

THE UNIVERSITY OF CALGARY

BLENDING FUNCTIONS AND FINITE ELEMENTS

by

DAVID S. WATKINS

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS, STATISTICS, AND COMPUTING SCIENCE

CALGARY, ALBERTA

APRIL, 1974

© David S. Watkins, 1974

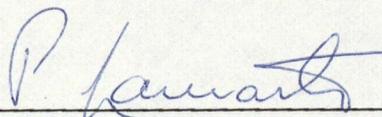
THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

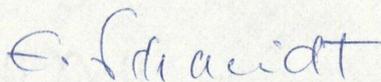
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled

"Blending Functions and Finite Elements"

submitted by David Scott Watkins in partial fulfillment of the requirements for the degree of Doctor of Philosophy.



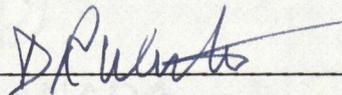
Dr. P. Lancaster (Chairman)
Department of Mathematics



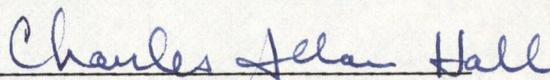
Dr. E. Schmidt
Department of Mathematics



Dr. J. Wong
Department of Mathematics



Dr. D.R. Westbrook
Division of Applied Mathematics



Dr. C.A. Hall (External Examiner)
University of Pittsburgh

April 26, 1974

date

Abstract

This thesis deals with blending-function methods, finite element methods, and some aspects of the interplay between them. In chapter two blending-function methods are introduced and asymptotic error bounds for blending-function interpolation are obtained for the case in which the blending functions are polynomials. The finite element method is described in chapter three, and in chapter four it is shown how blending-function methods can be used in the construction of finite elements. Error bounds for finite element interpolation and approximation are proven in chapter five by two different approaches. The first approach uses the theory of noninteger order Sobolev spaces to predict noninteger (as well as integer) powers of convergence, e.g. $O(h^{\frac{1}{2}})$, where h is a mesh parameter. This approach does not allow the estimation of the constants appearing in the error bounds, whereas the second approach does. However, the second approach gives only integer powers of convergence.

The key theorems for all of the error bounds of this thesis are the Sobolev lemma and the Bramble-Hilbert lemma, or variants thereof. To obtain noninteger rates of convergence the Bramble-Hilbert lemma must be generalized to noninteger Sobolev spaces. This generalization has been carried out in chapter one. It is noted that the Bramble-Hilbert lemma applies to operators as well as functionals. This observation allows some simplification in the proofs of error bounds. Also given in chapter one are variants of the Sobolev lemma and the Bramble-Hilbert lemma appropriate to blending-function methods.

In chapter six some numerical results are tabulated, and they are seen to agree with the theoretical results. The comparative cost of running various finite element programs is discussed.

Acknowledgements

I would like to thank my supervisor, Dr. P. Lancaster, for his valuable guidance these past three years. By suggesting interesting topics for investigation, while not usurping my intellectual independence, he has helped me to get a foothold in the field of numerical analysis. I would also like to thank the other members of the Department from whom I have taken courses and with whom I have discussed mathematics. I am indebted to Dr. R.E. Barnhill of the University of Utah for introducing me to blending-function methods, and to Dr. C.A. Hall of the University of Pittsburgh for discussing error bounds for blending-function interpolation with me. Finally, I would like to thank Ms. Betty Cline for quickly and accurately typing this thesis.

My research has been supported financially by the National Research Council of Canada and the University of Calgary.

Contents

Abstract	iii
Acknowledgements	v

Chapter One

Introduction, Notation, and Some Basic Theorems

(1.1) Introduction	1
(1.2) Notation; Definition of Sobolev Spaces	3
(1.3) The Sobolev Lemma and a Variant	9
(1.4) The Bramble-Hilbert Lemma	14
(1.5) A Variant of the Bramble-Hilbert Lemma	18

Chapter Two

Error Bounds for Blending-Function Methods

(2.1) Blending-Function Interpolation	23
(2.2) Boundedness of the Error Operator	28
(2.3) Sobolev Space Error Bounds	31
(2.4) Error Bounds for Continuously Differentiable Functions	33

Chapter Three

A Description of the Finite Element Method

(3.1) Elliptic Boundary Value Problems	35
(3.2) Finite Element Spaces	38
(3.3) The Finite Element Solution	43

Chapter Four

The Use of Blending-Function Methods in the Construction of Finite Elements

(4.1) Adini's Rectangle	46
(4.2) C^1 Elements	49
(4.3) An Element for Three-Dimensional Problems	57

Chapter Five

Error Bounds for Finite Element Methods

(5.1) Introduction	61
(5.2) Error Bounds for Fractional Sobolev Spaces	64
(5.3) Error Bounds with Computable Constants	75

Chapter Six

Numerical Results

(6.1) Finite Element Programs	85
(6.2) Confirmation of Rates of Convergence	90
(6.3) Comparative Cost of Running Finite Element Programs	92
(6.4) Pointwise Approximation	96

References	98
------------	----

Tables

5.1: Parameters for Theorem 5.2.6	72
6.1: $\ u-u^*\ _1$, Bilinear Element	91
6.2: $\ u-u^*\ _1$, Bilinear Element	91
6.3: $\ u-u^*\ _1$, Adini's Rectangle	92
6.4: $\ u-u^*\ _1$, 24 d.o.f. Element	92
6.5: Estimated and Actual Values of u_1	96
6.6: Estimated and Actual Values of $\frac{\partial u_1}{\partial x}$	97
6.7: Estimated and Actual Values of $\frac{\partial u_1}{\partial y}$	97

CHAPTER ONE

INTRODUCTION, NOTATION, AND SOME BASIC THEOREMS

(1.1) Introduction

The finite element method was originated almost twenty years ago by structural engineers as a method of structural analysis. As is usually the case, a lack of mathematical foundations for the procedure did not stop the engineers from using it and getting good results. After about ten years, when it had been realized that the finite element method is essentially a Ritz-Galerkin procedure, interest in the method spread to the mathematical community, and work was begun on securing the mathematical foundations of the method. Much work has been done by both engineers and mathematicians, and an extensive finite element literature now exists. Two good general references are the books [24] by O.C. Zienkiewicz (an engineer) and [21] by Strang and Fix (two mathematicians). Many further references can be found in each of these works.

From the mathematical standpoint the finite element method is a family of procedures for numerically solving differential equations. (In this thesis only elliptic partial differential equations will be considered.) Assuming that the problem is defined on some region Ω in the plane, Ω is divided into small triangular or rectangular "elements", and the solution of the equation is approximated by a function whose restriction to each element is a polynomial of low degree. It is the piecewise polynomial nature of the approximating functions which distinguishes the finite element method from other Ritz type procedures. A short introduction to the mathematical version of the finite element method is given in chapter three.

Blending-function methods are a more recent development and are less well-known. The theory was originated in 1964 by S.A. Coons [8] and has been advanced by Gordon, Hall, Barnhill, Birkhoff, Mansfield, and others. Blending-function methods are a class of methods for interpolating curves and surfaces. For example, if a continuous function v is defined on the boundary of some rectangle, blending-function methods can be used to define a "blended interpolant" q on the entire rectangle such that q equals v on the boundary of the rectangle. Of course, if v itself is defined throughout the rectangle, it is still possible to define q . In this case q is an approximant of v which interpolates v at the boundary and is completely determined by the boundary values of v . A good introduction to blending-function methods is Gordon's article [11]. However, enough information on blending-function methods for an understanding of this thesis is given in section 2.1 below.

Gordon and Hall [12] have given asymptotic bounds for the error between a function and its blended interpolant. It is assumed that the function being interpolated has a number of continuous derivatives, and the error is measured in the supremum norm. In this thesis (chapter two) similar error bounds are given in which the function has weak derivatives, and the error is measured in various Sobolev norms.

After the introduction to finite elements in chapter three, various finite element schemes are constructed in chapter four by the use of blending-function methods. The elements constructed are Adini's rectangle [1], a number of C^1 elements, and an element for three-dimensional problems. Gordon and Hall [12] and Barnhill and Gregory [3] have also used blending-function

methods to construct finite elements.

In chapter five are presented two methods for deriving bounds for the error between the exact solution and the finite element solution of an elliptic boundary value problem. The first method is essentially the method of Bramble and Zlamal [6], but here we have generalized the result by considering noninteger Sobolev spaces. An example is given to show the practical value of making such a generalization.

The bounds obtained by the first method are asymptotic error bounds containing a constant C , the value of which is generally unknown. The second method produces bounds of the same type in which the constants can be estimated. This method applies only to those elements which can be constructed by blending-function methods as in chapter four.

All of the error bounds of this thesis have been obtained by the use of the Sobolev lemma [21], and the Bramble-Hilbert lemma [21, 4] or variants of these theorems. In chapter one these two well-known theorems are stated so that they can readily be compared with their respective variants, which are stated and proved. Also, a proof of the Bramble-Hilbert lemma for noninteger Sobolev spaces is given.

In chapter six numerical results are given for comparison with the theoretical results. The comparative cost of running various finite element programs is discussed.

(1.2) Notation; Definition of Sobolev Spaces

Let Ω be a bounded domain in Euclidean n -space \mathbb{R}^n and let u be a

a smooth real-valued function on Ω . By $D_i u$ we will mean $\frac{\partial u}{\partial x_i}$, $i=1, \dots, n$. Given $\alpha=(\alpha_1, \dots, \alpha_n)$, where $\alpha_1, \dots, \alpha_n$ are nonnegative integers, we define the α th derivative of u to be $D^\alpha u = D_1^{\alpha_1} D_2^{\alpha_2} \dots D_n^{\alpha_n} u$. (The order in which the factors D_i appear in this expression is irrelevant if u is sufficiently smooth.) The *order* of α is $|\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_n$. We shall call α a *multiinteger*. The *sum* of two multiintegers α and β is $\alpha + \beta = (\alpha_1 + \beta_1, \dots, \alpha_n + \beta_n)$. The multiintegers can be partially ordered by the relation $\alpha \leq \beta$ if and only if $\alpha_i \leq \beta_i$, $i=1, \dots, n$. If $\alpha \leq \beta$, then $\beta - \alpha$ can be defined in the obvious manner. In sum, the multiintegers form a partially-ordered, commutative semigroup.

Given a multiinteger β , $C^\beta(\bar{\Omega})$ will denote the set of functions u on $\bar{\Omega}$ (the closure of Ω) such that for all $\alpha \leq \beta$, $D^\alpha u$ exists and is continuous on $\bar{\Omega}$. With the norm

$$\|u\|_{\beta, \infty} = \max_{\alpha \leq \beta} \max_{x \in \bar{\Omega}} |D^\alpha u(x)|$$

$C^\beta(\bar{\Omega})$ is a complete space. Given a nonnegative integer m we define $C^m(\bar{\Omega})$ to be the space of all functions u on $\bar{\Omega}$ such that for all α with $|\alpha| \leq m$, $D^\alpha u$ exists and is continuous on $\bar{\Omega}$. $C^m(\bar{\Omega})$ is complete with respect to the norm

$$\|u\|_{m, \infty} = \max_{|\alpha| \leq m} \max_{x \in \bar{\Omega}} |D^\alpha u(x)|.$$

The space $C^\infty(\bar{\Omega})$ is defined to be the intersection of the spaces $C^m(\bar{\Omega})$, i.e. $C^\infty(\bar{\Omega}) = \bigcap_{m=1}^{\infty} C^m(\bar{\Omega})$. We define $C_0^\infty(\Omega)$ to be the space of functions $u \in C^\infty(\bar{\Omega})$ for which there exists a compact set K in Ω such that $u(x) = 0$ for all x not in K .

We shall define *weak derivatives* as in [10]. A function u on Ω is

said to have a *weak* α th derivative v if for all $\phi \in C_0^\infty(\Omega)$

$$\int_{\Omega} u D^\alpha \phi = (-1)^{|\alpha|} \int_{\Omega} v \phi.$$

The weak derivative is at least as general as the classical derivative.

That is, if u has a continuous α th derivative $D^\alpha u$, then u has a weak α th derivative v , and $v = D^\alpha u$. This can be seen by performing $|\alpha|$ integrations by parts to obtain

$$\int_{\Omega} u D^\alpha \phi = (-1)^{|\alpha|} \int_{\Omega} (D^\alpha u) \phi.$$

The weak derivative is the same as the *distributional* derivative and is also sometimes known as the *generalized* derivative. Throughout this thesis the same notation will be used for weak derivatives as for classical derivatives. This should cause no confusion.

Let $1 \leq p \leq \infty$, and let β be a multiinteger. The space $W_p^\beta(\Omega)$ is defined to be the set of all functions $u \in L_p(\Omega)$ such that for $\alpha \leq \beta$, $D^\alpha u$ exists in the weak sense and is in $L_p(\Omega)$. We equip $W_p^\beta(\Omega)$ with the norm

$$\|u\|_{\beta,p} = \left(\sum_{\alpha \leq \beta} \int_{\Omega} |D^\alpha u|^p \right)^{1/p} \quad \text{if } p < \infty$$

$$\|u\|_{\beta,\infty} = \max_{\alpha \leq \beta} \|D^\alpha u\|_{L_\infty(\Omega)} \quad \text{if } p = \infty.$$

With this norm $W_p^\beta(\Omega)$ is a complete space. Except in the case $p = \infty$, the completeness depends on the fact that weak derivatives are admitted.

We also introduce the *Sobolev space* $W_p^{(m)}(\Omega)$. This is the space of all functions $u \in L_p(\Omega)$ such that for all α satisfying $|\alpha| \leq m$, $D^\alpha u$ exists in the weak sense and is in $L_p(\Omega)$. $W_p^{(m)}(\Omega)$ is a Banach space with the norm

$$\|u\|_{m,p} = \left(\sum_{|\alpha| \leq m} \int_{\Omega} |D^\alpha u|^p \right)^{1/p} \quad \text{if } p < \infty$$

$$\|u\|_{m,\infty} = \max_{|\alpha| \leq m} \|D^\alpha u\|_{L_\infty(\Omega)} \text{ if } p = \infty.$$

Throughout this chapter and the next chapter expressions involving the index p , where $1 \leq p \leq \infty$, will occur. In most cases a separate expression is needed for the case $p = \infty$. We have just seen two examples. From now on the expression for the case $p = \infty$ will not be explicitly stated but can be inferred from the expression for the case $p < \infty$. For example, $\left(\sum_{\alpha \leq \beta} |x_\alpha|^p \right)^{1/p}$ will mean $\max_{\alpha \leq \beta} |x_\alpha|$ if $p = \infty$, and $J^{1/p}$ will mean 1 if $p = \infty$.

Note that $W_p^{(0)}(\Omega)$ is just $L_p(\Omega)$. Accordingly, the notation $\|\cdot\|_{0,p}$ will be used to denote the L_p norm. We have

$$\|u\|_{\beta,p} = \left(\sum_{\alpha \leq \beta} \|D^\alpha u\|_{0,p}^p \right)^{1/p}$$

$$\|u\|_{m,p} = \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|_{0,p}^p \right)^{1/p}$$

The Sobolev space $\overset{\circ}{W}_p^{(m)}(\Omega)$ is defined to be the completion of $C_0^\infty(\Omega)$ in $W_p^{(m)}(\Omega)$. The space $\overset{\circ}{W}_p^{(m)}(\Omega)$ should be viewed as the set of functions in $W_p^{(m)}(\Omega)$ which satisfy in a generalized sense the Dirichlet boundary conditions $\frac{\partial^k u}{\partial n^k} = 0$ on $\partial\Omega$, $k=0, \dots, m-1$, where n is an outward normal to the boundary of Ω . The justification for this point of view is that if a function u is in $W_p^{(m)}(\Omega)$ and also is sufficiently differentiable in the classical sense, then u belongs to $\overset{\circ}{W}_p^{(m)}(\Omega)$ if and only if u satisfies the Dirichlet boundary conditions. See [10], page 39.

The case $p=2$ merits special attention. $\overset{\circ}{W}_2^{(m)}(\Omega)$ is a Hilbert space with the inner product

$$(u, v)_m = \sum_{|\alpha| \leq m} \int_{\Omega} D^{\alpha} u D^{\alpha} v.$$

Denoting the L_2 inner product by $(\cdot, \cdot)_0$, we have

$$(u, v)_m = \sum_{|\alpha| \leq m} (D^{\alpha} u, D^{\alpha} v)_0.$$

The space $W_2^{(m)}(\Omega)$ is usually denoted simply $W^{(m)}(\Omega)$, and its norm is denoted $\|\cdot\|_m$. Similar remarks apply to the spaces $\tilde{W}_2^{(m)}(\Omega)$ and $W_2^{\beta}(\Omega)$. The notation $H^m(\Omega) = W^{(m)}(\Omega)$ and $H_0^m(\Omega) = \tilde{W}^{(m)}(\Omega)$ is often seen in the literature.

As was indicated in the introductory section, it is possible to define Sobolev spaces $W_p^{(s)}(\Omega)$ for noninteger values of s . We now define these spaces for $1 \leq p < \infty$. (See [2], [19], and [13].) First suppose $0 < s = \sigma < 1$. Consider the seminorm

$$(1.2.1) \quad |u|_{\sigma, p} = \left(\int_{\Omega} \int_{\Omega} \frac{|u(x) - u(y)|^p}{\|x-y\|^{n+p\sigma}} dx dy \right)^{1/p}.$$

We define $W_p^{(\sigma)}(\Omega)$ to be the space of all functions $u \in L_p(\Omega)$ such that $|u|_{\sigma, p} < \infty$. We equip $W_p^{(\sigma)}(\Omega)$ with the norm

$$(1.2.2) \quad \|u\|_{\sigma, p} = (\|u\|_{0, p}^p + |u|_{\sigma, p}^p)^{1/p}$$

It can be shown that $W_p^{(\sigma)}(\Omega)$ is complete by using the same arguments as are used in showing that $L_p(\Omega)$ is complete.

For $s > 1$ we write $s = m + \sigma$, where m is a positive integer and $0 < \sigma < 1$.

We define a seminorm $|\cdot|_{s, p}$ by

$$(1.2.3) \quad |u|_{s, p} = \left(\sum_{|\alpha|=m} |D^{\alpha} u|_{\sigma, p}^p \right)^{1/p}.$$

$W_p^{(s)}(\Omega)$ is defined to be the space of all $u \in W_p^{(m)}(\Omega)$ such that $|u|_{s, p} < \infty$. An appropriate norm for $W_p^{(s)}(\Omega)$ is

$$(1.2.4) \quad \|u\|_{s,p} = (\|u\|_{m,p}^p + |u|_{s,p}^p)^{1/p}$$

$W_p^{(s)}(\Omega)$ is complete.

As an illustration we shall consider a simple example. Let $n=1$, $p=2$, and $\Omega=(-1,1)$. Let u be the Heaviside function

$$u(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

This function is in $W^{(0)}(-1,1) = L_2(-1,1)$ but not in $W^{(1)}(-1,1)$ (See [20], section 110). We shall verify that u is in $W^{(\sigma)}(-1,1)$ if and only if $\sigma < \frac{1}{2}$. By (1.2.1),

$$|u|_{\sigma}^2 = \int_{-1}^1 \int_{-1}^1 \frac{|u(t) - u(s)|^2}{|t - s|^{1+2\sigma}} dt ds.$$

We must show that this integral is finite if and only if $\sigma < \frac{1}{2}$. Clearly $|u(t) - u(s)|^2 = 1$ if s and t have opposite signs, and $|u(t) - u(s)|^2 = 0$ otherwise. Thus

$$|u|_{\sigma}^2 = \int_{-1}^0 \int_0^1 \frac{1}{(t-s)^{1+2\sigma}} dt ds + \int_0^1 \int_{-1}^0 \frac{1}{(s-t)^{1+2\sigma}} dt ds.$$

Either one of these integrals converges if and only if the other does. We shall examine the first integral. For fixed $s \in (-1,0)$,

$$\int_0^1 \frac{1}{(t-s)^{1+2\sigma}} dt = \frac{(t-s)^{-2\sigma}}{-2\sigma} \Big|_{t=0}^{t=1} = \frac{1}{2\sigma} [(-s)^{-2\sigma} - (1-s)^{-2\sigma}].$$

We must integrate this expression from -1 to 0 with respect to s . The integral $\int_{-1}^0 (1-s)^{-2\sigma} ds$ converges for all σ , so $|u|_{\sigma} < \infty$ if and only if $\int_{-1}^0 (-s)^{-2\sigma} ds$ converges. This integral, in turn, converges if and only if $\sigma < \frac{1}{2}$. Thus u is in $W^{(\sigma)}(\Omega)$ if and only if $\sigma < \frac{1}{2}$.

A different (but equivalent) approach to noninteger order Sobolev spaces

can be found in [16] and [13]. From this approach it is easily seen that if $s < t$, then $W_p^{(t)}(\Omega) \subseteq W_p^{(s)}(\Omega)$ with compact embedding. That is, the embedding operator $I: W_p^{(t)}(\Omega) \rightarrow W_p^{(s)}(\Omega)$ given by $Iu = u$ is a compact operator. This fact is needed in the proof of the Bramble-Hilbert lemma for noninteger (as well as integer) order Sobolev spaces.

We close this section with one last notational convention about Sobolev spaces. Situations will arise in which two domains Ω_1 and Ω_2 are being considered simultaneously. In these situations we will use notation such as $\|\cdot\|_{m,p,\Omega_1}$ and $\|\cdot\|_{m,p,\Omega_2}$ to distinguish the norms on $W_p^{(m)}(\Omega_1)$ and $W_p^{(m)}(\Omega_2)$.

(1.3) The Sobolev Lemma and a Variant

The Sobolev Lemma says that any function which has weak derivatives of high enough order is bounded and continuous. This theorem is an indispensable tool for proving error bounds for weakly differentiable functions. In this section we shall state the Sobolev lemma and state and prove a modified version which is more appropriate for blending-function methods than is the standard version.

In order to prove the Sobolev lemma one must make certain regularity assumptions about the domain Ω . Authors vary on the assumptions made. Indeed, the variation is considerable and is a source of confusion. To keep the situation simple: let us assume that Ω is a convex polyhedron. In the one application of the Sobolev lemma in this thesis, Ω is the unit square.

Theorem 1.3.1 (Sobolev Lemma): Suppose $1 \leq p < \infty$ and $m > \frac{n}{p}$. Then $W_p^{(m)}(\Omega) \subseteq C(\bar{\Omega})$ and there is a constant C such that for all $u \in W_p^{(m)}(\Omega)$,

$$\max_{x \in \bar{\Omega}} |u(x)| \leq C \|u\|_{m,p}.$$

For a proof which holds for integer values of m see [10], page 22.

A proof which holds also for noninteger m , but which is restricted to the case $p=2$, is given in [16], pages 45-46.

The Sobolev lemma has an obvious corollary, which will be considered to be part of the Sobolev lemma.

Corollary 1.3.2: Suppose $m > \frac{n}{p} + j$. Then $W_p^{(m)}(\Omega) \subseteq C^j(\bar{\Omega})$, and there is a constant C such that for all $u \in W_p^{(m)}(\Omega)$,

$$\max_{|\alpha| \leq j} \max_{x \in \bar{\Omega}} |D^\alpha u(x)| \leq C \|u\|_{m,p}.$$

Before stating and proving the modified version of the Sobolev lemma we must state a density theorem which will be used in the proof. A domain Ω is said to be *star-shaped* if there is a point x_0 in Ω such that for every $x \in \bar{\Omega}$ and $0 \leq \theta < 1$, the point $x_0 + \theta(x - x_0)$ lies in (the interior of) Ω .

Theorem 1.3.3: Let $1 \leq p < \infty$ and suppose Ω is a bounded star-shaped domain. Then for any multiinteger β , $C^\beta(\bar{\Omega})$ is dense in $W_p^\beta(\Omega)$ with respect to the norm $\|\cdot\|_{\beta,p}$.

For a proof see [20], page 328. The theorem remains true with $C^\beta(\bar{\Omega})$, $W_p^\beta(\Omega)$, and $\|\cdot\|_{\beta,p}$ replaced by $C^m(\bar{\Omega})$, $W_p^{(m)}(\Omega)$, and $\|\cdot\|_{m,p}$, where m is any nonnegative integer. The assumption that Ω be star-shaped can be relaxed considerably ([20], page 355).

A domain Ω in \mathbb{R}^n satisfies *cube condition* if there is a positive real number r such that for each $x \in \Omega$ there is a cube $K = \prod_{i=1}^n [a_i, a_i + r]$ lying in Ω such that x is one of the vertices of K . The modified Sobolev lemma is

valid for bounded, star-shaped domains satisfying cube condition. Admittedly this is a very restrictive set of conditions. However, the one domain on which we intend to apply the theorem is the unit hypercube, which does satisfy the conditions.

Theorem 1.3.4 (Modified Sobolev Lemma): *Let Ω be a bounded, star-shaped domain in \mathbb{R}^n satisfying cube condition with constant r , let $1 \leq p \leq \infty$, and let n be the multiinteger $(1,1,\dots,1)$. Then $W_p^n(\Omega) \subseteq C(\bar{\Omega})$, and for all $u \in W_p^n(\Omega)$,*

$$(1.3.1) \quad \max_{x \in \bar{\Omega}} |u(x)| \leq M \|u\|_{n,p}$$

where

$$M = \left(r^n \sum_{j=0}^n \binom{n}{j} \left(\frac{1}{r} \right)^{\frac{jp}{p-1}} \right)^{\frac{p-1}{p}}.$$

Proof:* For simplicity of notation we will prove only the case $n=2$. Let $t = (t_1, t_2) \in \bar{\Omega}$. Without loss of generality the cube $K = [t_1, t_1+r] \times [t_2, t_2+r]$ lies within $\bar{\Omega}$. As a first step we shall assume that $1 \leq p < \infty$ and show that inequality (1.3.1) holds for all functions $\phi \in C^n(\bar{\Omega})$. Let $g(x) = 1-x$. Then by the fundamental theorem of calculus

$$(1.3.2) \quad \phi(t) = - \int_{t_1}^{t_1+r} \frac{\partial}{\partial x_1} \left[g \left(\frac{x_1 - t_1}{r} \right) \phi(x_1, t_2) \right] dx_1.$$

Letting $\psi_{x_1}(t)$ denote the integrand in (1.3.2) we have

$$(1.3.3) \quad \psi_{x_1}(t) = - \int_{t_2}^{t_2+r} \frac{\partial}{\partial x_2} \left[g \left(\frac{x_2 - t_2}{r} \right) \psi_{x_1}(t_1, x_2) \right] dx_2.$$

*The author thanks Dr. D.R. Westbrook for simplifying this proof.

We combine (1.3.2) and (1.3.3) to get

$$(1.3.4) \quad \phi(t) = \int_K D_2 \left\{ g \left(\frac{x_2 - t_2}{r} \right) D_1 \left[g \left(\frac{x_1 - t_1}{r} \right) \phi(x_1, x_2) \right] \right\} dx.$$

Let I denote the integrand in (1.3.4). Then

$$I = \frac{1}{r} g' \left(\frac{x_2 - t_2}{r} \right) \frac{1}{r} g' \left(\frac{x_1 - t_1}{r} \right) \phi(x_1, x_2) + \frac{1}{r} g' \left(\frac{x_2 - t_2}{r} \right) g \left(\frac{x_1 - t_1}{r} \right) D_1 \phi(x_1, x_2) \\ + g \left(\frac{x_2 - t_2}{r} \right) \frac{1}{r} g' \left(\frac{x_1 - t_1}{r} \right) D_2 \phi(x_1, x_2) + g \left(\frac{x_2 - t_2}{r} \right) g \left(\frac{x_1 - t_1}{r} \right) D_1 D_2 \phi(x_1, x_2).$$

We apply Hölder's inequality to this last expression and use the bounds

$|g(x)| \leq 1$ and $|g'(x)| \leq 1$ to obtain

$$(1.3.5) \quad |I|^p \leq \left[\sum_{j=0}^2 \binom{2}{j} \left(\frac{1}{r} \right)^{\frac{jp}{p-1}} \right]^{p-1} \left[\sum_{\alpha \leq \eta} |D^\alpha \phi(x_1, x_2)|^p \right].$$

Remembering that I is the integrand in (1.3.4), we apply Hölder's inequality to (1.3.4) and use (1.3.5) to get

$$|\phi(t)| \leq \left(\text{vol}(K) \right)^{\frac{p-1}{p}} \left(\int_K |I|^p dx \right)^{1/p} \\ \leq (r^2)^{\frac{p-1}{p}} \left[\sum_{j=0}^2 \binom{2}{j} \left(\frac{1}{r} \right)^{\frac{jp}{p-1}} \right]^{\frac{p-1}{p}} \left[\sum_{\alpha \leq \eta} \int_K |D^\alpha \phi|^p \right]^{1/p}.$$

The constant in the inequality is just the constant M which appears in the statement of the theorem. Therefore

$$|\phi(t)| \leq M \left(\sum_{\alpha \leq \eta} \int_K |D^\alpha \phi|^p \right)^{1/p} \leq M \|\phi\|_{\eta, p, \Omega}.$$

Noting that t was arbitrary we have

$$(1.3.6) \quad \max_{t \in \bar{\Omega}} |\phi(t)| \leq M \|\phi\|_{\eta, p} \quad \forall \phi \in C^\eta(\bar{\Omega}).$$

This proves that (1.3.1) holds for all ϕ in $C^\eta(\bar{\Omega})$.

Now suppose $u \in W_p^\eta(\Omega)$, where $1 \leq p < \infty$. We are to show that u is continuous on $\bar{\Omega}$ and satisfies (1.3.1). More precisely, we shall show that u is equal almost everywhere (a.e.) to a function $\phi \in C(\bar{\Omega})$, and ϕ satisfies (1.3.1). By theorem 1.3.3 $C^\eta(\bar{\Omega})$ is dense in $W_p^\eta(\Omega)$, so there is a sequence (ϕ_j) of functions in $C^\eta(\bar{\Omega})$ which converges to u in the norm of $W_p^\eta(\Omega)$. In particular (ϕ_j) is a Cauchy sequence in $W_p^\eta(\Omega)$ and, as each $\phi_j - \phi_i$ satisfies (1.3.6), (ϕ_j) must also be a Cauchy sequence in $C(\bar{\Omega})$. As $C(\bar{\Omega})$ is complete, (ϕ_j) converges uniformly to some ϕ in $C(\bar{\Omega})$. Because $\bar{\Omega}$ is bounded, uniform convergence implies L_p convergence. Thus $\phi_j \rightarrow \phi$ in $L_p(\Omega)$. On the other hand, $\phi_j \rightarrow u$ in $W_p^\eta(\Omega)$, and W_p^η convergence also implies L_p convergence. Hence (ϕ_j) converges to both u and ϕ in $L_p(\Omega)$. This implies that $u = \phi$ a.e. Inequality (1.3.1), which holds for each ϕ_j , now follows for ϕ by continuity.

Finally we consider the case $p = \infty$. Suppose $u \in W_\infty^\eta(\Omega)$. Then, as Ω is bounded, $u \in W_q^\eta(\Omega)$ for $1 \leq q < \infty$. Therefore, as has already been proven, $u \in C(\bar{\Omega})$. Inequality (1.3.1) is now trivial. ||

Theorem 1.3.4, like the Sobolev lemma, has an obvious corollary.

Corollary 1.3.5: *Let Ω be a bounded, star-shaped domain satisfying cube condition. Let $1 \leq p \leq \infty$, let $\eta = (1, \dots, 1)$, and let β be a multiinteger such that $\eta \leq \beta$. Then $W_p^\beta(\Omega) \subseteq C^{\beta-\eta}(\bar{\Omega})$, and for all $u \in W_p^\beta(\Omega)$,*

$$\max_{\alpha \leq \beta - \eta} \max_{x \in \bar{\Omega}} |D^\alpha u(x)| \leq M \|u\|_{\beta, p}$$

where M is as in theorem 1.3.4.

(1.4) The Bramble-Hilbert Lemma

The Bramble-Hilbert lemma is a useful tool for proving error bounds in general. This theorem was popularized in the West by J.H. Bramble and S.R. Hilbert in their 1970 paper [4]. Similar results have appeared in the Soviet literature. An example is the theorem which is proved on page 354 of V. I. Smirnov's book [20]. This result was brought to my attention by Dr. Lois Mansfield.

Before we can state the Bramble-Hilbert lemma we must define a seminorm on $W_p^{(m)}(\Omega)$. Let

$$(1.4.1) \quad |u|_{m,p} = \left(\sum_{|\alpha|=m} \|D^\alpha u\|_{0,p}^p \right)^{1/p}$$

Here the sum is taken over all multiintegers with order exactly m , whereas in the norm $\|\cdot\|_{m,p}$ the sum also includes those α with order less than m .

As in the case of the Sobolev lemma some restrictions on the domain are required. The one major theorem upon which the Bramble-Hilbert lemma depends is the compact embedding theorem, the theorem which says that if $s_1 < s_2$ then $W_p^{(s_2)}(\Omega) \subseteq W_p^{(s_1)}(\Omega)$ with compact embedding. Various forms of this theorem are given in [10], [20], and [16], among other sources. The Bramble-Hilbert lemma holds on any domain for which the compact embedding theorem is valid. In particular, it holds if Ω is a convex polyhedron.

Theorem 1.4.1. (Bramble-Hilbert Lemma): *Let A be a bounded linear operator with domain $W_p^{(s)}(\Omega)$ ($1 \leq p < \infty$) and range in a normed linear space $(Y, \|\cdot\|)$.*

(Thus there exists a constant $\|A\|$ such that $\|Au\| \leq \|A\| \cdot \|u\|_{s,p}$ for all $u \in W_p^{(s)}(\Omega)$.) Suppose that A annihilates all polynomials of degree less than s . Then there is a constant C , depending on s and p but not on A ,

such that for all $u \in W_p^{(s)}(\Omega)$,

$$\|Au\| \leq C\|A\| \|u\|_{s,p}.$$

The original statement of the theorem referred to a functional F rather than an operator A . The switch to an operator makes the theorem easier to apply and does not in any way affect the proof of the theorem.

Bramble and Hilbert proved the result for integer values of s . The result also holds for noninteger values of s if we take the seminorm $|\cdot|_{s,p}$ to be the one given by (1.2.3). Here we present a proof of the noninteger case. The proof of the integer case is similar.

We begin by introducing some notation. Let $s = m + \sigma$, where m is a nonnegative integer and $0 < \sigma < 1$. Let P_m denote the space of polynomials of degree less than or equal to m , and let \mathcal{D} be the space

$$\mathcal{D} = \{u \in W_p^{(s)}(\Omega) \mid \int_{\Omega} D^{\alpha} u = 0 \ (\forall \alpha) \mid |\alpha| \leq m\}.$$

The theorem follows from two lemmas.

Lemma 1.4.2: $W_p^{(s)}(\Omega) = P_m \oplus \mathcal{D}$.

Lemma 1.4.3: There is a constant C such that for all $v \in \mathcal{D}$, $\|v\|_{s,p} \leq C|v|_{s,p}$.

The constant C appearing here is the same C as in the statement of the Bramble-Hilbert lemma. Before proving the two lemmas we show how the Bramble-Hilbert lemma follows from them.

Proof of Theorem 1.4.1: We are given a bounded linear operator $A: W_p^{(s)}(\Omega) \rightarrow Y$ which annihilates all polynomials of degree less than s . That is, $Ap = 0$ for all $p \in P_m$. Given $u \in W_p^{(s)}(\Omega)$ we can write $u = p + v$, where $p \in P_m$ and $v \in \mathcal{D}$, by lemma 1.4.2. We have $Au = Av$, so $\|Au\| = \|Av\| \leq \|A\| \cdot \|v\|_{s,p}$.

Therefore, by lemma 1.4.3,

$$(1.4.2) \quad \|Au\| \leq C\|A\| \cdot |v|_{s,p}.$$

This is almost the assertion of the Bramble-Hilbert lemma. We can complete the proof by showing that $|v|_{s,p} = |u|_{s,p}$. Recall that

$$|v|_{s,p} = \left(\sum_{|\alpha|=m} |D^\alpha v|_{\sigma,p}^p \right)^{1/p}$$

where $s = m + \sigma$ and

$$|D^\alpha v|_{\sigma,p} = \left(\int_{\Omega} \int_{\Omega} \frac{|D^\alpha v(x) - D^\alpha v(y)|^p}{\|x - y\|^{n+\sigma p}} dx dy \right)^{1/p}$$

Since $D^\alpha p$ is a constant if $|\alpha|=m$, we have $|D^\alpha v(x) - D^\alpha v(y)| = |D^\alpha u(x) - D^\alpha u(y)|$, so $|D^\alpha v|_{\sigma,p} = |D^\alpha u|_{\sigma,p}$, and therefore $|v|_{s,p} = |u|_{s,p}$. Combining this with (1.4.2) we get $\|Au\| \leq C\|A\| \cdot |u|_{s,p}$, which is the assertion of the Bramble-Hilbert lemma. ||

Proof of Lemma 1.4.2: We are to show that $W_p^{(s)}(\Omega) = P_m \oplus \mathcal{D}$. First we establish the fact that $P_m \cap \mathcal{D} = (0)$. This is equivalent to showing that if $p \in P_m$ and $\int_{\Omega} D^\alpha p = 0$ for all α such that $|\alpha| \leq m$, then $p=0$. It is a simple matter to prove this by induction on m . There is no need to include the argument here.

We now prove that $W_p^{(s)}(\Omega) = P_m + \mathcal{D}$. The dimension of the space P_m is the number of monomials $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$ such that $\alpha_1 + \dots + \alpha_n \leq m$. That is, it is just the number of multiintegers α such that $|\alpha| \leq m$. Call this number k . Define a linear transformation $T: P_m \rightarrow \mathbb{R}^k$ by

$$Tp = (\int_{\Omega} D^\alpha p, \int_{\Omega} D^\beta p, \dots, \int_{\Omega} D^\gamma p)$$

where $\alpha, \beta, \dots, \gamma$ are the k multiintegers of order not exceeding m . By the previous paragraph T is one-to-one. Therefore, as $\dim(P_m) = k = \dim(\mathbb{R}^k)$, T is also onto. Thus, for any constants c_α ($|\alpha| \leq m$) there is a unique

$p \in P_m$ such that $c_\alpha = \int_\Omega D^\alpha p$ for all $|\alpha| \leq m$. In particular, given $u \in W_p^{(s)}(\Omega)$ there is a unique $p \in P_m$ such that $\int_\Omega D^\alpha u = \int_\Omega D^\alpha p$ for all α such that $|\alpha| \leq m$. Letting $v = u - p$ we have $\int_\Omega D^\alpha v = 0$ if $|\alpha| \leq m$, so $v \in \mathcal{D}$. Thus $u = p + v$, where $p \in P_m$ and $v \in \mathcal{D}$. This proves that $W_p^{(s)}(\Omega) = P_m + \mathcal{D}$. \parallel

Proof of Lemma 1.4.3: We are to prove the existence of a constant C such that $\|v\|_{s,p} \leq C|v|_{s,p}$ for all $v \in \mathcal{D}$. Assume that no such C exists. Then there exists a sequence (v_j) of functions from \mathcal{D} such that

$$\|v_j\|_{s,p} > j|v_j|_{s,p} \quad j=1,2,3\dots$$

We may assume that $\|v_j\|_{s,p} = 1$ for all j . By the compact embedding theorem the sequence (v_j) has a subsequence $(w_i) = (v_{j_i})$ which converges in the norm of $W_p^{(m)}(\Omega)$. In particular $\|w_i - w_k\|_{m,p} \rightarrow 0$ as $i, k \rightarrow \infty$. Note also that

$$|w_i|_{s,p} = |v_{j_i}|_{s,p} < (j_i)^{-1} \|v_{j_i}\|_{s,p} = (j_i)^{-1}$$

Thus $|w_i|_{s,p} \rightarrow 0$ as $i \rightarrow \infty$. Therefore

$$\begin{aligned} \|w_i - w_k\|_{s,p} &= (\|w_i - w_k\|_{m,p}^p + |w_i - w_k|_{s,p}^p)^{1/p} \\ &\leq \left(\|w_i - w_k\|_{m,p}^p + (|w_i|_{s,p} + |w_k|_{s,p})^p \right)^{1/p} \rightarrow 0 \end{aligned}$$

as $i, k \rightarrow \infty$. Thus (w_i) is a Cauchy sequence in the complete space $W_p^{(s)}(\Omega)$, and therefore (w_i) converges to some w in $W_p^{(s)}(\Omega)$.

We shall prove that w is in P_m . First of all, $|w|_{s,p} = \lim_{i \rightarrow \infty} |w_i|_{s,p} = 0$. Therefore, as $|w|_{s,p}^p = \sum_{|\alpha|=m} |D^\alpha w|_{\sigma,p}^p$ (where $s = m + \sigma$) we see that $|D^\alpha w|_{\sigma,p} = 0$ for all α such that $|\alpha| = m$. By definition of $|\cdot|_{\sigma,p}$,

$$0 = |D^\alpha w|_{\sigma,p}^p = \int_\Omega \int_\Omega \frac{|D^\alpha w(x) - D^\alpha w(y)|^p}{\|x - y\|^{n+p\sigma}} dx dy.$$

Therefore, for almost all y in Ω ,

$$0 = \int_\Omega \frac{|D^\alpha w(x) - D^\alpha w(y)|^p}{\|x - y\|^{n+p\sigma}} dx.$$

Let y_0 be one value of y such that this integral is zero. Then, for almost all x in Ω ,

$$0 = \frac{|D^\alpha w(x) - D^\alpha w(y_0)|^p}{\|x - y_0\|^{n+p\sigma}}$$

Therefore $D^\alpha w(x) = D^\alpha w(y_0)$ for almost all x , and $D^\alpha u$ is (equivalent to) a constant function. This is true for all α such that $|\alpha|=m$, so w must be a polynomial of degree at most m , i.e. $w \in P_m$.

On the other hand, $w \in \mathcal{D}$, for w is the limit of the sequence (w_i) of functions in \mathcal{D} and, as is easily verified, \mathcal{D} is a closed subspace of $W_p^{(s)}(\Omega)$. Therefore, by lemma 1.4.2, $w=0$. But this contradicts the fact that $\|w\|_{s,p} = \lim_{i \rightarrow \infty} \|w_i\|_{s,p} = 1$.

(1.5) A Variant of the Bramble-Hilbert Lemma

As in the case of the Sobolev lemma, there is a variant of the Bramble-Hilbert lemma which is more appropriate for blending-function methods. We shall approach this theorem via a series of lemmas.

Let Ω be a bounded domain in \mathbb{R}^n , and let $1 \leq p \leq \infty$. We define linear operators $S_i: L_p(\Omega) \rightarrow L_p(\Omega)$, $i=1, \dots, n$, by $S_i u = D_i u$, where the domain of S_i is

$$\mathcal{D}(S_i) = \{u \in L_p(\Omega) \mid \exists \text{ weak } D_i u \in L_p(\Omega)\}.$$

It is easy to show that S_i is a closed operator.

Lemma 1.5.1: S_i has a closed, bijective restriction T_i .

Proof: Let $B = \prod_{j=1}^n [a_j, b_j]$ be a box which contains $\bar{\Omega}$. Given $f \in L_p(\Omega)$ we can extend f to all of B by setting f equal to zero on $B \setminus \Omega$. Define a function u_i on B by

$$(1.5.1) \quad u_i(x_1, \dots, x_n) = \int_{a_i}^{x_i} f(x_1, \dots, x_{i-1}, t_i, x_{i+1}, \dots, x_n) dt_i.$$

This integral is defined for all $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ in $\prod_{j \neq i} [a_j, b_j]$, except on a set Z_i of measure zero. Let $Z = [a_i, b_i] \times Z_i$. Then Z has measure zero in B and the integral (1.5.1) is well defined for all x outside of Z .

Clearly $D_i u_i = f$ in the weak sense, and therefore $S_i u_i = f$. We define a linear operator $A_i: L_p(\Omega) \rightarrow L_p(\Omega)$ by $A_i f = u_i|_{\Omega}$. It is easy to calculate from (1.5.1) that A_i is bounded (in fact, compact) and everywhere defined. If we can show that A_i is one-to-one, it will follow that $T_i \equiv A_i^{-1}$ is a closed, bijective restriction of S_i .

To prove that A_i is one-to-one we must show that if $u_i|_{\Omega} = 0$ then $f=0$ a.e. If $u_i|_{\Omega} = 0$, then $u_i=0$ on all of B because $f=0$ on $B \setminus \Omega$. Thus

$$(1.5.2) \quad 0 = \int_{a_i}^{x_i} f(x_1, \dots, x_{i-1}, t_i, x_{i+1}, \dots, x_n) dt_i; \quad a_i \leq x_i \leq b_i$$

for all $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \left[\prod_{j \neq i} [a_j, b_j] \right] \setminus Z_i$.

Let $B_i = \prod_{j \neq i} [a_j, b_j]$ and fix $\hat{x} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in B_i \setminus Z_i$. Let $S_{\hat{x}}$ be the set of all $t_i \in [a_i, b_i]$ such that $f(x_1, \dots, x_{i-1}, t_i, x_{i+1}, \dots, x_n) \neq 0$. By (1.5.2) $S_{\hat{x}}$ has Lebesgue measure zero, $m_1(S_{\hat{x}}) = 0$. Let S be the set of all x in B such that $f(x) \neq 0$. Then, by definition, the measure of S is

$$m_n(S) = \int_{B_i} m_1(S_{\hat{x}}) d\hat{x} = 0.$$

Thus $f=0$ a.e. This proves that A_i is one-to-one. ||

Lemma 1.5.2: The operators T_1, \dots, T_n which were constructed in the proof of lemma 1.5.1 commute with one another.

Proof: We defined T_i by $T_i = A_i^{-1}$, where

$$A_i f(x_1, \dots, x_n) = \int_{a_i}^{x_i} f(x_1, \dots, x_{i-1}, t_i, x_{i+1}, \dots, x_n) dt_i.$$

By Fubini's theorem A_1, \dots, A_n commute. It follows that their inverses T_1, \dots, T_n commute. ||

Lemma 1.5.3: Let $B = \prod_{j=1}^n [a_j, b_j]$ be any box containing Ω . Then for all u in the domain of T_i ,

$$\|u\|_{0,p} \leq (b_i - a_i) \|T_i u\|_{0,p} \quad i=1, \dots, n.$$

Proof: This is equivalent to saying that $\|A_i\| \leq b_i - a_i$, which can be verified by applying Hölder's inequality to (1.5.1) and integrating. ||

Let γ be a multiinteger, and define S_γ to be the operator given by $S_\gamma u = D^\gamma u$, where the domain of S_γ is

$$\mathcal{D}(S_\gamma) = \{u \in L_p(\Omega) \mid \exists \text{ weak } D^\gamma u \in L_p(\Omega)\}.$$

Lemma 1.5.4: S_γ has a restriction T_γ which is closed and bijective.

Proof: Let $T_\gamma = \prod_{i=1}^n T_i^{\gamma_i}$. By lemma 1.5.2 the order of the factors T_i is immaterial. T_γ clearly has the desired properties. ||

Lemma 1.5.5: Let T_γ be as defined in the proof of the previous lemma. Then for all u in the domain of T_γ ,

$$\|u\|_{0,p} \leq \prod_{i=1}^n (b_i - a_i)^{\gamma_i} \|T_\gamma u\|_{0,p}.$$

Proof: Apply lemma 1.5.3 repeatedly. ||

We are now prepared to state and prove the modified Bramble-Hilbert lemma.

Theorem 1.5.6: Let $1 \leq p \leq \infty$, let Ω be a bounded domain in \mathbb{R}^n , and let A be a bounded linear operator with domain $W_p^\beta(\Omega)$ and range in some normed linear space $(Y, \|\cdot\|)$. (Thus $\|Au\| \leq \|A\| \cdot \|u\|_{\beta,p}$ for all $u \in W_p^\beta(\Omega)$.) Suppose that A annihilates every function $\phi \in W_p^\beta(\Omega)$ such that $D^\beta \phi = 0$. Then for all

$u \in W_p^\beta(\Omega)$

$$\|Au\| \leq B\|A\| \cdot \|D^\beta u\|_{0,p}$$

where

$$B = \left(\sum_{\alpha \leq \beta} \left[\prod_{i=1}^n (b_i - \alpha_i)^{\beta_i - \alpha_i} \right]^p \right)^{1/p}$$

If Ω is the unit hypercube U , then $B = \left[\prod_{i=1}^n (\beta_i + 1) \right]^{1/p}$

Proof: Let $u \in W_p^\beta(\Omega)$. Since T_β is surjective (lemma 1.5.4), there exists v in the domain of T_β such that $T_\beta v = D^\beta u$. We have $D^\beta(u-v) = 0$, so $A(u-v) = 0$. Thus

$$(1.5.3) \quad \|Au\| = \|Av\| \leq \|A\| \cdot \|v\|_{\beta,p}$$

Let $\alpha \leq \beta$. Then, as the operators T_i commute (lemma 1.5.2), $T_\beta = T_{\beta-\alpha} T_\alpha$.

Therefore $T_\alpha v$ is in the domain of $T_{\beta-\alpha}$. Applying lemma 1.5.5 with u replaced by $T_\alpha v$ and γ replaced by $\beta-\alpha$, we have

$$\|D^\alpha v\|_{0,p} \leq \left[\prod_{i=1}^n (b_i - \alpha_i)^{\beta_i - \alpha_i} \right] \|D^\beta v\|_{0,p}$$

This holds for all $\alpha \leq \beta$, so

$$\begin{aligned} \|v\|_{\beta,p} &= \left[\sum_{\alpha \leq \beta} \|D^\alpha v\|_{0,p}^p \right]^{1/p} \leq \left\{ \sum_{\alpha \leq \beta} \left[\prod_{i=1}^n (b_i - \alpha_i)^{\beta_i - \alpha_i} \right]^p \right\}^{1/p} \cdot \|D^\beta v\|_{0,p} \\ &= B \|D^\beta v\|_{0,p} \end{aligned}$$

Combining this last inequality with 1.5.3 and recalling that $D^\beta v = D^\beta u$

we have

$$\|Au\| \leq B\|A\| \cdot \|D^\beta u\|_{0,p}$$

which is the assertion of the theorem.

If $\Omega=U$, then $b_i - \alpha_i = 1$ for all i , so $\sum_{\alpha \leq \beta} \left[\prod_{i=1}^n (b_i - \alpha_i)^{\beta_i - \alpha_i} \right]^p = \sum_{\alpha \leq \beta} 1 = \prod_{i=1}^n (\beta_i + 1)$.

Hence $B = \left[\prod_{i=1}^n (\beta_i + 1) \right]^{1/p}$.

It is interesting to note that the only hypothesis on Ω in theorem 1.5.6 is that Ω be bounded. There are no regularity conditions whatsoever.

The proof of theorem 1.5.6 is constructive and allows us to compute a specific constant B . By contrast, the standard Bramble-Hilbert lemma gives us no idea of the size of the constant C appearing in that theorem.

CHAPTER TWO

ERROR BOUNDS FOR BLENDING-FUNCTION METHODS

(2.1) Blending-Function Interpolation

In this chapter blending-function methods are introduced, some of their elementary properties are established, and asymptotic error bounds for blending-function interpolation are obtained. Blending-function methods are a class of methods of generating functions (blended interpolants) which interpolate curves or surfaces rather than points. The nature of a blended interpolant depends on the "blending functions" used. In this thesis we shall consider only two-point Hermite polynomials as blending functions. Other possible choices of blending functions are Lagrange polynomials, splines, trigonometric polynomials, etc. The reader is referred to Gordon's very general discussion of blending-function methods [11]. The decision to consider only Hermite polynomial blending functions in this thesis was dictated by a desire to keep the notation simple. The proofs given here can be applied to any blending-function scheme in which the blending functions are polynomials. Thus, for example, the methods of this chapter can be applied to Lagrange polynomial blending functions.

The error bounds of Gordon and Hall [12] mentioned in the introduction are stated for the case of Lagrange polynomial blending functions but could also be applied to other cases. The error bounds of [12] are applicable to continuously differentiable functions, whereas the main results given here apply to weakly differentiable functions. Error bounds for continuously differentiable functions are also given here.

The technique for obtaining error bounds is fundamentally the same as that used by Bramble and Zlamal [6]. Where they have used the Sobolev lemma and the Bramble-Hilbert lemma, we shall use the variants of these theorems which were presented in chapter one.

The developments of this chapter are stated for planar regions, but the entire theory carries over to n -space for arbitrary n . We stick to the special case $n=2$ for simplicity of presentation.

We begin the technical discussion by introducing the blending functions themselves, the two-point Hermite polynomials. Let k be a positive integer which will remain fixed throughout this chapter. Define $2k$ polynomials

$p_0, \dots, p_{k-1}, q_0, \dots, q_{k-1}$ as follows:

Let p_i and q_i , $i=0, \dots, k-1$, be the unique polynomials of degree less than $2k$ such that

$$(2.1.1) \quad \left. \begin{aligned} p_i^{(j)}(0) &= \delta_{ij} = q_i^{(j)}(1) \\ p_i^{(j)}(1) &= 0 = q_i^{(j)}(0) \end{aligned} \right\} i, j=0, 1, \dots, k-1.$$

Let U be the open unit square in \mathbb{R}^2 . We use the blending functions to define an operator P_1 on $C^{(k-1,0)}(\bar{U})$ by

$$(2.1.2) \quad P_1 u(x_1, x_2) = \sum_{j=0}^{k-1} \left[D_1^j u(0, x_2) p_j(x_1) + D_1^j u(1, x_2) q_j(x_1) \right].$$

It is clear from (2.1.2) that if $u \in C^{(k-1,i)}(\bar{U})$ for some i , then

$P_1 u \in C^{(\infty,i)}(\bar{U})$. Also from the properties (2.1.1), $P_1 u$ interpolates u and its first $k-1$ normal derivatives along the sides $x_1=0$ and $x_1=1$ of \bar{U} . In

fact, for each x_2 , $P_1 u$ (as a function of x_1) is the unique polynomial of degree less than $2k$ which interpolates u and its first $k-1$ derivatives (with

respect to x_1) at the endpoints $x_1=0$ and $x_1=1$. Therefore, if P_1 is applied to $P_1 u$ we will get back the same polynomial again. This holds for each x_2 , so $P_1^2 u = P_1 u$. That is, P_1 is a projector. The range of P_1 is the set of functions $v \in C^{(2k,0)}(\bar{U})$ such that for each fixed $x_2 \in [0,1]$, v is a polynomial in x_1 of degree less than $2k$. Equivalently, the range of P_1 is the set of all $v \in C^{(2k,0)}(\bar{U})$ such that $D^{(2k,0)} v \equiv D_1^{2k} v = 0$. The range of a projector Q is exactly the space of elements v such that $Qv=v$. We say that Q preserves or is exact for such functions. Thus P_1 is exact for the set $\{v \in C^{(2k,0)}(\bar{U}) \mid D^{(2k,0)} v = 0\}$.

The operator P_2 , the "mirror image" of P_1 , is defined on $C^{(0,k-1)}(\bar{U})$ by

$$(2.1.3) \quad P_2 u(x_1, x_2) = \sum_{j=0}^{k-1} \left[D_2^j u(x_1, 0) p_j(x_2) + D_2^j u(x_1, 1) q_j(x_2) \right].$$

Clearly $P_2 u$ interpolates u and its first $k-1$ normal derivatives on the sides $x_2=0$ and $x_2=1$ of \bar{U} . P_2 is a projector which preserves all $v \in C^{(0,2k)}(\bar{U})$ such that $D^{(0,2k)} v = 0$.

If $u \in C^{(k-1,k-1)}(\bar{U})$, then $P_1 P_2 u$ and $P_2 P_1 u$ are defined and equal. A direct computation shows that

$$(2.1.4) \quad P_1 P_2 u = P_2 P_1 u = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \left[D^{(i,j)} u(0,0) p_i(x_1) p_j(x_2) + D^{(i,j)} u(1,0) q_i(x_1) p_j(x_2) + D^{(i,j)} u(0,1) p_i(x_1) q_j(x_2) + D^{(i,j)} u(1,1) q_i(x_1) q_j(x_2) \right].$$

It is clear from this representation that $P_1 P_2 u \in C^\infty(\bar{U})$.

We define another projector on $C^{(k-1,k-1)}(\bar{U})$ by either of the two

equivalent definitions

$$(2.1.5) \quad P = P_1 \oplus P_2 \equiv P_1 + P_2 - P_1 P_2$$

$$I - P = (I - P_1)(I - P_2).$$

Because P_1 and P_2 commute, P is a projector. Pu is called the *blended interpolant* of u . Theorem 2.1.1 below gives the justification for the use of the term "interpolant." The *error projector* E is defined by $E = I - P$. If we define $E_1 = I - P_1$ and $E_2 = I - P_2$, then $E = E_1 E_2 = E_2 E_1$.

Theorem 2.1.1: Given $u \in C^{(k-1, k-1)}(\bar{U})$, Pu is completely determined by the boundary values of u and its first $k-1$ normal derivatives. Pu interpolates these values of u .

Proof: An inspection of (2.1.2), (2.1.3), and (2.1.4) shows that Pu is determined by the boundary values of u and its first $k-1$ normal derivatives. To prove that u interpolates these values we must show that along the sides $x_1=0$ and $x_1=1$,

$$(2.1.7) \quad D_1^j Eu = 0 \quad j=0, \dots, k-1$$

and along the sides $x_2=0$ and $x_2=1$,

$$(2.1.8) \quad D_2^j Eu = 0 \quad j=0, \dots, k-1.$$

We shall prove (2.1.7). An analogous argument proves (2.1.8). Let $v = E_2 u$. Then $Eu = E_1 v$. We know that $P_1 v$ interpolates v and its first $k-1$ normal derivatives along the sides $x_1=0$ and $x_1=1$. This means

$$D_1^j E_1 v = 0 \quad j=0, \dots, k-1$$

along these sides. This, together with the fact that $Eu = E_1 v$, proves (2.1.7).||

Theorem 2.1.2: Let β be a multiinteger satisfying $(k, k) \leq \beta \leq (2k, 2k)$. Suppose $u \in W_p^\beta(U)$ and $D^\beta u = 0$. Then $Eu=0$, i.e. u is preserved by P .

Proof: Starting with the equation $D^\beta u=0$ and performing $|\beta|$ integrations we see that u is of the form

$$u(x_1, x_2) = \sum_{j=0}^{\beta_1-1} \phi_j(x_2) x_1^j + \sum_{j=0}^{\beta_2-1} \psi_j(x_1) x_2^j$$

where $\phi_j \in W_p^{(\beta_2)}(0,1)$, $j=0, \dots, \beta_1-1$, and $\psi_j \in W_p^{(\beta_1)}(0,1)$, $j=0, \dots, \beta_2-1$. Letting

$$v(x_1, x_2) = \sum_{j=0}^{\beta_1-1} \phi_j(x_2) x_1^j$$

$$w(x_1, x_2) = \sum_{j=0}^{\beta_2-1} \psi_j(x_1) x_2^j$$

we have $u=v+w$, $E_1 v=0$, and $E_2 w=0$. Thus $Eu = Ev + Ew = E_2 E_1 v + E_1 E_2 w = 0$. ||

Theorem 2.1.2 indicates that it may be possible to apply the modified Bramble-Hilbert lemma (theorem 1.5.6) to the operator E or, more generally, $D^\alpha E$. Theorem 1.5.6 requires that $D^\alpha E$ be bounded in some sense. The boundedness of $D^\alpha E$ is proven in the following section.

Before proceeding to the next section we examine two special cases which will be used in the construction of finite elements. Consider the case $k=1$. The blending functions are the linear polynomials $p_0(x) = 1-x$ and $q_0(x) = x$. The function $P_1 u(x_1, x_2)$ interpolates linearly from $u(0, x_2)$ across to $u(1, x_2)$ for each x_2 . $P_1 P_2 u$ is the unique bilinear function which interpolates u at the four corners of \bar{U} . Pu is determined by and interpolates the boundary values of u . $E=I-P$ annihilates all functions $u \in W_p^{(2,2)}(U)$ such that $D^{(2,2)} u=0$. Among such u are all monomials of the form $x_1^i x_2^j$, where either $i \leq 1$ or $j \leq 1$. We state the following corollary for future reference.

Corollary 2.1.3: Let $k=1$. Then i) the blended interpolant Pu is completely

determined by the boundary values of u and interpolates u on the boundary of U , and ii) P preserves all monomials of the form $x_1^i x_2^j$, where $i \leq 1$ or $j \leq 1$. In particular, P preserves all cubic polynomials.

Now consider the case $k=2$. In this case the blending functions are cubic polynomials. For example $q_0(x) = 3x^2 - 2x^3$ and $q_1(x) = x^3 - x^2$. For each fixed x_1 , $P_2 u$ is the unique cubic polynomial in x_2 which interpolates u and its first derivative (with respect to x_2) at the endpoints $x_2=0$ and $x_2=1$. Pu is determined by and interpolates the boundary values of u and its normal derivative. The error operator E annihilates all $u \in W_p^{(4,4)}(U)$ such that $D^{(4,4)} u = 0$. Thus E annihilates all monomials of the form $x_1^i x_2^j$, where $i \leq 3$ or $j \leq 3$.

Corollary 2.1.4: Let $k=2$. Then i) the blended interpolant Pu is completely determined by the boundary values of u and its normal derivative. Pu interpolates u and its normal derivative on the boundary of u , and ii) P preserves all monomials of the form $x_1^i x_2^j$, where $i \leq 3$ or $j \leq 3$. In particular, P preserves all polynomials of degree seven or less.

(2.2) Boundedness of the Error Operator

In order to get a bound for $D^\alpha E$ we obtain bounds for $D^\alpha I$, $D^\alpha P_1$, $D^\alpha P_2$, and $D^\alpha P_1 P_2$. These will yield a bound for $D^\alpha E$, as $E = I - P_1 - P_2 + P_1 P_2$. The use of the modified Sobolev lemma (theorem 1.3.4) will be demonstrated in this section. We assume throughout that $1 \leq p \leq \infty$. The multiinteger $(1,1)$ will be denoted by η .

Lemma 2.2.1: Let α be any multiinteger, Then for all $u \in W_p^{\alpha+\eta}(U)$

$$\|D^\alpha u\|_{0,p} \leq M \|u\|_{\alpha+\eta,p}$$

where M is as in theorem 1.3.4.

Proof: This lemma is trivial because $M \geq 1$ in the case under consideration. ||

Lemma 2.2.2: Let α be any multiinteger. Then

$$\begin{aligned} \|D^{\alpha} P_1 u\|_{0,p} &\leq C_{\alpha_1} M \|u\|_{(k, \alpha_2+1), p} & \forall u \in W_p^{(k, \alpha_2+1)}(U) \\ \|D^{\alpha} P_2 u\|_{0,p} &\leq C_{\alpha_2} M \|u\|_{(\alpha_1+1, k), p} & \forall u \in W_p^{(\alpha_1+1, k)}(U) \end{aligned}$$

where M is as in theorem 1.3.4, and

$$C_{\alpha_i} = \max_{0 \leq x \leq 1} \left[\sum_{j=0}^{k-1} |p_j^{(\alpha_i)}(x)| + |q_j^{(\alpha_i)}(x)| \right] \quad i=1,2..$$

Proof: We shall prove only the first inequality. Let $u \in W_p^{(k, \alpha_2+1)}(U)$.

Then, by the corollary to the modified Sobolev lemma (corollary 1.3.5),

$u \in C^{(k-1, \alpha_2)}(\bar{U})$, so $P_1 u$ is well defined and contained in $C^{(\infty, \alpha_2)}(\bar{U}) \subseteq C^{\alpha}(\bar{U})$.

Applying D^{α} to equation (2.1.2) we have

$$D^{\alpha} P_1 u(x_1, x_2) = \sum_{j=0}^{k-1} \left[D^{(j, \alpha_2)} u(0, x_2) p_j^{(\alpha_1)}(x_1) + D^{(j, \alpha_2)} u(1, x_2) q_j^{(\alpha_1)}(x_1) \right].$$

Thus

$$|D^{\alpha} P_1 u(x_1, x_2)| \leq \max_{x \in \bar{U}} |D^{(j, \alpha_2)} u(x)| \left[\sum_{j=0}^{k-1} \left(|p_j^{(\alpha_1)}(x_1)| + |q_j^{(\alpha_1)}(x_1)| \right) \right] \quad 0 \leq j \leq k-1$$

It follows from corollary 1.3.5 that

$$\max_{x \in \bar{U}} |D^{(j, \alpha_2)} u(x)| \leq M \|u\|_{(k, \alpha_2+1), p} \quad 0 \leq j \leq k-1$$

so $|D^{\alpha} P_1 u(x_1, x_2)| \leq C_{\alpha_1} M \|u\|_{(k, \alpha_2+1), p}$. Now take p th powers, integrate the left hand side over U , and take p th roots to finish the proof. ||

Lemma 2.2.3: Let α be a multiinteger. Then for all $u \in W^{(k, k)}(U)$,

$$\|D^{\alpha} P_1 P_2 u\|_{0,p} \leq C_{\alpha_1 \alpha_2} M \|u\|_{(k, k), p}$$

where

$$C_{\alpha_1\alpha_2} = \max_{(x_1, x_2) \in \bar{U}} \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \left[\left(|p_i^{(\alpha_1)}(x_1)| + |q_i^{(\alpha_1)}(x_1)| \right) \left(|p_j^{(\alpha_2)}(x_2)| + |q_j^{(\alpha_2)}(x_2)| \right) \right].$$

The proof is similar to that of lemma 2.2.2, the representation (2.1.4) being used.

Theorem 2.2.4: Let α and β be multiintegers satisfying $\alpha + \eta \leq \beta$ and $(k, k) \leq \beta$. Then for all $u \in W_p^\beta(U)$,

$$(2.2.1) \quad \|D^\alpha Eu\|_{0,p} \leq C_\alpha M \|u\|_{\beta,p}$$

where $C_\alpha = 1 + C_{\alpha_1} + C_{\alpha_2} + C_{\alpha_1\alpha_2}$.

Proof: From the hypotheses we have $\alpha + \eta \leq \beta$, $(k, \alpha_2 + 1) \leq \beta$, $(\alpha_1 + 1, k) \leq \beta$, and $(k, k) \leq \beta$, so by the last three lemmas,

$$\begin{aligned} \|D^\alpha u\|_{0,p} &\leq M \|u\|_{\beta,p} \\ \|D^{\alpha P_1} u\|_{0,p} &\leq C_{\alpha_1} M \|u\|_{\beta,p} \\ \|D^{\alpha P_2} u\|_{0,p} &\leq C_{\alpha_2} M \|u\|_{\beta,p} \\ \|D^{\alpha P_1 P_2} u\|_{0,p} &\leq C_{\alpha_1 \alpha_2} M \|u\|_{\beta,p}. \end{aligned}$$

Application of the triangle inequality to $D^\alpha Eu = D^\alpha u - D^{\alpha P_1} u - D^{\alpha P_2} u + D^{\alpha P_1 P_2} u$ and addition of these four inequalities gives (2.2.1). ||

Theorem 2.2.5: Let α and β be multiintegers satisfying $\alpha + \eta \leq \beta$ and $(k, k) \leq \beta \leq (2k, 2k)$. Then for all $u \in W_p^\beta(U)$,

$$(2.2.2) \quad \|D^\alpha Eu\|_{0,p} \leq BC_\alpha M \|D^\beta u\|_{0,p}$$

where B is as in theorem 1.5.6, C_α is as in theorem 2.2.4, and M is as in theorem 1.3.4.

Proof: Let $A: W_p^\beta(U) \rightarrow L_p(U)$ be the operator given by $Au = D^\alpha Eu$. By theorem 2.2.4 A is bounded. By theorem 2.1.2 $Au=0$ whenever $D^\beta u=0$. Therefore we can apply theorem 1.5.6 to obtain (2.2.2). ||

(2.3) Sobolev Space Error Bounds

Let $R = (a_1, b_1) \times (a_2, b_2)$ be a rectangle in the w -plane. Let $h_1 = b_1 - a_1$ and $h_2 = b_2 - a_2$. The affine transformation

$$(2.3.1) \quad w_i = h_i x_i + a_i \quad i=1,2$$

is a one-to-one mapping of \bar{U} in the x -plane onto \bar{R} in the w -plane. There is an obvious correspondence between functions u defined on \bar{U} and functions \tilde{u} defined on \bar{R} . This correspondence is given by $u \leftrightarrow \tilde{u}$, where $\tilde{u}(w_1, w_2) = u(x_1, x_2)$. Here (w_1, w_2) and (x_1, x_2) are linked by (2.3.1). Obviously \tilde{u} retains the essential properties of u , and vice versa. For instance, $u \in W_p^\beta(U)$ if and only if $\tilde{u} \in W_p^\beta(R)$.

The projector P on $C^{(k-1, k-1)}(\bar{U})$ induces a projector \tilde{P} on $C^{(k-1, k-1)}(\bar{R})$ by $\tilde{P}\tilde{u} = \tilde{P}u$. It is easy to see that \tilde{P} retains the important characteristics of P . Specifically, theorems 2.1.1 and 2.1.2 hold for \tilde{P} . We define an error projector \tilde{E} by $\tilde{E}\tilde{u} = \tilde{u} - \tilde{P}\tilde{u}$. Clearly $\tilde{E}\tilde{u} = \tilde{E}u$.

We shall establish an asymptotic bound for the L_p norm of $D^{\alpha} \tilde{E}\tilde{u}$. In order to do this we shall first establish the link between the norm of $D^{\alpha} v$ in $L_p(U)$ and the norm of $D^{\alpha} \tilde{v}$ in $L_p(R)$ for any $v \in W_p^{\alpha}(U)$. Then we shall use the bound (2.2.2) on $D^{\alpha} E u$ to establish the asymptotic bound for $D^{\alpha} \tilde{E}\tilde{u}$.

In the following lemma $D^{\alpha} v$ means $\frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}$, and $D^{\alpha} \tilde{v}$ means $\frac{\partial^{|\alpha|} \tilde{v}}{\partial w_1^{\alpha_1} \partial w_2^{\alpha_2}}$.

Lemma 2.3.1: Let J be the Jacobian of the transformation (2.3.1), and let α be any multiinteger. Then for all $v \in W_p^{\alpha}(U)$,

$$(2.3.2) \quad \|D^{\alpha} v\|_{0,p,U} = h_1^{\alpha_1} h_2^{\alpha_2} J^{-1/p} \|D^{\alpha} \tilde{v}\|_{0,p,R}$$

Proof: The differentiation formula

$$D^\alpha v(x_1, x_2) = h_1^{\alpha_1} h_2^{\alpha_2} D^\alpha v(w_1, w_2)$$

is easily verified. Taking p th powers, integrating, and taking p th roots, we get (2.3.2). ||

Theorem 2.3.2: Let α and β be multiintegers satisfying $\alpha + \eta \leq \beta$ and $(k, k) \leq \beta \leq (2k, 2k)$. Then for all $\tilde{u} \in W_p^\beta(R)$,

$$(2.3.3) \quad \|D^{\alpha} \tilde{E} \tilde{u}\|_{0,p} \leq BC_\alpha M h_1^{\beta_1 - \alpha_1} h_2^{\beta_2 - \alpha_2} \|D^{\beta} \tilde{u}\|_{0,p}$$

where B , C_α , and M are as in theorems 1.5.6, 2.2.4, and 1.3.4, respectively.

Proof: By (2.3.2) with v replaced by Eu ,

$$\|D^{\alpha} \tilde{E} \tilde{u}\|_{0,p,R} = h_1^{-\alpha_1} h_2^{-\alpha_2} J^{1/p} \|D^{\alpha} Eu\|_{0,p,U}$$

and by (2.3.2) with v replaced by u and α replaced by β ,

$$\|D^{\beta} u\|_{0,p,U} = h_1^{\beta_1} h_2^{\beta_2} J^{-1/p} \|D^{\beta} \tilde{u}\|_{0,p,R}$$

We now sandwich inequality (2.2.2) between these two lines to obtain (2.3.3). ||

Theorem 2.3.3: Let $h = \max\{h_1, h_2\}$, and suppose $h \leq 1$. Let m be a nonnegative integer less than $2k$, and let β be a multiinteger satisfying $(k, k) \leq \beta \leq (2k, 2k)$ and $(m, m) + \eta \leq \beta$. Then for all $\tilde{u} \in W_p^\beta(R)$,

$$(2.3.4) \quad \|\tilde{E} \tilde{u}\|_{m,p} \leq BC_m M h^{|\beta| - m} \|D^{\beta} \tilde{u}\|_{0,p}$$

where $C_m = \left[\sum_{|\alpha| \leq m} C_\alpha^p \right]^{1/p}$. B , C_α , and M are as in theorems 1.5.6, 2.2.4, and 1.3.4 respectively.

Proof: Let α be such that $|\alpha| \leq m$. Then $\alpha + \eta \leq \beta$, so (2.3.3) holds. Therefore

$$\begin{aligned} \|D^{\alpha} \tilde{E} \tilde{u}\|_{0,p} &\leq BC_\alpha M h^{|\beta| - |\alpha|} \|D^{\beta} \tilde{u}\|_{0,p} \\ &\leq BC_\alpha M h^{|\beta| - m} \|D^{\beta} \tilde{u}\|_{0,p} \end{aligned}$$

as $h \leq 1$ and $|\alpha| \leq m$. Taking p th powers, summing over all α such that $|\alpha| \leq m$, and taking p th roots, we obtain (2.3.4). $\| \cdot \|$

(2.4) Error Bounds for Continuously Differentiable Functions

The procedure which has been used to obtain Sobolev space error bounds can be used also to obtain error bounds for continuously differentiable functions. If we are willing to work with continuously differentiable functions we do not have to use a Sobolev lemma type theorem, and the proofs are correspondingly easier. It is also possible to relax some of the hypotheses.

The modified Bramble-Hilbert lemma (theorem 1.5.6) is still required and remains true with $W_p^\beta(\Omega)$ replaced by $C^\beta(\bar{\Omega})$ and $\| \cdot \|_{0,p}$ replaced by $\| \cdot \|_{0,\infty}$. Indeed the proofs of some of the lemmas leading up to theorem 1.5.6 are technically simpler in this case.

Theorem 2.1.2 holds with $W_p^\beta(\Omega)$ replaced by $C^\beta(\bar{\Omega})$. The constraint on β can be relaxed to $(k-1, k-1) \leq \beta \leq (2k, 2k)$.

Theorems 2.2.4, 2.2.5, 2.3.2, and 2.3.3 have the following counterparts.

Theorem 2.4.1: *Let α and β be any multiintegers satisfying $\alpha \leq \beta$ and $(k-1, k-1) \leq \beta$. Then for all $u \in C^\beta(\bar{U})$,*

$$\|D^\alpha E u\|_{0,\infty} \leq C_\alpha \|u\|_{\beta,\infty}$$

where C_α is as in theorem 2.2.4.

Theorem 2.4.2: *Let α and β be multiintegers satisfying $\alpha \leq \beta$ and $(k-1, k-1) \leq \beta \leq (2k, 2k)$. Then for all $u \in C^\beta(\bar{U})$,*

$$\|D^\alpha E u\|_{0,\infty} \leq BC_\alpha \|D^\beta u\|_{0,\infty}$$

where B and C_α are as in theorems 1.5.6 and 2.2.4, respectively.

Theorem 2.4.3: Let α and β be multiintegers satisfying $\alpha \leq \beta$ and $(k-1, k-1) \leq \beta \leq (2k, 2k)$. Then for all $\tilde{u} \in C^\beta(\bar{R})$,

$$\|D^{\alpha} \tilde{E} \tilde{u}\|_{0,p} \leq BC_\alpha h_1^{\beta_1 - \alpha_1} h_2^{\beta_2 - \alpha_2} \|D^\beta \tilde{u}\|_{0,p}$$

where B and C_α are as in theorems 1.5.6 and 2.2.4, respectively.

Theorem 2.4.4: Let m be a nonnegative integer less than or equal to $2k$, and let β be a multiinteger satisfying $(k-1, k-1) \leq \beta \leq (2k, 2k)$ and $(m, m) \leq \beta$. Then for all $\tilde{u} \in C^\beta(\bar{R})$,

$$\|\tilde{E} \tilde{u}\|_{m,\infty} \leq BC_m h^{|\beta| - m} \|D^\beta \tilde{u}\|_{0,p}$$

where B and C_m are as in theorems 1.5.6 and 2.3.3, respectively.

CHAPTER THREE

(3.1) Elliptic Boundary Value Problems

This chapter, which contains no new material, is included in order to establish a general framework to be used in the chapters which follow. The most common application of the finite element method is the numerical solution of elliptic boundary value problems. Therefore, we begin with a description of problems of this type.

Let Ω be a bounded domain in \mathbb{R}^2 . For convenience we shall make some changes in notation. Points in \mathbb{R}^2 will be denoted (x,y) rather than (x_1,x_2) . For the derivatives of a function u we will use notation such as u_x for $\frac{\partial u}{\partial x}$ and u_{xy} for $\frac{\partial^2 u}{\partial x \partial y} = \frac{\partial^2 u}{\partial y \partial x}$.

The simplest example of an elliptic boundary value problem is the *Dirichlet problem* for Poisson's equation. Given a function f in some appropriate function space on Ω we seek a function u in some other function space such that

$$(3.1.1) \quad -\Delta u = f \quad \text{on } \Omega$$

$$(3.1.2) \quad u = 0 \quad \text{on } \partial\Omega$$

where Δ is the *Laplacian*: $\Delta u = u_{xx} + u_{yy}$. For our purposes it is desirable to transform this problem into a generalized form. If we multiply both sides of equation (3.1.1) by an arbitrary function $\psi \in C_0^\infty(\Omega)$ and integrate by parts on the left side, we obtain the equation

$$(3.1.3) \quad \int_{\Omega} (u_x \psi_x + u_y \psi_y) = \int_{\Omega} f \psi \quad \forall \psi \in C_0^\infty(\Omega).$$

Define a bilinear form $a(\cdot, \cdot)$ by

$$(3.1.4) \quad a(u, v) = \int_{\Omega} (u_x v_x + u_y v_y)$$

and recall that we have agreed to denote the L_2 inner product by $(\cdot, \cdot)_0$.

Also, note that, as $C_0^{\infty}(\Omega)$ is dense in $\overset{\circ}{W}^{(1)}(\Omega)$, (3.1.3) holds for all functions $v \in \overset{\circ}{W}^{(1)}(\Omega)$. Equation (3.1.3) can now be rewritten as

$$(3.1.5) \quad a(u, v) = (f, v)_0 \quad \forall v \in \overset{\circ}{W}^{(1)}(\Omega).$$

This is the generalized form of (3.1.1). The bilinear form $a(\cdot, \cdot)$ is well defined for u and v in $\overset{\circ}{W}^{(1)}(\Omega)$, whereas Δu is not well defined unless u is twice differentiable. Thus (3.1.5) admits a larger class of functions as potential solutions than does (3.1.1). We generalize the boundary condition (3.1.2) by requiring that the solution u lie in $\overset{\circ}{W}^{(1)}(\Omega)$ rather than the larger space $\overset{\circ}{W}^{(1)}(\Omega)$. The *generalized Dirichlet problem* is to find $u \in \overset{\circ}{W}^{(1)}(\Omega)$ such that (3.1.5) holds. The given function f is assumed to be in $L_2(\Omega)$.

The form $a(\cdot, \cdot)$ is obviously bounded on $\overset{\circ}{W}^{(1)}(\Omega)$. That is,

$$(3.1.6) \quad |a(u, v)| \leq \|u\|_1 \cdot \|v\|_1 \quad \forall u, v \in \overset{\circ}{W}^{(1)}(\Omega).$$

A less obvious fact is that $a(\cdot, \cdot)$ is *strongly elliptic* on $\overset{\circ}{W}^{(1)}(\Omega)$. By this we mean that there is a constant $\alpha > 0$ such that

$$(3.1.7) \quad a(v, v) \geq \alpha \|v\|_1^2 \quad \forall v \in \overset{\circ}{W}^{(1)}(\Omega).$$

This inequality is a consequence of *Friedrichs' inequality*, a statement and proof of which can be found in the article [23] by the author of this thesis.

See also [17]. Note that (3.1.7) holds only on $\overset{\circ}{W}^{(1)}(\Omega)$, not on $\overset{\circ}{W}^{(1)}(\Omega)$.

For example, the function $v(x, y) \equiv 1$ does not satisfy (3.1.7)

A consequence of (3.1.6), (3.1.7), and the symmetry of the form $a(\cdot, \cdot)$ is that $a(\cdot, \cdot)$ is an inner product on $\overset{\circ}{W}^{(1)}(\Omega)$ equivalent to the standard

inner product $(\cdot, \cdot)_1$. The functional $L(v) = (f, v)_0$ is a bounded linear functional on the Hilbert space $\overset{\circ}{W}^{(1)}(\Omega)$. Therefore, by a classical theorem of F. Riesz ([18], page 80), there is a unique function $u \in \overset{\circ}{W}^{(1)}(\Omega)$ such that $\alpha(u, v) = L(v) = (f, v)_0$ for all $v \in \overset{\circ}{W}^{(1)}(\Omega)$. Thus, the generalized Dirichlet problem has a unique solution.

We now propose a more general problem which retains the essential features of the generalized Dirichlet problem. Let m be a nonnegative integer, and let V be a closed subspace of $\overset{\circ}{W}^{(m)}(\Omega)$ containing $\overset{\circ}{W}^{(m)}(\Omega)$. Suppose that there exists a symmetric bilinear form $\alpha(\cdot, \cdot)$ which is bounded and strongly elliptic on V . Thus, there exist constants M and $\alpha > 0$ such that

$$(3.1.8) \quad |\alpha(u, v)| \leq M \|u\|_m \|v\|_m \quad \forall u, v \in V$$

$$(3.1.9) \quad \alpha(v, v) \geq \alpha \|v\|_m^2 \quad \forall v \in V.$$

Let L be a bounded linear functional on V . The general *elliptic boundary value problem* is to find $u \in V$ such that

$$(3.1.10) \quad \alpha(u, v) = L(v) \quad \forall v \in V.$$

By the previously cited theorem of F. Riesz, (3.1.10) has a unique solution in V .

In specific realizations of the elliptic boundary value problem the integer m is determined by the order of the equation to be solved. Given an equation of order $2m$ we work in the Sobolev space $\overset{\circ}{W}^{(m)}(\Omega)$. The space V is determined by the boundary conditions. If the boundary conditions for the problem are the Dirichlet boundary conditions $\frac{\partial^k u}{\partial n^k} = 0$ on $\partial\Omega$, $k = 0, \dots, m-1$, we take $V = \overset{\circ}{W}^{(m)}(\Omega)$. At the other extreme is the Neumann problem for which there are no essential boundary conditions. In this case

we take $V = W^{(m)}(\Omega)$.

The framework proposed here is far from the most general possible framework for elliptic boundary value problems. For instance, one might consider nonsymmetric forms $a(\cdot, \cdot)$. See Cea [7] and Friedman [10]. Also, we are considering here only homogeneous problems. Nonhomogeneous problems are considered in Lions and Magenes [16].

(3.2) Finite Element Spaces

The finite element method offers a means of numerically solving (3.1.10). The approximate solution u^* is taken from a finite dimensional space V^* of functions of a special type. To construct a space V^* of "trial functions" we first partition the domain Ω into triangular or quadrilateral *elements*. (In this thesis we will concentrate on rectangular elements.) A finite dimensional space S of bivariate functions, usually polynomials, is prescribed. A finite element space V^* satisfies three conditions: For each function $v^* \in V^*$ i) the restriction of v^* to each element coincides with a member of S , ii) v^* satisfies certain specified conditions at the interelement boundaries, and iii) v^* satisfies specified conditions at the boundary of Ω .

As an illustration we shall consider a specific example. Suppose Ω is a rectangle with sides parallel to the coordinate axes. We draw a grid of horizontal and vertical lines, partitioning Ω into small rectangular elements. The space S is defined to be the four-dimensional space of bilinear polynomials $a+bx+cy+dxy$, and V^* is defined to be the space of functions v^* such that i) the restriction of v^* to each element coincides with a bilinear polynomial, ii) v^* is continuous throughout Ω , and iii) v^* is

zero on the boundary of Ω . The meaning of condition ii) is that the bilinear functions which define v^* on two adjacent elements must be equal at the interelement boundary.

A bilinear polynomial is completely determined by its values at the four corners of a rectangle. Therefore a function $v^* \in V^*$ is uniquely determined by its values at the points of intersection of the lines of the grid. This is a statement of uniqueness. Simple arguments show that the related question of existence is also true. For any values which we might assign to the intersection points of the grid there exists a function v^* in V^* which takes on the prescribed value at each intersection point. This is not quite true. The value zero must be specified at those intersection points which lie on the boundary of Ω . Otherwise the boundary condition will not be satisfied.

The points of intersection of the grid lines will be referred to as *nodes*. A function $v^* \in V^*$ is uniquely determined by its nodal values. More specifically, v^* is completely determined on each element by its values at the nodes associated with that element.

Clearly the dimension of V^* is exactly the number of nodes in the interior of Ω . A useful basis for V^* is the set of functions which have the value one at one interior node and zero at all other nodes. A basis function of this type is identically zero outside of the patch of four elements surrounding the one node at which the function is not zero.

We now consider a second example. Again Ω is divided into small rectangular elements, but this time S is defined to be the space of all

polynomials of the form $p(x,y) + ax^3y + bxy^3$, where $p(x,y)$ is a cubic polynomial and a and b are real numbers. We define V^* to be the set of functions v^* such that i) the restriction of v^* to each element coincides with a polynomial in S , ii) v^* is continuous throughout Ω and C^1 at the intersection points of the grid, and iii) v^* is zero on the boundary of Ω .

The space S is twelve-dimensional, and it can be shown that a function $s \in S$ is uniquely determined by the values $s, s_x,$ and s_y at the four corners of a rectangle. (This will be proven in section 4.1). Thus a function $v^* \in V^*$ is completely determined by the values of $v^*, v_x^*,$ and v_y^* at the points of intersection of the lines of the grid. Again these points will be called nodes, but in this case they are *triple* nodes and will be viewed as three separate nodes in certain contexts. The values of $v^*, v_x^*,$ and v_y^* at the nodes will be called the *nodal values* of v^* . Every function of V^* is completely determined by its nodal values. Conversely it can be shown that, as in the previous example, for any prescribed nodal values compatible with the boundary conditions there exists a function in V^* which takes on those nodal values. Thus the dimension of V^* is just the number of nodes (counting triple nodes as three nodes) which are unaffected by the boundary conditions. An important basis for V^* is the set of functions which have one nodal value equal to one and all other nodal values zero. This basis is "local" in that each basis function is identically zero except on the elements associated with the node to which that function corresponds.

So far the term "element" has been used to refer to the small rectangular regions into which a domain has been partitioned. From this point on "element"

will also be used in an extended sense to mean a particular finite element scheme such as the two which have just been described. The element of the first example is known as the *bilinear element*. It is called a four *degree-of-freedom* element because the space S of bilinear polynomials is four-dimensional. The element of the second example, a twelve degree-of-freedom element, is known as *Adini's rectangle* [1].

We now return to the general discussion. We have postulated a space V^* which is made up of functions which are piecewise polynomials from some finite-dimensional space S , and which satisfy some sort of interelement continuity conditions and boundary conditions. The continuity condition need not be as strict as the requirement that the functions be continuous along all of each interelement boundary. For example, there are elements for which continuity is attained only at the midpoint of each interelement boundary. However, in the two examples considered above we did, in fact, have continuity along the interelement boundaries. As is well known ([20], page 327), this implies that $V^* \subseteq W^{(1)}(\Omega)$. Similarly, since the members of V^* (in the two examples) are zero on the boundary of Ω , the even stronger inclusion $V^* \subseteq \mathcal{H}^{(1)}(\Omega)$ holds. This inclusion indicates that the two elements which we have considered might be useful for the solution of the generalized Dirichlet problem (3.1.5), as $\mathcal{H}^{(1)}(\Omega)$ is the space in which this problem has been posed. More generally, if we wish to solve the elliptic boundary value problem (3.1.10), which has been posed in the space V , we might like to use an element which satisfies the inclusion $V^* \subseteq V$. Indeed, it might seem that this inclusion is essential to the success of the procedure. In fact it is not. Elements which fail to satisfy the inclusion $V^* \subseteq V$ are

called *nonconforming* elements. Nonconforming elements are widely used and often give good results. The theory of nonconforming elements is discussed in the chapter on "variational crimes" in Strang and Fix [21]. This thesis will consider only *conforming* elements, i.e. elements which satisfy the inclusion $V^* \subseteq V$. The bilinear element and Adini's rectangle are conforming elements for the generalized Dirichlet problem and second-order problems in general. However, these elements are nonconforming for fourth-order problems, which require the inclusion $V^* \subseteq V \subseteq W^{(2)}(\Omega)$.

Both of the examples which have been considered feature "nodes" such that any function $v^* \in V^*$ is completely determined on each element by its nodal values at the nodes associated with that element. Additionally, for any specified set of nodal values there is a function in V^* which takes on these nodal values. Finite element schemes having these properties are called *nodal* finite elements and are the subject of the *nodal finite element method*. Every nodal finite element space has a "local" basis consisting of functions which have one nodal value equal to one and all other nodal values zero. In this thesis only nodal finite element schemes will be considered.

A nodal finite element scheme is completely defined once the placement of the nodes on a typical element (say the unit square \bar{U}) has been specified and the polynomial space S has been defined. The space S must be compatible with the choice of nodes: For each specified set of nodal values there must be a unique function in S which takes on the given nodal values. An equivalent statement of the compatibility condition is that for each sufficiently

smooth function u on \bar{V} there exists a unique polynomial q in S such that q and u have the same nodal values. The map $u \rightarrow q$ defines a linear projection operator $Qu = q$. As the range of Q is S , Q and the placement of the nodes determine the finite element scheme. In chapter four a number of elements will be constructed. In each case the construction will consist of specifying the nodes and constructing the projector Q .

(3.3) The Finite Element Solution

Having determined the nature of the spaces V^* which we will be considering, we now address the problem of selecting an approximate solution $u^* \in V^*$ of the elliptic problem (3.1.10). A reasonable procedure would be to select that function u^* from V^* which is in some sense closest to the exact solution u . In the finite element method we take as our measure of closeness the *energy norm*, the norm induced by the inner product $\alpha(\cdot, \cdot)$. This norm is equivalent to the Sobolev norm $\|\cdot\|_m$ on the Hilbert space V . According to the elementary theory of Hilbert space there is a unique $u^* \in V^*$ which is closest to the exact solution u in the sense that

$$\alpha(u-u^*, u-u^*)^{\frac{1}{2}} = \inf_{v^* \in V^*} \alpha(u-v^*, u-v^*)^{\frac{1}{2}}$$

and u^* is characterized by the fact that the error function $u-u^*$ is orthogonal to V^* . That is

$$(3.3.1) \quad \alpha(u-u^*, v^*) = 0 \quad \forall v^* \in V^*.$$

Combining (3.3.1) with (3.1.10) we arrive at a characterization of u^* which does not involve the unknown function u :

$$(3.3.2) \quad \alpha(u^*, v^*) = L(v^*) \quad \forall v^* \in V^*.$$

Thus u^* is the Ritz-Galerkin approximation to u .

We now consider the problem of solving (3.3.2). It is assumed that V^* is a nodal finite element space. We number the nodes and define ψ_j to be the unique basis function of V^* whose j th nodal value is one and whose other nodal values are zero. Problem (3.3.2) is equivalent to

$$(3.3.3) \quad a(u^*, \psi_j) = L(\psi_j) \quad j=1, \dots, d$$

where d is the dimension of V^* . The function u^* is of the form $u^* = \sum_{i=1}^d \alpha_i \psi_i$, where $\alpha_1, \dots, \alpha_d$ are real coefficients to be determined. Thus (3.3.3) can be rewritten as a $d \times d$ matrix equation

$$(3.3.4) \quad \sum_{i=1}^d a(\psi_i, \psi_j) \alpha_i = L(\psi_j) \quad j=1, \dots, d.$$

The coefficient matrix $K = (a(\psi_i, \psi_j))$ is called the *stiffness matrix*. It is a Gram matrix based on linearly independent functions and is therefore symmetric and positive definite. These properties are obviously important to solving (3.3.4), but the real advantage of the finite element method lies in the "localness" of the basis functions. The support of ψ_j is restricted to the elements immediately surrounding the j th node. Therefore $a(\psi_i, \psi_j)$ will be zero unless one of the elements has both the i th and j th nodes associated with it. Also, those off-diagonal entries $a(\psi_i, \psi_j)$ which are not zero tend to be smaller than the main-diagonal entries. In other words, the functions ψ_1, \dots, ψ_d are "nearly orthogonal". In consequence the stiffness matrix K is sparse and well-conditioned, and (3.3.4) can be solved inexpensively and accurately. For a discussion of the condition of the stiffness matrix see Strang and Fix [21].

We can most easily take advantage of the sparseness of K by taking care in numbering the nodes. It is possible to number the nodes in such

a way that the nonzero elements of K are confined to a narrow band of diagonals on either side of the main diagonal. A matrix whose nonzero entries are restricted to such a band is called a *band matrix*. Its *band width* is the number of diagonals between (and including) the highest and lowest diagonals containing nonzero entries. By keeping the band width small we can significantly reduce the computer time and storage requirements for solving (3.3.4).

CHAPTER FOUR

THE USE OF BLENDING-FUNCTION METHODS IN THE CONSTRUCTION OF FINITE ELEMENT SCHEMES

(4.1) Adini's Rectangle

Blending-function methods, in the form in which they were presented in chapter two, are of no immediate use in numerical analysis because the blended interpolant of a function u depends on an infinity of data associated with u . In order to make blending-function methods useful numerically it is necessary to carry out a further discretization to obtain an interpolant which is determined by finitely many data. The end product of such a discretization is a finite element. Gordon and Hall [12] and Barnhill and Gregory [3] have produced finite elements in this manner. In this chapter we use blending-function methods to construct one well known element and several others which are evidently new. We begin by showing that Adini's rectangle, which was introduced in chapter three, can be arrived at by means of blending-function methods.

Let U denote, as before, the open unit square. Let u be a once continuously differentiable function on \bar{U} , and suppose we are given the values of u , u_x , and u_y at the four corners of \bar{U} . We shall construct a polynomial $q=Qu$ which interpolates these twelve nodal values and is completely determined by them. Define v along the bottom edge $\{(x,y) | 0 \leq x \leq 1, y=0\}$ of \bar{U} by $v(x,0)=s(x)$, where $s(x)$ is the unique cubic polynomial satisfying the interpolatory conditions

$$\begin{aligned} s(0) &= u(0,0) & s(1) &= u(1,0) \\ s'(0) &= u_x(0,0) & s'(1) &= u_x(1,0). \end{aligned}$$

Similarly, define v along the right edge of \bar{U} by $v(1,y)=r(y)$, where $r(y)$ is the unique cubic polynomial satisfying

$$\begin{aligned} r(0) &= u(1,0) & r(1) &= u(1,1) \\ r'(0) &= u_y(1,0) & r'(1) &= u_y(1,1). \end{aligned}$$

Define v along the other two edges of \bar{U} in an analogous manner. Obviously v interpolates the nodal values of u . Note that this process preserves bicubic polynomials (polynomials of the form $\sum_{i,j \leq 3} a_{ij} x^i y^j$) in the sense that if u is a bicubic polynomial then $v=u|_{\partial U}$.

The second step of the construction must now be obvious. Define the interpolant $q=Qu$ to be the blended interpolant Pv ,

$$(4.1.1) \quad q = Qu = Pv$$

where P is the blended interpolating operator based on linear blending functions. The important properties of this operator are summarized in corollary 2.1.3. One of these properties is that Pv is completely determined by the boundary values of v . Thus q is well defined by (4.1.1).

This stage of the construction preserves all monomials of the form $x^i y^j$, where either $i \leq 1$ or $j \leq 1$. That is, if $v=m|_{\partial U}$, where m is a monomial of the prescribed form, then $Pv=m$. P is linear, so all linear combinations of these monomials are preserved.

The interpolant q has the desired interpolatory properties: q is Pv , which interpolates (the boundary values of) v , which in turn interpolates the nodal values of u . Thus q interpolates the twelve nodal values of u . Likewise q is completely determined by these nodal values.

Both stages of the construction are linear, so the operator Q which

has been implicitly defined by the construction is linear. Q is also clearly a projector. Therefore the range of Q is exactly the set of functions which are preserved by Q . It is easily verified that the cubic polynomials and the monomials x^3y and xy^3 are preserved in both stages of the construction. Thus they are preserved by Q . The span of these polynomials is a twelve-dimensional space. The dimension of the range of Q cannot exceed twelve because Qu is determined by twelve parameters associated with u . Therefore the range of Q must be exactly the space spanned by x^3y , xy^3 , and the cubic polynomials.

We have specified a set of twelve nodes, the same nodes as for Adini's rectangle, and we have constructed an interpolating projector Q whose range S is the space of polynomials on which Adini's rectangle is based. Thus we have constructed Adini's rectangle. In the process we have fulfilled a promise made in chapter three. It has been shown that for any specified set of nodal values there is a unique $s \in S$ which interpolates these values. Existence is guaranteed by the construction. Uniqueness follows from existence because the dimension of S is the same as the number of nodes.

There is one other point worth discussing. The original definition of Adini's rectangle given in chapter three does not involve the concept of nodes. The trial space V^* was defined to be the set of functions which are elementwise members of the polynomial space S , which are continuous from one element to the next, and which satisfy appropriate boundary conditions. The alternate characterization in terms of nodes was then stated without proof. It was stated that for any given set of nodal values con-

sistent with the boundary conditions, there is a unique $v^* \in \mathcal{V}^*$ which interpolates these values. This is equivalent to saying that for each $u \in C^1(\bar{\Omega})$ which satisfies the boundary conditions there is a unique $v^* \in \mathcal{V}^*$ which has the same nodal values as u . The uniqueness is obvious: the restriction of v^* to an element e has to be the unique $s \in \mathcal{S}$ which interpolates the nodal values of $u|_e$. Thus $v^*|_e = s = Q_e(u|_e)$, where Q_e is just the interpolating projector Q scaled to the element e . The question of existence will have been answered in the affirmative once it has been shown that the unique function v^* defined by $v^*|_e = Q_e(u|_e)$ (for all elements e) is continuous at the interelement boundaries and satisfies the boundary conditions.

We shall demonstrate interelement continuity. Let e_1 and e_2 be two elements having a common edge. The interpolant $v^*|_{e_1} = Q_{e_1}(u|_{e_1})$ was defined along the common edge to be the unique cubic polynomial determined by four nodal values. The interpolant $v^*|_{e_2} = Q_{e_2}(u|_{e_2})$ was defined along the same edge to be the unique cubic polynomial determined by the same four nodal values. Thus $v^*|_{e_1}$ and $v^*|_{e_2}$ are equal at the element boundary, and v^* is continuous throughout $\bar{\Omega}$. The proof that v^* satisfies the boundary conditions is similar.

(4.2) C^1 Elements

In chapter three we considered a second-order problem as a model problem. For second-order problems it is convenient to work in the space $W^{(1)}(\Omega)$. Any finite element scheme which features interelement continuity is conforming in the sense that the trial space V^* lies within $W^{(1)}(\Omega)$. Such elements are called C^0 elements.

For fourth-order elliptic problems the conformity condition is $V^* \subseteq V \subseteq W^{(2)}(\Omega)$. The condition $V^* \subseteq W^{(2)}(\Omega)$ is satisfied if and only if every $v^* \in V^*$ is once continuously differentiable globally. Elements satisfying this condition are called C^1 elements. In this section we use blending-function methods to construct three C^1 elements. We could take a much more general approach and define large classes of C^k elements at one stroke. It is this author's opinion that a few specific constructions are worth as much.

One other remark is in order here. We have introduced C^1 elements on the grounds that they are useful for solving fourth-order problems. Obviously C^1 elements can be used to solve second-order problems as well.

The construction of C^1 elements can be carried out as follows. For each $u \in C^2(\bar{U})$ we define functions v and v_n (= normal derivative of v) on the boundary of U such that v and v_n are determined by a finite number of nodal values associated with u , and v and v_n interpolate these nodal values. We then define the interpolant $q=Qu$ by $q=Pv$, where P is the blended interpolating operator based on Hermite cubic blending functions. The important properties of P are summarized in corollary 2.1.4. One of the properties is that Pv is determined by the boundary values of v and its normal derivative, so the definition $q=Pv$ makes sense even though only the boundary values of v and v_n have been defined. Also Pv interpolates v and v_n , so q interpolates the nodal values of u .

The first step in defining a specific nodal finite element is the determination of the nodes. As a prelude to this first step it is worth-

while to examine one of this author's unsuccessful attempts at element construction and thereby discover a pitfall which must be avoided. In an attempt to construct a C^1 element with as few nodes as possible, I chose the same twelve nodes as for Adini's rectangle -- u , u_x , and u_y at the four corners of \bar{U} . The function v was defined on the boundary of U in the same manner as for Adini's rectangle. The normal derivative v_n was defined on (say) the bottom edge of \bar{U} by $v_n(x,0) = v_y(x,0) = r(x)$, where $r(x)$ is the unique linear polynomial such that

$$r(0) = u_y(0,0) \qquad r(1) = u_y(1,0).$$

Similarly v_n was defined on the left edge of \bar{U} by $v_n(0,y) = v_x(0,y) = s(y)$, where $s(y)$ is the unique linear polynomial such that

$$s(0) = u_x(0,0) \qquad s(1) = u_x(0,1).$$

The interpolant $q=Qu$ was defined to be the blended interpolant of v based on Hermite cubic blending functions.

On attempting to determine the elementary properties of this element I immediately encountered what appeared to be contradictions. After some thought I discovered the root of the problem. On one hand $v_y(x,0) = r(x)$, so $v_{yx}(0,0) = r'(0)$. On the other hand $v_x(0,y) = s(y)$, so $v_{xy}(0,0) = s'(0)$. In general $r'(0) \neq s'(0)$, so $v_{xy}(0,0) \neq v_{yx}(0,0)$. This inequality destroys the commutativity of P_1 and P_2 (see (2.1.2), (2.1.3), (2.1.4), with $k=2$), and consequently the interpolatory properties of P are lost.

The obvious way to avoid this pitfall is to include the corner values of $u_{xy} = u_{yx}$ as nodes and define v_n in such a way that $v_{xy} = v_{yx} = u_{xy}$ at the corners of \bar{U} . Other problems may occur unless the corner values of u , u_x ,

and u_y are included as nodal values. For example, suppose we define $v(x,0) = r(x)$ and $v_n(0,y) = v_x(0,y) = s(y)$, where r and s are certain specified polynomials. Then $r'(0) = v_x(0,0)$ and $s(0) = v_x(0,0)$. To have $r'(0) \neq s(0)$ would be fatal. We avoid this possibility by including $u_x(0,0)$ as a nodal value and defining $r(x)$ and $s(y)$ in such a way that $r'(0) = u_x(0,0) = s(0)$.

We are now committed to having at least sixteen nodes, namely u , u_x , u_y and u_{xy} at the four corners of \bar{U} . It is possible to construct an element having just these sixteen nodes using the method proposed here. This element is just the product two-point Hermite element, which can be obtained by a simpler construction. We shall not discuss this element in detail.

Having determined the minimum number of nodes which we are willing to tolerate, we might now ask what is the maximum number of nodes from which we can benefit. In order to obtain a high rate of convergence an element should be exact for all polynomials of as high a degree as possible. By corollary 2.1.4, the blending-function stage of our construction preserves all polynomials of degree seven or less but not all polynomials of degree eight. The monomial x^4y^4 is not preserved by P . Therefore we might profit by furnishing enough nodes to make the first stage of the construction exact for seventh degree polynomials. A C^1 element which preserves seventh degree polynomials can be constructed by the method proposed here with 44 nodes, a number which seems a bit high.

We shall describe three elements which lie between the two extremes. We begin with an element which is exact for quintic polynomials. Suppose

we take as nodes the values u_{xx} and u_{yy} at the corners, as well as u , u_x , u_y , and u_{xy} . Define v along the bottom edge of \bar{U} by $v(x,0) = r(x)$, where $r(x)$ is the unique quintic polynomial such that

$$\begin{aligned} r(0) &= u(0,0) & r(1) &= u(1,0) \\ r'(0) &= u_x(0,0) & r'(1) &= u_x(1,0) \\ r''(0) &= u_{xx}(0,0) & r''(1) &= u_{xx}(1,0). \end{aligned}$$

Define v in an analogous manner on the other three sides.

How should v_n be defined? If the element is to preserve quintic polynomials, v_n must be a quartic polynomial along each edge. In order to define v_n as a quartic polynomial we introduce four new nodes -- the normal derivatives u_n at the midpoints of the four sides of \bar{U} . These are $u_y(\frac{1}{2},0)$, $u_x(1,\frac{1}{2})$, $u_y(\frac{1}{2},1)$, and $u_x(0,\frac{1}{2})$. Define v_n on the bottom edge of \bar{U} by $v_n(x,0) = v_y(x,0) = s(x)$, where $s(x)$ is the unique quartic polynomial such that

$$\begin{aligned} s(0) &= u_y(0,0) & s(\frac{1}{2}) &= u_y(\frac{1}{2},0) & s(1) &= u_y(1,0) \\ s'(0) &= u_{yx}(0,0) & s'(1) &= u_{yx}(1,0). \end{aligned}$$

Define v_n analogously on the other sides of \bar{U} . The definitions of v and v_n are consistent in that the pitfalls mentioned above have been avoided. We can now complete the construction by defining $q=Qu=Pv$, where P is the hermite cubic blended interpolating operator.

The 28 degree-of-freedom (d.o.f.) element which we have just constructed is clearly a C^1 element. For suppose e_1 and e_2 are two elements having a common side Γ . Let v^* be any function in the trial space V^* , and let $s_1=v^*|_{e_1}$ and $s_2=v^*|_{e_2}$. Then s_1 and s_2 are defined on Γ to be quintic polynomials which are both determined by the same interpolation scheme, and

which interpolate the same nodal values. Thus $s_1 = s_2$ on Γ . By similar reasoning, the normal derivatives of s_1 and s_2 on Γ are equal. Thus v^* is once continuously differentiable at the element boundary.

In order to really "know" an element it is necessary to determine the space S of polynomials preserved by the interpolating projector Q . Therefore we shall now determine S . The element has been constructed in such a way that S contains all quintic polynomials. The dimension of the space of quintic polynomials is 21, whereas the dimension of S is 28. Therefore there are seven other linearly independent polynomials in S . The first stage of the construction preserves any biquintic polynomial whose normal derivative is quartic along each side of the boundary of U . In particular, the first stage preserves all biquartic polynomials. The second stage of the construction preserves (according to corollary 2.1.4) every monomial of the form $x^i y^j$, where either $i \leq 3$ or $j \leq 3$. Therefore the monomials $x^4 y^2$, $x^3 y^3$, $x^2 y^4$, $x^4 y^3$, and $x^3 y^4$ are monomials of degree greater than five which lie in S . This leaves only two more linearly independent polynomials in S to be determined. The two remaining polynomials turn out to be $x^5(3y^2 - 2y^3)$ and $(3x^2 - 2x^3)y^5$. We shall verify that $\psi(x, y) = x^5(3y^2 - 2y^3)$ is in S . Since ψ is cubic in y , it is preserved in the blending-function (second) stage of the construction. As for the first stage, the function values of ψ are preserved on the boundary of U because ψ is biquintic. For the normal derivatives we have on the vertical sides $\psi_n(x, y) = \psi_x(x, y) = 5x^4(3y^2 - 2y^3)$, which is (less than) quartic in y and is therefore preserved. On the horizontal sides we have $\psi_n(x, y) = \psi_y(x, y) = 6x^5(y - y^2)$, which is not quartic in x . However, for $y=0$ or $y=1$ we have $\psi_n(x, y) = 0$, which is certainly quartic.

This proves that $\psi \in S$. By symmetry the polynomial $(3x^2 - 2x^3)y^5$ is also in S .

The next element which we shall construct is a 24 d.o.f. C^1 element which is exact for all quartic but not all quintic polynomials. For our 24 nodal values we take u , u_x , u_y , u_{xx} , u_{xy} , and u_{yy} at the four corners. We define v on the boundary of U exactly as for the previous element. That is, we define v on (for instance) the right edge of \bar{U} by $v(1, y) = r(y)$, where $r(y)$ is the unique quintic polynomial such that

$$\begin{aligned} r(0) &= u(1,0) & r(1) &= u(1,1) \\ r'(0) &= u_y(1,0) & r'(1) &= u_y(1,1) \\ r''(0) &= u_{yy}(1,0) & r''(1) &= u_{yy}(1,1). \end{aligned}$$

The definition of v_n for this element differs from that of the previous element in that here we dispense with the midside nodes and require that v_n be cubic. Specifically, we define v_n along (say) the bottom edge of \bar{U} by $v_n(x, 0) = v_y(x, 0) = s(x)$, where $s(x)$ is the unique cubic polynomial such that

$$\begin{aligned} s(0) &= u_y(0,0) & s(1) &= u_y(1,0) \\ s'(0) &= u_{yx}(0,0) & s'(1) &= u_{yx}(1,0). \end{aligned}$$

This completes the first stage of the construction. The second stage is, of course, the same as in the construction of the 28 d.o.f. element.

The 24 d.o.f. element clearly preserves all quartic polynomials, for quartic polynomials have cubic normal derivatives. There are fifteen linearly independent quartic polynomials. The other nine linearly independent polynomials which are preserved by this element are i) the bicubics x^3y^2 , x^2y^3 , and x^3y^3 , ii) x^5 and y^5 , and iii) $x^4(3y^2 - 2y^3)$, $x^5(3y^2 - 2y^3)$, $(3x^2 - 2x^3)y^4$,

and $(3x^2 - 2x^3)y^5$.

The saving of four nodes in this element as compared to the previous element has a greater effect than one might at first suspect. Suppose we have a large grid of elements. Then the total number of edges is almost twice the number of vertices, and the ratio approaches two as the grid is made finer. Therefore the total saving in nodes realized by discarding the midside nodes is almost as great as would be gotten by eliminating two nodes at each vertex.

The third C^1 element which we shall construct is a 20 d.o.f. element which is exact for quartic polynomials. The nodal values for this element are u , u_x , u_y , and u_{xy} at the vertices, and the function value u at the midpoints of the sides. Define v along (for example) the left edge of \bar{U} by $v(0,y) = r(y)$, where $r(y)$ is the unique quartic polynomial such that

$$\begin{aligned} r(0) &= u(0,0) & r(\frac{1}{2}) &= u(0,\frac{1}{2}) & r(1) &= u(0,1) \\ r'(0) &= u_x(0,0) & & & r'(1) &= u_x(0,1). \end{aligned}$$

Define v_n to be a cubic polynomial exactly as for the 24 d.o.f. element. The second stage of the construction is exactly as it was for the other two C^1 elements.

The space S of polynomials preserved by this element is the space spanned by the quartic polynomials and x^3y^2 , x^2y^3 , x^3y^3 , $x^4(3y^2 - 2y^3)$, and $(3x^2 - 2x^3)y^4$.

The saving in total degrees of freedom realized by using this element as compared to the 24 d.o.f. element is small. The effect of eliminating

two nodes at each vertex in the 20 d.o.f. element is almost offset by the addition of one midside node.

We conclude this section by mentioning the possibility of mixing rectangular and triangular elements in a given problem. In many problems, such as problems in which the boundary is a polygon with angles other than right angles, it might be useful to use rectangular elements in the interior of the domain and triangular elements near the boundary. Both the 28 d.o.f. element and the 24 d.o.f. element are good candidates for such mixing. There is a well-known 21 d.o.f. C^1 element [25] having the nodal parameters u , u_x , u_y , u_{xx} , u_{xy} , and u_{yy} at the vertices of each triangle and u_n at the midpoint of each side. This nodal configuration is the same as that of the 28 d.o.f. rectangular element, so the two elements could share a common side. The 21 d.o.f. element is exact for quintic polynomials, as is the 28 d.o.f. element. The two elements are, in a word, *compatible*.

The 24 d.o.f. element also has a triangular counterpart -- the 18 d.o.f. element gotten from the 21 d.o.f. triangle by discarding those trial functions whose normal derivatives on the element boundaries are not cubic. This measure eliminates the need for the midside nodes and gives rise to a C^1 element which is exact for quartic polynomials and is compatible with the 24 d.o.f. rectangular element.

(4.3) An Element for Three-Dimensional Problems

To demonstrate the possibility of using blending-function methods in the construction of elements for three-dimensional problems we shall construct a three-dimensional "brick-shaped" C^0 element which preserves cubic polynomials.

We take as a standard "brick" the unit cube $U=(0,1)^3$. For nodal parameters we take $u, u_x, u_y,$ and u_z at the eight vertices of \bar{U} . Thus the element will have 32 degrees of freedom.

Given $u \in C^1(\bar{U})$ we are to construct an interpolant $q = Qu$ depending only on the nodal values of u . The construction will be carried out in three stages. The first stage consists of defining a function v on the edges of \bar{U} which interpolates the nodal parameters of u . Define v on (for example) the edge $\{(x,y,z) \mid 0 \leq x \leq 1, y=z=0\}$ by $v(x,0,0) = p(x)$, where $p(x)$ is the unique cubic polynomial such that

$$\begin{aligned} p(0) &= u(0,0,0) & p(1) &= u(1,0,0) \\ p'(0) &= u_x(0,0,0) & p'(1) &= u_x(1,0,0). \end{aligned}$$

Define v along the other eleven edges in an analogous manner. This completes the first stage of the construction. This stage is exact for tricubic polynomials, i.e. linear combinations of monomials of the form $x^i y^j z^k$, where $i, j,$ and k are all less than or equal to three.

The second phase of the construction consists of defining a function w on the faces of \bar{U} which interpolates v along the edges. Define w on each face to be the blended interpolant of v based on linear blending functions. Recall (corollary 2.1.3) that this operation preserves those monomials which are linear (or constant) in at least one of the variables. Thus the monomial $x^i y^j z^k$ is preserved on the faces $z=0$ and $z=1$ if $i \leq 1$ or $j \leq 1$. Similarly, $x^i y^j z^k$ is preserved on the faces $y=0$ and $y=1$ if $i \leq 1$ or $k \leq 1$, and on the faces $x=0$ and $x=1$ if $j \leq 1$ or $k \leq 1$. It follows that the monomial $x^i y^j z^k$ is preserved on all faces by the second stage of the construction if at least two of $i, j,$ and k are less than or equal to one.

Note that the first two stages could have been described as one: we define w on (for example) the face $z=0$ by considering the nodal values u , u_x , and u_y at the four vertices associated with this face and taking w to be the Adini's rectangle interpolant of these nodal values.

The third and final phase of the construction is the definition of the interpolant $q=Qu$ throughout \bar{U} . Define q to be the three-dimensional blended interpolant of w based on linear blending functions. That is, $q=Pw$, where P is defined by

$$P = P_1 + P_2 + P_3 - P_1P_2 - P_1P_3 - P_2P_3 + P_1P_2P_3$$

or, equivalently,

$$(I-P) = (I-P_1)(I-P_2)(I-P_3)$$

and P_1 , P_2 , and P_3 are given by

$$P_1w(x,y,z) = w(0,y,z)(1-x) + w(1,y,z)x$$

$$P_2w(x,y,z) = w(x,0,z)(1-y) + w(x,1,z)y$$

$$P_3w(x,y,z) = w(x,y,0)(1-z) + w(x,y,1)z.$$

The theory of three-dimensional blending-function methods is no different from the two-dimensional theory. The blended interpolant Pw is completely determined by the values of w on the faces (i.e. on the boundary) of \bar{U} , and Pw interpolates (the boundary values of) w . This phase of the construction preserves those monomials $x^i y^j z^k$ for which at least one of i , j , and k is less than or equal to one. All of these monomials were preserved in the second stage as well.

To determine the space S of polynomials which are preserved by this element we must determine which polynomials are preserved by all three

stages of the construction. The first stage preserves tricubic polynomials, and the second and third stages preserve all monomials which are of first degree in two of their three variables. Thus all monomials $x^i y^j z^k$ for which $i, j, k \leq 3$ and two of $i, j,$ and k are less than or equal to one are preserved in all three stages of the construction. A quick count shows that there are 32 such monomials, exactly the dimension of S . Therefore S is the space generated by these 32 monomials. Contained in S are all of the cubic polynomials but not, for example, the quartic monomials x^4 and $y^2 z^2$.

CHAPTER FIVE

ERROR BOUNDS FOR FINITE ELEMENT METHODS

(5.1) Introduction

In this chapter error bounds for finite element methods are obtained by two different approaches. The first approach is essentially that of Bramble and Zlamal [6]. Here we obtain some simplification by applying the Bramble-Hilbert lemma to linear operators with range in $L_2(U)$ rather than to linear functionals. Also we utilize fractional (i.e. noninteger) order Sobolev spaces to obtain a more general result. Aside from being aesthetically pleasing, the increased generality has practical importance, as will be shown by means of an example. This approach can be applied to arbitrary nodal finite elements, not just those which can be constructed using blending-function methods. Our treatment of the subject is correspondingly general.

The second approach uses the one-variable versions of the Sobolev and Bramble-Hilbert lemmas together with the error bounds already obtained for blending-function methods to derive finite element error bounds. This approach applies only to elements which can be constructed by the methods of chapter four. Its advantage over the first approach is that in this case it is possible to estimate the constants appearing in the error bounds. The second approach is more involved than the first, and rather than striving for great generality, we present complete proofs only for Adini's rectangle. We also indicate the procedure to be followed to obtain error bounds for the 24 d.o.f. element constructed in chapter four. This procedure can be applied

to the other C^1 elements as well.

The two approaches have much in common, and we shall cover their common points before examining them separately in sections 5.2 and 5.3, respectively.

In this chapter we return to the notation of chapters one and two in that points in a plane will be denoted $x = (x_1, x_2)$ and $w = (w_1, w_2)$, rather than (x, y) .

Let Ω be a bounded domain in the w -plane whose boundary is a polygon with sides parallel to the coordinate axes. Let m be a positive integer (which will remain fixed throughout this chapter) and let V be a complete subspace of $W^{(m)}(\Omega)$ which contains $\mathcal{H}^{(m)}(\Omega)$. Let $a(\cdot, \cdot)$ be a symmetric bilinear form which is bounded and strongly elliptic on V . Thus there exist positive constants M and α such that

$$(5.1.1) \quad |a(u, v)| \leq M \|u\|_m \|v\|_m \quad \forall u, v \in V$$

$$(5.1.2) \quad a(v, v) \geq \alpha \|v\|_m^2 \quad \forall v \in V.$$

Given a bounded linear functional L on V we wish to solve numerically the elliptic problem

$$(5.1.3) \quad a(u, v) = L(v) \quad \forall v \in V.$$

In section 3.1 it was shown that (5.1.3) has a unique solution $u \in V$.

There are infinitely many ways of partitioning $\bar{\Omega}$ into rectangular elements by drawing grids of horizontal and vertical lines. To each partition Π we assign two numbers, $h=h(\Pi)$ and $g=g(\Pi)$. The norm, h , of the partition is the maximum of the lengths of all sides of all elements in the partition, and g is the corresponding minimum. We pick a number $b \geq 1$ and

consider only those partitions for which $h/g \leq b$. In other words, we consider only those partitions for which the variation in size and shape of the elements is not too great. We shall also require that h not exceed one. We are primarily interested in what happens as h tends to zero.

We select a nodal finite element scheme. For each partition Π the finite element scheme gives rise to a finite element subspace $V^* \subseteq V$. There is a unique $u^* \in V^*$ such that

$$(5.1.4) \quad a(u^*, v^*) = L(v^*) \quad \forall v^* \in V^*.$$

We shall obtain asymptotic bounds for $\|u - u^*\|_m$ as the mesh norm h tends to zero.

In section 3.3 it was seen that u^* is the unique element of V^* which is closest to u with respect to the energy norm $a(\cdot, \cdot)^{1/2}$. Therefore

$$a(u - u^*, u - u^*) \leq a(u - Q'u, u - Q'u)$$

where $Q'u$ is the unique function in V^* which interpolates the nodal values of u . From this inequality and inequalities (5.1.1) and (5.1.2) we conclude that

$$(5.1.5) \quad \|u - u^*\|_m^2 \leq \frac{M}{\alpha} \|u - Q'u\|_m^2.$$

Thus, in order to estimate $\|u - u^*\|_m$ it suffices to estimate $\|u - Q'u\|_m$.

The estimate will be obtained in an element-by-element manner. That is, for each element R we will get a bound for $\|u - Q'u\|_{m,R}$. We will then sum up these bounds to obtain an estimate for $\|u - Q'u\|_{m,\Omega}$. To obtain a bound for $\|u - Q'u\|_{m,R}$ we shall set up an affine map between R and the unit square U , as in section 2.3. We shall then obtain a bound for the error between a function v on \bar{U} and its finite element interpolant Qv , and use

this bound together with the affine map to get a bound for $u-Q'u$ on R . The two approaches which we shall consider differ only in the way in which bounds for $v-Qv$ on U are obtained.

Suppose the dimensions of R are $h_1 \times h_2$, and the lower left corner of R is at the point (a_1, a_2) . Then the affine transformation

$$(5.1.6) \quad w_i = h_i x_i + a_i \quad i=1,2$$

maps \bar{U} in the x -plane onto \bar{R} in the w -plane. As in section 2.3 there is an obvious one-to-one correspondence between functions v defined on \bar{U} and functions \tilde{v} defined on \bar{R} . We associate the function \tilde{v} with v , where $\tilde{v}(w_1, w_2) = v(x_1, x_2)$. Lemma 2.3.1 is valid. For convenience we restate it here for the case $p=2$.

Lemma 5.1.1: *Let J be the Jacobian of the transformation (5.1.6), and let α be any multiinteger. Then for all $v \in W^\alpha(U)$*

$$\|D^\alpha v\|_{0,U} = h_1^{\alpha_1} h_2^{\alpha_2} J^{-1/2} \|D^\alpha \tilde{v}\|_{0,R}$$

Here $D^\alpha v$ means $\frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}$ and $D^\alpha \tilde{v}$ means $\frac{\partial^{|\alpha|} \tilde{v}}{\partial w_1^{\alpha_1} \partial w_2^{\alpha_2}}$.

(5.2) Error Bounds for Fractional Sobolev Spaces

We begin the first approach to error bounds by proving another lemma about the transformation from U to R . Let s be any positive real number, and recall the definition of the seminorm $|\cdot|_s = |\cdot|_{s,2}$ on $W^{(s)}(\Omega)$, where Ω is any bounded domain. If s is an integer the seminorm is defined by (1.4.1). Otherwise it is defined by (1.2.3) together with (1.2.1). Recall that $h_1 \leq h$ and $h_2 \leq h$, where $h_1 \times h_2$ are the dimensions of R , and h is the norm of the partition under consideration. Recall also the definitions of g and b . We have $\frac{h}{g} \leq b$.

Lemma 5.2.1: Let J be the Jacobian of the transformation (5.1.6), and let s be any positive number. Then for all $v \in W^{(s)}(U)$,

$$(5.2.1) \quad |v|_{s,U} \leq Ch^s J^{-1/2} |\tilde{v}|_{s,R}$$

where $C=1$ if s is an integer, and $C=b^{1/2}$ otherwise.

Proof: First suppose s is an integer. By lemma 5.1.1 we have, for all α such that $|\alpha|=s$,

$$\|D^\alpha v\|_{0,U}^2 = h_1^{2\alpha_1} h_2^{2\alpha_2} J^{-1} \|D^\alpha \tilde{v}\|_{0,R}^2 \leq h^{2s} J^{-1} \|D^\alpha \tilde{v}\|_{0,R}^2.$$

Thus

$$|v|_{s,U}^2 = \sum_{|\alpha|=s} \|D^\alpha v\|_{0,U}^2 \leq h^{2s} J^{-1} \sum_{|\alpha|=s} \|D^\alpha \tilde{v}\|_{0,R}^2 = h^{2s} J^{-1} |\tilde{v}|_{s,R}^2.$$

This proves (5.2.1) if s is an integer.

Now suppose s is not an integer, and let $s=s'+\sigma$, where s' is an integer and $0<\sigma<1$. Then by (1.2.3)

$$|v|_{s,U}^2 = \sum_{|\alpha|=s'} |D^\alpha v|_{\sigma,U}^2$$

and $|\tilde{v}|_{s,R}^2$ is given by an analogous equation. Thus it suffices to prove that

$$(5.2.2) \quad |D^\alpha v|_{\sigma,U}^2 \leq bh^{2s} J^{-1} |D^\alpha \tilde{v}|_{\sigma,R}^2$$

for all α such that $|\alpha|=s'$. Let $x=(x_1, x_2)$ and $y=(y_1, y_2)$ be two points in U , and let $w=(w_1, w_2)$ and $z=(z_1, z_2)$ be their respective images in R under the transformation (5.1.6). Recalling definition (1.2.1),

$$(5.2.3) \quad |D^\alpha v|_{\sigma,U}^2 = \int_U \int_U \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{\|x-y\|^{2+2\sigma}} dx dy.$$

We shall estimate separately each of the factors of the integrand in (5.2.3).

First, by the chain rule for differentiation,

$$(5.2.4) \quad \begin{aligned} |D^\alpha v(x) - D^\alpha v(y)|^2 &= h_1^{2\alpha_1} h_2^{2\alpha_2} |D^\alpha \tilde{v}(w) - D^\alpha \tilde{v}(z)|^2 \\ &\leq h^{2s} |D^\alpha \tilde{v}(w) - D^\alpha \tilde{v}(z)|^2. \end{aligned}$$

To estimate the term $\|x-y\|^{2+2\sigma}$ we shall assume, without loss of generality, that $h_1 \leq h_2$. Then $\|x-y\|^2 = (x_1-y_1)^2 + (x_2-y_2)^2 = h_1^{-2}(w_1-z_1)^2 + h_2^{-2}(w_2-z_2)^2 \geq h_2^{-2}\|w-z\|^2$. Thus

$$(5.2.5) \quad \|x-y\|^{2+2\sigma} \geq h_2^{-2+2\sigma} \|w-z\|^{2+2\sigma}$$

Finally

$$(5.2.6) \quad dx dy = J^{-2} dw dz.$$

Applying (5.2.4), (5.2.5), and (5.2.6) to (5.2.3) we get

$$(5.2.7) \quad |D^\alpha v|_{\sigma,U}^2 \leq h^{2s} h_2^{2+2\sigma} J^{-2} |D^\alpha \tilde{v}|_{\sigma,R}^2.$$

It is clear that $J=h_1 h_2$. Therefore $h_2^2 J^{-1} = \frac{h_2}{h_1} \leq \frac{h}{g} \leq b$, so by (5.2.7)

$$|D^\alpha v|_{\sigma,U}^2 \leq b h^{2s} h_2^{2\sigma} J^{-1} |D^\alpha \tilde{v}|_{\sigma,R}^2 \leq b h^{2s} J^{-1} |D^\alpha \tilde{v}|_{\sigma,R}^2.$$

This is just (5.2.2). ||

We shall make use of this lemma as soon as we have obtained bounds for $v-Qv$ on U , where $q=Qv$ is, as in previous sections, the finite element interpolant of v . Let k denote the number of degrees of freedom of the element under consideration. Then there exist points x^1, \dots, x^k (generally not distinct) in \bar{U} and formal differential operators $D^{\alpha^1}, \dots, D^{\alpha^k}$ such that

$$(5.2.8) \quad D^{\alpha^i} q(x^i) = D^{\alpha^i} v(x^i) \quad i=1, \dots, k.$$

As before, S will denote the range of Q , the space of polynomials preserved by the element. q is the unique member of S satisfying the interpolatory conditions (5.2.8). Let s_1, \dots, s_k be the unique members of S such that

$$D^{\alpha^i} s_j(x^i) = \delta_{ij} \quad i, j=1, \dots, k.$$

The functions s_1, \dots, s_k form a *canonical* basis for S . By the uniqueness of interpolation it is easy to verify that

$$q(x) = \sum_{i=1}^k D^{\alpha^i} q(x^i) s_i(x).$$

In view of (5.2.8) we can rewrite this expression as

$$(5.2.9) \quad Qv(x) = \sum_{i=1}^k D^{\alpha^i} v(x^i) s_i(x).$$

We shall use this representation of Q to obtain error bounds for $v - Qv$.

Theorem 5.2.2: Let $d = \max_{1 \leq i \leq k} |\alpha^i|$, where $\alpha^1, \dots, \alpha^k$ are as in (5.2.8) and (5.2.9). Let α be a multiinteger, and let s be a positive real number satisfying $|\alpha| \leq s$ and $d+1 < s$. Then there exists a constant C such that for all $v \in W^{(s)}(U)$,

$$(5.2.10) \quad \|D^\alpha(v - Qv)\|_{0,U} \leq C \|v\|_{s,U}$$

Proof: First we show that Qv is well defined if $v \in W^{(s)}(U)$. For this it is sufficient that the derivatives $D^{\alpha^i} v(x^i)$ appearing in (5.2.8) be well defined. This will certainly be the case if $v \in C^d(\bar{U})$, where $d = \max |\alpha^i|$. But by the corollary of the Sobolev lemma (corollary 1.3.2) with $p=n=2$, $W^{(s)}(U) \subseteq C^d(\bar{U})$. Thus Qv is well defined for all $v \in W^{(s)}(U)$.

The hypothesis $|\alpha| \leq s$ implies immediately that $\|D^\alpha v\|_{0,U} \leq \|v\|_{s,U}$. Thus the theorem will be proven if we can show that there is a constant C' such that $\|D^\alpha Qv\|_{0,U} \leq C' \|v\|_{s,U}$. From (5.2.9) we have

$$D^\alpha Qv(x) = \sum_{i=1}^k D^{\alpha^i} v(x^i) D^\alpha s_i(x).$$

Therefore

$$\|D^\alpha Qv\|_{0,U} \leq \left[\sum_{i=1}^k \|D^\alpha s_i\|_{0,U} \right] \max_{1 \leq i \leq k} |D^{\alpha^i} v(x^i)|.$$

The sum on the right hand side of this inequality is independent of v . We denote it C_1 . By corollary 1.3.2 (Sobolev lemma) with $p=n=2$, the max term is bounded by $C_2 \|v\|_{s,U}$, where C_2 is independent of v . Thus $\|D^\alpha Qv\|_{0,U} \leq C_1 C_2 \|v\|_{s,U}$. This proves the theorem. ||

The *degree* of a finite element scheme is the largest integer j such that the polynomial space S associated with the scheme contains all polynomials of degree j or less. Thus an element is of degree j if and only if it is exact for all polynomials of degree j or less but not all polynomials of degree $j+1$.

Theorem 5.2.3: *Let α , d , and s be as in the previous theorem. Thus $d = \max |\alpha^i|$, $|\alpha| \leq s$, and $d+1 < s$. Suppose that the element under consideration is of degree $t-1$, and suppose $s \leq t$. Then there exists a constant C such that, for all $v \in W^{(s)}(U)$,*

$$(5.2.11) \quad \|D^\alpha (v - Qv)\|_{0,U} \leq C |v|_{s,U}$$

Proof: Define a linear operator $A: W^{(s)}(U) \rightarrow L_2(U)$ by $Av = D^\alpha (v - Qv)$. By theorem 5.2.2 A is a bounded linear operator defined on all of $W^{(s)}(U)$. Since $Qv = v$ for all $v \in S$, and S contains all polynomials of degree less than s , A annihilates all polynomials of degree less than s . Therefore, by the Bramble-Hilbert lemma (theorem 1.4.1), there exists a constant C such that (5.2.11) holds. ||

A consequence of the hypotheses of theorem 5.2.3 is the inequality $d+1 < t$. Therefore this theorem is applicable to only those elements for which $d+1 < t$. (For n -dimensional elements the corresponding inequality is $d + \frac{n}{2} < t$.) However, this restriction does not cause any problems. This

author has never encountered an element which fails to satisfy $d + \frac{n}{2} < t$. In particular, each of the elements mentioned in this thesis satisfies the inequality, as can be easily checked.

Having obtained error bounds on U , we apply lemmas 5.1.1 and 5.2.1 to translate the results onto the element R . Recall that \bar{U} is mapped onto \bar{R} by the affine transformation (5.1.6)

$$w_i = h_i x_i + a_i \quad i=1,2.$$

A one-to-one correspondence between functions v on \bar{U} and functions \tilde{v} on \bar{R} is given by $\tilde{v}(w_1, w_2) = v(x_1, x_2)$. Let Q' denote, as before, the finite element interpolating projector on functions on $\bar{\Omega}$ (or their restrictions to \bar{R} .) One can easily verify that if $q = Qv$ then $\tilde{q} = Q'\tilde{v}$.

Theorem 5.2.4: *Let s be a real number satisfying $d+1 < s \leq t$, where $d = \max |\alpha^i|$ (cf. (5.2.8)) and $t-1$ is the degree of the element. Let α be a multiinteger such that $|\alpha| \leq s$. Then there exists a constant C such that for all $\tilde{v} \in W^{(s)}(R)$*

$$(5.2.12) \quad \|D^\alpha(\tilde{v} - Q'\tilde{v})\|_{0,R} \leq Ch^{s-|\alpha|} |\tilde{v}|_{s,R}.$$

Proof: Let w denote $v - Qv$. Then obviously $\tilde{w} = \tilde{v} - Q'\tilde{v}$. Applying lemma 5.1.1 with v replaced by $w = v - Qv$, we obtain

$$(5.2.13) \quad \|D^\alpha(\tilde{v} - Q'\tilde{v})\|_{0,R} = h_1^{-\alpha_1} h_2^{-\alpha_2} J^{\frac{1}{2}} \|D^\alpha(v - Qv)\|_{0,U}.$$

Recall that we have assumed the existence of a constant b such that $\frac{h}{h_i} \leq \frac{h}{g} \leq b$ ($i=1,2$). Therefore $h_i^{-1} \leq bh^{-1}$ ($i=1,2$). These inequalities applied to (5.2.13) imply

$$(5.2.14) \quad \|D^\alpha(\tilde{v} - Q'\tilde{v})\|_{0,R} \leq b^{|\alpha|} h^{-|\alpha|} J^{\frac{1}{2}} \|D^\alpha(v - Qv)\|_{0,U}.$$

Combining this inequality with inequalities (5.2.11) and (5.2.1) we get

(5.2.12). (Clearly the letter C has been used to denote different constants in different places.)

Theorem 5.2.5: Let s be a real number satisfying $d+1 < s \leq t$, where $d = \max |\alpha^i|$ (cf. (5.2.8)) and $t-1$ is the degree of the element. Suppose also that $m < s$. (Recall that m is the order of the Sobolev space in which the Elliptic problem is defined.) Then there exists a constant C such that for all $\tilde{v} \in \dot{W}^{(s)}(R)$,

$$(5.2.15) \quad \|\tilde{v} - Q'\tilde{v}\|_{m,R} \leq Ch^{s-m} |\tilde{v}|_{s,R}.$$

Proof: From theorem 5.2.4 we have, for all α such that $|\alpha| \leq m$,

$$\|D^\alpha(\tilde{v} - Q'\tilde{v})\|_{0,R} \leq C h^{s-|\alpha|} |\tilde{v}|_{s,R}.$$

The subscript α has been affixed to the constant to emphasize the constant's dependence on α . By definition

$$\|\tilde{v} - Q'\tilde{v}\|_{m,R}^2 = \sum_{|\alpha| \leq m} \|D^\alpha(\tilde{v} - Q'\tilde{v})\|_{0,R}^2$$

so

$$(5.2.16) \quad \|\tilde{v} - Q'\tilde{v}\|_{m,R}^2 \leq \left(\sum_{|\alpha| \leq m} C_\alpha^2 h^{2(s-|\alpha|)} \right) |\tilde{v}|_{s,R}^2.$$

We have assumed that $h \leq 1$. (This is the one point at which this assumption is used.) Therefore $h^{2(s-|\alpha|)} \leq h^{2(s-m)}$ for all α such that $|\alpha| \leq m$. It follows from this and (5.2.16) that

$$\|\tilde{v} - Q'\tilde{v}\|_{m,R}^2 \leq \left(\sum_{|\alpha| \leq m} C_\alpha^2 \right) h^{2(s-m)} |\tilde{v}|_{s,R}^2.$$

Letting $C^2 = \left(\sum_{|\alpha| \leq m} C_\alpha^2 \right)$ and taking square roots we get (5.2.15). ||

In order to use theorem 5.2.5 to predict convergence as h tends to zero, we must have $s > m$, as is obvious from (5.2.15). This requirement, together

with the hypothesis $s \leq t$, forces the inequality $m < t$. This is equivalent to $m \leq t-1$ because both m and t are integers. The interpretation of this inequality is that, in order to guarantee convergence of the finite element method for an elliptic problem in $W^{(m)}(\Omega)$ (e.g. a $2m$ -th order elliptic differential equation), one must use an element of degree at least m . This is a well known criterion.

We now present the main theorem of this section.

Theorem 5.2.6: *Let s be any real number satisfying $d+1 < s \leq t$, where $d = \max |\alpha_i|$ (cf. (5.2.8)) and $t-1$ is the degree of the element. Suppose also that $m < s$. Let C be the constant of theorem 5.2.5. Then for all $v \in W^{(s)}(\Omega)$,*

$$(5.2.17) \quad \|v - Q^s v\|_{m,\Omega} \leq Ch^{s-m} |v|_{s,\Omega}.$$

Proof: $\bar{\Omega}$ is the union of rectangular elements \bar{R}_i . Therefore

$$\|v - Q^s v\|_{m,\Omega}^2 = \sum_i \|v - Q^s v\|_{m,R_i}^2.$$

If s is an integer the similar equation

$$(5.2.18) \quad |v|_{s,\Omega}^2 = \sum_i |v|_{s,R_i}^2$$

holds, so (5.2.17) can be gotten by squaring (5.2.15), summing over all elements R_i , and taking square roots.

If s is not an integer (5.2.18) is not valid. However, we shall demonstrate the validity of

$$(5.2.19) \quad |v|_{s,\Omega}^2 \geq \sum_i |v|_{s,R_i}^2$$

which is sufficient to imply (5.2.17).

Recall that

$$|v|_{s,\Omega}^2 = \sum_{|\alpha|=s'} |D^\alpha v|_{\sigma,\Omega}^2$$

where $s=s'+\sigma$, s' is an integer, and $0<\sigma<1$. Therefore, to prove (5.2.19) it suffices to show that

$$|D^\alpha v|_{\sigma,\Omega}^2 \geq \sum_i |D^\alpha v|_{\sigma,R_i}^2$$

for all α satisfying $|\alpha|=s'$. But, letting

$$I = I(w, z) = \frac{|D^\alpha v(w) - D^\alpha v(z)|^2}{\|w - z\|^{2+2\sigma}}$$

we have $|D^\alpha v|_{\sigma,\Omega}^2 = \int_\Omega \int_\Omega I dw dz = \sum_i \sum_j \int_{R_i} \int_{R_j} I dw dz \geq \sum_i \int_{R_i} \int_{R_i} I dw dz = \sum_i |D^\alpha v|_{\sigma,R_i}^2$.

The following table shows the values of d and $t-1$ and the allowable values of s in theorem 5.2.6 for the two-dimensional elements considered in this thesis.

Table 5.1: Parameters for Theorem 5.2.6.

Element	d	$t-1$	s
Bilinear	0	1	$1 < s \leq 2$
Adini's Rectangle	1	3	$2 < s \leq 4$
20 d.o.f. and 24 d.o.f.	2	4	$3 < s \leq 5$
28 d.o.f.	2	5	$3 < s \leq 6$

Suppose for example that we are using Adini's rectangle to solve a second-order problem ($m=1$), and we know that the actual solution u is four times weakly differentiable, i.e. $u \in W^{(4)}(\Omega)$. Then, according to theorem 5.2.6, inequality (5.1.5), and table 5.1, the finite element solution u^* converges to the actual solution at the rate $O(h^3)$ as the mesh norm h tends to zero,

the error being measured with respect to the norm $\|\cdot\|_1$.

Bramble and Hilbert [5] have shown in the case in which s is an integer that it is sometimes possible (depending on the element) to strengthen inequality (5.2.17). We demonstrate this by means of an example. Suppose we wish to solve the generalized Dirichlet problem using the bilinear element. By table 5.1 the admissible values of s are $1 < s \leq 2$. We are interested in the case in which s is an integer, so we take $s=2$. Inequality (5.2.17) now takes the form

(5.2.20)

$$\|v-Qv\|_1 \leq Ch|v|_2 = Ch(\|D^{(2,0)}v\|_0^2 + \|D^{(1,1)}v\|_0^2 + \|D^{(0,2)}v\|_0^2)^{\frac{1}{2}}.$$

This inequality was obtained by utilizing the fact that the bilinear element is an element of degree one. That is, it preserves the monomials $1, x_1$, and x_2 . Nowhere was the fact that x_1x_2 is preserved used. According to theorem 2 of [5], the fact that x_1x_2 is preserved implies that (5.2.20) can be replaced by the stronger assertion

$$\|v-Qv\|_1 \leq Ch(\|D^{(2,0)}v\|_0^2 + \|D^{(0,2)}v\|_0^2)^{\frac{1}{2}}.$$

Similar improvements can be made for the other elements considered in this thesis.

We conclude this section with the promised example showing the value of considering noninteger Sobolev spaces. We consider the problem of the bending of a thin clamped plate under a point load at the point x_0 . The solution is the unique $u \in W^{0(2)}(\Omega)$ satisfying

$$a(u, v) = L(v) \quad \forall v \in W^{0(2)}(\Omega)$$

where (see Landau and Lifshitz [15])

$$\alpha(u, v) = \int_{\Omega} [\Delta u \Delta v + (1-\sigma)(2D_1 D_2 u D_1 D_2 v - D_1^2 u D_2^2 v - D_2^2 u D_1^2 v)]$$

and $L(v) = cv(x_0)$. Here Δ is the Laplacian, σ is Poisson's ratio [15], and c is a constant representing the force of the load. The boundary conditions on a clamped plate are $u|_{\partial\Omega} = \frac{\partial u}{\partial n}|_{\partial\Omega} = 0$. Thus the solution is required to be in $\dot{W}^{(2)}(\Omega)$.

The linear functional L is a distribution [16] and corresponds to the "generalized function" $c\delta(x)$, where δ is Dirac's " δ -function" concentrated at the point x_0 . By performing two integrations by parts on $\alpha(u, v)$ we see that the solution u satisfies the differential equation $\Delta^2 u = c\delta$ in the distributional sense.

By the Sobolev lemma (theorem 1.3.1) L is a bounded linear functional on $W^{(s)}(\Omega)$ for any $s > 1$. If we define $W^{(-s)}(\Omega)$ to be the dual space of $W^{(s)}(\Omega)$, then $L \in W^{(t)}(\Omega)$ for all $t < -1$. By a very general theorem in the book of Lions and Magenes ([16], pages 188-189) the solution u has four ($=2m$) more derivatives than L . Thus $u \in W^{(s)}(\Omega)$ for all $s < 3$. However, $u \notin W^{(3)}(\Omega)$. (There is a gap in the mathematics here. The theorem which has just been cited has as an hypothesis that the domain Ω has a smooth boundary, whereas the domains considered in this thesis have corners. This author doubts that this is a serious problem, especially for convex domains Ω .)

Suppose we wish to solve this problem numerically by the finite element method. We select an element of at least second degree, i.e. $t-1 \geq 2$. By theorem 5.2.6 and the ellipticity of the problem we have

$$\|u - u^*\|_2 \leq C_1 \|u - Q'u\|_2 \leq C_2 h^{s-2} |u|_s$$

provided that $u \in W^{(s)}(\Omega)$. We know that $u \in W^{(2)}(\Omega)$, but $u \notin W^{(3)}(\Omega)$, so we cannot predict convergence if we consider only integer order Sobolev spaces.

However, if we consider noninteger Sobolev spaces we can predict nearly $O(h)$ convergence because $u \in W^{(s)}(\Omega)$ for all $s < 3$.

(5.3) Error Bounds with Computable Constants

In this section the second approach to finite element error bounds is presented. Suppose that the situation is as stated in section 5.1. We know that to get a bound on $\|u - u^*\|_m$ it suffices to bound $\|u - Q'u\|_m$, where u is the actual solution of (5.1.3), u^* is the finite element solution, and $Q'u$ is the finite element interpolant of u . Bounds on $\|u - Q'u\|_m$ are, in turn, derived from bounds on $u - Qu$ on the unit square U , where u is any sufficiently smooth function on \bar{U} , and Qu is its interpolant.

As stated in section 5.1, we shall consider the case of Adini's rectangle in detail. For any $u \in C^1(\bar{U})$ let $q = Qu$ denote the Adini's rectangle interpolant of u . Thus q is the unique member of S such that $u = q$, $D_1 u = D_1 q$, and $D_2 u = D_2 q$ at the corners of \bar{U} , where S is the 12-dimensional polynomial space generated by the cubic polynomials and the monomials $x_1^3 x_2$ and $x_1 x_2^3$. In chapter four it was seen that the projector Q can be constructed using blending-function methods. Let us recall the construction explicitly. We first defined a function v around the perimeter of \bar{U} . v was defined on (for example) the bottom edge of \bar{U} to be the unique cubic polynomial $v(x_1, 0) = p(x_1)$ such that

$$\begin{aligned} p(0) &= u(0,0) & p(1) &= u(1,0) \\ p'(0) &= D_1 u(0,0) & p'(1) &= D_1 u(1,0). \end{aligned}$$

Let $p_0, p_1, q_0,$ and q_1 be the unique cubic polynomials such that

$$(5.3.1) \quad \left. \begin{aligned} p_i^{(j)}(0) &= \delta_{ij} = q_i^{(j)}(1) \\ p_i^{(j)}(1) &= 0 = q_i^{(j)}(0) \end{aligned} \right\} i, j=0,1.$$

Then clearly

$$v(x_1, 0) = \sum_{j=0}^1 \left[D_1^j u(0,0) p_j(x_1) + D_1^j u(1,0) q_j(x_1) \right].$$

We define projectors S_1 and S_2 by

$$(5.3.2) \quad \begin{aligned} S_1 u(x_1, x_2) &= \sum_{j=0}^1 \left[D_1^j u(0, x_2) p_j(x_1) + D_1^j u(1, x_2) q_j(x_1) \right] \\ S_2 u(x_1, x_2) &= \sum_{j=0}^1 \left[D_2^j u(x_1, 0) p_j(x_2) + D_2^j u(x_1, 1) q_j(x_2) \right]. \end{aligned}$$

Then clearly $v(x_1, 0) = S_1 u(x_1, 0)$ and $v(x_1, 1) = S_1 u(x_1, 1)$. Also $v(0, x_2) = S_2 u(0, x_2)$ and $v(1, x_2) = S_2 u(1, x_2)$.

Having defined v around the perimeter of \bar{U} , we then defined $q=Qu$ to be the blended interpolant of v based on linear blending functions. Thus

$$(5.3.3) \quad Qu = Pv$$

where P is defined by

$$(5.3.4) \quad P = P_1 + P_2 - P_1 P_2$$

and P_1 and P_2 are given by

$$P_1 v(x_1, x_2) = v(0, x_2)(1-x_1) + v(1, x_2)x_1$$

$$P_2 v(x_1, x_2) = v(x_1, 0)(1-x_2) + v(x_1, 1)x_2.$$

Clearly $P_1 v$ is completely determined by $v(0, x_2)$ and $v(1, x_2)$. It was seen above that $v(j, x_2) = S_2 u(j, x_2)$ ($j=0,1$), so $P_1 v = P_1 S_2 u$. Similarly $P_2 v = P_2 S_1 u$.

$P_1 P_2 v$ is determined by the values of v at the corners of \bar{U} . v interpolates u at the corners of \bar{U} , so $P_1 P_2 v = P_1 P_2 u$. We can therefore rewrite (5.3.3) and (5.3.4) as

$$(5.3.5) \quad Qu = P_1 S_2 u + P_2 S_1 u - P_1 P_2 u.$$

A straightforward computation shows that P_1 commutes with S_2 and P_2 commutes with S_1 . Thus (5.3.5) can be rewritten as

$$(5.3.6) \quad Qu = S_2 P_1 u + S_1 P_2 u - P_1 P_2 u$$

(cf. [12], page 117, equation (20)). Our aim is to obtain bounds on $u - Qu = (I - Q)u$. We already have bounds on $(I - P)u$ (theorem 2.2.5 with $k=1$), so it will suffice to derive bounds for $(P - Q)u$. By (5.3.4) and (5.3.6),

$$(5.3.7) \quad (P - Q)u = (I - S_2)P_1 u + (I - S_1)P_2 u$$

(cf. [12], page 117, (19)). We therefore begin by obtaining bounds for $(I - S_2)v$ and $(I - S_1)v$ for sufficiently smooth functions v on \bar{U} .

Lemma 5.3.1: *Let i and s be integers with $i \leq s$ and $2 \leq s \leq 4$. Then*

$$(5.3.8) \quad \|D_1^i (I - S_1)v\|_0 \leq B_{si} \|D^{(s,0)}v\|_0 \quad \forall v \in W^{(s,0)}(U)$$

$$(5.3.9) \quad \|D_2^i (I - S_2)v\|_0 \leq B_{si} \|D^{(0,s)}v\|_0 \quad \forall v \in W^{(0,s)}(U)$$

where $B_{si} = (s+1)^{\frac{1}{2}} (1 + K_i (5/2)^{\frac{1}{2}})$ and

$$K_i = \left(\int_0^1 \left[\sum_{j=0}^1 (|p_j^{(i)}(x)| + |q_j^{(i)}(x)|)^2 dx \right]^{\frac{1}{2}} \right)^{\frac{1}{2}}.$$

Here p_j, q_j are defined by (5.3.1).

Proof: We shall prove (5.3.8). By (5.3.2)

$$D_1^i S_1 v(x_1, x_2) = \sum_{j=0}^1 (D_1^j v(0, x_2) p_j^{(i)}(x_1) + D_1^j v(1, x_2) q_j^{(i)}(x_1)).$$

Therefore

$$(5.3.10) \quad |D_1^i S_1 v(x_1, x_2)| \leq \left| \sum_{j=0}^1 (|p_j^{(i)}(x_1)| + |q_j^{(i)}(x_1)|) \right| \max_{j=0,1} \{ |D_1^j u(0, x_2)|, |D_1^j u(1, x_2)| \}.$$

By corollary 1.3.5 (modified Sobolev lemma) with $n=1$ and $p=2$ we have

$$(5.3.11) \quad \max_{j=0,1} \{ |D_1^j v(0, x_2)|, |D_1^j v(1, x_2)| \} \leq \left(\frac{5}{2} \right)^{\frac{1}{2}} \left[\int_0^1 \left(|v(x_1, x_2)|^2 + |D_1 v(x_1, x_2)|^2 + |D_1^2 v(x_1, x_2)|^2 \right) dx_1 \right]^{\frac{1}{2}}$$

Substituting (5.3.11) into (5.3.10), squaring both sides of the inequality, integrating with respect to x_1 and x_2 , and taking square roots, we get

$$\|D_1^i S_1 v\|_0 \leq \left(\frac{5}{2} \right)^{\frac{1}{2}} K_i \left(\|v\|_0^2 + \|D_1 v\|_0^2 + \|D_1^2 v\|_0^2 \right)^{\frac{1}{2}} = \left(\frac{5}{2} \right)^{\frac{1}{2}} K_i \|v\|_{(2,0)}.$$

It follows immediately that for any integer $s \geq 2$, $\|D_1^i S_1 v\|_0 \leq \left(\frac{5}{2} \right)^{\frac{1}{2}} K_i \|v\|_{(s,0)}$. Also, it is trivially true that for any $i \leq s$, $\|D_1^i v\|_0 \leq \|v\|_{(s,0)}$. Therefore,

by the triangle inequality,

$$(5.3.12) \quad \|D_1^i (I-S_1)v\|_0 \leq \left(1 + \left(\frac{5}{2} \right)^{\frac{1}{2}} K_i \right) \|v\|_{(s,0)}$$

if $i \leq s$ and $2 \leq s$.

By the construction of S_1 , $S_1 v = v$ for all v which are cubic in x_1 for each fixed x_2 . Thus $S_1 v = v$ if $D^{(4,0)} v = 0$. This implies that $D_1^i (I-S_1)v = 0$ if $D^{(4,0)} v = 0$. For $s=2,3$, or 4 and $i \leq s$ define a linear operator $A: W^{(s,0)}(U) \rightarrow L_2(U)$ by $Av = D_1^i (I-S_1)v$. By (5.3.12) A is bounded, and we have seen that A annihilates all v for which $D^{(s,0)} v = 0$. Therefore, applying the modified Bramble-Hilbert lemma (theorem 1.5.6) with $\beta = (s,0)$ and $p=2$ we have, for $i \leq s$ and $2 \leq s \leq 4$,

$$\|D_1^i (I-S_1)v\|_0 \leq (s+1)^{\frac{1}{2}} \left(1 + \left(\frac{5}{2} \right)^{\frac{1}{2}} K_i \right) \|D^{(s,0)} v\|_0$$

for all $v \in W^{(s,0)}(U)$.

Lemma 5.3.2: Let $\alpha = (\alpha_1, \alpha_2)$ be a multiinteger and let $s=2,3$, or 4. Then for all $u \in W^{(s,1)}(U)$

$$(5.3.13) \quad \|D^\alpha (I-S_1)P_2 u\|_0 \leq B_{s\alpha_1} M_{\alpha_2} \left(\frac{5}{2}\right)^{\frac{1}{2}} \left(\|D^{(s,0)} u\|_0^2 + \|D^{(s,1)} u\|_0^2 \right)^{\frac{1}{2}} \text{ if } \alpha_1 \leq s.$$

For all $u \in W^{(1,s)}(U)$,

$$(5.3.14) \quad \|D^\alpha (I-S_2)P_1 u\|_0 \leq B_{s\alpha_2} M_{\alpha_1} \left(\frac{5}{2}\right)^{\frac{1}{2}} \left(\|D^{(0,s)} u\|_0^2 + \|D^{(1,s)} u\|_0^2 \right)^{\frac{1}{2}} \text{ if } \alpha_2 \leq s.$$

Here M_{α_i} is given by $M_0=1$, $M_1=2$, and $M_i=0$ for $i \geq 2$. $B_{s\alpha_i}$ is as defined in lemma 5.3.1.

Proof: We shall prove (5.3.13). It is easy to show that D_2 commutes with $I-S_1$. Thus $D^\alpha (I-S_1)P_2 u = D_1^{\alpha_1} (I-S_1)D_2^{\alpha_2} P_2 u$, and it follows from lemma 5.3.1 that

$$\|D^\alpha (I-S_1)P_2 u\|_0 \leq B_{s\alpha_1} \|D^{(s,\alpha_2)} P_2 u\|_0.$$

Therefore we will be done if we can show that

$$(5.3.15) \quad \|D^{(s,\alpha_2)} P_2 u\|_0 \leq M_{\alpha_2} \left(\frac{5}{2}\right)^{\frac{1}{2}} \left(\|D^{(s,0)} u\|_0^2 + \|D^{(s,1)} u\|_0^2 \right)^{\frac{1}{2}}.$$

Recall that P_2 is given by

$$P_2 u(x_1, x_2) = u(x_1, 0)(1-x_2) + u(x_1, 1)x_2.$$

Therefore

$$D^{(s,\alpha_2)} P_2 u(x_1, x_2) = D_1^s u(x_1, 0) D_2^{\alpha_2} (1-x_2) + D_1^s u(x_1, 1) D_2^{\alpha_2} x_2.$$

It follows that

$$(5.3.16) \quad |D^{(s,\alpha_2)} P_2 u(x_1, x_2)| \leq M_{\alpha_2} \max\{|D_1^s u(x_1, 0)|, |D_1^s u(x_1, 1)|\}.$$

We apply corollary 1.3.5 (modified Sobolev lemma) with $n=1$ and $p=2$ to get

(5.3.17)

$$\max\{|D_1^s u(x_1, 0)|, |D_1^s u(x_1, 1)|\} \leq \left(\frac{5}{2}\right)^{\frac{1}{2}} \left[\int_0^1 \left(|D_1^s u(x_1, x_2)|^2 + |D_2 D_1^s u(x_1, x_2)|^2 \right) dx_2 \right]^{\frac{1}{2}}.$$

Substituting (5.3.17) into (5.3.16), squaring, integrating, and taking square roots, we get (5.3.15). ||

Theorem 5.3.3: Let $s=2, 3$, or 4 , and let α be a multiinteger such that $\alpha \leq (s, s)$. Then for all $u \in W^{(s, 1)}(U) \cap W^{(1, s)}(U)$,

$$\begin{aligned} \|D^\alpha (P-Q)u\|_0 &\leq F_{s\alpha} \left(\|D^{(s, 0)}u\|_0^2 + \|D^{(s, 1)}u\|_0^2 \right)^{\frac{1}{2}} \\ &\quad + G_{s\alpha} \left(\|D^{(0, s)}u\|_0^2 + \|D^{(1, s)}u\|_0^2 \right)^{\frac{1}{2}} \end{aligned}$$

where $F_{s\alpha} = B_{s\alpha_1} M_{\alpha_2} (5/2)^{\frac{1}{2}}$ and $G_{s\alpha} = B_{s\alpha_2} M_{\alpha_1} (5/2)^{\frac{1}{2}}$.

Proof: This theorem is an immediate consequence of lemma 5.3.2 and the representation (5.3.7). ||

To simplify matters we take $s=4$ from now on. F_α and G_α will denote the constants $F_{4\alpha}$ and $G_{4\alpha}$ defined in the statement of theorem 5.3.3.

Theorem 5.3.4: Let $\alpha=(0, 0)$, $(1, 0)$, or $(0, 1)$. Let $N_\alpha=30$ if $|\alpha|=0$, $N_\alpha=45$ if $|\alpha|=1$. Then for all $u \in W^{(4, 1)}(U) \cap W^{(1, 4)}(U) \cap W^{(2, 2)}(U)$,

$$\begin{aligned} \|D^\alpha (I-Q)u\|_0 &\leq N_\alpha \|D^{(2, 2)}u\|_0 \\ &\quad + F_\alpha \left(\|D^{(4, 0)}u\|_0^2 + \|D^{(4, 1)}u\|_0^2 \right)^{\frac{1}{2}} \\ &\quad + G_\alpha \left(\|D^{(0, 4)}u\|_0^2 + \|D^{(1, 4)}u\|_0^2 \right)^{\frac{1}{2}} \end{aligned}$$

Proof: By theorem 2.2.5 with $E=I-P$, $k=1$, $\beta=(2, 2)$, and $p=2$,

$$\|D^\alpha (I-P)u\|_0 \leq N_\alpha \|D^{(2, 2)}u\|_0.$$

Theorem 5.3.4 now follows from the representation $D^\alpha (I-Q) = D^\alpha (I-P) + D^\alpha (P-Q)$ and theorem 5.3.3. ||

Now let R be a typical element, as in section 5.1. Suppose R has the dimensions $h_1 \times h_2$ with $h_i \leq h$ ($i=1,2$), and R is connected to U by the affine transformation (5.1.6). There is a number b , independent of R , such that $\frac{h}{h_i} \leq b$ ($i=1,2$).

Theorem 5.3.5: Let $\alpha=(0,0)$, $(1,0)$, or $(0,1)$, and let N_α , F_α , and G_α be as in theorem 5.3.4. Then for all $\tilde{u} \in W^{(4,1)}(R) \cap W^{(1,4)}(R) \cap W^{(2,2)}(R)$,

$$(5.3.18) \quad \begin{aligned} \|D^\alpha(I-Q')\tilde{u}\|_{0,R} &\leq b^{|\alpha|} h^{4-|\alpha|} \left[N_\alpha \|D^{(2,2)}\tilde{u}\|_{0,R} \right. \\ &\quad + F_\alpha \left(\|D^{(4,0)}\tilde{u}\|_{0,R}^2 + h^2 \|D^{(4,1)}\tilde{u}\|_{0,R}^2 \right)^{\frac{1}{2}} \\ &\quad \left. + G_\alpha \left(\|D^{(0,4)}\tilde{u}\|_{0,R}^2 + h^2 \|D^{(1,4)}\tilde{u}\|_{0,R}^2 \right)^{\frac{1}{2}} \right]. \end{aligned}$$

Proof: Apply lemma 5.1.1 to each term. ||

The appearance of the fifth order derivatives $D^{(4,1)}$ and $D^{(1,4)}$ is a weakness of the theory. It should be possible to eliminate them. At least their influence becomes small as h tends to zero.

Theorem 5.3.6: Theorem 5.3.5 remains true with R replaced by Ω .

Proof: $\bar{\Omega}$ is the union of elements \bar{R}_i . We square (5.3.18), sum over all elements R_i , and take square roots to get the desired result. ||

Theorem 5.3.7: Let $F = \left(\sum_{|\alpha| \leq 1} F_\alpha^2 \right)^{\frac{1}{2}}$, $G = \left(\sum_{|\alpha| \leq 1} G_\alpha^2 \right)^{\frac{1}{2}}$, and $N = (4950)^{\frac{1}{2}}$. Then for all $u \in W^{(4,1)}(\Omega) \cap W^{(1,4)}(\Omega) \cap W^{(2,2)}(\Omega)$,

$$(5.3.19) \quad \begin{aligned} \|u-Q'u\|_1 &\leq bh^3 \left[N \|D^{(2,2)}u\|_0 \right. \\ &\quad + F \left(\|D^{(4,0)}u\|_0^2 + h^2 \|D^{(4,1)}u\|_0^2 \right)^{\frac{1}{2}} \\ &\quad \left. + G \left(\|D^{(0,4)}u\|_0^2 + h^2 \|D^{(1,4)}u\|_0^2 \right)^{\frac{1}{2}} \right] \end{aligned}$$

Proof: According to the previous theorem, (5.3.18) holds with R replaced by Ω . We square (5.3.18), sum on α , and take square roots to get (5.3.19). The inequality $h \leq 1$ (or at least $h \leq b$) is required.

The constants N , F , and G can be readily computed with the aid of a calculator. We have $N \cong 70.36$ and $F = G \cong 29.19$. These constants are undoubtedly far from optimal, but they do indicate at least that the constants which appear in finite element error bounds are not so large as to make the error bounds worthless from a practical standpoint.

We now consider the problem of obtaining error bounds for the 24 d.o.f. element defined in chapter four. We shall pursue the problem only to the point of producing a representation analogous to (5.3.7).

Let us recall in detail the construction of this element. Given $u \in C^2(\bar{U})$ we defined the finite element interpolant $q = Qu$ in two stages. The first stage involved defining functions v and v_n on the boundary of \bar{U} to represent the boundary values of q and its normal derivative, respectively. Recall that $v(x_1, 0)$ (for instance) was defined to be the unique quintic polynomial $p(x_1) = v(x_1, 0)$ such that

$$\begin{aligned} p(0) &= u(0,0) & p(1) &= u(1,0) \\ p'(0) &= D_1 u(0,0) & p'(1) &= D_1 u(1,0) \\ p''(0) &= D_1^2 u(0,0) & p''(1) &= D_1^2 u(1,0) \end{aligned}$$

Let r_j, t_j ($j=0,1,2$) be the unique quintic polynomials satisfying

$$\left. \begin{aligned} r_j^{(i)}(0) &= \delta_{ij} = t_j^{(i)}(1) \\ r_j^{(i)}(1) &= 0 = t_j^{(i)}(0) \end{aligned} \right\} i, j=0,1,2..$$

If we define operators T_1 and T_2 by

$$T_1 u(x_1, x_2) = \sum_{j=0}^2 \left(D_1^j u(0, x_2) r_j(x_1) + D_1^j u(1, x_2) t_j(x_1) \right)$$

$$T_2 u(x_1, x_2) = \sum_{j=0}^2 \left(D_2^j u(x_1, 0) r_j(x_2) + D_2^j u(x_1, 1) t_j(x_2) \right)$$

then, as is easily seen,

$$(5.3.20) \quad \left. \begin{aligned} v(x_1, k) &= T_1 u(x_1, k) \\ v(k, x_2) &= T_2 u(k, x_2) \end{aligned} \right\} k=0,1.$$

The normal derivative $v_n(x_1, 0)$ was defined to be the unique cubic polynomial $m(x_1) = v_n(x_1, 0)$ such that

$$\begin{aligned} m(0) &= D_2 u(0, 0) & m(1) &= D_2 u(1, 0) \\ m'(0) &= D_1 D_2 u(0, 0) & m'(1) &= D_1 D_2 u(1, 0). \end{aligned}$$

Let S_1 and S_2 be the operators defined by (5.3.2). Then

$$(5.3.21) \quad \left. \begin{aligned} v_n(x_1, k) &\equiv D_2 v(x_1, k) = S_1 D_2 u(x_1, k) = D_2 S_1 u(x_1, k) \\ v_n(k, x_2) &\equiv D_1 v(k, x_2) = S_2 D_1 u(k, x_2) = D_1 S_2 u(k, x_2) \end{aligned} \right\} k=0,1.$$

Once we have defined the boundary values of v and v_n , we define q to be the blended interpolant of v based on Hermite cubic blending functions.

Thus

$$(5.3.22) \quad q = Qu = Sv = S_1 v + S_2 v - S_1 S_2 v$$

where S_1 and S_2 are defined by (5.3.2). It is convenient to write S_1 as a sum $S_1 = S_{10} + S_{11}$, where

$$S_{10} u(x_1, x_2) = u(0, x_2) p_0(x_1) + u(1, x_2) q_0(x_1)$$

$$S_{11} u(x_1, x_2) = D_1 u(0, x_2) p_1(x_1) + D_1 u(1, x_2) q_1(x_1).$$

The operator S_2 has an analogous decomposition $S_2 = S_{20} + S_{21}$. By (5.3.20)

we have

$$S_{10}v = S_{10}T_2u \quad S_{20}v = S_{20}T_1u.$$

By (5.3.21) we have

$$S_{11}v = S_{11}S_2u \quad S_{21}v = S_{21}S_1u.$$

Also $S_1S_2v = S_1S_2u$, as S_1S_2u is determined by the corner values of u , u_x , u_y , and u_{xy} , and v interpolates these values of u . Therefore (5.3.22) can be rewritten as

$$\begin{aligned} Qu &= S_{10}v + S_{11}v + S_{20}v + S_{21}v - S_1S_2v \\ &= S_{10}T_2u + S_{11}S_2u + S_{20}T_1u + S_{21}S_1u - S_1S_2u. \end{aligned}$$

It is easy to check that each pair of operators in the above sum commutes.

For instance $S_{10}T_2u = T_2S_{10}u$. Therefore we can subtract this sum from the sum $Su = S_{10}u + S_{11}u + S_{20}u + S_{21}u - S_1S_2u$ to obtain

$$(S-Q)u = (I-T_2)S_{10}u + (I-S_2)S_{11}u + (I-T_1)S_{20}u + (I-S_1)S_{21}u.$$

This expression can be used to obtain bounds for $S-Q$ in the same way that (5.3.7) was used to obtain bounds for $P-Q$ for the Adini element. Bounds for $u-Qu = (I-Q)u$ are then gotten by writing $I-Q = (I-S) + (S-Q)$ and using the bound for $(I-S)u$ provided by theorem 2.2.5 with $E = I-S$, $k=2$, $\beta=(3,3)$ and $p=2$.

CHAPTER SIX

NUMERICAL RESULTS

(6.1) Finite Element Programs

The subroutines of a finite element program fall into four categories depending on their function. The categories are preprocessing, assembly of the matrix equation (3.3.4), solution of the matrix equation, and interpretation of the solution. We shall consider these points one by one.

The main tasks involved in preprocessing are the subdivision of the given domain into elements, the numbering of the elements, and the numbering of the nodes. The preprocessing programs which this author has written are rather primitive. The domain Ω is assumed to be a rectangle, and the program reads in data telling how many rows and columns of elements there will be and where the horizontal and vertical mesh lines which define the elements are to be placed. The elements are numbered from left to right starting at the bottom of the rectangular domain. The main task is the numbering of the nodes. There are two numbering schemes. One is local and is determined by the element type. For a given d degree-of-freedom finite element scheme, each element has d nodes associated with it, and we number these nodes locally. The job of the preprocessing program is to give each node a number in the global numbering scheme. The nodes, like the elements, are numbered from left to right, starting at the bottom of the domain. The global number of the j th node of the i th element is stored in the (j,i) entry of a two-dimensional integer array named NODE. (The programs are written in FORTRAN.) Thus, if the j th node of the i th element is the k th node globally, we have $\text{NODE}(j,i) = k$. For nodes on

the boundary of the domain which are forced by boundary conditions to be zero, we set $\text{NODE}(J,I) = 0$. The boundary conditions can be altered by changing two short subroutines.

After the nodes have been numbered, the program reports the total number n of nodes unaffected by the boundary conditions. The order of the matrix equations to be solved is n .

The final task of the preprocessing program is to compute a *band parameter* which indicates the band width of the stiffness matrix which will be generated. The band parameter is not exactly the same as the band width, which was defined at the end of chapter three. The *band parameter* is the number of diagonals from the main diagonal to the last nonzero diagonal (in either direction, since the stiffness matrix is symmetric). Thus the band width is one less than twice the band parameter.

The second phase of the finite element program is the assembly of the matrix equation. We shall first consider the assembly of the stiffness matrix K . The (i,j) entry of K is $\alpha(\psi_i, \psi_j)$, where ψ_i is the basis function whose i th nodal value (in the global numbering scheme) is one and whose other nodal values are zero, and $\alpha(\cdot, \cdot)$ is a symmetric, bounded, strongly elliptic bilinear form. Let us assume for definiteness that $\alpha(\cdot, \cdot)$ is given by (cf. (3.1.4))

$$(6.2.2) \quad \alpha(u,v) = \int_{\Omega} (u_x v_x + u_y v_y) \quad u, v \in W^{\Omega(1)}(\Omega).$$

(Here we have returned to the x,y notation of chapters three and four. We will use this notation from now on.) Let e_1, \dots, e_s be the elements into which Ω has been partitioned, and define $\alpha_p(\cdot, \cdot)$ by

$$a_r(u,v) = \int_{a_r} (u_x v_x + u_y v_y) \quad r=1, \dots, s.$$

Then

$$a(u,v) = \sum_{r=1}^s a_r(u,v).$$

The stiffness matrix is evaluated on an element by element basis. The integrals $a_1(\psi_i, \psi_j)$ are evaluated for all i and j , then the integrals $a_2(\psi_i, \psi_j)$ are evaluated, and so on. Most of these integrals will be zero trivially by the localness of the basis functions. Those integrals which are not zero are added to the appropriate entry of the stiffness matrix array, which was originally set equal to zero. The integrals

$$(6.2.2) \quad a_r(\psi_i, \psi_j) = \int_{e_r} (\psi_{i_x} \psi_{j_x} + \psi_{i_y} \psi_{j_y})$$

are integrals of polynomials and can therefore be integrated analytically.

However, it is convenient to integrate by numerical quadrature [14], [22] instead. For rectangular elements, product Gauss quadrature rules are the obvious choice. This author has used them exclusively. The integrals (6.1.2) can be evaluated exactly by a quadrature formula of sufficiently high degree, but this can prove expensive for elements of high degree.

It is better to "cheat" and use a lower-degree formula. This is one of the "variational crimes" discussed by Strang and Fix [21]. In section 6.3 we shall discuss the specific quadrature rules used for various elements.

Because the stiffness matrix K is symmetric, only the main diagonal and the lower half of K need be stored. In fact, if the band parameter of K is m , only m diagonals need be stored. If the order of K is n , and the m relevant diagonals of K are stored one after the other in a one-dimensional array, the storage requirement is only $nm - \frac{1}{2}m(m-1)$.

The *load vector*, i.e. the right hand side of the matrix equation, must also be computed. The entries of the load vector are generally of the form

$$L(\psi_i) = (f, \psi_i)_0 = \int_{\Omega} f \psi_i$$

and can be evaluated by numerical quadrature in an element-by-element manner.

The matrix equation, once assembled, is solved by the *Choleski* or *square root* method [9]. In this method a lower triangular matrix G such that $K = GG^T$ is computed. The matrix equation is then solved by forward elimination and back substitution. The matrix G inherits the band structure of K , so G takes no more storage space than does K . The order of computation of the entries of G is such that once G_{ij} has been computed, the value of K_{ij} is no longer needed. Thus it is possible to store G over K .

In practice there is some overlap between the matrix assembly phase and the equation solving phase of the program. Once the stiffness matrix has been computed, it is immediately decomposed into GG^T . G is stored where K was; K is lost. In practice there may be more than one load vector. We may wish to solve several equations $GG^T x^i = y^i$. The first load vector is computed, the forward elimination and back substitution are carried out, and the first solution vector is obtained. This process stores the solution vector over the load vector but does not destroy G . Once a solution vector is obtained, the interpretive subroutines are called, and the solution is translated into usable data, which is printed out. Only after the solution

has been interpreted is the next load vector computed.

The interpretive phase of the problem is the least clear cut and most difficult phase. There are many possibilities, the simplest of which is to merely print out the solution vector. This is not a total loss, as the i th entry of the solution vector is just the i th nodal value of the finite element solution. That is, it is the actual value of the finite element solution or one of its derivatives at a specific point. The obvious next step is to write a subroutine which evaluates the finite element solution and/or certain of its derivatives at points other than the nodes. This procedure is not very satisfactory because it causes the generation of great tables of numbers which are not easily interpreted. A better idea is to display the data graphically with the aid of a plotter.

I have done none of this (except print out the solution vector), as my main intention was to measure the error in the Sobolev norm and determine whether the rates of convergence predicted in chapter five are attained in practice. Accordingly, I have chosen problems whose actual solution is known and can be programmed. For such problems it is possible to calculate the Sobolev norm of the error between the actual solution and the finite element solution. The program which I have written evaluates the norm of the error by numerically integrating over each element using the 5x5 product Gauss. rule. This rule gives exact integrals for polynomials of degree as high as nine in each variable and can be expected to give accurate results for smooth functions. Accurate results will certainly be obtained on fine meshes, as the rate of convergence of the rule is $O(h^{10})$ as $h \rightarrow 0$, where h is the mesh norm.

There is one other subroutine which has not been mentioned. This routine should be classified as a preprocessor, even though it is called at the end of the program. Its function is to refine the mesh by inserting new mesh lines midway between the existing mesh lines, thus halving the mesh norm h and quadrupling the number of elements. Once the refinement is carried out, the entire program is run again.

(6.2) Confirmation of rates of Convergence

The model problem which we shall consider is the Dirichlet problem for Poisson's equation:

$$(6.2.1) \quad \begin{aligned} -\Delta u &= f & \text{on } \Omega \\ u &= 0 & \text{on } \partial\Omega \end{aligned}$$

where Ω is the unit square. Two load functions

$$\begin{aligned} f_1(x,y) &= (6x-10)(y^4-y^3+y^2-y) + (x^3-5x^2+4x)(12y^2-6y+2) \\ f_2(x,y) &= e^x(\sin\pi x[(2\pi^2-1)y\sin\pi y-2\pi\cos\pi y] - 2\pi y\cos\pi x\sin\pi y) \end{aligned}$$

will be considered. The respective solutions of (6.2.1) are

$$\begin{aligned} u_1(x,y) &= (x^3-5x^2+4x)(y-y^2+y^3-y^4) \\ u_2(x,y) &= (e^x\sin\pi x)(y\sin\pi y). \end{aligned}$$

I have implemented three elements -- the bilinear element, Adini's rectangle, and the 24 d.o.f. element constructed in chapter four. Each of these elements has been used to calculate approximate solutions of (6.2.1) for the two load functions f_1 and f_2 , using a variety of mesh sizes. The programs were run on the CDC 6400 computer of the University of Calgary Data Centre.

Tables 6.1 through 6.4 below indicate the Sobolev norm $\|u-u^*\|_1$ of the difference between the actual solution u and the finite element solution u^* for the three elements. Only regular meshes with square elements were used, so each mesh is uniquely determined by its mesh norm h . To give an idea of the scale of the functions involved, the approximate norms of the solutions u_1 and u_2 are

$$\|u_1\|_1 \cong .714$$

$$\|u_2\|_1 \cong 2.276$$

From theorem 5.2.6 and table 5.1 we expect convergence at the rate $O(h)$ for the bilinear element. This expectation is confirmed by the data in tables 6.1 and 6.2.

Table 6.1: $\|u-u^*\|_1$, Bilinear Element

mesh norm load function	.5	.25	.125	.0625
f_1	.412	.207	.103	.0516
f_2	1.52	.717	.355	.177

Table 6.2: $\|u-u^*\|_1$, Bilinear Element

mesh norm load function	.2	.1	.05
f_1	.165	.0826	.0413
f_2	.571	.283	.142

With Adini's rectangle we expect $O(h^3)$ convergence by theorem 5.2.6 and table 5.1, or by theorem 5.3.7. Table 6.3 indicates that this rate of convergence is attained.

Table 6.3: $\|u-u^*\|_1$, Adini's Rectangle

mesh norm load function	.5	.25	.125
f_1	$.903 \times 10^{-1}$	$.122 \times 10^{-1}$	$.163 \times 10^{-2}$
f_2	.429	$.538 \times 10^{-1}$	$.750 \times 10^{-2}$

The 24 d.o.f. element should give $O(h^4)$ convergence. Table 6.4 indicates that it does.

Table 6.4: $\|u-u^*\|_1$, 24 d.o.f. Element

mesh norm load function	.5	.25
f_1	$.220 \times 10^{-2}$	$.114 \times 10^{-3}$
f_2	$.209 \times 10^{-1}$	$.981 \times 10^{-3}$

(6.3) Comparative Cost of Running the Programs

Comparing two different finite element programs is difficult because the relative significance of the various phases of the program is different for elements of low degree than for elements of high degree. For example, the time required for the assembly of the stiffness matrix is virtually

negligible for the bilinear element, whereas it is significant for the other two elements. There are three reasons for this. The first is that in the element-by-element process by which the stiffness matrix is assembled, the number of integrals $a_r(\psi_i, \psi_j)$ to be evaluated in each element is $\frac{1}{2}d(d+1)$, where d is the number of degrees of freedom of the element. Thus the number of integrals to be evaluated in each element grows quadratically with d . The second reason is that each of the function evaluations required for numerical quadrature takes comparatively little time for a low-degree element. This is because low-degree polynomials can be evaluated more quickly than can high-degree polynomials. Specifically, a typical function evaluation for the bilinear element requires one multiplication, whereas a function evaluation for Adini's rectangle or the 24 d.o.f. element requires about four or eight multiplications, respectively. The third reason is that a quadrature rule of low degree can be used to evaluate the integrals for low-degree elements. Thus only a few function evaluations are needed to evaluate each integral. For example, for the bilinear element the 2×2 product Gauss rule integrates the integrals $a_r(\psi_i, \psi_j) = \int_{e_r} (\psi_{i_x} \psi_{j_x} + \psi_{i_y} \psi_{j_y})$ exactly. In fact, the 1×1 rule gives results which are almost as good. The rate of convergence is still $O(h)$, as predicted by Strang and Fix [21]. However, the reduction of cost realized in changing from the four-point rule to the one-point rule is insignificant. It was this observation which convinced this author that the assembly time of the stiffness matrix for the bilinear element is negligible. The same is not true for the two more complex elements. In the case of Adini's rectangle costs can be cut considerably by using the 3×3 Gauss rule rather than the 4×4 rule which would be required to calculate the integrals exactly. The results obtained

using 3x3 quadrature are as good as those given by the 4x4 rule. On the other hand, the results gotten using 2x2 quadrature were poor. A similar situation holds for the 24 d.o.f. element. In this case the 6x6 rule would be needed to integrate the terms exactly. This is out of the question; thirty-six points are too many. It was found that the 4x4 and 5x5 rules give equally good results, whereas the 3x3 rule gives poor results. The cost of running the program is reduced considerably if the 4x4 rule is used instead of the 5x5 rule. Interestingly, the theory of [21] predicts that the use of the 4x4 rule instead of the 5x5 rule will cause a reduction in the rate of convergence from $O(h^4)$ to $O(h^3)$. This worsening of the convergence rate was not observed.

So far it appears that the low-degree elements have the upper hand, but we have not yet taken into account the fact that to attain a given accuracy a much finer mesh is required for a low-degree element than for a high-degree element. The use of a coarser mesh for high-degree elements partially compensates for the slowness of the stiffness matrix assembly. Other benefits are derived from the fact that the stiffness matrix is of relatively low order if the mesh is coarse. We shall consider a specific example. From tables 6.2 and 6.3 we see that even with a mesh norm of $h=.05$ the bilinear element gives worse solutions than does Adini's rectangle with $h=.25$. The stiffness matrix for the bilinear element with $h=.05$ is of order 361 with a band parameter of 21. By contrast, the stiffness matrix for Adini's rectangle with $h=.25$ is of order 39 with band parameter 17. As was mentioned previously, the storage requirement for the stiffness matrix is $nm - \frac{1}{2}n(m-1)$, or roughly nm , where n is its order and m is its

band parameter. For the bilinear element $nm = 7581$, whereas for Adini's rectangle $nm = 663$.

Storage space is not the only problem. Obviously a large matrix equation takes longer to solve than a small matrix equation. The most significant step is the decomposition $K=GG^T$. This decomposition involves about nm^2 multiplications [9]. Thus the decomposition time for Adini's rectangle in the example under consideration is only about seven percent the decomposition time for the bilinear element.

Having established the fact that it is difficult to compare finite element programs by analyzing the subroutines, we must resort to a very crude method of comparison: we compare the actual cost of running the programs. It happens that the run in which the data of table 6.2 was compiled cost just slightly more than the run for table 6.3, which, in turn, cost just a bit more than did the run for table 6.4. In each case the bulk of the time was spent on the most refined mesh. The results of table 6.4 (24 d.o.f. element) are significantly better than those of table 6.3 (Adini's rectangle), which are much better than those of table 6.2 (bilinear element). On this basis we can conclude that the 24 d.o.f. element is best for the given problems, and the bilinear element is worst. We should be extremely cautious about concluding that the 24 d.o.f. element is better than the others. We have considered only one problem with two sets of data. In both cases the data and the solutions are smooth. The lower-degree elements might fare better in a competition in which there are singularities in the data.

(6.4) Pointwise Approximation

In section 6.1 it was noted that the solution vector of the finite element matrix equation (3.3.4) gives the nodal values of the finite element solution. Some of this information has been compiled in tables 6.5 through 6.7. All data in these tables pertains to the load function f_1 and the corresponding solution u_1 . In view of the conclusions drawn in the previous section, the only surprise in these tables is the relatively poor showing of the 24 d.o.f. element in table 6.6.

Table 6.5:

Estimated and Actual Values of u_1 .

point	Bilinear Element $h=.0625$	Adini's Rectangle $h=.125$	24 d.o.f. Element $h=.25$	Actual Value
$(\frac{1}{4}, \frac{1}{4})$.1405	.140088	.140079	.140076
$(\frac{1}{2}, \frac{1}{4})$.1748	.174325	.174318	.174316
$(\frac{3}{4}, \frac{1}{4})$.1217	.121406	.121398	.121401
$(\frac{1}{4}, \frac{1}{2})$.2204	.219750	.219730	.219727
$(\frac{1}{2}, \frac{1}{2})$.2743	.273455	.273440	.273438
$(\frac{3}{4}, \frac{1}{2})$.1910	.190444	.190432	.190430
$(\frac{1}{4}, \frac{3}{4})$.2068	.206045	.205997	.205994
$(\frac{1}{2}, \frac{3}{4})$.2572	.256389	.256350	.256348
$(\frac{3}{4}, \frac{3}{4})$.1792	.178560	.178530	.178528

Table 6.6: Estimated and Actual Values of $\frac{\partial u_1}{\partial x}$.

point	Adini's Rectangle $h=.125$	24 d.o.f. Element $h=.25$	Actual Value
$(\frac{1}{4}, \frac{1}{4})$.33615	.33607	.33618
$(\frac{1}{2}, \frac{1}{4})$	-.04986	-.04979	-.04805
$(\frac{3}{4}, \frac{1}{4})$	-.36115	-.36096	-.36108
$(\frac{1}{4}, \frac{1}{2})$.52729	.52721	.52734
$(\frac{1}{2}, \frac{1}{2})$	-.07822	-.07811	-.07813
$(\frac{3}{4}, \frac{1}{2})$	-.56652	-.56625	-.56641
$(\frac{1}{4}, \frac{3}{4})$.49430	.49427	.49438
$(\frac{1}{2}, \frac{3}{4})$	-.07344	-.07323	-.07324
$(\frac{3}{4}, \frac{3}{4})$	-.53124	-.53087	-.53101

Table 6.7: Estimated and Actual Values of $\frac{\partial u_1}{\partial y}$.

point	Adini's Rectangle $h=.125$	24 d.o.f. Element $h=.25$	Actual Value
$(\frac{1}{4}, \frac{1}{4})$.43950	.43947	.43945
$(\frac{1}{2}, \frac{1}{4})$.54691	.54689	.54688
$(\frac{3}{4}, \frac{1}{4})$.38089	.38086	.38087
$(\frac{1}{4}, \frac{1}{2})$.17619	.17578	.17578
$(\frac{1}{2}, \frac{1}{2})$.21910	.21875	.21875
$(\frac{3}{4}, \frac{1}{2})$.15261	.15234	.15234
$(\frac{1}{4}, \frac{3}{4})$	-.35095	-.35158	-.35156
$(\frac{1}{2}, \frac{3}{4})$	-.43697	-.43752	-.43750
$(\frac{3}{4}, \frac{3}{4})$	-.30428	-.30470	-.30469

References

1. A. Adini and R.W. Clough, Analysis of Plate Bending by the Finite Element Method, NSF Report G. 7337, 1961.
2. I. Babuška and A.K. Aziz, Survey Lectures on the Mathematical Foundations of the Finite Element Method, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, A.K. Aziz Ed., pp. 3-359, Academic Press, New York, 1972.
3. R.E. Barnhill and J.A. Gregory, Blending Function Interpolation to Boundary Data on Triangles, Technical Report TR/14, Department of Mathematics, Brunel University, 1972.
4. J.H. Bramble and S.R. Hilbert, Estimation of Linear Functionals on Sobolev Spaces with Applications to Fourier Transforms and Spline Interpolation, *SIAM J. Numer. Anal.*, v. 7, 1970, pp. 113-124.
5. J.H. Bramble and S.R. Hilbert, Bounds for a Class of Linear Functionals with Applications to Hermite Interpolation, *Numer. Math.*, v. 16, 1971, pp. 362-369.
6. J.H. Bramble and M. Zlamal, Triangular Elements in the Finite Element Method, *Math. of Comp.*, v. 24, 1970, pp. 809-820.
7. J. Céa, Approximation Variationnelle des Problèmes aux Limites, *Ann. Inst. Fourier (Grenoble)*, v. 14, 1964, pp. 345-444.
8. S.A. Coons, Surfaces for Computer Aided Design of Space Forms, Project MAC, Design Div., Department of Mechanical Engineering, MIT, 1964. Revised to MAC-TR-41, 1967.
9. G. Forsythe and C. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, 1967.
10. A. Friedman, *Partial Differential Equations*, Holt, Rinehart, and Winston, New York, 1969.
11. W. J. Gordon, Blending-Function Methods of Bivariate and Multivariate Interpolation and Approximation, *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 158-177.
12. W.J. Gordon and C.A. Hall, Transfinite Element Methods: Blending-Function Interpolation over Arbitrary Curved Element Domains, *Numer. Math.*, v. 21, 1973, pp. 109-129.
13. S.G. Krein and Yu. I. Petunin, Scales of Banach Spaces, *Russ. Math. Surveys*, v. 21, 1966, pp. 85-159.
14. V.I. Krylov, *Approximate Calculation of Integrals*, Macmillan, New York, 1962.

15. L.D. Landau and E.M. Lifshitz, *Theory of Elasticity*, Pergamon, Addison-Wesley, Reading, 1959.
16. J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications I*, Springer-Verlag, New York, 1972.
17. S.G. Mikhlin and K.L. Smolitskiy, *Approximate Methods for Solution of Differential and Integral Equations*, Elsevier, New York, 1967.
18. W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
19. L.N. Slobodetskiy, Generalized Sobolev Spaces and Their Applications to Boundary Problems for Partial Differential Equations, A.M.S. Translations, Ser: 2, v. 57, 1966, pp. 207-275.
20. V.I. Smirnov, *A Course of Higher Mathematics*, v. 5, *Integration and Functional Analysis*, Pergamon, Addison-Wesley, Reading, 1964.
21. G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, 1973.
22. A.H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall, Englewood Cliffs, 1971.
23. D.S. Watkins, Analysis, in Lecture Notes on the Preliminary Programme, NATO/NRC International Research Seminar on the Theory and Application of Finite Element Methods, The University of Calgary, 1973.
24. O.C. Zienkiewicz, *The Finite Element Method in Engineering Science*, McGraw-Hill, Ltd., London, 1971.
25. M. Zlámal, On the Finite Element Method, Numer. Math., v. 12, 1968, pp. 394-409.