

A LEAST-SQUARES APPROACH BASED ON A DISCRETE MINUS ONE INNER PRODUCT FOR FIRST ORDER SYSTEMS

JAMES H. BRAMBLE, RAYTCHO D. LAZAROV, AND JOSEPH E. PASCIAK

June 1994

ABSTRACT. The purpose of this paper is to develop and analyze a least-squares approximation to a first order system. The first order system represents a reformulation of a second order elliptic boundary value problem which may be indefinite and/or nonsymmetric. The approach taken here is novel in that the least-squares functional employed involves a discrete inner product which is related to the inner product in $H^{-1}(\Omega)$ (the Sobolev space of order minus one on Ω). The use of this inner product results in a method of approximation which is optimal with respect to the required regularity as well as the order of approximation even when applied to problems with low regularity solutions. In addition, the discrete system of equations which needs to be solved in order to compute the resulting approximation is easily preconditioned, thus providing an efficient method for solving the algebraic equations. The preconditioner for this discrete system only requires the construction of preconditioners for standard second order problems, a task which is well understood.

1. INTRODUCTION.

Substantial progress in the finite element methods and in the solution techniques for solving the corresponding systems of algebraic equations in the last three decades has resulted in the development of mathematical formulations that introduce physically meaningful quantities as new dependent variables (fluxes, velocity, vorticity, strains and stresses, etc.). These problems can be posed in a weak sense and approximated by finite element methods. In many cases (for example, Stokes equations), this procedure leads to a saddle point problem. Due largely to Babuška [3] and Brezzi [10], it is now well understood that the finite element spaces approximating different physical quantities (pressure and velocity, or temperature and flux, or displacement and stresses, etc.) cannot be chosen independently and have to satisfy the so-called inf-sup condition of Ladyzhenskaya-Babuška-Brezzi [25], [3], [10]. Although substantial progress in approximation and solution methods for saddle point problems has been achieved, these problems may still be difficult and expensive to solve.

1991 *Mathematics Subject Classification.* Primary 65N30; Secondary 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS-9007185 and by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University.

In recent years there has been significant interest in least-squares methods, considered as an alternative to the saddle point formulations and circumventing the inf-sup condition. Examples of application of the least-squares to potential flows, convection-diffusion problems, Stokes and Navier-Stokes equations can be found in [8], [9], [16], [17], [20], [21], [23], [28]. In general, the corresponding problem is written as a system of partial differential equations of first order with possibly additional compatibility conditions. For example, $-\nabla \cdot \mathbf{u} = f$, $\mathbf{u} = \nabla p$ provides a first order system for the Poisson equation $-\Delta p = f$ which can be augmented by the compatibility equation $\mathbf{curl} \mathbf{u} = \mathbf{0}$ (see [24], [28]). Alternatively, the system $\mathbf{curl} \mathbf{u} = \mathbf{0}$ and $\nabla \cdot \mathbf{u} + f = 0$ has been used (cf. Chen and Fix in [16], [17]) for fluid flow computations.

There are two main approaches for studying least-squares methods for systems of first order. The first approach introduced by Aziz, Kellogg and Stephens in [2] uses the general theory of elliptic boundary value problems of Agmon-Douglis-Nirenberg (ADN) and reduces the system to a minimization of a least-squares functional that consists of a weighted sum of the residuals occurring in the equations and the boundary conditions. The weights occurring in the least-squares functional are determined by the indices that enter into the definition of the ADN boundary value problem. See also the paper of Chang [15]. This approach generalizes both the least-squares method of Jespersen [22], which is for the Poisson equation written as a *grad - div* system, and the method of Wendland [34], which is for elliptic systems of Cauchy-Riemann type. Recently, Bochev and Gunzburger [8], [9], have extended the ADN approach to velocity-vorticity-pressure formulation of Stokes and Navier-Stokes equations and have produced some very interesting theoretical and computational results.

The second approach, mostly used for second order elliptic problems written as systems of first order, introduces a least-squares functional and studies the resulting minimization problem in the framework of the Lax-Milgram theory establishing the boundness and the coercivity of the corresponding bilinear form in an appropriate space. Interesting computational experiments in this setting have been done by Chen and Fix in [17] and by Carey and Shen in [14] that were a basis for the theoretical analysis of Pehlivanov, Carey and Lazarov in [30] for selfadjoint and of Cai, Lazarov, Manteuffel and McCormick in [12] for non-selfadjoint second order elliptic equations. The main result in [12], [30] is that the least-squares functional generates a bilinear form that is continuous and coercive in a properly defined subspace of $H_{div}(\Omega) \times H^1(\Omega)$ and, therefore, any finite element approximation of $H_{div}(\Omega)$ can be used since the approximating space need not to satisfy the inf-sup condition. A recent paper by Pehlivanov, Carey and Vassilevski [32] considers a least-squares method non-selfadjoint problems.

One problem with the above mentioned least-squares methods is that the error estimates require relatively smooth solutions. The known estimates do not guarantee any convergence when the methods are applied to problems with low regularity solutions. The least-squares method developed in this paper will be stable and convergent as long as the solution belongs to the Sobolev space $H^{1+\beta}(\Omega)$, for any positive β .

In this paper, we introduce and study a new least-squares norm for systems arising from splitting convection-diffusion and reaction-diffusion equations into a

system of equations of first order. The problem may be indefinite and nonsymmetric as long as it has a unique solution. We introduce a least-squares functional that involves a discrete inner product that is related to the inner product in the Sobolev space $H^{-1}(\Omega)$. The use of this inner product results in a method which is optimal with respect to the required regularity as well as the order of approximation and extends to problems with low regularity solutions. In addition, the discrete system of equations which needs to be solved in order to compute the resulting approximation is easily preconditioned thus providing an efficient method for solving the algebraic equations. The preconditioner for the algebraic system corresponding to the new least-squares system only requires the construction of preconditioners for standard second order problems, a task which is well understood.

We only treat second order problems in this paper. Future papers will consider the H^{-1} technique applied to elliptic systems, for example, the equations of Stokes and elasticity.

The paper is organized as follows. In Section 2 we describe the least-squares approach using a discrete $H^{-1}(\Omega)$ inner product. We then discuss some of the properties of more standard least-squares methods already studied in the literature and show how this inner product results in a more balanced quadratic form. Next we define the computational algorithm and study its properties. In Section 3 we derive an error estimate for the least-squares finite element approximation, in Section 4 we discuss the issues of implementation of the iteration methods and finally in Section 5 we provide the results of numerical experiments on some model problems.

2. THE DISCRETE H^{-1} LEAST-SQUARES APPROACH.

In this section, we describe the least-squares approach using a discrete $H^{-1}(\Omega)$ inner product. We start by defining the second order boundary value problem which we shall be approximating. We next give some notation for norms and Sobolev spaces. We then discuss some of the properties of more standard least-squares methods already studied in the literature and show how the use of the inner product in $H^{-1}(\Omega)$ in the least square functional provides a more balanced quadratic form. Finally, we define the computational algorithm by introducing a discrete version of the $H^{-1}(\Omega)$ inner product.

We shall consider least-squares approximations to the solutions for the following second order elliptic boundary value problem. Let Ω be a domain in d dimensional Euclidean space with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ and let u satisfy

$$(2.1) \quad \begin{aligned} \mathcal{L}u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma_D, \\ \frac{\partial u}{\partial \nu} &= 0 && \text{on } \Gamma_N. \end{aligned}$$

Here $\frac{\partial u}{\partial \nu}$ denotes the co-normal derivative on Γ_N and the operator \mathcal{L} is given by

$$\mathcal{L}u = - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^d b_i(x) \frac{\partial u}{\partial x_i} + c(x)u.$$

We assume that the matrix $\{a_{ij}(x)\}$ is symmetric, uniformly positive definite and bounded. We further assume that $b_i \in L^\infty(\Omega)$, for $i = 1, \dots, d$.

To describe and analyze the least-squares method, we shall use Sobolev spaces. For non-negative integers s , let $H^s(\Omega)$ denote the Sobolev space of order s defined on Ω (see, e.g., [19], [26], [29]). The norm in $H^s(\Omega)$ will be denoted by $\|\cdot\|_s$. For $s = 0$, $H^s(\Omega)$ coincides with $L^2(\Omega)$. In this case, the norm and inner product will be denoted by $\|\cdot\|$ and (\cdot, \cdot) respectively. The space W is defined to be the closure of

$$\{v \in C^\infty(\Omega) \mid v = 0 \quad \text{on } \Gamma_D\}.$$

with respect to the norm in $H^1(\Omega)$. In the case where $\Gamma_D = \emptyset$, we define W to be the set of functions in $H^1(\Omega)$ with zero mean value. The space $H^{-1}(\Omega)$ is defined by duality and consists of the functionals v for which the norm

$$(2.2) \quad \|v\|_{-1} = \sup_{\phi \in W} \frac{(v, \phi)}{\|\phi\|_1}$$

is finite. Here (v, ϕ) also is the value of the functional v at ϕ . For noninteger values of s , $H^s(\Omega)$ is defined by the real method of interpolation (cf., [26]) between consecutive integers. We use the same notation for the norms of vector valued functions. Thus, if δ is a vector valued function with each component $\delta^i \in H^s(\Omega)$, then

$$\|\delta\|_s^2 \equiv \sum_{i=1}^d \|\delta^i\|_s^2.$$

Let $A(\cdot, \cdot)$ be the form corresponding to the operator \mathcal{L} , i.e., for $u, v \in H^1(\Omega)$,

$$\begin{aligned} A(u, v) &= \sum_{i,j=1}^d \int_{\Omega} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx \\ &\quad + \sum_{i=1}^d \int_{\Omega} b_i(x) \frac{\partial u}{\partial x_i} v dx + \int_{\Omega} c(x) uv dx. \end{aligned}$$

The weak formulation of (2.1) is given by the problem: Given $f \in L^2(\Omega)$, find $u \in W$ satisfying

$$(2.3) \quad A(u, \theta) = (f, \theta) \quad \text{for all } \theta \in W.$$

We assume that the solution of (2.3) is unique. This means that if $v \in W$ and satisfies $A(v, \theta) = 0$ for all $\theta \in W$ then $v = 0$. As usual (cf., [18], [26]), the uniqueness assumption implies the existence of solutions as well.

The particular space $H^{-1}(\Omega)$ chosen above is related to the boundary conditions used in our boundary value problem (2.1). We consider the symmetric problem

$$(2.4) \quad \begin{aligned} w - \Delta w &= f && \text{in } \Omega, \\ w &= 0 && \text{on } \Gamma_D, \\ \frac{\partial w}{\partial n} &= 0 && \text{on } \Gamma_N. \end{aligned}$$

Let $T : H^{-1}(\Omega) \mapsto W$ denote the solution operator for the above problem, i.e., for $f \in H^{-1}(\Omega)$, $Tf = w$ is the solution to (2.4). The following lemma provides the relationship between T and the norm in $H^{-1}(\Omega)$. Its proof is a simple consequence of the definition of T .

Lemma 2.1. *For all $v \in H^{-1}(\Omega)$,*

$$(2.5) \quad (v, Tv) = \sup_{\theta \in W} \frac{(v, \theta)^2}{\|\theta\|_1^2} = \|v\|_{-1}^2$$

and thus the inner product on $H^{-1}(\Omega) \times H^{-1}(\Omega)$ is given by (v, Tw) , for $v, w \in H^{-1}(\Omega)$. For v and w in $L^2(\Omega)$, $(v, Tw) = (Tv, w)$.

To define the least-squares approximation to (2.1) we start by considering the following reformulation of (2.1) into a system of first order equations. Let u be the solution of (2.1) and define $\sigma = -\mathcal{A}\nabla u$ where $\mathcal{A} = \mathcal{A}(x)$ is the matrix with entries $\{a_{ij}(x)\}$, $i, j = 1, \dots, d$. In addition, for $\theta \in H^1(\Omega)$, define

$$\mathcal{X}\theta = \sum_{i=1}^d b_i(x) \frac{\partial \theta}{\partial x_i} + c(x)\theta.$$

Then, (2.1) can be rewritten as

$$(2.6) \quad \begin{aligned} \sigma + \mathcal{A}\nabla u &= 0 && \text{in } \Omega, \\ \nabla \cdot \sigma + \mathcal{X}u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma_D, \\ \sigma \cdot n &= 0 && \text{on } \Gamma_N. \end{aligned}$$

Here n denotes the outward normal on Γ_N .

In order to motivate our new least-squares formulation, we shall consider for the moment a standard least-squares approach to (2.6). In particular, we shall point out some undesirable features which are not present in our new method. To this end, let $H_{div}(\Omega)$ denote the linear space of vector functions δ whose components δ_i , for $i = 1, \dots, d$, are in $L^2(\Omega)$ and whose divergence is also in $L^2(\Omega)$. The corresponding norm $\|\cdot\|_{H_{div}}$ is defined by

$$\|\delta\|_{H_{div}}^2 = \|\delta\|^2 + \|\nabla \cdot \delta\|^2.$$

The subset $H_{div}^0(\Omega)$ consisting of functions with vanishing normal component on Γ_N will be denoted $H_{div}^0(\Omega)$. The solution (σ, u) of (2.6) obviously minimizes the quadratic functional

$$(2.7) \quad Q_1(\delta, v) = \|\nabla \cdot \delta + \mathcal{X}v - f\|^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2$$

for all $\delta \in H_{div}^0(\Omega)$ and $v \in W$. It is known that for some positive numbers C_0, C_1 ,

$$(2.8) \quad \begin{aligned} C_0(\|\delta\|_{H_{div}}^2 + \|v\|_1^2) &\leq \|\nabla \cdot \delta + \mathcal{X}v\|^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2 \\ &\leq C_1(\|\delta\|_{H_{div}}^2 + \|v\|_1^2) \end{aligned}$$

for all $v \in W$ and $\delta \in H_{div}^0(\Omega)$. The one dimensional case was proved in [31] and the case of higher dimensions was proved in [12].

Numerical approximations are defined by introducing spaces of approximating functions $V_h \subseteq H_{div}^0(\Omega)$ and $W_h \subseteq W$. The discrete approximations are defined to be the pair $\sigma_h \in V_h$ and $u_h \in W_h$ which minimize (2.7) over all pairs (δ, v) in $V_h \times W_h$. It follows from (2.8) (cf. [12], [30]) that the errors $e_\sigma = \sigma - \sigma_h$ and $e_u = u - u_h$ are quasi-optimal with respect to the norm appearing in (2.8), i.e.,

$$(2.9) \quad \|e_\sigma\|_{H_{div}} + \|e_u\|_1 \leq C \inf_{(\delta, v) \in V_h \times W_h} \left\{ \|\sigma - \delta\|_{H_{div}} + \|u - v\|_1 \right\}.$$

Although (2.9) is optimal with respect to this norm, it does not provide an optimal estimate with respect to regularity of the solution. Consider, for example, the case when V_h and W_h consist of standard conforming piecewise linear finite element approximation subspaces on a triangulation of size h . In that case, (2.9) gives rise to the estimate

$$\|e_\sigma\|_{H_{div}} + \|e_u\|_1 \leq Ch \|u\|_3.$$

Thus, to get first order convergence in $L^2(\Omega)^d$ for σ (or $H^1(\Omega)$ for u), we need three Sobolev derivatives on the solution. Moreover, there is no theoretical convergence in the case when f is only in $L^2(\Omega)$ or u is only in $H^2(\Omega)$. In addition to this deficiency there is no obvious efficient way to solve the resulting algebraic equations.

The problem with the above least-squares formulation is that there are too many derivatives on σ , i.e., the $L^2(\Omega)$ norm is too strong in the first term on the right hand side of (2.7). This suggests the use of a weaker norm. Consider the least-squares method based on the following functional:

$$(2.10) \quad Q_2(\delta, v) = \|\nabla \cdot \delta + \mathcal{X}v - f\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2.$$

The above functional makes sense for $\delta \in H_{div}^0(\Omega)$ (and in fact somewhat more generally as we will see in Section 3) and $v \in W$. The solution pair (σ, u) is its minimum. The following lemma will be proved in the next section.

Lemma 2.2. *There are positive numbers c_0 and c_1 such that*

$$(2.11) \quad c_0(\|\delta\|^2 + \|v\|_1^2) \leq \|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2 \\ \leq c_1(\|\delta\|^2 + \|v\|_1^2),$$

for all $(\delta, v) \in H_{div}^0(\Omega) \times W$. The constants c_0 and c_1 above depend on $\{a_{ij}\}$, $\{b_j\}$ and c .

We now consider least-squares approximation based on Q_2 . Let $V_h \subseteq H_{div}^0(\Omega)$ and $W_h \subseteq W$ and let (σ_h, u_h) minimize (2.10) over $V_h \times W_h$. It follows from Lemma 2.2 that the resulting errors $e_\sigma = \sigma - \sigma_h$ and $e_u = u - u_h$ are quasi-optimal with respect to the norm appearing in (2.11); i.e.,

$$(2.12) \quad \|e_\sigma\| + \|e_u\|_1 \leq C \inf_{(\delta, v) \in V_h \times W_h} \left\{ \|\sigma - \delta\| + \|u - v\|_1 \right\}.$$

This method gives rise to estimates which are optimal with respect to the the order of approximation as well as the required regularity. Let us consider the case when V_h consists of standard conforming piecewise linear finite element approximation subspaces and W_h consists of the lowest order Raviart-Thomas spaces on a triangulation of size h . In that case, (2.12) gives rise to the estimate

$$\|e_\sigma\| + \|e_u\|_1 \leq Ch \|u\|_2.$$

Thus, we get first order convergence in $L^2(\Omega)^d$ on σ (or in $H^1(\Omega)$ on u) when u is only in $H^2(\Omega)$. This estimate is optimal both with respect to the order of approximation as well as the required regularity.

Although minimization with respect to the functional $Q_2(\cdot, \cdot)$ appears attractive from the point of view of stability and accuracy, it is unfortunately not computationally feasible. This is because the evaluation of the operator T defining the inner product in $H^{-1}(\Omega)$ involves the solution of the boundary value problem (2.4).

To make the method computationally feasible, we will replace the operator T appearing in (2.10). Note that the first term of (2.10) can be rewritten

$$(2.13) \quad (T(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \delta + \mathcal{X}v)$$

Our goal is to replace T by an operator \mathcal{T}_h which is computable and is equivalent to T in the sense that there are positive constants c_2, c_3 not depending on h such that

$$(2.14) \quad \begin{aligned} c_2(T(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \delta + \mathcal{X}v) &\leq (\mathcal{T}_h(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \delta + \mathcal{X}v) \\ &\leq c_3(T(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \delta + \mathcal{X}v), \end{aligned}$$

for all $(\delta, v) \in V_h \times W_h$.

We construct \mathcal{T}_h from a preconditioner for the finite element approximation of (2.4). Let $T_h : H^{-1}(\Omega) \mapsto W_h$ be defined by $T_h f = w$ where w is the unique element of W_h satisfying

$$D(w, \theta) = (f, \theta) \quad \text{for all } \theta \in W_h.$$

Here $D(\cdot, \cdot)$ denotes the form on $H^1(\Omega)$ and is defined by

$$D(v, w) = \int_{\Omega} (\nabla v \cdot \nabla w + vw) \, dx.$$

A preconditioner $B_h : W_h \mapsto W_h$ is a symmetric, positive definite operator with respect to the $L^2(\Omega)$ inner product. A good preconditioner is one which is computationally easy to evaluate and is spectrally equivalent to T_h in the sense that there are positive constants c_4, c_5 not depending on h and satisfying

$$(2.15) \quad c_4(T_h w, w) \leq (B_h w, w) \leq c_5(T_h w, w) \quad \text{for all } w \in W_h.$$

Remark 2.1. The symmetry of B_h with respect to the inner product on $L^2(\Omega)$ implies that $B_h = B_h Q_h$, where Q_h is the $L^2(\Omega)$ orthogonal projection onto W_h .

Similarly, $T_h = T_h Q_h$. Thus (2.15) holds for all w in $L^2(\Omega)$ if it is satisfied for all w in W_h .

We will assume that B_h is a good preconditioner. We then define $\mathcal{T}_h = h^2 I + B_h$ where I denotes the identity operator on W_h . The purpose of this paper is to analyze least-squares approximation based on the functional

$$(2.16) \quad Q_3(\delta, v) = (\mathcal{T}_h(\nabla \cdot \delta + \mathcal{X}v - f), \nabla \cdot \delta + \mathcal{X}v - f) + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2.$$

The quadratic form $Q_3(\cdot, \cdot)$ shares many of the properties of $Q_2(\cdot, \cdot)$ when restricted to the approximation subspaces. For example, the inequality analogous to (2.11) holds under reasonable assumptions (the norm $\|\cdot\|_{-1}$ is replaced by $(\mathcal{T}_h \cdot, \cdot)^{1/2}$). This allows us to construct efficient iterative methods for the solution of the resulting discrete equations. We will discuss this more fully in Section 4. Inequalities analogous to (2.11) also enable one to prove error estimates which are optimal both in order of approximation and required regularity (see, Section 3).

Remark 2.2. Note that we require that the operator \mathcal{T}_h be equivalent to T on functions of the form appearing in (2.14). The $h^2 I$ term is necessary since the operator B_h alone may fail to satisfy the lower inequality in (2.14).

Remark 2.3. The exact weighting in the definition of \mathcal{T}_h is not critical. For example, one could take

$$\mathcal{T}_h = \alpha h^2 I + \beta B_h$$

for fixed positive constants α and β . These parameters could be used to tune the iterative convergence rate. The order of convergence is not changed.

3. ERROR ANALYSIS.

We provide in this section an analysis of the least-squares approximation based on the functional $Q_3(\cdot, \cdot)$ defined in (2.16). First we prove Lemma 2.2 and then establish a stability estimate involving the norms corresponding to the quadratic functional $Q_3(\cdot, \cdot)$. We then prove error estimates which are optimal in order and regularity for e_σ in $L^2(\Omega)^d$ and e_u in $H^1(\Omega)$ and conclude this section by proving an optimal $L^2(\Omega)$ estimate for e_u .

In the remainder of this paper, C , with or without subscript will denote a generic positive constant. These constants will take on different values in different occurrences but are always independent of the mesh parameter h .

Proof of Lemma 2.2. We will use an additional function space for the proof. Define the boundary norm, for $\theta \in L^2(\Gamma_N)$, by

$$(3.1) \quad \|\theta\|_{-1/2, \Gamma_N} = \sup_{\phi \in \mathcal{W}} \frac{\langle \theta, \phi \rangle}{\|\phi\|_1}$$

where $\langle \cdot, \cdot \rangle$ denotes the $L^2(\Gamma_N)$ inner product. We consider the norm

$$(3.2) \quad \|\delta, v\| \equiv (\|\delta \cdot n\|_{-1/2, \Gamma_N}^2 + \|\delta\|^2 + \|v\|_1^2)^{1/2}$$

and let \mathcal{H} be the closure of $H_{div}(\Omega) \times W$ with respect to this norm.

We first prove that there is a constant C satisfying

$$(3.3) \quad \begin{aligned} \|\delta, v\|^2 &\leq C \left(\|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2 \right. \\ &\quad \left. + \|\delta \cdot n\|_{-1/2, \Gamma_N}^2 + \|v\|^2 \right) \end{aligned}$$

for all $(\delta, v) \in \mathcal{H}$. Indeed, for smooth δ and v ,

$$\begin{aligned} (\mathcal{A}\nabla v, \nabla v) &= -(\delta, \nabla v) + (\delta + \mathcal{A}\nabla v, \nabla v) \\ &= (\nabla \cdot \delta + \mathcal{X}v, v) - \langle \delta \cdot n, v \rangle \\ &\quad + (\mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v), \mathcal{A}^{1/2}\nabla v) - (\mathcal{X}v, v). \end{aligned}$$

By the Poincaré inequality,

$$\|v\|_1^2 \leq C(\mathcal{A}\nabla v, \nabla v).$$

Thus, the Schwarz inequality and obvious manipulations imply that

$$\begin{aligned} \|v\|_1^2 &\leq C \left(\|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2 \right. \\ &\quad \left. + \|\delta \cdot n\|_{-1/2, \Gamma_N}^2 + \|v\|^2 \right). \end{aligned}$$

The inequality (3.3) immediately follows for smooth δ and v . We clearly have that

$$(3.4) \quad \|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2 + \|\delta \cdot n\|_{-1/2, \Gamma_N}^2 + \|v\|^2 \leq C\|\delta, v\|^2.$$

It follows that inequality (3.3) holds for all $(\delta, v) \in \mathcal{H}$ by continuity.

We next show that

$$(3.5) \quad \begin{aligned} \|\delta, v\|^2 &\leq C \left(\|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 \right. \\ &\quad \left. + \left\| \mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v) \right\|^2 + \|\delta \cdot n\|_{-1/2, \Gamma_N}^2 \right) \end{aligned}$$

for all $(\delta, v) \in \mathcal{H}$ by applying a standard compactness argument. This argument is by contradiction. Assume that (3.5) does not hold for any constant $C > 0$. Then there is a sequence $\{(\delta_i, v_i)\}$, for $i = 1, 2, \dots$, with $\{(\delta_i, v_i)\} \in \mathcal{H}$, $\|\delta_i, v_i\| = 1$ and

$$(3.6) \quad \|\nabla \cdot \delta_i + \mathcal{X}v_i\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta_i + \mathcal{A}\nabla v_i) \right\|^2 + \|\delta_i \cdot n\|_{-1/2, \Gamma_N}^2 \leq \frac{1}{i}.$$

Since W is compactly contained in $L^2(\Omega)$, we may assume without loss of generality that v_i converges in $L^2(\Omega)$. It immediately follows from (3.3) and (3.6) that the sequence $\{(\delta_i, v_i)\}$ is a Cauchy sequence with respect to the norm $\|\cdot, \cdot\|$. Let (δ, v) converge to (δ, v) in \mathcal{H} .

For any $\phi \in W$, we then have

$$\begin{aligned} A(v_i, \phi) &= (\mathcal{A}\nabla v_i, \nabla\phi) + (\mathcal{X}v_i, \phi) \\ &= -(\delta_i, \nabla\phi) + (\mathcal{X}v_i, \phi) + (\delta_i + \mathcal{A}\nabla v_i, \nabla\phi) \\ &= (\nabla \cdot \delta_i + \mathcal{X}v_i, \phi) + (\delta_i + \mathcal{A}\nabla v_i, \nabla\phi) - \langle \delta_i \cdot n, \phi \rangle. \end{aligned}$$

Hence,

$$\begin{aligned} |A(v, \phi)| &= \lim_{i \rightarrow \infty} |A(v_i, \phi)| \\ &\leq \lim_{i \rightarrow \infty} \left(\|\nabla \cdot \delta_i + \mathcal{X}v_i\|_{-1}^2 + \left\| \mathcal{A}^{-1/2}(\delta_i + \mathcal{A}\nabla v_i) \right\|^2 \right. \\ &\quad \left. + \|\delta_i \cdot n\|_{-1/2, \Gamma_N}^2 \right)^{1/2} \|\phi\|_1 = 0. \end{aligned}$$

By our assumption that solutions of (2.3) are unique, it follows that $v = 0$. In addition, using (3.3),

$$\begin{aligned} \|\delta\|^2 + \|\delta \cdot n\|_{-1/2, \Gamma_N}^2 &\leq C \lim_{i \rightarrow \infty} \left(\|\nabla \cdot \delta_i + \mathcal{X}v_i\|_{-1}^2 \right. \\ &\quad \left. + \left\| \mathcal{A}^{-1/2}(\delta_i + \mathcal{A}\nabla v_i) \right\|^2 + \|\delta_i \cdot n\|_{-1/2, \Gamma_N}^2 \right) = 0. \end{aligned}$$

This contradicts the assumption that

$$\llbracket \delta, v \rrbracket = \lim_{i \rightarrow \infty} \llbracket \delta_i, v_i \rrbracket = 1$$

and hence completes the proof of (3.5). The lemma follows by restricting (3.4) and (3.5) to $H_{div}^0(\Omega) \times W$ and hence completes the proof.

We next state some hypotheses which we shall require to hold for the approximation subspaces. It is well known that these properties hold for typical finite element spaces consisting of piecewise polynomials with respect to quasi-uniform triangulations of the domain Ω (cf., [1], [7], [11], [13], [33]). Let r be an integer greater than or equal to one.

(H.1) The subspace V_h has the following approximation property: For any $\eta \in H^r(\Omega)^d \cap H_{div}^0(\Omega)$,

$$(3.7) \quad \inf_{\delta \in V_h} \left\{ \|\eta - \delta\| + h \|\nabla \cdot (\eta - \delta)\| \right\} \leq Ch^r \|\eta\|_r.$$

(H.2) The subspace W_h has the following approximation property: For any $w \in H^{r+1}(\Omega) \cap W$,

$$(3.8) \quad \inf_{v \in W_h} \left\{ \|w - v\| + h \|w - v\|_1 \right\} \leq Ch^{r+1} \|w\|_{r+1}.$$

(H.3) We assume that W_h is such that Q_h , the $L^2(\Omega)$ orthogonal projection operator onto W_h , is a bounded operator with respect to the norm in W , i.e.,

$$(3.9) \quad \|Q_h u\|_1 \leq C \|u\|_1 \quad \text{for all } u \in W.$$

Remark 3.1. It follows from [6] that if (H.1) and (H.2) hold for $r = \tilde{r}$ then they hold for $r = 1, 2, \dots, \tilde{r}$. The property (H.3) is studied in [7].

We note some properties implied by the above assumptions. It follows from (3.8) that

$$(3.10) \quad \begin{aligned} \|(I - Q_h)v\|_{-1} &= \sup_{\theta \in W} \frac{(v, (I - Q_h)\theta)}{\|\theta\|_1} \\ &\leq Ch \|v\| \quad \text{for all } v \in L^2(\Omega). \end{aligned}$$

It follows from (3.9) that Q_h is defined and bounded on $H^{-1}(\Omega)$. In addition, (3.9) implies that

$$(3.11) \quad \|Q_h u\|_{-1}^2 \leq C_2(u, T_h u) \leq C_2 \|u\|_{-1}^2 \quad \text{for all } u \in H^{-1}(\Omega).$$

Indeed, the upper inequality follows from (2.5) and the analogous equality

$$(v, T_h v) = \sup_{\theta \in W_h} \frac{(v, \theta)^2}{\|\theta\|_1^2}.$$

For the lower inequality of (3.11), (3.9) implies that

$$\begin{aligned} (TQ_h v, Q_h v) &= \sup_{\theta \in W} \frac{(v, Q_h \theta)^2 \|Q_h \theta\|_1^2}{\|Q_h \theta\|_1^2 \|\theta\|_1^2} \\ &\leq C \sup_{\phi \in W_h} \frac{(v, \phi)^2}{\|\phi\|_1^2} = C(v, T_h v). \end{aligned}$$

We next prove a result analogous to Lemma 2.2 for the functional $Q_3(\cdot, \cdot)$. For convenience, we define the corresponding form

$$(3.12) \quad \begin{aligned} [\delta, v; \eta, w] &= (T_h(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \eta + \mathcal{X}w) \\ &\quad + (\mathcal{A}^{-1}(\delta + \mathcal{A}\nabla v), \eta + \mathcal{A}\nabla w), \end{aligned}$$

for all $\delta, \eta \in H_{div}^0(\Omega)$ and $v, w \in W$. The corresponding norm will be denoted by $|||\cdot, \cdot|||$ and is defined by

$$|||\delta, v||| = [\delta, v; \delta, v]^{1/2}.$$

We then have the following lemma.

Lemma 3.1. *Assume (H.1) and (H.3) hold and that T_h is constructed as described in Section 2 with a preconditioning operator B_h satisfying (2.15) with constants c_4 and c_5 not depending on h . Then, for all $\delta \in H_{div}^0(\Omega)$ and $v \in W$,*

$$(3.13) \quad C_0(\|\delta\|^2 + \|v\|_1^2) \leq |||\delta, v|||^2 \leq C_1(h^2 \|\nabla \cdot \delta\|^2 + \|\delta\|^2 + \|v\|_1^2)$$

holds with C_0 and C_1 which are independent of h .

Proof. By Lemma 2.2, the lower inequality of (3.13) will follow if we can show that

$$(3.14) \quad \|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 \leq C(T_h(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \delta + \mathcal{X}v),$$

for all $\delta \in H_{div}^0(\Omega)$ and $v \in W$. For any $w \in L^2(\Omega)$,

$$\|w\|_{-1} \leq \|(I - Q_h)w\|_{-1} + \|Q_h w\|_{-1}$$

and hence (3.10) and (3.11) imply that

$$\|w\|_{-1}^2 \leq C(h^2 \|w\|^2 + (T_h w, w)) \leq C(\mathcal{T}_h w, w).$$

We used Remark 2.1 for the last inequality above. This verifies (3.14) and completes the proof of the lower inequality in (3.13).

For the upper inequality in (3.13), we note that by (3.11), for $w \in L^2(\Omega)$,

$$(3.15) \quad (\mathcal{T}_h w, w) \leq C(h^2 \|w\|^2 + (T_h w, w)) \leq C(h^2 \|w\|^2 + \|w\|_{-1}^2).$$

The upper inequality of (3.13) follows from Lemma 2.2 and (3.15). This completes the proof of Lemma 3.1.

Remark 3.2. If V_h satisfies an inverse inequality of the form

$$(3.16) \quad \|\nabla \cdot \delta\|^2 \leq Ch^{-2} \|\delta\|^2$$

then the upper inequality of (3.13) can be replaced by

$$(3.17) \quad \|\|\delta, v\|\|^2 \leq C_1(\|\delta\|^2 + \|v\|_1^2),$$

for all $\delta \in V_h$ and $v \in W$.

The following theorem gives estimates for the least-squares approximation using the functional $Q_3(\cdot, \cdot)$.

Theorem 3.1. *Assume that (H.2) and the hypotheses of Lemma 3.1 are satisfied. Let (σ_h, u_h) in $V_h \times W_h$ be the unique minimizer of $Q_3(\cdot, \cdot)$ over all (δ, v) in $V_h \times W_h$. If the solution u of (2.1) is in $H^{r+1}(\Omega)$ then the errors $e_\sigma = \sigma - \sigma_h$ and $e_u = u - u_h$ satisfy the inequality*

$$(3.18) \quad \|e_\sigma\| + \|e_u\|_1 \leq Ch^r \|u\|_{r+1}.$$

The constant C above is independent of the mesh size h .

Proof. It is immediate from the definition of the approximation scheme that the error (e_σ, e_u) satisfies

$$(3.19) \quad [e_\sigma, e_u; \delta, v] = 0 \quad \text{for all } (\delta, v) \in V_h \times W_h.$$

Thus, by Lemma 3.1,

$$C_0(\|e_\sigma\|^2 + \|e_u\|_1^2) \leq [e_\sigma, e_u; e_\sigma, e_u] = \inf_{(\delta, v) \in V_h \times W_h} [e_\sigma, e_u; \sigma - \delta, u - v].$$

By the Schwarz inequality and (3.13), using (H.1) and (H.2), it follows that

$$\begin{aligned} \|e_\sigma\|^2 + \|e_u\|_1^2 &\leq C \inf_{(\delta,v) \in V_h \times W_h} \|\sigma - \delta, u - v\|^2 \\ &\leq C \inf_{(\delta,v) \in V_h \times W_h} (h^2 \|\nabla \cdot (\sigma - \delta)\|^2 + \|\sigma - \delta\|^2 + \|u - v\|_1^2) \\ &\leq Ch^{2r} \|u\|_{r+1}^2. \end{aligned}$$

This completes the proof of the theorem.

Often solutions to elliptic boundary value problems may not be in the Sobolev space $H^{r+1}(\Omega)$ for any $r \geq 1$. Problems on nonconvex polygonal domains, problems with discontinuous coefficients and problems with arbitrary Γ_N all give rise to solutions which are only in $H^{1+\beta}(\Omega)$ for some positive β strictly less than one. The following corollary shows that the least-squares method with a simple modification will still give stable and accurate approximation.

Corollary 3.1. *Assume that the hypotheses of Theorem 3.1 are satisfied and that the inverse inequality*

$$(3.20) \quad \|v\|_1 \leq Ch^{-1} \|v\| \quad \text{for all } v \in W_h,$$

holds. Let $Q_4(\cdot, \cdot)$ denote the functional which is defined by replacing f by $Q_h f$ in (2.16). Let (σ_h, u_h) in $V_h \times W_h$ be the unique minimizer of $Q_4(\cdot, \cdot)$ over all (δ, v) in $V_h \times W_h$. If the solution u of (2.1) is in $H^{1+\beta}(\Omega)$ with $0 \leq \beta \leq 1$ then the errors $e_\sigma = \sigma - \sigma_h$ and $e_u = u - u_h$ satisfy the inequality

$$(3.21) \quad \|e_\sigma\| + \|e_u\|_1 \leq Ch^\beta \|u\|_{1+\beta}.$$

The constant C above is independent of the mesh size h .

Proof. Note that

$$\|\sigma_h, u_h\|^2 = [\sigma_h, u_h; \sigma_h, u_h] = (\mathcal{T}_h Q_h f, \nabla \cdot \sigma_h + \mathcal{X} u_h).$$

By the Schwarz inequality

$$\|\sigma_h, u_h\|^2 \leq (\mathcal{T}_h Q_h f, Q_h f)^{1/2} (\mathcal{T}_h (\nabla \cdot \sigma_h + \mathcal{X} u_h), \nabla \cdot \sigma_h + \mathcal{X} u_h)^{1/2}.$$

It easily follows that

$$\|\sigma_h, u_h\|^2 \leq (\mathcal{T}_h Q_h f, Q_h f).$$

By (3.20) and duality,

$$h^2 \|Q_h f\|^2 \leq C \|Q_h f\|_{-1}^2$$

and hence (3.11) implies that

$$(\mathcal{T}_h Q_h f, Q_h f) \leq C \|f\|_{-1}^2.$$

Using (2.2) and (2.3), it follows that

$$\|f\|_{-1}^2 \leq C \|u\|_1^2.$$

Combining the above inequalities with Lemma 3.1 gives

$$\|\sigma_h\|^2 + \|u_h\|_1^2 \leq C \|u\|_1^2$$

and hence

$$(3.22) \quad \|e_\sigma\|^2 + \|e_u\|_1^2 \leq C \|u\|_1^2.$$

We next show that (3.21) holds for $\beta = 1$ and discrete solutions resulting from the functional $Q_4(\cdot, \cdot)$. Assume that $u \in H^2(\Omega)$. Let (σ'_h, u'_h) in $V_h \times W_h$ be the unique minimizer of $Q_3(\cdot, \cdot)$ over all (δ, v) in $V_h \times W_h$. Then,

$$\begin{aligned} \|\sigma_h - \sigma'_h, u_h - u'_h\|^2 &= [\sigma_h - \sigma'_h, u_h - u'_h; \sigma_h - \sigma'_h, u_h - u'_h] \\ &= -(\mathcal{T}_h(I - Q_h)f, \nabla \cdot (\sigma_h - \sigma'_h) + \mathcal{X}(u_h - u'_h)). \end{aligned}$$

As above,

$$\|\sigma_h - \sigma'_h, u_h - u'_h\|^2 \leq (\mathcal{T}_h(I - Q_h)f, (I - Q_h)f) = h^2 \|(I - Q_h)f\|^2.$$

By (3.8),

$$\|\sigma_h - \sigma'_h, u_h - u'_h\|^2 \leq Ch^2 \|f\|_0^2 \leq Ch^2 \|u\|_2^2.$$

The above estimate and Lemma 3.1 imply that

$$\|\sigma_h - \sigma'_h\|^2 + \|u_h - u'_h\|_1^2 \leq Ch^2 \|u\|_2^2.$$

Thus by Theorem 1 and the triangle inequality, it follows that

$$(3.23) \quad \|e_\sigma\|^2 + \|e_u\|_1^2 \leq Ch^2 \|u\|_2^2.$$

The corollary follows interpolating (3.22) and (3.23).

We conclude this section by proving an improved error estimate for e_u in $L^2(\Omega)$. For this result, we need somewhat stronger assumptions on the operator B_h used in the definition of \mathcal{T}_h . We assume that B_h is such that there is a positive number c_6 satisfying

$$(3.24) \quad c_6 \|B_h^{-1}v\| \leq \|T_h^{-1}v\| \quad \text{for all } v \in W_h.$$

In contrast to (2.15), there are far fewer examples of operators B_h known to satisfy (3.24). If the operator T gives rise to full elliptic regularity, then it is known that the W-cycle multigrid algorithm with sufficiently many smoothings on each level gives rise to an operator B_h which satisfies (3.24) (cf., [4]). Another example of an operator B_h which satisfies (3.24) is the variable V-cycle introduced in [5]. It seems that in this case also sufficiently many smoothings are required on the finest level. Though results of numerical calculations indicate that (3.24) holds also for the usual V-cycle, as far as we know there is no proof of this in the literature.

Improved $L^2(\Omega)$ estimates depend upon elliptic regularity. We consider the adjoint boundary value problem in weak form: Given $g \in L^2(\Omega)$ find $v \in W$ such that

$$(3.25) \quad A(\phi, v) = (\phi, g) \quad \text{for all } \phi \in W.$$

Solutions of (3.25) exist and are unique since we have assumed uniqueness and existence for solutions to (2.3). We assume full elliptic regularity, i.e., solutions to (2.3), (2.4) and (3.25) are in $H^2(\Omega) \cap W$ and satisfy the inequalities

$$(3.26) \quad \begin{aligned} \|u\|_2 &\leq C \|f\|, \\ \|Tf\|_2 &\leq C \|f\|, \\ \|v\|_2 &\leq C \|g\|. \end{aligned}$$

Theorem 3.2. *Assume that the hypotheses for Theorem 3.1 are satisfied. In addition, assume that B_h also satisfies (3.24) and that solutions of (2.1), (2.4) and (3.25) satisfy (3.26). Then,*

$$\|e_u\| \leq Ch \| |e_\sigma, e_u| \|.$$

Proof. The proof is by duality. Let v solve (3.25) with $g = e_u$. Then,

$$(3.27) \quad \begin{aligned} \|e_u\|^2 &= A(e_u, v) \\ &= (\mathcal{T}_h(\nabla \cdot e_\sigma + \mathcal{X}e_u), \mathcal{T}_h^{-1}v) + (\mathcal{A}^{-1}(e_\sigma + \mathcal{A}\nabla e_u), \mathcal{A}\nabla v). \end{aligned}$$

We want to define w and η so that

$$(3.28) \quad \mathcal{T}_h^{-1}v = \nabla \cdot \eta + \mathcal{X}w \quad \text{and} \quad \mathcal{A}\nabla v = \eta + \mathcal{A}\nabla w.$$

To this end let w be the solution of

$$(3.29) \quad A(w, \phi) = (\mathcal{T}_h^{-1}v - \nabla \cdot \mathcal{A}\nabla v, \phi) \quad \text{for all } \phi \in W.$$

By (3.26), v is in $H^2(\Omega)$ and hence the data appearing in (3.29) are in $L^2(\Omega)$. Thus, w is in $H^2(\Omega)$ and satisfies

$$\mathcal{L}w = \mathcal{T}_h^{-1}v - \nabla \cdot \mathcal{A}\nabla v.$$

Setting $\eta = \mathcal{A}\nabla(v - w)$ we see that (3.28) is satisfied. Hence (3.27) and (3.28) give

$$\begin{aligned} \|e_u\|^2 &= (\mathcal{T}_h(\nabla \cdot e_\sigma + \mathcal{X}e_u), \nabla \cdot \eta + \mathcal{X}w) + (\mathcal{A}^{-1}(e_\sigma + \mathcal{A}\nabla e_u), \eta + \mathcal{A}\nabla w) \\ &= [e_\sigma, e_u; \eta, w]. \end{aligned}$$

By (3.19) and the Schwarz inequality,

$$(3.30) \quad \|e_u\|^2 \leq \| |e_\sigma, e_u| \| \| \eta - \delta, w - \theta \|$$

for all $(\delta, \theta) \in V_h \times W_h$. Applying Lemma 3.1 and Remark 3.1 implies that there exists $(\delta, \theta) \in V_h \times W_h$ such that

$$\begin{aligned}
 (3.31) \quad \|\eta - \delta, w - \theta\|^2 &\leq C(h^2 \|\nabla \cdot (\eta - \delta)\|^2 + \|\eta - \delta\|^2 + \|w - \theta\|_1^2) \\
 &\leq Ch^2(\|\eta\|_1^2 + \|w\|_2^2) \leq Ch^2(\|w\|_2^2 + \|v\|_2^2) \\
 &\leq Ch^2(\|\mathcal{T}_h^{-1}v\|^2 + \|v\|_2^2).
 \end{aligned}$$

The last inequality above follows from (3.26). By the triangle inequality,

$$\begin{aligned}
 \|\mathcal{T}_h^{-1}v\| &= \|\mathcal{T}_h^{-1}T(v - \Delta v)\| \\
 &\leq \|\mathcal{T}_h^{-1}(T - T_h)(v - \Delta v)\| + \|\mathcal{T}_h^{-1}T_h(v - \Delta v)\|.
 \end{aligned}$$

It is easy to see from eigenfunction expansions with respect to the operator B_h that

$$\|\mathcal{T}_h^{-1}\zeta\| \leq \|B_h^{-1}\zeta\| \quad \text{for all } \zeta \in W_h$$

and

$$\|\mathcal{T}_h^{-1}\zeta\| \leq h^{-2} \|\zeta\| \quad \text{for all } \zeta \in L^2(\Omega).$$

It follows from (3.26), Remark 3.1 and standard finite element theory that

$$\|(T - T_h)f\| \leq Ch^2 \|f\| \quad \text{for all } f \in L^2(\Omega).$$

Combining the above estimates gives

$$(3.32) \quad \|\mathcal{T}_h^{-1}v\| \leq C(\|v\|_2 + \|B_h^{-1}T_h(v - \Delta v)\|) \leq C \|v\|_2.$$

We used (3.24) for the last inequality above. Combining (3.30), (3.31) (3.32) and (3.26) completes the proof of the theorem.

4. IMPLEMENTATION AND THE ITERATIVE SOLUTION OF THE LEAST-SQUARES SYSTEM.

In this section we consider the implementation aspects of the least square method corresponding to $Q_3(\cdot, \cdot)$ described in Section 2 and analyzed in Section 3. The resulting equations are solved by preconditioned iteration. There are two major aspects involved in the implementation of a preconditioned iteration, the operator evaluation and the evaluation of the preconditioner. These tasks will be considered in detail in this section.

Our goal is to solve the equations which result from the minimization of the functional $Q_3(\cdot, \cdot)$ over the space $V_h \times W_h$. The solution pair (σ_h, u_h) satisfies the equations

$$(4.1) \quad [\sigma_h, u_h; \delta, v] = (\mathcal{T}_h f, \nabla \cdot \delta + \mathcal{X}v)$$

for all pairs $(\delta, v) \in V_h \times W_h$. As a model application, we will consider the case when V_h and W_h consist of continuous piecewise linear functions with respect to a quasi uniform triangulation of Ω of size h . We only consider the case when $\Gamma_N = \emptyset$

and when Ω is a subset of R^2 . The functions in W_h vanish on $\partial\Omega$ while those in V_h are piecewise linear vector functions without any imposed boundary conditions. We also let \bar{W}_h denote the set of continuous piecewise linear functions on Ω (not satisfying any boundary conditions). Extensions to higher dimensional finite element subspaces are straightforward. We shall avoid the question of quadrature in the case of variable coefficient problems. Instead, we shall assume that all coefficients are piecewise constant with respect to the triangulation defining the mesh. We will also replace f by its interpolant $\tilde{f} \in \bar{W}_h$.

Remark 4.1. For general polygonal domains with $\Gamma_N \neq \emptyset$, it is not suitable to use spaces of continuous piecewise linear functions for V_h . This is because the boundary condition in $H_{div}^0(\Omega)$ may force all components of continuous piecewise linear vector functions to vanish on Γ_N resulting in a loss of accuracy. For these problems we could define V_h to be the spaces designed for mixed finite approximation such as the Raviart-Thomas or the Brezzi-Douglas-Marini spaces. In such spaces, the boundary conditions can be easily satisfied while retaining the desired approximation properties.

Let $\{\delta^i\}$ and $\{\phi^i\}$ be the nodal bases for the spaces V_h and W_h respectively. The two bases provide a natural basis for the product space $V_h \times W_h$ which we shall denote by $\{\zeta^i\}$. Each basis function ζ^i is of the form (ζ_v^i, ζ_m^i) where ζ_v^i is a basis element for V_h or ζ_m^i is a basis element for W_h .

As usual, one writes the solution

$$(\sigma_h, u_h) = \sum_i D_i \zeta^i$$

in terms of this basis and replaces (4.1) by the matrix problem

$$(4.2) \quad MD = F$$

where M is the matrix with entries

$$M_{ij} = [\zeta_v^i, \zeta_m^i; \zeta_v^j, \zeta_m^j]$$

and F is the vector with entries

$$F_i = (\mathcal{T}_h \tilde{f}, \nabla \cdot \zeta_v^i + \mathcal{X} \zeta_m^i).$$

Clearly, M is symmetric and is also positive definite by Lemma 3.1. Although, the implementation involves the solution matrix system (4.2), the matrix itself is never assembled. In fact, because of the operator \mathcal{T}_h appearing in the first term of (3.12), M is a dense matrix. Instead, one solves (4.2) by preconditioned iteration.

The implementation of a preconditioned iteration for solving (4.2) involves three distinct steps. First, we must compute the vector F . Second, we must be able to compute the action of the matrix M applied to arbitrary vectors $G \in R^m$, where m is the dimension of $V_h \times W_h$. Finally, we must be able to compute the action of a suitable preconditioner applied to arbitrary vectors $G \in R^m$. As we shall see, all three steps involve the preconditioner B_h .

In previous sections in this paper, we defined B_h as a symmetric positive definite operator on W_h . In terms of the implementation, the preconditioner can be more naturally thought of in terms of a $n \times n$ matrix N where n is the dimension of W_h . The operator B_h is defined in terms of this matrix as follows. Fix $v \in W_h$ and expand

$$B_h v = \sum_i G_i \phi^i.$$

Then,

$$(4.3) \quad NG = \tilde{G}$$

where

$$(4.4) \quad \tilde{G}_i = (v, \phi^i).$$

The operator B_h is a good preconditioner for T_h provided that the matrix $N^{-1}\tilde{N}$ has small condition number. Here \tilde{N} is the stiffness matrix for the form $D(\cdot, \cdot)$ defined in Section 2. The matrix N need not explicitly appear in the computation of the action of the preconditioner. Instead, one often has a process or algorithm which acts on the vector \tilde{G} and produces the vector G , i.e., computes $N^{-1}\tilde{G}$. Thus, the practical application of the preconditioner on a function v reduces to a predefined algorithm and the evaluation of the vector \tilde{G} defined by (4.4).

We now outline the steps for computing F . We first compute the nodal values of \tilde{f} by evaluating f at the nodes. The data (for application of B_h)

$$\tilde{G}_i = (\tilde{f}, \phi^i)$$

can be analytically calculated since the product $\tilde{f}\phi^i$ is piecewise quadratic with respect to the mesh triangulation. The coefficients of $B_h\tilde{f}$ result from application of the preconditioning algorithm. The remaining quantities can be analytically computed since they only involve integration of piecewise quadratic functions.

The next action required for the preconditioning iteration is the application of M to arbitrary vectors $G \in R^m$. The vector G represents the coefficients of a function pair

$$(\delta, v) = \sum_i G_i (\zeta_v^i, \zeta_m^i)$$

and we are required to evaluate

$$(4.5) \quad \begin{aligned} (MG)_j = [\delta, v; \zeta_v^j, \zeta_m^j] &= (\mathcal{T}_h(\nabla \cdot \delta + \mathcal{X}v), \nabla \cdot \zeta_v^j + \mathcal{X}\zeta_m^j) \\ &+ (\mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v), \zeta_v^j + \mathcal{A}\nabla\zeta_m^j), \end{aligned}$$

for $j = 1, \dots, n$. The quantity $(\nabla \cdot \delta + \mathcal{X}v)$ is a discontinuous piecewise linear function with respect to the mesh triangulation. The data for the preconditioner solve

$$((\nabla \cdot \delta + \mathcal{X}v), \phi^i)$$

can be computed since it reduces to integrals of piecewise quadratic functions over the triangles. After application of the preconditioning process, the function $\mathcal{T}_h(\nabla \cdot \delta + \mathcal{X}v)$ is known. Since both $\mathcal{T}_h(\nabla \cdot \delta + \mathcal{X}v)$ and $\mathcal{A}^{-1/2}(\delta + \mathcal{A}\nabla v)$ are known, piecewise linear (discontinuous) functions, the remaining integrals required for the computation of $(MG)_j$ in (4.5) reduce to local linear combinations of integrals of piecewise quadratic functions over triangles.

The final step required for a preconditioned iteration is the action of a preconditioner for M . Let $G \in R^m$ and let G_v and G_m denote the coefficients of G which correspond to basis functions for V_h and W_h respectively. From Lemma 3.1 and (3.17), it follows that there are positive numbers C_0 and C_1 not depending on h which satisfy

$$(4.6) \quad \begin{aligned} C_0 [(\tilde{I}G_v) \cdot G_v + (\tilde{N}G_m) \cdot G_m] \\ \leq (MG) \cdot G \leq C_1 [(\tilde{I}G_v) \cdot G_v + (\tilde{N}G_m) \cdot G_m] \end{aligned}$$

for all $G \in R^m$. Here \tilde{I} is the mass matrix $\tilde{I}_{ij} = (\delta_i, \delta_j)$. It is not difficult to see that the matrix \tilde{I} is spectrally equivalent to the h^2 times the identity matrix I . It follows that the matrix M is spectrally equivalent to the block matrix

$$(4.7) \quad \begin{pmatrix} h^2 I & 0 \\ 0 & N \end{pmatrix}.$$

The blocks above correspond to the partitioning of the basis functions into those from V_h and W_h respectively. We use the inverse of the block matrix of (4.7) as a preconditioner for M . Thus, the application of the preconditioner to a vector $G \in R^n$ involves multiplying the V_h components of G by h^{-2} and applying the preconditioning process (N^{-1}) to the W_h components of G .

We now consider the amount of computational work involved in the above steps. Each step involves the computational effort required to evaluate the action of the preconditioner. The additional computations for each step require a fixed amount of work per node since the subsequent nodal computation only involves integration over the local support of basis functions. The work per step in the preconditioned iteration for (4.2) consists of the work for two B_h preconditioner evaluations plus work on the order of the number of unknowns n .

5. NUMERICAL EXPERIMENTS.

In this section, we report the results of numerical experiments involving the least-squares method developed earlier. In all of these experiments, the operator B_h was defined in terms of one multigrid V-cycle iteration. We first consider the rate of convergence for preconditioned iterative methods for computing the minimizer over the approximation subspace for (2.16). This convergence rate can be bounded in terms of the condition number of the precondition system which we shall report for three sets of coefficients. Subsequently, we will report the error in the approximation when the least-squares approach is applied to a problem with known solution.

We consider problem (2.1) when Ω is the unit square in two dimensional Euclidean space. For our reported results we shall only consider the case of constant

coefficients and take $\Gamma_N = \emptyset$. The unit square is first partitioned into a regular $n \times n$ mesh of smaller squares of size $h = 1/n$. The triangulation is then defined by breaking each of these squares into two triangles by connecting the lower left hand corner with the upper right. The approximation space W_h is defined to be the finite element space consisting of the continuous functions on Ω which are piecewise linear with respect to the triangulation and vanish on $\partial\Omega$. The approximation space V_h is defined to be the continuous vector valued functions which are piecewise linear with respect to the triangulation. Since $\Gamma_N = \emptyset$, no boundary conditions are imposed on V_h . This construction agrees with that discussed in Section 4.

In all of our examples, we shall use B_h to be the preconditioner for T_h corresponding one sweep of the multigrid V-cycle algorithm. We shall take n to be a power of two. To define the multigrid algorithm, one requires a sequence of coarser grid spaces. These spaces are defined by successively doubling the mesh size. Since the resulting sequence of triangulations is nested, so is the sequence of spaces,

$$M^1 \subset M^2 \subset \dots \subset M^j = W_h.$$

We use the point Gauss Seidel smoothing iteration on all spaces except the first (with mesh size $1/2$) on which we solve directly. The resulting multigrid iterative procedure is described in, for example, [27]. The multigrid preconditioner results from applying one step of the iterative procedure with zero starting iterate, [5]. The V-cycle uses one pre and post Gauss Seidel iteration sweep where the directions of the sweeps are reversed in the pre and post smoothing iterations. This results in a symmetric preconditioning operator B_h which satisfies

$$(5.1) \quad .74(T_h v, v) \leq (B_h v, v) \leq (T_h v, v) \quad \text{for all } v \in W_h.$$

The above bound was computed numerically and holds for $h = 1/n$ for $n = 4, 8, 16, \dots, 128$. The evaluation of B_h (i.e., N^{-1} applied to a vector where N is given by (4.3)) can be done in $O(n^2)$ operations and hence is proportional to the number of grid points on the mesh defining W_h .

Table 5.1.
Condition number of $\tilde{M}^{-1}M$ for three problems

h	Problem (a)	Problem (b)	Problem (c)
1/8	11.8	12.7	205
1/16	11.9	13.3	318
1/32	12.2	13.5	367
1/64	12.3	13.7	383
1/128	12.4	13.7	387

We first report condition numbers for the preconditioned system. As noted in Remark 2.3, we have some freedom in choosing the definition of \mathcal{T}_h . In all of the reported calculations, we used $\alpha = 1/3$ and $\beta = 2$ (see, Remark 2.3). Let M be

the stiffness matrix for the least-squares approximation as defined in Section 4. We replace the Gram matrix \tilde{I} in (4.6) by the diagonal matrix \bar{I} with diagonal entries given by

$$(\bar{I})_{ii} = 2\tilde{I}_{ii}.$$

Thus, the preconditioner for M involves the inversion of the block matrix

$$\tilde{M} = \begin{pmatrix} \bar{I} & 0 \\ 0 & N \end{pmatrix}.$$

Remark 3.2 and Lemma 3.1 show that the condition number of $\tilde{M}^{-1}M$ is bounded independently of the mesh size h . We report the actual condition numbers in Tables 5.1. We give the condition numbers for three different problems. For the first problem (a), the operator is given by the Laplacian, i.e.,

$$a_{ii} = 1, \quad a_{ij} = 0 \text{ for } i \neq j, \quad b_i = 0, \quad \text{and } c = 0.$$

The second column (b) corresponds to an operator with the coefficients

$$a_{ii} = 1, \quad a_{ij} = 0 \text{ for } i \neq j, \quad b_1 = 2, b_2 = 3, \quad \text{and } c = 0.$$

Finally, the third column (c) corresponds to (2.1) with coefficients

$$a_{ii} = 1, \quad a_{ij} = 0 \text{ for } i \neq j, \quad b_i = 0, \quad \text{and } c = -25.$$

We note a significant increase in the condition numbers in the case of Problem (c). The reason for this increase is that this problem is more singular than the other two. Let $v \in H_0^1(\Omega)$ be arbitrary and set $\delta = -\mathcal{A}\nabla v$ then Lemma 2.2 implies that

$$(5.2) \quad \|\nabla v\|^2 \leq c_0^{-1}(-\Delta v + \mathcal{X}v, v).$$

By Fourier analysis, it is straightforward to see that (5.2) holds for $c_0 = 1$ for Problem (a) whereas we must take $c_0 < 1/14$ in the case of Problem (c). This suggests that the condition number of Problem (c) should be at least 14 times larger than that of Problem (a). This explains much of the increase in condition number reported for Problem (c).

We next considered applying the least-squares method to approximately solve problems with a known analytic solution. We do this by starting with the solution

$$u = x(x - 1) \sin(\pi y).$$

This obviously satisfies the zero Dirichlet boundary condition. We generate the right hand side data by applying the differential operator to the solution. This resulting right hand side function is then interpolated and used as data in the least-squares algorithm as discussed in Section 4. We consider the Problems (a), (b) and (c) described above. We report the discrete $L^2(\Omega)$ norms of the errors e_σ

Table 5.2.
Error and iteration counts for Problem (a)

h	e_u	e_σ	Iterations
1/8	6.9×10^{-3}	2.9×10^{-2}	47
1/16	1.8×10^{-3}	8.5×10^{-3}	47
1/32	4.5×10^{-4}	2.7×10^{-3}	47
1/64	1.1×10^{-4}	9.2×10^{-4}	46
1/128	2.8×10^{-5}	3.1×10^{-4}	45

Table 5.3.
Error and iteration counts for Problem (b)

h	e_u	e_σ	Iterations
1/8	6.4×10^{-3}	2.9×10^{-2}	48
1/16	1.6×10^{-3}	8.6×10^{-3}	48
1/32	4.1×10^{-4}	2.8×10^{-3}	48
1/64	1.0×10^{-4}	9.2×10^{-4}	47
1/128	2.6×10^{-5}	3.2×10^{-4}	46

Table 5.4.
Error and iteration counts for Problem (c)

h	e_u	e_σ	Iterations
1/8	6.2×10^{-2}	3.7×10^{-1}	63
1/16	2.4×10^{-2}	1.3×10^{-1}	66
1/32	6.9×10^{-3}	3.7×10^{-2}	67
1/64	1.8×10^{-3}	9.5×10^{-3}	66
1/128	4.5×10^{-4}	2.4×10^{-3}	66

and e_u as well as the number of iterations required for numerical convergence of the preconditioned iteration.

The errors and iteration counts for Problem (a) are given in Table 5.2. The discrete $L^2(\Omega)$ convergence appears to be second order for u . In this case the method is behaving somewhat better than predicted by the theory since it is not known whether the preconditioner B_h satisfies (3.24) with c_0 independent h . The error in σ appears also to be converging somewhat faster than the first order rate guaranteed by Theorem 3.1.

The error and iteration counts for Problems (b) and (c) are reported in Tables 5.3 and 5.4 respectively. The results of Problem (a) and (b) are really rather similar. This suggests that the effect of the non-symmetric terms in Problem (b) is relatively small. We note that the moderate increase in the number of iterations for

convergence for Problem (c) does not reflect the large increase in condition number observed in column (c) of Table 5.1. One expects that the large condition numbers are due to a few small eigenvalues which correspond to eigenvalues of the continuous problem with small absolute value. In such a situation, the preconditioned conjugate gradient algorithm is known to perform much better than predicted by the worst case bound involving the condition number. The errors in Problem (c) are also larger by about a factor of ten than those observed for Problem (a) and (b).

REFERENCES

1. A.K. Aziz and I. Babuška, *Part I, survey lectures on the mathematical foundations of the finite element method*, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (A.K. Aziz, ed.), Academic Press, New York, NY, 1972, pp. 1–362.
2. A. K. Aziz, R. B. Kellogg, and A.B. Stephens, *Least-squares methods for elliptic systems*, Math. Comp. **44** (1985), 53–70.
3. I. Babuška, *On the Schwarz algorithm in the theory of differential equations of mathematical physics*, Tcheosl. Math. J. **8** (1958), 328–342 (in Russian).
4. R.E. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Math. Comp. **36** (1981), 35–51.
5. J.H. Bramble and J.E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
6. J.H. Bramble and R. Scott, *Simultaneous approximation in scales of Banach spaces*, Math. Comp. **32** (1978), 947–954.
7. J.H. Bramble and J. Xu, *Some estimates for weighted L^2 projections*, Math. Comp. **56** (1991), 463–476.
8. P. B. Bochev and M. D. Gunzburger, *Accuracy of least-squares methods for the Navier–Stokes equations*, Comput. Fluids **22** (1993), 549–563.
9. P. B. Bochev and M. D. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp. **63** (1994), 479–505.
10. F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, R.A.I.R.O. (1974), 129–151.
11. F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
12. Z. Cai, R. Lazarov, T. Manteuffel, and S. McCormick, *First-order system least-squares for second-order partial differential equations: Part I*, SIAM J. Numer. Anal. **31** (1994), 1785–1799.
13. P.G. Ciarlet, *Basic error estimates for elliptic problems*, Finite Element Methods : Handbook of Numerical Analysis, (P.G. Ciarlet and J.L. Lions, eds.), vol. II, North-Holland, New York, 1991, pp. 18–352.
14. G. F. Carey and Y. Shen, *Convergence studies of least-squares finite elements for first order systems*, Comm. Appl. Numer. Meth. **5** (1989), 427–434.
15. C. L. Chang, *Finite element approximation for grad-div type of systems in the plane*, SIAM J. Numerical Analysis **29** (1992), 590–601.
16. T. F. Chen, *On the least-squares approximations to compressible flow problems*, Numer. Meth. PDE's **2** (1986), 207–228.
17. T. F. Chen and G. J. Fix, *Least-squares finite element simulation of transonic flows*, Appl. Numer. Math. **2** (1986), 399–408.
18. M. Dauge, *Elliptic Boundary Value Problems on Corner Domains*, Lecture Notes in Mathematics, 1341, Springer-Verlag, 1988.
19. P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
20. T. J. R. Hughes and L. P. Franca, *A new finite element formulation for computational fluid dynamics. VII. The Stokes problems with various well-posed boundary conditions: symmetric*

- formulation that converges for all velocity pressure spaces*, Comput. Meth. Appl. Mech. Engrg. **65** (1987), 85–96.
21. T. J. R. Hughes, L. P. Franca, and M. Bulestra, *A new finite element formulation for computational fluid dynamics. V. Circumventing the Babuška–Brezzi condition: a stable Petrov–Galerkin formulation of the Stokes problem accomodating equal–order interpolations*, Comput. Meth. Appl. Mech. Engrg. **59** (1986), 85–99.
 22. D. C. Jespersen, *A least-square decomposition method for solving elliptic systems*, Math. Comp. **31** (1977), 873–880.
 23. B. N. Jiang and C. Chang, *Least-squares finite elements for the Stokes problem*, Comput. Meth. Appl. Mech. Engrg. **81** (1990), 13–37.
 24. B. N. Jiang and L. A. Povinelli, *Optimal least-squares finite element method for elliptic problems*, Comput. Meth. Appl. Mech. Engrg. **102** (1993), 199–212.
 25. O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flows*, Gordon and Breach, London, 1969.
 26. J.L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, vol. 1, Dunod, Paris, 1968.
 27. J. Mandel, S. McCormick and R. Bank, *Variational multigrid theory*, Multigrid Methods (S. McCormick, ed.), SIAM, Philadelphia, Penn., 1987, pp. 131–178.
 28. P. Neittaanmäki and J. Saranen, *On finite element approximation of the gradient for the solution to Poisson equation*, Numer. Math. **37** (1981), 131–148.
 29. J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
 30. A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least squares mixed finite elements for second order elliptic problems*, SIAM J. Numer. Anal. **31** (1994), 1368–1377.
 31. A. I. Pehlivanov, G. F. Carey, R. D. Lazarov, and Y. Shen, *Convergence analysis of least-squares mixed finite elements*, Computing **51** (1993), 111–123.
 32. A. I. Pehlivanov, G. F. Carey and P. S. Vassilevski, *Least-squares mixed finite element methods for non-selfadjoint elliptic problems: I. Error estimates*, University of Texas at Austin **CNA-271** (1994).
 33. P.A. Raviart and J.M. Thomas, *A mixed finite element method for 2-nd order elliptic problems*, Mathematical Aspects of Finite Element Methods, Lecture Notes in Mathematics, #606 (Eds. I. Galligani and E. Magenes), Springer-Verlag, New York, 1977, pp. 292–315.
 34. W. L. Wendland, *Elliptic Systems in the Plane*, Pitman, London, 1979.

DEPARTMENT OF MATHEMATICS
 CORNELL UNIVERSITY
 ITHACA, NY 14853
 AND DEPARTMENT OF MATHEMATICS
 TEXAS A&M UNIVERSITY
 COLLEGE STATION, TX 77843-3404
 E-MAIL : BRAMBLE@MATH.TAMU.EDU

DEPARTMENT OF MATHEMATICS
 TEXAS A&M UNIVERSITY
 COLLEGE STATION, TX 77843-3404
 E-MAIL : LAZAROV@MATH.TAMU.EDU

DEPARTMENT OF MATHEMATICS
 TEXAS A&M UNIVERSITY
 COLLEGE STATION, TX 77843-3404
 E-MAIL : PASCIAK@MATH.TAMU.EDU