FINITE ELEMENT ERROR ANALYSIS: AN ELEMENTARY APPROACH

Thesis submitted in partial fulfillment of the requirements

for the award of the degree of

MASTER OF SCIENCE

IN

(MATHEMATICS AND COMPUTING)

submitted by

Rinkal Sachdeva

Roll No: 301103017

Under the supervision of

Dr. Vivek

SMCA,

Thapar University, Patiala

SCHOOL OF MATHEMATICS AND COMPUTER APPLICATIONS
THAPAR UNIVERSITY
PATIALA-147001 (PUNJAB)
JULY, 2013.

CERTIFICATE

I hereby certify that the dissertation entitled ***"FINITE ELEMENT ERROR ANALYSIS: AN ELEMENTARY APPROACH ",*** *which is being submitted by* ***Ms. Rinkal Sachdeva (Roll no. 301103017),*** *in the partial fulfillment of the requirements for the award of degree of* ***MASTER OF SCIENCE*** *in* ***"Mathematics and Computing"****, to the School of Mathematics and Computer Applications (SMCA), Thapar University, Patiala, comprises of candidate's authentic record of the work studied under the supervision of Dr. Vivek, Assistant Professor, SMCA, Thapar University, Patiala, during the period from January 2013 to June 2013.*

The part of the work presented in this dissertation has not been submitted either in part or in full to this or any other university for the award of any degree by the author.

Rinkal.
(Rinkal Sachdeva)

This is to certify that the above statement made by the candidate is correct and true to the best of our knowledge.

V. Sandhu 15-7-13

Dr. Vivek
Assistant Professor,
SMCA, Thapar University,
Patiala

Countersigned by:

Dr. Rajesh Kumar
Associate Professor and Head,
SMCA, Thapar University, Patiala

Dr. S.K. Mohapatra
Dean of Academic Affairs,
Thapar University, Patiala

DEDICATED

TO

GOD, MY TEACHERS AND MY PARENTS

Contents

Acknowledgements

I wish to express my deep sense of gratitude to my supervisor, Dr. Vivek, Assistant Professor, School of Mathematics and Computer Applications, Thapar University, Patiala, for his support and valuable guidance throughout the course of this dissertation.

I am also grateful to Dr. Rajesh Kumar, Associate Professor and Head, School of Mathematics and Computer Applications, Thapar University, Patiala, for providing all necessary facilities in the department.

I am deeply thankful to Dr. P.K Bajpai, Dean, Research and Sponsored Projects and Dr. S.K. Mohapatra, Dean, Academic Affairs, Thapar University, Patiala, for the support and needful help during the various stages of this dissertation.

Special thanks are gratefully given to my parents for their encouragement and blessings to me at all stages during my studies.

Above all, I pay my reverence to the almighty of GOD.

Rinkal Sachdeva
301103017

ABSTRACT

The present dissertation entitled, **"Finite Element Error Analysis: An Elementary Approach"**, embodies a brief account of investigations carried out by various authors on finite element error analysis under the supervision of Dr. Vivek, Assistant Professor, School of Mathematics and Computer Applications, Thapar University, Patiala. The work presented in this dissertation has been divided into three chapters. The aim of this work is to study some results on error bounds in finite element methods. In the very first chapter there is a brief introduction about the finite element methods. Differential equations arise in the mathematical modelling of many physical, chemical and biological phenomena and many more areas of science and engineering. Asymptotic and numerical are two principle approaches for solving differential equations. In the present study, the numerical techniques have been used to approximate the solution. As there are many numerical methods to find an approximate solution such as finite difference methods(FDM), finite element methods(FEM), finite volume methods(FVM), boundary element methods(BEM), etc. The present work consists of finite element analysis. In the first chapter, finite element method has been discussed. The finite element method is one of the most powerful method known for finding the numerical solutions. In the first chapter, the method is outlined with the help of a simple example of finding the area of a circle. Then finite element errors are discussed. The errors can be categorized into two types. One is a priori error estimates and the other is a posteriori error estimates. Then the one dimensional model problem

$$L\phi = -a\phi_{xx} + b\phi_x + c\phi = f(\mathrm{x}) \quad \text{in} \quad \Omega =]0, 1[$$

with homogeneous Dirichlet boundary conditions, i.e.

$$\phi(0) = g_0 = 0$$

$$\phi(L) = g_L = 0$$

has been discussed for which error bounds have been discussed. For suitable choices of the coefficients, the model problem presented in Section 6 of Chapter 1 allows us to consider different types of operators, including the advection-diffusion and Helmholtz operators.

In the second chapter, a priori error estimates are discussed in detail. A priori estimates are not computable because they are calculated before the computed solution is known. Also results on error bounds for different operators are discussed. The chapter is again decomposed into two sections namely, a priori error estimates in energy norm (for symmetric and nonsymmetric operators) and a priori estimates in the L_2 norm (for positive definite and indefinite operators). In chapter 2, firstly the simplest case -the energy norm for positive definite operators has been discussed. Also, both kind of symmetric and nonsymmetric operators are considered. A priori error estimates, which can be obtained using different techniques for these types of operators, are presented in this chapter. Again, a priori estimates require different techniques for these two cases. Positive-definite operators are considered in Section 3.1, and the indefinite case-specifically the Helmholtz equation-is considered in Section 3.2 of Chapter 2.

In the third chapter, a posteriori error estimates, which do not depend on symmetry of the operator, are discussed. A posteriori estimates refers for those error estimates which are obtained once the computed solution is known and hence are computable. This chapter is again decomposed into two sections namely, a posteriori error estimates in energy norm and a posteriori estimates in the L_2 norm. The L_2 norm is then considered for both positive definite and indefinite operators. A posteriori estimates, which do not

depend on the definiteness of the operator, have been derived in Section 3. In all cases, only the Galerkin method is analyzed.

Chapter 1

INTRODUCTION

1.1 Differential Equations

A differential equation can be described as a relation between dependent variables, independent variables and the derivatives of the dependent variables with respect to the independent variables. Differential equations arise in the mathematical modelling of many physical, chemical and biological phenomena and many more areas of science and engineering such as fluid dynamics, electromagnetism, material science, astrophysics, economy etc.

Differential equations can be categorized in two categories:

(1) Ordinary differential equations

(2) Partial differential equations

An ordinary differential equation (ODE) is a differential equation in which the unknown function is a function of a single independent variable. A partial differential equation (PDE) is a differential equation in which the unknown function is a function of more than one independent variables and their partial derivatives.

1.2 Solution Methodology

Differential equations are mathematically studied from several different perspectives, mostly concerned with their solutions, functions that make the equation hold true. For most of the differential equations with complex or transcidental variable coefficients, non-linear differential equations, it is very much difficult to find the exact solution. Many properties of solutions of these differential equations may be determined without finding their exact form. If a self-contained formula for the solution is not available, the solution may be numerically approximated. Thus we can say, there are two principle approaches for finding the solutions of differential equations. One is the asymptotic approach and the other one is numerical approach.

In the asymptotic approach, one finds a solution of a differential equation by using the properties and the nature of the problem. These methods require the problem solver to have some apriori knowledge of the solution expected. But very frequently the differential equations under consideration are so complicated that finding their solutions by purely analytical means (e.g. by Laplace and Fourier transform methods, or in the form of a power series etc.) is either impossible or impracticable, and one has to seek for numerical approximations to find the approximate solution of the problem. Therefore, for nonlinear or differential equations having complex coefficient or differential equations governing the real life phenomenon or problems having complex domains, it is very much difficult to find the solution using asymptotic approach. In this regard, numerical approaches have benefits over asymptotic approaches. In numerical approach, one finds a solution of a differential equation numerically. These methods does not require the problem solver much information of the solution expected. Also, these methods are comparatively very easy to implement as against to asymptotic methods. Some of the numerical techniques include Finite Difference Methods, Finite Element Method, Finite Volume Method etc.

1.3 Brief History of Finite Element Methods

Though, there are many numerical methods to solve differential equations, among these techniques, a particular class of numerical techniques namely "Finite Element Methods" is widely used for for approximating exact solution of differential equations. These methods have been applied to number of physical problems for which the governing differential equations are available. The finite element method is a good choice for solving differential equations over complex domains (like cars and oil pipelines), when the domain changes (as during a solid state reaction with a moving boundary), or when the desired precision varies over the entire domain, or when the solution lacks smoothness. Around 40 to 50 years back, people use analytical methods to find the solution of a mathematical problem. But together with the overall development, the governing differential equations modelling the real life phenomenon also got complicated. The researchers have to move towards the numerical approaches to get approximate solutions of the differential equations. At that time finite element method came into existence. Finite element method was originally suggested by the famous mathematician Courant in 1943. But it was later developed by the engineers in early 60's. Since then finite element methods have been developed into one of the most general and powerful class of techniques for the numerical solution of differential equations and are widely used in engineering design and analysis. The term finite element was first coined by Clough in 1960. In the early 1960's, engineers used finite element methods(FEM) for approximating the solutions of problems in stress analysis, fluid flow, heat transfer, and in other areas. The first book on the FEM by Zienkiewicz and Chung was published in 1967. In the late 1960's and early 1970's, the FEM was applied to a wide variety of engineering problems. Most commercial FEM software packages originated in the 1970's like Abaqus, Adina, Ansys, etc.

1.4 Finite Element Methods:

The basic steps of finite element methods in computing an approximate solution of any problem consists :

1. Finite element discretization
2. Element equations
3. Assembly of element equations and solution
4. Convergence and error estimate

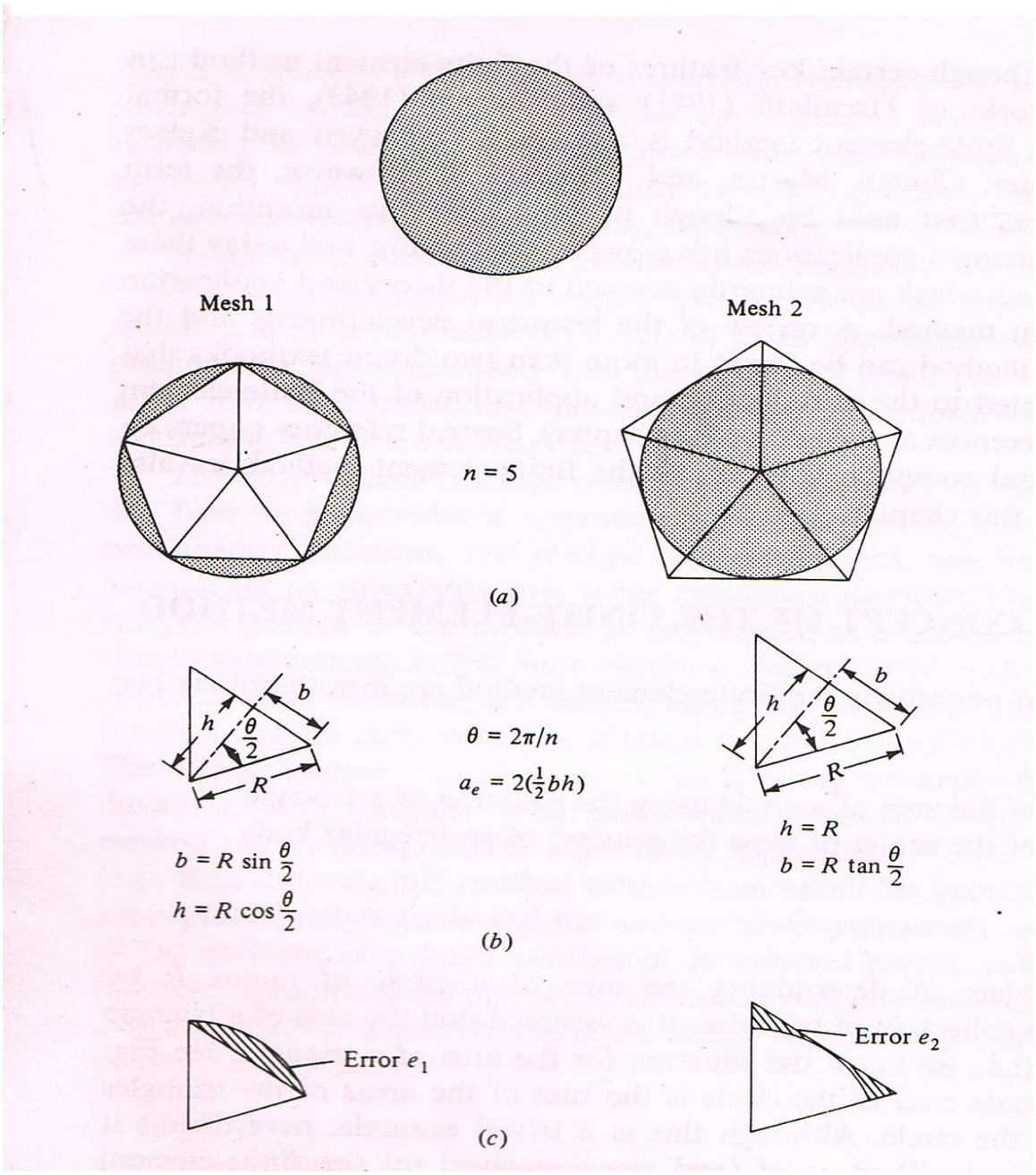
The basic ideas underlying the finite element method are introduced via one example [25]:

Consider the problem of determining the area of a circle of radius R by discretizing it as a collection of triangles. It is assumed that the area of the triangle can be calculated. The approximate area of the circle is the sum of the areas of the triangles used to represent the circle. Here we discuss the basic steps involved while finding finite element solution of any problem.

1. **Finite element discretization**: First, the continuous region is represented as a collection of n subregions, say triangles. This is called discretization of the domain. Each subregion is called an element. The collection of elements is called the finite element mesh. In our case, we discretize the circle into a mesh of five triangles. Since all the elements are of same size, the mesh is said to be uniform. We consider two discretizations of the circle say mesh 1 and mesh 2.

2. **Element equations**: A typical element from the discretized mess is selected and its properties (i.e. area in the present case) or equations are computed. It is here that we bring in the governing equation (i.e., the equation for computing the area) of the element to calculate the required property. Let a_e be the area of element e in mesh 1 and $\overline{a}_e$ be the area of element e in mesh 2. Thus, for the for element e,

-


Mesh 1
Mesh 2
n = 5
(a)
h
b
θ/2
R
θ = 2π/n
a_e = 2(½bh)
b = R sin θ/2
h = R cos θ/2
h = R
b = R tan θ/2
(b)
Error e1
Error e2
(c)

[h]

we have

$$a_e = \frac{R^2}{2} \sin\left(\frac{2\pi}{n}\right) \tag{1.4.1}$$

$$\overline{a}_e = R^2 \tan\left(\frac{\pi}{n}\right) \tag{1.4.2}$$

where R is the radius of the circle. The above equations are called element equations.

3. **Assembly of element equations and solutions**: The approximate area of the circle is obtained by putting together the element property of all the elements in the mesh, this is called assembly of the element equations. The assembly is based, in the present case, on the simple idea that the total area of the assembled elements is equal to the sum of the areas of all the individual elements of the mesh:

$$A_1 = \sum_{e=1}^{n} a_e \tag{1.4.3}$$

$$A_2 = \sum_{e=1}^{n} \overline{a}_e \tag{1.4.4}$$

Since the mesh is uniform, a_e *or* $\overline{a}_e$ is the same for each of the elements in the mesh, and we have

$$A_1^{(n)} = \left\{ n\frac{R^2}{2} \ \sin\left(\frac{2\pi}{n}\right) \right\} \tag{1.4.5}$$

$$A_2^{(n)} = nR^2 \ \tan\left(\frac{\pi}{n}\right) \tag{1.4.6}$$

4. **Convergence and error estimate**: For this problem we know the exact solution of the problem: $A_0 = \pi R^2$. We can estimate the error in the approximation and show that the approximate solution converges to the exact solution in the limiting case as $n \to \infty$. Consider the typical element e. The error in the approximation is equal to the difference between the area of the sector and that of the triangle (see fig 1).

$$e_1 = |S_e - a_e| \tag{1.4.7}$$

$$e_2 = |S_e - \overline{a}_e| \tag{1.4.8}$$

where $S_e = \frac{1}{2}R^2\theta$ is the area of the sector. Thus, the error estimates for an element in the meshes 1 and 2 are given by

$$e_1 = R^2\left(\frac{\pi}{n} - \frac{1}{2}\sin\left(\frac{2\pi}{n}\right)\right) \tag{1.4.9}$$

$$e_2 = R^2\left(\tan\left(\frac{\pi}{n}\right) - \frac{\pi}{n}\right) \tag{1.4.10}$$

The total error (called global error) is given by multiplying the $e_i\,'s$ by n:

$$\begin{aligned} E_1^{(n)} &= R^2(\pi - \frac{n}{2}\sin\frac{2\pi}{n}) \\ &= \pi R^2 - A_1^{(n)} \end{aligned} \tag{1.4.11}$$

$$\begin{aligned} E_2^{(n)} &= R^2\left(n\tan(\frac{\pi}{n} - \pi\right) \\ &= A_2^{(n)} - \pi R^2 \end{aligned} \tag{1.4.12}$$

We now show that E_1 and E_2 go to zero as $n \to \infty$. Letting $x = 2/n$, we have

$$\begin{aligned} A_1^{(n)} &= \frac{R^2 n}{2}\sin\left(\frac{2\pi}{n}\right) \\ &= R^2\sin\left(\frac{\pi x}{n}\right) \end{aligned} \tag{1.4.13}$$

$$\begin{aligned} \lim_{n\to\infty} A_1^{(n)} &= \lim_{n\to 0} R^2\sin\left(\frac{\pi x}{n}\right) \\ &= \lim_{n\to 0} \pi R^2\cos(\pi x) \end{aligned} \tag{1.4.14}$$

Similarly, letting $y = 1/n$, we have

$$\lim_{n\to\infty} A_2^{(n)} = \lim_{y\to 0} R^2\frac{\tan(\pi y)}{y} \tag{1.4.15}$$

$$= \lim_{y\to 0} \pi R^2\sec^2(\pi y) \tag{1.4.16}$$

$$= \pi R^2 \tag{1.4.17}$$

Hence, $E_1^{(n)}$ *and* $E_2^{(n)}$ are zero as $n \to \infty$. This completes the proof the convergence.

Also, one can use either rectangles or any other polygon. The approximation error in each case is different and therefore so is the solution. Note that the equation for the element area is used in its exact form. Therefore, no approximation error in the solution of the equations is introduced.

1.4.1 Some Remarks on the Example

Although the above example gives the basic idea of the finite element method, there are several other features which are discussed below[25]:

1. One can discretize the domain, depending on the shape of the domain, into a mesh of more than one type of element. For example, in the approximation of the circle, one can use either rectangles only or a combination of triangles and rectangles. Note that a triangular mesh or a mesh of combined triangles and rectangles represents the circle more closely than a corresponding number of rectangular elements.

2. If more than one type of element are used in the representation of the domain, one of the each kind should be isolated and its properties developed.

3. The governing equations are generally more complex than those considered in the above example. They are usually differential equations. In many cases, the equations cannot be solved over an element for two reasons. First, the equation do not permit the exact solution. It is here that the variational methods come into play. Second, the discrete equations obtained in the variational methods cannot be solved independent of the remaining elements because the assemblage of the elements is subjected to certain boundary or initial conditions.

4. There are two main differences in the form of the approximate solution used in the finite element method and that used in the variational method (i.e. variational method applied to whole domain). First, instead of representing the solution u as a linear combination ($u = \sum c_j \phi_j$) in terms of arbitrary parameter c_j in the variational methods, in the finite element method the solution is represented as a linear combination ($u = \sum u_j \psi_j$) in terms of the values of u_j of u at the nodal points. Second, the approximate functions in the finite element method are polynomials that are derived using interpolation theory.

5. The number and the location of the nodes in an element depend on

   (a) the geometry of the element,

   (b) the degree of the approximation(i.e. the degree of the polynomials), and

   (c) the variational form of the equation or equations. By representing the required solution in terms of its values at the nodes, one obtains directly the approximate solution at the nodes.

6. The assembly of elements, in general, is based on the idea that the solution is continuous at the inter element boundaries. In the example, the continuity conditions were not present because the equations used were algebraic equations.

7. In general the assemblage of elements is subjected to boundary and / or initial conditions. The discrete equations associated with the finite element mesh are solved only after the boundary and/or initial conditions have been imposed.

8. There are three sources of error in the finite element solution:

   (a) errors due to the approximation of the domain (this was the only error present in the example),

   (b) errors due to the approximation of the solution, and

   (c) errors due to numerical computation (e.g. numerical integration and round-off errors in a computer).

   The estimation of these errors is not a simple matter. However, under certain conditions they can be estimated over a given element and hence for the problem.

9. The accuracy and convergence of the finite element solution depends on the differential equation solved and the element used. Accuracy means the difference between the exact solution and the finite element solution, and convergence refers to the accuracy as the number of elements in the mesh increased. The nature of the convergence depends on the formulation of the governing equations.

In the work presented, the error estimates for the finite element method has been discussed. Finite element error estimates, in general can be classified into two categories namely a priori error estimates and a posteriori error estimates. A priori error estimates refers for those kind of error estimates which are obtained before the solution is known. This type of estimate is useful for finite element method designing, and is valuable for determining the asymptotic rate of convergence of the method as well as for finding dependencies on problem parameters. A priori estimates are not computable because these are obtained before the exact solution is known. The second broad class of error estimates is a posteriori error estimates which refers for those kind of error estimate which are obtained after the solution is known. These types of estimates are computable since they are expressed in terms of the known finite element solution. They are most useful for adaptivity and for controlling the solution error.

A posteriori estimators can, in turn, be categorized into two main types. The first of these is stress recovery, which is also referred to as a postprocessing or flux-projection technique. It was proposed in the context of linear elliptic problems by Zienkiewicz and Zhu [31, 32] (this is commonly referred to as the Z^2 estimator). The second is the class of residual-based estimators, which has two subtypes: Explicit and Implicit. The literature covering these topics is vast. The implicit type of a posteriori error estimators are described in [1, 2, 3, 8, 21, 22, 23, 24], the explicit type error estimators are described in [7, 10, 11, 19, 17, 18, 20, 27, 30].

Because it is analogous in form to a priori error estimates, the explicit residual-based a posteriori error estimator is presented in the chapters to follow. Explicit residual-based a posteriori error estimates and a priori error estimates each consist of an upper bound on the error measured in some norm, one might expect that the derivation of these types of error estimators is also analogous, and this is indeed the case. In the chapters to follow, the structure of these proofs, focusing on the similarities between the a priori

and a posteriori bounds has been presented. The emphasis is on presenting a simplified structure of the proofs. The error bounds can be derived, for the most part, by following a few basic steps. As will be shown, the steps depend on the properties of the differential operator and the norm chosen for the error.

Let e be the finite element error, and $|||\cdot|||$ denote a norm on $\Omega =]0, L[$. The general form of a priori error estimates is

$$|||e||| \leqslant C(\Omega, p) h^{f_1(p,r,s)} \parallel \phi \parallel_{H^q(\Omega)} \tag{1.4.18}$$

where h is nondimensional measure of element length, p is the highest order complete polynomial in the element shape functions, C is a constant depending on p and the domain, ϕ is the exact solution with regularity r, $\quad s$ is determined by the choice of norm $|||\cdot|||$ and $H^q(\Omega)$ denotes a sobolev space with q a function of p and r.

The general form of the explicit residual based a posteriori error estimates is

$$|||e||| \leq ||C_1(\Omega, p) h^{f_2(s)} r^h||_{L_2(\Omega)} + ||C_2(\Omega, p) h^{f_2(s)-1} R^h||_{L_2(\Omega)} \tag{1.4.19}$$

where instead of the exact solution ϕ appearing, residuals of the Euler lagrange equation appear. The residual on the element interior $\widetilde{\Omega}$ is denoted by r^h, and the residual on element boundaries(involving 'jumps' in the normal derivatives of ϕ) is denoted by R^h. C_1 *and* C_2 are constants, which depend on p and the domain. Although the similarities between (1.4.18) and (1.4.19) are obvious, there are some important technical differences, which are discussed below.

REMARKS:

1. The a posteriori estimates are computable, since they are expressed in terms of residuals. On the other hand, a priori estimates are expressed in terms of the exact solution, which is generally not known.

2. The two terms on the right-hand side of (1.4.19) emanate from the fundamentally different form the residuals take on element interiors and element boundaries. The residual on the element boundaries is due to the C^0 continuous finite element shape functions.

3. The norms on the right-hand side of (1.4.19) are specifically L_2 norms. Furthermore, they are computed only on the union of element interiors, donated by Ω. This is even true for the boundary term, since R^h is the element boundary residual lumped onto the element interior.

4. The exponent of h in (1.4.19) depends (explicitly) only on the choice of the $|||\cdot|||$ norm. However, the residual (in the L_2 norm) depends asymptotically on h raised to some power $\alpha(p,r)$. Therefore, the exponent of h is implicitly a function s, p and r.

5. The constants C_1, C_2 and the h terms appear inside the norms inside the a posteriori estimate. This is due to the nature of the norms and the way in which they are computed. Since they only involve element interiors, they are computed by summing the element contributions. In general, adaptive meshes contain elements with different h and p, which must be accounted for in computing the norms. This distinction plays an important role in the design of hp adaptive strategies. For h refinement only, the constants can be pulled out of the norms.

The first choice one needs to make is the norm ($|||\cdot|||$) in which to estimate the error. For positive definite operators, the energy norm is a natural and convenient choice. For indefinite operators (as well as definite ones), L_p norms and H^q norms may be used, where $1 \leqslant p \ \leqslant \infty \ and \ 0 \leqslant q \leqslant \infty$. The simplest of these is the L_2 norm (which corresponds to the H^0 norm). It is also possible to have $q < 0$, which defines certain dual norms. Other norms can also be defined, for example, to include terms appearing in the discretization

formulation which must be controlled, or to account for a specific part of the operator, such as the skew symmetric part. We take $|||\cdot|||$ to represent either the energy norm (for positive definite operators only) or the L_2 norm (for any operator).

1.5 Some Important Definations

$$
\begin{aligned}
L_p - norm \qquad & ||v||_p = \Big(\int_0^L |v(x)|^p \Big)^{1/p}, \qquad 0 < p < \infty \\
L_2 - norm \qquad & ||v||_2 = \Big(\int_0^L |v(x)|^2 \Big)^{1/2}, \qquad 0 < p < \infty \\
L_\infty - norm \qquad & ||v||_\infty = \sup_{x \in [0,L]} |v(x)| \\
Energy - norm \qquad & ||v||_E = \Big(\int_0^1 a(x) |v\prime(x)|^2 \Big)^{1/2}.
\end{aligned}
$$

1.6 Model Problem

In this Section, a one-dimensional model problem has been presented and its finite element formulation has been discussed. This model problem will used in the next two Chapters to follow for deriving the a priori and a posteriori error estimates.

Consider the following one-dimensional model problem [28]:

$$L\phi = -a\phi_{xx} + b\phi_x + c\phi = f(x) \quad \text{in} \quad \Omega =]0, L[\tag{1.6.1}$$

with homogeneous Dirichlet boundary conditions

$$\phi(0) = g_0 = 0 \tag{1.6.2}$$

$$\phi(L) = g_L = 0. \tag{1.6.3}$$

It is assumed $f \in L_2$, which implies $\phi \in H^2(\Omega)$. The coefficients a, b and c are real constants, where $a > 0,\ b \geq 0$ and c can be of any sign. Note that complex solutions are admitted for certain choices of a, b, c, f and boundary conditions.
The model problem can also be written in Sturm-Liouville form. Dividing (1.6.1) by a and multiplying by the integration factor $P(x) = e^{-(b/a)x}$, we get

$$-(P(x)\phi_x)_x + Q(x)\phi = F(x) \tag{1.6.4}$$

where

$$Q(x) = \frac{c}{a}P(x)$$
$$F(x) = \frac{1}{a}f(x)P(x).$$

This gives an operator which is self-adjoint.

The weak formulation of (1.6.1) is given by
Find $\phi \in S$ such that $\forall\ \ w \in V$,

$$B(w, \phi) = L(w) \tag{1.6.5}$$

where

$$B(w, \phi) = a(w_x, \phi^x) - b(w_x, \phi) + c(w, \phi) \tag{1.6.6}$$

$$L(w) = (w, f), \tag{1.6.7}$$

where $(\cdot, \cdot) : V \times S \longrightarrow C$ is the $L_2(\Omega)$ inner product, with the first argument conjugated if necessary. S and V are the spaces of trial solutions and weighting functions respectively and are defined as

$$V = \{\phi : \phi \in H^1(\Omega) : \phi \tag{1.6.8}$$

where ϕ satisfies homogeneous conditions on Dirichlet boundaries.

$$V = \{\phi : \phi \in H^1(\Omega) : \phi \tag{1.6.9}$$

where ϕ satisfies Dirichlet boundaries conditions. that is, these spaces consist of functions in $H^1(\Omega)$; members of S satisfy the Dirichlet boundary conditions, while those in V satisfy homogeneous conditions on Dirichlet boundaries (in the present case, due to (1.6.2) and (1.6.3), S = V). Note that if the model problem included Neumann or Robin boundary conditions, in that case $B(w, \phi)$ and $L(w)$ operators would need to be appropriately modified.

The Galerkin formulation is obtained from (1.6.5) by restricting the trial solutions and weighting functions to suitable finite dimensional subspaces $S^h \subset S$ and $V^h \subset V$ respectively. The finite dimensional Galerkin weak formulation is defined as

Find $\phi^h \in S^h$ such that $\forall \;\; w^h \in V^h$,

$$B(w^h, \phi^h) = L(w^h) \tag{1.6.10}$$

Here only the Galerkin method is considered for estimating the error bounds. Clearly, if $c \geq 0$, $B(w, \phi)$ is positive definite, in which case (1.6.5) (equivalently (1.6.1)) has a unique solution. Furthermore, if $b = 0$, $B(w, \phi)$ is symmetric. Positivity is lost if $c < 0$. In this case, for c an eigenvalue and if certain boundary conditions are imposed, nonunique solutions are admitted. For the homogeneous problem $i.e.(f = 0)$, two examples of such boundary conditions are: (1) the ordinary case, with homogeneous Robin boundary conditions, i.e.

$$a_1\phi(0) + a_2\phi_x(0) = 0$$
$$b_1\phi(L) + b_2\phi_x(L) = 0$$

where a_1, a_2, b_1 and b_2 are constants, and
(2) the periodic case, i.e.

$$\phi(0) = \phi(L)$$
$$\phi_x(0) = \phi_x(L).$$

These conditions (both cases) lead to solutions which are orthogonal with respect to the function $1/a\ P(x)$. We will assume that (1.6.1) always has a unique solution, which is equivalent to assuming $|c|$ is not an eigenvalue. This implies the following stability condition:

$$||\phi||_2 \leq C||f|| \tag{1.6.11}$$

where C is a constant, $||\cdot||_2$ denotes the H_2 norm, and $||\cdot||$ denotes the $L_2(\Omega)$ norm. Therefore, if $f = 0$, only the trivial solution $\phi = 0$ is allowed. This stability condition plays a critical role in the proofs of error estimates in the L_2 norm.

The Helmholtz equation is obtained from (1.6.1) by the substitutions $a = 1, b = 0$ and $c = -k^2$, where k is the wave number. This leads to

$$-(\phi_{xx} + k^2\phi) = f(x) \quad in\ \Omega =]0, L[. \tag{1.6.12}$$

If $k^2 > 0 (i.e.\ c < 0)$, propagating solutions are obtained. Note that in this case the operator is indefinite. If $k^2 < 0$, the wave number is imaginary and the solution is decaying. Comparison of (1.6.12) to (1.6.4) indicates that the Helmholtz operator (again, ignoring boundary conditions) is self-adjoint.

Nonsymmetric operators are obtained from (1.6.1) by choosing $b \neq 0$. For example, the advection-diffusion model equation is obtained by the substitutions $a = \kappa, b = u$ and $c = 0$, where κ is the diffusivity and u is the velocity. This leads to, after rearranging the terms,

$$u\phi_x = \kappa\phi_{xx} + f \quad in\ \ \Omega =]0, L[. \tag{1.6.13}$$

The advection term represents the skew symmetric part of the operator, while the diffusion term is symmetric. Although this operator is positive definite, ill-conditioning (i.e. virtual loss of positivity) may occur for large values of the Peclet number, $Pe = uL/\kappa$. Such a situation corresponds to advection dominated flow, in which the skew symmetric part of the operator dominates the symmetric part. The skew symmetric part contributes

nothing to positivity or, equivalently, numerical stability, in a Galerkin or central difference context.

By suitable choices of the coefficients, the model problem presented in Section 6 allows us to consider different types of operators, including the advection-diffusion and Helmholtz operators which has been discussed above. In Chapter 2, firstly the simplest case -the energy norm for positive definite operators has been discussed. Also, symmetric and nonsymmetric operators are considered for both. A priori estimates, which can be obtained using different techniques for these types of operators, are presented in this chapter. Again, a priori estimates require different techniques for these two cases. Positive-definite operators are considered in Section 3.1, and the indefinite case-specifically the Helmholtz equation-is considered in Section 3.2 of Chapter 2.

A posteriori estimates, which do not depend on symmetry of the operator, are presented in Chapter 3. The L_2 norm is then considered for both positive definite and indefinite operators. A posteriori estimates, which do not depend on the definiteness of the operator, are presented in Section 3 for the L_2 norm. In all cases, only the Galerkin method is analyzed.

Chapter 2

A PRIORI ERROR ESTIMATES

2.1 Introduction

In this chapter, the goal is to find the bounds for the error $u - u_h$ in the finite element approximation of the solution u to the general problem. In such estimates the error analysis gives information about the size of the error, depending on the (unknown) exact solution u, before any computational steps. The first step is how to measure the error, that is, which norm to use. It has already demonstrated that the energy norm $||.||$ is more convenient than the other norms.

2.2 A Priori Estimates in Energy norm

In this Section, the a priori bounds on the error in the computed solution and the exact solution has been derived in the energy norm. The objective has been achieved by categorizing the operator as symmetric and nonsymmetric operators.

2.2.1 Symmetric operators

Symmetric operators are obtained from the model equation (1.6.1) defined in Chapter 1 by setting $b = 0$. For an energy norm, the operator must be positive definite. Therefore, we consider $c \leqslant 0$. The energy norm is defined by

$$||w||_E^2 = |B(w, w)| \tag{2.2.1}$$

Now using Cauchy-Schwarz inequality on this norm, we get

$$|B(w, v)| \leqslant ||w||_E ||v||_E \tag{2.2.2}$$

REMARK. The Cauchy-Schwarz inequality is valid in this form since $B(w, v)$ is assumed to be symmetric.

Now introduce the following splitting of the error, which will be useful for the proofs. Let $\widetilde{\phi}_h$ be the nodal interpolant of the exact solution ϕ. Let e denote the finite element solution error,

$$\begin{aligned} e &= \phi_h - \phi && (2.2.3)\\ &= \phi_h - \widetilde{\phi}_h + \widetilde{\phi}_h - \phi \\ &= e_h + \eta && (2.2.4) \end{aligned}$$

where

$$e_h = \phi_h - \widetilde{\phi}_h \tag{2.2.5}$$

is the portion of the error in V_h, and

$$\begin{aligned} \eta &= \widetilde{\phi}_h - \phi \\ &= e - e_h && (2.2.6) \end{aligned}$$

is the interpolation error and is a member of V.

From Galerkin Orthogonality property,

$$a(u, v) = f(v) \tag{2.2.7}$$

Since $v^h \subseteq V$, so (2.2.7) will also hold for $\forall\ u_h \in V^h$, i.e

$$a(u, v_h) = f(v_h) \tag{2.2.8}$$

Also, from finite element weak formulation, we have

$$a(u_h, v_h) = f(v_h) \qquad \forall\ \ v_h \in V^h \tag{2.2.9}$$

Subtracting (2.2.9) from (2.2.8), we get

$$a(u - u_h, v_h) = 0 \tag{2.2.10}$$

Thus, Galerkin orthogonality property is,

$$B(e, e_h) = 0, \tag{2.2.11}$$

Using the definition of $||\cdot||_E$, given by (2.2.1), and inserting $e \in V$ into both slots of $B(\cdot, \cdot)$ and using (2.2.4), we get

$$\begin{aligned} ||e||_E^2 &= |B(e, e)| \\ &= |B(e, e^h + \eta)| \\ &= |B(e, e^h) + B(e, \eta)| \\ &= |B(e, \eta)| \\ &\leqslant ||e||_E ||\eta||_E \qquad \text{(Cauchy-Schwarz inequality)} \end{aligned} \tag{2.2.12}$$

$$\Rightarrow ||e||_E \leqslant ||\eta||_E \tag{2.2.13}$$

which states the best approximation property with respect to the energy norm. Expanding the right-hand side of (2.2.13) in terms of the model equation (1.6.5) (with $b = 0$ and

$c \geqslant 0$),

$$\begin{aligned} ||\eta||_E &\leq (|B(\eta,\eta)|)^{1/2} \\ &= (|a(\eta_x,\eta_x) + c(\eta,\eta)|)^{1/2} \\ &= (a||\eta_x||^2 + c||\eta||^2)^{1/2} \\ &= \left(\frac{a}{L^2}|\eta|_1^2 + c||\eta||^2\right)^{1/2} \\ &\leq \overline{C}(|\eta|_1^2 + (\eta^2)^{1/2} \\ &= \overline{C}||\eta||_1 \end{aligned} \tag{2.2.14}$$

where $||\cdot||$ denotes the $L_2(\Omega)$ norm, and $||\cdot||_1$ *and* $|\cdot|_1$ denote the $H^1(\Omega)$ norm and seminorm respectively. The constant in (2.2.14) is given by

$$\overline{C} = max\left\{\frac{\sqrt{a}}{L}, \sqrt{c}\right\}.$$

Now applying the standard interpolation estimate (see e.g. [26, 9]) of the form

$$||\eta||_1 \leqslant C_i(p,\Omega)h^p||\phi||_{p+1} \tag{2.2.15}$$

where it is assumed that ϕ has regularity $r \geq p+1$. Substituting into (2.2.14) and letting $C = \overline{C}C_i(p,\Omega)$, we get

$$||e||_E \leqslant Ch^p||\phi||_{p+1} \tag{2.2.16}$$

which is in the same form as (1.4.18).

SUMMARY:

The derivation of (2.2.16) for symmetric, positive definite operators consists of three basic steps:

1. Use the definition of $||e||_E^2$, split the error in the solution slot, then use Galerkin consistency and the Cauchy-Schwarz inequality to cancel the error term on the right-hand side (see (2.2.12)). This leaves a bound in terms of the interpolation error. This is the best approximation error estimates in the energy norm.

2. Again using the definition of $||\eta||_E$, expand the right-hand side to obtain a bound in terms of the $||\eta||_1$ norm (see (2.2.14)).

3. Apply the standard interpolation estimate (2.2.15).

2.2.2 Nonsymmetric operators

Nonsymmetric operators can be systematically decomposed into a symmetric part and a skew symmetric part, i.e.

$$B(w,v) = B_{symm}(w,v) + B_{skew}(w,v) \tag{2.2.17}$$

where

$$B_{symm}(w,v) = \frac{1}{2}(B(w,v) + B(w,v)) \tag{2.2.18}$$

$$B_{skew}(w,v) = \frac{1}{2}(B(w,v) - B(w,v)) \tag{2.2.19}$$

A nonsymmetric operator is obtained from the model equation by taking $b > 0$. Note that we have assumed $c \geqslant 0$ to ensure positivity. Substituting (1.6.6) into (2.2.18) and (2.2.19), we get

$$B_{symm}(w,\phi) = a(w_x,\phi_x) + c(w,\phi) \tag{2.2.20}$$

$$B_{skew}(w,\phi) = b(w,\phi_x) \tag{2.2.21}$$

The proof of the previous Section breaks down for nonsymmetric operators since the Cauchy-Schwarz inequality does not hold in this case. This is because the skew symmetric part of the operator does not contribute to the energy norm, so that if $B(\cdot,\cdot)$ is skew

symmetric,

$$\begin{aligned} ||w||_E^2 &= |B(w,w)| \\ &= |B_{skew}(w,w)| \\ &= \int_0^L w\frac{\partial w}{\partial x}dx \\ &= 0 \end{aligned} \tag{2.2.22}$$

because

$$\begin{aligned} \int_0^L w\frac{\partial w}{\partial x}dx &= w.w|_0^L - \int_0^L \frac{\partial w}{\partial x}.wdx \\ \Rightarrow 2\int_0^L w.\frac{\partial w}{\partial x}dx &= 0. \end{aligned}$$

The approach taken here is to define a dual norm to accommodate the skew symmetric part. This definition is canonical, and allows the proof to proceed in a natural way, along the lines of the preceding derivation for symmetric operators. The skew norm is defined by

$$||w||_{skew} = \sup_{v\in V} \frac{|B_{skew}(w,v)|}{||v||_E}. \tag{2.2.23}$$

Clearly, from (2.2.23),

$$|B_{skew}(w,v)| \leqslant ||w||_{skew}||v||_E. \tag{2.2.24}$$

This inequality is derived directly from the definition of the skew norm. It is similar in form to the Cauchy-Schwarz inequality (2.2.2); however, note the appearance of the symmetric norm (i.e. the energy norm) on the right-hand side.

For finding the a priori error estimates, we begin as in the previous section, with the definition of $||\cdot||_E$, inserting e into both slots of $B(\cdot,\cdot)$. Using (2.2.4) and noting that

from (2.2.11) $B(e^h, e) = 0$ (consistency), we get

$$\begin{aligned}
||e||_E^2 &= |B(e,e)| \\
&= |B(e^h + \eta, e)| \\
&= |B(e^h, e) + B(\eta, e)| \\
&= B|(\eta, e)| \\
&= |B_{symm}(\eta, e) + B_{skew}(\eta, e)| \\
&\leqslant |B_{symm}(\eta, e)| + |B_{skew}(\eta, e)| \\
&\leqslant ||\eta||_E ||e||_E + ||\eta||_{skew} ||e||_{skew} \qquad (2.2.25)
\end{aligned}$$

Therefore,

$$||e||_E \leqslant ||\eta||_E + ||\eta||_{skew} ||e||_E. \qquad (2.2.26)$$

which is identical to (2.2.13) except for the extra skew term. Now using the results from the previous Section for $||\eta||_E$, which allows us to focus attention here on the skew term. Expanding $||\eta||_E$ in terms of the model equation (1.6.1) (with $b > 0$ and $c > 0$, and noting that $|B_{skew}(\eta, v)| = |B_{skew}(v, \eta)|$),

$$\begin{aligned}
||\eta|| &= \sup_{v \in V} \frac{|B_{skew}(\eta, , v)|}{||v||_E} \\
&= \sup_{v \in V} \frac{|B_{skew}(\eta, , v)|}{||v||_E} \\
&= \sup_{v \in V} \frac{|b(v, \eta_x|}{(a(v_x, v_x) + c(v, v))^1/2} \\
&\leqslant \sup_{v \in V} \frac{b||v||||\eta_x||}{(a||v_x||^2 + c||v||^2)^{1/2}} \\
&\leqslant \sup_{v \in V} \frac{b||v||||\eta_x||}{(aC_{PF}^{-2} + c)^{1/2})||v||} \\
&= \overline{C_n}||\eta_x|| \\
&= \frac{\overline{C_n}}{L}|\eta|_1 \\
&\leqslant C_n||\eta||_1 \qquad (2.2.27)
\end{aligned}$$

where C_{PF} is a domain-dependent constant arising from the Poincare-Friedrich inequality

$$||v|| \leqslant C_{PF}||v_x|| \qquad (2.2.28)$$

which holds for functions in V. The constant in (2.2.27) is

$$C_n = \frac{b}{L(aC_{PF}^{-2} + c)^{1/2}}. \tag{2.2.29}$$

Applying the interpolation estimate (2.2.11) to (2.2.23),

$$||\eta||_{skew} \leqslant C_n C_i(p, \Omega) h^p ||\phi||_{p+1} \tag{2.2.30}$$

Finally, substituting (2.2.30) into (2.2.26), and using the estimate for $||\eta||_E$ derived in (2.2.14)-(2.2.16), we get

$$||e||_E \leqslant C_N h^p ||\phi||_{p+1} \tag{2.2.31}$$

where

$$C_N = C + C_n C_i \tag{2.2.32}$$

Comparing (2.2.31) to (2.2.16), it is seen that the skew symmetric part does not destroy the error estimate unless C_n is large. Let us discuss such a situation in the following example.

EXAMPLE: *Model advection-diffusion equation*: If $c = 0$, the above proof proceeds in a slightly different manner, without the need for the Poincare-Friedrich inequality. This case corresponds to the advection-diffusion model equation, upon the substitutions $a = \kappa$ and $b = u$, where κ is the diffusivity and u is the velocity. We now repeat the derivation of (2.2.27).

$$\begin{aligned} ||\eta|| &= \sup_{v \in V} \frac{|B_{skew}(\eta, v)}{||v||_E} \\ &= \sup_{v \in V} \frac{|u(\eta, v_x)}{(\kappa(v_x, v_x)^{1/2})} \\ &\leqslant \sup_{v \in V} \frac{u||\eta||\ ||v_x||}{(\kappa^{1/2}||v_x||} \\ &= \frac{u}{\kappa^{1/2}} ||\eta|| \end{aligned} \tag{2.2.33}$$

Here, notice that a bound in terms of η is obtained, instead of a bound in terms of η_x, when $c \neq 0$. At first glance it appears this may lead to a higher rate of convergence.

However, we still need to consider the symmetric part of the operator. For $c = 0$,

$$||\eta||_E = \frac{\eta^{1/2}}{L}|\eta|_1 \tag{2.2.34}$$

and

$$||e||_E = \frac{\eta^{1/2}}{L}|e|_1 \tag{2.2.35}$$

Thus, the energy norm is equivalent to the H^1-seminorm. Substituting (2.2.33) - (2.2.35) into (2.2.26) and rearranging

$$|e|_1 \leqslant |\eta|_1 + Pe||\eta||$$

where $Pe = uL/\kappa$ is the (nondimensional) Peclet number. Applying the interpolation estimates (2.2.15),

$$|e|_1 \leqslant C_n \leqslant h^P||\phi||_{P+1} \tag{2.2.36}$$

where

$$C_n = C_i[1 + Pe\ h]$$

Note that the term involving the Peclet number comes from the skew symmetric part of the operator. Clearly, to control the error, h needs to be restricted such that

$$hPe = O(1)$$

If h is held fixed, the error will grow as the Peclet number grows, which happens when the flow becomes advection dominated. The Galerkin method will fail in this situation. Stabilized methods (e.g. Galerkin Least-Squares methods [12, 15]) have been developed to maintain accuracy and consistency independent of the Peclet number. The derivation of error estimates for these types of methods can proceed in a similar fashion as shown here; however, additional steps are required.

SUMMARY:

The derivation of (2.2.31) for nonsymmetric, positive definite operators involves four basic steps:

1. Split the operator into its symmetric and skew symmetric parts, and define an appropriate dual norm for the skew symmetric term (see (2.2.17)-(2.2.23)).

2. Using the definition of the skew norm, bound the error (in the energy norm) in terms of the interpolation error η. This bound will have a term due to the symmetric part and a term due to the skew symmetric part (see (2.2.26)).

3. Again using the definition of the skew norm, bound the skew symmetric part in terms of an appropriate H^q norm of the interpolation error (see (2.2.27) for $c \neq 0$, or (2.2.33) for $c = 0$).

4. Combine the result with the bound obtained for the symmetric part, and apply a standard interpolation estimate.

2.3 A Priori Estimates in the L_2 norm

2.3.1 Positive Definite Operators

The procedure for proving error estimates in the L_2 norm involves first proving the error estimates in the energy norm, then performing additional steps to convert them to an L_2 bound. Therefore we will utilize the estimates already derived for the energy norm in the previous Section. Because it is assumed the operator is positive definite, the energy norm exists. Referring to the model equation (1.6.1), positive-definiteness is attained for $c \geqslant 0$ (note that we have already placed the restrictions $a > 0$ and $b \geqslant 0$).

2.3.1.1 Nitsche trick

The procedure for converting an energy norm estimate to an L_2 estimate is known as the 'Nitsche trick. It is described in [26], where the conversion to other H^q estimates ($q > 0$) is also discussed. In general, the Nitsche trick involves posing an auxiliary problem, referred

to here in as the *dual problem*. The dual problem consists of the adjoint operator with the error as the source term. Thus, from (2.2.20), the *dual problem* is stated as follows: Find $\delta \in V$ such that $\forall \;\; w \in V$,

$$B(\delta, w) = (e, w) \tag{2.3.1}$$

Note that the weighting function w appears in the right slot, and that (2.3.1) includes homogeneous Dirichlet boundary conditions.
Noting that $e = \phi^h - \phi \in V$, we can replace w by e in (2.3.1), which gives

$$\begin{aligned}(e, e) &= ||e||^2 \\ &= B(\delta, e)\end{aligned} \tag{2.3.2}$$

Applying the strong stability condition (1.6.11) to the dual problem,

$$||\delta||_2 \leqslant C_s||e|| \tag{2.3.3}$$

This condition is fundamental in proving the L_2 error bounds. Let $\iota \in V^h$ be the *interpolant* of δ. From the symmetry of the operator and consistency of the Galerkin methods,

$$B(\iota, e) = 0 \tag{2.3.4}$$

Subtracting (2.3.4) from (2.3.2) and applying the Cauchy-Schwarz inequality,

$$\begin{aligned}||e||^2 &= B(\delta - \iota, e) \\ &= |B(\eta_\delta, e)| \\ &= |B_{symm}(\eta_\delta, e) + |B_{skew}(\eta_\delta, e)| \\ &\leqslant ||\eta_\delta||_E + ||\eta_\delta||_{skew}||e||_E\end{aligned} \tag{2.3.5}$$

where

$$\eta = \iota - \delta \tag{2.3.6}$$

is the *interpolation error* of the solution to the dual problem. Substituting the energy norm error bound (2.2.16) (or (2.2.31) for nonsymmetric $B(\cdot,\cdot)$), (2.3.5) becomes

$$||e||^2 \leqslant (||\eta_\delta||_E + ||\eta_\delta||_{skew})Ch^p||\phi||_{p+1} \tag{2.3.7}$$

We now focus attention on bounding the η_δ term. In general, interpolation estimates for η_δ can be stated in the following form:

$$||\eta_\delta||_s \leqslant C_i(\Omega,p)h^{\alpha-s}||\delta||_\alpha \tag{2.3.8}$$

where α is usually taken to be

$$\alpha = \min(p+1, r)$$

and r denotes the regularity of δ (which in our case is $r = 3$, since the data of the dual problem is e which lies in $H^1(\Omega)$). Obviously, bounds can be obtained in terms of lower norms $||\delta||_\beta$, where $\beta < \alpha$, since $H^\alpha(\Omega) \subset H^\beta(\Omega)$. Noting the stability condition (2.3.3) involves the H^2 norm of δ, we apply (2.3.8) with $\beta = 2$ in place of α:

$$||\eta_\delta||_s \leqslant C_i(\Omega,p)h^{2-s}||\delta||_2 \tag{2.3.9}$$

The cost of bounding η_δ in the lower H_β norm is a reduced power of h. However, because of the relationship of δ to e through (2.3.3), this is the best we can do (note for piecewise linears, $p = 1$ and $\alpha = 2$; thus the above argument is not required). Expanding $||\eta_\delta||_E$ and applying (2.3.9), we get

$$\begin{aligned}
||\eta_\delta||_E &= (B_{symm}(\eta_\delta,\eta_\delta))^{1/2} \\
&= (a(\eta_{\delta x},\eta_{\delta x}) + c(\eta_\delta,\eta_\delta))^{1/2} \\
&= (a||\eta^{\delta x}||^2 + c||\eta_\delta||^2)^{1/2} \\
&\leqslant ([ah^2 + ch^4]C_i^2||\delta||_2^2)^{1/2} \\
&= \overline{C}C_i h||\delta||_2
\end{aligned} \tag{2.3.10}$$

where

$$\overline{C} = (a + ch^2)^{1/2}$$

Similarly for the skew term,

$$\begin{aligned}
||\eta_\delta|| &= \sup_{v \in V} \frac{|B_{skew}(\eta_\delta, v)|}{||v||_E} \\
&= \sup_{v \in V} \frac{|b(\eta_\delta, v_x)|}{||v||_E} \\
&= \sup_{v \in V} \frac{b||\eta_\delta||\ ||v_x)||}{(a||v||_x||^2 + c||v||^2)^{1/2}} \\
&\leqslant \frac{b}{a^{1/2}} ||\eta_\delta|| \\
&\leqslant \frac{b}{a^{1/2}} C_i h^2 ||\delta||_2 \\
&= \overline{\overline{C}} C_i h ||\delta||_2 \qquad (2.3.11)
\end{aligned}$$

where

$$\overline{\overline{C}} = \frac{bh}{a^{1/2}}$$

Substituting (2.3.10) and (2.3.11) into (2.3.7) and applying the stability condition (2.3.3),

$$\begin{aligned}
||e||^2 &\leqslant (\overline{C} + \overline{\overline{C}}) C_i C h^{p+1} ||\delta||_2\ ||\phi||_{p+1} \\
&\leqslant (\overline{C} + \overline{\overline{C}}) C_i C C_s h^{p+1} ||e||\ ||\phi||_{p+1} \qquad (2.3.12)
\end{aligned}$$

Therefore, the a priori error estimate in the L_2, norm is given by

$$||e|| \leqslant C_L h^{p+1} ||\phi||_{p+1} \qquad (2.3.13)$$

where

$$C_L = (\overline{C} + \overline{\overline{C}}) C_i C C_s$$

REMARK: The statement $||e|| \leqslant C_D ||e||_E$ (where C_D is a constant) follows directly from the definition of the norms. However, convergence in the energy norm is $O(h_p)$; thus,

such a statement implies $O(h_p)$ convergence in the L_2 norm as well. The purpose of the Nitsche trick is to extract the extra power of h, showing explicitly that convergence in the L_2 norm is $O(h_{p+1})$.

SUMMARY: The derivation of (2.3.13) begins with an error bound in the energy norm (see previous section) followed by the application of the Nitsche trick. It is noted that the stability of the dual problem (see (2.3.3)) plays a fundamental role in the proof of the estimate.

2.3.2 Indefinite operators (the Helmholtz equation)

For simplicity we analyze only the Helmholtz equation, obtained from the model equation (1.6.1) by putting $a = 1,\ b = 0$ and $c = -k_2$, where $k \in \Re$ is the wave number. Since $c < 0$, this operator is, in general, indefinite. As in the previous cases, only Dirichlet boundary conditions at both ends of the domain (see (1.6.2) and (1.6.3)) has been considered. Since these conditions admit real eigenvalues, it is necessary to assume that $g_0,\ g_L,\ f$ and k are chosen such that a unique solution exists. For the first time in this work an energy norm does not exist, which leads to a slightly more complicated proof. It is particularly important to keep close track of the constants which appear. A general outline of proofs for indefinite operators of this type (where the indefiniteness is due to a lower order term) is given in [6]. Ihlenburg and Babuska have analyzed the Helmholtz equation in detail in [13, 16], where a nonreflecting Robin boundary condition at $x = L(= 1)$ is considered. The Helmholtz equation has also been analyzed by Bayliss et al. [5] and by Aziz et al. [4].

2.3.2.1 Nitsche trick

Since there is no energy norm, the Nitsche trick is interpreted differently than in the previous Section. Instead of converting an energy norm estimate to an L_2 estimate, the Nitsche trick is used to derive the following bound:

$$||e|| \leqslant C_1||e_x|| \tag{2.3.14}$$

The idea is to obtain a bound on $||e_x||$ (which is equivalent to the energy norm in the positive definite case), then use (2.3.14) to convert to a bound on the L_2 error. Note the similarity of (2.3.14) with the Poincare-Friedrich inequality, given by (2.2.28). The important difference is that the constant C_1 in (2.3.14) depends on h. As mentioned in the remark in the previous section, the purpose of the Nitsche trick is to extract this extra power of h, leading to a convergence rate in the L_2 norm that is one order higher than the convergence rate in the H_1-seminorm.

The Nitsche trick begins with the same dual problem as before (see (2.3.1)):

Find $\delta \in V$ such that $\forall\ w \in V$,

$$B(\delta, w) = (e, w) \tag{2.3.15}$$

Applying the strong stability condition (1.6.11) to the dual problem,

$$||\delta||_2 \leqslant \widetilde{C}_s||e|| \tag{2.3.16}$$

Note that $\widetilde{C}_s$ is a different constant than the one appearing in (2.3.3). In particular, $\widetilde{C}_s$ is a function of the wave number:

$$\widetilde{C}_s = C_e(1 + k) \tag{2.3.17}$$

where C_e is a constant. This condition is proved in [13], and is fundamental in proving the L_2 error bounds. Substituting $w = e \in V$ into (2.3.15),

$$\begin{aligned}(e, e) &= ||e||^2 \\ &= B(\delta, e)\end{aligned} \tag{2.3.18}$$

As in the previous section, let $\iota \in V^h$ be the interpolant of δ. From the symmetry of the operator and consistency of the Galerkin method,

$$B(\iota, e) = 0 \tag{2.3.19}$$

Subtracting (2.3.19) from (2.3.18) and expanding the operator and applying the Cauchy-Schwarz inequality, we get

$$\begin{aligned} ||e||^2 &= B(\delta - \iota, e) \\ &= |B(\eta_\delta, e| \\ &\leqslant |(\eta_{\delta x}, e_x) - k^2(\eta_\delta, e)| \\ &\leqslant |(\eta_{\delta x}, e_x) + k^2(\eta_\delta, e)| \\ &\leqslant ||\eta_{\delta x}||||e_x|| + k^2||\eta_\delta||||e|| \end{aligned} \tag{2.3.20}$$

where

$$\eta_\delta = \iota - \delta \tag{2.3.21}$$

is the interpolation error with respect to the dual problem. Using the same arguments as in the previous section (see the discussion prior to (2.3.9)), the applicable interpolation estimates are of the following form (see Remark 3 below for an alternative procedure for higher-order elements):

$$||\eta||_s \leqslant C_i(\Omega, p)h^{2-s}||\delta||_2 \tag{2.3.22}$$

Applying (2.3.22) and the stability condition (2.3.16) to (2.3.20), we get

$$||e||^2 \leqslant C_i\widetilde{C}_s h||e||\ ||e_x|| + C_i\widetilde{C}_s k^2h^2||e||^2$$

Dividing through by $||e||$ and rearranging, we arrive at the form given by (2.3.14), where

$$C_1 = \frac{C_i\widetilde{C}_s h}{1 - C_i\widetilde{C}_s k^2h^2} \tag{2.3.23}$$

For C_1 to be positive,

$$\widetilde{C}_s k^2h^2 < \frac{1}{C_i}. \tag{2.3.24}$$

Note that this development does not make sense unless C_1 is positive. Hence, it is assumed that (2.3.14) holds.

2.3.2.2 Convergence proof:

We first show *coercivity* of the indefinite operator with respect to $e \in V$:

$$\begin{aligned} B(e,e) &= ||e_x||^2 - k^2||e||^2 \\ &\geqslant ||e_x||^2 - C_1^2 k^2 ||e_x||^2 \\ &= (1 - C_1^2 k^2)||e_x||^2. \end{aligned} \tag{2.3.25}$$

For the right-hand side to be positive,

$$1 - C_1^2 k^2 > 0$$

Now, beginning with (2.3.25), convergence in the H_1-seminorm is established as follows:

$$\begin{aligned} 1 - C_1^2 k^2 ||e_x||^2 &\leqslant B(e,e) \\ &= B(e, e^h + \eta) \qquad (Errorsplitting) \\ &= B(e,\eta) \qquad (Symmetry\ and\ consistency) \\ &= (e_x, \eta_x) - k^2(e,\eta) \\ &\leqslant |(e_x,\eta_x)| + k^2|(e,\eta)| \\ &\leqslant ||(e_x||\ ||\eta_x|| + k^2||e||\ ||\eta)|| \qquad (Cauchy - Schwarz) \\ &\leqslant ||(e_x||\ ||\eta_x|| + C_1 k^2 ||e_x||\ ||\eta)|| \qquad (from((2.3.14))) \end{aligned} \tag{2.3.26}$$

Dividing throughout by $||e_x||$,

$$(1 - C_1^2 k^2)||e_x|| \leqslant ||\eta_x|| + C_1 k^2 ||\eta||. \tag{2.3.27}$$

Assuming the exact solution ϕ is smooth (i.e. $\phi \in H^{p+1}(\Omega)$), interpolation estimates take the form

$$||\eta||_s \leqslant C_i(\Omega, p) h^{p+1-s} ||\phi||_{p+1} \tag{2.3.28}$$

Applying these estimates to (2.3.27) and dividing throughout by $(1 - C_1^2 k^2)$, we get

$$||e_x|| \leqslant C_2 h^p ||\phi||_{p+1} \tag{2.3.29}$$

where C_2 depends on k and h:

$$C_2 = \frac{C_i(1 + C_1 k^2 h)}{1 - c_1^2 k^2} \tag{2.3.30}$$

The L_2 bound can now be obtained from (2.3.14), which leads to convergence of $O(h^{p+1})$. The above error bound is valid only if C_2 is positive. Note that the condition for positivity of C_1 is given by (2.3.24), which implies that the numerator of C_2 is positive. For the denominator of C_2

$$\begin{aligned} 0 &< 1 - C_1^2 k^2 \\ &= 1 - \frac{C_i^2 \widetilde{C}_s^2 k^2 h^2}{(1 - C_i \widetilde{C}_s k^2 h^2)^2} \\ &= \frac{(1 - C_i \widetilde{C}_s k^2 h^2)^2 - C_i^2 \widetilde{C}_s^2 k^2 h^2}{(1 - C_i \widetilde{C}_s k^2 h^2)^2} \end{aligned} \tag{2.3.31}$$

The denominator of (2.3.31) is positive, which implies (examining the numerator)

$$C_i^2 \widetilde{C}_s^2 k^2 h^2 < (1 - C_i \widetilde{C}_s k^2 h^2)^2 \tag{2.3.32}$$

Rearranging the terms, we get

$$2 C_i \widetilde{C}_s k^2 h^2 + C_i^2 \widetilde{C}_s^2 k^2 h^2 < 1 + C_i^2 \widetilde{C}_s^2 k^4 h^4. \tag{2.3.33}$$

REMARKS:

1. Note that if $\widetilde{C}_s$ was independent of k, then (2.3.24) and (2.3.33) imply a need to control kh, which is a measure of the number of elements per wave. However, this is not the case. Due to the dependency of $\widetilde{C}_s$ on k (see (2.3.17)), the condition on the mesh size h is actually much stronger. Substituting for $\widetilde{C}_s$, using (2.3.17), condition (2.3.24) becomes

$$k^2 h^2 + k^3 h^2 < \frac{1}{C_e C_i}. \tag{2.3.34}$$

Condition (2.3.33), which is stronger, becomes

$$(1+\frac{1}{2}C_eC_i)k^2h^2+(1+C_eC_i)k^3h^2+\frac{1}{2}C_eC_ik^4h^2<\frac{1}{2C_eC_i}+\frac{C_i^2\widetilde{C}_s^2k^4h^4}{2C_eC_i}$$
$$=\frac{1}{C_eC_i},$$

where the second line was obtained using (2.3.24). Implied is the need to control not only kh, but also k^3h^2 (from both conditions) and k^2h (from (2.3.33)).

2. Assume $p=1$ (linear). The error bound (2.3.29) becomes (using the strong stability condition for the ϕ term):

$$\begin{aligned}||e_x|| &\leqslant C_2h||\phi||_2\\ &\leqslant C_2\widetilde{C}_sh||f||\\ &= C_2C_e(1+k)h||f|| \end{aligned} \tag{2.3.35}$$

As mentioned in [13], controlling the size of k^2h (necessary only to prove the error estimate) leads to overkill. Assuming k^2h is bounded, then C_2 is also bounded as shown above. Substituting $h\leqslant C/k^2$ into(2.3.35), we get

$$||e_x||\leqslant CC_2C_e\frac{(1+k)}{k^2}||f||. \tag{2.3.36}$$

Thus, as k gets very large, $||e_x||$ tends to zero. Ihlenburg and Babuska [13] showed that a finite error can be maintained with the less stringent condition that k^3h^2 be controlled. However, the extension to multidimensions of these ideas is not clear. The need to control k^3h^2 is also concluded by Bayliss et al. [5].

3. An alternative analysis for $p>1$ (i.e. for higher order elements) can proceed along the lines of Ihlenburg and Babuska [16]. Here, the Nitsche trick is structurally different. Noting that the solution δ to the dual problem (2.3.14) lies in $H^3(\Omega)$ (since $e\in H^1(\Omega)$), the applicable interpolation estimate is

$$||\eta_\delta||_s\leqslant C_ih^{3-s}||\delta||_{3-s} \tag{2.3.37}$$

where (recall) η_δ is the interpolation error with respect to δ. Ihlenburg and Babuska derive a stability condition for $|\delta|_3$ which is given as

$$|\delta|_3 \leqslant (1 + 4k)||e||_1 \tag{2.3.38}$$

These results change the form of the constant C_1 in (2.3.14). Namely, the condition (2.3.24) is no longer required. However, the constant C_2 still appears in the convergence proof with the same stringent requirement on k^2h. Thus, the basic conclusions do not change.

4. It is noted that for higher p, the interpolation constant C_i decreases (see e.g. [21]), which relaxes the restricti]]ons on h (see ((2.3.24) and (2.3.33)).

SUMMARY:

The derivation of the a priori error estimate for the Helmholtz equation consists of the following basic steps:

1. Apply the Nitsche trick to bound $||e||$ in terms of $||e_x||$ (see (2.3.14) and (2.3.23)).
2. Establish coercivity of $B(\cdot, \cdot)$, noting the restriction on h to ensure positivity (see (2.3.25)).
3. Using the result from Step 2, derive a bound on $||e_x||$ (see (2.3.26)-(2.3.29)). Use the result from Step 1 to convert this bound on the L_2 norm of e.

The critical role of the strong stability result is emphasized (See (2.3.16)). In particular, the stability constant $\widetilde{C}_s$ depends on k. This leads to a restriction on the magnitude of K^2h in order to prove convergence.

Chapter 3

A POSTERIORI ERROR ESTIMATES

3.1 Introduction

A posteriori error estimate are those estimates which measure the error in some norm once the exact solution is known. As against the a priori estimates, the a posteriori error estimates are computable as they use the exact solution and the computed solution. Therefore, a posteriori error estimates give quantitative information about the size of the error after that the approximate solution $u_h(x)$ has been computed. In the sections to follow, the a posteriori error estimates have been derived in some of the norms [28].

3.2 A Posteriori Estimates in energy norm

The methodology for proving a posteriori error estimates is almost similar for symmetric operators and nonsymmetric operators, since the Cauchy-Schwarz inequality (valid for symmetric operators only) is not needed. In the model equation we need only to assume $c \geqslant 0$ to ensure positivity (also there are restrictions $a > 0$ and $b > 0$). The derivation begins the same way as for the a priori estimates. Note that $||w||_E = B(w, w)$ regardless

of the symmetry properties of $B(\cdot,\cdot)$. Let e denote the finite element solution error, then

$$\begin{aligned} e &= \phi^h - \phi \qquad (3.2.1)\\ &= \phi^h - \widetilde{\phi}^h + \widetilde{\phi}^h - \phi \\ &= e^h + \eta. \qquad (3.2.2)\end{aligned}$$

where

$$e_h = \phi_h - \widetilde{\phi}_h \qquad (3.2.3)$$

is the portion of the error in V_h, and

$$\begin{aligned} \eta &= \widetilde{\phi}_h - \phi \\ &= e - e_h \qquad (3.2.4)\end{aligned}$$

is the interpolation error and is a member of V.

Using the error splitting (3.2.1) and Galerkin consistency, i.e.

$$B(e, e^h) = 0$$

we get,

$$\begin{aligned} ||e||_E^2 &= |b(e,e)| \\ &= |B(e^h + \eta, e)| \\ &= |B(e^h, e) + B(\eta, e)| \\ &= |B(\eta, e)|. \qquad (3.2.5)\end{aligned}$$

Here, we diverge from the a priori proofs. For the a priori estimates, the objective is to cancel the error term on the right with one on the left, and then apply interpolation estimates to bound the remaining η term as a function of the exact solution ϕ. For the a posteriori estimates, the goal is to form residuals, and then apply interpolation estimates to bound the remaining η term as a function of the error. (Recall from (3.2.4) that η has

two different interpretations.) This error term then cancels with one on the left, leaving a bound in terms of only residuals.

Continuing from (3.2.5), now split the error term in the right slot of $B(\cdot,\cdot)$ using (3.2.1), we get

$$\begin{aligned} ||e||_E^2 &= |B(\eta, \phi^h - \phi)| \\ &= |B(\eta, \phi^h) - B(\eta, \phi)| \\ &= |B(\eta, \phi^h - L(\eta)|, \end{aligned} \tag{3.2.6}$$

where $L(\eta)$ is obtained using weak formulation of the model equation. The effect of the third line is to remove the exact solution ϕ from the expression, leaving the Galerkin solution ϕ^h. (Note that the right-hand side of (3.2.6) is not zero since $\eta \notin V^h$.)

The next step is to form residuals, which is always done through integration by parts. Furthermore, the integration by parts must be performed separately on each element K, since ϕ^h is only C^0 continuous across element interfaces (which are simply nodes in one dimension).

We introduce the following notation. Let element K be defined by

$$\widetilde{\Omega}_K =]x_{K-1}, x_K[\qquad K = 1, ..., n_{e1} \tag{3.2.7}$$

where n_{e1} is the number of elements, $x_0 = 0$ and $x_{n_{e1}} = L$. Thus, the elements are numbered sequentially from left to right. In addition, let

$$\begin{aligned} x_K^+ &= \lim_{\epsilon \to 0}(x_K + \epsilon) \\ x_K^- &= \lim_{\epsilon \to 0}(x_K - \epsilon) \end{aligned}$$

where $\epsilon > 0$. Substituting (1.6.6) into (3.2.6) and integrating by parts, we get

$$\begin{aligned}
||e||_E^2 &= |a(\eta_x, \phi_x^h) - b(\eta_x, \phi^h) + c(\eta, \phi^h) - (\eta, f)| \\
&= |\sum_{K=1}^{n_{e1}} \Big\{ -a(\eta, \phi_{xx}^h)_{L_2(\Omega_K)} + a\eta\phi_x^h|_{x_{K^+-1}}^{x_{K^-}} + b(\eta, \phi_x^h)_{L_2(\Omega_K)} - b\eta\phi^h|_{x_{K-1}}^{x_K} \\
&\quad + c(\eta, \phi^h)_{L_2(\Omega_K)} - (\eta, f)_{L_2(\Omega_K)} \Big\}| \\
&= \left| \sum_{K=1}^{n_{e1}} \{(\eta, r^h)_{L_2(\Omega_K)}\} + \sum_{K=1}^{n_{e1}-1} \{\eta(x_K)[[a\phi_x^h(x_K)]]\} \right| \\
&\leqslant |(\eta, r^h)_{L_2(\Omega)}| + \left| \sum_{K=1}^{n_{e1}-1} \{\eta(x_K)[[a\phi_x^h(x_K)]]\} \right|
\end{aligned} \tag{3.2.8}$$

where r^h is the residual on element interiors, given by

$$r^h = -a\phi_{xx}^h + b\phi_x^h + c\phi^h - f, \tag{3.2.9}$$

and $[[\cdot]]$ is a jump operator, i.e.

$$[[a\phi_x^h(x_K)]] = a\phi_x^h(x_K^+) - a\phi_x^h(x_K^-). \tag{3.2.10}$$

The union of element interiors is denoted by a. Thus,

$$|| \cdot ||_{L^2(\Omega)} = \left(\sum_{K=1}^{n_{e1}} || \cdot ||_{L_2(\Omega_K)}^2 \right)^{1/2}$$

Note that the second term in the last two lines of (3.2.8) is a sum over the interior element interfaces (which are nodes in one dimension); it is usually assumed that $\eta = 0$ (thus the jump term vanishes) on any Dirichlet boundaries, which includes both $x = 0$ and $x = L$ in the model problem.

REMARK:

In our simplified one-dimensional setting, $\eta = 0$ at every element interface. Therefore, the jump term in (3.2.8) vanishes for all K. This is not true in multidimensions, where the jump term must be retained. (However, the assumption $\eta = 0$ on Dirichlet boundaries is still made, which simply implies that the Dirichlet data is assumed to be exactly satisfied by all functions in S^h.) Since the purpose here is to show the structure of the derivation

in a simplified setting, we will retain this term and carry it along in the proof. The generalization to higher dimensions will therefore be more straightforward. It should finally be noted that in multidimensions, the jump term appears as an integral over the element boundary. The multidimensional case (in the context of the Helmholtz equation), including Neumann and DtN boundary conditions, is considered in [27] .

Before continuing, we introduce appropriate interpolation estimates,

$$||\eta||_{L^2(\partial\Omega_K)} \leqslant C_1 h_K ||e||_{E.\Omega_K} \tag{3.2.11}$$

$$||\eta||_{L_2(\partial\Omega_K)} \leqslant \overline{C_1} h_K^{1/2} ||e||_{E.\Omega_K} \tag{3.2.12}$$

where

$$C_1 = \sup_{v\in V} \frac{h_K^{-1}||\widetilde{v}^h - v||_{L^2(\Omega_K)}}{||v||_{E.\Omega_K}} \tag{3.2.13}$$

$$\widetilde{C}_1 = \sup_{v\in V} \frac{h_K^{-1/2}||\widetilde{v}^h - v||_{L_2(\partial\Omega_K)}}{||v||_{E.\Omega_K}}. \tag{3.2.14}$$

and $\widetilde{v^h}$ is the interpolant of v. These estimates are given in [19] .

Next, we will treat the element interior term and the element boundary term (the jump term) separately. The element interior term (the first term in the last line of (3.2.8)) can be bounded as follows:

$$\begin{aligned} |(\eta, r^h)_{L_2(\Omega)}| &= |(h^{-1}\eta, hr^h)_{L_{2(\Omega)}}| \\ &\leqslant ||h^{-1}\eta||\ ||hr^h||_{L_{2(\Omega)}} \\ &\leqslant ||e||_E ||C_1 hr^h||_{L_{2(\Omega)}}. \end{aligned} \tag{3.2.15}$$

The error on element boundaries can be rewritten as a sum over the elements:

$$\left|\sum_{K=1}^{n_{e1}-1} \eta(x_K)[[a\phi_x^h(x_K)]]\right| = \left|\sum_{K=1}^{n_{e1}-1} \left\{ \sum_{S=K-1}^{K} \eta(x_s) h_K R_h(x_S) \right\}\right| \tag{3.2.16}$$

where

$$h_K R^h(x_s) = \begin{cases} \frac{1}{2}[[a\phi_x^h(x_s)]], & \text{if S is an interior node,} \\ 0, & \text{if S is a Drichlet boundary node} \end{cases} \tag{3.2.17}$$

The second summation in (3.2.16) is a sum over the two endpoints of element K. In multidimensions, this sum appears as an integral over the element boundary. In this example, $\eta(x_s)[[a\phi_x^h(x_s)]]$ is a constant (actually, it is zero, as noted in the remark above). Applying the Cauchy-Schwarz inequality and the interpolation estimate (3.2.12), we get

$$\left|\sum_{K=1}^{n_{e1}}\left\{\sum_{S=K-1}^{K}\eta(x_s)h_K R^h(x_S)\right\}\right| \leqslant \sum_{K=1}^{n_{e1}}\left\{\left(\sum_{S=K-1}^{K}|\eta(x_s)|\right)\left(\sum_{S=K-1}^{K}|h_K R^h(x_S)|\right)\right\} \tag{3.2.18}$$

$$\leqslant \sum_{K=1}^{n_{e1}}\left\{(\overline{C}_1 h_K^{1/2}||e||_{E.\Omega_K})\left(\sum_{S=K-1}^{K}|h_K R^h(x_S)|\right)\right\} \tag{3.2.19}$$

$$\leqslant \sum_{K=1}^{n_{e1}}\left\{\left(\overline{C}_1 h_K^{1/2}\sum_{S=K-1}^{K}|h_K R^h(x_S)|\right)^2\right\}^{1/2} ||e||_E \tag{3.2.20}$$

$$\equiv ||\overline{C}_1 h R^h||_{L_2(\Omega)}||e||_E \tag{3.2.21}$$

Note that in (3.2.21) R^h has been lumped onto the element interiors where it is taken as constant.

REMARKS:

1. In treating the boundary terms as element (interior) quantities, we can proceed to obtain the general form of the error estimator given by (1.4.19). This is a convenient form for visualizing the global error as a summation of element contributions.

2. The treatment of the boundary residuals follows that of Johnson and Hansbo [19].

3. The factor of $\frac{1}{2}$ appearing in front of the jump term in (3.2.17) is an approximation. It simply means that the residual on element interfaces is split evenly among the two elements sharing that interface. A more intelligent flux splitting would likely increase the local accuracy of the error estimates.

We may now combine the interior and boundary terms to obtain the final a posteriori

error estimate. Inserting (3.2.15) and (3.2.21) into (3.2.8), we get

$$||e||_E \leqslant ||C_1 h r^h||_{L_2(\Omega)} + ||\widetilde{C}_1 h R^h||_{L_2(\Omega)} \tag{3.2.22}$$

SUMMARY:

The derivation of (3.2.22) for positive definite operators consists of three basic steps:

1. Use the definition of $||e||_E^2$, split the error in both slots, using (2.2.3) in one slot and (2.2.4) in the other. Use Galerkin consistency to get rid of some terms, and then get rid of the exact solution ϕ by substituting from the variational equation (see (3.2.6)).

2. Integrate by parts (on each element separately) to form residuals on element boundaries(see (3.2.9) and (3.2.17), respectively).

3. Bound the residual terms (seperately) using Cauchy-Schwarz inequalities and appropriate interpolation estimates. Combine these bounds to obtain the final error estimate.

3.3 A Posteriori Estimates in the L_2 norm

The procedure for a posteriori estimates in the L_2 norm is the same regardless of symmetry or definiteness of the operator. Many of the steps mimic those already presented in Section 2 for the energy norm, so we will refer back to that section. However, the interpretation of some of the terms will change.

3.3.1 Nitsche trick

Recall the application of the Nitsche trick for proving a priori estimates in the L_2 norm. Derivation of a posteriori estimates in the L_2 norm has an analogous procedure. We begin

by introducing the dual problem:

Find $\delta \in V$ such that $\forall \;\; w \in V$, we get

$$B(\delta, w) = (e, w). \tag{3.3.1}$$

Recall the strong stability condition (2.3.3), which plays a fundamental role in the proof of a posteriori estimates. Substituting $e \in V$ for w, (3.3.1) becomes

$$\begin{aligned}(e, e) = ||e||^2 &= B(\delta, e) \\ &= B(\delta, \phi^h - \phi) \\ &= B(\delta, \phi^h) - B(\delta, \phi) \\ &= B(\delta, \phi^h) - L(\delta). \end{aligned} \tag{3.3.2}$$

The error splitting allows the substitution $B(\delta, \phi^h) = L(\delta)$ to be made. Compare this step to the a priori proof, where the error splitting was not performed (see (2.3.5)).

Let $\iota \in V$ be the interpolant of δ. Using the Galerkin method, we get

$$B(\iota, \phi^h) - L(\iota) = 0. \tag{3.3.3}$$

Subtracting (3.3.3) from (3.3.2),

$$\begin{aligned}||e||^2 &= B(\delta - \iota, \phi^h) - L(\delta - \iota) \\ &= |B(\eta_\delta, \phi^h) - L(\eta_s) \end{aligned} \tag{3.3.4}$$

where, as in the a priori proofs,

$$\eta_\delta = \iota - \delta \tag{3.3.5}$$

is the interpolation error with respect to the dual problem.

Now, compare (3.3.4) to (3.2.6), which corresponds to a similar point in the a posteriori energy-norm proof. The two expressions are identical in form; only the interpolation

error is different. From this point, the L_2-norm proof follows the same lines as the energy-norm proof (see Section 2); therefore the steps are not repeated here. The interpolation estimates given by (3.2.11) and ((3.2.12) will not suffice, however, and others are needed. The applicable estimates here are (see [19])

$$||\eta_\delta||_{L_2(\Omega_K)} \leqslant C_i h^2 ||\delta||_2 \tag{3.3.6}$$

$$||\eta_\delta||_{L_2(\partial\Omega_K)} \leqslant \widetilde{C}_i h^{3/2} ||\delta||_2 \tag{3.3.7}$$

The estimate over element interiors (3.3.6) is simply (2.2.13) (which was used for the a priori proof) with $s = 0$. Substituting the strong stability condition (2.2.3) into the above estimates, we get

$$||\eta_\delta||_{L_2(\Omega_K)} \leqslant C_i C_s h^2 ||e|| \tag{3.3.8}$$

$$||\eta_\delta||_{L_2(\partial\Omega_K)} \leqslant \widetilde{C}_i C_s h^{3/2} ||e||. \tag{3.3.9}$$

The above changes lead to the following modifications of the energy-norm proof:
The element interior term in (3.2.15), becomes

$$\begin{aligned} |(\eta, r^h)_{L_2(\Omega)} &= |h^{-2}\eta, h^2 r^h)_{L_2(\Omega)}| \\ &\leqslant ||h^{-2}\eta||\ ||h^2 r^h)||_{L_2(\Omega)} \\ &\leqslant C_s||e||\ ||C_i h^2 r^h||_{L_2(\Omega)}. \end{aligned} \tag{3.3.10}$$

The element boundary term in (3.2.21), becomes

$$\left| \sum_{K=1}^{n_{e1}} \left\{ \sum_{S=K-1}^{K} \eta(x_s) h_K R^h(x_S) \right\} \right| \leqslant C_s||e||\ ||\overline{C}_i h^2 R^h||_{L_2(\Omega)}. \tag{3.3.11}$$

The definitions of the residuals r^h and R^h (given, respectively, by (3.2.9) and (3.2.17))do not change. As in last Section, (3.3.10) and (3.3.11) may now be combined to form an upper bound on $||e||^2$. Canceling the $||e||$ term appearing on both sides, we obtain the final a posteriori error estimate in the L_2 norm:

$$||e|| \leqslant C_s[||C_i h^2 r^h||_{L_2(\Omega)}]. \tag{3.3.12}$$

Noting that the stability constant C_s is a global scaling factor. For h-adaptivity, where p does not change, the interpolation constants C_i and $\widetilde{C}_i$ can be pulled outside of their respective norms. Comparing (3.3.12) to the energy-norm estimate, given by (3.2.22), we see that convergence in the L_2 norm is one order higher with respect to h. The use of the dual problem in the proof extracted this extra power of h. In [27], the preceding proof is carried out for the Helmholtz equation in multidimensions.

SUMMARY:

The derivation of (3.3.12) consists of five basic steps:

1. Introduce a continuous dual problem (the Nitsche trick), involving the adjoint operator with the error e as the source term (see (3.3.1)).

2. Insert e as the weighting function in the dual problem, which forms immediately the L_2 norm of the error. Then expand the operator, splitting the error in the right slot. Get rid of the exact solution ϕ by substituting from the variational equation (see ((3.3.2)).

3. From the resulting expression, subtract a Galerkin variational formulation with $\iota \in V^h$ as the weighting function, where ι is taken to be the interpolant of the dual problem solution. This step allows formation of interpolation errors with respect to the dual problem.

4. Integrate by parts (on each element separately) to form residuals on element interiors and on element boundaries (these steps were shown explicitly for the energy-norm proofs in last Section).

5. Bound the residual terms (separately) using Cauchy-Schwarz inequalities, appropriate interpolation estimates, and the strong stability condition of the dual problem. Combine these bounds to obtain the final error estimates.

3.4 Conclusions

The work provides a tutorial on how to derive a priori and a posteriori Galerkin finite element error bounds for general linear elliptic operators. The a posteriori error bounds fall into the class of explicit residual-based a posteriori error estimators. The a priori and a posteriori estimates have very similar structures, and their derivations follow along similar lines. The connections between the two were emphasized and put in a new light. However, important technical differences do exist, and these were noted.

Although all the proofs herein were carried out in one dimension, extensions to multidimensions are also possible. The steps involved in deriving the error bounds depend on the error norm. Both the energy norm and L_2 norm were considered. In general, the L_2 bounds were derived by starting with the energy norm estimates and applying some form of the Nitsche trick. The derivation of the a priori bounds also depends on the symmetry and definiteness of the operator. By appropriately bounding coefficients of a general operator, we were able to analyze symmetric, nonsymmetric, positive definite and indefinite operators [28].

Two special operators were also considered. First, as an example of a (positive definite) nonsymmetric operator, the advection-diffusion equation was analyzed. A novel approach, which involved the introduction of the skew norm for the skew symmetric part of the operator, was used to derive the a priori error bound. The use of the skew norm made the proof very similar to that for the symmetric positive definite operator. This is a general approach and allows nonsymmetric operators to be handled in a very simple and straightforward fashion. The second special operator considered was the Helmholtz operator, which is indefinite. In deriving the a priori error bound for this operator, special attention had to be paid to the constants which appeared. Also, a critical role was played by the strong stability of the continuous operator. In particular, the stability constant

depends on the wave number k, and this directly leads to the stringent requirement of controlling k^2h (where h is the nondimensional element length) in order to complete the proof.

Bibliography

[1] M. Ainsworth and J.T. Oden, A unified approach to a posteriori error estimation using element residual methods, Numer. Math. 65:23-50, 1993.

[2] M. Ainsworth and J.T. Oden, A posteriori error estimators for second order elliptic systems: Part 1. Theoretical foundations and a posteriori error analysis, Comput. Math. Applic. 25(2):101-l13, 1993.

[3] M. Ainsworth and J.T. Oden, A posteriori error estimators for second order elliptic systems: Part 2. An optimal order process for calculating self-equilibrating fluxes, Comput. Math. Applic. 26(9):75-87, 1993.

[4] A.K. Aziz, R.B. Kellogg and A.B. Stephens, A two point boundary value problem with a rapidly oscillating solution, Numer. Math. 53:107-121, 1988.

[5] A. Bayliss, Cl. Goldstein and E. Turkel, On accuracy conditions for the numerical computation of waves, J. Comput. Phys. 59:396-404, 1995.

[6] A.H. Schatz, An observation concerning Ritz-Galerkin methods with indefinite bilinear forms, Math. Comput. 28(128):959-962, 1974.

[7] I.M. BabuSka and W.C. Rheinboldt, A posteriori error estimates for the finite element method, Int. J. Numer. Methods Engrg. 12:1597-1615, 1978.

[8] R.E. Bank and A. Weiser, Some a posteriori error estimators for elliptic partial differential equations, Math. Comput. 44:283-301, 1985.

[9] P.G. Ciarlet, The Finite Element Method for Elliptic Problems (North-Holland, amsterdam, 1978.

[10] K. Eriksson and C. Johnson, Error estimates and automatic time step control for non-linear parabolic problems, I. SlAM J. Numer. Anal. 24:12-23, 1987.

[11] K. Eriksson and C. Johnson, An adaptive finite element method for linear elliptic problems, Math. Comput. 50:361-383, 1988.

[12] L.P. Franca and T.J.R. Hughes, Convergence analysis of Galerkin Least-Squares methods for symmetric advective-diffusive forms of the Stokes and incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg. 105:285-298, 1993.

[13] F. Ihlenburg and I. BabuSka, Finite element solution to the Helmholtz equation with high wave number, Part I: The h-version of the FEM, Technical Note BN-1159, Institute for Physical Science and Technology, University of Maryland at College Park, 1993.

[14] T.J.R. Hughes, L.P. Franca and G.M. Hulbert, A new finite element formulation for computational fluid dynamics: VIII. The Galerkin Least-Squares method for advective-diffusive equations, Comput. Methods Appl. Mech. Engrg. 73:173-189, 1989.

[15] T.J.R. Hughes, G. Hauke and K. Jansen, Stabilized finite element methods in fluids: Inspirations, origins, status and recent developments, in: T.J.R. Hughes, E. Otiate and O.C. Zienkiewicz, eds., Recent Developments in Finite Element Analysis, A book dedicated to Robert L. Taylor (CIMNE, Barcelona, Spain, 272-292, 1994.

[16] 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, Institute for Physical Science and Technology, University of Maryland at College Park, Technical Note BN-1 173, 1994.

[17] C. Johnson, Adaptive finite element methods for diffusion and convection problems, Comput. Methods Appl. Mech. Engrg. 82:301-322, 1990.

[18] C. Johnson, Finite element methods for flow problems, in: AGARD Report 787 (AGARD, 7 Rue Ancelle, 92299 Neuilly sur Seine, France, 1-1-l-47, 1992.

[19] C. Johnson and P. Hansbo, Adaptive finite element methods in computational mechanics, Comput. Methods Appl. Mech. Engrg. 101:143-181, 1992.

[20] R. Miicke and J.R. Whiteman, A posteriori error estimates and adaptivity for finite element solutions in finite elasticity, Int. J. Numer. Methods Engrg. 38:775-795, 1995.

[21] J.T. Oden, L. Demkowicr, W. Rachowicz and T.A. Westermann, Toward a universal h-p adaptive finite element strategy, Part 2. A posteriori error estimation, Comput. Methods Appl. Mech. Engrg. 77:113-180, 1989.

[22] J.T. Oden, L.Demkowicz, W. Rachowicz and T.A. Westermann, A posteriori error analysis in finite elements: The element residual method for symmetrizable problems with applications to compressible Euler and Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg. 82:183-203, 1990.

[23] J.T. Oden, Error estimation and control in computational fluid dynamics, The O.C. Zienkiewicz Lecture, in: J.R. Whiteman. ed., MAFELAP VIII: Mathematics of Finite Elements with Applications, Brunei University, Uxbridge, England, 1993.

[24] J.T. Oden, W. Wu and M. Ainsworth, An a posteriori error estimate for hnite element approximations of the Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg. 1:185-202, 1994.

[25] J.N. Reddy, An introduction to finite element methods, McGraw-Hill, New York, 2006.

[26] Cl. Strang and G. Fix, An Analysis of the Finite Element Method (Prentice-Hall, Englewood Cliffs, NJ, 1973.

[27] J.R. Stewart and T.J.R. Hughes, An a posteriori error estimator and hp-adaptive strategy for finite element discretizations of the Helmholtz equation in exterior domains, Finite Elem. Anal. Des. 25:l-26, 1997.

[28] J.R. Stewart and T.J.R. Hughes, A tutorial in elementary finite element error analysis: A systemic presentation of a priori and a posteriori error estimates, , Comput. Methods Appl. Mech. Engrg., 158:1-22, 1998.

[29] J.R. Stewart, T.J.R. Hughes I Comput. Methods Appl. Ma-h. Engrg. 158 (1998) i-22 [7] D.W. Kelly, The self-equilibration of residuals and complementary a posteriori error estimates in the finite element method, Int. J.Numer. Methods Engrg. 20:1491-1506, 1984.

[30] P. Wriggers, 0. Scherf and C. Carstensen, Adaptive techniques for the contact of elastic bodies, in: T.J.R. Hughes, E. Otiate and O.C.Zienkiewicz, eds., Recent Developments in Finite Element Analysis, A book dedicated to Robert L. Taylor (CIMNE, Barcelona, Spain, 78-86, 1994.

[31] O.C. Zienkiewicz and J.Z. Zhu, A simple error estimator in the finite element method, Int. J. Numer. Methods Engrg. 24:337-357, 1987.

[32] O.C. Zienkiewicz and J.Z. Zhu, adaptivity and mesh generation, Int. J. Numer. Methods Engrg. 32:783-810, 1991.