INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106

76-27,501

JESPERSEN, Dennis Charles, 1949-
A LEAST SQUARES DECOMPOSITION METHOD FOR THE NUMERICAL SOLUTION OF ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS.

The University of Michigan, Ph.D., 1976
Mathematics

Xerox University Microfilms, Ann Arbor, Michigan 48106

A LEAST SQUARES DECOMPOSITION METHOD FOR THE NUMERICAL SOLUTION OF ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS

by
Dennis Charles Jespersen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mathematics)
in The University of Michigan
1976

Doctoral Committee:

Professor George J. Fix, Co-Chairman, Carnegie-Mellon University
Assistant Professor Jeffrey B. Rauch, Co-Chairman
Professor Peter L. Duren
Assistant Professor Dianne P. O'Leary
Professor Wei H. Yang

DEDICATION

To my parents

ACKNOWLEDGMENTS

I thank George Fix for his support and inspiration, and Gerard Richter for his constant encouragement.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGMENTS iii

CHAPTER

I. INTRODUCTION 1

II. PRELIMINARIES 6

III. ERROR ESTIMATES 14

IV. FURTHER REMARKS 35

The Condition Number 42
Further Error Bounds 49
The One-Dimensional Case 53
Other Boundary Conditions 57
The Difference Equations 64

V. A GENUINE ELLIPTIC SYSTEM 67

BIBLIOGRAPHY 82

CHAPTER I. INTRODUCTION

The study of numerical methods for solving partial differential equations has undergone tremendous growth in the past few decades. Before the 1940's the subject was almost nonexistent, but with the advent of electronic computers in that decade computations involving a large number of arithmetic steps became routinely possible, and the subject "took off."

The numerical methods first proposed and analyzed were finite difference methods. In these methods derivatives in the differential equation are replaced by difference quotients and a system of algebraic equations results, which one is (hopefully) able to solve on one's computer. Book-length treatments of finite-difference methods may be found in [12] and [19].

In recent years finite element methods have come to the fore. The basic idea here is to replace a partial differential equation by an equivalent variational problem and then to compute a function which approximately satisfies the variational principle (the Rayleigh-Ritz-Galerkin idea). The approximations one works with are piecewise polynomials, this choice turning out to be both computationally attractive and mathematically tractable. This idea was already mentioned in 1943 by

Courant but not systematically put into practice until the 1960's, when engineers began to use the method for their problems. The last ten years has seen the appearance of a flood of literature on the finite element method, and the subject has matured enough to have received several book-length treatments [22], [24], [3].

The method of least squares approximation to be considered here has its roots in the following idea. Suppose one wishes to approximate the solution u of the equation $Lu = f$, where L is a partial differential operator. Then one idea is to choose a finite linearly independent set of trial functions $\{\phi_1,\ldots,\phi_M\}$ and seek an approximate solution $\tilde{u}$ in the span of the ϕ_i, $\tilde{u} = \Sigma c_i\phi_i$, where the c_i are found by requiring that

$$\|L\tilde{u} - f\|_{L_2}^2 = \inf_{\tilde{w}\in \mathrm{span}\{\phi_1,\ldots,\phi_M\}} \|L\tilde{w} - f\|_{L_2}^2 .$$

This requirement leads to a linear system of equations which can be solved to yield $c_1,\ldots,c_M$. The first rigorous treatment of this idea, including the question of handling the boundary conditions, was given by Bramble and Schatz [7]. They showed that if $\Omega \subset R^N$ is a bounded region with smooth boundary Γ and if L is a linear second-order elliptic operator, then the solution u of $Lu = f$ in Ω, $u = g$ on Γ, is "optimally" approximated (this notion will be clarified later) by the solution $\tilde{u}$ of the variational problem:

minimize $\|L\tilde{w} - f\|^2_{L_2(\Omega)} + h^{-3}\|\tilde{w} - g\|^2_{L_2(\Gamma)}$ over all trial functions $\tilde{w} \in \text{span}\{\phi_1,\ldots,\phi_M\}$ where the functions $\phi_1,\ldots,\phi_M$ span a finite element space. They also extended the analysis to the case of a general $2m^{th}$ order elliptic equation. Their method has the feature that the trial functions need satisfy no boundary conditions, which can be convenient if Ω has a non-polyhedral boundary. Also, the linear equations which arise have a symmetric positive definite matrix, which is convenient for numerical work. On the other hand, the trial functions are required to be smooth enough so that L may be applied to any of them and still produce an L_2 function. Also, the matrix that eventually arises has a condition number which is the square of the "expected" condition number (for a $2m^{th}$ order problem and N basis functions the expected condition number is $O(N^m)$). This difficulty can be overcome, at the expense of some additional complications [6]. Here the condition number of a symmetric positive definite matrix is defined as the largest eigenvalue divided by the smallest eigenvalue, and is relevant to the practical problem of solving the matrix equation which arises. That is, error bounds for the solutions of linear systems involve the condition number of the matrix, and the larger the condition number of the matrix the more sensitive is the numerical procedure to perturbations in the data and to rounding

errors [13].

The method to be considered later is also related to a variant of the finite element method known as the mixed method. Here the basic idea is to introduce the derivatives of the solution as new independent variables and to approximate the solution and its derivatives simultaneously. This idea has been studied in [10], [17], and [18]. A nice feature of this method is that one automatically obtains approximations to the derivatives of the solution. This is useful in some problems; for example in fluid mechanics the equation $\Delta\psi = \zeta$ arises, where the physical quantities, the velocities, are the derivatives of the stream function ψ. Another feature of this method is that less smoothness may be required of the trial functions. A possible disadvantage is that the matrices which arise are generally nonsymmetric, and difficult to prove nonsingular. Indeed, extra hypotheses may have to be introduced to prove the existence of a solution to the finite-dimensional problem. This situation is encountered in [10] and is further studied in [17]. Also, error bounds for mixed methods are difficult to obtain. Mixed finite element methods are currently being studied by several authors.

The numerical method to be considered here can be thought of as an attempt to combine the least squares

method and the mixed method, hopefully preserving the best features of each. Thus we hope the method will require no boundary conditions to be satisfied by the trial functions, will result in a symmetric positive definite matrix which has the "expected" condition number, and we hope to simultaneously find "optimal" approximations to both the solution and its derivatives. We note that the essential features of the method were apparently first seen in the engineering literature [15].

The plan of this work is as follows. We first briefly review the necessary tools from functional analysis and partial differential equations. Then we introduce the numerical method to be considered and prove L_2 error estimates. Then we consider variations on the basic theme, one of which involves weighting the terms differently, and again prove error estimates. We prove higher-order error estimates, assuming a certain hypothesis of "inverse" type. We also estimate the condition number of the finite-dimensional problem. We consider a problem with Neumann boundary conditions and prove error estimates. Finally we consider a true system of equations, the equations of two-dimensional elasticity, and again prove error estimates.

CHAPTER II. PRELIMINARIES

The purpose of this chapter is to introduce the mathematical machinery which will be necessary in the body of this work. Also, much of the notation is introduced. No attempt is made to keep the presentation self-contained. The basic references for this material are [14], [1], [5], [4].

Let $\Omega \subset R^N$ be a bounded open set with closure $\bar{\Omega}$. All functions to be considered will be real-valued. Let $C^\infty(\bar{\Omega}) = \{\phi:\Omega \to R \mid \phi$ has derivatives of all orders on Ω and all its derivatives are continuously extendable to $\bar{\Omega}\}$. For $\alpha = (\alpha_1,\ldots,\alpha_N)$ a multi-index of length $|\alpha| = \alpha_1 + \ldots + \alpha_N$, define $D^\alpha = \partial_1^{\alpha_1}\cdots\partial_N^{\alpha_N}$ where $\partial_i = \partial/\partial x_i$. For s a nonnegative integer define a norm on $C^\infty(\bar{\Omega})$ by

$$\|\phi\|_s^2 = \sum_{|\alpha|\leq s} \|D^\alpha\phi\|_{L_2(\Omega)}^2$$

where $L_2(\Omega)$ is the usual space of Lebesgue-square-integrable functions on Ω. Define $H^s(\Omega)$ as the completion of $C^\infty(\bar{\Omega})$ in the norm $\|\cdot\|_s$.

For s positive but non-integral, $H^s(\Omega)$ is defined by the technical process of interpolation. Then for each $s\geq 0$ $H^s(\Omega)$ is a Hilbert space with norm which will be denoted $\|\cdot\|_s$ and inner product to be denoted by $(\cdot,\cdot)_s$. In particular, $H^0(\Omega) = L_2(\Omega)$ and $\|\cdot\|_0 = \|\cdot\|_{L_2(\Omega)}$.

The inner product in $H^0(\Omega)$ will be denoted simply $(\cdot,\cdot)$.

We will frequently use the following facts. For $t > s \geq 0$, $H^t(\Omega) \subset H^s(\Omega)$ and if $u \in H^t(\Omega)$, $\|u\|_t \geq \|u\|_s$. If $u \in H^{s+1}(\Omega)$ then $\|\partial_i u\|_s \leq \|u\|_{s+1}$ $(i = 1,\ldots,N)$. We also have the interpolation inequality $\|u\|_t \leq \|u\|_s^{1-\theta} \|u\|_k^{\theta}$ for $0 \leq s \leq t \leq k$, $s \neq k$, and $\theta = (t-s)/(k-s)$.

For $s < 0$, define a norm on $C^\infty(\bar{\Omega})$ by

$$\|u\|_s = \sup_{\substack{\phi \in C^\infty(\bar{\Omega}) \\ \phi \neq 0}} \frac{(u,\phi)}{\|\phi\|_{-s}}$$

and let $H^s(\Omega)$ be the completion of $C^\infty(\bar{\Omega})$ in this norm. (This differs from the definition in [14] but is more suited to our purposes.) Our space $H^s(\Omega)$ for $s < 0$ is isomorphic to the space of all distributions in $H^s(R^N)$ with support in $\bar{\Omega}$, but we will not need this fact in the sequel. By the way we have defined the space $H^s(\Omega)$ we have the interpolation inequality $\|u\|_t \leq \|u\|_s^{1-\theta} \|u\|_k^{\theta}$ for $0 \geq s \geq t \geq k$, $s \neq k$, and $\theta = (t-s)/(k-s)$.

We will also need the Sobolev spaces on the boundary Γ of Ω. We now assume Γ is smooth, with outward normal vector n. Then for $s \geq 0$ a norm $|\cdot|_s$ and inner product $\langle\cdot,\cdot\rangle_s$ can be defined on $C^\infty(\Gamma)$ so that $H^s(\Gamma)$ is the completion of $C^\infty(\Gamma)$ in the norm $|\cdot|_s$. (For the details see [14].) For $s < 0$ we define $H^s(\Gamma)$ as the dual of

$H^{-s}(\Gamma)$, so we have

$$|\phi|_s = \sup_{\substack{\tau \in H^{-s}(\Gamma) \\ \tau \neq 0}} \frac{\langle \phi, \tau \rangle}{|\tau|_{-s}}$$

for $s < 0$, where we write $\langle \cdot, \cdot \rangle$ for $\langle \cdot, \cdot \rangle_0$.

The first result we wish to cite is a trace and extension theorem. Here and throughout the rest of this work C will denote a generic positive constant, not necessarily the same in any two different places. The proof of Theorem 2.1 may be found in [14].

Theorem 2.1.

(a) For each $s > 1/2$ there exists a bounded linear map $\gamma_0 : H^s(\Omega) \to H^{s-1/2}(\Gamma)$ (so $|\gamma_0 w|_{s-1/2} \leq C \|w\|_s$) with $\gamma_0 w = w|_\Gamma$ for all $w \in C^\infty(\overline{\Omega})$. Here C depends on s and Ω.

(b) For each $s > 3/2$ there exists a bounded linear map $\gamma_1 : H^s(\Omega) \to H^{s-3/2}(\Gamma)$ (so $|\gamma_1 w|_{s-3/2} \leq C \|w\|_s$) with $\gamma_1 w = \partial w / \partial n$ for all $w \in C^\infty(\overline{\Omega})$. Here C depends on s and Ω.

(c) For each $s > 0$ there exists a bounded linear map $\ell : H^s(\Gamma) \to H^{s+1/2}(\Omega)$ with $\gamma_0 \circ \ell$ = identity map on $H^s(\Gamma)$.

By a slight abuse of notation we will often write $\gamma_0 w$ as "$w|_\Gamma$" or "w on Γ" for $w \in H^s(\Omega)$ $(s > 1/2)$ and $\gamma_1 w$ as $\partial w / \partial n$ for $w \in H^s(\Omega)$ $(s > 3/2)$.

We next introduce the Sobolev spaces $H_0^s(\Omega)$. Let $C_0^\infty(\Omega) = \{\phi \in C^\infty(\Omega) \mid \text{supp}(\phi) \text{ is a compact subset of } \Omega\}$, where $\text{supp}(\phi) = \overline{\{x \mid \phi(x) \neq 0\}}$. For $s \geq 0$, define $H_0^s(\Omega)$ to be the closure of $C_0^\infty(\Omega)$ in $H^s(\Omega)$. In particular,

$H_0^1(\Omega)$ turns out to be the kernel of $\gamma_0: H^1(\Omega) \to H^{1/2}(\Gamma)$.

Finally, for any real s define the space $\underset{\sim}{H}^s(\Omega) = (H^s(\Omega))^N$ with the product norm and inner product. Functions in $\underset{\sim}{H}^s(\Omega)$ will be distinguished with an undertilde, $\underset{\sim}{v}$, and the norm and inner product will be denoted $\|\underset{\sim}{v}\|_s$ and $(\underset{\sim}{v},\underset{\sim}{\psi})_s$, respectively. In particular, if $u \in H^{s+2}(\Omega)$ where $s \geq 0$, then
$\underset{\sim}{v} = \nabla u = (\partial_1 u, \ldots, \partial_N u) \in \underset{\sim}{H}^{s+1}(\Omega)$ and
$\operatorname{div} \underset{\sim}{v} = \Sigma\, \partial_i v_i \in H^s(\Omega)$, and $\|\operatorname{div} \underset{\sim}{v}\|_s \leq C\, \|\underset{\sim}{v}\|_{s+1} \leq C\|u\|_{s+2}$.

The next result we cite is the divergence theorem extended to Sobolev spaces.

Lemma 2.2. For all $u \in H^1(\Omega)$ and $\underset{\sim}{\psi} \in \underset{\sim}{H}^1(\Omega)$,

$$(\nabla u, \underset{\sim}{\psi}) + (u, \operatorname{div} \underset{\sim}{\psi}) = \langle \gamma_0 u, n \cdot \underset{\sim}{\psi} \rangle,$$

where $n \cdot \underset{\sim}{\psi}$ is interpreted as $n_1\gamma_0\psi_1 + \ldots + n_N\gamma_0\psi_N$ ($n = (n_1, \ldots, n_N)$ is the normal to Γ, $\underset{\sim}{\psi} = (\psi_1, \ldots, \psi_N)$).

Proof. The formula holds for $u \in C^\infty(\bar{\Omega})$, $\underset{\sim}{\psi} \in (C^\infty(\bar{\Omega}))^N$ by the classical divergence theorem and extends by continuity to $u \in H^1(\Omega)$, $\underset{\sim}{\psi} \in \underset{\sim}{H}^1(\Omega)$.

We will often abuse the notation slightly and write the formula as simply $(\nabla u, \underset{\sim}{\psi}) + (u, \operatorname{div} \underset{\sim}{\psi}) = \langle u, n \cdot \underset{\sim}{\psi} \rangle$.

Finally we need one slightly nonstandard function space, seen in [4] and [5]. Define a norm on $C^\infty(\overline{\Omega})$ by

$$\|\phi\|^2_{\tilde{H}^{s,2}(\Omega)} = \|\phi\|^2_s + |\phi|^2_{s-1/2} + \left|\frac{\partial\phi}{\partial n}\right|^2_{s-3/2}$$

and let $\tilde{H}^{s,2}(\Omega)$ be the completion of $C^\infty(\overline{\Omega})$ in this norm. The utility of this space will become clear later. Right now we note it is easy to see there is a bounded linear map $N:\tilde{H}^{1,2}(\Omega) \to H^{-1/2}(\Gamma)$ such that $N\phi = \frac{\partial\phi}{\partial n}$ for all $\phi \in C^\infty(\overline{\Omega})$.

We will later need to estimate the boundary values of functions in Sobolev spaces by their values in the interior. To be precise, we have

Lemma 2.3. There exists a constant $C > 0$ such that for all $w \in H^1(\Omega)$ and all $\varepsilon > 0$,

$$|w|_0 \le \varepsilon \|w\|_1 + (C/\varepsilon)\|w\|_0.$$

Proof.

The proof is in the literature [1], [8], and we merely repeat it here. Let $\underset{\sim}{f}$ be a C^1 vector field on $\overline{\Omega}$ with $\underset{\sim}{f} = n$ on Γ. Then

$$\begin{aligned}
|w|^2_0 &= \langle w,w\rangle = \langle w, n\cdot(\underset{\sim}{f}w)\rangle \\
&= (w,\operatorname{div}(\underset{\sim}{f}w)) + (\nabla w, \underset{\sim}{f}w) \quad \text{by Lemma 2.2} \\
&= (w, w(\operatorname{div}\underset{\sim}{f})) + 2(\nabla w, w\underset{\sim}{f}) \\
&\le C\|w\|^2_0 + 2C\|w\|_0\|w\|_1 \\
&\le C\|w\|^2_0 + \varepsilon^2\|w\|^2_1 + (C/\varepsilon^2)\|w\|^2_0
\end{aligned}$$

using the inequality $2ab \leq \delta a^2 + \delta^{-1} b^2$,

valid for all $\delta > 0$

$\leq \varepsilon^2 \|w\|_1^2 + (C/\varepsilon^2) \|w\|_0^2$

$\leq (\varepsilon \|w\|_1 + (C/\varepsilon) \|w\|_0)^2$, completing the proof.

With the function space preliminaries settled, let us turn to the topic of differential operators. We immediately specialize to the case of major interest in the following, the case of second order elliptic partial differential operators. Let L be a second order partial differential operator given by

$$Lu = -\sum_{i=1}^{N} \partial_i \Big(\sum_{j=1}^{N} a_{ij}(x) \partial_j u \Big) + \sum_{i=1}^{N} a_i(x) \partial_i u + c(x) u$$

where the $a_{ij}(x)$, $a_i(x)$, $c(x)$ are smooth functions on $\overline{\Omega}$. We may (upon replacing a_{ij} by $(a_{ij} + a_{ji})/2$ and changing some lower-order terms if necessary) assume the matrix $A(x) = [a_{ij}(x)]$ is symmetric. We will then write L in the form

$$Lu = -\mathrm{div}(A \nabla u) + \underset{\sim}{b} \cdot \nabla u + cu$$

where $\underset{\sim}{b}(x)$, $c(x)$ are smooth functions on $\overline{\Omega}$.

Define the operator L to be elliptic if there exists a $c_0 > 0$ such that for any real numbers $\xi_1, \ldots, \xi_N$ and any $x \in \overline{\Omega}$, $\sum_{i,j} a_{ij}(x) \xi_i \xi_j \geq c_0 \sum_i \xi_i^2$. Thus the ellipticity condition asserts that the symmetric matrix A is (uniformly) positive definite on $\overline{\Omega}$.

The adjoint L^* of the operator L is defined by $(L\phi,\psi) = (\phi,L^*\psi)$ for all $\phi,\psi \in C_0^\infty(\Omega)$. In particular, for $Lu = -\mathrm{div}(A\nabla u) + \underset{\sim}{b}\cdot\nabla u + cu$ where A is symmetric, we have $L^*w = -\mathrm{div}(A\nabla w) - \mathrm{div}(\underset{\sim}{b}w) + cw$, and L^* is elliptic if L is.

We will be interested in solving problems of the form $Lu = f$ in Ω, $u = g$ on Γ, and we will now cite a result which will be one of our main tools. To avoid complications assume that the only solution to the problem $Lu = 0$ in Ω, $u = 0$ on Γ, is the zero solution, and similarly for the problem $L^*w = 0$ in Ω, $w = 0$ on Γ. Then we have

Theorem 2.4. Consider the problem $Lu = f$ in Ω, $u = g$ on Γ, where L is as above.

(a) If $f \in H^s(\Omega)$, $g \in H^{s+3/2}(\Gamma)$ $(s \geq 0)$, there exists a unique $u \in H^{s+2}(\Omega)$ solving the above problem, and $\|u\|_{s+2} \leq C\,(\|f\|_s + |g|_{s+3/2})$.

(b) If $f \in H^{-s}(\Omega)$ $(s \geq 0,\ s \neq \text{integer} + 1/2)$ and $g \in H^{-s+3/2}(\Gamma)$, there exists a unique $u \in \tilde{H}^{s+2,2}(\Omega)$ solving the above problem, and

$$\|u\|_{\tilde{H}^{s+2,2}(\Omega)} \leq C\,(\|f\|_s + |g|_{-s+3/2}).$$

(Here $Lu = f$ is understood as meaning $(Lu,\phi) = (f,\phi)$ for all $\phi \in H^s(\Omega)$, where a meaning is assigned to (Lu,ϕ) by integration by parts.)

The above theorem is a fundamental result in the theory of elliptic partial differential equations. Any of the references cited on p. 6 may be consulted for more details. We remark that analogous results hold for the problem $L^*w = f$ in Ω, $w = g$ on Γ.

This concludes Chapter II.

CHAPTER III. ERROR ESTIMATES

We begin this chapter by defining the finite-dimensional problem to be considered. As before, let L be a second order elliptic partial differential operator given by $Lu = -\mathrm{div}(A\nabla u) + \underset{\sim}{b}\cdot\nabla u + cu$ where the coefficients are smooth functions on Ω and A is a symmetric positive definite matrix on Ω. We assume the only solution of the problem $Lu = 0$ in Ω, $u = 0$ on Γ, is the zero solution, and similarly for L^*.

Now consider the boundary value problem

$$\begin{aligned} &Lu = f \text{ in } \Omega,\ f \in H^k(\Omega),\ k \geq 0 \\ &u = g \text{ on } \Gamma,\ g \in H^{\ell}(\Gamma),\ \ell \geq 3/2. \end{aligned} \tag{3.1}$$

We will need the basic a priori estimates for the operator L and its formal adjoint L^*, where L^* is given by $L^*w = -\mathrm{div}(A\nabla w) - \mathrm{div}(\underset{\sim}{b}w) + cw$. The estimates are

$$\begin{aligned} \|u\|_{k+2} &\leq C(\|Lu\|_k + |u|_{k+3/2}) \\ \|w\|_{k+2} &\leq C(\|L^*w\|_k + |w|_{k+3/2}). \end{aligned}$$

Now we introduce the finite-dimensional spaces in which we will work. For $0 < h \leq 1$ let $S_1^h = S_1^h(\Omega) \subset H^1(\Omega)$ and $S_2^h = S_2^h(\Omega) \subset H^1(\Omega)$ be finite-dimensional spaces, and define $\underset{\sim}{S}_2^h = \underset{\sim}{S}_2^h(\Omega) = (S_2^h)^N$. More will be said about these spaces later.

Since the matrix A is symmetric positive definite

it has a symmetric positive definite square root S, $S^2 = A$. We want to consider least-squares approximation methods for (3.1), and as indicated in Chapter I we wish to consider mixed methods. So consider the formulation of (3.1),

$$
\begin{aligned}
&-S\nabla u = \underset{\sim}{v} \quad \text{in } \Omega, \\
&\operatorname{div}(S\underset{\sim}{v}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{v} + cu = f \quad \text{in } \Omega, \\
&u = g \quad \text{on } \Gamma,
\end{aligned}
\tag{3.2}
$$

and now consider the variational problem

$$
\begin{aligned}
&\text{Find } (u^h, \underset{\sim}{v}^h) \in S_1^h \times \underset{\sim}{S}_2^h \text{ such that} \\
&J_h(u^h, \underset{\sim}{v}^h) = \inf_{(w^h, \underset{\sim}{z}^h) \in S_1^h \times \underset{\sim}{S}_2^h} J_h(w^h, \underset{\sim}{z}^h),
\end{aligned}
\tag{3.3}
$$

where $J_h(w,\underset{\sim}{z}) \equiv \|S\nabla w + \underset{\sim}{z}\|_0^2 + \|\operatorname{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw - f\|_0^2 + h^{-1}|w - g|_0^2$

for $(w,\underset{\sim}{z}) \in H^1(\Omega) \times \underset{\sim}{H}^1(\Omega)$.

Thus $(u^h, \underset{\sim}{v}^h)$ is the best least squares approximation to $(u,\underset{\sim}{v})$, the solution of (3.2), in $S_1^h \times \underset{\sim}{S}_2^h$ in the sense that

$$
J_h(u^h - u, \underset{\sim}{v}^h - \underset{\sim}{v}) = \inf_{(w^h, \underset{\sim}{z}^h) \in S_1^h \times \underset{\sim}{S}_2^h} J_h(w^h - u, \underset{\sim}{z}^h - \underset{\sim}{v}),
$$

which holds because $J_h(w,\underset{\sim}{z}) = J_h(w-u,\underset{\sim}{z}-\underset{\sim}{v})$ for all $(w,\underset{\sim}{z})$. A necessary condition for the quadratic functional J_h to be minimized is given by setting the first variation equal to zero. If we do so we are led to define a

bilinear form for $(w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi}) \in H^1(\Omega) \times \underset{\sim}{H}^1(\Omega)$ by

$$
\begin{aligned}
B_h((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) &= (S\nabla w + \underset{\sim}{z}, S\nabla\phi + \underset{\sim}{\psi}) \\
&+ (\operatorname{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw,\ \operatorname{div}(S\underset{\sim}{\psi}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{\psi} + c\phi) \\
&+ h^{-1}\langle w,\phi\rangle .
\end{aligned} \tag{3.4}
$$

Generally we will omit the subscript h and write simply B for B_h. It is easy to see that $(u^h,\underset{\sim}{v}^h)$ solves (3.3) if and only if

$$
\begin{aligned}
B((u^h,\underset{\sim}{v}^h),(\phi_h,\underset{\sim}{\psi}_h)) &= B((u,\underset{\sim}{v}),(\phi_h,\underset{\sim}{\psi}_h)) \\
&= (f,\ \operatorname{div}(S\underset{\sim}{\psi}_h) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{\psi}_h + c\phi_h) \\
&+ h^{-1}\langle g,\phi_h\rangle
\end{aligned} \tag{3.5}
$$

for all $(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$.

Once a basis for $S_1^h \times \underset{\sim}{S}_2^h$ has been chosen, (3.5) gives a system of linear equations to be solved to obtain $(u^h,\underset{\sim}{v}^h)$.

Define the B-norm of a pair $(w,\underset{\sim}{z})$ by

$$\|(w,\underset{\sim}{z})\|_B = B((w,\underset{\sim}{z}),(w,\underset{\sim}{z}))^{1/2}.$$

Then we have the Cauchy-Schwarz inequality for B,

$$|B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi}))| \le \|(w,\underset{\sim}{z})\|_B \, \|(\phi,\underset{\sim}{\psi})\|_B,$$

and we also have the following bounds for $\|(w,\underset{\sim}{z})\|_B$, where we use the inequality $x + y + z \le \sqrt{3}\,(x^2 + y^2 + z^2)^{1/2} \le \sqrt{3}\,(x + y + z)$, valid for all nonnegative x, y, z.

$$
\begin{aligned}
&\|S\nabla w + \underset{\sim}{z}\|_0 + \|\operatorname{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw\|_0 \\
&\qquad + h^{-1/2}|w|_0 \\
&\le \sqrt{3}\,\|(w,\underset{\sim}{z})\|_B \\
&\le \sqrt{3}\,(\,\|S\nabla w + \underset{\sim}{z}\|_0 + \|\operatorname{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw\|_0 \\
&\qquad + h^{-1/2}|w|_0)\,.
\end{aligned}
\tag{3.6}
$$

Now, using Lemma 2.3 with $\varepsilon = h^{1/2}$ we have

$$
\begin{aligned}
h^{-1/2}|w|_0 &\le h^{-1/2}(Ch^{-1/2}\,\|w\|_0 + h^{1/2}\,\|w\|_1) \\
&= Ch^{-1}\,\|w\|_0 + \|w\|_1\,.
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\|(w,\underset{\sim}{z})\|_B &\le C(\,\|S\nabla w + \underset{\sim}{z}\|_0 \\
&\quad + \|\operatorname{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw\|_0 \\
&\quad + h^{-1}\|w\|_0 + \|w\|_1) \\
&\le C(h^{-1}\,\|w\|_0 + \|w\|_1 + \|\underset{\sim}{z}\|_1)\,.
\end{aligned}
\tag{3.7}
$$

Proposition 3.1. There exists a unique $(u^h,\underset{\sim}{v}^h) \in S_1^h \times \underset{\sim}{S}_2^h$ which solves (3.5), and

$$
\begin{aligned}
\|(u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h)\|_B &= \inf_{(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h\times\underset{\sim}{S}_2^h} \|(u-\phi_h,\underset{\sim}{v}-\underset{\sim}{\psi}_h)\|_B \\
&\le C \inf_{(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h\times\underset{\sim}{S}_2^h} (\,h^{-1}\,\|u-\phi_h\|_0 + \|u-\phi_h\|_1 \\
&\qquad + \|\underset{\sim}{v}-\underset{\sim}{\psi}_h\|_1)\,.
\end{aligned}
\tag{3.8}
$$

Proof. The existence follows because $(u^h,\underset{\sim}{v}^h)$ is the

projection of $(u,\underset{\sim}{v})$ on $S_1^h \times \underset{\sim}{S}_2^h$ in the B-inner product. For the uniqueness, suppose $(u^h,\underset{\sim}{v}^h)$ and $(\hat{u}^h,\underset{\sim}{\hat{v}}^h)$ both solve (3.5). Then

$$B((u^h-\hat{u}^h,\underset{\sim}{v}^h-\underset{\sim}{\hat{v}}^h),(\phi_h,\underset{\sim}{\psi}_h)) = 0 \text{ for all } (\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h,$$

and in particular $\|(u^h-\hat{u}^h,\underset{\sim}{v}^h-\underset{\sim}{\hat{v}}^h)\|_B^2 = 0$.

Now suppose $\|(w,\underset{\sim}{z})\|_B^2 = 0$. Then we have

$$S\nabla w + \underset{\sim}{z} = 0 \quad \text{in } \Omega,$$
$$\operatorname{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw = 0 \quad \text{in } \Omega, \text{ and}$$
$$w = 0 \text{ on } \Gamma.$$

The first two conditions yield $Lw = 0$ in Ω, and this combined with $w = 0$ on Γ implies $w = 0$. Hence $\underset{\sim}{z} = -S\nabla w = 0$, and we have shown that $\|(w,\underset{\sim}{z})\|_B = 0$ implies $w = 0$ and $\underset{\sim}{z} = 0$. Thus our two solutions $(u^h,\underset{\sim}{v}^h)$ and $(\hat{u}^h,\underset{\sim}{\hat{v}}^h)$ satisfy $u^h = \hat{u}^h$ and $\underset{\sim}{v}^h = \underset{\sim}{\hat{v}}^h$, so we have uniqueness.

The inequality follows from

$$\|(u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h)\|_B^2 = B((u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h),(u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h))$$
$$= B((u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h),(u-\phi_h,\underset{\sim}{v}-\underset{\sim}{\psi}_h))$$
$$\text{for all } (\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$$
$$\leq \|(u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h)\|_B \, \|(u-\phi_h,\underset{\sim}{v}-\underset{\sim}{\psi}_h)\|_B.$$

Now using the inequality (3.7) and taking the infimum over all $(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$ completes the proof.

Proposition 3.1 gives a bound for the error $(u-u^h, \underset{\sim}{v}-\underset{\sim}{v}^h)$, but the norm involved is somewhat unsatisfying. What we would like are error bounds for the quantities $u-u^h$ and $\underset{\sim}{v}-\underset{\sim}{v}^h$ separately in the usual Sobolev norms. Obtaining such bounds is the next order of business. First we introduce some approximation-theoretic hypotheses on the spaces S_1^h and S_2^h.

Approximation Hypothesis

There are integers $r_1, r_2 \geq 2$ such that if $w \in H^s(\Omega)$, $1 \leq s \leq r_i$, then there exists a $w_h^i \in S_i^h$ with

$$\|w-w_h^i\|_0 + h\,\|w-w_h^i\|_1 \leq Ch^s\,\|w\|_s \tag{3.9}$$

where $C > 0$ is independent of w and h.

We remark here that the hypothesis (3.9) coupled with (3.8) gives us the error bound

$$\|(u-u^h, \underset{\sim}{v}-\underset{\sim}{v}^h)\|_B \leq C(\,h^{s-1}\,\|u\|_s + h^{t-1}\,\|\underset{\sim}{v}\|_t)$$

if $u \in H^s(\Omega)$, $\underset{\sim}{v} \in \underset{\sim}{H}^t(\Omega)$, $1 \leq s \leq r_1$, $1 \leq t \leq r_2$. In particular, if $u \in H^{r_1}(\Omega)$ and $\underset{\sim}{v} \in \underset{\sim}{H}^{r_2}(\Omega)$, then

$$\|(u-u^h, \underset{\sim}{v}-\underset{\sim}{v}^h)\|_B \leq C(\,h^{r_1-1}\,\|u\|_{r_1} + h^{r_2-1}\,\|\underset{\sim}{v}\|_{r_2}).$$

Much effort has gone into the study of the approximation properties of finite element spaces. Some of

the results can be described as follows. For simplicity think of the two-dimensional case, $\Omega \subset R^2$. Let $\Omega \subset [a,b] \times [c,d]$, and let S^h consist of all bivariate Hermite polynomials of degree $2m - 1$ ($m \geq 1$) defined on a rectangular mesh covering $[a,b] \times [c,d]$. Let h = maximum side length of all rectangles, $\underline{h}$ = minimum side length. Assume $\underline{h} \geq ch$ for some fixed $c > 0$ as $h \to 0$. Let S_i^h consist of the restrictions of elements of S^h to Ω. Then S_i^h satisfies (3.9) with $r_i = 2m$ [23]. For another example, let S^h consist of all bivariate splines of degree $2m - 1$ defined on the same type of mesh and let S_i^h again consist of the restrictions of elements of S^h to Ω. Then (3.9) is satisfied with $r_i = 2m$. As a last example, suppose $\{T_n(h)\}$ is a collection of triangles for each $h > 0$ such that $\bigcup_n T_n(h) \supset \Omega$, any two triangles which intersect have an entire edge or only a vertex in common, the maximum side length of all the triangles $T_n(h)$ is h and the minimum angle of all the triangles is bounded below as $h \to 0$. Let S^h consist of all functions which are continuous on $\bigcup T_n(h)$ and linear on each triangle. Let S_i^h consist of the restriction of elements of S^h to Ω. Then (3.9) holds with $r_i = 2$. For further information on this topic, see [22], [23], [3], [9], or [11].

It is known that for piecewise polynomials of

degree $r_i - 1$ the exponent of h, r_i, is sharp in that it cannot be increased without invalidating (3.9). Thus if we are able to prove that our approximate solution u^h satisfies $\|u^h-u\|_0 = O(h^{r_1})$ we will have shown that (asymptotically) u^h is as good an approximation to u as there is in the space S_1^h.

The basic goal now is to derive estimates for the errors $e_u \equiv u-u^h$ and $\underset{\sim}{e}_v \equiv \underset{\sim}{v}-\underset{\sim}{v}^h$. For the sake of clarity, in what follows we will first give the complete proof for the case when $Lu = -\Delta u$, and then indicate what changes must be made to carry the proof over to the general case. We will constantly use (3.5) in the form $B((e_u,\underset{\sim}{e}_v),(\phi_h,\underset{\sim}{\psi}_h)) = 0$ for all $(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$.

Lemma 3.2. $|e_u|_{-1/2} \leq Ch \|(e_u,\underset{\sim}{e}_v)\|_B$.

Proof. First consider the case $L = -\Delta$. Then the decomposition is $-\nabla u = \underset{\sim}{v}$, div $\underset{\sim}{v} = f$ and the bilinear form is

$$B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) = (\nabla w + \underset{\sim}{z},\nabla\phi + \underset{\sim}{\psi}) + (\text{div } \underset{\sim}{z},\text{div } \underset{\sim}{\psi}) + h^{-1} \langle w,\phi\rangle.$$

By definition, $|e_u|_{-1/2} = \sup_{\substack{\tau \in H^{1/2}(\Gamma) \\ \tau \neq 0}} \frac{\langle e_u,\tau\rangle}{|\tau|_{1/2}}$.

Given $\tau \in H^{1/2}(\Gamma)$, define α as $\ell(\tau)$ (see Theorem 2.1), so $\alpha \in H^1(\Omega)$, $\|\alpha\|_1 \leq C\,|\tau|_{1/2}$, and $\gamma_0\alpha = \tau$. Then

$$\langle e_u, \tau \rangle = h\left(h^{-1} \langle e_u, \alpha \rangle \right)$$

$$= h\left(h^{-1} \langle e_u, \alpha - \alpha_h \rangle - (\nabla e_u + \underset{\sim}{e}_v, \nabla \alpha_h) \right)$$

for any $\alpha_h \in S_1^h$ by the orthogonality condition with $\underset{\sim}{\psi}_h = 0$

$$\leq h\left(h^{-1/2} |e_u|_0 \; h^{-1/2} |\alpha - \alpha_h|_0 + \|\nabla e_u + \underset{\sim}{e}_v\|_0 \, \|\nabla \alpha_h\|_0 \right)$$

$$\leq Ch \, \|(e_u, \underset{\sim}{e}_v)\|_B \left(h^{-1/2} |\alpha - \alpha_h|_0 + \|\nabla \alpha_h\|_0 \right)$$

$$\leq Ch \, \|(e_u, \underset{\sim}{e}_v)\|_B \left(C' h^{-1} \|\alpha - \alpha_h\|_0 + \|\alpha - \alpha_h\|_1 + \|\alpha_h\|_1 \right)$$

$$\leq Ch \|(e_u, \underset{\sim}{e}_v)\|_B \left(h^{-1} \|\alpha - \alpha_h\|_0 + \|\alpha - \alpha_h\|_1 + \|\alpha\|_1 \right)$$

$$\leq Ch \, \|(e_u, \underset{\sim}{e}_v)\|_B \left(C' \|\alpha\|_1 \right)$$

upon taking the infimum over all $\alpha_h \in S_1^h$ and using (3.9)

$$\leq Ch \, \|(e_u, \underset{\sim}{e}_v)\|_B \, |\tau|_{1/2}, \text{ as desired.}$$

In the general case we again define α as $\ell(\tau)$. Then

$$\langle e_u, \tau \rangle = h \left(h^{-1} \langle e_u, \alpha \rangle \right)$$

$$= h \left(h^{-1} \langle e_u, \alpha - \alpha_h \rangle - (S\nabla e_u + \underset{\sim}{e}_v, S\nabla \alpha_h) \right.$$

$$\left. - (\operatorname{div}(S\underset{\sim}{e}_v) - \underset{\sim}{b} \cdot S^{-1} \underset{\sim}{e}_v + c e_u, c\alpha_h) \right)$$

$$\leq Ch \, \|(e_u, \underset{\sim}{e}_v)\|_B \left(h^{-1/2} |\alpha - \alpha_h|_0 + \|S\nabla \alpha_h\|_0 + \|c\alpha_h\|_0 \right)$$

$$\leq Ch \, \| (e_u, \underset{\sim}{e}_v) \|_B \, (h^{-1} \, \| \alpha - \alpha_h \|_0 + \| \alpha - \alpha_h \|_1 + \| \alpha_h \|_1)$$

and the proof proceeds as before.

Corollary 3.3. For $0 \leq s \leq 1/2$,

$$|e_u|_{s-1/2} \leq Ch^{1-s} \, \| (e_u, \underset{\sim}{e}_v) \|_B .$$

Proof. We have $|e_u|_0 \leq Ch^{1/2} \, \| (e_u, \underset{\sim}{e}_v) \|_B$ and $|e_u|_{-1/2} \leq Ch \, \| (e_u, \underset{\sim}{e}_v) \|_B$. The proof follows by interpolation theory. Let $\theta = 1 - 2s$; then

$$\begin{aligned} |e_u|_{s-1/2} &\leq |e_u|^{\theta}_{-1/2} \, |e_u|^{1-\theta}_0 \\ &\leq C \, h^{\theta} h^{(1-\theta)/2} \, \| (e_u, \underset{\sim}{e}_v) \|_B \\ &= C \, h^{(1+\theta)/2} \, \| (e_u, \underset{\sim}{e}_v) \|_B \\ &= C \, h^{1-s} \, \| (e_u, \underset{\sim}{e}_v) \|_B . \end{aligned}$$

Lemma 3.4. For $0 \leq t \leq \min(r_1 - 1, r_2 - 1)$,

$$\| \operatorname{div}(S \underset{\sim}{e}_v) - \underset{\sim}{b} \cdot S^{-1} \underset{\sim}{e}_v + c e_u \|_{-t} \leq Ch^t \, \| (e_u, \underset{\sim}{e}_v) \|_B .$$

Proof. First consider the case $L = -\Delta$. Then the left-hand side is just

$$\| \operatorname{div} \underset{\sim}{e}_v \|_{-t} = \sup_{\substack{\phi \in H^t(\Omega) \\ \phi \neq 0}} \frac{(\operatorname{div} \underset{\sim}{e}_v, \phi)}{\| \phi \|_t} .$$

Given $\phi \in H^t(\Omega)$, let α solve $-\Delta \alpha = \phi$ in Ω, $\alpha = 0$ on Γ,

and let $\underset{\sim}{\beta} = -\nabla\alpha$. Then $\|\alpha\|_{t+2} + \|\underset{\sim}{\beta}\|_{t+1} \leq C\,\|\phi\|_t$. Now,

$$(\operatorname{div} \underset{\sim}{e}_v, \phi) = (\operatorname{div} \underset{\sim}{e}_v, \operatorname{div} \underset{\sim}{\beta})$$

$$= B((e_u, \underset{\sim}{e}_v), (\alpha, \underset{\sim}{\beta}))$$

$$= B((e_u, \underset{\sim}{e}_v), (\alpha-\alpha_h, \underset{\sim}{\beta}-\underset{\sim}{\beta}_h)) \qquad \text{for all } (\alpha_h, \underset{\sim}{\beta}_h) \in S_1^h \times \underset{\sim}{S}_2^h$$

$$\leq \|(e_u, \underset{\sim}{e}_v)\|_B \, \|(\alpha-\alpha_h, \underset{\sim}{\beta}-\underset{\sim}{\beta}_h)\|_B$$

$$\leq \|(e_u, \underset{\sim}{e}_v)\|_B \; C \; (h^{-1} \|\alpha-\alpha_h\|_0 + \|\alpha-\alpha_h\|_1 + \|\underset{\sim}{\beta}-\underset{\sim}{\beta}_h\|_1) \quad \text{by (3.7)}$$

$$\leq \|(e_u, \underset{\sim}{e}_v)\|_B \; C \; (h^t \|\alpha\|_{t+1} + h^t \|\underset{\sim}{\beta}\|_{t+1})$$

by taking the infimum over all $(\alpha_h, \underset{\sim}{\beta}_h)$ and using (3.9)

$$\leq C\, h^t \, \|(e_u, \underset{\sim}{e}_v)\|_B \, \|\phi\|_t \text{ , as desired.}$$

For the general case, solve $L\alpha = \phi$ in Ω, $\alpha = 0$ on Γ, and put $\underset{\sim}{\beta} = -S\nabla\alpha$. Then

$\|\underset{\sim}{\beta}\|_{t+1} + \|\alpha\|_{t+2} \leq C\,\|\phi\|_t$, and

$$(\operatorname{div}(S\underset{\sim}{e}_v) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{e}_v + c e_u, \phi)$$

$$= (\operatorname{div}(S\underset{\sim}{e}_v) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{e}_v + c e_u, \operatorname{div}(S\underset{\sim}{\beta}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{\beta} + c\alpha)$$

$$= B((e_u, \underset{\sim}{e}_v), (\alpha, \underset{\sim}{\beta})),$$

and the rest of the argument is unchanged.

Theorem 3.5. Let $(u^h, \underset{\sim}{v}^h)$ be the solution of (3.5), and let the approximation hypothesis (3.9) hold. Then for $0 \leq s < 1/2$,

$$\|e_u\|_s \leq C\, h^{1-s}\, \|(e_u, \underset{\sim}{e}_v)\|_B .$$

Proof. The idea is to use the duality argument of Nitsche [16]. We have

$$\|e_u\|_s = \sup_{\substack{\phi \in H^{-s}(\Omega) \\ \phi \neq 0}} \frac{(e_u, \phi)}{\|\phi\|_{-s}} .$$

First consider the case $L = -\Delta$. Given $\phi \in H^{-s}(\Omega)$, let α solve $-\Delta\alpha = \phi$ in Ω, $\alpha = 0$ on Γ, and put $\underset{\sim}{\beta} = -\nabla\alpha$. Then $\|\underset{\sim}{\beta}\|_{1-s} + \|\alpha\|_{2-s} \leq C\, \|\phi\|_{-s}$. Now,

$$\begin{aligned}
(e_u, \phi) &= (e_u, -\Delta\alpha) \\
&= (\nabla e_u, \nabla\alpha) - \left\langle e_u, \frac{\partial\alpha}{\partial n} \right\rangle \\
&= (\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha) - (\underset{\sim}{e}_v, \nabla\alpha) - \left\langle e_u, \frac{\partial\alpha}{\partial n} \right\rangle \\
&= (\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha) + (\operatorname{div} \underset{\sim}{e}_v, \alpha) - \left\langle e_u, \frac{\partial\alpha}{\partial n} \right\rangle \\
&\qquad \text{since } \alpha = 0 \text{ on } \Gamma.
\end{aligned}$$

These terms can be bounded as follows. We have

$$\begin{aligned}
\left| \left\langle e_u, \frac{\partial\alpha}{\partial n} \right\rangle \right| &\leq |e_u|_{s-1/2} \left| \frac{\partial\alpha}{\partial n} \right|_{-s+1/2} \\
&\leq C\, h^{1-s}\, \|(e_u, \underset{\sim}{e}_v)\|_B\, \|\alpha\|_{2-s} \\
&\qquad \text{by Corollary 3.3 and Theorem 2.1;}
\end{aligned}$$

$$|(\operatorname{div} \underset{\sim}{e}_v, \alpha)| \leq \|\operatorname{div} \underset{\sim}{e}_v\|_{L_{1+s}} \|\alpha\|_{1-s}$$

$$\leq C\, h^{1-s} \|(e_u, \underset{\sim}{e}_v)\|_B \|\alpha\|_{2-s}$$

by Lemma 3.4.

For the first term,

$$(\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha) = (\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha - \nabla\alpha_h) + h^{-1} \langle e_u, \alpha - \alpha_h \rangle$$

for all $\alpha_h \in S_1^h$ by orthogonality and the fact that $\alpha = 0$ on Γ

$$\leq C \|(e_u, \underset{\sim}{e}_v)\|_B \left(\|\nabla(\alpha - \alpha_h)\|_0 + h^{-1} \|\alpha - \alpha_h\|_0 + \|\alpha - \alpha_h\|_1 \right)$$

$$\leq C \|(e_u, \underset{\sim}{e}_v)\|_B\, h^{1-s} \|\alpha\|_{2-s}$$

upon taking the infimum over α_h and using (3.9).

Collecting the pieces yields

$$|(e_u, \phi)| \leq C\, h^{1-s} \|(e_u, \underset{\sim}{e}_v)\|_B \|\alpha\|_{2-s}$$

$$\leq C\, h^{1-s} \|(e_u, \underset{\sim}{e}_v)\|_B \|\phi\|_{L_s}, \text{ as desired.}$$

For the general case, we solve $L^*\alpha = \phi$ in Ω, $\alpha = 0$ on Γ, and note that $\|\alpha\|_{2-s} \leq C \|\phi\|_{L_s}$ and further that

$$L^*\alpha = -\operatorname{div}(A\nabla\alpha) - \operatorname{div}(\underset{\sim}{b}\alpha) + c\alpha$$

$$= -\operatorname{div}(A\nabla\alpha) - \underset{\sim}{b}\cdot\nabla\alpha - \alpha \operatorname{div} \underset{\sim}{b} + c\alpha.$$

Let $\underset{\sim}{\beta} = -S\nabla\alpha$, so $\operatorname{div}(S\underset{\sim}{\beta}) + \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{\beta} + \alpha(c - \operatorname{div} \underset{\sim}{b}) = \phi$.

Then

$$(e_u,\phi) = (e_u,L^*\alpha)$$

$$= (e_u,-\mathrm{div}(A\nabla\alpha) - \mathrm{div}(\underset{\sim}{b}\alpha) + c\alpha)$$

$$= (\nabla e_u,\ A\nabla\alpha) - \langle e_u,\ N(\alpha)\rangle - (e_u,\mathrm{div}(\underset{\sim}{b}\alpha) - c\alpha)$$

where N is an operator satisfying

$$|N(\alpha)|_{k-3/2} \leq C\,\|\phi\|_k \quad \text{for } k > 3/2$$

$$= (S\nabla e_u, S\nabla\alpha) - \langle e_u,N(\alpha)\rangle - (e_u,\mathrm{div}(\underset{\sim}{b}\alpha) - c\alpha)$$

$$= (S\nabla e_u + \underset{\sim}{e}_v, S\nabla\alpha) - (\underset{\sim}{e}_v, S\nabla\alpha) - \langle e_u,N(\alpha)\rangle - (e_u,\mathrm{div}(\underset{\sim}{b}\alpha) - c\alpha)$$

$$= (S\nabla e_u + \underset{\sim}{e}_v,\ S\nabla\alpha) + (\mathrm{div}(S\underset{\sim}{e}_v),\ \alpha) - \langle e_u,N(\alpha)\rangle - (e_u,\mathrm{div}(\underset{\sim}{b}\alpha) - c\alpha)$$

$$= (S\nabla e_u + \underset{\sim}{e}_v,\ S\nabla\alpha) - \langle e_u,N(\alpha)\rangle + (\mathrm{div}(S\underset{\sim}{e}_v) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{e}_v + ce_u,\alpha) + (\underset{\sim}{b}\cdot S^{-1}\underset{\sim}{e}_v - ce_u,\alpha) - (e_u,\mathrm{div}(\underset{\sim}{b}\alpha) - c\alpha)$$

$$= (S\nabla e_u + \underset{\sim}{e}_v,\ S\nabla\alpha) - \langle e_u,N(\alpha)\rangle + (\mathrm{div}(S\underset{\sim}{e}_v) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{e}_v + ce_u,\alpha) + (S^{-1}\underset{\sim}{e}_v + \nabla e_u,\ \underset{\sim}{b}\alpha).$$

The first three terms here can be bounded as in the previous argument. The fourth term can be written as $(S\nabla e_u + \underset{\sim}{e}_v, S^{-1}(\alpha\underset{\sim}{b}))$. This is handled by the following lemma. Collecting all the results finishes the proof,

Lemma 3.6. For $1 \le r \le \min(r_1-1, r_2-1)$,

$$\|S\nabla e_u + \underset{\sim}{e}_v\|_{-r} \le C\, h^{r-1}\, \|(e_u, \underset{\sim}{e}_v)\|_B .$$

Proof. We have

$$\|S\nabla e_u + \underset{\sim}{e}_v\|_{-r} = \sup_{\substack{\underset{\sim}{\psi} \in \underset{\sim}{H}^r(\Omega) \\ \underset{\sim}{\psi} \neq 0}} \frac{(S\nabla e_u + \underset{\sim}{e}_v, \underset{\sim}{\psi})}{\|\underset{\sim}{\psi}\|_r} .$$

Given $\underset{\sim}{\psi} \in \underset{\sim}{H}^r(\Omega)$, let w solve $Lw = -\mathrm{div}(S\underset{\sim}{\psi}) + \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{\psi}$ in Ω, $w = 0$ on Γ. Then $\|w\|_{r+1} \le C\,\|\underset{\sim}{\psi}\|_r$. Now let $\underset{\sim}{z} = \underset{\sim}{\psi} - S\nabla w$, so $\mathrm{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw = 0$ and $\|\underset{\sim}{z}\|_r \le C\,\|\underset{\sim}{\psi}\|_r$. Then

$$(S\nabla e_u + \underset{\sim}{e}_v, \underset{\sim}{\psi}) = (S\nabla e_u + \underset{\sim}{e}_v, S\nabla w + \underset{\sim}{z})$$

$$= B((e_u, \underset{\sim}{e}_v), (w, \underset{\sim}{z}))$$

$$= B((e_u, \underset{\sim}{e}_v), (w-w_h, \underset{\sim}{z}-\underset{\sim}{z}_h)) \quad \text{for all } (w_h, \underset{\sim}{z}_h)$$

$$\le \|(e_u, \underset{\sim}{e}_v)\|_B\, \|(w-w_h, \underset{\sim}{z}-\underset{\sim}{z}_h)\|_B$$

$$\le \|(e_u, \underset{\sim}{e}_v)\|_B\, C\, (\|w-w_h\|_1 + h^{-1}\|w-w_h\|_0 + \|\underset{\sim}{z}-\underset{\sim}{z}_h\|_1) \quad \text{by (3.7)}$$

$$\le \|(e_u, \underset{\sim}{e}_v)\|_B\, C\, (h^{r-1}\|w\|_r + h^{r-1}\|\underset{\sim}{z}\|_r)$$

by taking the infimum over all $(w_h, \underset{\sim}{z}_h)$ and using (3.9)

$$\le \|(e_u, \underset{\sim}{e}_v)\|_B\, C\, h^{r-1}\, \|\underset{\sim}{\psi}\|_r, \text{ completing the proof.}$$

Some remarks on this theorem are in order. Suppose that $u \in H^r(\Omega)$, $\underset{\sim}{v} \in \underset{\sim}{H}^r(\Omega)$, and $r_1 = r_2 = r$. Then $\| (e_u, \underset{\sim}{e}_v) \|_B \leq C\, h^{r-1} (\|u\|_r + \|\underset{\sim}{v}\|_r)$, so Theorem 3.5 implies $\|e_u\|_s \leq C\, h^{r-s} (\|u\|_r + \|\underset{\sim}{v}\|_r)$ for $0 \leq s < 1/2$. Thus u^h is asymptotically as good an approximation to u in the H^s norm as exists in S_1^h.

Also, the range $0 \leq s < 1/2$ has been seen in the literature in error bounds for least squares methods [7], [4]. We are unfortunately unable to include the case $s = 1/2$ because of our lack of a trace inequality for $\gamma_1 : H^{3/2}(\Omega) \to H^0(\Omega)$; in fact the inequality $|\gamma_1 \alpha|_0 \leq C \|\alpha\|_{3/2}$ is false in general [14].

We now turn our attention to the decomposition (3.2) itself. From a practical point of view this may be an unsatisfactory system to consider, because if the matrix A is not constant or diagonal then S could be difficult to compute. When one sets up the matrix formulation (3.5) of the problem one has to choose bases for the spaces S_1^h, $\underset{\sim}{S}_2^h$ and then compute various inner products of the basis elements, for example $(S\nabla\phi_i + \underset{\sim}{\psi}_j,\ S\nabla\phi_k + \underset{\sim}{\psi}_m)$ where $\phi_i, \phi_k \in S_1^h$, $\underset{\sim}{\psi}_j, \underset{\sim}{\psi}_m \in \underset{\sim}{S}_2^h$, and thus the matrix S complicates the formation of these inner products. It seems natural (because we directly obtain an approximation to the quantity $A\nabla u$, which in physical problems corresponds to the flux) to consider the decomposition

$$
\begin{aligned}
-A\nabla u &= \underset{\sim}{v} \quad \text{in } \Omega \\
\text{div } \underset{\sim}{v} - \underset{\sim}{b}\cdot A^{-1}\underset{\sim}{v} + cu &= f \quad \text{in } \Omega \\
u &= 0 \quad \text{on } \Gamma,
\end{aligned} \tag{3.10}
$$

and the associated bilinear form, which we again denote B:

$$
\begin{aligned}
B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) &= (A\nabla w + \underset{\sim}{z},\ A\nabla\phi + \underset{\sim}{\psi}) \\
&+ (\text{div } \underset{\sim}{z} - \underset{\sim}{b}\cdot A^{-1}\underset{\sim}{z} + cw,\ \text{div } \underset{\sim}{\psi} - \underset{\sim}{b}\cdot A^{-1}\underset{\sim}{\psi} + c\phi) \\
&+ h^{-1} < w,\phi > .
\end{aligned} \tag{3.11}
$$

If we try to work with this bilinear form we find that most of the preceding arguments carry through. In fact, Proposition 3.1, Lemmas 3.2 and 3.4, and Corollary 3.3 may be proved as before. Unfortunately we are unable to prove Theorem 3.5 in this case, as the proper terms do not appear when we integrate by parts (the first chain of equalities in the proof of Theorem 3.5). It remains unclear whether or not we can give quasi-optimal error bounds for $\|e_u\|_s$ in this case.

There is one important special case where we can find quasi-optimal error bounds for $\|e_u\|_s$. This is the case when A is a scalar matrix, $A = a(x)I$, where $a(x)$ is a smooth function satisfying $a(x) \geq a_0 > 0$ on $\overline{\Omega}$. So let us consider the decomposition (3.10) and the bilinear form (3.11) in this case.

Theorem 3.7. With $A = a(x)I$, $a(x) \geq a_0 > 0$ in Ω, consider the decomposition (3.10) and bilinear form (3.11), and let $(u^h, \underset{\sim}{v}^h) \in S_1^h \times \underset{\sim}{S}_2^h$ solve the problem $B((u^h, \underset{\sim}{v}^h), (\phi_h, \underset{\sim}{\psi}_h)) = B((u, \underset{\sim}{v}), (\phi_h, \underset{\sim}{\psi}_h))$ for all $(\phi_h, \underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$. Then for $0 \leq s < 1/2$, $\|e_u\|_s \leq C\, h^{1-s}\, \|(e_u, \underset{\sim}{e}_v)\|_B$.

Proof. As remarked above, Proposition 3.1, Lemmas 3.2 and 3.4, and Corollary 3.3 hold in this case, and their proofs are unchanged. We begin as in the proof of Theorem 3.5. Given $\phi \in H^{-s}(\Omega)$, $0 \leq s < 1/2$, let w solve $L^*w = \phi$ in Ω, $w = 0$ on Γ, i.e., $-\mathrm{div}(a\nabla w) - \mathrm{div}(\underset{\sim}{b}w) + cw = \phi$ in Ω, $w = 0$ on Γ. Then $\|w\|_{2-s} \leq C\, \|\phi\|_{-s}$. We have

$$
\begin{aligned}
(e_u, \phi) &= (e_u,\ -\mathrm{div}(a\nabla w) - \mathrm{div}(\underset{\sim}{b}w) + cw) \\
&= (\nabla e_u,\ a\nabla w) - \langle e_u, n\cdot(a\nabla w)\rangle + (\nabla e_u,\ \underset{\sim}{b}w) + (e_u,\ cw) \\
&= (a\nabla e_u + \underset{\sim}{e}_v,\ \nabla w) + (\mathrm{div}\ \underset{\sim}{e}_v, w) - \langle e_u, n\cdot a\nabla w\rangle \\
&\quad + (\nabla e_u,\ \underset{\sim}{b}w) + (ce_u,\ w) \\
&= (a\nabla e_u + \underset{\sim}{e}_v,\ a\nabla(a^{-1}w)) - (a\nabla e_u + \underset{\sim}{e}_v,\ aw\ \nabla a^{-1}) \\
&\quad - \langle e_u, n\cdot a\nabla w\rangle + (\mathrm{div}\ \underset{\sim}{e}_v - \underset{\sim}{b}\cdot a^{-1}\underset{\sim}{e}_v + ce_u,\ w) \\
&\quad + (\underset{\sim}{b}\cdot a^{-1}\underset{\sim}{e}_v,\ w) + (\nabla e_u,\ \underset{\sim}{b}w) \\
&= (a\nabla e_u + \underset{\sim}{e}_v,\ a\nabla(a^{-1}w)) + (a\nabla e_u + \underset{\sim}{e}_v,\ a^{-1}w\nabla a) \\
&\quad - \langle e_u, n\cdot a\nabla w\rangle + (\mathrm{div}\ \underset{\sim}{e}_v - \underset{\sim}{b}\cdot a^{-1}\underset{\sim}{e}_v + ce_u,\ w) \\
&\quad + (a\nabla e_u + \underset{\sim}{e}_v,\ a^{-1}\underset{\sim}{b}w)
\end{aligned}
$$

$$= (a\nabla e_u + \underset{\sim}{e}_v,\ a\nabla(a^{-1}w)) + (a\nabla e_u + \underset{\sim}{e}_v,\ a^{-1}w(\underset{\sim}{b} + \nabla a))$$
$$- \langle e_u,\ n\cdot a\nabla w\rangle + (\operatorname{div} \underset{\sim}{e}_v - \underset{\sim}{b}\cdot a^{-1}\underset{\sim}{e}_v + ce_u,\ w).$$

By Lemma 3.6, the second term here is bounded by $C\ h^{1-s}\ \|(e_u,\underset{\sim}{e}_v)\|_B\ \|w\|_{2-s}$. By Lemma 3.4, the fourth term is bounded by $C\ h^{1-s}\ \|(e_u,\underset{\sim}{e}_v)\|_B\ \|w\|_{1-s}$ which is in turn bounded by $C\ h^{1-s}\ \|(e_u,\underset{\sim}{e}_v)\|_B\ \|w\|_{2-s}$. Corollary 3.3 bounds the third term by $C\ h^{1-s}\ \|(e_u,\underset{\sim}{e}_v)\|_B\ \|w\|_{2-s}$. For the first term, write

$$(a\nabla e_u + \underset{\sim}{e}_v,\ a\nabla(a^{-1}w))$$

$$= (a\nabla e_u + \underset{\sim}{e}_v,\ a\nabla(a^{-1}w - \phi_h))$$
$$- (\operatorname{div} \underset{\sim}{e}_v - \underset{\sim}{b}\cdot a^{-1}\underset{\sim}{e}_v + ce_u, c\phi_h) \qquad (3.12)$$
$$- h^{-1}\ \langle e_u, a^{-1}w - \phi_h\rangle\ ,$$

where we used the orthogonality relation plus the vanishing of $a^{-1}w$ on Γ. Now use the approximation hypothesis (3.9) to choose ϕ_h so that

$$\|a^{-1}w - \phi_h\|_0 + h\,\|a^{-1}w - \phi_h\|_1 \le C\ h^{2-s}\ \|a^{-1}w\|_{2-s}$$
$$\le C\ h^{2-s}\ \|w\|_{2-s}\ .$$

Then the first term of (3.12) is bounded by

$$C\ \|a\nabla e_u + \underset{\sim}{e}_v\|_0\ \|a^{-1}w - \phi_h\|_1 \le C\ \|(e_u,\underset{\sim}{e}_v)\|_B\ h^{1-s}\ \|w\|_{2-s},$$

the third term is bounded by

$$h^{-1/2}|e_u|_0\ h^{-1/2}|a^{-1}w - \phi_h|_0$$

$$\leq \|(e_u, \underset{\sim}{e}_v)\|_B\ C\ (\ h^{-1}\|a^{-1}w - \phi_h\|_0 + \|a^{-1}w - \phi_h\|_1)$$

$$\leq \|(e_u, \underset{\sim}{e}_v)\|_B\ C\ h^{1-s}\ \|w\|_{2-s},$$

and the second term is bounded by

$$\|\text{div}\ \underset{\sim}{e}_v - \underset{\sim}{b}\cdot a^{-1}\underset{\sim}{e}_v + ce_u\|_{-1+s}\ \|c\phi_h\|_{1-s}$$

$$\leq C\ h^{1-s}\ \|(e_u, \underset{\sim}{e}_v)\|_B\ \|w\|_{1-s}$$

by Lemma 3.4 and (3.9)

$$\leq C\ h^{1-s}\ \|(e_u, \underset{\sim}{e}_v)\|_B\ \|w\|_{2-s}.$$

Collecting all the terms, we find

$$|(e_u, \phi)| \leq C\ h^{1-s}\ \|(e_u, \underset{\sim}{e}_v)\|_B\ \|w\|_{2-s}$$

$$\leq C\ h^{1-s}\ \|(e_u, \underset{\sim}{e}_v)\|_B\ \|\phi\|_{-s}\ ,$$

which implies $\|e_u\|_s \leq C\ h^{1-s}\ \|(e_u, \underset{\sim}{e}_v)\|_B$ for $0 \leq s < 1/2$, and the proof is complete.

We conclude this chapter by giving a general theorem on the decomposition method. Let L, a second order elliptic operator, be written as $L = L_1^* L_1 + L_2 L_1 + L_3$ where L_1 is a first order differential operator and L_2, L_3 are scalar operators. For example, with $L = -\text{div}(A\nabla u) + \underset{\sim}{b}\cdot\nabla u + cu$ we can take $L_1 u = -S\nabla u$, $L_2 \underset{\sim}{v} = -\underset{\sim}{b}\cdot S^{-1}\underset{\sim}{v}$, $L_3 u = cu$, where $S^T S = A$.

Consider the decomposition scheme

$$
\begin{aligned}
L_1 u &= \underset{\sim}{v} \quad \text{in } \Omega \\
L_1^* \underset{\sim}{v} + L_2 \underset{\sim}{v} + L_3 u &= f \quad \text{in } \Omega \\
u &= g \quad \text{on } \Gamma,
\end{aligned} \tag{3.13}
$$

and the associated bilinear form

$$
\begin{aligned}
B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) &= (L_1 u - \underset{\sim}{v},\ L_1 \phi - \underset{\sim}{\psi}) \\
&\quad + (L_1^* \underset{\sim}{z} + L_2 \underset{\sim}{z} + L_3 w,\ L_1^* \underset{\sim}{\psi} + L_2 \underset{\sim}{\psi} + L_3 \phi) \\
&\quad + h^{-1} \langle w, \phi \rangle \ .
\end{aligned} \tag{3.14}
$$

We also assume that there exists a bounded map $N : H^{k+3/2}(\Omega) \to H^k(\Gamma)$ $(k > 0)$ such that $(\phi, L_1^* L_1 w) = (L_1 \phi, L_1 w) + \langle \phi, N(w) \rangle$ for all $\phi \in H^1(\Omega)$ and $w \in H^k(\Omega)$ $(k > 3/2)$. Then we have

Theorem 3.8. Under the above hypotheses, all the previous theorems of this chapter hold. In particular, $\| e_u \|_s \leq C\, h^{1-s} \| (e_u, \underset{\sim}{e_v}) \|_B$ for $0 \leq s < 1/2$.

Proof. One can check that all the previous proofs carry through unchanged.

This concludes Chapter 3.

CHAPTER IV. FURTHER REMARKS

In this chapter we consider further questions relating to the numerical scheme introduced in Chapter 3. We will consider certain modifications of the bilinear form, further error bounds, and other types of boundary conditions. We will look at a special one-dimensional problem, and also consider the condition number of the finite-dimensional problem (3.5).

To begin with, notice that the a priori error bound of Chapter 3,

$$\| (e_u, \underset{\sim}{e}_v) \|_B \leq C \ (h^{s-1} \|u\|_s + h^{t-1} \|\underset{\sim}{v}\|_t),$$

yields the following if $r_1 = r_2 = r$ and $s, t = r$:

$$\begin{aligned} \| (e_u, \underset{\sim}{e}_v) \|_B &\leq C\, h^{r-1} \ (\|u\|_r + \|\underset{\sim}{v}\|_r) \\ &\leq C\, h^{r-1} \ (\|u\|_r + \|u\|_{r+1}) \\ &\leq C\, h^{r-1} \ (\|f\|_{r-1} + |g|_{r+1/2}), \end{aligned}$$

and hence

$$\|e_u\|_0 \leq C\, h \, \| (e_u, \underset{\sim}{e}_v) \|_B \leq C\, h^r \ (\|f\|_{r-1} + |g|_{r+1/2}).$$

Notice this has the "right" power of h on the right-hand side, but the rest of the expression is "wrong" in the sense that we should be able to bound $\|e_u\|_0$ by

$$C\, h^r \|u\|_r \le C\, h^r \left(\|f\|_{r-2} + |g|_{r-1/2} \right).$$

Our expression requires more smoothness on f and g than one would think necessary. Looking at this more closely leads one to consider the following modified bilinear form, again denoted by B. (For simplicity we stick to the case $L = -\Delta$; the general case is similar.)

$$B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) = (\nabla w + \underset{\sim}{z},\ \nabla\phi + \underset{\sim}{\psi}) + h^2\ (\operatorname{div} \underset{\sim}{z},\ \operatorname{div} \underset{\sim}{\psi}) + h^{-1} \langle w,\phi \rangle . \tag{4.1}$$

Then we have the a priori bound

$$\|(w,\underset{\sim}{z})\|_B \le C \left(h^{-1} \|w\|_0 + \|w\|_1 + h\, \|\underset{\sim}{z}\|_1 + \|\underset{\sim}{z}\|_0 \right),$$

and so if $(u^h,\underset{\sim}{v}^h)$ solves the problem

$$B((u^h,\underset{\sim}{v}^h),(\phi_h,\underset{\sim}{\psi}_h)) = B((u,\underset{\sim}{v}),(\phi_h,\underset{\sim}{\psi}_h)) \quad \text{for all } (\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h, \tag{4.2}$$

we have the bound

$$\begin{aligned} \|(e_u,\underset{\sim}{e}_v)\|_B &\le C \inf_{\phi_h,\underset{\sim}{\psi}_h} \left(h^{-1} \|u-\phi_h\|_0 + \|u-\phi_h\|_1 + \|\underset{\sim}{v}-\underset{\sim}{\psi}_h\|_0 + h\, \|\underset{\sim}{v}-\underset{\sim}{\psi}_h\|_1 \right) \\ &\le C \inf_{\phi_h,\underset{\sim}{\psi}_h} \left(\|u-\phi_h\|_1 + \|\underset{\sim}{v}-\underset{\sim}{\psi}_h\|_0 \right) \text{ by (3.9)} \\ &\le C \left(h^{s-1} \|u\|_s + h^t \|\underset{\sim}{v}\|_t \right) \end{aligned}$$

if $u \in H^s(\Omega)$, $1 \le s \le r_1$, $v \in \underset{\sim}{H}^t(\Omega)$, $1 \le t \le r_2$.

Now suppose we have $u \in H^{r_1}(\Omega)$ and $r_2 = r_1 - 1$.

Then

$$\begin{aligned} \|(e_u, \underset{\sim}{e}_v)\|_B &\le C \left(h^{r_1-1} \|u\|_{r_1} + h^{r_2} \|\underset{\sim}{v}\|_{r_2} \right) \\ &\le C\, h^{r_1-1} \|u\|_{r_1} \quad \text{since } \underset{\sim}{v} = -\nabla u \\ &\le C\, h^{r_1-1} \left(\|f\|_{r_1-2} + |g|_{r_1-1/2} \right), \end{aligned}$$

so the "right" amount of smoothness is now required of f and g.

The question now arises, is $\|e_u\|_0$ quasi-optimal for this new scheme? A sufficient condition for this is provided by the next theorem.

Theorem 4.1. If $r_1 \ge 4$ and $r_2 = r_1 - 1$ then

$$\|e_u\|_0 \le C\, h\, \|(e_u, \underset{\sim}{e}_v)\|_B.$$

Proof. We are restricting attention to the case of $\|e_u\|_0$ for simplicity; the proof that

$$\|e_u\|_s \le C\, h^{1-s} \|(e_u, \underset{\sim}{e}_v)\|_B \quad \text{for } 0 \le s < 1/2$$

is similar.

As in Lemma 3.2, we verify $|e_u|_{-1/2} \le C\, h\, \|(e_u, \underset{\sim}{e}_v)\|_B$. Lemma 3.4 requires revision, however. Indeed, we revise the Lemma to read

$$\|\operatorname{div} \underset{\sim}{e}_v\|_{L_2} \le C\, h\, \|(e_u, \underset{\sim}{e}_v)\|_B.$$

For a proof, let $\phi \in H^2(\Omega)$ and let α solve $-\Delta\alpha = \phi$ in Ω, $\phi = 0$ on Γ, and let $\underset{\sim}{\beta} = -\nabla\alpha$. Then $\|\alpha\|_4 + \|\underset{\sim}{\beta}\|_3 \le C\, \|\phi\|_2$, and

$$
\begin{aligned}
(\operatorname{div} \underset{\sim}{e}_v,\ \phi) &= (\operatorname{div} \underset{\sim}{e}_v,\ \operatorname{div} \underset{\sim}{\beta}) \\
&= h^{-2}\ (\ h^2\ (\operatorname{div} \underset{\sim}{e}_v,\ \operatorname{div} \underset{\sim}{\beta})) \\
&= h^{-2}\ B((e_u, \underset{\sim}{e}_v), (\alpha, \underset{\sim}{\beta})) \\
&= h^{-2}\ B((e_u, \underset{\sim}{e}_v), (\alpha-\alpha_h, \underset{\sim}{\beta}-\underset{\sim}{\beta}_h)) \text{ for all } (\alpha_h, \underset{\sim}{\beta}_h) \\
&\le h^{-2}\ \|(e_u, \underset{\sim}{e}_v)\|_B\ C\ (\ \|\alpha-\alpha_h\|_1 + \|\underset{\sim}{\beta}-\underset{\sim}{\beta}_h\|_0) \\
&\le h^{-2}\ \|(e_u, \underset{\sim}{e}_v)\|_B\ C\ (\ h^3\ \|\alpha\|_4 + h^3\ \|\underset{\sim}{\beta}\|_3) \\
&\qquad \text{since } r_1 \ge 4 \text{ and } r_2 = r_1 - 1 \\
&\le C\ h\ \|(e_u, \underset{\sim}{e}_v)\|_B\ \|\phi\|_2\ ,
\end{aligned}
$$

and we are done.

The rest of the proof proceeds as in Theorem 3.5.

Thus if we use the modified bilinear form (4.1) we can ensure the optimal convergence rate for $\|e_u\|_0$ by using higher-degree polynomials as our trial functions. The functions in S_1^h must be (at least) piecewise cubic, those in S_2^h must be piecewise quadratic. This may be undesirable in practice, for if we are going to use high-order polynomials as elements of S_1^h we might just as well stick to the least squares method applied to the original equation $Lu = f$.

It is natural to ask if there is a connection between the solution $(u^h, \underset{\sim}{v}^h)$ of the problem (4.2) and the solution $\hat{u}^h$ of the problem

(4.3) Find $\hat{u}^h \in S^h \subset H^2(\Omega)$ such that

$$\hat{B}(\hat{u}^h,\phi_h) \equiv (\Delta\hat{u}^h,\Delta\phi_h) + h^{-3}\langle \hat{u}^h,\phi_h\rangle$$

$$= \hat{B}(u,\phi_h)$$

$$= -(f,\Delta\phi_h) + h^{-3}\langle g,\phi_h\rangle \text{ for all } \phi_h \in S^h.$$

Thus problem (4.3) is the original least squares formulation. A priori there need be no connection between the two problems, since the space S_1^h need not be contained in $H^2(\Omega)$. So let us suppose $S^h = S_1^h \subset H^2(\Omega)$ and $\nabla\phi_h \in \underset{\sim}{S}_2^h$ for all $\phi_h \in S_1^h$. Then we have the following result.

Proposition 4.2. If $\underset{\sim}{v}^h = -\nabla u^h$ then $u^h = \hat{u}^h$, so the original least squares formulation and the decomposition method give the same solution.

Proof. If $\underset{\sim}{v}^h = -\nabla u^h$ then for all $(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$,

$$-h^2\,(\Delta u^h,\ \text{div}\ \underset{\sim}{\psi}_h) + h^{-1}\langle u^h,\phi_h\rangle$$

$$= h^2\,(f,\text{div}\ \underset{\sim}{\psi}_h) + h^{-1}\langle g,\phi_h\rangle\ .$$

In particular, for $\underset{\sim}{\psi}_h = -\nabla\phi_h$ this says

$$(\Delta u^h,\Delta\phi_h) + h^{-3}\langle u^h,\phi_h\rangle = -(f,\Delta\phi_h) + h^{-3}\langle g,\phi_h\rangle$$

for all $\phi_h \in S_1^h$, that is,

$$\hat{B}(u^h,\phi_h) = \hat{B}(u,\phi_h) \text{ for all } \phi_h \in S_1^h = S^h.$$

By uniqueness of the solution to (4.3), $u^h = \hat{u}^h$ and we are done.

Unfortunately, the hypothesis $\underset{\sim}{v}^h = -\nabla u^h$ of the Proposition is a strong one and we are unable to verify or disprove it.

We consider one more idea for modifying the bilinear form B. For simplicity consider the case $L = -\Delta$ and $\Omega \subset R^2$. If we are given Dirichlet boundary conditions $u = g$ on Γ and g is smooth, then $-\nabla u = \underset{\sim}{v}$ implies $t \cdot \underset{\sim}{v} = -dg/ds$ on Γ where t is the unit tangent vector on Γ and d/ds denotes differentiation with respect to arc length. Thus one could consider minimizing the functional

$$J_h(w,\underset{\sim}{z}) = \|\nabla w + \underset{\sim}{z}\|_0^2 + \|\text{div } \underset{\sim}{z} - f\|_0^2 + h^{-1}|w-g|_0^2 + \omega|t\cdot\underset{\sim}{z} + dg/ds|_0^2$$

where $\omega = \omega(h)$ is some weighting factor. The bilinear form to consider is then

$$B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) = (\nabla w + \underset{\sim}{z}, \nabla\phi + \underset{\sim}{\psi}) + (\text{div } \underset{\sim}{z}, \text{div } \underset{\sim}{\psi}) + h^{-1}\langle w,\phi\rangle + \omega\langle t\cdot\underset{\sim}{z}, t\cdot\underset{\sim}{\psi}\rangle .$$

Let us choose $\omega = h^{-1}$. Then we have the bound

$$\|(w,\underset{\sim}{z})\|_B \le C\,(h^{-1}\|w\|_0 + \|w\|_1 + h^{-1}\|\underset{\sim}{z}\|_0 + \|\underset{\sim}{z}\|_1)$$

since $h^{-1/2}|t\cdot\underset{\sim}{z}|_0 \le h^{-1/2}|\underset{\sim}{z}|_0 \le C\,(h^{-1}\|\underset{\sim}{z}\|_0 + \|\underset{\sim}{z}\|_1)$.

Hence $\|(e_u, \underset{\sim}{e}_v)\|_B \le C \inf_{(\phi_h, \underset{\sim}{\psi}_h)} (\|u-\phi_h\|_1 + \|\underset{\sim}{v}-\underset{\sim}{\psi}_h\|_1)$.

What about $\|e_u\|_s$? Let it suffice to note that all the proofs in Chapter 3 can be carried over to this case without essential change, so
$\|e_u\|_s \le C\, h^{1-s} \|(e_u, \underset{\sim}{e}_v)\|_B$ for $0 \le s < 1/2$.

We finally consider one point concerning the degrees of the polynomials in the spaces S_1^h and S_2^h. We expect the approximate solution $(u^h, \underset{\sim}{v}^h)$ of (3.5) to mimic the behavior of the solution $(u, -\nabla u)$, so we expect $\underset{\sim}{v}^h \approx -\nabla u^h$. Thus it seems reasonable for S_2^h to consist of piecewise polynomials of one degree lower than S_1^h, for example if we use piecewise quadratic functions in S_1^h we would use piecewise linear functions in S_2^h. Then $r_2 = r_1 - 1$. Is it now the case that $\|e_u\|_0$ is quasi-optimal? The answer seems to be no, for the following reason. We still (of course) have $\|e_u\|_0 \le C\, h \|(e_u, \underset{\sim}{e}_v)\|_B$ by Theorem 3.5, but the best bound on $\|(e_u, \underset{\sim}{e}_v)\|_B$ is (assuming for simplicity $u \in H^{r_1}(\Omega)$)

$$\|(e_u, \underset{\sim}{e}_v)\|_B \le C\, (h^{r_1-1} \|u\|_{r_1} + h^{r_2-1} \|\underset{\sim}{v}\|_{r_2})$$

$$\le C\, (h^{r_1-1} \|u\|_{r_1} + h^{r_1-2} \|u\|_{r_2+1})$$

$$\le C\, (h^{r_1-1} + h^{r_1-2}) \|u\|_{r_1}$$

$$\le C\, h^{r_1-2} \|u\|_{r_1},$$

and hence $\|e_u\|_0 \leq C\, h^{r_1-1} \|u\|_{r_1}$, which is non-optimal.

It should be pointed out this is not a proof, but it it does suggest that $\|e_u\|_0$ is non-optimal in this case.

On the other hand, we have shown that if one works with the modified bilinear form (4.1) then one can take S_2^h to consist of piecewise polynomials of degree one less than the polynomials of S_1^h and still get quasi-optimal error estimates for $\|e_u\|_0$, provided S_1^h contains (at least) cubic polynomials.

The Condition Number

The numerical solution of a system of linear equations will generally be affected by roundoff errors, and the condition number is one way of estimating how severe the errors can be. In our situation we have three functions to consider: the true solution u, the approximate solution u^h computed in exact arithmetic, and the solution $\tilde{u}^h$ actually found on the computer. We have in Theorem 3.5 an error bound for $u-u^h$; an estimate for the condition number enables us to bound $u^h-\tilde{u}^h$ and hence to bound $u-\tilde{u}^h$. Specifically, let $\{\phi_i\}$ be a basis for S_1^h, and write $u^h = \Sigma\, \alpha_i\phi_i$, $\tilde{u}^h = \Sigma\, \tilde{\alpha}_i\phi_i$. Assume that for all $\{\xi_i\}$, $\|\xi_i\phi_i\|_0^2 \leq C\, h^N\, \Sigma\xi_i^2$. (This condition is satisfied in practice; see [22] and also (4.7) below.) Then, putting $\varepsilon_i = \alpha_i-\tilde{\alpha}_i$, $\|u^h-\tilde{u}^h\|_0^2 = \|\Sigma\varepsilon_i\phi_i\|_0^2 \leq C\, h^N\, \Sigma\varepsilon_i^2$. Also,

$(\Sigma\varepsilon_i^2)^{1/2} \leq C\ \text{macheps}\cdot\text{cond}(B)\cdot(\Sigma\alpha_i^2)^{1/2}$, where B is the matrix of the finite-dimensional problem (3.5) and macheps is the smallest positive floating-point number such that $1 + \text{macheps} \neq 1$ [13]. Hence, assuming $r_1 = r_2 = r$,

$$\|u-\tilde{u}^h\|_0 \leq \|u-u^h\|_0 + \|u^h-\tilde{u}^h\|_0$$

$$\leq C\ h^r(\ \|f\|_{r-1} + |g|_{r-1/2}) + \|u^h-\tilde{u}^h\|_0$$

by Theorem 3.5

$$\leq C\ h^r(\ \|f\|_r + |g|_{r-1/2})$$
$$+ C\ h^{N/2}\cdot\text{macheps}\cdot\text{cond}(B)\cdot(\Sigma\alpha_i^2)^{1/2}.$$

Thus for the error due to roundoff not to contaminate the theoretical error bound, it suffices that $\text{macheps}\cdot\text{cond}(B) \leq C\ h^{r-N/2}$. This gives a relation between the number of digits carried in the calculation and the condition number of the matrix. We will show below that (under suitable hypotheses) $\text{cond}(B) \leq C\ h^{-2}$; hence it suffices that $\text{macheps} \leq C\ h^{r+2-N/2}$. Thus for fixed r and macheps, if h is too small there is a danger of roundoff error destroying the accuracy of the computed approximation.

The key tool in proving the condition number estimate is an a priori inequality to which we now turn. Recall that $\tilde{H}^{1,2}(\Omega)$ is the completion of $C^\infty(\bar{\Omega})$ in the norm

$$\|w\|^2_{\tilde{H}^{1,2}(\Omega)} = \|w\|_1^2 + |w|^2_{1/2} + \left|\frac{\partial w}{\partial n}\right|^2_{-1/2}.$$

Proposition 4.3. For all $(w,\underset{\sim}{z}) \in H^1(\Omega) \times \underset{\sim}{H}^1(\Omega)$,

$$(4.4) \qquad \|w\|_1 + \|\underset{\sim}{z}\|_0 \le C \left(\|S\nabla w + \underset{\sim}{z}\|_0 + \|\mathrm{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw\|_{-1} + |w|_{1/2} \right).$$

Proof. Given $\phi \in H^{-1}(\Omega)$, let α solve $L^*\alpha = \phi$ in Ω, $\alpha = 0$ on Γ. Then $\alpha \in \tilde{H}^{1,2}(\Omega)$ and $\|\alpha\|_{\tilde{H}^{1,2}(\Omega)} \le C \|\phi\|_{-1}$. Now, by the same computation as in Theorem 3.5,

$$(w,\phi) = (w,L^*\alpha)$$

$$= (S\nabla w+\underset{\sim}{z}, S\nabla\alpha) + (\mathrm{div}(S\underset{\sim}{z}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z} + cw, \alpha) - \langle w, N(\alpha)\rangle + (S\nabla w+\underset{\sim}{z}, S^{-1}(\alpha\underset{\sim}{b}))$$

where $|N(\alpha)|_{-1/2} \le C \|\alpha\|_{\tilde{H}^{1,2}(\Omega)}$. Taking absolute values gives the bound for $\|w\|_1$. For $\|\underset{\sim}{z}\|_0$, write $\underset{\sim}{z} = (S\nabla w+\underset{\sim}{z}) - S\nabla w$ and take norms. This completes the proof.

It is easy to see that $\|\phi\|_{-1} \le \|\phi\|_0$ for $\phi \in H^0(\Omega)$. Hence (4.4) almost gives a lower bound for $\|(w,\underset{\sim}{z})\|_B$, $\|(w,\underset{\sim}{z})\|_B \ge c \left(\|w\|_1 + \|\underset{\sim}{z}\|_0 \right)$. It fails to do so only because of the discrepancy in norms on the boundary. This is a fundamental discrepancy; one can find examples of sequences $\{w_n\}$ and $\{\underset{\sim}{z}_n\}$ where (for any fixed $h > 0$), $\|(w_n,\underset{\sim}{z}_n)\|_B \to 0$ but $\|w_n\|_1 + \|\underset{\sim}{z}_n\|_0 = 1$. (Example: Let Ω = unit disk, $L = -\Delta$, $w_n = \mathrm{Re}(I_n(r)e^{in\theta})$ where I_n is the n^{th} modified Bessel function, $\underset{\sim}{z}_n = -\nabla w_n$, normalized so $\|\underset{\sim}{z}_n\|_0 + \|w_n\|_1 = 1$.) Thus we can only

hope to find a lower bound on the subspace $S_1^h \times \underset{\sim}{S}_2^h$. To do this we make the following assumption.

(4.5) There exists a $C > 0$ such that

$$|w_h|_{1/2} \leq C\, h^{-1/2}\, |w_h|_0 \quad \text{as } h \to 0 \text{ for all } w_h \in S_1^h.$$

This is an assumption of "inverse" type (bounding a higher Sobolev norm by a lower one) on the boundary. To get an idea of the implications of this assumption, consider the case $\Omega \subset R^2$. Then Γ is one-dimensional. For polynomials of degree at most k in one space dimension on an interval J_i of length h_i, it is known that $|p|_{1/2,J_i} \leq C\, h_i^{-1/2} |p|_{0,J_i}$, where C depends on k and the notation $|\cdot|_{s,J_i}$ denotes the $H^s(J_i)$ norm. Let $J = \bigcup J_i$. Then $|p|_{1/2,J} \leq C\, (\Sigma |p|^2_{0,J_i} h_i^{-1})^{1/2}$. For this to be bounded by $C\, h^{-1/2}\, |p|_{0,J}$ where $h = \max_i \{h_i\}$ it suffices that $c h_i^{-1} \leq h^{-1}$ for all i, that is, $h_i/h \geq c > 0$ as $h \to 0$, i.e., the mesh in J is "quasi-uniform." So (heuristically) in the two-dimensional case a sufficient condition for (4.5) is that the boundary mesh be "quasi-uniform."

Hypotheses of inverse type have been seen before. They are the usual means of deriving "higher-order" error bounds [4] and we will see this role later.

We now study the matrix of the finite-dimensional problem (3.5). Its entries are various inner products of basis elements of the spaces S_1^h and $\underset{\sim}{S}_2^h$. Suppose a basis for S_1^h is $\{w_1,\ldots,w_n\}$ and $\{\underset{\sim}{z}_1,\ldots,\underset{\sim}{z}_m\}$ is a basis for $\underset{\sim}{S}_2^h$. Call the matrix B. Then B is a symmetric positive definite (m+n) × (m+n) matrix whose entries b_{ij} are given by

$$b_{ij} = B((w_i,0),(w_j,0))$$
$$= (S\nabla w_i, S\nabla w_j) + (cw_i, cw_j) + h^{-1}\langle w_i, w_j\rangle$$
$$\text{for } 1 \le i,j \le n$$

$$b_{ij} = B((w_i,0),(0,\underset{\sim}{z}_{j-n}))$$
$$= (S\nabla w_i, \underset{\sim}{z}_{j-n}) + (cw_i, \operatorname{div}(S\underset{\sim}{z}_{j-n}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z}_{j-n}) = b_{ji}$$
$$\text{for } 1 \le i \le n,\ n+1 \le j \le m+n$$

$$b_{ij} = B((0,\underset{\sim}{z}_{i-n}),(0,\underset{\sim}{z}_{j-n}))$$
$$= (\underset{\sim}{z}_{i-n}, \underset{\sim}{z}_{j-n}) + (\operatorname{div}(S\underset{\sim}{z}_{i-n}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z}_{i-n},$$
$$\operatorname{div}(S\underset{\sim}{z}_{j-n}) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{z}_{j-n})$$
$$\text{for } n+1 \le i,j \le m+n.$$

(Of course a different ordering of the unknowns will greatly decrease the bandwidth of the matrix B, but for conceptual clarity we stick to this formulation.)

We now assume the bases are chosen so that the following conditions hold [6]. Recall $\Omega \subset R^N$, and the notation $(\cdot,\cdot)_1$ denotes the $H^1(\Omega)$ inner product.

For all real numbers $\xi_1,\ldots,\xi_n,\eta_1,\ldots,\eta_m$

$$\text{(a)}\quad ch^N \Sigma \xi_i^2 \le \Sigma (w_i,w_j)\xi_i\xi_j \le Ch^N \Sigma \xi_i^2$$

$$\text{(b)}\quad ch^N \Sigma \eta_i^2 \le \Sigma (\underset{\sim}{z}_i,\underset{\sim}{z}_j)\eta_i\eta_j \le Ch^N \Sigma \eta_i^2 \qquad (4.6)$$

$$\text{(c)}\quad \Sigma (w_i,w_j)_1 \xi_i\xi_j \le C h^{N-2} \Sigma \xi_i^2$$

$$\text{(d)}\quad \Sigma (\underset{\sim}{z}_i,\underset{\sim}{z}_j)_1 \eta_i\eta_j \le C h^{N-2} \Sigma \eta_i^2$$

The conditions (4.6) assert that the basis functions are independent in a certain sense. For the usual finite element bases the conditions hold, provided the geometry of the elements does not become degenerate [22].

Now we have

Theorem 4.4. Assume the boundary inverse hypothesis (4.5) and the basis conditions (4.6). Then the condition number of the matrix B is $O(h^{-2})$, i.e., $\text{cond}(B) \le C h^{-2}$.

Proof. Since B is symmetric positive definite, $\text{cond}(B) = \lambda_{max}(B)/\lambda_{min}(B)$, the maximum eigenvalue of B divided by the minimum eigenvalue. We can estimate the eigenvalues by considering the Rayleigh quotient, $\xi^T B\xi/\xi^T\xi$. So let $\xi = (\xi_1,\ldots,\xi_n,\eta_1,\ldots,\eta_m)$ be an (m+n)-component vector. Define $w_h = \Sigma \xi_i w_i$, $\underset{\sim}{z}_h = \Sigma \eta_i \underset{\sim}{z}_i$. Then by (4.5) and (4.6),

$$C\, B((w_h,\underset{\sim}{z}_h),(w_h,\underset{\sim}{z}_h)) \ge \|w_h\|_1^2 + \|\underset{\sim}{z}_h\|_0^2$$

$$\ge \|w_h\|_0^2 + \|\underset{\sim}{z}_h\|_0^2$$

$$= \Sigma \xi_i \xi_j (w_i, w_j) + \Sigma \eta_i \eta_j (\underset{\sim}{z}_i, \underset{\sim}{z}_j)$$

$$\geq c\, h^N \, (\Sigma \xi_i^2 + \Sigma \eta_i^2)$$

$$= c\, h^N \, \xi^T \xi \, .$$

But also

$$B((w_h, \underset{\sim}{z}_h), (w_h, \underset{\sim}{z}_h)) \leq C \, (h^{-2} \, \|w_h\|_0^2 + \|w_h\|_1^2 + \|\underset{\sim}{z}_h\|_0^2 + \|\underset{\sim}{z}_h\|_1^2)$$

$$\leq C \, (h^{-2} \, \|w_h\|_0^2 + \|w_h\|_1^2 + \|\underset{\sim}{z}_h\|_1^2)$$

$$\leq C \, (h^{-2} \, \Sigma \xi_i \xi_j (w_i, w_j) + \Sigma \xi_i \xi_j (w_i, w_j)_1$$
$$+ \Sigma \eta_i \eta_j (\underset{\sim}{z}_i, \underset{\sim}{z}_j)_1)$$

$$\leq C \, (h^{N-2} \, \Sigma \xi_i^2 + h^{N-2} \, \Sigma \xi_i^2 + h^{N-2} \, \Sigma \eta_i^2)$$
by (4.6)

$$\leq C \, h^{N-2} \, \xi^T \xi \, .$$

To complete the proof simply note that $B((w_h, \underset{\sim}{z}_h), (w_h, \underset{\sim}{z}_h)) = \xi^T B \xi$.

We remark here that the proof would have worked for any decomposition of the form considered at the end of Chapter II, $L = L_1^* L_1 + L_2 L_1 + L_3$. Also, suppose we were working with the bilinear form as modified in (4.1). Then the preceding results hold provided an additional assumption is made; it suffices to assume $\|\text{div } \underset{\sim}{z}_h\|_{L_1} \leq C\, h \, \|\text{div } \underset{\sim}{z}_h\|_0$ for all $\underset{\sim}{z}_h \in \underset{\sim}{S}_2^h$. This assumption coupled with the assumption (4.5) is suf-

ficient to prove Theorem 4.4 in this case. However, the meaning of the additional assumption $\|\text{div } \underset{\sim}{z}_h\|_{-1} \leq C\, h\, \|\text{div } \underset{\sim}{z}_h\|_0$ is unclear.

Further Error Bounds

We now turn our attention to the derivation of bounds for $\|e_u\|_s$ ($s \geq 1/2$) and $\|\underset{\sim}{e}_v\|_0$. We will find an inverse hypothesis of one type or another essential.

Proposition 4.5. Suppose the boundary inverse hypothesis (4.5) holds. Then

$$|e_u|_{1/2} \leq C \left(\|(e_u, \underset{\sim}{e}_v)\|_B + \inf_{\phi_h \in S_1^h} (h^{-1} \|u-\phi_h\|_0 + \|u-\phi_h\|_1) \right).$$

Proof. We have

$$\begin{aligned} |u-u^h|_{1/2} &\leq |u-\phi_h|_{1/2} + |\phi_h - u^h|_{1/2} \quad \text{for all } \phi_h \in S_1^h \\ &\leq |u-\phi_h|_{1/2} + C\, h^{-1/2} |\phi_h - u^h|_0 \quad \text{by (4.5)} \\ &\leq C \|u-\phi_h\|_1 + C\, h^{-1/2} (|\phi_h - u|_0 + |u-u^h|_0) \\ &\qquad \text{by Theorem 2.1} \\ &\leq C \|u-\phi_h\|_1 + C\, h^{-1} \|\phi_h - u\|_0 + C \|\phi_h - u\|_1 \\ &\qquad + C \|(e_u, \underset{\sim}{e}_v)\|_B \quad \text{by Lemma 2.3,} \end{aligned}$$

and the proof is complete.

Corollary 4.6. If the approximation hypothesis (3.9) also holds, then $|e_u|_{1/2} \le C\ (\ h^{s-1}\,\|u\|_s + h^{t-1}\,\|\underset{\sim}{v}\|_t)$ if $u \in H^s(\Omega)$, $\underset{\sim}{v} \in \underset{\sim}{H}^t(\Omega)$, $1 \le s \le r_1$, $1 \le t \le r_2$.

The proof of the Corollary is immediate.

Theorem 4.7. If the boundary inverse hypothesis (4.5) holds, then

$$\|e_u\|_1 + \|\underset{\sim}{e}_v\|_0 \le C\ (\|(e_u,\underset{\sim}{e}_v)\|_B + \inf_{\phi_h \in S_1^h} (h^{-1}\|u-\phi_h\|_0 + \|u-\phi_h\|_1)\,.$$

Proof. By Propositions 4.3 and 4.5,

$$\begin{aligned}\|e_u\|_1 + \|\underset{\sim}{e}_v\|_0 &\le C\ (\ \|S\nabla e_u + \underset{\sim}{e}_v\|_0 + \|\mathrm{div}(S\underset{\sim}{e}_v) - \underset{\sim}{b}\cdot S^{-1}\underset{\sim}{e}_v + ce_u\|_{-1} \\ &\qquad + |e_u|_{1/2}) \\ &\le C\ (\ \|(e_u,\underset{\sim}{e}_v)\|_B + \inf_{\phi_h} (h^{-1}\|u-\phi_h\|_0 + \|u-\phi_h\|_1))\end{aligned}$$

and the proof is complete.

Corollary 4.8. If the approximation hypothesis (3.9) also holds, then

$$\|e_u\|_1 + \|\underset{\sim}{e}_v\|_0 \le C\ (\ h^{s-1}\,\|u\|_s + h^{t-1}\,\|\underset{\sim}{v}\|_t)$$

if $u \in H^s(\Omega)$, $\underset{\sim}{v} \in \underset{\sim}{H}^t(\Omega)$, $1 \le s \le r_1$, $1 \le t \le r_2$.

Thus if $u \in H^{r_1}(\Omega)$, $\underset{\sim}{v} \in \underset{\sim}{H}^{r_2}(\Omega)$ and $r_1 = r_2$ the Corollary gives an optimal error bound for $\|e_u\|_1$ but a suboptimal error bound for $\|\underset{\sim}{e}_v\|_0$. It remains unclear whether or not this suboptimal bound is sharp.

Let us now look at another hypothesis of inverse type, one that has been seen in the literature [7].

(4.7) For each open set G with $\bar{G} \subset \Omega$, there exists an open set H, $\bar{G} \subset H \subset \bar{H} \subset \Omega$ and a constant C such that for all sufficiently small h,

$$\|w_h\|_{1,H} \leq C\, h^{-1} \|w_h\|_{0,H} \quad \text{for all } w_h \in S_1^h.$$

(Here we have adopted the notation $\|\cdot\|_{s,G}$ for the $H^s(G)$ norm.) This hypothesis is a classical inverse hypothesis and is verified for some special choices of finite element spaces in [7].

Proposition 4.9. Under the hypotheses (3.9) and (4.7), for each open set G with $\bar{G} \subset \Omega$ we have

$$\|e_u\|_{1,G} + \|\underset{\sim}{e}_v\|_{0,G} \leq C \left(\|(e_u, \underset{\sim}{e}_v)\|_B + \inf_{w_h \in S_1^h} \left(h^{-1} \|u-w_h\|_{0,G} + \|u-w_h\|_{1,G} \right) \right).$$

Proof. $\|u-u^h\|_{1,G} \leq \|u-w_h\|_{1,G} + \|w_h-u^h\|_{1,G}$ for all w_h

$$\leq \|u-w_h\|_{1,G} + C\, h^{-1} \|w_h-u^h\|_{0,H} \quad \text{by (4.7)}$$

$$\leq \|u-w_h\|_1 + C\, h^{-1} \left(\|w_h-u\|_0 + \|u-u^h\|_0 \right)$$

$$\leq \|u-w_h\|_1 + C\, h^{-1} \|w_h-u\|_0 + C\, \|(e_u, \underset{\sim}{e}_v)\|_B,$$

the last step following from Theorem 3.5. This proves the inequality for e_u. For $\underset{\sim}{e}_v$, write $\underset{\sim}{e}_v = (\nabla e_u + \underset{\sim}{e}_v) - \nabla e_u$ and take norms.

Corollary 4.10. If the inverse hypothesis (4.7) and the approximation hypothesis (3.9) hold, then for any open G with $\bar{G} \subset \Omega$ we have

$$\|e_u\|_{1,G} + \|\underset{\sim}{e}_v\|_{0,G} \leq C \left(h^{s-1} \|u\|_{s,\Omega} + h^{t-1} \|\underset{\sim}{v}\|_{t,\Omega} \right)$$

if $u \in H^s(\Omega)$, $\underset{\sim}{v} \in \underset{\sim}{H}^t(\Omega)$, $1 \leq s \leq r_1$, $1 \leq t \leq r_2$.

The proof of the Corollary is immediate. Again, the estimate for $\|e_u\|_{1,G}$ is quasi-optimal (if $r_1 = r_2$) but the estimate for $\|\underset{\sim}{e}_v\|_{0,G}$ is suboptimal.

Let us briefly consider the modified bilinear form (4.1).

Proposition 4.11. Assume the boundary inverse hypothesis (4.5), and assume $r_1 \geq 4$, $r_2 = r_1 - 1$. Then for the modified bilinear form we have

$$\|e_u\|_1 + \|\underset{\sim}{e}_v\|_0 \leq C \left(\|(e_u, \underset{\sim}{e}_v)\|_B + \inf_{\phi_h \in S_1^h} (h^{-1} \|u - \phi_h\|_0 + \|u - \phi_h\|_1) \right).$$

Proof. Begin by noting that Proposition 4.5 holds unchanged for this case. It is easy to prove that $\|\text{div } \underset{\sim}{e}_v\|_{L_1} \leq C \|(e_u, \underset{\sim}{e}_v)\|_B$. Hence the proof of Theorem 4.7 is applicable.

Notice this result gives optimal bounds for both $\|e_u\|_1$ and $\|\underset{\sim}{e}_v\|_0$.

The One - Dimensional Case

We now turn to the special case where the differential equation is posed in a one-dimensional domain, so we are dealing with a two-point boundary value problem. It turns out in this case we can find somewhat stronger results than the above. Our model problem will be $Lu \equiv -(pu')' + qu' + ru = f$ on $(0,1)$, $u(1) = u(0) = 0$. For ellipticity we assume $p(x) \geq p_0 \geq 0$ on $[0,1]$. To avoid complications assume p, q, r and f are smooth functions, and assume the problem $Lu = 0$ on $(0,1)$, $u(0) = u(1) = 0$ has only the zero solution, and similarly for the problem $L^*u = 0$, $u(0) = u(1) = 0$.

We will decompose L as in Chapter III. Since in one dimension the boundary conditions are easy to satisfy with our trial functions, we will force the boundary conditions to hold. So consider the equivalent problem

$$-p^{1/2}u' = v, \quad (p^{1/2}v)' - qp^{-1/2}v + ru = f,$$
$$u(0) = u(1) = 0.$$

Choose $S_1^h \subset H_0^1(0,1)$, $S_2^h \subset H^1(0,1)$ and seek $(u^h, v^h) \in S_1^h \times S_2^h$ which minimizes the functional

$$J(w,z) = \| p^{1/2}w' + z \|_0^2 + \| (p^{1/2}z)' - qp^{-1/2}z + rw - f \|_0^2.$$

Define a bilinear form via

$$B((w,z),(\phi,\psi)) = (p^{1/2}w' + z, p^{1/2}\phi' + \psi)$$
$$+ ((p^{1/2}z)' - qp^{-1/2}z + rw, (p^{1/2}\psi)' - qp^{-1/2}\psi + r\phi)$$

and seek a solution $(u^h, v^h) \in S_1^h \times S_2^h$ of the finite-dimensional problem

$$B((u^h,v^h),(\phi_h,\psi_h)) = B((u,v),(\phi_h,\psi_h))$$

$$= (f,(p^{1/2}\psi_h)' - qp^{-1/2}\psi_h + r\phi_h)$$

$$\text{for all } (\phi_h,\psi_h) \in S_1^h \times S_2^h.$$

Again make the approximation assumption (3.9). Then the analogs of Proposition 3.1, Lemmas 3.4 and 3.6, and Theorem 3.5 carry over, so a unique solution (u^h, v^h) exists and $\|e_u\|_0 = \|u-u^h\|_0$ is quasi-optimal. The interesting result in the one-dimensional case is a quasi-optimality result for $\|v-v^h\|_0$.

Proposition 4.12. Under the above hypotheses,

$$\|e_v\|_0 \le C\, h\, \|(e_u,e_v)\|_B.$$

Proof. We begin with a lemma whose proof will be deferred.

Lemma 4.13. Given $g \in H^1(0,1)$, there exist functions $\phi \in H_0^1(0,1)$, $\psi \in H^1(0,1)$ with $(p^{1/2}\psi)' = r\phi$, $\psi - p^{1/2}\phi' - qp^{-1/2}\phi = g$ on $(0,1)$, and $\|\psi\|_2 + \|\phi\|_1 \le C\,\|g\|_0$.

Let us take such ϕ,ψ for the case $g = e_v$. Then

$$\|e_v\|_0^2 = (e_v, \psi - p^{1/2}\phi' - qp^{-1/2}\phi) + (e_u, r\phi - (p^{1/2}\psi)')$$

$$= (e_v,\psi) + ((p^{1/2}e_v)',\phi) - (qp^{-1/2}e_v,\phi) + (re_u,\phi)$$

$$+ (p^{1/2}e_u',\psi) \quad \text{upon integrating by parts}$$

and noticing all the boundary terms vanish

$$= (p^{1/2}e_u' + e_v,\psi) + ((p^{1/2}e_v)' - qp^{-1/2}e_v + re_u,\phi).$$

Now let α solve the problem

$$L\alpha = \phi - (p^{1/2}\psi)' + qp^{-1/2}\psi \quad \text{on } (0,1),\ \alpha(0) = \alpha(1) = 0,$$

and put $\beta = \psi - p^{1/2}\alpha'$. Then

$$\|\alpha\|_3 + \|\beta\|_2 \le C(\|\phi\|_1 + \|\psi\|_2) \le C\|e_v\|_0$$

and $(p^{1/2}\beta)' - qp^{-1/2}\beta + r\alpha = \phi$. Hence

$$\|e_v\|_0^2 = (p^{1/2}e_u' + e_v,\psi) + ((p^{1/2}e_v)' - qp^{-1/2}e_v + re_u,\phi)$$

$$= B((e_u,e_v),(\alpha,\beta))$$

$$= B((e_u,e_v),(\alpha-\alpha_h,\beta-\beta_h)) \text{ for all } (\alpha_h,\beta_h) \in S_1^h \times S_2^h$$

$$\le \|(e_u,e_v)\|_B\ C\ (\|\alpha-\alpha_h\|_1 + \|\beta-\beta_h\|_1)$$

$$\le \|(e_u,e_v)\|_B\ C\ h\ (\|\alpha\|_2 + \|\beta\|_2) \text{ upon taking infimum}$$

$$\le \|(e_u,e_v)\|_B\ C\ h\ \|e_v\|_0\ , \text{ completing the proof.}$$

Proof of Lemma 4.13.

Write $\sigma = p^{1/2}\psi$ for convenience. We are trying to solve $\sigma' = r\phi$, $\sigma - p\phi' - q\phi = p^{1/2}g$ on $(0,1)$ with

$\phi(0) = \phi(1) = 0$ and $\|\sigma\|_2 + \|\phi\|_1 \le C\|e_v\|_0$. If we differentiate the second equation we get a second-order differential equation for ϕ:

$$-(p\phi')' - (q\phi)' + r\phi = (p^{1/2}g)', \text{ i.e., } L^*\phi = (p^{1/2}g)',$$
$$\phi(0) = \phi(1) = 0.$$

We know a unique solution ϕ exists; furthermore the inequality $\|\phi\|_1 \le C\|(p^{1/2}g)'\|_{-1}$ holds, where here the norm on the right-hand side is defined by duality with H_0^1,

$$\|w\|_{-1} = \sup_{\alpha \in H_0^1(0,1)} \frac{(w,\alpha)}{\|\alpha\|_1}.$$ (See [14] or [5].)

In particular,

$$\|(p^{1/2}g)'\|_{-1} = \sup_{\alpha \in H_0^1(0,1)} \frac{((p^{1/2}g)',\alpha)}{\|\alpha\|_1}$$

$$= \sup_{\alpha} \frac{-(p^{1/2}g,\alpha')}{\|\alpha\|_1}$$

$$\le C\|g\|_0 .$$

Hence $\|\phi\|_1 \le C\|g\|_0$. Now let $\sigma = p\phi' + q\phi + p^{1/2}g$; then $\|\sigma\|_0 \le C\|g\|_0$. Since $g \in H^1(0,1)$, ϕ is actually in $H^2(0,1)$, by virtue of $L^*\phi = (p^{1/2}g)'$. Hence we may differentiate the equation defining σ to find

$$\sigma' = (p\phi')' + (q\phi)' + (p^{1/2}g)' = r\phi \in H^1(0,1),$$

and $\|\sigma'\|_0 \le C\|\phi\|_0$, $\|\sigma''\|_0 = \|(r\phi)'\|_0 \le C\|\phi\|_1 \le C\|g\|_0$.

Hence $\sigma \in H^2(0,1)$ and $\|\sigma\|_2 \leq C \|g\|_0$, completing the proof.

Thus in one dimension both the solution u and the derivative $p^{1/2}u'$ are quasi-optimally approximated by the solution of the finite-dimensional problem.

One other word about Proposition 4.12 is in order. If one tries to carry the proof over to the two-dimensional case (assuming one forces the boundary conditions to hold) one encounters a problem in Lemma 4.13. In the two-dimensional case we have been unable to prove that $\underset{\sim}{\psi}$ is in $\underset{\sim}{H}^2(\Omega)$; instead we have $\underset{\sim}{\psi} \in \underset{\sim}{H}^1(\Omega)$ and $\operatorname{div} \underset{\sim}{\psi} \in H^1(\Omega)$, but these do not imply $\underset{\sim}{\psi} \in \underset{\sim}{H}^2(\Omega)$.

Finally, let us mention that the proof was for the splitting $-(p^{1/2}u') = v$, $(p^{1/2}v)' - qp^{-1/2}v + ru = f$, but the splitting $-u' = v$, $(pv)' - qv + ru = f$ can also be handled in the same manner, with the same result of quasi-optimal error bounds for both $\|e_u\|_0$ and $\|e_v\|_0$.

Other Boundary Conditions

In the preceding we always considered Dirichlet boundary conditions $u = g$ on Γ. What can be done with other boundary conditions? Let us consider the model problem

$$Lu \equiv -\Delta u + \underset{\sim}{b}\cdot\nabla u + cu = f \quad \text{in } \Omega$$

$$\frac{\partial u}{\partial n} = g \quad \text{on } \Gamma.$$

Assume as usual that the coefficients of L are smooth and that both L and its adjoint L* satisfy a uniqueness theorem ($Lu = 0$ in Ω, $\frac{\partial u}{\partial n} = 0$ on Γ implies $u \equiv 0$, and similarly for L* and the adjoint boundary conditions). The decomposition to consider is

$$\begin{aligned} -\nabla u &= \underset{\sim}{v} \quad \text{in } \Omega, \\ \operatorname{div} \underset{\sim}{v} - \underset{\sim}{b}\cdot\underset{\sim}{v} + cu &= f \quad \text{in } \Omega, \\ n\cdot\underset{\sim}{v} &= -g \text{ on } \Gamma. \end{aligned}$$

Define a bilinear form via

$$B((w,\underset{\sim}{z}),(\phi,\underset{\sim}{\psi})) = (\nabla w+\underset{\sim}{z},\nabla\phi+\underset{\sim}{\psi}) + (\operatorname{div}\underset{\sim}{z} - \underset{\sim}{b}\cdot\underset{\sim}{z}+cw, \operatorname{div}\underset{\sim}{\psi} - \underset{\sim}{b}\cdot\underset{\sim}{\psi}+c\phi) + \langle n\cdot\underset{\sim}{z}, n\cdot\underset{\sim}{\psi}\rangle .$$

Notice that now there is no weighting factor multiplying the boundary inner product.

Choose spaces $S_1^h \subset H^1(\Omega)$, $\underset{\sim}{S}_2^h \subset \underset{\sim}{H}^1(\Omega)$ as in Chapter III, and consider the problem

(4.8) Find $(u^h,\underset{\sim}{v}^h) \in S_1^h \times \underset{\sim}{S}_2^h$ such that

$$B((u^h,\underset{\sim}{v}^h),(\phi_h,\underset{\sim}{\psi}_h)) = B((u,\underset{\sim}{v}),(\phi_h,\underset{\sim}{\psi}_h)) = (f,\operatorname{div}\underset{\sim}{\psi}_h - \underset{\sim}{b}\cdot\underset{\sim}{\psi}_h+c\phi_h) - \langle g, n\cdot\underset{\sim}{\psi}_h\rangle$$

for all $(\phi_h,\underset{\sim}{\psi}_h) \in S_1^h \times \underset{\sim}{S}_2^h$.

For $\underset{\sim}{z} \in \underset{\sim}{H}^1(\Omega)$ we have

$$|n\cdot\underset{\sim}{z}|_0 \le |\underset{\sim}{z}|_0 \le C\,(\|\underset{\sim}{z}\|_0 + \|\underset{\sim}{z}\|_1) \text{ by Lemma 2.3.}$$

Hence, defining $\|(w,\underset{\sim}{z})\|_B^2 = B((w,\underset{\sim}{z}),(w,\underset{\sim}{z}))$, we have the bound $\|(w,\underset{\sim}{z})\|_B \leq C\ (\|w\|_1 + \|\underset{\sim}{z}\|_1)$. Thus we have

Proposition 4.14. A unique solution $(u^h,\underset{\sim}{v}^h)$ of (4.8) exists, and

$$\|(u-u^h,\underset{\sim}{v}-\underset{\sim}{v}^h)\|_B \leq C \inf_{(w_h,\underset{\sim}{z}_h)\,\in\, S_1^h \times \underset{\sim}{S}_2^h} (\|u-w_h\|_1+\|\underset{\sim}{v}-\underset{\sim}{z}_h\|_1).$$

The proof is easy, and will be omitted.

As usual, we are interested in bounds for $u-u^h$ and $\underset{\sim}{v}-\underset{\sim}{v}^h$ separately. For convenience, write $e_u = u-u^h$, $\underset{\sim}{e}_v = \underset{\sim}{v}-\underset{\sim}{v}^h$.

Theorem 4.15. For $0 \leq s \leq 1$, $\|e_u\|_s \leq C\, h^{1-s}\, \|(e_u,\underset{\sim}{e}_v)\|_B$.

Proof. We will prove the result for $s = 0$ and $s = 1$; then interpolating gives the theorem. All the following holds for both $s = 0$ and $s = 1$.

We have $\|e_u\|_s = \sup_{\substack{w\,\in\, H^{-s}(\Omega)\\ w\neq 0}} \dfrac{(e_u,w)}{\|w\|_{-s}}$.

Given $w \in H^{-s}(\Omega)$, let α solve $L^*\alpha = w$ in Ω, $\dfrac{\partial\alpha}{\partial n} + n\cdot(\underset{\sim}{b}\alpha) = 0$ on Γ (these are the adjoint boundary conditions). Then $\alpha \in \tilde{H}^{2-s,2}(\Omega)$ ($\tilde{H}^{2,2}(\Omega)$ is just $H^2(\Omega)$ with an equivalent norm), and $\|\alpha\|_{\tilde{H}^{2-s,2}} \leq C\,\|w\|_{-s}$.

Now,

$$(e_u,w) = (e_u,L^*\alpha)$$

$$= (e_u,-\Delta\alpha - \operatorname{div}(\underset{\sim}{b}\alpha) + c\alpha)$$

$$= (\nabla e_u, \nabla\alpha) - \langle e_u, N(\alpha)\rangle + (\nabla e_u, \underset{\sim}{b}\alpha) - \langle e_u, n\cdot(\underset{\sim}{b}\alpha)\rangle + (e_u, c\alpha),$$

where $N(\alpha) \in H^{-1/2}(\Gamma)$ if $\alpha \in \tilde{H}^{1,2}(\Omega)$, $N(\alpha) \in H^{1/2}(\Gamma)$ if $\alpha \in \tilde{H}^{2,2}(\Omega)$, and $N(\alpha) = \frac{\partial\alpha}{\partial n}$ for $\alpha \in C^\infty(\bar{\Omega})$. The last expression is equal to

$$(\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha) + (\operatorname{div} \underset{\sim}{e}_v, \alpha) - \langle n\cdot\underset{\sim}{e}_v, \alpha\rangle + (\nabla e_u + \underset{\sim}{e}_v, \underset{\sim}{b}\alpha) - (\underset{\sim}{e}_v, \underset{\sim}{b}\alpha) + (ce_u, \alpha)$$

$$= (\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha) + (\operatorname{div} \underset{\sim}{e}_v - \underset{\sim}{b}\cdot\underset{\sim}{e}_v + ce_u, \alpha) + (\nabla e_u + \underset{\sim}{e}_v, \underset{\sim}{b}\alpha) - \langle n\cdot\underset{\sim}{e}_v, \alpha\rangle .$$

We now bound these terms as in Theorem 3.5. Start with the second term. We can prove as in Lemma 3.4 that $\|\operatorname{div} \underset{\sim}{e}_v - \underset{\sim}{b}\cdot\underset{\sim}{e}_v + ce_u\|_{-t} \le C\, h^t \|(e_u, \underset{\sim}{e}_v)\|_B$ for $0 \le t \le \min(r_1 - 1, r_2 - 1)$. Hence the second term is bounded by $C\, h^{2-s} \|(e_u, \underset{\sim}{e}_v)\|_B \|\alpha\|_{2-s}$. For the first term, we have

$$(\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha) = (\nabla e_u + \underset{\sim}{e}_v, \nabla(\alpha - \alpha_h)) - (\operatorname{div} \underset{\sim}{e}_v - \underset{\sim}{b}\cdot\underset{\sim}{e}_v + ce_u, c\alpha_h)$$

$$= B((e_u, \underset{\sim}{e}_v), (\alpha - \alpha_h, 0)) - (\operatorname{div} \underset{\sim}{e}_v - \underset{\sim}{b}\cdot\underset{\sim}{e}_v + ce_u, c\alpha)$$

and hence $|(\nabla e_u + \underset{\sim}{e}_v, \nabla\alpha)| \le \|(e_u, \underset{\sim}{e}_v)\|_B\, C\, h^{1-s} \|\alpha\|_{2-s} + C\, h^{2-s} \|(e_u, \underset{\sim}{e}_v)\|_B \|c\alpha\|_{2-s}$

$$\le C\, h^{1-s} \|(e_u, \underset{\sim}{e}_v)\|_B \|\alpha\|_{2-s} .$$

For the third term, solve $\nabla\phi + \underset{\sim}{\psi} = \underset{\sim}{b}\alpha$ in Ω, $\operatorname{div} \underset{\sim}{\psi} - \underset{\sim}{b}\cdot\underset{\sim}{\psi} + c\phi = 0$ in Ω, $\frac{\partial\phi}{\partial n} = 0$ on Γ, i.e., solve $L\alpha = -\operatorname{div}(\underset{\sim}{b}\alpha) + |\underset{\sim}{b}|^2\alpha$ in Ω, $\frac{\partial\phi}{\partial n} = 0$ on Γ, and put $\underset{\sim}{\psi} = \underset{\sim}{b}\alpha - \nabla\phi$, and proceed as in Lemma 3.6.

For the fourth term, we have

$$|\langle n\cdot\underset{\sim}{e}_v,\alpha\rangle| \le |n\cdot\underset{\sim}{e}_v|_{s-3/2}\,|\alpha|_{-s+3/2} \le C|n\cdot\underset{\sim}{e}_v|_{s-3/2}\,\|\alpha\|_{2-s}.$$

Now,

$$|n\cdot\underset{\sim}{e}_v|_{s-3/2} = \sup_{\substack{\tau\in H^{-s+3/2}(\Gamma)\\ \tau\neq 0}} \frac{\langle n\cdot\underset{\sim}{e}_v,\tau\rangle}{|\tau|_{-s+3/2}}.$$

Given $\tau \in H^{-s+3/2}(\Gamma)$, let ϕ solve $L\phi = 0$ in Ω, $\frac{\partial\phi}{\partial n} = -\tau$ on Γ, and put $\underset{\sim}{\psi} = -\nabla\phi$. Then $\|\phi\|_{3-s} + \|\underset{\sim}{\psi}\|_{2-s} \le C\,|\tau|_{-s+3/2}$, and

$$\begin{aligned}
\langle n\cdot\underset{\sim}{e}_v,\tau\rangle &= \langle n\cdot\underset{\sim}{e}_v, n\cdot\underset{\sim}{\psi}\rangle \\
&= B((e_u,\underset{\sim}{e}_v),(\phi,\underset{\sim}{\psi})) \\
&= B((e_u,\underset{\sim}{e}_v),(\phi-\phi_h,\underset{\sim}{\psi}-\underset{\sim}{\psi}_h)) \\
&\le \|(e_u,\underset{\sim}{e}_v)\|_B\, C\,(\|\phi-\phi_h\|_1 + \|\underset{\sim}{\psi}-\underset{\sim}{\psi}_h\|_1) \\
&\le \|(e_u,\underset{\sim}{e}_v)\|_B\, C\, h^{1-s}\,|\tau|_{-s+3/2},
\end{aligned}$$

proving $|n\cdot\underset{\sim}{e}_v|_{s-3/2} \le C\,h^{1-s}\,\|(e_u,\underset{\sim}{e}_v)\|_B$. Putting all the pieces together proves the result.

Remarks: (1.) We did not prove the theorem directly because a certain pathology would have entered in the case s = 1/2 [4]. The a priori bound for α would have been affected; otherwise all the arguments would have gone through.

(2.) It is interesting that this result gives a bound for $\|e_u\|_1$ without any inverse hypothesis, whereas the problem with Dirichlet boundary conditions was unable to go beyond $\|e_u\|_{1/2}$ without an inverse hypothesis. As an immediate corollary we have

Corollary 4.16. $\|\underset{\sim}{e}_v\|_0 \leq C \|(e_u, \underset{\sim}{e}_v)\|_B$.

Proof. Write $\underset{\sim}{e}_v = (\nabla e_u + \underset{\sim}{e}_v) - \nabla e_u$ and take norms.

We will now study the condition number of the finite-dimensional problem (4.8). We first need an analog of Proposition 4.3.

Proposition 4.17. For all $(w, \underset{\sim}{z}) \in H^1(\Omega) \times \underset{\sim}{H}^1(\Omega)$,
$\|w\|_1 + \|\underset{\sim}{z}\|_0 \leq C \left(\|\nabla w + \underset{\sim}{z}\|_0 + \|\operatorname{div} \underset{\sim}{z} - \underset{\sim}{b} \cdot \underset{\sim}{z} + cw\|_{-1} + |n \cdot \underset{\sim}{z}|_{-1/2} \right)$.

Proof. We first bound $\|w\|_1$. Given $\phi \in H^{-1}(\Omega)$, let α solve $L^*\alpha = \phi$ in Ω, $\frac{\partial \alpha}{\partial n} + n \cdot (\underset{\sim}{b}\alpha) = 0$ on Γ. Then $\alpha \in \underset{\sim}{H}^{1,2}(\Omega)$ and $\|\alpha\|_{\underset{\sim}{H}^{1,2}(\Omega)} \leq C \|\phi\|_{-1}$.

Now, by the same calculation as in Theorem 4.15,

$$(w,\phi) = (w,L^*\alpha)$$

$$= (\nabla w+\underset{\sim}{z},\nabla\alpha+\underset{\sim}{b}\alpha) + (\operatorname{div}\underset{\sim}{z} - \underset{\sim}{b}\cdot\underset{\sim}{z} + cw,\alpha) - \langle n\cdot\underset{\sim}{z},\alpha\rangle \ .$$

Hence

$$|(w,\phi)| \le (\|\nabla w + \underset{\sim}{z}\|_0 + \|\operatorname{div}\underset{\sim}{z} - \underset{\sim}{b}\cdot\underset{\sim}{z} + cw\|_{L_1} + |n\cdot\underset{\sim}{z}|_{-1/2}) \, C\|\alpha\|_1$$

$$\le (\|\nabla w + \underset{\sim}{z}\|_0 + \|\operatorname{div}\underset{\sim}{z} - \underset{\sim}{b}\cdot\underset{\sim}{z} + cw\|_{L_1} + |n\cdot\underset{\sim}{z}|_{-1/2}) \, C\|\phi\|_{L_1}$$

since $\|\alpha\|_1 \le \|\alpha\|_{\underset{\sim}{H}^{1,2}(\Omega)} \le C\|\phi\|_{L_1}$.

This proves the result for w. For $\underset{\sim}{z}$, write $\underset{\sim}{z} = (\nabla w + \underset{\sim}{z}) - \nabla w$ and take norms.

Now we can prove

Theorem 4.18. Assume the basis functions for S_1^h and $\underset{\sim}{S}_2^h$ satisfy the hypothesis (4.6). Then the condition number of the matrix of the finite-dimensional problem (4.8) is $O(h^{-2})$.

Proof. The proof is the same as the proof of Theorem 4.8. It suffices to note that $\|\phi\|_{L_1} \le \|\phi\|_0$ for all $\phi \in H^0(\Omega)$ and $|\tau|_{-1/2} \le |\tau|_0$ for all $\tau \in H^0(\Gamma)$. Then the proof of Theorem 4.8 can be carried over word for word.

Thus for our model problem all the previous results carry over without the inverse hypotheses that were brought in before.

The Difference Equations

It is interesting to see exactly what difference equations are produced by the least squares decomposition method. To simplify matters we will consider only the simplest one-dimensional problem, $-u'' = f$ on $(0,1)$, $u(0) = u(1) = 0$. For n a positive integer let S_1^h, S_2^h be spaces of continuous piecewise linear functions defined on a uniform mesh of width $h = 1/(n+1)$. If we let $x_j = jh$ $(j = 0,1,\ldots,n+1)$ a convenient basis for S_1^h is given by $\{\phi_1,\ldots,\phi_n\}$ where

$$\phi_j(x) = \begin{cases} (x_j - x)/h & \text{if } x_{j-1} \leq x \leq x_j, \\ (x - x_j)/h & \text{if } x_j \leq x \leq x_{j+1}, \\ 0 & \text{otherwise .} \end{cases}$$

A basis for S_2^h is formed by adjoining to the basis for S_1^h the functions

$$\phi_0(x) = \begin{cases} (x_1 - x)/h & \text{if } 0 \leq x \leq x_1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\phi_{n+1}(x) = \begin{cases} (x - x_n)/h & \text{if } x_n \leq x \leq x_{n+1}, \\ 0 & \text{otherwise .} \end{cases}$$

The splitting to consider is $-u' = v$, $v' = f$, $u(0) = u(1) = 0$.

Write the solution u^h, v^h as

$$u^h(x) = \sum_{j=1}^{n} u_j\phi_j(x),\ v^h(x) = \sum_{j=0}^{n+1} v_j\phi_j(x).$$

Then the coefficients $u_1,\ldots,u_n,v_0,\ldots,v_{n+1}$ are determined by the equation

$$(u^{h'} + v^h,\alpha_h' + \beta_h) + (v^{h'},\beta_h') = (f,\beta_h') \text{ all } (\alpha_h,\beta_h),$$

i.e.,

$$(u^{h'} + v^h,\phi_j') = 0 \quad \text{for } j = 1,\ldots,n \quad \text{and}$$

$$(u^{h'} + v^h,\phi_j) + (v^{h'},\phi_j') = (f,\phi_j'),\ j = 0,\ldots,n+1.$$

These yield

$$\frac{1}{h}(2u_1 - u_2) + \frac{1}{2}(v_0 - v_2) = 0$$

$$\frac{1}{h}(-u_{j-1} + 2u_j - u_{j+1}) + \frac{1}{2}(v_{j-1} - v_{j+1}) = 0 \quad (2 \le j \le n-1)$$

$$\frac{1}{h}(-u_{n-1} + 2u_n) + \frac{1}{2}(v_{n-1} - v_{n+1}) = 0$$

and

$$\frac{1}{2}u_1 + \frac{h}{6}(2v_0 + v_1) + \frac{1}{h}(v_0 - v_1) = (f,\phi_0')$$

$$\frac{1}{2}u_2 + \frac{h}{6}(v_0 + 4v_1 + v_2) + \frac{1}{h}(-v_0 + 2v_1 - v_2) = (f,\phi_1')$$

$$\frac{1}{2}(-u_{j-1}+u_{j+1})+\frac{h}{6}(v_{j-1}+4v_j+v_{j+1})+\frac{1}{h}(-v_{j-1}+2v_j-v_{j+1}) = (f,\phi_j')$$

$$\text{for } j = 2,\ldots,n-1$$

$$-\frac{1}{2}u_{n-1} + \frac{h}{6}(v_{n-1} + 4v_n + v_{n+1}) + \frac{1}{h}(-v_{n-1} + 2v_n + v_{n+1}) = (f,\phi_n')$$

$$-\frac{1}{2}u_n + \frac{h}{6}(v_n + 2v_{n+1}) + \frac{1}{h}(-v_n + v_{n+1}) = (f,\phi_{n+1}') \ .$$

The first group of equations here is a finite difference approximation to the equation $-u'' = v'$, while the second group is an approximation to $u' + v - v'' = -f'$. The fact that these differential equations have higher order than the equations $-u' = v$, $v' = f$, is characteristic of the least squares method, and is reflected in the fact that the error bounds involve higher norms on f than one would think necessary (cp. the comments at the beginning of this chapter). With this we conclude Chapter IV.

CHAPTER V. A GENUINE ELLIPTIC SYSTEM

We have concentrated on second order problems in the preceding because many applications involve such problems. The extension of the least squares decomposition method to higher order problems is unwieldy, and analyzing it is probably not worth the effort. Also, the physical meaning of the various partial derivatives of the solution of a higher-order elliptic problem may be obscure. Thus we feel that extending the previous analysis to higher-order equations is of little interest.

The purpose of this chapter is to analyze a genuine system of partial differential equations. The least squares decomposition method of the preceding chapters began with a single equation and wrote it as a system of equations by introducing new independent variables, a somewhat artificial procedure in some cases. In this chapter we shall analyze a least squares method applied to a genuine system of partial differential equations; there will be no decomposition step. The essential theoretical tool will be the theory of elliptic systems of partial differential equations due to Agmon, Douglis, and Nirenberg [2]. This paper will be referred to in what follows as ADN.

We want to work with the equations of two-dimensional elasticity, the equations governing the small displacement behavior of plane stress and plane strain problems. These equations can be written as

$$\partial_1\sigma_x + \partial_2\tau_{xy} = -X$$
$$\partial_1\tau_{xy} + \partial_2\sigma_y = -Y$$
$$E\partial_1 u = \sigma_x - \nu\sigma_y$$
$$E\partial_2 v = \sigma_y - \nu\sigma_x$$
$$E(\partial_2 u + \partial_1 v) = 2(1 + \nu)\tau_{xy} .$$

Here u and v are displacements in the x and y directions, respectively; σ_x, τ_{xy}, σ_y are the stress components; $\partial_1 = \partial/\partial x$ and $\partial_2 = \partial/\partial y$; E is Young's modulus, ν is Poisson's ratio (both constants); and X and Y are body forces. The first two equations are the "equilibrium" equations, the last three the "constitutive" equations. There are two common sets of boundary conditions which may be employed. First, we may specify u and v on the boundary. Second, we may specify $n_1\sigma_x + n_2\tau_{xy}$ and $n_1\tau_{xy} + n_2\sigma_y$ on the boundary, where $n = (n_1,n_2) = (n_1(x,y),n_2(x,y))$ is the outward normal.

We now rewrite the problem in a slightly more convenient form. Define new variables by $u_1 = Eu$, $u_2 = Ev$, $u_3 = \sigma_x$, $u_4 = \tau_{xy}$, $u_5 = \sigma_y$. Then the above equations

can be written as a matrix system $Lu = f$ where $u = (u_1,u_2,u_3,u_4,u_5)$, $f = (f_1,f_2,f_3,f_4,f_5) = (-X,-Y,0,0,0)$ and L is a matrix with operator entries,

$$(5.1) \qquad L = \begin{pmatrix} 0 & 0 & \partial_1 & \partial_2 & 0 \\ 0 & 0 & 0 & \partial_1 & \partial_2 \\ \partial_1 & 0 & -1 & 0 & \nu \\ \partial_2 & \partial_1 & 0 & -2(1+\nu) & 0 \\ 0 & \partial_2 & \nu & 0 & -1 \end{pmatrix}.$$

The equation $Lu = f$ is to hold in a bounded region $\Omega \subset R^2$ with smooth boundary Γ and outward normal $n = (n_1,n_2)$. We will consider (to ensure uniqueness of solutions) the case of specified displacements on the boundary, so the boundary conditions are $u_1 = g_1$, $u_2 = g_2$ on Γ.

The first step in verifying the conditions of ADN for this system is to choose constants $s_1,\ldots,s_5$ and $t_1,\ldots,t_5$ such that degree$(L_{ij}) \leq s_i + t_j$, where L_{ij} is the operator in the (i,j) position of L. We choose $s_1 = s_2 = 0$, $s_3 = s_4 = s_5 = -1$, $t_1 = t_2 = 2$, $t_3 = t_4 = t_5 = 1$, and the condition is satisfied.

Now let L'_{ij} be the principal part of L_{ij}, the terms which are exactly of degree $s_i + t_j$. We must check that det $L'(\xi)$ is nonzero for all nonzero vectors $\xi = (\xi_1,\xi_2)$, where L' has (i,j) entry L'_{ij} and $L'(\xi)$ is the result of substituting ξ_1 and ξ_2 for ∂_1,∂_2, respect-

ively. If we perform the computation we find $\det L'(\xi) = -(\xi_1^2 + \xi_2^2)^2$, so this condition is satisfied.

Next we must check the "supplementary condition" on L. We find the roots (in τ) of the equation $\det L'(\xi + \tau\xi') = 0$ are given by

$$\tau = \frac{-\xi\cdot\xi' \pm i(|\xi|^2|\xi'|^2 - |\xi\cdot\xi'|^2)^{1/2}}{|\xi'|^2},$$

so there are exactly two roots (both equal) with positive imaginary part.

We have now verified the system (5.1) is elliptic in the sense of ADN. Now we must consider the boundary conditions. They can be written $Bu = g$ on Γ where

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

and $g = (g_1, g_2, 0, 0, 0)$. First we have to find constants ρ_1, ρ_2 such that $\text{degree}(B_{ij}) \leq \rho_i + t_j$ and $B_{ij} = 0$ if $\rho_i + t_j < 0$. We take $\rho_1 = \rho_2 = -2$ and this is satisfied. Now let ξ be a tangent vector to Γ and compute the roots of the equation $\det L'(\xi + \tau n) = 0$; they are easily found to be $\tau_+ = i|\xi|$, $\tau_- = -i|\xi|$. Then define $M^+(\xi,\tau) = \tau - \tau_+ = \tau - i|\xi|$. The boundary conditions are <u>complementing</u> if the rows of the matrix $B(L'(\xi + \tau n))^T$, considered as polynomials in τ, are

linearly independent modulo $M^+(\xi,\tau)$. In our case we have

$$B(L'(\xi + \tau n))^T = \begin{pmatrix} 0 & 0 & \xi_1 + \tau n_1 & \xi_2 + \tau n_2 & 0 \\ 0 & 0 & 0 & \xi_1 + \tau n_1 & \xi_2 + \tau n_2 \end{pmatrix},$$

and the complementing condition is satisfied.

Finally it is known that a unique solution to the system $Lu = f$ in Ω, $Bu = g$ on Γ exists [21]. Then the ADN theory gives the following a priori estimate.

(5.2) For any $k \geq 0$,

$$\|u_1\|_{k+2} + \|u_2\|_{k+2} + \|u_3\|_{k+1} + \|u_4\|_{k+1} + \|u_5\|_{k+1}$$
$$\leq C \left(\|f_1\|_k + \|f_2\|_k + \|f_3\|_{k+1} + \|f_4\|_{k+1} + \|f_5\|_{k+1} \right.$$
$$\left. + |g_1|_{k+3/2} + |g_2|_{k+3/2} \right).$$

By the symmetry of the operator matrix L a similar result holds for the adjoint problem. We also remark that we have the Green's formula

$$(Lu,v) - (u,L^*v) = \langle u_1, n_1 v_3 + n_2 v_4 \rangle + \langle u_2, n_1 v_4 + n_2 v_5 \rangle$$
$$+ \langle n_1 u_3 + n_2 u_4, v_1 \rangle + \langle n_1 u_4 + n_2 u_5, v_2 \rangle .$$

Here L^* is the formal adjoint of L, defined by $(Lu,v) = (u,L^*v)$ for all $u,v \in (C_0^\infty(\Omega))^5$ and the inner products in the equation defining L^* are taken in $(L_2(\Omega))^5$.

Now we are ready to define a least squares approximation scheme. Choose finite-dimensional spaces $S_1^h = S_1^h(\Omega) \subset H^1(\Omega)$ and $S_2^h = S_2^h(\Omega) \subset H^1(\Omega)$. We will seek an approximate solution in the space $S^h \equiv S_1^h \times S_1^h \times S_2^h \times S_2^h \times S_2^h$.

We will minimize the following quadratic functional over S^h.

$$\begin{aligned} J_h(u_1,u_2,u_3,u_4,u_5) &\equiv h^2 \| \partial_1 u_3 + \partial_2 u_4 - f_1 \|_0^2 \\ &+ h^2 \| \partial_1 u_4 + \partial_2 u_5 - f_2 \|_0^2 + \| \partial_1 u_1 - u_3 + \nu u_5 - f_3 \|_0^2 \\ &+ \| \partial_2 u_1 + \partial_1 u_2 - 2(1+\nu) u_4 - f_4 \|_0^2 \\ &+ \| \partial_2 u_2 + \nu u_3 - u_5 - f_5 \|_0^2 + h^{-1} | u_1 - g_1 |_0^2 + h^{-1} | u_2 - g_2 |_0^2 . \end{aligned}$$

We are led to define a bilinear form on $(H^1(\Omega))^5 \times (H^1(\Omega))^5$ by

$$\begin{aligned} (5.3)\quad & B((u_1,u_2,u_3,u_4,u_5),(\phi_1,\phi_2,\phi_3,\phi_4,\phi_5)) \\ &= h^2 (\partial_1 u_3 + \partial_2 u_4, \partial_1 \phi_3 + \partial_2 \phi_4) \\ &+ h^2 (\partial_1 u_4 + \partial_2 u_5, \partial_1 \phi_4 + \partial_2 \phi_5) \\ &+ (\partial_1 u_1 - u_3 + \nu u_5, \partial_1 \phi_1 - \phi_3 + \nu \phi_5) \\ &+ (\partial_2 u_1 + \partial_1 u_2 - 2(1+\nu) u_4, \partial_2 \phi_1 + \partial_1 \phi_2 - 2(1+\nu) \phi_4) \\ &+ (\partial_2 u_2 + \nu u_3 - u_5, \partial_2 \phi_2 + \nu \phi_3 - \phi_5) \\ &+ h^{-1} \langle u_1, \phi_1 \rangle + h^{-1} \langle u_2, \phi_2 \rangle . \end{aligned}$$

Then the finite-dimensional problem is

(5.4) Find $u^h = (u_1^h, u_2^h, u_3^h, u_4^h, u_5^h) \in S^h$ such that

$$B(u^h, \phi_h) = B(u, \phi_h) \text{ for all } \phi_h \in S^h.$$

Define a norm by $\|w\|_B^2 = B(w,w)$ for $w \in (H^1(\Omega))^5$. Then we have

Proposition 5.1. A unique solution u^h to (5.4) exists, and

$$\begin{aligned}\|u-u^h\|_B &= \inf_{\phi^h \in S^h} \|u-\phi^h\|_B \\ &\le C \inf_{\phi^h \in S^h} \big(\|u_1-\phi_1^h\|_1 + h^{-1} \|u_1-\phi_1^h\|_0 \\ &\quad + \|u_2-\phi_2^h\|_1 + h^{-1} \|u_2-\phi_2^h\|_0 + \|u_3^h-\phi_3^h\|_0 \\ &\quad + h \|u_3^h-\phi_3^h\|_1 + \|u_4^h-\phi_4^h\|_0 + h \|u_4^h-\phi_4^h\|_1 \\ &\quad + \|u_5^h-\phi_5^h\|_0 + h \|u_5^h-\phi_5^h\|_1 \big).\end{aligned}$$

The proof is similar to previous proofs, and will be omitted.

We now want to get error bounds on the separate components of $u - u^h$. Again we make the approximation hypothesis (3.9): There are integers $r_i \ge 1$ $(i = 1,2)$ such that if $w \in H^s(\Omega)$, $1 \le s \le r_i$, then there exists $w_i^h \in S_i^h$ with $\|w-w_i^h\|_0 + h \|w-w_i^h\|_1 \le C h^s \|w\|_s$, where $C > 0$ is independent of w and h. Then by Proposition 5.1,

$$\|u-u^h\|_B \le C \left(h^{s_1-1} \|u\|_{s_1} + h^{s_2-1} \|u\|_{s_2} + h^{s_3} \|u\|_{s_3} + h^{s_4} \|u_4\|_{s_4} + h^{s_5} \|u_5\|_{s_5} \right)$$

if $u_i \in H^{s_i}(\Omega)$, $1 \le s_1, s_2 \le r_1$, $1 \le s_3, s_4, s_5 \le r_2$.
We make the further assumption that $r_1 \ge 4$ and $r_2 = r_1 - 1$.

Write $u - u^h = e = (e_1, \ldots, e_5)$. Then we have

Lemma 5.2. Under the above hypotheses,

(a) $|e_1|_{-1/2} + |e_2|_{-1/2} \le C\, h\, \|e\|_B$

(b) $\|\partial_1 e_3 + \partial_2 e_4\|_{L_s} + \|\partial_1 e_4 + \partial_2 e_5\|_{L_s} \le C\, h^{s-1} \|e\|_B$

for $1 \le s \le 2$

(c) $\|\partial_1 e_1 - e_3 + \nu e_5\|_{L_t} + \|\partial_2 e_1 - \partial_1 e_2 - 2(1+\nu) e_4\|_{L_t}$

$+ \|\partial_2 e_2 + \nu e_3 - e_5\|_{L_t} \le C\, h^t \|e\|_B$ for $0 \le t \le 1$.

Proof. Begin by noticing that if $w_1, w_2 \in H^{s+1}(\Omega)$, $w_3, w_4, w_5 \in H^s(\Omega)$, $1 \le s \le r_1 - 1$, then

$$\inf_{w^h \in S^h} \|w - w^h\|_B \le C\, h^s \left(\|w_1\|_{s+1} + \|w_2\|_{s+1} + \|w_3\|_s + \|w_4\|_s + \|w_5\|_s \right)$$

by the approximation hypothesis. Now we proceed to the separate proofs.

(a) Given $\tau \in H^{1/2}(\Gamma)$, let $\alpha = \ell(\tau)$ (see Theorem 2.1), so $\alpha \in H^1(\Omega)$, $\gamma_0 \alpha = \tau$ and $\|\alpha\|_1 \le C\, |\tau|_{1/2}$. Then

$$\langle e_1, \tau \rangle = h \left(h^{-1} \langle e_1, \alpha \rangle \right)$$

$$= h\,(\,h^{-1}\,\langle e_1, \alpha-\alpha_h\rangle - (\partial_1 e_1 - e_3 + \nu e_5, \partial_1\alpha_h)$$
$$- (\partial_2 e_1 + \partial_1 e_2 - 2(1+\nu)e_4, \partial_2\alpha_h)) \quad \text{for all } \alpha_h \in S_1^h,$$
since $B(e,(\alpha_h,0,0,0,0)) = 0$

$$\leq h\,\|e\|_B\, C\,(\,h^{-1}\,\|\alpha-\alpha_h\|_0 + \|\alpha-\alpha_h\|_1 + \|\alpha_h\|_1)$$

$$\leq C\,h\,\|e\|_B\|\alpha\|_1, \text{ upon taking the infimum over } \alpha_h$$

$$\leq C\,h\,\|e\|_B\,|\tau|_{1/2},$$

so $|e_1|_{-1/2} \leq C\,h\,\|e\|_B$. The proof for e_2 is similar.

(b) Given $\phi \in H^s(\Omega)$, let $\alpha = (\alpha_1,\ldots,\alpha_5)$ solve $L\alpha = (\phi,0,0,0,0)$ in Ω, $\alpha_1 = \alpha_2 = 0$ on Γ. Then
$\|\alpha_1\|_{s+2} + \|\alpha_2\|_{s+2} + \|\alpha_3\|_{s+1} + \|\alpha_4\|_{s+1} + \|\alpha_5\|_{s+1} \leq C\,\|\phi\|_s$
by (5.2), and

$$(\partial_1 e_3 + \partial_2 e_4, \phi) = h^{-2}\,(\,h^2\,(\partial_1 e_3 + \partial_2 e_4, \partial_1\alpha_3 + \partial_2\alpha_4))$$
$$= h^{-2}\,B(e,\alpha)$$
$$= h^{-2}\,B(e,\alpha-\alpha^h) \quad \text{for all } \alpha^h \in S^h$$
$$\leq h^{-2}\,\|e\|_B\,\|\alpha-\alpha^h\|_B$$
$$\leq h^{-2}\,\|e\|_B\,C\,h^{s+1}\,(\,\|\alpha_1\|_{s+2} + \|\alpha_2\|_{s+2}$$
$$+ \|\alpha_3\|_{s+1} + \|\alpha_4\|_{s+1} + \|\alpha_5\|_{s+1})$$
using the remark at the beginning of the proof and the hypothesis $r_1 \geq 4$
$$\leq C\,h^{s-1}\,\|e\|_B\,\|\phi\|_s, \text{ as desired.}$$

The proof for $\partial_1 e_4 + \partial_2 e_5$ is similar.

(c) Given $\phi \in H^t(\Omega)$, let $\alpha = (\alpha_1,\ldots,\alpha_5)$ solve $L\alpha = (0,0,\phi,0,0)$ in Ω, $\alpha_1 = \alpha_2 = 0$ on Γ. Then $\|\alpha_1\|_{t+1} + \|\alpha_2\|_{t+1} + \|\alpha_3\|_t + \|\alpha_4\|_t + \|\alpha_5\|_t \le C\|\phi\|_t$, and

$$\begin{aligned}(\partial_1 e_1 - e_3 + \nu e_5, \phi) &= B(e,\alpha)\\ &= B(e,\alpha-\alpha^h) \text{ for all } \alpha^h \in S^h\\ &\le \|e\|_B \|\alpha-\alpha^h\|_B\\ &\le \|e\|_B\, C\, h^t\, (\|\alpha_1\|_{t+1} + \|\alpha_2\|_{t+1}\\ &\quad + \|\alpha_3\|_t + \|\alpha_4\|_t + \|\alpha_5\|_t)\\ &\le C\, h^t \|e\|_B \|\phi\|_t,\end{aligned}$$

as desired. The proof for the other terms is similar.

Theorem 5.3. Under the above hypotheses,

$$\|e_1\|_0 + \|e_2\|_0 \le C\, h\, \|e\|_B.$$

Proof. Begin by solving the auxiliary problem $L^*w = (e_1,e_2,0,0,0)$ in Ω, $w_1 = w_2 = 0$ on Γ. By the ADN theory a unique solution w exists and we have the a priori bound

$$\|w_1\|_2 + \|w_2\|_2 + \|w_3\|_1 + \|w_4\|_1 + \|w_5\|_1 \le C\,(\|e_1\|_0 + \|e_2\|_0).$$

Now,

$$\|e_1\|_0^2 + \|e_2\|_0^2 = (e,L^*w)$$

$$= (Le,w) - \langle e_1, n_1 w_3 + n_2 w_4 \rangle - \langle e_2, n_1 w_4 + n_2 w_5 \rangle$$

$$= (\partial_1 e_3 + \partial_2 e_4, w_1) + (\partial_1 e_4 + \partial_2 e_5, w_2)$$
$$+ (\partial_1 e_1 - e_3 + \nu e_5, w_3) + (\partial_2 e_1 - \partial_1 e_2 - 2(1+\nu) e_4, w_4)$$
$$+ (\partial_2 e_2 + \nu e_3 - e_5, w_5) - \langle e_1, n_1 w_3 + n_2 w_4 \rangle$$
$$- \langle e_2, n_1 w_4 + n_2 w_5 \rangle$$

$$\leq C\, h\, \|e\|_B \, (\|w_1\|_2 + \|w_2\|_2)$$
$$+ C\, h\, \|e\|_B \, (\|w_3\|_1 + \|w_4\|_1 + \|w_5\|_1)$$
$$+ C\, h\, \|e\|_B \, (|n_1 w_3 + n_2 w_4|_{1/2} + |n_1 w_4 + n_2 w_5|_{1/2})$$

by Lemma 5.2

$$\leq C\, h\, \|e\|_B \, (\|e_1\|_0 + \|e_2\|_0), \text{ completing the proof.}$$

Thus the L_2 errors in the displacements, $\|u_1 - u_1^h\|_0$ and $\|u_2 - u_2^h\|_0$ are of optimal order in h, provided $r_1 \geq 4$ and $r_2 = r_1 - 1$. This condition has been seen before, in Theorem 4.1. It is a sufficient condition for optimality; whether or not it is necessary is unclear.

We now turn to the question of bounds for the other components of e. An inverse hypothesis will come in here, as in Chapter IV. We will assume the boundary inverse hypothesis (4.5):

$|w_h|_{1/2} \leq C\, h^{-1/2}\, |w_h|_0$ as $h \to 0$ for all $w_h \in S_1^h$, where $C > 0$ is independent of h and w_h.

Before proceeding we need to take note of an extension of the ADN theory due to Roitberg and Seftel' [20].

They proved, for a certain class of ADN-elliptic systems, a priori estimates of the type (5.2) for all k, not just nonnegative k. Their hypotheses are not satisfied by the system (5.1) but an analysis of their proof reveals it still is valid for (5.1), for reasons having to do with the symmetry of the operator matrix L and the interplay between the matrices L and B. (To be more specific would involve going into too many technical details.) Define the function space $\tilde{H}^{0,1}(\Omega)$ as the completion of $C^{\infty}(\bar{\Omega})$ in the norm

$$\|u\|^2_{\tilde{H}^{0,1}(\Omega)} = \|u\|^2_0 + |u|^2_{-1/2} .$$

Then their extension of the ADN theory gives the following result for our system (5.1).

(5.5) If $f_1, f_2 \in H^{-1}(\Omega)$, $f_3, f_4, f_5 \in H^0(\Omega)$, there exists a unique solution w to the problem $Lw = f$ in Ω, $w_1 = w_2 = 0$ on Γ, we have $w_1, w_2 \in H^1(\Omega)$, $w_3, w_4, w_5 \in \tilde{H}^{0,1}(\Omega)$, and we have the a priori bound

$$\|w_1\|_1 + \|w_2\|_1 + \|w_3\|_{\tilde{H}^{0,1}(\Omega)} + \|w_4\|_{\tilde{H}^{0,1}(\Omega)} + \|w_5\|_{\tilde{H}^{0,1}(\Omega)} \leq C \left(\|f_1\|_{-1} + \|f_2\|_{-1} + \|f_3\|_0 + \|f_4\|_0 + \|f_5\|_0 \right).$$

By the symmetry of L, a similar result holds for the problem $L^*w = f$ in Ω, $w_1 = w_2 = 0$ on Γ.

With this, we can prove

Theorem 5.4. Assuming the boundary inverse hypothesis (4.5), we have the error bound

$$(5.6)\qquad \|e_1\|_1 + \|e_2\|_1 + \|e_3\|_0 + \|e_4\|_0 + \|e_5\|_0$$
$$\le C \Big(\|e\|_B + \inf_{v_h, w_h \in S_1^h} \big(\|u_1 - v_h\|_1 + h^{-1} \|u_1 - v_h\|_0 + \|u_2 - w_h\|_1 + h^{-1} \|u_2 - w_h\|_0 \big) \Big).$$

Proof. Begin by noting that the calculations of Proposition 4.5 can be repeated to give

$|e_1|_{1/2} + |e_2|_{1/2} \le$ the right hand side of (5.5),

assuming the boundary inverse hypothesis (4.5).

We start with e_1. Given $\phi \in H^{-1}(\Omega)$, let w solve $L^*w = (\phi,0,0,0,0)$ in Ω, $w_1 = w_2 = 0$ on Γ. Then

$$(e_1,\phi) = (e,L^*w)$$
$$= (Le,w) - \langle e_1, n_1 w_3 + n_2 w_4 \rangle + \langle e_2, n_1 w_4 + n_2 w_5 \rangle$$
$$\le \|\partial_1 e_3 + \partial_2 e_4\|_{L_1} \|w_1\|_1 + \|\partial_1 e_4 + \partial_2 e_5\|_{L_1} \|w_2\|_1$$
$$+ \|\partial_1 e_1 - e_3 + \nu e_5\|_0 \|w_3\|_0$$
$$+ \|\partial_2 e_1 + \partial_1 e_2 - 2(1+\nu) e_4\|_0 \|w_4\|_0$$
$$+ \|\partial_2 e_2 + \nu e_3 - e_5\|_0 \|w_5\|_0$$
$$+ \big(|e_1|_{1/2} + |e_2|_{1/2} \big) C \big(\|w_3\|_{\tilde{H}^{0,1}(\Omega)} + \|w_4\|_{\tilde{H}^{0,1}(\Omega)} + \|w_5\|_{\tilde{H}^{0,1}(\Omega)} \big)$$
$$\le C \big(\|e\|_B + |e_1|_{1/2} + |e_2|_{1/2} \big) \|\phi\|_{-1},$$

by Lemma 5.2 and the a priori bound for w.

This gives the inequality for e_1. The proof for e_2 is similar.

For e_3, let w solve $L^*w = (0,0,e_3,0,0)$ in Ω, $w_1 = w_2 = 0$ on Γ. Then (5.5) holds for w, and

$$\|e_3\|_0^2 = (e,L^*w)$$

$$= (Le,w) - \langle e_1, n_1 w_3 + n_2 w_4 \rangle + \langle e_2, n_1 w_4 + n_2 w_5 \rangle$$

$$\leq C \left(\|e\|_B + |e_1|_{1/2} + |e_2|_{1/2} \right) \|e_3\|_0$$

by the same computations as above. The proofs for e_4 and e_5 are similar.

Thus if the boundary inverse hypothesis (4.5) and the approximation hypothesis (3.9) with $r_1 \geq 4$ hold, then the errors $\|e_i\|_0$ $(i = 1,\ldots,5)$ are all of optimal order in h.

We conclude this chapter by mentioning possible directions for future work. The error bounds of Chapters III and IV were generally of optimal order for the solution $u - u^h$ but of suboptimal order in the derivatives $\underset{\sim}{v} - \underset{\sim}{v}^h$; are these bounds sharp? Numerical experimentation is possible here, as a guide to the true state of affairs, but will have to be carried out on a fully two-dimensional problem, in light of Proposition 4.12. This thesis has not touched on the crucial matter of efficient methods for solving the finite-dimen-

sional problems that arise. Such considerations are very important for practical implementation of numerical schemes. Finally, the extension of decomposition and mixed methods to nonlinear equations and non-elliptic equations is an inviting area for future research.

BIBLIOGRAPHY

1. S. Agmon, Lectures on Elliptic Boundary Value Problems, Van Nostrand, Princeton, N. J., 1965.

2. S. Agmon, A. Douglis, and L. Nirenberg, "Estimates Near the Boundary for Solutions of Elliptic Partial Differential Equations Satisfying General Boundary Conditions II," Comm. Pure Appl. Math. 17 (1964), pp. 35-92.

3. Jean-Pierre Aubin, Approximation of Elliptic Boundary-Value Problems, Wiley-Interscience, New York, 1972.

4. Ivo Babuska 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," ed. by A. K. Aziz, Academic Press, New York, 1972.

5. Ju. M. Berezanskii, Expansions in Eigenfunctions of Selfadjoint Operators, Translations of Mathematical Monographs 17, American Mathematical Society, Providence, R. I., 1968.

6. J. H. Bramble and J. A. Nitsche, "A Generalized Ritz-Least-Squares Method for Dirichlet Problems," SIAM J. Numer. Anal. 10 (1973), pp. 81-93.

7. J. H. Bramble and A. H. Schatz, "Least Squares Methods for 2mth Order Elliptic Boundary Value Problems," Math. Comp. 25 (1971), pp. 1-32.

8. J. H. Bramble and V. Thomee, "Semidiscrete Least Squares Methods for a Parabolic Boundary Value Problem," Math. Comp. 26 (1972), pp. 633-648.

9. J. H. Bramble and M. Zlamal, "Triangular Elements in the Finite Element Method," Math. Comp. 24 (1970), pp. 809-821.

10. F. Brezzi, "On the Existence, Uniqueness, and Approximation of Saddle-Point Problems Arising from Lagrangian Multipliers," R.A.I.R.O., ser.3, 1974, pp. 129-151.

11. P. G. Ciarlet and P.-A. Raviart, "General Lagrange and Hermite Interpolation in R^n with Applications to Finite Element Methods," Arch. Rational Mech. Anal. 46 (1972), pp. 177-199.

12. G. E. Forsythe and W. R. Wasow, Finite Difference Methods for Partial Differential Equations, John Wiley, New York, 1960.

13. George Forsythe and Cleve B. Moler, Computer Solution of Linear Algebraic Systems, Prentice-Hall, Englewood Cliffs, N. J., 1967.

14. J. L. Lions and E. Magenes, Non-Homogeneous Boundary Value Problems and Applications I, Springer-Verlag, New York, 1972.

15. Paul Lynn and Santosh Arya, "Use of the Least Squares Criterion in the Finite Element Formulation," Int. J. Num. Meth. Engr. 6 (1973), pp. 75-88.

16. J. A. Nitsche, "Ein Kriterium für die Quasi-Optimalität des Ritsched Verfahrens," Numer. Math. 11 (1968), pp. 346-348.

17. P.-A. Raviart and J. M. Thomas, "Primal Hybrid Finite Element Methods for Second Order Elliptic Problems," Centre de Mathematiques Appliquees, Rapport Interne No. 2, 1975.

18. J. N. Reddy and J. T. Oden, "Mixed Finite Element Approximations of Linear Boundary-Value Problems," Quart. Appl. Math. 33 (1975), pp. 255-280.

19. Robert D. Richtmyer and K. W. Morton, Difference Methods for Initial Value Problems, Wiley-Interscience, New York, 1967.

20. Ja. A. Roitberg and Z. G. Seftel', "A Theorem on Homeomorphisms for Elliptic Systems and its Applications," Mat. Sb. 78 (1964), translated in Math. USSR Sb. 7 (1969), pp. 439-465.

21. I. S. Sokolnikoff, Mathematical Theory of Elasticity, McGraw-Hill, New York, 1956.

22. Gilbert Strang and George J. Fix, An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, N. J., 1973.

23. R. S. Varga, "Functional Analysis and Approximation Theory in Numerical Analysis," Regional Conference Series in Applied Mathematics 3, SIAM, Philadelphia, Pennsylvalia, 1971.

24. O. C. Zienkiewicz, The Finite Element Method in Engineering Science, McGraw-Hill, New York, 1971.