$ such that

$$\mathsf{h}(\hat{v}, \lambda_h^u) - a(u_h, \hat{v}) = -\ell(\hat{v}) \quad \forall \hat{v} \in \widehat{\mathcal{U}}_h, \tag{4.21}$$

2. Find $\lambda_h^{\psi} \in \Lambda_h$ such that

$$\mathsf{h}(\hat{v}, \lambda_h^{\psi}) + a(\psi_h, \hat{v}) = -\ell^{\circ}(\hat{v}) \quad \forall \hat{v} \in \widehat{\mathcal{U}}_h, \tag{4.22}$$

where the approximation spaces used here have been defined in (4.7)-(4.9).

4.2.3 Local Dual Sub-problems

In order to apply the same ideas as in the energy bounds where a finite dimensional space was used to bound the infinite dimensional problem, we need to first consider the local contributions of the Lagrangian and bound the energy term with a complementary energy functional and treat all other terms as forcing terms. The local Lagrangian in the

appropriate form necessary for obtaining bounds, has the disposition

$$\begin{aligned}\mathcal{L}_T^{\pm}(\omega^{\pm};\pm\tilde{\psi},\frac{\kappa}{2}\tilde{\lambda}^u \pm \tilde{\lambda}^{\psi}) \equiv \frac{\kappa}{2}\int_T \nabla\omega^{\pm}\cdot\nabla\omega^{\pm}\,d\Omega \\ -\frac{\kappa}{2}\left\{\int_T (f-\Delta\tilde{u})\omega^{\pm}\,d\Omega + \int_{\partial T}(\sigma_T\tilde{\lambda}^u + \nabla\tilde{u}\cdot\mathbf{n})\omega^{\pm}\,d\Gamma + \int_T f\tilde{u}\,d\Omega\right\} \\ \mp\left\{\int_T (f^{\circ}-\Delta\tilde{\psi})\omega^{\pm}\,d\Omega + \int_{\partial T}(\sigma_T\tilde{\lambda}^{\psi} + \nabla\tilde{\psi}\cdot\mathbf{n})\omega^{\pm}\,d\Gamma + \int_T f\tilde{\psi}\,d\Omega\right\},\end{aligned}$$

where Green's identity: $-\int_T \nabla u\cdot\nabla\omega\,d\Omega = \int_T \Delta u\omega\,d\Omega - \int_{\partial T}\nabla u\cdot\mathbf{n}\omega\,d\Gamma$ is used to ensure that no term other than the dissipative term $\frac{\kappa}{2}\int_T \nabla\omega\cdot\nabla\omega\,d\Omega$ involves derivatives of $\omega^{\pm}$. Recall $\tilde{u}$ is any element belonging to the space $\mathcal{U}$, and $\tilde{\lambda}^u$,$\tilde{\lambda}^{\psi}$, and $\tilde{\psi}$ are candidate Lagrange multipliers. The Lagrangian can now be written as

$$\mathcal{L}_T^{\pm}(\omega^{\pm};\pm\tilde{\psi},\frac{\kappa}{2}\tilde{\lambda}^u \pm \tilde{\lambda}^{\psi}) = -\frac{\kappa}{2}\int_T f\tilde{u}\,d\Omega \mp \int_T f\tilde{\psi}\,d\Omega + J_T^{\pm}(\omega^{\pm}), \tag{4.23}$$

where

$$J_T^{\pm}(\omega^{\pm}) \equiv \frac{\kappa}{2}\int_T \nabla\omega^{\pm}\cdot\nabla\omega^{\pm}\,d\Omega - \int_T f^{\pm}\omega^{\pm}\,d\Omega - \int_{\partial T} g^{\pm}\omega^{\pm}\,d\Gamma \tag{4.24}$$

for

$$\begin{aligned} f^{\pm} &\equiv \frac{\kappa}{2}\{f-\Delta\tilde{u}\} \pm \{f^{\circ}-\Delta\tilde{\psi}\} \\ g^{\pm} &\equiv \frac{\kappa}{2}\{\sigma_T\tilde{\lambda}^u + \nabla\tilde{u}\cdot\mathbf{n}\} \pm \{\sigma_T\tilde{\lambda}^{\psi} + \nabla\tilde{\psi}\cdot\mathbf{n}\}.\end{aligned}$$

For the positive definite functional

$$J_T^c(\mathbf{q}) \equiv \frac{1}{2}\int_T \mathbf{q}\cdot\mathbf{q}\,d\Omega$$

where $\mathbf{q}\in\mathcal{H}(div;T)$ it is shown [58] that

$$J_T(\omega^{\pm}) \geq -\frac{1}{\kappa}J_T^c(\mathbf{q}^{\pm}), \tag{4.25}$$

provided that $\mathbf{q}^{\pm}$ is chosen from the space of functions

$$\begin{aligned}Q^{\pm}(T) \equiv \Bigg\{\mathbf{q}\in H(\mathrm{div};T)\Bigg| \int_T \nabla\cdot\mathbf{q}\,v\,d\Omega - \int_{\partial T}\mathbf{q}\cdot\mathbf{n}\,v\,d\Gamma \\ = -\int_T f^{\pm}v\,d\Omega - \int_{\partial T} g^{\pm}v\,d\Gamma,\ \forall v\in H^1(T)\Bigg\}.\end{aligned} \tag{4.26}$$

From (4.20) we can write

$$\inf_{\hat{\omega}^- \in \widehat{\mathcal{U}}} \mathcal{L}_T^-(\hat{\omega}^-, \tilde{\psi}^-, \tilde{\lambda}^-) \leq s_T \leq - \inf_{\hat{\omega}^+ \in \widehat{\mathcal{U}}} \mathcal{L}_T^+(\hat{\omega}^+, \tilde{\psi}^+, \tilde{\lambda}^+),$$

where s_T are the local contributions to the output and the upper and lower bounds to this output occur as a consequence of (4.23) and (4.25) which procure bounds to these bounds. With the choice $\tilde{u} = u_h \in \mathcal{U}_h \subset \mathcal{U}$, $\tilde{\psi} = \psi_h$, $\tilde{\lambda}^u = \lambda_h^u$, and $\tilde{\lambda}^\psi = \lambda_h^\psi$, we obtain the non-trivial upper and lower bounds

$$\mp s_T^\pm = \frac{\kappa}{2} \int_T f u_h \, d\Omega \pm \int_T f \psi_h \, d\Omega + \sup_{\mathbf{q}^\pm \in Q^\pm(T)} -\frac{1}{\kappa} J_T^c(\mathbf{q}^\pm). \tag{4.27}$$

Global output bounds result then from the aggregate of these local contribution.

Sub-problem Approximation

We recall that the data u_h, ψ_h, and λ_h in the right hand side of the constraint (4.26) are polynomial approximations of order p, based on the local basis functions used. Therefore a suitable approximation space for the dual feasibility constraint is chosen to be

$$Q^{\pm,q} \equiv Q^\pm \cap (\mathbb{P}^q(T))^3,$$

however, the polynomial order q must chosen so that the right hand side of Equation (4.26) belong to the range of operators on the left. The decomposition $\mathbf{q}_h = \kappa \nabla u_h + \frac{\kappa}{2} \mathbf{q}_h^u \pm \mathbf{q}_h^\psi$, in the minimization (4.25) leads to the two κ independent subproblem calculations:

$$\mathbf{q}_h^u = \arg \inf_{\mathbf{q}_h \in Q_h^u(T)} J^c(\mathbf{q}_h), \tag{4.28}$$

$$\mathbf{q}_h^\psi = \arg \inf_{\mathbf{q}_h \in Q_h^\psi(T)} J^c(\mathbf{q}_h), \tag{4.29}$$

for the dual feasibility approximation sets:

$$\mathcal{Q}_h^u(T) \equiv \left\{ \mathbf{q} \in (\mathbb{P}^q(T))^3 \middle| \int_T \nabla \cdot \mathbf{q} \, v \, d\Omega - \int_{\partial T} \mathbf{q} \cdot \mathbf{n} \, v \, d\Omega = \right. \tag{4.30}$$

$$\left. - \int_T (f + \Delta u_h) \, v \, d\Omega - \int_{\partial T} (\sigma_T \lambda_h^u - \nabla u_h \cdot \mathbf{n}) \, v \, d\Gamma \right\}$$

$$\mathcal{Q}_h^{\psi}(T) \equiv \left\{ \mathbf{q} \in (\mathbb{P}^q(T))^3 \,\middle|\, \int_T \nabla \cdot \mathbf{q}\, v \, d\Omega - \int_{\partial T} \mathbf{q} \cdot \mathbf{n}\, v \, d\Omega = \right. \tag{4.31}$$
$$\left. - \int_T (f^\circ - \Delta\psi_h)\, v \, d\Omega - \int_{\partial T} (\sigma_T \lambda_h^{\psi} + \nabla\psi_h \cdot \mathbf{n})\, v \, d\Gamma \right\}.$$

In order to determine the degrees of freedom of the dual variables $\mathbf{q}$, let us as an example consider re-expressing the restrictions given by (4.30) in its equivalent form

$$\nabla \cdot \mathbf{q}_h^u = -f - \Delta u_h \quad \text{(inside the subdomain)} \tag{4.32}$$

$$\mathbf{q}_h^u \cdot \mathbf{n} = \sigma_T \lambda_h^u - \nabla u_h \cdot \mathbf{n} \quad \text{(on the contour of the subdomain).} \tag{4.33}$$

It is clear that each component of the dual variables $\mathbf{q}$ are to be chosen to be polynomial of degree q. For $u_h|_T \in \mathbb{P}^p(T)$, implies $\Delta u_h|_T \in \mathbb{P}^{p-2}(T)$. In order for this equation to be solvable the forcing function $f|_T$ must be approximated with polynomial of degree $\mathbb{P}^r(T)$ for $q > r$. A similar analysis follows with the constraint in (4.31). In this work, we approximate the forcing and output vectors with the same order as the variables $u_h^{(s)}$, $\psi_h^{(s)}$, and the hybrid fluxes, and so we take $r = p$ with $q > p$ sufficing. On a further note, the two sets of equations (4.32)-(4.33) do not uniquely determine, and moreover there is one equation that is linearly dependent of the others. In order to find $\mathbf{q}$ minimizing the complementary energy, Lagrange multipliers are used where the restrictions are the two sets of the equations previously described. These equations can be imposed as constraints either in their weak form or strong form. Here we impose them weakly.

It can be verified from the equilibration equations (4.21) and (4.22) that by choosing the test function $v \in \mathcal{U}_h$ will result in the vanishing of the right hand side of (4.30) and (4.31), and thus impose the orthogonality condition

$$\sum_{T \in \mathcal{T}_h} \int_T \mathbf{q} \cdot \nabla v \, d\Omega = 0. \tag{4.34}$$

The right hand side of (4.30) and (4.31) are localized residuals forms, and are non-zero

only for the appropriate choice of polynomial approximations of both the primal and dual variables. These set of polynomial fields then allows for the certification of the bounds.

Remark 4.1 *To clarify this last statement, we mention here that by imposing the constraints weakly, we must consider the appropriate space to choose the test function v. If we consider $v \in \mathcal{U}_h$, then $v \in \mathbb{P}^p$, and since all of $u_h, \psi_h . \lambda_h^u, \lambda_h^\psi$ are of polynomial order p as well, then from the equilibration equation,*

$$\begin{aligned} \int_{\partial T} \sigma_T \lambda_h^u v \, d\Gamma - \int_T \nabla u_h \cdot \nabla v \, d\Omega + \int_T f v \, d\Omega &= 0 \\ \int_{\partial T} \sigma_T \lambda_h^\psi v \, d\Gamma + \int_T \nabla \psi_h \cdot \nabla v \, d\Omega + \int_T f^\circ v \, d\Omega &= 0 \end{aligned}$$

as the FETI method applied to this equation, satisfies this over each subdomain exactly. Thus as mentioned previously, the orthogonality condition emerges, and will imply that the dual component for this choice of test function will be the trivial solution $\mathbf{q} = 0$ and thus will not produce bounds. Hence, we choose test functions $v \in \mathbb{P}^q(T)$ so that the right hand side of the subdomain equations (4.30) and (4.31) will have a non-zero residual. However, this implies that as the hybrid fluxes and the primitive variables are of a lower order of approximation, these quantities must be interpolated so that discretely they are defined on the nodal points corresponding to $\mathbb{P}^q(T)$ space.

Now we consider the last step of the procedure, and that is the formulation of the bounds themselves after obtaining the dual variables $\mathbf{q}_h^u$ and $\mathbf{q}_h^\psi$

4.2.4 Output bounds

With the splitting described above, we expand $J_T^c(\mathbf{q}^\pm) = J_T^c(\kappa \nabla u_h + \frac{\kappa}{2}\mathbf{q}_h^u \pm \mathbf{q}_h^\psi)$ in (4.27) and summing the local contributions gives

$$s_h^\pm = -\int_\Omega f\psi_h + \frac{1}{2}\int_\Omega \hat{\mathbf{q}}_h^u \cdot \hat{\mathbf{q}}_h^\psi \, d\Omega \pm \frac{\kappa}{4} J^c(\hat{\mathbf{q}}_h^u) \pm \frac{1}{\kappa} J^c(\hat{\mathbf{q}}_h^\psi)$$

where we have invoked the orthogonality condition (4.34). Letting

$$\bar{s}_h = \frac{1}{2}\int_\Omega \hat{\mathbf{q}}_h^u \cdot \hat{\mathbf{q}}_h^\psi \, d\Omega - \int_\Omega f\psi_h \, d\Omega, \; z_h^u = \frac{1}{4}J^c(\hat{\mathbf{q}}_h^u), \; z_h^\psi = J^c(\hat{\mathbf{q}}_h^\psi)$$

and optimizing with respect to κ, where the optimum $\kappa = \sqrt{\frac{z_h^\psi}{z_h^u}}$, the bounds emerge as

$$s_h^\pm = \bar{s}_h \pm 2\sqrt{z_h^u z_h^\psi} \tag{4.35}$$

where $\bar{s}_h$ is the bound average.

It is shown in [58] that the theoretical convergence rate of both the upper and lower bounds is that of the finite element solution, and that they both approach the exact solution at the same rate, that is

$$s - s_h^- \leq C|u - u_h|_{H^1} \, |\psi - \psi_h|_{H^1} \tag{4.36}$$

$$s_h^+ - s \leq C|u - u_h|_{H^1} \, |\psi - \psi_h|_{H^1}. \tag{4.37}$$

Thus, only when the finite element approximations are in the asymptotic regime, will the bounds converge at the optimal rate.

4.3 Implementation

We have described the FETI procedure in the previous chapter, therefore we will not cover this here. In this section we briefly comment on the interpolation of the hybrid fluxes which is needed as right hand data in the computation of the sub-problems.

4.3.1 Interpolation of the Hybrid Fluxes

In discrete form, the approximation to the inter-partition connections λ_h^u and λ_h^ψ are given by the solution to the system

1. find $(\hat{u}_h, \lambda_h^u) \in \hat{\mathcal{U}} \times \Lambda$, such that

$$B^{\mathrm{T}(\mathrm{s})}\lambda_h^{u(\mathrm{s})} - A^{(\mathrm{s})}\hat{u}_h^{(\mathrm{s})} = -M^{(\mathrm{s})}f^{(\mathrm{s})} \tag{4.38}$$

$$\sum_{k=1}^{N_s} B^{(\mathrm{s})}\hat{u}_h^{(\mathrm{s})} = 0 \tag{4.39}$$

2. find $(\hat{\psi}_h, \lambda_h^{\psi}) \in \hat{\mathcal{U}} \times \Lambda$, such that

$$B^{\mathrm{T}(\mathrm{s})}\lambda_h^{\psi(\mathrm{s})} + A^{(\mathrm{s})}\hat{\psi}_h^{(\mathrm{s})} = -M^{(\mathrm{s})}f^{\mathrm{o}(\mathrm{s})} \tag{4.40}$$

$$\sum_{k=1}^{N_s} B^{(\mathrm{s})}\hat{u}_h^{(\mathrm{s})} = 0 \tag{4.41}$$

where in three-dimensions, A is the stiffness matrix, M is the mass matrix, and B is the signed Boolean matrix described in the previous chapter. For test functions $v \in \mathbb{P}^q(T)$, the hybrid fluxes in the right hand side of constraints (4.30) and (4.31) must be interpolated. The FETI approach solves for the fluxes on each face, which in continuous is given by the quantity

$$\int_{\overline{\Gamma}} \sigma_T \lambda v \, d\overline{\Gamma} \quad \overline{\Gamma} = \text{face of subdomain.}$$

In the discrete form, this corresponds to a product of a two-dimensional mass matrix and the nodal values of the hybrid fluxes which we write as $M^{(f)}\lambda$ (where $M^{(f)}$ is the 2D mass matrix for each face of the elemental subdomains) on each face. Before interpolation the λ's must first be obtained on each face by multiplying by ${M^{(f)}}^{-1}$. The right-hand side of the feasibility constraint now involves integrals where the test functions are approximated using $\mathbb{P}^q(T)$ polynomials and the hybrid fluxes are approximated using only $\mathbb{P}^p(T)$ polynomials. One way to do the interpolation is to introduces a 2D mixed mass matrix $M^{(f)}_{\mathbb{P}^q \times \mathbb{P}^p}$, then the product $M^{(f)}_{\mathbb{P}^q \times \mathbb{P}^p} \cdot {M^{(f)}_{\mathbb{P}^p \times \mathbb{P}^p}}^{-1} \cdot \lambda_h^{(k)}$, gives the interpolated hybrid fluxes over each face of the elemental subdomains.

Figure (4.3), details the interpolation procedure, with both structured and unstructured subdomains which we use in this study to illustrate our results. The tetrahedral subdomains

are just elemental, while the cubic subdomains are comprised of six tetrahedrons. The difference in subdomains is that while the tetrahedrons offer more flexibility in the geometry, the constraint

$$\mathbf{q} \cdot \mathbf{n} = \sigma_T \lambda_h$$

is simpler to implement when the normals are orthogonal to the faces of each subdomain. Interpolating the subdomains include certain numerical errors which affect the accuracy of the bounds. Therefore, our choice for a more structured mesh is to avoid interpolation errors. For illustration purposes, Figure (4.3) depicts the case where we use tetrahedral subdomains with $u_h^{(s)}, \psi_h^{(s)}, \lambda_h \in \mathbb{P}^1$ and $\mathbf{q}_h^u, \mathbf{q}_h^\psi \in \mathbb{P}^2$. Our experience has shown that going to higher order on the unstructured meshes will produce errors which we believe is due to interpolation of the hybrid fluxes. While the finite element solution $u_h^{(s)}, \psi_h^{(s)}, \lambda_h$ converges at their theoretical rate, the bounds lose convergence for the tetrahedral mesh at higher than $p = 1$, $q = 2$. Therefore, in this work we mainly consider a fully structured mesh, where for the case of cubic subdomains we demonstrate results when $p = 2$, $q = 3$. We see from the results section that for this order the bounds retain the theoretical convergence rate of $\mathrm{O}(h^4)$.

4.4 Discrete forms and Sub-problems Computation

As mentioned previously, the minimization statements in Equations (4.28) and (4.29) can be solved using Lagrange multipliers. For the statement (4.28) we construct the Lagrangian,

$$L_T^u(\mathbf{q}_h^u; v) \equiv \frac{1}{2}\int_T \mathbf{q}_h^u \cdot \mathbf{q}_h^u \, d\Omega + \int_T \nabla \cdot \mathbf{q}_h^u \, v \, d\Omega - \int_{\partial T} \mathbf{q}_h^u \cdot \mathbf{n} \, v \, d\Gamma$$
$$+ \int_T (f + \Delta u_h) \, v \, d\Omega + \int_{\partial T} (\sigma_T \lambda_h^u - \nabla u_h \cdot \mathbf{n}) \, v \, d\Gamma$$

where at optimality $\nabla_{\mathbf{q}} L_T^u = 0$, and $\nabla_v L_T^u = 0$ gives the set of equations

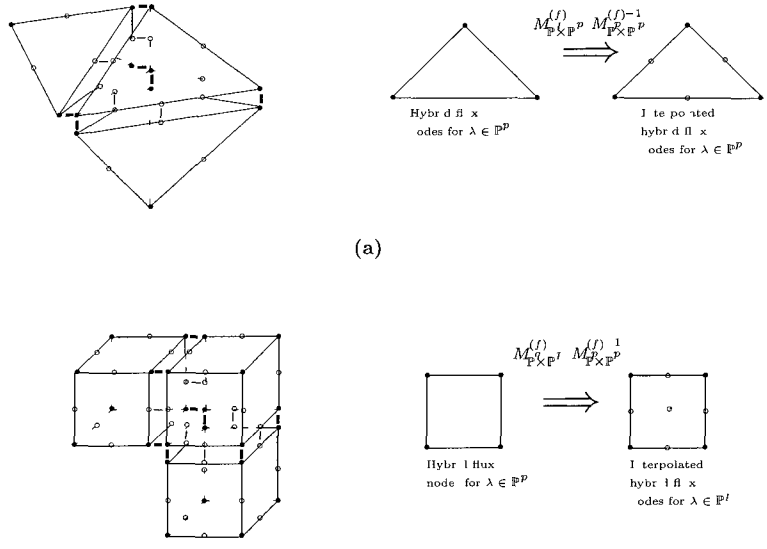


Figure 4.3 **(a)** Illustration of the hybrid flux interpolation on the faces of the elemental tetrahedral subdomains. This thick lines signify the hybrid fluxes at the vertices, calculated from the FETI, and the thinner lines represent the interpolated values given by the operation to the right of it. **(b)** Illustration of the hybrid flux interpolation on the faces of the cubic subdomains.

$$\int_T \mathbf{q}_h^u \cdot \mathbf{r}_h \, d\Omega - \int_T \mathbf{r}_h \cdot \nabla v_h^u \, d\Omega = 0 \qquad \forall \mathbf{r}_h \in \mathcal{P}_h$$

$$-\int_T \mathbf{q}_h^u \cdot \nabla v' \, d\Omega = \int_T \nabla u_h \cdot \nabla v' \, d\Omega - \int_T f v' \, d\Omega - \int_{\partial T} \sigma_T \lambda_h^u v' \, d\Gamma \qquad \forall v' \in \mathbb{P}^q,$$

which has the matrix representation

$$\begin{bmatrix} M^{(s)} & 0 & 0 & -D_1^{(s)} \\ 0 & M^{(s)} & 0 & -D_2^{(s)} \\ 0 & 0 & M^{(s)} & -D_3^{(s)} \\ -D_1^{t(s)} & -D_2^{t(s)} & -D_3^{t(s)} & 0 \end{bmatrix} \begin{bmatrix} (q_h^u)_1 \\ (q_h^u)_2 \\ (q_h^u)_3 \\ V_h^u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathcal{R}_h^u, \end{bmatrix}$$

Here, the superscript t is used to designate the transpose of the matrix distinguishing it from T which represents the subdomains. The matrices D_1, D_2, D_3 arise from the discretization

of each of the terms in

$$\int_T \mathbf{q} \cdot \nabla v \, d\Omega = \int_T q_1 \frac{\partial v}{\partial x_1} \, d\Omega + \int_T q_2 \frac{\partial v}{\partial x_2} \, d\Omega + \int_T q_3 \frac{\partial v}{\partial x_3} \, d\Omega$$

respectively. Furthermore, $\mathcal{R}_h^u$ is the residual defined to be

$$\mathcal{R}_h^u \equiv -M^{(\mathrm{k})} f^{(\mathrm{s})} + A^{(\mathrm{s})} u_h^{(\mathrm{s})} - B^{\mathrm{T}(\mathrm{s})} \lambda_h^{u^{(\mathrm{s})}}. \tag{4.42}$$

The minimization statement for (4.29) is similar and we will not repeat the process. There are a number of these small size subdomain calculations which involve solving the block system matrix given above, one for $\mathbf{q}_h^u|_T$ and one for $\mathbf{q}_h^\psi|_T$. This system is solved using the Uzawa algorithm which is a two step conjugate gradient method for each subdomain. Accumulation of these quantities over all subdomains lead to the bounds as described previously.

4.4.1 Numerical Examples

The remainder of this section, verifies the results in a cube geometry where two different forcing functions are considered. The output vector $f^\circ = 1.0$ for all cases. Results are obtained on a tetrahedral mesh and are presented in Table 4.1 , where the effectivity index is defined as $\theta_h^- = \frac{|s - s_h^-|}{|s - s_h|}$ and $\theta_h^+ = \frac{|s - s_h^+|}{|s - s_h|}$. Moreover, for the second case, a constant forcing is applied, where we have for for cubic subdomains two different approximations in which we consider linear approximation for the finite element solution $u_h^{(s)}, \psi_h^{(s)}, \lambda_h^{(s)} \in \mathbb{P}^1$ and $\mathbf{q}, v \in \mathbb{P}^2$, and a higher order approximation where $u_h^{(s)}, \psi_h^{(s)}, \lambda_h^{(s)} \in \mathbb{P}^2$ and $\mathbf{q}, v \in \mathbb{P}^3$.

Constructed Exact Solution

In this case we consider as a test for validation the exact solution

$$u(x, y, z) = \sin(\pi x) \sin(\pi y) \sin(\pi z)$$

satisfying the homogeneous Dirichlet boundary conditions. This leads to the forcing function

$$f(x,y,z) = 3\pi^2 \sin(\pi x)\sin(\pi y)\sin(\pi z)$$

where the exact output of interest which we use as a reference solution is calculated to be $s = 8/\pi^3$. Figure (4.4(b)) shows the convergence rate of the bounds and the finite element solution for the linear approximation where $p = 1$, $q = 2$. The bounds asymptotically approach the optimal rate of $O(h^2)$ which is also the asymptotic rate of convergence of the finite element approximation. Figure (4.4(a)) shows the bounding property, that is, it shows that the bounds hold for all levels of refinement and approach the exact solution (dashed line). In addition, it shows the predicted output s_h^{pre}, namely, the bound average $s_h^{pre} = s_{av} = (s_h^+ + s_h^-)/2$. If we consider the normalized error of the bounds: $\Delta s_h^{\pm}/s = |s - s_h^{\pm}|/s$ and $\Delta s_{av}/s = |s - s_{av}|/s$, from Table (4.1) we see that for the mesh size $h = 1/16$, the relative error of the upper and lower bounds are 13% and 4.1% respectively. On the other hand, the bound average is 0.269482, and has a relative error of 4% which compares well with the error in the finite element solution of about 3%. Table (4.1) also gives the effectivity of the bounds which compares how well the bounds converge with respect to the finite element solution. As is observed for this case, the convergence of the lower bound has a much better effectivity than the upper bound. Typically, the bound average will yield sharper results than both the upper and lower bounds, however the good effectivity of the lower bound causes it to have a slightly lower percentage relative error. As for the effectivity of the upper bounds, we point out that for engineering practice, these higher effectivities are acceptable.

Uniformly Forced Domain

Here we take $f = -2.0$ and compare our results with the exact solution

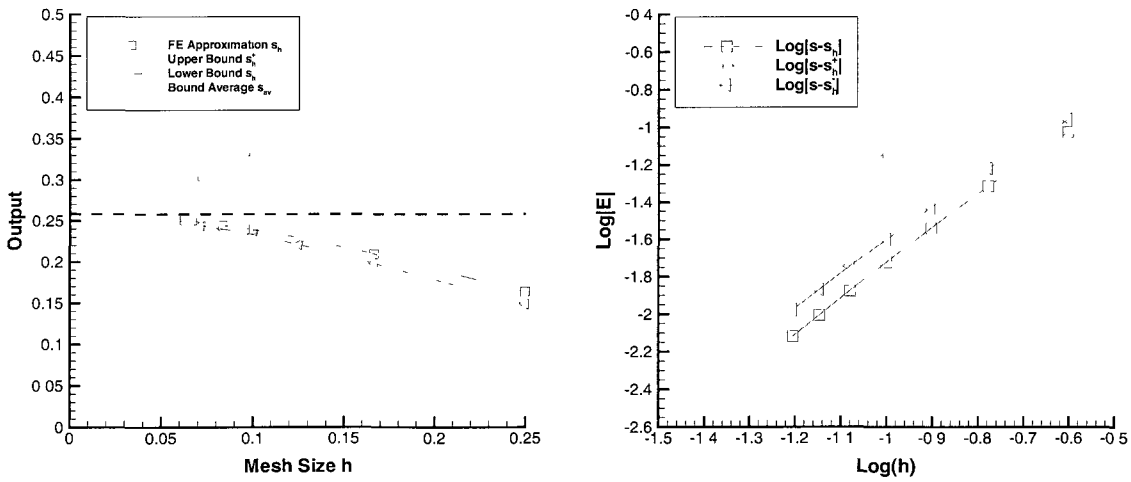
$$s = -\frac{8}{\pi^7}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{[1-(-1)^n]^2\,[1-(-1)^m]^2}{n^2m(n^2+m^2)^{3/2}} \times\left\{\frac{4e^{-\pi\sqrt{n^2+m^2}}}{1-e^{-2\pi\sqrt{n^2+m^2}}}-\frac{2(1+e^{-2\pi\sqrt{n^2+m^2}})}{1-e^{-2\pi\sqrt{n^2+m^2}}}+\pi\sqrt{n^2+m^2}\right\}\approx -0.0403405$$

We note that for the optimal parameter κ, choosing a constant forcing will cause the finite element solution to be equal to either the upper bound or the lower bound depending on the sign of the forcing term. In this case, the upper bound is equal to the finite element approximation. For $p = 1$, $q = 2$, Figure(4.5) again shows plots of the bounds and an asymptotic convergence rate $O(h^2)$. When comparing the effictivities for this case in Table (4.1) we see that as the finite element solution is equal to the upper bound, the effectivity for the upper bound is 1.0 at all the refinements. This is referred to as compliance. The upper bound has a relative error of almost 16% and the lower bound has a relative error of 2.8%. The sharpness of the bounds can be increased with higher-order approximations. A comparison of this with that of Figure(4.6) where higher-order polynomial approximation, $p = 2$, $q = 3$ are used, demonstrates that by using the higher-order polynomials results in significantly sharper bounds. The asymptotic rate of convergence for both the output bounds and the finite element approximations approaches the theoretical rate of $O(h^4)$ for a structured mesh.

$f = 3\pi^2 \sin(\pi x)\sin(\pi y)\sin(\pi z)$ $s = 8/\pi^3 \approx 0.2580122$					
h	$\mathbf{s}_h^-$	$\mathbf{s}_h^+$	$\mathbf{s}_h$	θ_h^-	θ_h^+
1/2	0.038018	0.380210	0.055517	1.0864	0.6035
1/4	0.148498	0.495173	0.163741	1.1617	2.5157
1/6	0.197912	0.416287	0.209552	1.2402	3.2661
1/8	0.220851	0.363288	0.229237	1.2914	3.6585
1/10	0 232964	0.331964	0.239116	1.3256	3.9136
1/12	0.240049	0.312650	0.244702	1.3496	4.1049
1/14	0.244523	0.300063	0.248149	1.3676	4.2633
1/16	0.247519	0.291445	0.250418	1.3817	4.4024

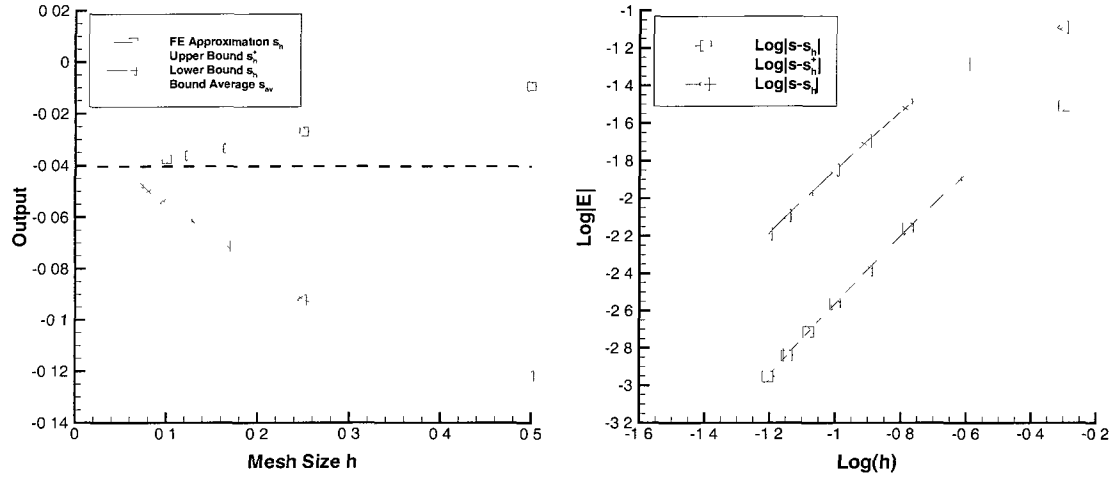
$f = -2.0$ $s \approx -0.0403405$					
h	$\mathbf{s}_h^-$	$\mathbf{s}_h^+$	$\mathbf{s}_h$	θ_h^-	θ_h^+
1/2	−0.121818	−0.009375	−0.009375	2.6312	1.0
1/4	−0.092205	−0 026951	−0.026951	3.8735	1.0
1/6	−0.071412	−0.033440	−0.033440	4.5028	1.0
1/8	−0.060518	−0.036211	−0.036211	4.8862	1.0
1/10	−0.054446	−0.037610	−0.037610	5.1659	1.0
1/12	−0.050774	−0.038408	−0.038408	5.3990	1.0
1/14	−0.048397	−0.038902	−0.038902	5.6006	1.0
1/16	−0.046772	−0.039229	−0.039229	5.7863	1.0

Table 4.1: Tabulated bounds results and their effectivities obtained for the Poisson problem using tetrahedral elemental subdomains with $p = 1, q = 2$.



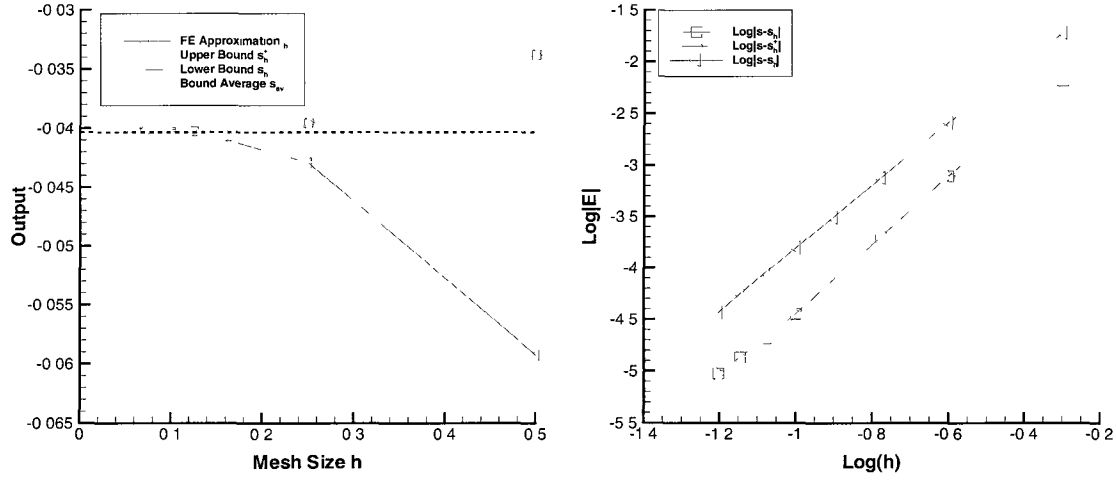
(a) (b)

Figure 4.4: Output bounds obtained for the constructed solution using a tetrahedral subdomains. (a) upper and lower bounds. (b) convergence of the bounds.



(a) (b)

Figure 4.5: Output bounds obtained for a constant forcing (f=-2 0) with $p = 1, q = 2$. (a) upper and lower bounds (b) convergence of the bounds.



(a) (b)

Figure 4.6: Output bounds obtained for a constant forcing (f=-2.0) with $p = 2, q = 3$. (a) upper and lower bounds. (b) convergence of the bounds

4.5 Some Results on Stokes Problem

4.5.1 Model Problem

Here we only present the main results for the Stokes problem in order to demonstrate the effectiveness of the bounds method for coercive bilinear forms where in particular the energy term has an intrinsic minimization principle as in the Poisson case. For details of the bounds calculation and the type of polynomial approximations used, we refer the reader to [15]. The problem considered is a steady, incompressible (density ρ is constant), creeping flow driven by a forcing term in an endless square channel with an array of rectangular obstacles in the center. The flow has a constant dynamic viscosity μ and is assumed to be Newtonian. We let Ω represent the geometry of the domain where the coordinate system is given by (x_1, x_2, x_3), with corresponding unit vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. The pressure gradient in the $\mathbf{x}_3$ direction is the driving force and is expressed as $\frac{\Delta P}{L}$, where L is a scaling length of the channel section, and ΔP is the pressure difference between two reference points with the distance L in the $\mathbf{x}_3$ direction. The fluid velocity and pressure perturbations are periodic in the $\mathbf{x}_3$ direction. We let the fluid velocity be $\mathbf{u} = (u_1, u_2, u_3)$ with u_i being the corresponding component int the $\mathbf{x}_i$ direction and p be the pressure fluctuation field divided by the viscosity μ. The governing equation for the incompressible Stokes flow can be written in indicial notation as $\mathbf{x}_3$ direction.

$$-\frac{\partial^2 u_i}{\partial x_j \partial x_j} + \frac{\partial p}{\partial x_i} = f_i \quad \text{in } \Omega,\ i = 1, 2, 3 \tag{4.43}$$

$$\frac{\partial u_i}{\partial x_i} = 0 \quad \text{in } \Omega \tag{4.44}$$

with boundary conditions:

$$u_i|_{\Gamma_1} = u_i|_{\Gamma_2} \tag{4.45}$$

$$u_i = 0 \quad \text{on the other boundaries.} \tag{4.46}$$

Here $f_i = \frac{\Delta P}{\mu L}$ is the prescribed forcing term in the $\mathbf{x}_i$ direction. For our numerical

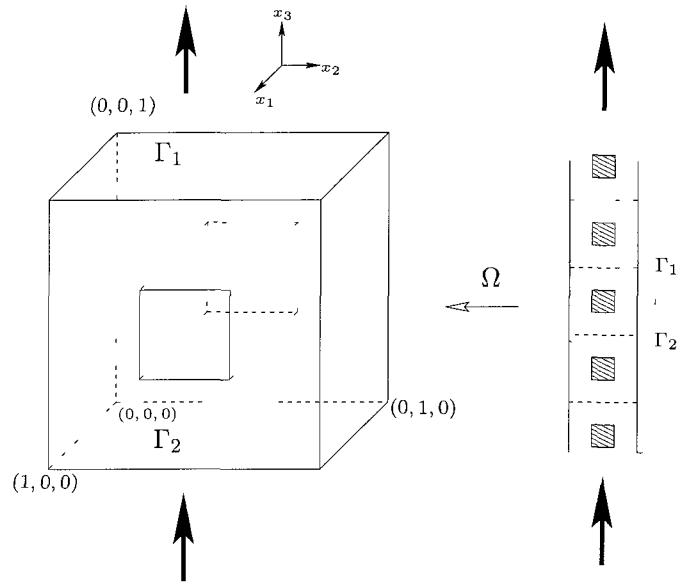


Figure 4.7: Geometry for the Stokes problem. Periodic boundary conditions are imposed on Γ_1 and Γ_2, and homogenous boundary conditions are enforced on $\partial\Omega\backslash(\Gamma_1 \cup \Gamma_2)$

experiments we choosed $\mu = 1$ for simplicity and as a benchmark problems we select $f_1 = f_2 = 0$ and $f_3 = 1$. The geometry then, is the cube Ω (Figure 4.7) is the bounded cubic domain $]0,1[\times]0,1[\times]0,1[$ with rectangular obstacle $]0,1[\times]\frac{1}{3},\frac{2}{3}[\times]\frac{1}{3},\frac{2}{3}[$ inside. Thus $\Omega = \{\,]0,1[\times]0,1[\times]0,1[\,\} - \{\,]0,1[\times\frac{1}{3},\frac{2}{3}[\times]\frac{1}{3},\frac{2}{3}[\,\}$. The periodic boundaries are $\Gamma_1 =]0,1[\times]0,1[$ at $x_3 = 0$ and $\Gamma_2 =]0,1[\times]0,1[$ at $x_3 = 1$.

In order to ensure a unique solution for the pressure, we further impose the additional requirement that

$$\int_\Omega p \, d\Omega = 0. \tag{4.47}$$

The weak formulation of the Stokes problem in terms of its bilinear forms for a given $(f_1, f_2, f_3) \in (H^{-1}(\Omega))^3$ is then written as: Find $(\mathbf{u}, p) \in Y$, such that:

$$a^s(\mathbf{u}, \mathbf{w}) + b^s(\mathbf{w}, p) = \ell^s(\mathbf{w}) \quad \forall \mathbf{w} \in X \tag{4.48}$$

$$b^s(\mathbf{u}, r) = 0 \quad \forall r \in Q \tag{4.49}$$

where these forms are defined explicitly as

$$a^s(\mathbf{v}, \mathbf{w}) = \int_\Omega \frac{\partial v_i}{\partial x_j} \frac{\partial w_i}{\partial x_j} \, d\Omega, \qquad \forall (\mathbf{v}, \mathbf{w}) \in X \times X \tag{4.50}$$

$$b^s(\mathbf{v}, r) = -\int_\Omega r \frac{\partial v_i}{\partial x_i} \, d\Omega, \qquad \forall (\mathbf{v}, r) \in Y \tag{4.51}$$

$$\ell^s(\mathbf{w}) = \int_\Omega f_i w_i \, d\Omega, \qquad \forall \mathbf{w} \in X, \tag{4.52}$$

and the function spaces X, Q, and Y are defined as

$$X = H_0^1(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega) \tag{4.53}$$

$$Q = L^2(\Omega) \tag{4.54}$$

$$Y = X \times Q. \tag{4.55}$$

4.5.2 Output functional

The quantity of interest in this problem is a normalized flow rate where the flow rate is the output values divided by the volume of the computational domain which is $\frac{8}{9}$. The output functional is given as

$$s(\mathbf{u}) = \int_\Omega (\boldsymbol{\alpha} \cdot \mathbf{u}) \, d\Omega \tag{4.56}$$

where $\boldsymbol{\alpha}$ is a unit vector that indicates the direction in which the displacement is evaluated; and is a user defined coefficient where we take this to be $\boldsymbol{\alpha} = (0, 0, 1)$. Therefore, the output can be expressed as

$$\ell^\circ(\mathbf{u}) = \int_\Omega u_3 \, d\Omega. \tag{4.57}$$

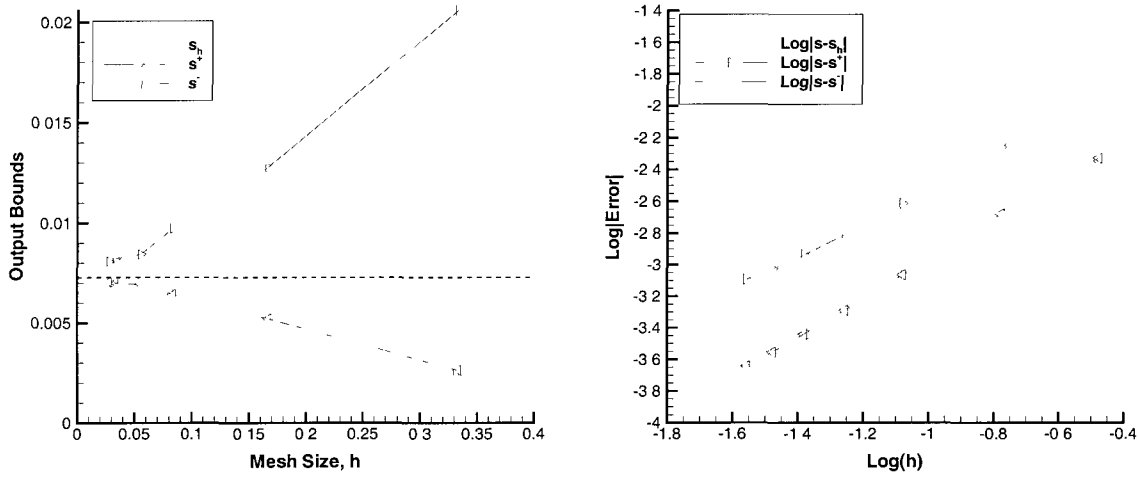
4.5.3 Numerical Example of Bound Calculation for Stokes

Table (4.2) shows the lower bound (s_h^-), upper bound (s_h^+), the average of the bounds (s_{av}), and the discrete output to the solution (s_h) obtained on the mesh size h using the exact

$f_1 = f_2 = 0,\ f_3 = 1.0$ $s \approx 0.007283$				
h	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	$\mathbf{s_{av}}$	$\mathbf{s_h}$
1/3	0.002616	0.020594	0.011605	0.002616
1/6	0.005199	0.012781	0.008990	0.005199
1/12	0.006425	0.009708	0.008067	0.006425
1/18	0.006770	0.008828	0.007799	0.006770
1/24	0.006921	0.008437	0.007679	0.006921
1/30	0.007004	0.008221	0.007613	0.007004
1/36	0.007056	0.008087	0.007572	0.007056

Table 4.2: Tabulated upper bound s_h^+, lower bound s_h^-, finite element output soution s_h, and the bound average s_{av} obtained for the Stokes problem using cubic elemental subdomains.

Stokes bounds method. For the finest mesh, a half bound gap 7.1% is reached. The half bound gap is the bound gap divided by 2 i.e. $(s_h^+ - s_h^-)/2$ and normalized with the most accurate output value. Moreover, the error between the finite element output value on



(a) (b)

Figure 4.8: Bounds for Stokes output: (a) output bounds, where the dashed line represents the exact solution. (b) Convergence of the bounds.

the finest mesh and the most accurate output value is 3.1%, which shows that the bounds are sharp. The most accurate output value is $s \approx \tilde{s} = 0.007283$, and is obtained on a Crouzeix-Raviart finite element space [20] with mesh size $h = 0.02$. We further note that

h	Θ_h^-	Θ_h^+	τ_h	τ_h^-	τ_h^+
1/3	1 0	2 85	−0 93	0	−1 35
1/6	1 0	2 64	−0 82	0	−1 38
1/12	1 0	2 83	−0 91	0	−1 35
1/18	1 0	3 01	−1 01	0	−1 33
1/24	1 0	3 19	−1 09	0	−1 31
1/30	1 0	3 36	−1 18	0	−1 30
1/36	1 0	3 54	−1 27	0	−1 28

Table 4.3: Tabulated results of the effectivities obtained for the Stokes problem using cubic elemental subdomains

for a relatively coarse mesh $h = 1/12$ say, the bounds are quite large, with a half bound gap of 22.5%, but so is the finite element solution for this mesh with a half bound gap of 11.8%.

Figure (4.8) (a) shows the upper and lower bounds $s_h^{\pm}$, and the finite element solution s_h. We note that the lower bounds output value is the same as the finite element output value because the forcing term in the Stokes equations and the output functional are constants.

Figure (4.8) (b) shows the convergence rate for each bound. Here we assume the most accurate numerical solution $\tilde{s}$ has a negligent difference from the exact solution s, and is used for the convergence rate estimations: $e_h^- = |s_h^- - \tilde{s}|$, $e_h = |s_h - \tilde{s}|$, and $e_h^+ = |s_h^+ - \tilde{s}|$. The corresponding convergence rates are 1.22, 1.22, and 1.13, respectively. The convergence rates are consistent with the predictions of the Stokes bounds method.

Again we define the effectivity of the bounds consistent with [58], where the lower and upper effectivities are given by: $\Theta_h^- = |(\tilde{s} - s_h^-|)/(|\tilde{s} - s_h|)$ and $\Theta_h^+ = |(\tilde{s} - s_h^+|)/(|\tilde{s} - s_h|)$ respectively. These indicate the sharpness of the bounds by comparing the error in the bounds with the error in the finite element approximation. The effectivity of the bounds on different mesh sizes are given in Table (4.3) and range from 2.85-3.54. These values indicate that the bounds are sharp, being that they are only about three times the error of finite element solution This is reflected in Figure (4.8) (b), where the convergence of the exact

bounds, performs well compared to the convergence of the finite element output solution. We further point out that for the Stokes output, the effectivities show that the exact bounds perform even better than for the Poisson output. We also consider the predicted finite element output solution denoted by s_h^{pre}, which is defined as the average of the lower and upper bounds, i.e. $s_h^{pre} = s_{av} = (s_h^+ + s_h^-)/2$, and define the error effectivity $\tau_h = (\tilde{s} - s_h^{pre})/(\tilde{s} - s_h)$. This index indicates that the performance of the bound average is almost the same as the finite element approximation of the output. Moreover, by defining yet another effectivity index as: $\tau_h^{\pm} = (s_h^{\pm} - s_h)/(\tilde{s} - s_h)$ we can measure the quality of the error estimator $s_h^{\pm} - s_h$ as compared with the error $\tilde{s} - s_h$. It is seen that the effectivity of the lower bound is 0 which verifies that the finite element output value is equal to the lower bound. However the magnitude of effectivity of the upper bound τ_h^+ is close to 1.3 which indicates that the finite element output solution on each mesh is about half way between the upper bound and the exact output solution.

Chapter 5

Bounds for the Helmholtz Equation

In this chapter we present the formulation for obtaining rigorous upper and lower bounds for exact outputs of interest to the Helmholtz equation. The method resembles the approach undertaken for the Poisson equation in the sense that the bounds are obtained solely on computations done on local spaces where the aggregate of several independent sub-problems lead to the upper and lower bounds. However, the approach here differs from the Poisson case, in that there does not exist an intrinsic minimization principle for the Helmholtz problem where the exact energy term can be bounded by a complementary energy functional. Nevertheless, the problem can be transformed into an equivalent constrained maximization principle that can be shown to guarantee computable bounds which preserve the bounding property. That is to say, it guarantees the bounds to hold for all levels of refinement on any polygonal domain with piecewise polynomial forcing.

For the particular case of the Helmholtz equation, we know from the discussion of Chapter 2, that the Helmholtz sesquilinear form loses ellipticity for higher wave numbers, and thus the principle of convex minimization generally cannot be applied. However, as the Helmholtz operator is H^1-coercive satisfying a Gårding inequality, then in the asymptotic regime where the number of degrees of freedom in the discrete model surpasses a critical

number, will the problem begin to have a positive definite structure and the bounding property of the method applies. We present our findings using high order nodal spectral element approximations in an attempt to reduce the pollution error and in obtaining more accurate bounds for higher wave numbers.

5.1 Problem Statement and Approximation Spaces

5.1.1 Governing Equation

In this section we continue with the exposition of the method as is applied to the Helmholtz equation. We are interested in bounding outputs which are functionals of the weak solution to the Helmholtz equation, and so we work with the Equation (2.6) in which we re-express the problem in the convenient form: find $u \in Z_D$ so that

$$b(u, v) = \ell^N(v) \quad \forall v \in Z \tag{5.1}$$

where the test space Z and the functions space Z_D have already been defined in (2.4)-(2.5). Consistent with our notation from Chapter 2, b designates indefinite forms, where here it is defined as

$$\begin{aligned} b(u, v) &= \int_\Omega \nabla u \cdot \nabla \overline{v}\, d\Omega - k^2 \int_\Omega u \overline{v}\, d\Omega, \text{ and} \\ \ell^N(v) &= \int_\Omega f \overline{v}\, d\Omega. \end{aligned}$$

5.1.2 Approximation Spaces

We use the structured cubic subdomains comprised of six tetrahedrons described in the previous chapter. As before, we have elemental subdomains T of a partioning $\mathcal{T}_h$ of Ω. Associated to this mesh we define a regular piecewise continuous finite element subspaces,

namely, the approximation subspaces,

$$Z_h(T) = \left\{v = v^R + iv^I : v^R|_T \in \mathbb{P}^p(T), v^I|_T \in \mathbb{P}^p(T), \forall T \in \mathcal{T}_h\right\} \cap Z, \tag{5.2}$$

$$Z_h^D(T) = \left\{v = v^R + iv^I : v^R|_T \in \mathbb{P}^p(T), v^I|_T \in \mathbb{P}^p(T), \forall T \in \mathcal{T}_h\right\} \cap Z_D, \tag{5.3}$$

where $\mathbb{P}^p(T)$ denotes the space of polynomials of order p over T, and v^R, v^I representing the real and imaginary parts of v respectively. In order to apply the method of bounds, additional spaces and sesquilinear forms must be defined. First, the global representation of $Z_h(T)$ are defined, which are the broken spaces with respect to the domain decomposition given as

$$\widehat{Z}_h = \left\{v|_T \in \widetilde{Z}(T), \forall T \in \mathcal{T}_h : v^R|_T \in \mathbb{P}^p(T), v^I|_T \in \mathbb{P}^p(T), \forall T \in \mathcal{T}_h\right\}. \tag{5.4}$$

where $\widetilde{Z}$ is defined in (2.3). Secondly, we designate $\mathsf{E}(\mathcal{T}_h)$ to be the edge space which consists of the set of open faces and edges γ between different subdomains of the partitioning $\mathcal{T}_h$, and we introduce the space defined over $\gamma \in \mathsf{E}(\mathcal{T}_h)$ by

$$\Lambda_h = \left\{\lambda = \lambda^R + i\lambda^I : \lambda^R|_\gamma \in \mathbb{P}^p(\gamma), \lambda^I|_\gamma \in \mathbb{P}^p(\gamma), \forall \gamma \in \mathsf{E}(\mathcal{T}_h)\right\}. \tag{5.5}$$

Finally, we define the continuity sesquilinear form $\mathsf{h} : \widehat{Z}_h \times \Lambda_h \mapsto \mathbb{C}$ formally as

$$\mathsf{h}(v, \lambda) = \sum_{\gamma \in \mathsf{E}(\mathcal{T}_h)} \int_\gamma [\overline{v}]_\gamma \lambda|_\gamma \, d\Gamma, \tag{5.6}$$

where $[v]_\gamma$ is the jump in v across γ when γ is in the interior face, and the trace of v on γ when γ is on the boundary Γ. The relation between the spaces Z_h and $\widehat{Z}_h$ become eminent when we exploit the form $\mathsf{h}(.,.)$ in order to enforce continuity of the hat functions as

$$Z_h = \left\{v \in \widehat{Z}_h : \mathsf{h}(v, \lambda) = 0, \forall \lambda \in \Lambda_h\right\}. \tag{5.7}$$

Conversely, for a continuous test function v satisfying the homogeneous essential boundary conditions, and any function $\tilde{\lambda} \in \Lambda_h$, $\mathsf{h}(v, \tilde{\lambda}) = 0$.

5.2 Error Bound Formulation

5.2.1 Output Functional

The goal here is to provide upper and lower estimators that provide upper and lower bounds to some exact output $S(u)$. As previously mentioned, we do this based solely on decoupled calculations on the broken spaces $\widehat{Z}_h$. As we are interested in real outputs s, that are functionals of the solution $u = u^R + iu^I$, we set $s = \Re(S(u))$, where $S(u) : \widetilde{Z}(\Omega) \mapsto \mathbb{C}$. In our numerical examples we present the bounds on outputs which are expressed as linear functionals of the exact solution u, although one can formulate the problem with non-linear outputs. For example, we can express our output as in [57] in the form

$$S(u_h + v) = S(u_h) + \ell^\circ(v) + \mathcal{M}(v, v)$$

where we have expanded a non-linear functional in a form which considers both linear and quadratic outputs. Here $\ell^\circ : \tilde{Z}(\Omega) \mapsto \mathbb{C}$ and $\mathcal{M}(w, w) : \tilde{Z}(\Omega) \times \tilde{Z}(\Omega) \mapsto \mathbb{C}$ are the linear an bilinear contributions to the output respectively. $\mathcal{M}$ is required to be L^2-continuous, with the requirement that $|\mathcal{M}(v, v)| \leq C\|v\|_{L^2}^2$. In this work we take $\mathcal{M} = 0$ and consider

$$S(u) = S(u_h + v) = S(u_h) + \ell^\circ(v).$$

More preciesely, one can express the linear output functional as

$$S(u) = \ell^\circ(u) = \int_{\Omega^\circ} f^\circ u \, d\Omega = \int_{\Omega^\circ} f^\circ u_h \, d\Omega + \int_{\Omega^\circ} f^\circ v \, d\Omega \tag{5.8}$$

where f° is some prescribed forcing data mentioned in the previous chapter. As we will see, here v represents the error $e = u - u_h$.

5.2.2 Error Formulation

We begin by writing the Helmholtz equation as a minimization statement where the functional that is to be minimized is in terms of the error defined above as $e = u - u_h$, where

we recall that u_h is the spectral element approximation. Since $u - u_h \in Z$ then

$$b(u, u - u_h) - \ell^N(u - u_h) = 0 \tag{5.9}$$

which implies through the sesquilinearity of $b(.,.)$, that

$$b(u - u_h, u - u_h) - \ell^N(u - u_h) + b(u_h, u - u_h) = 0.$$

Now defining the residual

$$\ell^E(v) \equiv \ell^N(v) - b(u_h, v) \;\; \forall v \in \widehat{Z}$$

and the energy equality can be rewritten as

$$b(e, e) - \ell^E(e) = 0. \tag{5.10}$$

5.2.3 Lagrange Multiplier Approximation

Now we introduce the set of functions $\mathcal{S} \subset \widehat{Z}$ given by

$$\mathcal{S} = \left\{ v \in \widehat{Z} \middle| b(u_h + v, \omega) = \ell^N(\omega), \forall \omega \in Z;\, \mathsf{h}(v, q) = 0, \forall t \in \Lambda \right\} \tag{5.11}$$

where the first constraint enforces $v = e$, and the second enforces continuity and the homogeneous essential conditions. This suggest the formulation of the quadratic Lagrangian functional $\mathcal{L} : \widehat{Z} \times Z \times \Lambda \mapsto \mathbb{C}$ given by

$$\mathcal{L}^{\pm}(v, \mu, t) = \mp \ell^{\circ}(u_h + v) + \frac{\kappa}{2}\left(b(v, v) - \ell^E(v)\right) + \ell^N(\mu) - b(u_h + v, \mu) + \mathsf{h}(v, t) \tag{5.12}$$

for some $\mu \in Z_D$, and $q \in \Lambda$, where the constraints given by the equilibrium equation, and continuity enforcement between subdomains are included in the Lagrangian.

Modified Lagrangian

We know from the discussion of Chapter 3, that a subdomain boundary value problem may become ill-posed if it is equipped with the same boundary conditions as in the Poisson

case, namely Equations (3 4)-(3 5) Instead, we replace these boundary conditions with the Robin boundary conditions at the interface between two subdomains given by (3 27)-(3 28), keeping in mind that we must utilize the special checkerboard like signing of the mesh as explained in Chapter3 Therefore, under the broken spaces, where these boundary conditions are satisfied at the interfaces, the sesquilinear form in Equation (5 1) is replaced and the equation is re-expressed as find $u \in \widehat{Z}_D$ such that

$$b_m(u,v) = b(u,v) + \imath p(u,v) = \ell^N(v), \quad \forall v \in Z \tag{5 13}$$

where between the two subdomains $\gamma^{(s\ q)} = \Omega^s \cap \Omega^q$, $p(u,v)$ is defined as

$$p(u,v) = k \sum_{\gamma \in \mathsf{E}(\mathcal{T}_h)} \sum_{\gamma^{(s\ q)} \neq \emptyset} \int_{\gamma^{(s\ q)}} \left((-1)^{\delta_{\mathsf{n}_s\ q}} \overline{v} u - (-1)^{\delta_{\mathsf{n}_q\ s}} \overline{v} u \right) d\Gamma \tag{5 14}$$

Moreover, this expressed in terms of an error equation gives like before,

$$b_m(e,e) = \ell_m^E(e) \tag{5 15}$$

where

$$\ell_m^E(v) \equiv \ell^N(v) - b_m(u_h, v), \quad \forall v \in \widehat{Z}$$

Now defining the set $\mathcal{S}_m \subset \widehat{Z}$ given by

$$\mathcal{S}_m = \left\{ v \in \widehat{Z} \middle| b_m(u_h + v, \omega) = \ell^N(\omega), \forall \omega \in Z,\ \mathsf{h}(v,t) = 0, \forall t \in \Lambda \right\} \tag{5 16}$$

where again, the first constraint enforces $v = e$, the error, and the second enforces continuity of the subdomains at the interfaces Thus, the modified Lagrangian is now expressed as

$$\mathcal{L}_m^{\pm}(v,\mu,t) = \mp \ell^{\circ}(u_h + v) + \frac{\kappa}{2} \left(b_m(v,v) - \ell_m^E(v) \right) + \ell^N(\mu) - b_m(u_h + v, \mu) + \mathsf{h}(v,t) \tag{5 17}$$

where the gradient condition of this Lagrangian solves for $\psi_h^{\pm}, \lambda_h^{\pm}$ for the bounds Thus we have at optimality

$$\frac{\partial}{\partial v} \mathcal{L}_m^{\pm}(v,\mu,t)\Big|_{(\mu - \psi_h^{\pm}\ t = \lambda_h^{\pm})} = 0, \quad \frac{\partial}{\partial \mu} \mathcal{L}_m^{\pm}(v,\mu,t)\Big|_{v = e_h^{\pm}} = 0, \quad \frac{\partial}{\partial t} \mathcal{L}_m^{\pm}(v,\mu,t)\Big|_{v = e_h^{\pm}} = 0,$$

which lead to finding $e_h^{\pm} \in \widehat{Z}_h$, $\psi_h^{\pm} \in Z_h^D$ and $\lambda_h^{\pm} \in \Lambda_h$ respectively, such that the following set of equations are satisfied:

$$\mathsf{h}(\omega, \lambda_h^{\pm}) = -\left\{ \mp \ell^{\circ}(\omega) + \frac{\kappa}{2}\left(2b_m(e_h^{\pm}, \omega) - \ell_m^E(\omega)\right) - b_m(\omega, \psi_h^{\pm}) \right\}, \quad \forall \omega \in \widehat{Z}_h \tag{5.18}$$

$$b_m(u_h + e_h^{\pm}, \omega) = \ell^N(\omega), \quad \forall \omega \in \widehat{Z}_h \tag{5.19}$$

$$\mathsf{h}(e_h^{\pm}, t) = 0, \qquad \forall t \in \Lambda_h. \tag{5.20}$$

Equation (5.20) forces $e_h^{\pm} \in Z_h$, which combined with the original primal problem i.e $b(u_h, v) = \ell^N(v)$, $\forall v \in Z_h$, and Equation (5.19), implies $e_h^{\pm} = 0$. Moreover, Equation (5.18) must be satisfied for all $\omega \in \widehat{Z}_h$ and thus for all $\omega \in Z_h \subset \widehat{Z}_h$. For the restriction $\omega \in Z_h$, $\ell_m^E(\omega) = \ell^E(\omega) = 0$ (i.e. the residual $\ell^N(\omega) - b(u_h, \omega) = 0$ in the global mesh). Therefore, by exploiting the fact that in the continuous (global) mesh $\ell^E(\omega) = 0$, and $\mathsf{h}(\ ,\ ) = 0$, yields the adjoint equation

$$b(\omega, \psi_h) = -\ell^{\circ}(\omega), \text{ or equivalently } \overline{b(\omega, \psi_h)} = -\overline{\ell^{\circ}(\omega)} \quad \forall \omega \in Z_h \tag{5.21}$$

with $\psi_h^{\pm} = \pm\psi_h$.

Remark 5.1 *We note that in the continuous global mesh, the modified term used in order to make each local problem well-posed becomes zero since the mesh is no longer broken between subdomains, and the $\pm$ contributions of the Robin boundary conditions cancel out.*

Under the broken spaces, the residual $R^{\psi}(\omega) = -\overline{\ell^{\circ}(\omega)} - \overline{b_m(\omega, \psi_h)} \neq 0$, in fact from (5.18) it is evident that the equilibration equation can be expressed as

$$\mathsf{h}(\omega, \lambda_h^{\pm}) = \frac{\kappa}{2}\ell_m^E(\omega) \mp \overline{\ell^{\circ}(\omega)} - \overline{b_m(\omega, \psi_h)} \quad \forall \omega \in \widehat{Z}_h. \tag{5.22}$$

Now as before, we make the decompositions

$$\lambda_h^{\pm} = \frac{\kappa}{2}\lambda_h^u \pm \lambda_h^{\psi} \quad \text{and recall} \quad \psi_h^{\pm} = \pm\psi_h,$$

which transforms the equilibration equation into two κ independent equations. Therefore, the solution to u_h, ψ_h, and λ_h which are necessary data for obtaining bounds, can be arrived at through the following set of equations

Step 1: (The original primal problem) Find $u_h \in Z_h^D$ such that

$$b(u_h, v) = \ell^N(v) \quad \forall v \in Z_h$$

Step 2: (The adjoint problem) Find $\psi_h \in Z_h^D$ such that

$$b(v, \psi_h) = -\ell^\circ(v) \quad \forall v \in Z_h$$

Step 3: (Equilibration equation) Find $\lambda_h^u \in \Lambda_h$ such that

$$\mathsf{h}(v, \lambda_h^u) = \ell^N(v) - b_m(u_h, v) \quad \forall v \in \widehat{Z}_h \tag{5.23}$$

Step 4: (Equilibration equation) Find $\lambda_h^\psi \in \Lambda_h$ such that

$$\mathsf{h}(v, \lambda_h^\psi) = -\overline{\ell^\circ(v)} - \overline{b_m(v, \psi_h)} \quad \forall v \in \widehat{Z}_h. \tag{5.24}$$

In the current work we do not solve the two global problems in **Step 1** and **Step 2**. The approach in obtaining the hybrid fluxes and the decoupled solutions is the same as in the FETI iterative procedure used for the Poisson problem in Chapter 4. The GCR method used in solving equations (5.23)-(5.24) in **Step 3** and **Step 4** gives the hybrid flux approximations, and ultimately, the approximation to the decoupled solutions $u^{(s)}$, and $\psi^{(s)}$ which then in turn leads to the corresponding global solutions by the aggregate of these quantities. We recall that the discrete forms of these equations is given in (3.31)-(3.32), which is consistent with the way the FETI-H procedure solves the interface problem with the inclusion of the regularizing term.

5.2.4 Local Problems

The Lagrange multiplier approximations, are achieved through the first variation of the modified Lagrangian, which is used to replace (5.12) in order to avoid singular solution in the system matrix. However, after obtaining these quantities, we pursue with the presentation of the bounds using the original Lagrangian (5.12), as the extra term does not affect the global solution nor the bounding property. It is premature to discuss minimization principle for the Lagrangian just described, as we are dealing with complex spaces, and minimum (or maximum) have no meaning here. We point out that although are numerical tests and validation are only in the real space, we have implemented our code as complex not only due to the additional regularizing term, but also to allow any future potential applications which may involve imaginary boundary conditions and complex forcing. Therefore, we present the method in the complex space setting. Here we only point out that as the problem involves bounding exact outputs, it is necessary that the Lagrangian so constructed will be such that the energy terms which are expressed in terms of the error will vanish in some limiting case as the exact error goes to zero. Consequently, in such a limit we approach the exact output from above or below which are the upper and lower bounds. In the previous subsection, the stationarity conditions of the Lagrangian $\mathcal{L}_m$, are restricted to the subspaces $\widehat{Z}_h \subset \widehat{Z}$, $\widehat{Z}_h^D \subset \widehat{Z}^D$and $\Lambda_h \subset \Lambda$, in which Equation (5.19) and (5.20) indicate that the continuity between the subdomains resulted in $e_h = 0$ and hence $b_m(e_h^{\pm}, \omega) = 0$. Now we consider the Lagrangian $\mathcal{L}^{\pm}(v, \psi_h^{\pm}, \lambda_h^{\pm})$, where the error $\hat{e} = \hat{e}^R + i\hat{e}^I$ will satisfy the gradient condition (see Equation (5.18))

$$\kappa b(\hat{e}^{\pm}, \omega) = \pm \ell^{\circ}(\omega) + \frac{\kappa}{2}\ell^E(\omega) + b(\omega, \psi_h^{\pm}) - \mathsf{h}(\omega, \lambda_h^{\pm}), \tag{5.25}$$

which corresponds to a number of local symmetric Neumann (or Robin) subdomain problems. By taking $\omega = \hat{e} \in \widehat{Z}$ we have

$$\kappa b(\hat{e}^{\pm}, \hat{e}^{\pm}) = \frac{\kappa}{2}\ell^N(\hat{e}^{\pm}) - \frac{\kappa}{2}b(u_h, \hat{e}^{\pm}) \pm \ell^\circ(\hat{e}^{\pm}) + b(\hat{e}^{\pm}, \psi_h^{\pm}) - \mathsf{h}(\hat{e}^{\pm}, \lambda_h^{\pm}),$$

and substituting this expression into the Lagrangian and invoking the primal problem with $v = \psi_h \in Z_h \subset Z_h^D (\Rightarrow b(u_h, \psi_h^{\pm}) = \ell^N(\psi_h^{\pm}))$ we have

$$\begin{aligned}
\mathcal{L}^{\pm}(e^{\pm}, \psi_h^{\pm}, \lambda_h^{\pm}) &= \frac{\kappa}{2}b(\hat{e}^{\pm}, \hat{e}^{\pm}) - \frac{\kappa}{2}\ell^N(\hat{e}^{\pm}) + \frac{\kappa}{2}b(u_h, \hat{e}^{\pm}) \mp \ell^\circ(u_h) \pm \ell^\circ(\hat{e}^{\pm}) \\
&+ \ell^N(\psi_h^{\pm}) - b(u_h, \psi_h^{\pm}) - b(\hat{e}^{\pm}, \psi_h^{\pm}) + (\hat{e}^{\pm}, \lambda_h^{\pm}) \\
&= -\frac{\kappa}{2}b(\hat{e}^{\pm}, \hat{e}^{\pm}) \mp \ell^\circ(u_h).
\end{aligned}$$

Now it is apparent that one simply obtains from the above that

$$s^+ = -\mathcal{L}^+(e^+, \psi_h^+, \lambda_h^+) = \Re\{\ell^\circ(u_h)\} + \frac{\kappa}{2}b(\hat{e}^+, \hat{e}^+) \tag{5.26}$$

$$s^- = \mathcal{L}^-(e^-, \psi_h^-, \lambda_h^-) = \Re\{\ell^\circ(u_h)\} - \frac{\kappa}{2}b(\hat{e}^-, \hat{e}^-), \tag{5.27}$$

which can be shown to be asymptotic upper and lower bounds respectively for the exact output $s = \Re\{\ell^\circ(u)\}$. For example, consider the gradient condition

$$\kappa b(\hat{e}^-, v) - \frac{\kappa}{2}\ell^N(v) + \frac{\kappa}{2}b(u_h, v) + \overline{\ell^\circ(v)} + \overline{b(v, \psi_h^{\pm})} + \mathsf{h}(v, \lambda_h^-) = 0 \quad \forall v \in \widehat{Z},$$

and impose the restriction that $v \in Z \subset \widehat{Z}$, in particular $v = e$. Then since in this case $\mathsf{h}(e, \lambda_h^-) = 0$, we have

$$\kappa b(\hat{e}, e) - \frac{\kappa}{2}\ell^N(e) + \frac{\kappa}{2}b(u_h, e) + \overline{\ell^\circ(e)} + \overline{b(e, \psi_h^{\pm})} = 0$$

or by the fact that $u_h = u - e$, gives

$$\kappa b(\hat{e}, e) - \frac{\kappa}{2}\ell^N(e) + \frac{\kappa}{2}b(u, e) - \frac{\kappa}{2}b(e, e) + \overline{\ell^\circ(e)} + \overline{b(e, \psi_h^{\pm})} = 0$$

which then simplifies to

$$\kappa b(\hat{e}^-, e) - \frac{\kappa}{2} b(e,e) + \overline{\ell^\circ(e)} = 0$$

after invoking $a(u,e) = \ell^N(e)$. This is the same as

$$\overline{\kappa b(\hat{e}^-, e)} - \frac{\kappa}{2} b(e,e) + \ell^\circ(e) = 0,$$

and in particular

$$\Re\{\overline{\kappa b(\hat{e}^-, e)} - \frac{\kappa}{2} b(e,e) + \ell^\circ(e) = 0\},$$

or

$$\kappa \Re\{\overline{b(\hat{e}^-, e)}\} - \frac{\kappa}{2} b(e,e) + \ell^\circ(e) = 0.$$

Adding this to (5.27), noting that $u = u_h + e$ and that

$$\Re\{b(\hat{e}^-, e)\} = \overline{b(\hat{e}^-, e)} + b(\hat{e}^-, e) = b(e, \hat{e}^-) + b(\hat{e}^-, e),$$

we have

$$\begin{aligned}
s^- &= \Re\{\ell^\circ(u)\} - \frac{\kappa}{2} b(\hat{e}^-, \hat{e}^-) - \frac{\kappa}{2} b(e,e) + \kappa \Re\{b(\hat{e}^-, e)\} \\
&= \Re\{\ell^\circ(u)\} - \frac{\kappa}{2} b(\hat{e}^-, \hat{e}^-) - \frac{\kappa}{2} b(e,e) + \frac{\kappa}{2} b(e, \hat{e}^-) + \frac{\kappa}{2} b(\hat{e}^-, e) \\
&= \Re\{\ell^\circ(u)\} - \frac{\kappa}{2} b(\hat{e}^- - e, \hat{e}^-) + \frac{\kappa}{2} b(\hat{e}^-, e) - \frac{\kappa}{2} b(e,e) \\
&= \Re\{\ell^\circ(u)\} - \frac{\kappa}{2} b(\hat{e}^- - e, \hat{e}^- - e).
\end{aligned}$$

The same approach will also lead to a similar expression for the upper bound. Finally,

$$s^+ = \Re\{\ell^\circ(u)\} + \frac{\kappa}{2} b(\hat{e}^+ - e, \hat{e}^+ - e) \tag{5.28}$$

$$s^- = \Re\{\ell^\circ(u)\} - \frac{\kappa}{2} b(\hat{e}^- - e, \hat{e}^- - e) \tag{5.29}$$

and asymptotically approach the exact solution when the error $e \to 0$. Therefore, we see that while the bounds (5.26) and (5.27) are asymptotic bounds to the exact solution,

they are un-computable as it requires the exact error. We will now formulate the exact bounds method for the Helmholtz problem based on the work presented in [59] for the 2D advection-diffusion-reaction equation; however, we generalize the approach to the complex space setting in the same spirit as in [57]. In real spaces, duality theory of convex minimization is the standard proof for the bounds method. As in the case of obtaining exact bounds for equations where the energy terms does not have an intrinsic minimization principle, the ideas of duality theory are applied to transform an unconstrained minimization statement to a constrained maximization problem (dual problem) where any approximation of the dual variables satisfying the optimality conditions can be used to provide non-trivial bounds on the Helmholtz energy term.

5.3 Localized Lagrangian

Now we write the local contributions of the unconstrained Lagrangian on an arbitrary element T of the partitioning $\mathcal{T}_h$ as

$$\mathcal{L}_T^{\pm}(\omega, \psi_h^{\pm}, \lambda^{\pm}) = \frac{\kappa}{2} b_T(\omega, \omega) + \ell_T^{\pm}(\omega) + C_T^{\pm} \tag{5.30}$$

where

$$\begin{aligned} \ell_T^{\pm}(v) &= \frac{\kappa}{2}\left\{-\ell_T^N(v) + b_T(u_h, v) + \mathsf{h}_T(v, \lambda_h^u)\right\} \\ &\pm \left\{-\overline{\ell^\circ(v)} - \overline{b_T(v, \psi_h)} - \mathsf{h}(v, \lambda_h^\psi)\right\} \end{aligned}$$

and

$$C_T^{\pm} = \mp \ell^\circ(u_h) + \ell_T^N(\psi_h^{\pm}) - b_T(u_h, \psi_h^{\pm}).$$

Here, the definitions of the sesquilinear form $b_T(\omega, v)$, the anti linear functional $\ell_T^N(v)$, and the boundary flux term $\mathsf{h}_T(v, \lambda)$ over each elemental subdomain are defined as

$$b_T(\omega, v) = \int_T (\nabla w \quad \nabla \overline{v} - k^2 w \overline{v})\, d\Omega \tag{5 31}$$

$$\ell_T^N(v) = \int_T f \overline{v}\, d\Omega \tag{5 32}$$

$$\mathsf{h}_T(\omega, \lambda) = \int_{\partial T} \sigma_T \overline{\omega} \lambda\, d\Gamma \tag{5 33}$$

with the special signing between neighboring faces σ_T as defined in Chapter 3

5.4 An Equivalent Lagrangian Formulation

The local unconstrained problem is un-computable in general since it must be performed over an infinite-dimensional space in order to guarantee the bounding property However, we can transform the problem to an equivalent constrained problem which effectively dualizes the minimization of the local energy term $b_T(v, v)$ to an equivalent maximization problem with equality constraints We do this by first introducing two auxiliary variables $\Pi^{\pm} \in (L^2(T))^3$, and $\rho^{\pm} \in L^2(T)$ satisfying [1]

$$\kappa \int_T (\nabla \omega^{\pm} - \Pi^{\pm}) \quad \overline{\mathbf{p}}\, d\Omega = 0, \qquad \forall \mathbf{p} \in (L^2(T))^3 \tag{5 34}$$

$$\kappa k^2 \int_T (\rho^{\pm} - \omega^{\pm}) \overline{v}\, d\Omega = 0, \qquad \forall v \in L^2(T) \tag{5 35}$$

Define Ξ to be the set of triples $(\omega^{\pm}, \Pi^{\pm}, \rho^{\pm})$ in $H^1(T) \times (L^2(T))^3 \times L^2(T)$ that satisfy the constraints (5 34)-(5 35) Then we can write the unconstrained Lagrangian (5 30) to the equivalent constrained problem by introducing the functional $\mathcal{J}_T^{\pm}(\omega^{\pm}, \Pi^{\pm}, \rho^{\pm}, \mathbf{q}^{\pm}, r^{\pm})$ given

[1] Here $L^2(T)$ is defined as the space of square integrable complex-valued functions $v \quad T \mapsto \mathbb{C}$ with the usual inner product for complex spaces The general $H^m(T)$ spaces are also defined analogously in complex spaces

by

$$\begin{aligned}\mathcal{J}_T^{\pm}(\omega^{\pm},\Pi^{\pm},\rho^{\pm},\mathbf{q}^{\pm},r^{\pm}) &= \frac{\kappa}{2}\left(\int_T(\Pi^{\pm})\cdot(\overline{\Pi}^{\pm})\,d\Omega - k^2\int_T\rho^{\pm}\overline{\rho}^{\pm}\,d\Omega\right)\\ &+ \kappa\left(\int_T\nabla\omega\cdot\overline{\mathbf{q}}^{\pm}\,d\Omega - \int_T\Pi^{\pm}\cdot\overline{\mathbf{q}}^{\pm}\,d\Omega\right)\\ &- \kappa k^2\left(\int_T\omega^{\pm}\overline{r}^{\pm}\,d\Omega - \int_T\rho^{\pm}\overline{r}^{\pm}\,d\Omega\right) + \ell_T^{\pm}(\omega^{\pm}) + C_T^{\pm}\end{aligned}$$

which reduces to the expression

$$\begin{aligned}\mathcal{J}_T^{\pm}(\omega^{\pm},\Pi^{\pm},\rho^{\pm},\mathbf{q}^{\pm},r^{\pm}) = \kappa&\left\{\int_T\left(\frac{1}{2}\Pi^{\pm}-\mathbf{q}^{\pm}\right)\cdot\overline{\Pi}^{\pm} - k^2\left(\frac{1}{2}\rho^{\pm}-r^{\pm}\right)\overline{\rho}^{\pm}\,d\Omega\right\}\\ &+\kappa\left\{\int_T\nabla\omega^{\pm}\cdot\overline{\mathbf{q}}^{\pm} - k^2\omega^{\pm}\overline{r}^{\pm}\,d\Omega\right\} + \ell_T^{\pm}(\omega^{\pm}) + C_T^{\pm}.\end{aligned} \tag{5.36}$$

The gradient condition with respect to the variables $(\omega^{\pm},\Pi^{\pm},\rho^{\pm})$ are:

$$\frac{\partial\mathcal{J}_T^{\pm}}{\partial\omega^{\pm}} = 0, \qquad \frac{\partial\mathcal{J}_T^{\pm}}{\partial\Pi^{\pm}} = 0, \qquad \frac{\partial\mathcal{J}_T^{\pm}}{\partial\rho^{\pm}} = 0$$

and are

$$\kappa\int_T(\nabla\overline{v}\cdot\mathbf{q}^{\pm *} - k^2\overline{v}r^{\pm *})\,d\Omega = -\ell_T^{\pm}(v) \quad \forall v\in H^1(T) \tag{5.37}$$

$$\kappa\int_T(\Pi^{\pm *}-\mathbf{q}^{\pm *})\cdot\overline{\mathbf{p}}\,d\Omega = \quad \forall\mathbf{p}\in(L^2(T))^3 \tag{5.38}$$

$$\kappa k^2\int_T(\rho^{\pm *}-r^{\pm *})\overline{v}\,d\Omega = 0 \quad \forall v\in L^2(T) \tag{5.39}$$

respectively. Here, the superscript (*) indicates the value corresponding to the vanishing of the gradient. Substituting (5.38) and (5.39) into the functional $\mathcal{J}_T^{\pm}(\omega^{\pm},\Pi^{\pm},\rho^{\pm},\mathbf{q}^{\pm},r^{\pm})$ above, we have

$$\mathcal{J}_{\mathcal{T}}{}^{\pm}(\mathbf{q}^{\pm},r^{\pm}) = -\frac{\kappa}{2}\left\{\int_T\mathbf{q}^{\pm}\cdot\overline{\mathbf{q}}^{\pm}\,d\Omega - k^2\int_T r^{\pm}\overline{r}^{\pm}\,d\Omega\right\} + C_T^{\pm}$$

Subject to the constraint in Equation (5.37)

As the objective function is a real quantity, this will lead to a maximization problem where in fact it can be shown that

$$\begin{aligned} \mp s_T^{\pm} &= \sup_{\substack{\mathbf{q}^{\pm}\in(L^2(T))^3 \\ r^{\pm}\in L^2(T)}} -\frac{\kappa}{2} a_T^{du}((\mathbf{q}^{\pm}, r^{\pm}), (\mathbf{q}^{\pm}, r^{\pm})) + \Re\{C_T^{\pm}\} \\ &\text{s.t.} \quad \kappa c_T^{du}((\mathbf{q}^{\pm}, r^{\pm}), v) = -\ell_T^{\pm}(v) \qquad \forall v \in H^1(T), \end{aligned} \tag{5.40}$$

where the form $a_T^{du} : (L^2(T))^3 \times H^1(T) \times (L^2(T))^3 \times H^1(T) \mapsto \mathbb{C}$ is defined as

$$a_T^{du}((\mathbf{q}, r), (\mathbf{p}, v)) = \int_T \mathbf{q} \cdot \overline{\mathbf{p}}\, d\Omega - k^2 \int_T r\overline{v}\, d\Omega$$

and the form $c_T^{du} : (L^2(T))^3 \times H^1(T) \times H^1(T) \to \mathbb{C}$ is defined as

$$c_T^{du}((\mathbf{q}, r), v) = \int_T \mathbf{q} \cdot \nabla\overline{v}\, d\Omega - k^2 \int_T r\overline{v}\, d\Omega.$$

5.5 Elemental Sub-problems

The gradient conditions corresponding to the maximization problem (5.40) can now be obtained in the usual way of constructing the Lagrangian $L_T^{du} : (L^2(T))^3 \times (L^2(T)) \times H^1(T)$ as

$$\begin{aligned} \mathcal{L}_T^{du}(\mathbf{q}^{\pm}, r^{\pm}; \xi) \equiv & -\frac{\kappa}{2}\left\{\int \mathbf{q}^{\pm} \cdot \overline{\mathbf{q}}^{\pm}\, d\Omega - k^2 \int_T r^{\pm}\overline{r}^{\pm}\, d\Omega\right\} \\ & -\kappa\left\{\int_T \mathbf{q}^{\pm} \cdot \nabla\overline{\xi}\, d\Omega - k^2 \int_T r^{\pm}\overline{\xi}\, d\Omega\right\} - \ell_T^{\pm}(\xi) + C_T^{\pm}. \end{aligned}$$

The gradient conditions

$$\frac{\delta\mathcal{L}_T^{du}}{\delta\mathbf{q}^{\pm}} = 0, \qquad \frac{\delta\mathcal{L}_T^{du}}{\delta r^{\pm}} = 0, \qquad \frac{\delta\mathcal{L}_T^{du}}{\delta\xi} = 0,$$

now reveals the set of equations

$$-\kappa \int_T \mathbf{q}^{\pm} \cdot \overline{\mathbf{p}}\, d\Omega - \kappa \int_T \nabla\xi \cdot \overline{\mathbf{p}}\, d\Omega = 0 \qquad \forall \mathbf{p} \in (L^2(T))^3 \tag{5.41}$$

$$\kappa k^2 \int_T r^{\pm}\overline{v}\, d\Omega + \kappa k^2 \int_T \xi\overline{v}\, d\Omega = 0 \qquad \forall v \in H^1(T) \tag{5.42}$$

$$-\kappa\left\{\int_T \mathbf{q}^{\pm} \cdot \nabla\overline{v}\, d\Omega - k^2 \int_T r^{\pm}\overline{v}\, d\Omega\right\} = \ell_T^{\pm}(v) \qquad \forall v \in H^1(T) \tag{5.43}$$

in which upon adding (5.41) and (5.42) we express these equations in the compact form,

$$\kappa a_T^{du}((\mathbf{q}^{\pm}, r^{\pm}), (\mathbf{p}, v)) + \kappa c_T^{du}((\mathbf{p}, v), \xi) = 0 \qquad \forall (\mathbf{p}, v) \in (L^2(T))^3 \times H^1(T) \quad (5.44)$$

$$-\kappa c_T^{du}((\mathbf{q}^{\pm}, r^{\pm}), v) = \ell_T^{\pm}(v) \qquad \forall v \in H^1(T). \quad (5.45)$$

We make the decomposition

$$(\mathbf{q}^{\pm}, r^{\pm}, \xi^{\pm}) = \left(-\frac{1}{2}\mathbf{q}^u \pm \frac{1}{\kappa}\mathbf{q}^{\psi}, -\frac{1}{2}r^u \pm \frac{1}{\kappa}r^{\psi}, -\frac{1}{2}\xi^u \pm \frac{1}{\kappa}\xi^{\psi}\right)$$

which we then substitute into (5.44) and (5.45) to obtain two κ- independent problems:

1. Find $(\mathbf{q}^u, r^u, \xi^u) \in (L^2(T))^3 \times H^1(T) \times H^1(T)$ such that

$$a_T^{du}((\mathbf{q}^u, r^u), (\mathbf{p}, v)) + c_T^{du}((\mathbf{p}, v), \xi^u) = 0 \qquad \forall (\mathbf{p}, v) \in (L^2(T))^3 \times H^1(T)$$

$$c_T^{du}((\mathbf{q}^u, r^u), v) = \mathcal{R}_T^u(v) \qquad \forall v \in H^1(T)$$

2. Find $(\mathbf{q}^{\psi}, r^{\psi}, \xi^{\psi}) \in (L^2(T))^3 \times H^1(T) \times H^1(T)$ such that

$$a_T^{du}((\mathbf{q}^{\psi}, r^{\psi}), (\mathbf{p}, v)) + c_T^{du}((\mathbf{p}, v), \xi^{\psi}) = 0 \qquad \forall (\mathbf{p}, v) \in (L^2(T))^3 \times H^1(T)$$

$$c_T^{du}((\mathbf{q}^{\psi}, r^{\psi}), v) = \mathcal{R}_T^{\psi}(v) \qquad \forall v \in H^1(T),$$

where we have defined the residuals $\mathcal{R}^u(v)$ and $\mathcal{R}^{\psi}(v)$ as:

$$\mathcal{R}_T^u(v) = -\ell_T^N(v) + b_T(u_h, v) + \mathsf{h}_T(v, \lambda_h^u) \quad (5.46)$$

$$\mathcal{R}_T^{\psi}(v) = -\overline{\ell_T^{\circ}(v)} - \overline{b_T(v, \psi_h)} - \mathsf{h}_T(v, \lambda_h^{\psi}). \quad (5.47)$$

We notice that for $v \in Z_h$, the residuals will be zero from the equilibration equations given in Equations (5.23) and (5.24), and consequently,

$$a_T^{du}((\mathbf{q}^u, r^u), (\nabla v, v)) = c_T^{du}((\mathbf{q}^u, r^u), v) = 0 \qquad \forall v \in Z_h(T),$$

$$a_T^{du}((\mathbf{q}^{\psi}, r^{\psi}, (\nabla v, v)) = c_T^{du}((\mathbf{q}^{\psi}, r^{\psi}), v) = 0 \qquad \forall v \in Z_h(T).$$

Invoking the decomposition into (5.40) now leads to the local output bounds

$$\mp s_T^{\pm} = \Re\{C_T^{\pm}\} \tag{5.48}$$
$$-\frac{\kappa}{8}a_T^{du}((\mathbf{q}^u, r^u), (\mathbf{q}^u, r^u)) - \frac{1}{2\kappa}a_T^{du}((\mathbf{q}^\psi, r^\psi), (\mathbf{q}^\psi, r^\psi)) \pm \frac{1}{2}a_T^{du}((\mathbf{q}^u, r^u), (\mathbf{q}^\psi, r^\psi)).$$

5.6 Bounding Property

To show that the constrained maximization problem possesses the bounding property, we first consider the constraint equation in the form (5.40) and write

$$\begin{aligned} \kappa c_T^{du}((\mathbf{q}^{\pm}, r^{\pm}), v) \quad &+ \quad \frac{\kappa}{2}\{-\ell_T^N(v) + b_T(u_h, v) + \mathsf{h}_T(v, \lambda_h^u)\} \\ &\pm \quad \{-\overline{\ell^\circ(v)} - \overline{b_T(v, \psi_h)} - \mathsf{h}_T(v, \lambda_h^\psi)\} = 0. \end{aligned}$$

Accumulating the local contributions and designating $a^{du} = \sum_{T\in\mathcal{T}_h} a_T^{du}$ and $c^{du} = \sum_{T\in\mathcal{T}_h} c_T^{du}$, and once again use the diacritical hat to indicate that the function is broken across the element, we write the constraint equation on the broken domain as

$$\begin{aligned} \kappa c^{du}((\widehat{\mathbf{q}}^{\pm}, \widehat{r}^{\pm}), \widehat{v}) \quad &+ \quad \frac{\kappa}{2}\{-\ell^N(\widehat{v}) + b(u_h, \widehat{v}) + \mathsf{h}(\widehat{v}, \lambda_h^u)\} \\ &\pm \quad \{-\overline{\ell^\circ(\widehat{v})} - \overline{b(\widehat{v}, \psi_h)} - \mathsf{h}(v, \lambda_h^\psi)\} = 0. \end{aligned}$$

Recall $\widehat{v} \in \widehat{Z}$, and $Z \subset \widehat{Z}$. Restricting $\widehat{v} = e \in Z$, and employing the relations

$$\begin{aligned} \mathsf{h}(e, \lambda_h^u) &= \mathsf{h}(e, \lambda_h^\psi) = 0 \\ b(u_h, e) &= b(u, e) - b(e, e) \quad \text{using } u = u_h + e \\ b(u, e) &= \ell^N(e) \quad \text{for } e \in Z \text{ (primal problem)} \\ b(e, \psi_h) &= 0 \quad \text{for } e \in Z \text{ (from orthogonality of the error)} \end{aligned}$$

reduces the constraint to

$$\kappa c^{du}((\widehat{\mathbf{q}}^{\pm}, \widehat{r}^{\pm}), e) - \frac{\kappa}{2}b(e, e) \mp \{\overline{\ell^\circ(e)}\} = 0$$

or

$$\kappa\overline{c^{du}((\widehat{\mathbf{q}}^{\pm},\widehat{r}^{\pm}),e)} - \frac{\kappa}{2}b(e,e) \mp \ell^{\circ}(e) = 0$$

which will also satisfy,

$$\Re\{\kappa\overline{c^{du}((\widehat{\mathbf{q}}^{\pm},\widehat{r}^{\pm}),e)} - \frac{\kappa}{2}b(e,e) \mp \ell^{\circ}(e)\} = 0.$$

Again invoking the relationship

$$\Re\{c^{du}((\widehat{\mathbf{q}}^{\pm},\widehat{r}^{\pm}),e)\} = \frac{1}{2}\left[\overline{c^{du}((\widehat{\mathbf{q}}^{\pm},\widehat{r}^{\pm}),e)} + c^{du}((\widehat{\mathbf{q}}^{\pm},\widehat{r}^{\pm}),e)\right]$$

and employing the gradient conditions (5.44) and (5.45) where the values at this point are given by $(\mathbf{q}_T^{\pm,*}, r_T^{\pm,*}) = (\nabla\phi_T^{\pm}, \phi_T^{\pm})$, we have that $c_T^{du}((\nabla\phi,\phi),e) = b_T(\phi,e)$ and hence,

$$\kappa\Re\{c^{du}((\widehat{\mathbf{q}}^{\pm,*},\widehat{r}^{\pm,*}),e)\} = \frac{\kappa}{2}[b(e,\widehat{\phi}) + b(\phi,e)].$$

Here we have in addition used the fact that $\overline{b(\widehat{\phi},e)} = b(e,\widehat{\phi})$. Thus, the constraint equation reduces to

$$\begin{aligned} & \frac{\kappa}{2}b(e,\widehat{\phi}^{\pm}) + \frac{\kappa}{2}b(\widehat{\phi}^{\pm},e) - \frac{\kappa}{2}b(e,e) \mp \Re\{\ell^{\circ}(e)\} \\ = \; & \frac{\kappa}{2}b(e,\widehat{\phi}^{\pm} - e) + \frac{\kappa}{2}b(\widehat{\phi}^{\pm},e) \mp \Re\{\ell^{\circ}(e)\} = 0 \end{aligned} \tag{5.49}$$

from the sesquilinearity of the operator $b(.,.)$. Accumulating the contributions of (5.40) over all subdomains, where we have $s^{\pm} = \sum_{T\in\mathcal{T}_h} s_T^{\pm}$ we can now write

$$\mp s^{\pm} = -\frac{\kappa}{2}b(\widehat{\phi}^{\pm},\widehat{\phi}^{\pm}) \mp \Re\{\ell^{\circ}(u_h)\}, \tag{5.50}$$

where it is to be noted that $a^{du}((\nabla\widehat{\phi},\phi),(\nabla\widehat{\phi},\phi)) = b(\widehat{\phi},\widehat{\phi})$. Summing Equation (5.49) with (5.50) we obtain

$$\begin{aligned} \mp s^{\pm} &= -\frac{\kappa}{2}b(e,\widehat{\phi}^{\pm} - e) + \frac{\kappa}{2}b(\widehat{\phi}^{\pm},e) - \frac{\kappa}{2}b(\widehat{\phi}^{\pm},\widehat{\phi}^{\pm}) \mp \Re\{\ell^{\circ}(u)\} \\ &= \frac{\kappa}{2}b(e,\widehat{\phi}^{\pm} - e) + \frac{\kappa}{2}b(\widehat{\phi}^{\pm},e - \widehat{\phi}^{\pm}) \mp \Re\{\ell^{\circ}(u)\} \\ &= \frac{\kappa}{2}b(e,\widehat{\phi}^{\pm} - e) - \frac{\kappa}{2}b(\widehat{\phi}^{\pm},\widehat{\phi}^{\pm} - e) \mp \Re\{\ell^{\circ}(u)\} \\ &= -\frac{\kappa}{2}b(\widehat{\phi}^{\pm} - e,\widehat{\phi}^{\pm} - e) \mp \Re\{\ell^{\circ}(u)\}. \end{aligned}$$

Therefore, it now follows, that

$$s^{\pm} = \Re\{\ell^{\circ}(u)\} \pm \frac{\kappa}{2} b(\widehat{\phi}^{\pm} - e, \widehat{\phi}^{\pm} - e) \tag{5.51}$$

We will show later that indeed in the approximations, the second term above asymptotically decreases to zero as the mesh is refined and thus the bounds preserve the bounding property where the bounds converge from above and below to the upper and lower bounds respectively at the theoretical rate.

5.7 Sub-problem Computations

In order to understand the computation of the subdomain computations, we re-write the constraint equations using Green's identity:

$$\int_T \mathbf{q} \cdot \nabla \overline{v} \, d\Omega = -\int_T \nabla \cdot \mathbf{q} \, \overline{v} \, d\Omega + \int_{\partial T} \mathbf{q} \cdot \mathbf{n} \, \overline{v} \, d\Gamma$$

and choose an appropriate finite dimensional space which is rich enough such that the traces of the functions $\mathbf{q}$ on each elemental subdomain will at least contain the continuity multipliers. Therefore, by writing $\mathbf{q}^u \cdot \mathbf{n} = \sigma_T \lambda_h^u + \nabla u_h \cdot \mathbf{n}$ on ∂T, the constraint equation specifying the solution pair $(\mathbf{q}^u, r^u)$ is now expressed equivalently as

$$\int_T \left(-\nabla \cdot \mathbf{q}^u - k^2 r^u\right) \overline{v} \, d\Omega = \int_T \left(-f - \Delta u_h + k^2 u_h\right) \overline{v} \, d\Omega \tag{5.52}$$

for all $v \in \mathbb{P}^q(T)$. In addition, we recall that the spectral element approximation for the field variables and hybrid fluxes are polynomials of order p over the subdomain, namely, $u_h, \psi_h, \lambda \in \mathbb{P}^p(T)$. We see that for solvability, the data on the right must be in the range of the operator on the left, and therefore there is no problem for the variables $(\mathbf{q}^u, r^u) \in \mathbb{P}^q$, and the constraint equation is solvable for $q > p$ where we have also taken the forcing data to be polynomial of order $\mathbb{P}^p(T)$. An analogous reasoning applies to the solution pair $(\mathbf{q}^{\psi}, r^{\psi})$. We now formulate the pair of computable sup-problems as follows:

1. Find $\mathbf{q}_h^u, r_h^u, \xi_h^u \in \mathcal{Q}_h^u(T) \times \mathbb{P}^q(T) \times \mathbb{P}^q(T)$ such that

$$a_T^{du}((\mathbf{q}_h^u, r_h^u), (\mathbf{p}, v)) + c_T^{du}((\mathbf{p}, v), \xi^u) = 0 \quad \forall (\mathbf{q}, v) \in (\mathbb{P}^q(T))^3 \times \mathbb{P}^q(T), \tag{5.53}$$

$$c_T^{du}((\mathbf{q}_h^u, r_h^u), v) = \mathcal{R}_T^u(v) \quad \forall v \in \mathbb{P}^q(T); \tag{5.54}$$

2. Find $\mathbf{q}_h^\psi, r_h^\psi, \xi_h^\psi \in \mathcal{Q}_h^\psi(T) \times \mathbb{P}^q(T) \times \mathbb{P}^q(T)$ such that

$$a_T^{du}((\mathbf{q}_h^\psi, r_h^\psi), (\mathbf{p}, v)) + c_T^{du}((\mathbf{p}, v), \xi^\psi) = 0 \quad \forall (\mathbf{q}, v) \in (\mathbb{P}^q(T))^3 \times \mathbb{P}^q(T), \tag{5.55}$$

$$c_T^{du}((\mathbf{q}_h^\psi, r_h^\psi), v) = \mathcal{R}_T^\psi(v) \quad \forall v \in \mathbb{P}^q(T); \tag{5.56}$$

where we have defined the finite dimensional spaces $\mathcal{Q}_h^u, \mathcal{Q}_h^\psi$ as:

$$\mathcal{Q}_h^u = \left\{ \mathbf{q} \in (\mathbb{P}^q(T))^3 | \mathbf{q} \cdot \mathbf{n} = \sigma_T \lambda_h^u + \nabla u_h \cdot \mathbf{n} \text{ on } \partial T \right\} \tag{5.57}$$

$$\mathcal{Q}_h^\psi = \left\{ \mathbf{q} \in (\mathbb{P}^q(T))^3 | \mathbf{q} \cdot \mathbf{n} = -\sigma_T \lambda_h^\psi - \nabla \psi_h \cdot \mathbf{n} \text{ on } \partial T \right\}. \tag{5.58}$$

We note that since the polynomial approximations to $u_h^{(s)}, \psi_h^{(s)}, \lambda$ are less than the approximations for $(\mathbf{q}_h^u, r_h^u)$ and $(\mathbf{q}_h^\psi, r_h^\psi)$, these right hand side data must be interpolated in the same manner as is described in Chapter 4.

5.7.1 Discrete Forms and Approximations

As mentioned, in our implementation, we use real forcing with real boundary conditions, therefore the discrete forms corresponding to the above set of computable constraints are given for $\mathbf{q}_h^u = \mathbf{q}_h^{Ru} + i\mathbf{q}_h^{Iu} = \mathbf{q}_h^{Ru}$ and $r_h^\psi = r^{R\psi} + ir_h^{I\psi} = r_h^{R\psi}$ where the imaginary parts are equal to zero, and just express the discrete forms corresponding to the real part of the systems. Therefore, we have

1. For the system (5.53)-(5.54) the block matrix

$$\begin{bmatrix} M^{(s)} & 0 & 0 & -D_1^{t(s)} \\ 0 & M^{(s)} & 0 & -D_2^{t(s)} \\ 0 & 0 & M^{(s)} & -D_3^{t(s)} \\ -D_1^{(s)} & -D_2^{(s)} & -D_3^{(s)} & -k^2 M^{(s)} \end{bmatrix} \begin{bmatrix} (q_h^u)_1 \\ (q_h^u)_2 \\ (q_h^u)_3 \\ r_h^u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ g_h^u \end{bmatrix}$$

2. For the system (5.55)-(5.56) the block matrix

$$\begin{bmatrix} M^{(s)} & 0 & 0 & -D_1^{t(s)} \\ 0 & M^{(s)} & 0 & -D_2^{t(s)} \\ 0 & 0 & M^{(s)} & -D_3^{t(s)} \\ -D_1^{(s)} & -D_2^{(s)} & -D_3^{(s)} & -k^2 M^{(s)} \end{bmatrix} \begin{bmatrix} (q_h^\psi)_1 \\ (q_h^\psi)_2 \\ (q_h^\psi)_3 \\ r_h^\psi \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ g_h^\psi \end{bmatrix}$$

where

$$g_h^u = -M^{(s)} f + A^{(s)} u_h + B^t \lambda_h^u \tag{5.59}$$

$$g_h^\psi = -M^{(s)} f^\circ - A^{(s)} \psi_h - B^t \lambda_h^\psi. \tag{5.60}$$

The Lagrange multipliers ξ^u,and ξ^ψ are just an artifact of solving the quadratic programing problem with constraints, and are not required in the bounds calculations. Therefore, we have eliminated them from the system matrix, by noting that Equations (5.53) and (5.55) impose the conditions that $r_h^u = -\xi_h^u$ and $r_h^\psi = -\xi_h^\psi$. Here again, $M^{(s)}$, $A^{(s)}$ are the mass and stiffness matrices respectively, and the $D_i^{(s)}$ for $i = 1, 2, 3$ are the same as that used in Chapter 4.

Remark 5.2 *The square system matrices above, can be written in the general form*

$$\begin{bmatrix} \mathbf{M} & \mathbf{D}^t \\ \mathbf{D} & -\mathbf{C} \end{bmatrix}$$

where $\mathbf{M}$ *is the* 3×3 *block matrix with the subdomain mass matrix on the diagonals, and thus it is in general symmetric positive definite. The matrix* $\mathbf{D}$ *is a block matrix composed of the matrices* $D_i^{(s)}$ *and has full rank. The matrix* $C = k^2 M^{(s)}$ *and is positive definite. A theorem on saddle point matrices(see [10]) guarantees that such a matrix is invertible provided* $\ker(\mathbf{M}) \cap \ker(\mathbf{D}) = \{\mathbf{0}\}$. *Since the null space of the mass matrix is zero, and the null spaces of the* $\mathbf{D}$ *matrix are the space of constants (as it is the discrete form of* $\int_T \nabla \cdot \mathbf{q}\, v\, d\Omega$ *which involves the divergence operator) their intersection is the zero vector. Therefore, invertibility is not a problem, and we solve the system for* $(\mathbf{q}_h^u, r_h^u)$ *and* $(\mathbf{q}_h^\psi, r_h^\psi)$ *which will be used in the computation of the bounds.*

5.8 Computation of the Output Bounds

After the computation of the bounds variables, namely, $(\mathbf{q}^u, r^u)$ and $(\mathbf{q}^\psi, r^\psi)$, we may compute the bounds based on the formulation in (5.48) by summing these elemental contributions. We note that the elemental sub-problems are computed independently and in parallel, so by summing the contributions, one will arrive at the global bounds procedure. It is important to point out that the formulation in (5.48) required knowledge of the exact solution solution. However we adhere to the fact that the variables $(\mathbf{q}^u, r^u)$ and $(\mathbf{q}^\psi, r^\psi)$ result from a maximization procedure, so that any approximation to these from a suitably chosen finite dimensional set satisfying (5.53)-(5.54) will be bounded above by the exact solution, i.e. $a^{du}((\widehat{\mathbf{q}}_h^{u,*}, \widehat{r}_h^{u,*}), (\widehat{\mathbf{q}}_h^{u,*}, \widehat{r}_h^{u,*})) \leq a^{du}((\widehat{\mathbf{q}}^{u,*}, \widehat{r}^{u,*}), (\widehat{\mathbf{q}}^{u,*}, \widehat{r}^{u,*}))$. Hence the approximations $(\mathbf{q}_h^u, r_h^u)$ and $(\mathbf{q}_h^\psi, r_h^\psi)$, through bounded projection will provide the bounds to the

exact output. If we now define

$$z_h^u = \frac{1}{8}\sum_{T\in\mathcal{T}_h} a_T^{du}((\mathbf{q}_h^u, r_h^u), (\mathbf{q}_h^u, r_h^u)), \qquad z_h^\psi = \frac{1}{2}\sum_{T\in\mathcal{T}_h} a_T^{du}((\mathbf{q}_h^\psi, r_h^\psi), (\mathbf{q}_h^\psi, r_h^\psi)),$$
$$\bar{z}_h = \frac{1}{2}\sum_{T\in\mathcal{T}_h} a_T^{du}((\mathbf{q}_h^u, r_h^u), (\mathbf{q}_h^\psi, r_h^\psi)), \tag{5.61}$$

then we can write the total output bound expression as

$$s_h^{\pm} = \Re\{C^u\} - \Re\{\bar{z}_h\} \pm \left\{\kappa z_h^u + \frac{1}{\kappa} z_h^\psi\right\} \tag{5.62}$$

where

$$C^u = \ell^\circ(u_h) + a(u_h, \psi_h) - \ell^N(\psi_h)\} = \ell^\circ(u_h),$$

since $a(u_h, \psi_h) = \ell^N(\psi_h)$ from the primal problem. Optimizing with respect to κ we have $\kappa = \sqrt{\frac{z_h^\psi}{z_h^u}}$, and the upper and lower bounds become:

$$s_h^{\pm} = \Re\{\bar{s}_h\} \pm 2\sqrt{z_h^u z_h^\psi}, \tag{5.63}$$

where we have the defined the bound average $\Re\{\bar{s}_h\} = \Re\{-\bar{z}_h + C^u\}$.

5.9 Convergence Properties

In this section we would like to verify that the bounds converge asymptotically at the required optimal rate. We have seen that calculation of the bounds which depends on several smaller sub-problem computations, require knowledge of the decoupled solutions in which their aggregate leads to the global finite element solutions u_h, and ψ_h. These approximations have already been achieved using higher order nodal spectral element method, where the GCR method is used to converge the solution to $|u^{(s)} - u_h^{(s)}| \leq 10^{-6}$, likewise for $\psi_h^{(s)}$. Therefore at this stage concerns about the loss of ellipticity of the Helmholtz operator have already been addressed as u_h and ψ_h have been obtained under sufficient hp refinement.

However, the approximations $(\widehat{\mathbf{q}}_h^u, \widehat{r}_h^u)$ and $(\widehat{\mathbf{q}}_h^\psi, \widehat{r}_h^\psi)$ also suffer from the lack of stability of the forms $a^{du}((\widehat{\mathbf{q}}_h^u, \widehat{r}_h^u), (\widehat{\mathbf{q}}_h^u, \widehat{r}_h^u))$ and $a^{du}((\widehat{\mathbf{q}}_h^u, \widehat{r}_h^u), (\widehat{\mathbf{q}}_h^u, \widehat{r}_h^u))$ for higher wave numbers and so conditions for which these forms retain their ellipticity must be considered in the convergence proofs. In order to proceed with the proof, we use the result of a proof by Maday and Patera [40] which shows that the Lagrange multipliers λ_h^u, and λ_h^ψ satisfy the following estimates,

$$\sum_{T\in\mathcal{T}_h} h^{\frac{1}{2}} \|\lambda^u - \lambda_h^u\|_{\partial T} \leq C|u - u_h|_{H^1}, \tag{5.64}$$

$$\sum_{T\in\mathcal{T}_h} h^{\frac{1}{2}} \|\lambda^\psi - \lambda_h^\psi\|_{\partial T} \leq C|\psi - \psi_h|_{H^1} \tag{5.65}$$

where λ^u and λ^ψ are the exact Lagrange multipliers. Now we prove a proposition which gurantees the convergence of the bounds for the Helmholtz operator. The proof is similar to [59] but we modify it for the case of indefinte forms.

Proposition 5.3 *Suppose that u_h and ψ_h are the spectral element approximations to the primal and adjoint pair u and ψ respectively, and that λ_h^u and λ_h^ψ are the solutions to the equilibration equations, then*

$$s - s_h^- \leq C(k)\|u - u_h\|_{H^1}\|\psi - \psi_h\|_{H^1}$$

$$s_h^+ - s \leq C(k)\|u - u_h\|_{H^1}\|\psi - \psi_h\|_{H^1}$$

provided that λ_h^u and λ_h^ψ satisfy the prerequisites (5.64)-(5.65), and the constant C depends on the wave number k.

proof:

First we re-write the gradient conditions solving for $(\mathbf{q}^u, r^u) \in (L^2(T))^3 \times H^1(T)$ as

$$\int_T \mathbf{q}^u \cdot \overline{\mathbf{p}}\, d\Omega - k^2 \int_T r^u \overline{v}\, d\Omega + \int_T \overline{\mathbf{p}} \cdot \nabla \xi^u\, d\Omega - k^2 \int_T \overline{v}\xi^u\, d\Omega = 0 \tag{5.66}$$

$$\int_T \mathbf{q}^u \cdot \nabla \overline{v}\, d\Omega - k^2 \int_T r^u \overline{v}\, d\Omega = \mathcal{R}_T^u(v), \tag{5.67}$$

where we observe that the first equation holding for all $(\mathbf{p}, v)$ will lead to the conditions

$$r^{u,*} = -\xi^u, \quad \mathbf{q}^{u,*} = -\nabla\xi^u,$$

and substituting these conditions into the second equation we can write

$$\int_T \mathbf{q}^{u,*} \cdot \nabla\overline{\phi}^u \, d\Omega - k^2 \int_T r^{u,*}\overline{\phi}^u \, d\Omega = -\mathcal{R}_T^u(\phi^u) \quad \forall \phi^u \in H^1(T). \tag{5.68}$$

At the gradient conditions we have $(\mathbf{q}_T^{u,*}, r_T^{u,*}) = (\nabla\phi_T^{u,*}, \phi_T^{u,*})$ and where in addition to the fact that $a_T^{du}((\mathbf{q}, r), (\nabla v, v)) = c_T^{du}((q, r), v)$, one obtains from Equation (5.68) that

$$a^{du}((\nabla\widehat{\phi}^{u,*}, \widehat{\phi}^{u,*}), (\nabla\widehat{\phi}^{u,*}, \widehat{\phi}^{u,*})) = \ell_T^N(\widehat{\phi}^{u,*}) - b(u_h, \widehat{\phi}^{u,*}) - \mathsf{h}(\widehat{\phi}^{u,*}, \lambda_h^u), \tag{5.69}$$

where we have again used the diacritical hats to designate functions on the broken domain. Just as the approximate solution u_h has the associated Lagrange multipliers λ_h^u, the exact solution u also has the associated exact Lagrange mulitpliers λ^u that satisfy the equilibration condition

$$0 = \ell_T^N(\widehat{v}) - b(u, \widehat{v}) - \mathsf{h}(\widehat{v}, \lambda^u) \quad \forall \widehat{v} \in \widehat{Z} \tag{5.70}$$

Letting $\widehat{v} = \widehat{\phi}^{u,*}$ in Equation (5.70) and subtracting it from (5.69) we have

$$\begin{aligned}
&a^{du}((\nabla\widehat{\phi}^{u,*}, \widehat{\phi}^{u,*}), (\nabla\widehat{\phi}^{u,*}, \widehat{\phi}^{u,*})) \\
&\quad = b(u - u_h, \widehat{\phi}^{u,*}) + \mathsf{h}(\widehat{\phi}^{u,*}, \lambda^u - \lambda_h^u) \\
&\quad \leq C\|u - u_h\|_{H^1}\|\widehat{\phi}^{u,*}\|_{H^1} + \sum_{T\in\mathcal{T}_h} C\|\lambda^u - \lambda_h^u\|_{\partial T}\|\widehat{\phi}^{u,*}\|_{\partial T} \\
&\quad \leq C\|u - u_h\|_{H^1}\|\widehat{\phi}^{u,*}\|_{H^1} + \sum_{T\in\mathcal{T}_h} Ch^{\frac{1}{2}}\|\lambda^u - \lambda_h^u\|_{H^1}\|\widehat{\phi}^{u,*}\|_{H^1} \\
&\quad \leq C\|u - u_h\|_{H^1}\|\widehat{\phi}^{u,*}\|_{H^1} + C|u - u_h|_{H^1}\|\widehat{\phi}^{u,*}\|_{H^1}
\end{aligned}$$

where we have in addition, employed the continuity of bilinear forms, Equation (2.41) for $V_1 = V_2 = H^1$, Equation (5.64) and the inequality $\|\omega\|_{\partial T} \leq h^{\frac{1}{2}}\|\omega\|_{H^1(T)}$. Now we point out that the operator a_T^{du} is not coercive in the usual sense of Equation (2.39), but rather

is H^1-coercive satisfying a Gårding inequality (2.44). However from the inf – sup stability condition for indefinite forms mentioned in Chapter 2, we know that

$$\exists \beta > 0 : \quad \beta \leq \sup_{0 \neq \widehat{\phi}^{u,*} \in v_2} \frac{|a^{du}(.,.)|}{\|\widehat{\phi}^{u,*}\|_{V_1} \|\widehat{\phi}^{u,*}\|_{V_2}} \quad \forall 0 \neq \widehat{\phi}^{u,*} \in V_1, \tag{5.71}$$

and thus for $V_1 = V_2 = H^1$ yields

$$\|\widehat{\phi}^{u,*}\|_{H^1} \leq \frac{1}{\sqrt{\beta}} \left\{ |a^{du}((\nabla \widehat{\phi}^{u,*}, \widehat{\phi}^{u,*}), (\nabla \widehat{\phi}^{u,*}, \widehat{\phi}^{u,*}))| \right\}^{\frac{1}{2}}. \tag{5.72}$$

Here β is the stability constant, which as mentioned previously, is typically larger for indefinite forms. For the Helmholtz sesquilinear forms in one dimension, $\beta = \frac{\text{constant}}{k}$. Therefore, $\frac{1}{\beta}$ is at least directly proportional to k for the three-dimensional problem. Employing (5.72) and knowing that the H^1 semi-norm is equivalent to the H^1 norm, then we have

$$a^{du}((\nabla \widehat{\phi}^{u,*}, \widehat{\phi}^{u,*}), (\nabla \widehat{\phi}^{u,*}, \widehat{\phi}^{u,*})) \leq C_1 \|u - u_h\|^2_{H^1}, \tag{5.73}$$

where $C_1 = \frac{C}{\sqrt{\beta}}$ and will be some function of wave number. We note that this will cause no difficulty in the convergence of the bounds. Similar estimates hold for the bound variables $(\mathbf{q}^{\psi,*}, r^{\psi,*})$. Again, by bounded projection which is a consequence of the maximization problem, the approximation will be bounded from above and converge to the maximum values. More precisely, after substituting back the relationship $(\nabla \phi_T^{u,*}, \phi_T^{u,*}) = (\mathbf{q}_h^{u,*}, r_h^{u,*})$ and recalling that $a^{du}((\mathbf{q}_h^{u,*}, r_h^{u,*}), (\mathbf{q}_h^{u,*}, r^{u,*})) \leq a^{du}((\mathbf{q}^{u,*}, r^{u,*}), (\mathbf{q}^{u,*}, r^{u,*}))$, we may write

$$a^{du}((\mathbf{q}_h^{u,*}, r_h^{u,*}), (\mathbf{q}_h^{u,*}, r_h^{u,*})) \leq C_1 \|u - u_h\|^2_{H^1}, \tag{5.74}$$

and analogously,

$$a^{du}((\mathbf{q}_h^{\psi,*}, r_h^{u,*}), (\mathbf{q}_h^{\psi,*}, r_h^{\psi,*})) \leq C_1 \|\psi - \psi_h\|^2_{H^1}. \tag{5.75}$$

It is clear from the bound computation (5.63), that

$$|s - s_h| \leq s_h^+ - s_h^- \leq 4\sqrt{z^u z^\psi} \leq C \|u - u_h\|_{H^1} \|\psi - \psi_h\|_{H^1}$$

where s_h is the nodal spectral element approximation of s. □

Chapter 6

Numerical Verification and Results

In this chapter we verify the results of the bounds to the exact outputs of the Helmholtz equation for wave numbers $k = 1, 3, 5$. We use a higher order nodal spectral element method where the approximations $\lambda_h, \psi_h^{(s)}, u_h^{(s)} \in \mathbb{P}^p$, and the bound components $\mathbf{q}_h, r_h \in \mathbb{P}^q$ with $q > p$ consistent with the method as presented in the previous chapter. We consider the numerical examples in a cube geometry $[0,1] \times [0,1] \times [0,1]$ with homogenous Dirichlet boundary conditions, where $u\Big|_{\partial\Omega} = 0$. We then verify our results against a constructed exact solution where we take the forcing function to be

$$f(x,y,z) = k^2(3\pi^2 - 1)\sin(k\pi x)\sin(k\pi y)\sin(k\pi z)$$

which would yield the exact solution $u = \sin(k\pi x)\sin(k\pi y)\sin(k\pi z)$, satisfying the boundary conditions. Also taking $f^\circ = 1.0$ gives an exact output of $s = \frac{8}{(k\pi)^3}$. For the different wave numbers, we do a convergence study where we plot the logarithm of the errors $|s - s_h^\pm|$ and $|s - s_h|$, versus the logarithm of the mesh size, where $s_h^\pm$ are the computed upper and lower bounds and s_h is the spectral element solution to the output. Moreover, we also consider the effectivity of the bounds for some cases, where we recall from Chapter 4 that the effectivity index is defined as $\theta_h^- = \frac{|s - s_h^-|}{|s - s_h|}$ and $\theta_h^+ = \frac{|s - s_h^+|}{|s - s_h|}$ for both the lower and upper bounds respectively. This is an indication of how well the bounds perform compared to the

spectral element solution.

6.1 Discussion of Results

6.1.1 For approximation polynomials of order $p = 2, q = 3$

Case I: $k = 1$

The theoretical rate of convergence for the nodal spectral element solution is consistent with $u \in \mathbb{P}^2$ for structured meshes, and is $O(h^4)$. By the discussion in Section (5.9), it is expected that the upper and lower bounds converge at the same theoretical rate of $O(h^4)$ as we refine the mesh. We see in Figure (6.1) that the spectral element solution converges faster, than the output bounds, which is generally the case, however for this case the convergence rate and results are very nearly the same as the Poisson problem of Chapter 4 because of the low wave number. We see from Table (6.1) that the effectivity of the bounds for $k = 1$ imply that the bounds are sharp and compare well with the spectral element approximation as they are very near equal to 1.0. Figure (6.2) depicts the bounds for this case.

Case II: $k = 3$

When we increase the wave number to $k = 3$ we see from the convergence study in Figure (6.3) that the gap between the Spectral element approximation and the bounds is widened which indicates loss in the sharpness of the bounds. The fact that the upper bound and the lower bound overlap indicates that the bounds are symmetric about the exact solution as seen in Figure (6.4). However from the figures and from Table (6.1) it is clear the a higher wave number causes the effectivity of the bounds to deteriorate, although for practical engineering applications, higher effectivities can also be acceptable. For improving the effectivities, one can used adaptive strategies and polynomial approximation of higher than $p = 2, q = 3$. We calculate the relative error defined by $\frac{|s - s_h^{\pm}|}{s}$ for a mesh of $h = 1/18$ to be approximately 11% error for both the upper bound and the lower bound. A mesh of

1/18 is relatively coarse for three-dimensional problems and an 11% error is considered to be within the acceptable range for engineering applications. However, we note that higher order polynomials can drastically reduce this error.

$f = k^2(3\pi^2 - 1)\sin(k\pi x)\sin(k\pi y)\sin(k\pi z)$ For $k = 1$, $s = 8/\pi^3 \approx 0\ 2580122$					$f = k^2(3\pi^2 - 1)\sin(k\pi x)\sin(k\pi y)\sin(k\pi z)$ For $k = 3$, $s \approx 0\ 00955601$				
h	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	θ_h^-	θ_h^+	**h**	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	θ_h^-	θ_h^+
1/2	0 21209	0 33598	1 21	2 06	1/4	−0 08295	0 10359	21 42	21 77
1/4	0 25233	0 26810	1 78	3 16	1/6	−0 027301	0 04764	26 20	27 07
1/8	0 25752	0 25895	2 30	4 35	1/8	−0 00620	0 02589	29 76	30 86
1/16	0 25797	0 25811	2 88	6 83	1/16	0 00799	0 01118	40 02	41 62

Table 6.1: Tabulated bounds results and their effectivities obtained with the cubic elemental subdomains for both $k = 1$ and $k = 3$ In both case the results are reported using $p = 2$, $q = 3$

Case III: $k = 5$

For this case we see from the widening between the convergence lines of the corresponding to the errors $|s - s_h|$ and $|s - s_h^{\pm}|$ that the effectivities deteriorate for the higher wave number in Figure (6.5). As mentioned before, the effectivity can be improved through adaptive procedures and more accurate numerical methods such as higher order polynomial basis. However, we further observe that as we refine the mesh for higher wave numbers, the convergence of the bounds will reach a plateau. This is an anomaly that we believe can be attributed to the interpolation of the hybrid-fluxes where we recall from Chapter 4 consists of interpolating from a lower order polynomial to a higher-order polynomial (with more discrete points on the tetrahedron) space. In this case, accurate interpolation at higher frequencies is not achieved and has a deleterious effect on the convergence of the bounds. Figure (6.6) depicts the bounds for this case which verifies that both the upper and lower bounds approach the exact solution at the same rate and are symmetric about the exact solution, although not as sharp as for the $k = 1$ and $k = 3$ case.

6.1.2 For approximation polynomials of order $p = 4, q = 5$

Case I: $k = 3$

Higher order polynomials are used in order to improve the bounds. We do this for $p = 4, q = 5$ and observe in Figure (6.7) that for the coarsest mesh of $h = 1/4$, the higher-order has tightened the gap between the convergence lines where the relative error for the upper bound is approximately 13 5% and the lower bound is approximately 12%. Figure (6.8) demonstrates the bounding property for this case, where the value of the bounds is tabulated in Table (6.2). We point out that the relative error at $h = 1/16$ for the upper bound is 0.59% and the lower bound 0.57%. The bound average is comparable with the spectral element solution. Although when an exact solution is unavailable, the bounds will provide certainty information, where in this case, we see that the bound average gives a relative error of 0.009%. We then compare this to the relative errors at $h = 1/16$ of 17% and 16% for the upper and lower bounds respectively, when $p = 2, q = 3$. Indeed the higher-order has sharpened the bounds.

It is also observed that as we refine the mesh, we reach a plateau. We again believe this is a result of interpolation of the hybrid-fluxes which is also affected by the interpolation of the hybrid-fluxes at higher polynomial orders.

$p = 4, q = 5$ For $k = 3$, $s \approx 0\ 00955601$				
h	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	$\mathbf{s_{av}}$	$\mathbf{s_h}$
1/4	0 0084193	0 0108484	0 0096338	0 0096314
1/8	0 0094563	0 0096642	0 0095602	0 0095590
1/12	0 0094919	0 0096229	0 0095574	0 0095563
1/16	0 0095017	0 0096121	0 0095569	0 0095561

Table 6.2: Tabulated bounds results obtained for the the cubic elemental subdomains for $k = 3$ using $p = 4, q = 5$

Case II: $k = 5$

Here we report the results for $k = 5$ using different polynomial orders in Table (6.3) where we demonstrate that the higher order approximation gives very improved bounds. For a mesh size of $h = 1/20$, using $p = 4, q = 5$ we obtain a relative error of 19% for both the upper and lower bounds. The bound average for this case is very near to the spectral element solution and has a relative error of 0.014%. Therefore, when no exact solution is available, one can use the bound average in order to provide very sharp estimates of the output. The bounds for this case is depicted in Figure (6.10). Again, we can expect from previous results that the convergence rate will be affected by an interpolation error of the hybrid-fluxes. However, as seen in Figure (6.5) for $p = 2, q = 3$ when $h = 1/20$ the log of the error is approximately -2.5 before it plateaus out, while for $p = 4, q = 5$ the log of the error is approximately -3.5. This implies that going to higher-order results in a significant improvement in the sharpness of the bounds despite the loss of convergence. Finally, Figure (6.11) depicts the solution u, when $k = 5$, propagating along the x direction.

$p = 2, q = 3$ For $k = 5$ $s \approx 0\ 002064098$				
h	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	$\mathbf{s_{av}}$	$\mathbf{s_h}$
1/8	−0 1497625	0 1542674	0 0022525	0 0014725
1/12	−0 0496803	0 0539782	0 0021489	0 0018986
1/16	−0 0193310	0 0235316	0 0021003	0 0020051
1/20	−0 0084544	0 01261805	0 0020818	0 0020385

$p = 3$ $q = 4$ For $k = 5$, $s \approx 0\ 002064098$				
h	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	$\mathbf{s_{av}}$	$\mathbf{s_h}$
1/8	−0 0156826	0 0198758	0 0020966	0 0020493
1/12	−0 0044378	0 0085799	0 0020711	0 00206798
1/16	−0 0009765	0 0051102	0 0020669	0 0020663
1/20	0 0002274	0 0039038	0 0020656	0 0020652

$p = 4$ $q = 5$ For $k = 5$ $s \approx 0\ 002064098$				
h	$\mathbf{s_h^-}$	$\mathbf{s_h^+}$	$\mathbf{s_{av}}$	$\mathbf{s_h}$
1/5	−0 0121749	0 0164506	0 0021379	0 00210508
1/10	0 00091761	0 0032158	0 0020667	0 0020671
1/15	0 0015841	0 0025453	0 0020647	0 0020644
1/20	0 0016673	0 0024615	0 0020644	0 0020642

Table 6 3 Tabulated results for the bounds and the spectral element solution s_h obtained with the cubic elemental subdomains for $k = 5$ using different polynomial orders

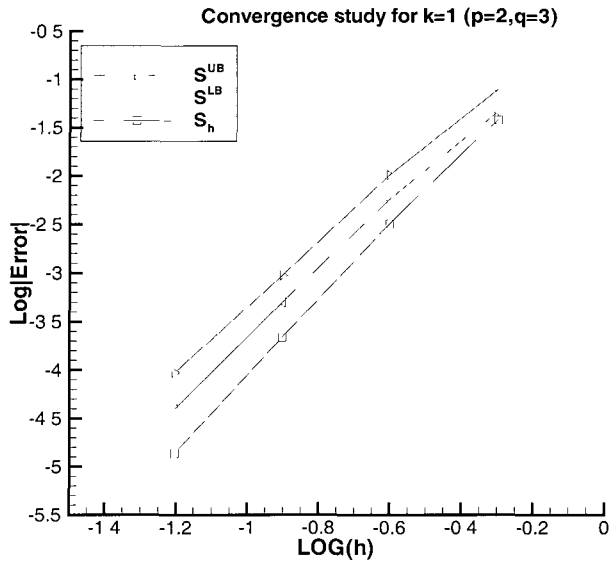


Figure 6.1: Convergence study for $k = 1$ which shows the reduction of the error in the bounds and the spectral element solution as we refine the mesh. The error in the spectral solution converges at a faster rate, however this is typical and is reflected in the effectivity (Table 6.1). The theoretical convergence rate here is $O(h^4)$.

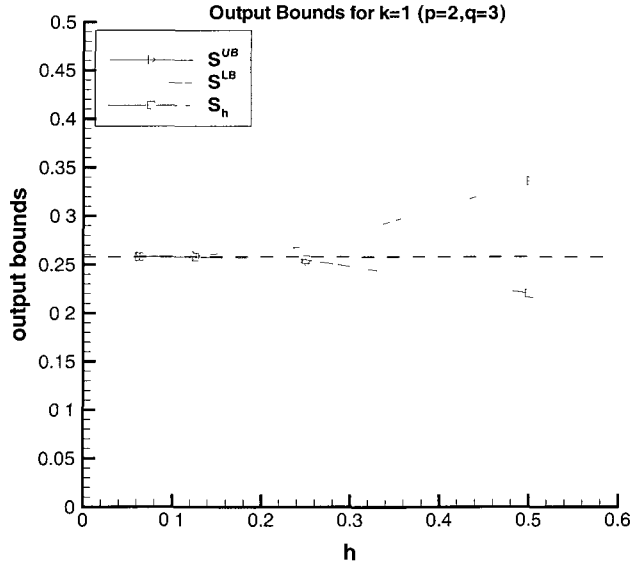


Figure 6.2: Figure demonstrating the bounding property of the bounds for $k = 1$. The dashed lines represent the exact solution, and we see that for all levels of refinement, the bounds hold.

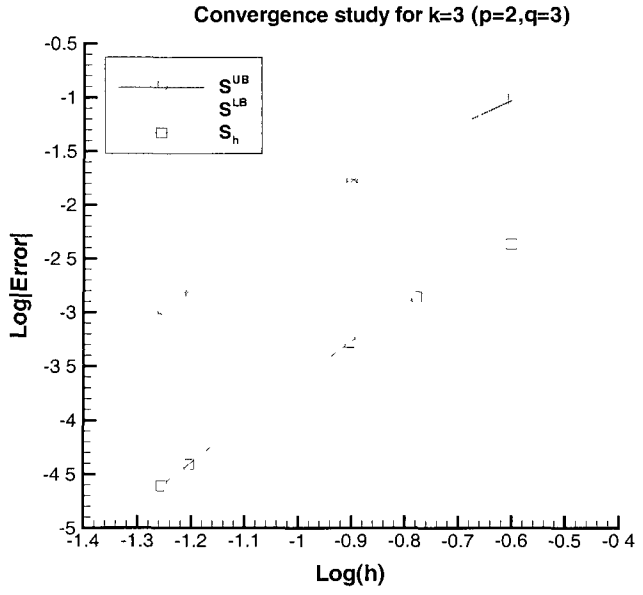


Figure 6.3: Convergence study for $k = 3$, where the theoretical convergence rate is $O(h^4)$.

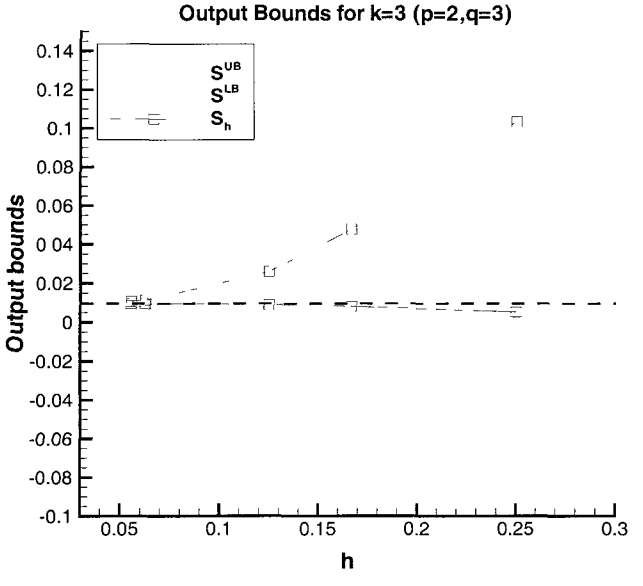


Figure 6.4: Upper and lower bounds for $k = 3$. The dashed lines indicate the exact solution.

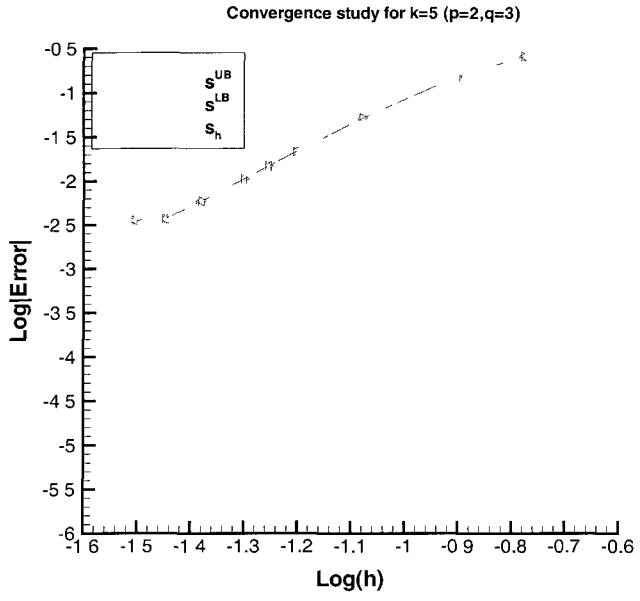


Figure 6.5: Convergence study for $k = 5$, where the theoretical convergence rate is again $O(h^4)$. The widening of the convergence lines indicates a loss of effectivity in the bounds.

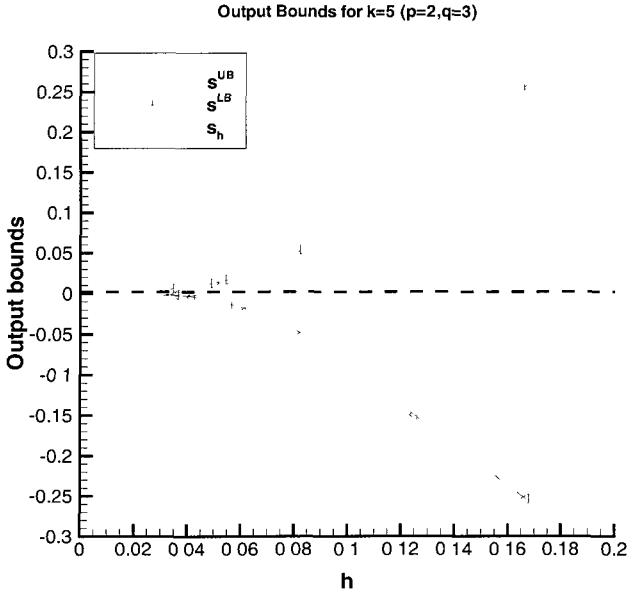


Figure 6.6: Upper and lower bounds for $k = 5$. The dashed lines indicate the exact solution. The bounds are symmetric about the exact solution and approach the exact solution at the same rate.

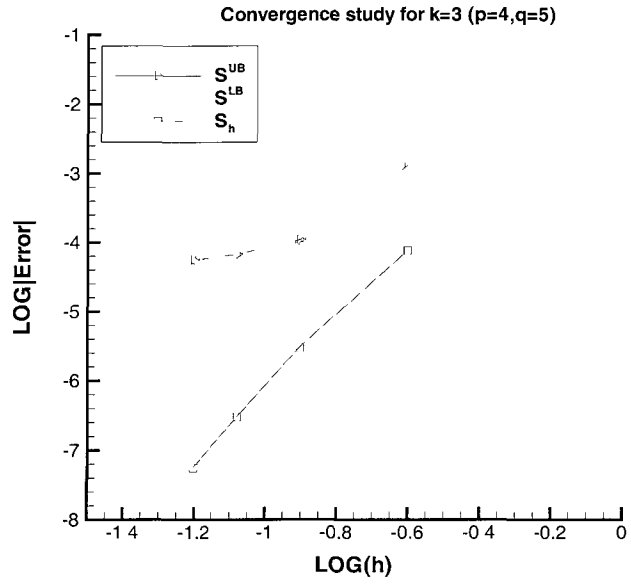


Figure 6.7: Convergence study for $k = 3$ using higher order elements, where the theoretical convergence rate is of $O(h^6)$. The bounds overlap which again indicates symmetry about the exact solution.

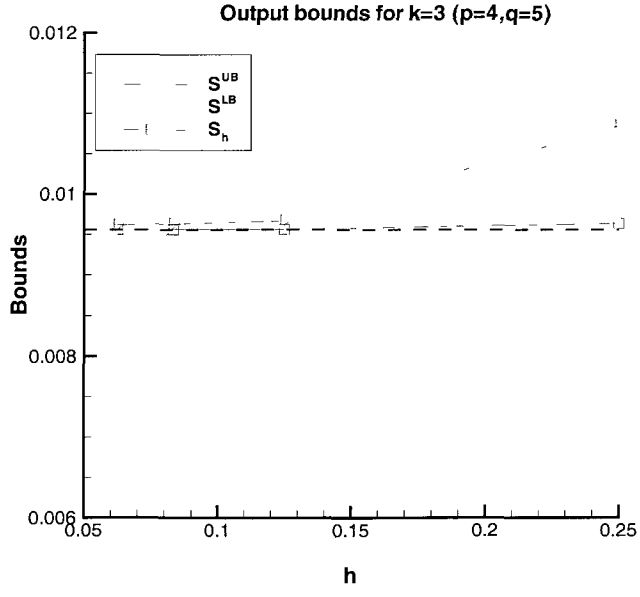


Figure 6.8: Figure demonstrating the bounding property of the bounds for $k = 3$ using higher order elements.

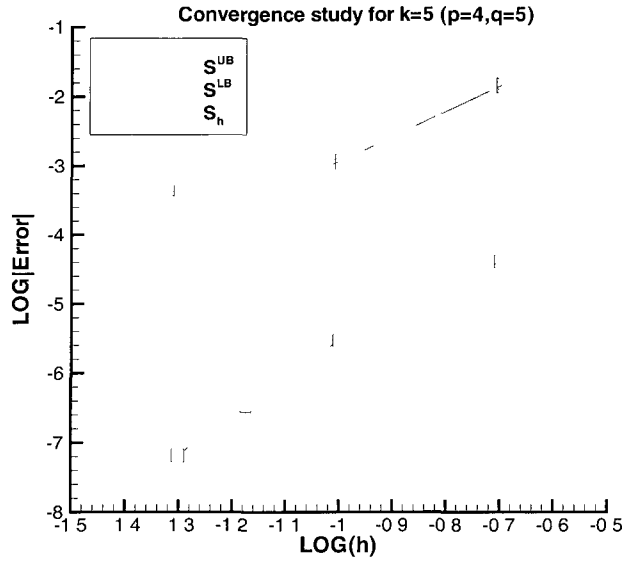


Figure 6 9 Convergence study for $k = 5$ using higher order elements Again we observe that the bounding lines overlap which implies symmetry in the bounds The bounds converge well until the point where it begins to plateau out

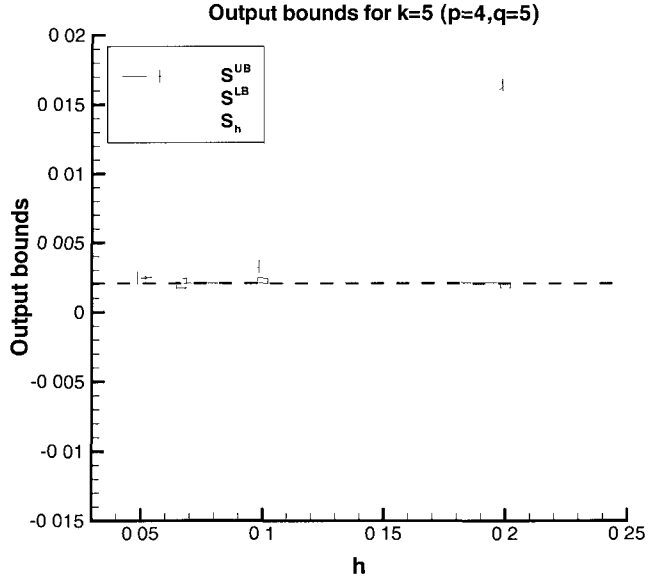


Figure 6 10 Upper and lower bounds for $k = 5$ using higher order We do see that bounds are much tighter than the $p = 2, q = 3$ case despite the loss of convergence For a comparison see Table 6 3

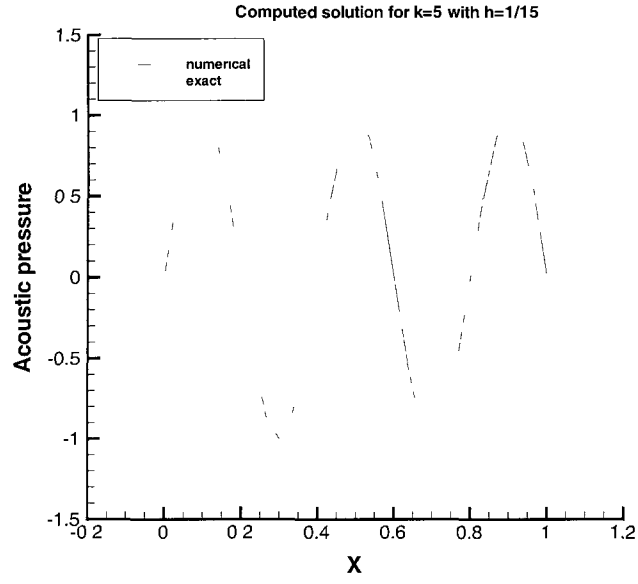


Figure 6.11: Propagation of acoustic wave in the x-direction with $k = 5$. A comparison of the numerical and exact solution and the convergence rate of the numerical approximation s_h in the above figure shows the precision of the spectral element solution.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have developed the method of exact bounds for the three-dimensional Helmholtz equation, where we have formulated the proofs in a complex space setting. We report that these bounds are rigorous and provide a certificate of precision for a predicted output with full certainty.

In Chapter 2 we gave a review of many of the difficulties associated with the numerical solution of the Helmholtz equation. As the bounds method adheres to a domain decomposition procedure where the global bounds are obtained via several local subdomain computations, we have invoked a FETI like procedure for the calculation of the hybrid fluxes. The loss of ellipticity for the Helmholtz equation will render the problem as ill-posed which will then cause the discrete system matrix to become indefinite. We have followed Farhat et al. [23] and implemented an additional regularizing term to the system matrix to prevent the problem from becoming singular. Such a procedure eliminates resonance between the subdomains while leaving the global nature of the numerical solution unaffected by the additional term. Furthermore, we have mentioned that in order to achieve a reliable result when approximating an oscillatory function u (solution of the Helmholtz equation) by finite element interpolating polynomials, the resolution of the mesh should be

adjusted to the wave number (the "rule of thumb"). However, while the "rule of thumb" controls the discretization errors, it cannot control the pollution error which is inherent in the Garlerkin FEM. Such an error is due to a loss of stability, and that in order to obtain asymptotic convergence of the numerical solution, one must restrict the size of the mesh to satisfy $hk^2 < 1$ which will render the standard h-version of the Galerkin FEM as very expensive, especially for practical engineering problems. The hp-version, on the other hand, uses higher order polynomials and is shown in [35] to always lead to a reduction of the pollution error. Moreover, a study done in [41], has compared degrees of freedom per wavelength for polynomial orders of $p = 2$, and $p = 5$ which indicates that going to higher order can significantly reduce the number of degrees of freedom (DOF) of the problem, and consequently, reduce the CPU time. For higher frequencies, a greater reduction in DOF can be achieved by going to higher order.

In Chapter 6 we present our results of the method applied to the three-dimensional Helmholtz problem, where we demonstrate the bounding properties. Equation 5.51, namely,

$$s^{\pm} = \Re\{\ell^{\circ}(u)\} \pm \frac{\kappa}{2} b(\widehat{\phi}^{\pm} - e, \widehat{\phi}^{\pm} - e)$$

indicates that the bounds preserve the bounding property as $e \rightarrow 0$ provided that the Helmholtz operator $b(.,.)$ has a positive definite structure. As mentioned before, this is only achieved under sufficient hp-refinement. Thus, when we mention strict guaranteed bounds, we are referring to bounds on the exact solution which hold for all levels of refinement. However in the case of the Helmholtz problem, the bounds are asymptotic in the sense that in order for the bounds to satisfy the bounding property the operator $b(.,.)$ must be in the coercive regime. Our results indicate that by incorporating the nodal spectral element method with the FETI approach leads to excellent convergence results for the primal and adjoint pairs u_h and ψ_h respectively. The theoretical convergence rate of the upper and lower

bounds is that of the spectral element solution. The results of $k = 1$ clearly demonstrate this, however we observe that for higher wave numbers the effectivity of the bounds will begin to deteriorate. This implies that while the bounds approach their theoretical convergence rate, compared to the convergence rate of the spectral element solution, they perform worse for higher k.

All in all, the current work contains several new contribution which are proven useful not only to certifying the precision of the approximation to specified outputs, but also in a potentially highly accurate domain decomposition scheme for solving the Helmholtz problem. These new ingredients include:

1. The development of the exact bounds method for the complex Helmholtz equation
2. The application of FETI-H to the exact bounds method
3. The use of FETI-H with high-order nodal spectral element method applied to the exact bounds method.

7.2 Future Work

We have solved the problem using higher-order polynomials on both structured and unstructured meshes. It is observed that invoking higher-order polynomials significantly reduces the bound gap Δ which is defined from Equation (5.63) to be

$$\Delta = s^+ - s^- = 4\sqrt{z_h^u z_h^{\psi}}.$$

For example, from Table (6.3) it is observed that for $k = 5$ with a mesh size of $h = 1/20$ we have

$$\Delta_{(p=2,q=3)} \approx 0.02107245, \quad \Delta_{(p=3,q=4)} \approx 0.0036764, \quad \Delta_{(p=4,q=5)} \approx 0.0007942,$$

which shows a reduction in bound gap of more than six times when going from $p = 2, q = 3$ to $p = 3, q = 4$ and almost five times when increasing order from $p = 3, q = 4$ to $p = 4, q = 5$. However, convergence results have shown that when we increase the wave number or go to higher order polynomials, the bounds begin to lose their convergence rate and reach a plateau. Therefore, as we refine the mesh for the higher-order polynomials, after a specific mesh refinement, we cannot expect any significant reduction in the bound gap. Our numerical experiments have shown that in the case of tetrahedral subdomains, where $p = 1$, $q = 2$ the bounds converge at the theoretical rate of $\mathrm{O}(h^2)$ while polynomial orders higher than this, i.e $p > 1$, $q > 2$ have demonstrated a loss of convergence for the bounds. For the structured mesh where we use cubic subdomains we don't lose the convergence properties until $p > 2$, $q > 3$. For polynomial orders $p > 2$, $q > 3$, then we get a similar behavior in convergence rate as in the unstructured mesh. It is believed that the plateauing out of the convergence rate, is a result of the interpolation of the hybrid-fluxes (see Chapter 4), where in the interpolating process from a lower polynomial space to a higher we tend to lose important information of the fluxes. As hinted earlier, using a flux-free approach may be deemed necessary in order to improve the method.

Ultimately, we would like to apply the method to more practical engineering problems. Moreover, the method as is presented can be readily extended to more complicated boundary conditions. It can also be extended to more complicated geometries where we have piecewise straight boundaries. Such restriction of the computational domain avoids accounting for the geometrical error arising from the finite discretization of curved boundaries. A method could be developed however, so that the error contribution to the output for curved geometries is calculated and added to the existing bounds.

In summary, in order to improve the method and its effectiveness our current aim will be directed as follows:

1. **Improvement of the method** First step is to alleviate the convergence issues described in the previous chapter. One potential remedy, can be to use the flux-free approach instead of the FETI like procedure in calculating the Lagrange multipliers.

2. **Application** Apply the method to more complicated geometries (with piecewise straight edge boundaries). In particular, investigate its potential application in calculating sharp, quantitative and cost effective upper and lower bounds for the transmission loss (TL) in mufflers and silencers, i.e. $TL^{LB} < TL < TL^{UB}$, where TL for a given frequency is the most important acoustic parameter.

3. **Minor enhancements**

   (a) Possible use of adaptive methods can be invoked to improve the effectivity of the bounds.

   (b) Implementation of the coarse mesh preconditioner in the FETI-H method in order to accelerate convergence of the spectral element solution.

Bibliography

[1] R. A. Adams, J. J.F. Fournier, *Sobolev Spaces*, Academic Press, New York, 2003.

[2] M. Ainsworth and J.T. Oden, "A Unified Approach to A Posteriori Error Estimation Based on Element Residual Methods", *Numer. Math.*, **65**, 23-50, 1993.

[3] M. Ainsworth and J.T. Oden, *A posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, 2000.

[4] I. Babuska, "Error bounds for finite element method", *Numer. Math*, **16**, 322-333, 1971.

[5] I. Babuska, F. Ihlenburg, T. Strouboulis and S. K. Gangaraj, "A posteriori error estimators for finite element solutions of Helmholtz equation, Part II. Estimation of the pollution error", *Int. J. Numer. Methods Engrg.*, **40**, p. 3883-3900, 1997.

[6] I. Babuska and A. Miller,"The Post-Processing Approach in the Finite Element Method-Part 3: A posteriori Error Estimates and Adaptive Mesh Selection", *Inter. J. Numer. Methods Engrg.*, **20**, 2311-2324, 1984.

[7] I. Babuska and W. Rheinboldt,"A posteriori Error Estimates for the Finite Element Method", *Int. J. Numer. Methods Engrg.*, **12**, 1597-1615, 1978.

[8] I Babuska and S. A. Sauter, "Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers", *SIAM J. Num. Anal.*, **34**(6), 2392-2423, 1997.

[9] R. Bank and A. Weiser, "Some *A posteriori* Error Estimates for Elliptic Partial Differential Equation", *Math Comp.*, **44**, 283-301, 1985.

[10] M. Benzi, G. H. Golub and J. Liesen, "Numerical Solution of Saddle Point Problems", *Acta Numerica.* **14**, 1-137, 2005.

[11] A. de La Bourdonnaye, C. Farhat, A. Macedo, F. Magoules and F. X. Roux, "A non overlapping domain decomposition method for the exterior Helmholtz problem", *Contemporary Mathematics*, **218**, 42-66, 1998.

[12] A. Bunse-Gerstner, and Ronald Stöver, "On a conjugate gradient type method for solving complex symmetric linear systems", *Linear Algebra and its Applications*, **287**, 105-123, 1999.

[13] L. Chamoin and P. Ladevèze, "Bounds on History Dependent or Independent Local Quantities in Viscoelasticity Problem Solved by Approximate Methods", *Int. J. Numer. Methods Engrg.*, **71**, 1387-1411, 2007.

[14] Z. Cheng, *A posteriori Finite Element Bounds: Application to Outputs of the Three Dimensional Navier-Stokes Equations*, PhD. Dissertation, University of Toronto, 2003.

[15] Z. Cheng, S. Ghomeshi and M. Paraschivoiu, "Bounds on Outputs of the Exact Weak Solution of the Three-Dimensional Stokes Problem",*Int. J. Numer. Meth. Fluids*, **61**(10), 1098-1131, 2009.

[16] H. W. Choi, *A posteriori finite element bounds with adaptive mesh refinement: application to outputs of the three dimensional convection-diffusion equation*, MSc. thesis, University of Toronto, 2001.

[17] H. W. Choi and M. Paraschivoiu, "Adaptive computations of A posteriori Finite Element Output Bounds· a Comparison of the 'hybrid flux' approach and the 'flux-free' approach", *Comp. Methods Appl. Mech. Engrg.*, **193**(36-38), p. 4001-4033, 2004.

[18] P. G. Ciarlet, *The Finite Element Method For Elliptic Problems*, North Holland, Amsterdam, 1978.

[19] R. Cottereau, P. Díez, and A. Huerta, "Strict Error Bounds for Linear Solid Mechanics Problems using a Subdomain-based Flux-Free Method", *Comput. Mech.*, **44**, 533-547, 2009.

[20] M. Crouzeix and P-A. Raviart, "Conforming, nonconforming finite element methods for solving the stationary Stokes equations," *RAIRO-Analyse Numerique*, **7**, 33-75, 1973.

[21] P. Díez and G. Calderón,"Goal-Oriented Error Estimation for Transient Parabolic Problems", *Comput. Mech.*, **39**(5), 631-646, 2007.

[22] C. Farhat, "A Lagrange multiplier based on divide and conquer finite element algorithm," *J. Comput. System Engrg.*, **2**, 149-156, 1991.

[23] C Farhat, P.-S. Chen, F. Risler and F. X. Roux, "A Unified Framework for Accelerating the Convergence of Iterative Substructuring Methods with Lagrange Multipliers", *Int. J. Numer. Methods Engrg.*, **42**, 257-288, 1998.

[24] C. Farhat, I. Harari and L. P. Franca, "The Discontinuous Enrichment Method", *Comput. Methods Appl. Mech. Engrg.*, **190**, 6455-6479, 2001.

[25] C. Farhat, A. Macedo and M. Lesoinne, "A Two-Level Domain Decomposition Method for the Iterative Solution of High Frequency Exterior Helmholtz Problems", *Numer. Math.*, **85**, 283-308, 2000.

[26] C. Farhat, A. Macedo, M. Lesoinne, F. X. Roux, R. Magoulès and A. de La Bourdonnaie, "Two-Level Domain Decomposition Methods with Lagrange Multipliers for the Fast Iterative Solution of Acoustic Scattering Problems", *Comput. Methods Appl. Mech. Engrg.*, **184**, 213-239, 2000.

[27] C. Farhat, K. Pierson and M. Lesoinne, "The second generation FETI methods and theri applications to parallel solution of large-scale linear and geometrically non-linear structural analysis problems," *Comput. Methods Appl. Mech. Engrg.*, **184**, 333-374, 2000.

[28] C. Farhat and F. X. Roux, "A Method of Finite Element Tearing and Interconnecting and its Parallel Solution Algorithm", *Int. J. Numer. Methods Engrg.*,**32**, 1205-1227(1991).

[29] C. Farhat and F.-X. Roux, "Implicit Parallel Processing in Structural Mechanics", *Computational Mechanics Advances*, North Holland, Vol. 2, No. 1 1994.

[30] P. Filippi, D. Habault, JP. Lefebvre and A. Bergassoli, *Acoustics: Basic physics, theory and methods*, Academic Press, San Diego, 1999.

[31] R. W. Freund, "Conjugate Gradient-Type Methods For Linear Systems with Complex Symmetric Coefficient Matrices", *SIAM J. Sci. Stat. Comput.*, **13**(1), 425-448, 1992.

[32] W. Hackbusch, *Elliptic Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[33] J. S. Hesthaven and T. Warburton, "Nodal High-Order Methods on Unstructured Grids", *J. Comput. Phys.*, **181**, 186-221, 2002.

[34] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Volume 132 of *Applied Mathematical Sciences* Springer-Verlag, New York, 1998.

[35] F. Ihlenburg and I. Babuska, "Finite Element Solution to the Helmholtz Equation with High Wave Number Part I: The h-version of the FEM.", *Comput. Math. Appli.*, **30**(9), 9-37, 1995.

[36] F. Ihlenburg and I. Babuska, "Finite Element Solution to the Helmholtz Equation with High Wave Number-Part II: The h-p version of the FEM", *SIAM J. Numer. Anal.*, **34**(1), 315-358, 1997.

[37] G. M. Karniadakis and S. J Sherwin, *Spectral/hp element methods for Computational fluid dynamics, Second Edition*, Oxford University Press, Oxford, England, 2005.

[38] P. Ladevéze and D. Leguillon,"Error estimation procedures in the finite element method and applications",*SIAM J. Numer. Anal.*, **20**, 485-509, 1983.

[39] P. Ladevéze and JP. Pelle, *Mastering calculations in linear and nonlinear mechanics.*, Mechanical engineering , Springer-Verlag, Heidelberg, 2005.

[40] Y. Maday and T. Patera, "Numerical analysis of a posteriori finite element bounds for linear functional outputs", *Math. Models Methods Appl. Sci.*, **10**, 785-799, 2000.

[41] O. Z. Mehdizadeh, *Simulation of Acoustic Performance of Mufflers and Silencers Using Spectral Element Methods*, PhD. Dissertation, University of Toronto, 2005.

[42] W. L. Oberkampf and T. G. Trucano, "Verification and Validation in Computational Fluid Dynamics", *Progress in Aerospace Sciences*, **38**, 209-272, 2002.

[43] H. Melbø and T. Kvamsdal,"Goal-Oriented Error Estimators for the Stokes Equation base on Variationally Consistent Post-processing", *Comp. Methods Appl. Mech. Engrg.*, **192**, 613-633, 2003.

[44] J. T. Oden and S. Prudhomme, "Goal-Oriented Error Estimation and Adaptivity for the Finite Element Method", *Computers and Mathematics with Applications*, **41**, 735-756, 2001.

[45] J. T. Oden and S. Prudhomme, "New Approaches to Error Estimation and Adaptivity for the Stokes and Oseen Equations", *Int. J. Numer. Fluids*, **31**, 3-15, 1999.

[46] J. T. Oden, S. Prudhomme and L. Demkowicz, "A Posteriori Error Estimation for Acoustic Wave Propagation Problems", *Arch. Comput. Meth. Engrg.*, **12**(4), 343-389, 2005.

[47] M. Paraschivoiu, "A posteriori Finite Element Output Bounds in Three Space Dimensions using the FETI method", *Comput. Methods Appl. Mech. Engrg.*, **190**, 6629-6640, 2001.

[48] M. Paraschivoiu and A.T. Patera, "A Hierarchical Duality Approach to Bounds for the Outputs of Partial Differential Equations",*Comput. Methods Appl. Mech. Engrg*, **158**, p. 389-407, 1998.

[49] M. Paraschivoiu and A.T. Patera, "A Posteriori Bounds for Linear-Functional Outputs of Crouzeix-Raviart Finite Element Discretization of the Incompressible Stokes Problem", *Int. J. Numer. Fluids*, **32**(7), 823-849, 2000.

[50] M Paraschivoiu, J Peraire and A T Patera, "A posteriori Finite Element Bounds for Linear-Functional Outputs of Elliptic Partial Differential Equations",*Comput Methods Appl Mech Engrg*, **150**, p 289-312, 1997

[51] N Pares, J Bonet, A Huerta and J Peraire, "The Computation of Bounds for Linear Functional Outputs of Weak Solutions to the Two-Dimensional Elasticity Equations", *Comput Methods in Appl Mech Engrg* , **195**(4-6), p 406-429, 2006

[52] N Pares, H Santos and P Diez,"Guaranteed Energy Error Bounds for the Poisson equation using a Flux-Free Approach Solving the Local Problems in Subdomains", *Int J Numer Methods Engrg* , **79**(10), p 1203 1244, 2009

[53] S Prudhomme and J T Oden,"On Goal-Oriented Error Estimation for Elliptic Problems application to the control of pointwise errors", *Comput Methods Appl Mech Engrg*, **176**, 313-331, 1999

[54] D R Raichel, *The Science and Applications of Acoustics*, second ed , Springer Verlag, New York, 2000

[55] B D Reddy, *Introductory Functional Analysis with applications to Boundary Value Problems and Finite Elements*, Springer-Verlag, New York, 1998

[56] Y Saad, *Iterative Methods for Sparse Linear Systems*, second ed , SIAM, Philadelphia, PA, 2003

[57] J Sarrate, J Peraire and A Patera, "*A Posteriori* Finte Element Error Bounds For Non-Linear Outputs of the Helmholtz Equation",*Int J Meth Fluids*, **31**, 17-36, 1999

[58] A. M. Sauer-Budge, J. Bonet,A. Huerta and J. Peraire, "Computing Bounds for Linear Functionals of Exact Weak Solutions to Poisson's Equation", *SIAM Journal on Numerical Analysis*, **42**(4), 1610-1630, 2004.

[59] A. M. Sauer-Budge and J. Peraire, "Computing Bounds for Linear Functionals of Exact Weak Solutions to the Advection-Diffusion-Reaction Equation",*SIAM Journal on Scientific Computing*, **26**(2), 636-652, 2004.

[60] A. H. Schatz, "An Observation Concerning Ritz-Galerkin Methods with Indefinite Bilinear Forms", *Math. Comp.*, **128**, 959-962, 1974.

[61] J. R. Stewart, "Adaptive Finite Element Method for the Helmholtz equation in Exterior Domains", Ph.D. Dissertation, Division of Applied Mechanics, Stanford University, Stanford, CA, 1995.

[62] B. Fraeijs De Veubeke, "Displacement and equilibrium models in the finite element method" In Zienkiewicz and Holister, editors, *Stress Analysis*, Wiley London, 1965.

[63] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.

[64] NE. Wiberg and P. Díez, "Adaptive modeling and simulation", *Comp. Methods Appl. Mech. Engrg.*, **195**(4-6), p. 205-480 (2006).

[65] OC. Zienkiewicz and JZ. Zhu, "A simpler error estimator and adaptive procedure for practical engineering analysis", *Int. J. Numer. Methods Engrg.*, **24**, p. 337-357, 1987.