Least-Squares for Second Order Elliptic Problems

James H. Bramble,* Raytcho D. Lazarov,* and Joseph E. Pasciak *

January 30, 1997

Abstract

In this paper, we introduce and analyze two least-squares methods for second order elliptic differential equations with mixed boundary conditions. These methods extend to problems which involve oblique derivative boundary conditions as well as nonsymmetric and indefinite problems as long as the original problem has a unique solution. With the methods to be developed, Neumann and oblique boundary conditions are imposed weakly and thus avoid compatiblity conditions on the finite element subspaces. The resulting least-squares approximations are unconditionally stable (no conditions on the step-size h) and will be shown to converge at an optimal rate.

The first least-squares method involves a discrete, computable H^{-1}-norm of the residual and stabilization terms consisting of the jumps at the interelement boundaries and a weighted elementwise L^2-norm of the residual over the finite elements. This method is developed without the introduction of additional problem variables.

The second method involves the use of the flux as an additional unknown. Although this method is similar to the least-squares method for first order systems introduced in [7], it differes in that discontinuous finite elements are allowed. It is also more general in that it extends to the oblique boundary problem.

1 Introduction.

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 [4], [3], [10], [14], [15], [17], [18]. In general, the corresponding problem is written as a system of partial differential equations of first order with possibly additional compatibility conditions. There is a variety of different approaches for introducing and studying least-squares methods for systems of first order.

Aziz, Kellogg and Stephens in [2] applied the general theory of elliptic boundary value problems of Agmon-Douglis-Nirenberg (ADN) and reduced

---

*Department of Mathematics, Texas A&M University, College Station, TX 77843–3368

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. This approach generalizes both the least-squares method of Jespersen [16], which is for the Poisson equation written as a $grad - div$ system, and the method of Wendland [24], which is for elliptic systems of Cauchy-Riemann type. Recently, Bochev and Gunzburger [3], [4], have extended the ADN approach to velocity-vorticity-pressure formulation of Stokes and Navier-Stokes equations.

Another 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. This approach has been used by Pehlivanov, Carey and their collaborators in [19] [20] and Cai, et al in [8], [9].

Recently, Bramble, Lazarov, and Pasciak in [7] have introduced and studied 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 least-squares functional studied there involved a discrete inner product related to the inner product in the Sobolev space $H^{-1}(\Omega)$. The use of such an 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.

In fact, the first computable H^{-1}-norm was used by R. Falk in [12] to treat in a weak form the incompressibility condition $\nabla \cdot \mathbf{u} = 0$ for Stokes problems. The essence of this early result is that the incompressibility condition is represented in the bilinear form by the sum of its L^2-norm and weighted by the factor h^{-2} times the discrete H^{-1}-norm. This leads to a stable scheme of optimal convergence order for linear finite elements.

In this paper, we provide a different approach from that of [7] for deriving least-squares methods for second order problems. This approach is more closely coupled to the Galerkin method. To contrast these two approaches, we study a somewhat more general problem.

Let Ω be a bounded polygonal or polyhedral domain in d dimensional Eu-

clidean space (for $d = 2$ or $d = 3$) with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$. We shall consider the following second order elliptic boundary value problem.

$$Lu = f \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \Gamma_D, \tag{1.1}$$
$$u_\nu + \alpha(x)u_{\mathbf{t}} + \beta(x)u = 0 \quad \text{on } \Gamma_N.$$

Here u_ν denotes the outward co-normal derivative on $\partial\Omega$ and $u_{\mathbf{t}}$ denotes the derivative along a tangential direction $\mathbf{t}$. The operator L is given by

$$Lu = -\nabla \cdot \mathcal{A}\nabla u + b \cdot \nabla u + cu$$
$$\equiv -\nabla \cdot \mathcal{A}\nabla u + \mathcal{X}u.$$

We assume that the matrix $\mathcal{A}(x)$ is symmetric, uniformly positive definite, and bounded. We also assume that $b \in (L^\infty(\Omega))^d$ and that β and $\alpha_{\mathbf{t}}$ belong to $L^\infty(\partial\Omega)$.

The development of least-squares methods is inherently coupled to various *a priori* inequalities. The methods of [7] are based on the inequality (see, Lemma 2.11 of [7])

$$C_0(\|\delta\|_0^2 + \|v\|_1^2) \leq \|\nabla \cdot \delta + \mathcal{X}v\|_{-1}^2 + \left\|\mathcal{A}^{-1/2}(\delta + \mathcal{A}v)\right\|_0^2 \tag{1.2}$$

The norms above are Sobolev norms (see Section 2). The inequality (1.2) as stated in [7] only holds for functions δ in $H_{div}(\Omega)$ satisfying $\delta \cdot n = 0$ on Γ_N. Here n is the outward unit normal vector. This means that the methods of [7] do not apply when either α or β is not identically zero.

The inequality (1.2) is related to the first order system which is equivalent to (1.1), i.e.,

$$\theta + \mathcal{A}\nabla u = 0$$
$$\nabla \cdot \theta + \mathcal{X}u = f. \tag{1.3}$$

In contrast, the methods developed in this paper are derived from the original second order operator and are based on the *a priori* inequality

$$\|v\|_1^2 \leq C \left\|\mathcal{L}v\right\|_{-1}^2.$$

The operator $\mathcal{L}$ above involves the boundary conditions as well as the differential part L. The resulting least-squares methods have a number of advantages. First, they extend to Robin and oblique derivative boundary conditions without complications, i.e., nonzero α and β. Secondly, since the *a priori* inequality only involves one variable, it is possible to develop a least-squares method without introducing the "flux" variable θ in (1.3). Least-squares methods involving the flux variable θ will also be developed for applications when θ is of

3

interest in itself. In contrast to those of [7], the flux approximation subspaces need not satisfy any boundary conditions on Γ_N.

The remainder of this paper is organized as follows. In Section 2, we define notation, place precise assumptions on the problem which we are studing and present some preliminary results. In Section 3 we develop the least-squares method without any new variables. The method involving the flux variable is developed in Section 4. Finally, the results of some numerical experiments are given in Section 5.

2 Preliminaries and further assumptions.

We assume that each edge (respectively face when $d = 3$) of Γ_N on which $\alpha(x)$ is not identically zero is completely surrounded by Γ_D. In addition, we assume that the tangential direction $\mathbf{t}$ varies smoothly on these regions. Finally, when $\alpha(x)$ is not identically zero, we assume that the boundary $\partial\Omega$ is Lipschitz continuous. This makes the problem variational (cf., §4.4.3) of [13]).

To describe and analyze the least-squares method, we shall use certain Sobolev spaces. For nonnegative integers s, let $H^s(\tilde{\Omega})$ denote the Sobolev space of order s defined on a domain $\tilde{\Omega}$ (see, e.g., [7]). The norm in $H^s(\tilde{\Omega})$ will be denoted by $\|\cdot\|_{s,\tilde{\Omega}}$. When $\tilde{\Omega} = \Omega$, we simplify the notation to $\|\cdot\|_s$. In the case of $L^2(\tilde{\Omega})$ the norm and the inner product will be denoted by $\|\cdot\|_{\tilde{\Omega}}$ and $(\cdot,\cdot)_{\tilde{\Omega}}$, respectively. The subscript will be dropped in the case when $\tilde{\Omega} = \Omega$.

For noninteger values of s, $H^s(\tilde{\Omega})$ is defined to be the functions in $H^{\underline{s}}(\tilde{\Omega})$ for which the norm

$$\|v\|_s := \left(\|v\|_{\underline{s}}^2 + \sum_{|\alpha|=\underline{s}} \int_\Omega \int_\Omega \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x-y|^{d+2\sigma}} \, dx \, dy \right)^{1/2}$$

is finite. Here $\underline{s}$ is the largest integer less than s and $\sigma = s - \underline{s}$. We use the same notations for the norms and inner products for vector valued functions. For example, if δ is a vector valued functions with component $\delta_i \in H^s(\Omega)$, then

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

Similarly,

$$(\delta, \eta) = \sum_{i=1}^d (\delta_i, \eta_i).$$

The weak solution of (1.1) is in the space W which is defined to be the closure of

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

with respect to the norm in $H^1(\Omega)$. In the case where Γ_D is empty and $\alpha \equiv \beta \equiv 0$, 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

$$\|v\|_{-1} = \sup_{\varphi \in W} \frac{(v, \varphi)}{\|\varphi\|_1} \tag{2.1}$$

is finite, where (v, φ) is the value of the functional at φ. If $v \in L^2(\Omega)$ then $(\cdot, \cdot)$ is identified with the $L^2(\Omega)$-inner product.

For any $u, v \in W$, we define the bilinear form

$$A(u, v) = (\mathcal{A}\nabla u, \nabla v) + (\mathcal{X}u, v) + \langle \alpha u_{\mathbf{t}} + \beta u, v \rangle_{\Gamma_N}. \tag{2.2}$$

Here $\langle \cdot, \cdot \rangle_{\Gamma_N}$ denotes the inner product in $L^2(\Gamma_N)$ or the pairing of certain appropriate Sobolev spaces with their duals.

In this paper, we shall repeatedly use the fact that the bilinear form A is bounded with respect to the norm in $H^1(\Omega)$. This follows from the §4.4.3 of [13] and is stated in the following lemma.

Lemma 2.1 *The bilinear form $A(\cdot, \cdot)$ is continuous on $W \times W$.*

Note that the oblique derivative term in (2.2) is not a compact perturbation. This term has the same strength as $(\mathcal{A}\nabla u, \nabla v)$.

The weak formulation of (1.1) is: Given $f \in H^{-1}(\Omega)$, find $u \in W$ satisfying

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

The adjoint weak formulation of (1.1) is: Given $f \in H^{-1}(\Omega)$, find $u \in W$ satisfying

$$A(\theta, u) = (f, \theta) \quad \text{for all } \theta \in W. \tag{2.4}$$

We assume that the solutions of (2.3) and (2.4) are unique. This means that if $v \in W$ and satisfies $A(v, \theta) = 0$ or $A(\theta, v) = 0$ for all $\theta \in W$, then $v = 0$.

The particular space $H^{-1}(\Omega)$ chosen above is related to the boundary conditions used in (1.1). Following [7], we give an alternative characterization of the norm in $H^{-1}(\Omega)$. Let $D(\cdot, \cdot)$ denote the inner product in W, i.e.,

$$D(u, \theta) = (u, \theta) + (\nabla u, \nabla \theta), \quad \text{for all } u, \theta \in W. \tag{2.5}$$

Let $T : H^{-1}(\Omega) \mapsto W$ be defined by $Tf = u$ where $u \in W$ is the unique function satisfying

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

As observed in [7],

$$(u, Tu) = \|u\|_{-1}^2 \quad \text{for all } u \in H^{-1}(\Omega).$$

Let the operator $\mathcal{L} : H^1(\Omega) \to H^{-1}(\Omega)$ be defined by the identity

$$(\mathcal{L}u, \varphi) = A(u, \varphi) \quad \text{for all } \varphi \in W. \tag{2.6}$$

The following lemma plays a fundamental role in the least-squares methods which will be developed in this paper.

Lemma 2.2 *There exists a constant C independent of $v \in W$ such that*

$$\|v\|_1 \le C \sup_{\varphi \in W} \frac{A(v, \varphi)}{\|\varphi\|_1} = C\|\mathcal{L}v\|_{-1} \tag{2.7}$$

and

$$\|v\|_1 \le C \sup_{\varphi \in W} \frac{A(\varphi, v)}{\|\varphi\|_1}. \tag{2.8}$$

Proof: First, we note that for v in W,

$$(\mathcal{X}v, v) = (b \cdot \nabla v, v) + (cv, v) \le C\|v\|\|v\|_1. \tag{2.9}$$

In addition,

$$\langle \alpha v_{\mathbf{t}} + \beta v, v \rangle_{\Gamma_N} = \langle \beta v, v \rangle_{\Gamma_N} - \frac{1}{2}\langle \alpha_{\mathbf{t}} v, v \rangle_{\Gamma_N}$$
$$\le C\|v\|_{0,\Gamma_N}^2 \le C\|v\|_{0,\Gamma_N}\|v\|_1. \tag{2.10}$$

We will prove the inequality

$$\|v\|_1 \le C \|\mathcal{L}v\|_{-1} \tag{2.11}$$

by contradiction. If (2.11) does not hold then for every integer $i > 0$ there is a $v_i \in W$ such that $\|v_i\|_1 = 1$ and $\|\mathcal{L}v_i\|_{-1} \le 1/i$. It follows from (2.9), (2.10) and obvious manipulations that for any $v \in W$,

$$C\|v\|_1^2 \le (\mathcal{A}\nabla u, \nabla v) = A(v, v) - (\mathcal{X}v, v) - \langle \alpha v_{\mathbf{t}} + \beta v, v \rangle_{\Gamma_N}$$
$$\le A(v, v) + C \left(\|v\|_{0,\Gamma_N} + \|v\| \right) \|v\|_1. \tag{2.12}$$

Thus,

$$\|v\|_1 \le C \left(\|\mathcal{L}v\|_{-1} + \|v\|_{0,\Gamma_N} + \|v\| \right) \le C \left(\|\mathcal{L}v\|_{-1} + \|v\|_s \right), \tag{2.13}$$

6

for any fixed s with $1/2 < s < 1$. In the last inequality we have used a well known trace inequality (cf. [1]). Since $\{v_i\}$ is a bounded sequence in $H^1(\Omega)$ and $H^1(\Omega)$ is compactly imbedded in $H^s(\Omega)$ for $s < 1$, there is a subsequence denoted again by $\{v_i\}$ which is convergent in $H^s(\Omega)$ to v. The inequality (2.13) applied to $v_j - v_i$ shows that v_i is a Cauchy sequence in W and therefore $v_j \to v$ in $W \subset H^1(\Omega)$. Clearly, by Lemma 2.1, for any $\varphi \in W$,

$$0 = \lim_{i \to \infty} A(v_i, \varphi) = A(v, \varphi).$$

Thus by uniqueness, $v = 0$ which contradicts

$$\|v\|_1 = \lim_{i \to \infty} \|v_i\|_1 = 1.$$

The argument for the second inequality of the lemma is analogous. This completes the proof of the Lemma 2.2. $\square$

Lemmas 2.1 and 2.2 together with the generalized Lax-Milgram Theorem imply the existence of the solution u of (2.3).

3 A Least-Squares Finite Element Method Without Additional Unknowns.

To approximately solve (1.1), we introduce the subspace $W_h \subset W$ indexed by h in the interval $0 < h < 1$. We do this by partitioning the domain Ω into a set of triangles or tetrahedra $\mathcal{T} = \{\tau\}$. For convenience, we will use the term triangle to refer to either a triangle when $d = 2$ or a tetrahedron when $d = 3$. Let h_τ denote the diameter of the triangle τ. The mesh parameter h is defined to be

$$h = \max_{\tau_i \in \mathcal{T}} h_{\tau_i}.$$

As usual, the boundaries of two triangles or tetrahedra will intersect at either a vertex, an entire edge or an entire face. We assume that the triangulations are locally quasi-uniform. By this we mean that there is a constant $0 < c < 1$ such that each triangle contains a ball of radius ch_{τ_i}. Spaces defined with respect to rectangular or parallelepiped partitioning of Ω pose no additional difficulty. For some integer $r \geq 2$, let W_h denote the functions which are piecewise polynomials of degree less than r with respect to the triangles, continuous on Ω, and vanish on Γ_D. There is a nodal basis associated with these spaces (see., e.g., [11]) and a corresponding nodal interpolation operator.

The following low order approximation and boundedness result can be proved. Given $\varphi \in W$, there exist $\varphi_h \in W_h$ and a constants C_2 not dependent on h and v such that

$$\sum_{\tau_i \in \mathcal{T}} \left\{ h_{\tau_i}^{-2} \|\varphi - \varphi_h\|_{\tau_i}^2 + \|\varphi - \varphi_h\|_{1,\tau_i}^2 \right\} \leq C_2 \|\varphi\|_1^2. \tag{3.1}$$

The arguments which can be used to prove this result in the quasi-uniform case are given in [6] and the general case is given in the proof of Lemma 4.1.1 of [25].

To develop the least-squares method, we shall need additional discrete norms and inner products. For $v \in W_h$, the discrete negative norm is given by

$$\|v\|_{-1,h} = \sup_{\varphi_h \in W_h} \frac{(v, \varphi_h)}{\|\varphi_h\|_1}. \tag{3.2}$$

This norm extends to a semi-norm on $H^{-1}(\Omega)$ which is bounded by the norm $\|\cdot\|_{-1}$. In addition, we define a weighted L^2 norm,

$$\|v\|_h = \left(\sum_i h_{\tau_i}^2 \|v\|_{\tau_i}^2 \right)^{1/2}. \tag{3.3}$$

The inner product corresponding to this norm shall be denoted by $(\cdot, \cdot)_h$. This norm will often be applied to derivatives of functions which are piecewise smooth with respect to the triangulation. In such cases, the differentiation will be done on an element by element basis.

We will also need edge norms and inner products. Let $\{\epsilon_i\}$ be the collection of the interior edges (respectively, faces) in the partitioning of Ω into triangles (respectively, tetrahedra). We introduce the bilinear form

$$\langle u, v \rangle_{h,I} = \sum_i h_{\tau(\epsilon_i)} \int_{\epsilon_i} uv \, ds \tag{3.4}$$

where the summation is over the set of all interior edges (respectively, faces) $\{\epsilon_i\}$. Here $\tau(\epsilon_i)$ is a triangle or tetrahedron which has ϵ_i as a edge or face. Similarly,

$$\langle u, v \rangle_{h,\Gamma_N} = \sum_i h_{\tau(\epsilon_i)} \int_{\epsilon_i} uv \, ds \tag{3.5}$$

where the summation is over the set of all edges (respectively, faces) on Γ_N. The corresponding seminorms are denoted

$$\|v\|_{h,I} = \langle u, v \rangle_{h,I}^{1/2} \quad \text{and} \quad \|v\|_{h,\Gamma_N} = \langle u, v \rangle_{h,\Gamma_N}^{1/2}. \tag{3.6}$$

Let the operator $\mathcal{L}_h : W \to W_h$ be defined by

$$(\mathcal{L}_h v, \varphi_h) = A(v, \varphi_h) \quad \text{for all } \varphi_h \in W_h. \tag{3.7}$$

We then have the following a priori inequality.

Lemma 3.1 *There exists a constant C not depending on h such that for any $v \in W_h$*

$$\|v\|_1^2 \le C \left\{ \|\mathcal{L}_h v\|_{-1,h}^2 + \|[v_\nu]\|_{h,I}^2 + \|Lv\|_h^2 \right.$$
$$\left. + \|v_\nu + \alpha v_{\mathbf{t}} + \beta v\|_{h,\Gamma_N}^2 \right\}. \tag{3.8}$$

Here $[u_\nu]$ denotes the jump in the co-normal derivative u_ν across an interior edge.

Proof: The argument is almost the same as the proof of Theorem 1 of [5]. By Lemma 2.2, it suffices to show that for $v \in W_h$,

$$|A(v,\phi)| \le C \left\{ \|\mathcal{L}_h v\|_{-1,h} + \|[v_\nu]\|_{h,I} + \|Lv\|_h \right.$$
$$\left. + \|v_\nu + \alpha v_{\mathbf{t}} + \beta v\|_{h,\Gamma_N} \right\} \|\varphi\|_1. \tag{3.9}$$

For $\varphi \in W$ let $\varphi_h \in W_h$ be a function which satisfies the inequalities (3.1). It follows that

$$|A(v,\varphi)| = |A(v,\varphi - \varphi_h) + A(v,\varphi_h)|$$
$$\le |A(v,\varphi - \varphi_h)| + \|\mathcal{L}_h v\|_{-1,h} \|\varphi_h\|_1 \tag{3.10}$$
$$\le |A(v,\varphi - \varphi_h)| + C\|\mathcal{L}_h v\|_{-1,h} \|\varphi\|_1.$$

Integrating by parts element-by-element gives

$$A(v,\varphi - \varphi_h) = \sum_{\tau_i \in \mathcal{T}} \left\{ \int_{\tau_i} (-\nabla \mathcal{A} \nabla v + \mathcal{X} v)(\varphi - \varphi_h)\, dx + \int_{\partial \tau_i} v_\nu (\varphi - \varphi_h)\, ds \right\}$$
$$+ \sum_{\epsilon_k \subset \Gamma_N} \int_{\epsilon_k} (\alpha v_{\mathbf{t}} + \beta v)(\varphi - \varphi_h)\, ds$$
$$= \sum_{\tau_i \in \mathcal{T}} \int_{\tau_i} Lv\, (\varphi - \varphi_h)\, dx + \sum_j \int_{\epsilon_j} [v_\nu](\varphi - \varphi_h)\, ds \tag{3.11}$$
$$+ \sum_{\epsilon_k \subset \Gamma_N} \int_{\epsilon_k} (v_\nu + \alpha v_{\mathbf{t}} + \beta v)(\varphi - \varphi_h)\, ds$$

We bound the terms on the right hand side of (3.11) separately. For the first,

$$\sum_{\tau_i \in \mathcal{T}} \left| \int_{\tau_i} Lv(\varphi - \varphi_h)\, dx \right| \le C \sum_{\tau_i \in \mathcal{T}} h_{\tau_i} \|Lv\|_{\tau_i} \|\varphi\|_{1,\tau_i} \tag{3.12}$$
$$\le C\|Lv\|_h \|\varphi\|_1.$$

For the second, we use the well-known inequality

$$\int_{\partial \tau_i} |\theta|^2\, ds \le C \left(h_{\tau_i}^{-1} \|\theta\|_{\tau_i}^2 + h_{\tau_i} \|\theta\|_{1,\tau_i}^2 \right). \tag{3.13}$$

Combining this with (3.1) gives

$$\sum_j \left| \int_{\epsilon_j} [v_\nu] \, (\varphi - \varphi_h) \, ds \right| \le C \sum_j h_{\tau(\epsilon_j)}^{1/2} \, |\|[v_\nu]\||_{\epsilon_j} \, \|\varphi\|_{1,\tau(\epsilon_j)}.$$
$$\le C \|[v_\nu]\|_{h,I} \|\varphi\|_1. \tag{3.14}$$

Similarly,

$$\sum_{\epsilon_k \subset \Gamma_N} \left| \int_{\epsilon_k} (v_\nu + \alpha v_{\mathbf{t}} + \beta v)(\varphi - \varphi_h) \, ds \right|$$
$$\le C \sum_{\epsilon_k \subset \Gamma_N} h_{\tau(\epsilon_k)}^{1/2} \|v_\nu + \alpha v_{\mathbf{t}} + \beta v\|_{\epsilon_k} \|\varphi\|_{1,\tau(\epsilon_k)} \tag{3.15}$$
$$\le C \|v_\nu + \alpha v_{\mathbf{t}} + \beta v\|_{h,\Gamma_N} \|\varphi\|_1.$$

Combining (3.10)–(3.15) proves (3.9) and hence completes the proof of the lemma. $\square$

Remark 3.1 The last three terms on the right hand side of (3.8) are stablizing terms. It is known that the Galerkin method is stable in $H^1(\Omega)$ (see, [21],[23]) if $h \le h_0$ is sufficiently small. This means that the Galerkin solution $V \in W_h$ satisfying

$$A(\theta, V) = D(\theta, u) \quad \text{for all} \ \theta \in W_h \tag{3.16}$$

can be bounded by
$$\|V\|_1 \le C\|v\|_1 \le C\|u\|_1$$

where v is the continuous solution of (3.16) ($v \in W$ satisfies (3.16) for all $\theta \in W$). If $u \in W_h$ with $h \le h_0$ then

$$\|u\|_1 = \frac{A(u,V)}{\|u\|_1} \le C\frac{A(u,V)}{\|V\|_1} \le C\|\mathcal{L}_h u\|_{-1,h}.$$

Thus, for $h \le h_0$, the stablizing terms are not necessary.

Before describing the least-squares method suggested by the previous lemma we provide an equivalent discrete negative norm. As in the continuous case, the discrete negative norm can be alternatively characterized in terms of a certain operator. Specifically, 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

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

Note that T_h is the finite element analogue of the operator T and that

$$\|v\|_{-1,h}^2 = (v, T_h v) \quad \text{for all } v \in H^{-1}(\Omega).$$

Note also that for $v \in L^2(\Omega)$, $(v, T_h v) = (T_h v, v)$.

Although one can develop the least-squares algorithms in terms of T_h, it is often more computationally efficient to replace this operator by a preconditioner B_h. To this end, we assume that we are given an operator $B_h : W_h \mapsto W_h$ which is a symmetric, positive definite operator with respect to the $L^2(\Omega)$ inner product and spectrally equivalent to T_h. This means that there are positive constants C_4, C_5 not depending on h and satisfying

$$C_4(T_h w, w) \le (B_h w, w) \le C_5(T_h w, w) \quad \text{for all } w \in W_h. \tag{3.17}$$

A good preconditioner is also computationally less expensive to evaluate. Examples of good preconditioners result from multigrid and domain decomposition algorithms.

We define the first least-squares bilinear form on $W_h \times W_h$ by

$$\begin{aligned} Q_1(u, v) \equiv &(B_h \mathcal{L}_h u, \mathcal{L}_h v) + (Lu, Lv)_h + \langle [u_\nu], [v_\nu] \rangle_{h,I} \\ &+ \langle u_\nu + \alpha u_\mathbf{t} + \beta u, v_\nu + \alpha v_\mathbf{t} + \beta v \rangle_{h,\Gamma_N}. \end{aligned} \tag{3.18}$$

The corresponding least-squares method follows.

Least-Squares Method without Additional Unknowns: Find $U \in W_h$ such that

$$(\mathcal{L}_h U - f, B_h \mathcal{L}_h V) + (LU - f, LV)_h + \langle [U_\nu], [V_\nu] \rangle_{h,I}$$

$$+ \langle U_\nu + \alpha U_\mathbf{t} + \beta U, V_\nu + \alpha V_\mathbf{t} + \beta V \rangle_{h,\Gamma_N} = 0, \quad \text{for all } V \in W_h, \tag{3.19}$$

or

$$Q_1(U, V) = (f, B_h \mathcal{L}_h V) + (f, LV)_h, \quad \text{for all } V \in W_h. \tag{3.20}$$

Note that the term $(f, LV)_h$ in the right hand side represents a sum of weighted L^2-inner products of f and LV over the finite elements $\tau \in \mathcal{T}$ and therefore it makes sense only for $f \in L^2(\Omega)$. If f does not belong to $L^2(\Omega)$ then we can reformulate the least-squares method by dropping the term $(f, LV)_h$ in (3.20). Thus, in the case of nonsmooth f, (3.19) is reduced to the following least-squares method: Find $U \in W_h$ such that

$$Q_1(U, V) = (f, B_h \mathcal{L}_h V), \quad \text{for all } V \in W_h. \tag{3.21}$$

As we shall show later this truncated formulation has order of convergence $O(h^{\gamma-1})$ for $u \in H^\gamma(\Omega)$ for $1 \le \gamma \le 2$.

Remark 3.2 The least-squares method (3.19) can be thought of as a stablized Galerkin method. By Remark 3.1, we can drop the stablizing terms when h is sufficiently small and use the least-squares method

$$(\mathcal{L}_h U - f, B_h \mathcal{L}_h V) = 0 \quad \text{for all} \ \ V \in W_h.$$

Since $B_h \mathcal{L}_h$ is invertible, this is the same as

$$(\mathcal{L}_h U - f, V) = 0 \quad \text{for all} \ \ V \in W_h,$$

which is the Galerkin method applied to (1.1).

The following Lemma is a consequence of Lemmas 2.1 and 3.1 and obvious manipulations.

Lemma 3.2 *The bilinear form $Q_1(u, v)$ is symmetric, coercive, and bounded in the H^1-norm on the finite element space W_h.*

3.1 Error Analysis

The above stability results lead to error estimates as we shall now demonstrate. We first introduce the following lemma.

Lemma 3.3 *Let $\hat{\tau}$ be a reference triangle of unit size and w be in $H^s(\hat{\tau})$ for $s > 3/2$. Then for any edge $\hat{\epsilon}$ of $\hat{\tau}$,*

$$\|w_\nu\|_{0,\hat{\epsilon}} \le C \|w\|_{s,\hat{\tau}},$$
$$\|w_{\mathbf{t}}\|_{0,\hat{\epsilon}} \le C \|w\|_{s,\hat{\tau}}.$$

Proof: We cannot simply apply trace theory since $\partial\hat{\tau}$ is not a $C^{1,1}$ boundary. Let $\tilde{\Omega}$ be a domain with $C^{1,1}$ boundary with $\hat{\epsilon} \subset \partial\tilde{\Omega}$. Let $\bar{w}$ be an H^s bounded extension of w in $H^s(R^d)$. It then follows that

$$\|w_\nu\|_{0,\hat{\epsilon}} \le \|\bar{w}_\nu\|_{0,\partial\tilde{\Omega}} \le C \|\bar{w}\|_{s,R^d} \le C \|w\|_{s,\tau}.$$

The second inequality follows in a similar manner. This completes the proof of the lemma. $\square$

We now state and prove an error estimate for the least-squares method in the case of a quasi-uniform mesh. The case of sufficently regular solutions and data is discussed in Theorem 3.1, while the case of solutions in $H^\gamma(\Omega)$ for $1 \le \gamma \le 2$ is discussed in Theorem 3.2.

Theorem 3.1 *Let $u \in H^\gamma(\Omega)$ for $2 \le \gamma \le r$ and U be the solution of the least-squares method defined by (3.19). Assume that the triangulation is quasi-uniform and set $h = \max_i h_{\tau_i}$. Then there exists a positive constant C not depending on u or h such that*

$$\|U - u\|_1 \le Ch^{\gamma-1} \|u\|_\gamma . \tag{3.22}$$

Theorem 3.2 *Assume that $u \in H^\gamma(\Omega)$ for $1 \le \gamma \le 2$, the triangulation is quasi-uniform with $h = \max_i h_{\tau_i}$, and U is the solution of the reduced least-squares method defined by (3.21). Then there exists a positive constant C not depending on u or h such that*

$$\|U - u\|_1 \le Ch^{\gamma-1} \|u\|_\gamma . \tag{3.23}$$

Let $E = U - V$ and $e = u - V$ where U is the solution to (3.19) and $V \in W_h$ is the interpolant (in W_h) of u. Before proving the above theorems, we introduce the following lemma based on the Bramble-Hilbert Lemma. Its proof uses Lemma 3.3 and is standard.

Lemma 3.4 *Let τ be a mesh triangle and assume that the solution u is in $H^s(\tau)$ for some s in $[2, r]$. Then,*

$$
\begin{aligned}
\|e\|_{l,\tau} &\le Ch_\tau^{s-l}\|u\|_{s,\tau}, \quad l = 0, 1, 2, \\
\|e_t\|_{0,\partial\tau} &\le Ch_\tau^{s-3/2}\|u\|_{s,\tau} \\
\|e_\nu\|_{0,\partial\tau} &\le Ch_\tau^{s-3/2}\|u\|_{s,\tau} \\
\|e\|_{0,\partial\tau} &\le Ch_\tau^{s-1/2}\|u\|_{s,\tau}.
\end{aligned}
\tag{3.24}
$$

The constant C above can be chosen independent of h_τ and τ.

Remark 3.3 For an edge ϵ_i, let $\bar\tau(\epsilon_i)$ denote the union of triangles which have ϵ_i as an edge. Then the above lemma implies that

$$\|[e_\nu]\|_{0,\epsilon_i} \le Ch_\tau^{s-3/2} \|u\|_{s,\bar\tau(\epsilon_i)} .$$

Proof (of Theorem 3.1): Clearly,

$$\|U - u\|_1 \le \|E\|_1 + \|e\|_1 .$$

By Lemma 3.1,

$$\|E\|_1 \le CQ_1(E, E)^{1/2} \le C\left[Q_1(U - u, U - u)^{1/2} + Q_1(e, e)^{1/2}\right].$$

It follows from (3.19) that

$$Q_1(U - u, V) = 0 \quad \text{for all } V \in W_h$$

13

and hence
$$Q_1(U - u, U - u) \leq Q_1(e, e).$$
Thus,
$$\|E\|_1 \leq CQ_1(e, e)^{1/2}.$$
Combining the above inequalities gives
$$\|U - u\|_1 \leq CQ_1(e, e)^{1/2} + \|e\|_1. \tag{3.25}$$

The estimate of $\|e\|_1$ follows immediately from the estimate (3.24) with $l = 1$ and summing over all triangles.

We estimate the first term in (3.25) by examining each term in (3.18). For the first, we have

$$
\begin{aligned}
(B_h \mathcal{L}_h e, \mathcal{L}_h e)^{1/2} &\leq C \|\mathcal{L}_h e\|_{-1,h} = C \sup_{\phi \in W_h} \frac{(\mathcal{L}_h e, \phi)}{\|\phi\|_1} \\
&= C \sup_{\phi \in W_h} \frac{A(e, \phi)}{\|\phi\|_1} \leq C \|e\|_1.
\end{aligned}
\tag{3.26}
$$

For the second, using (3.24) with $l = 2$ we get

$$(Le, Le)_h \leq C \sum_{\tau_i \in \mathcal{T}} h^2 \|e\|_{2,\tau}^2 \leq C h^{2(\gamma-1)} \|u\|_{\gamma,\Omega}^2. \tag{3.27}$$

Finally, since u is in $H^\gamma(\Omega)$ for $\gamma \in [2, r]$, it immediately follows from Lemma 3.4 and Remark 3.3 that for interior edges,

$$h_{\tau(\epsilon_i)}^{1/2} \|[e_\nu]\|_{0,\epsilon_i} \leq C h^{\gamma-1} \|u\|_{\gamma,\bar{\tau}(\epsilon_i)} \tag{3.28}$$

and for boundary edges,

$$h_{\tau(\epsilon_i)}^{1/2} \|e_\nu + \alpha e_{\mathbf{t}} + \beta e\|_{0,\epsilon_i} \leq C h^{\gamma-1} \|u\|_{\gamma,\bar{\tau}(\epsilon_i)}. \tag{3.29}$$

Theorem 3.1 follows from (3.26)–(3.29), Lemma 3.4 and summation. $\square$

Proof (of Theorem 3.2): To prove the result for the reduced least-squares method (3.21) when the solution $u \in H^\gamma(\Omega)$ for $1 \leq \gamma \leq 2$, we first prove stability in $H^1(\Omega)$, i.e.,

$$\|U\|_1 \leq C \|u\|_1. \tag{3.30}$$

By Lemma 3.1 and (3.21),

$$\|U\|_1^2 \leq CQ_1(U, U) = C(f, B_h \mathcal{L}_h U). \tag{3.31}$$

By (3.17),

$$(f, B_h \mathcal{L}_h U) \leq C \|f\|_{-1} \|\mathcal{L}_h U\|_{-1} \leq C \|u\|_1 \|U\|_1. \tag{3.32}$$

14

This proves (3.30) and the result for $\gamma = 1$ easily follows.

For $\gamma = 2$, it is enough to show that $(f, LU)_h \leq Ch\|u\|_2 \|U\|_1$. This is an immediate consequence of the definition of $(\cdot, \cdot)_h$, the inverse inequality and the obvious inequality $\|f\| \leq C\|u\|_{2,\Omega}$. Since $W \cap H^r(\Omega)$, for $1 \leq \gamma \leq 2$, is a Hilbert scale, it follows by interpolation that the result holds for $1 < \gamma < 2$. This completes the proof of Theorem 3.2 $\square$

We next illustrate that the stability estimate of Lemma 3.1 can be used to derive error estimates in the case of a refinement example. We illustrate this by considering a simple example involving piecewise linear finite element approximation. Specifically, we consider the problem: given $f \in L^2(\Omega)$ find u such that

$$\begin{aligned} -\Delta u &= f &&\text{in } \Omega, \\ u &= 0 &&\text{on } \partial\Omega. \end{aligned} \tag{3.33}$$

where Ω is defined to be the points in the square $(-1, 1) \times (-1, 1)$ which make an angle of absolute value less than $\pi - \omega$ with the positive x-axis. For $0 < \omega < \pi/2$, such a problem results in a singular solution of the form (see Theorem 4.4.3.7 of [13])

$$u = c(f)r^{\bar{\gamma}} \sin(\bar{\gamma}\phi)\eta(r) + w \tag{3.34}$$

where (r, ϕ) are the polar coordinates in the plane R^2, η is a C^∞ cutoff function which is zero for $r \geq 1$, $w \in H^2(\Omega)$ and $\bar{\gamma} = [2(1 - \omega/\pi)]^{-1}$. The constant $c(f)$ above satisfies

$$|c(f)| \leq C \|f\|_0 .$$

It follows that for $\gamma < \bar{\gamma}$

$$\|u\|_{1+\gamma} \leq C \|f\|_{-1+\gamma} \text{ and } \|u\|_{2,\Omega_r^*} \leq Cr^{\bar{\gamma}-1} \|f\|_0 .$$

Here

$$\Omega_r^* = \Omega \cap \{x \mid |x| > r\}.$$

We now define a mesh refinement strategy in terms of a "coarse" mesh parameter h_c. Such a strategy has been used, e.g., in [22]. All triangles shall be assumed to be shape regular (they satisfy a minimal angle condition). We introduce $R_j = 2^{-j}$ for $j = 0, \ldots, k$, where the positive integer k will be determined later. We define Ω_k to be the union of all triangles τ in $\mathcal{T}$ such that

$$\bar{\tau} \cap \{x \in \Omega \mid |x| \leq R_k\} \neq \emptyset.$$

Roughly speaking Ω_k is an R_k-neighborhood of the singular point $(0, 0)$. Next for $j = k - 1, \ldots, 0$ we define Ω_j to be the union of the triangles τ in $\mathcal{T}$ such that

$$\bar{\tau} \cap \{x \in \Omega \mid R_{j+1} \leq |x| \leq R_j\} \neq \emptyset.$$

We add all unassigned triangles of $\mathcal{T}$ to Ω_0.

We assume that the grid partition $\mathcal{T}$ satisfies the following property: There is a β in the interval $(1 - \bar{\gamma}, 1)$ such that for $j = 0, 1, \ldots, k$ and each triangle $\bar{\tau} \cap \Omega_j \neq \emptyset$,

$$C_0 R_j^{\beta} h_c \leq h_{\tau} \leq C_1 R_j^{\beta} h_c. \tag{3.35}$$

with some constants C_0 and C_1 independent of h_c and k. For example, $\beta = 0.5$ would be a good choice for all angles $0 < \omega < \pi/2$.

Theorem 3.3 *The least-squares method for problem (3.33), using a triangulation satisfying the above conditions, gives rise to an error estimate of the form*

$$\|u - U\|_1 \leq C h_c \|f\|_0$$

provided that k is chosen such that

$$R_k \leq h_c^{\frac{1-\gamma}{\beta\gamma}}, \quad i.e. \quad k \geq \frac{1-\gamma}{\beta\gamma} \cdot \frac{|\ln(h_c)|}{\ln(2)}. \tag{3.36}$$

Remark 3.4 It is easy to check that since $\beta < 1$, the number of triangles in such a mesh is bounded by a constant multiple of h_c^{-2}. Thus the theorem provides a quasi-optimal result in the sense that increasing the work by a fixed constant leads to the global accuracy one would expect on a smooth problem and a quasi-uniform mesh of size h_c.

Proof of Theorem 3.3. Let V, E and e be defined as in the proof of Theorem 3.1. Again, we need to bound the two terms on the right hand side of (3.25). By (3.26), this reduces to bounding

$$\|e\|_1^2 + (\Delta e, \Delta e)_h + \langle [e_\nu], [e_\nu] \rangle_{h,I}.$$

We first provide an estimate on Ω_j, for $j = 0, \ldots, k - 1$. Let $\mathcal{N}(\Omega_j)$ denote the union of triangles in $\mathcal{T}$ which have an edge which intersects Ω_j. Applying Lemma 3.4 and Remark 3.3 shows that

$$\begin{aligned}
\|e\|_{1,\Omega_j}^2 + \sum h_{\epsilon_i} \|[e_\nu]\|_{0,\epsilon_i}^2 &+ \sum_{\tau \cap \Omega_j \neq \emptyset} h_\tau^2 \|\Delta e\|_{0,\tau}^2 \\
&\leq C R_j^{2\beta} h_c^2 \|u\|_{2,\mathcal{N}(\Omega_j)}^2 \\
&\leq C R_j^{2(\beta+\bar{\gamma}-1)} h_c^2 \|f\|_0^2.
\end{aligned} \tag{3.37}$$

The first sum above is taken over the interior edges (with respect to Ω) in Ω_j.

In the case of Ω_k, we have

$$\|e\|_{1,\Omega_k}^2 + \sum h_{\epsilon_i} \|[e_\nu]\|_{0,\epsilon_i}^2 + \sum_{\tau \subset \Omega_k} h_\tau^2 \|\Delta e\|_{0,\tau}^2$$
$$\leq C R_k^{2\beta\gamma} h_c^{2\gamma} (\|u\|_{1+\gamma}^2 + \|f\|_0^2) \qquad (3.38)$$
$$\leq C R_k^{2\beta\gamma} h_c^{2\gamma} \|f\|_0^2.$$

The first sum above is taken over the interior edges (with respect to Ω) in Ω_k. Summing (3.37) over j and using (3.38) and the fact that $\beta > 1 - \bar{\gamma}$ gives

$$\|u - U\|_1^2 \leq C(h_c^2 + R_k^{2\beta\gamma} h_c^{2\gamma}) \|f\|_0^2.$$

The theorem follows taking k such that (3.36) holds. $\square$

4 A Least-Squares Finite Element Method involving the Flux

In this section, we develop a least-squares method for a first order system which is equivalent to (1.1) and involves the flux variable $\theta = -\mathcal{A}\nabla u$. To do this, we will prove a relevant *a priori* inequality.

Let δ be in $(L^2(\Omega))^d$. It follows from Lemma 2.2 that

$$\|v\|_1 \leq C \sup_{\varphi \in W} \frac{A(v, \varphi)}{\|\varphi\|_1}$$
$$= C \sup_{\varphi \in W} \frac{(\mathcal{A}\nabla v + \delta, \nabla\varphi) - (\delta, \nabla\varphi) + (\mathcal{X}v, \varphi) + \langle \alpha v_t + \beta v, \varphi \rangle_{\Gamma_N}}{\|\varphi\|_1}. \quad (4.1)$$

It is thus natural to introduce an operator $\mathcal{F} : (L^2(\Omega))^d \times W \mapsto H^{-1}(\Omega)$ defined by

$$(\mathcal{F}(\delta, v), \varphi) = -(\delta, \nabla\varphi) + (\mathcal{X}v, \varphi) + \langle \alpha v_t + \beta v, \varphi \rangle_{\Gamma_N}, \quad \text{for all} \ \ \varphi \in W.$$

Then by (4.1) and the triangle inequality

$$\|v\|_1 \leq C \left\{ \|\delta + \mathcal{A}\nabla v\|_0 + \|\mathcal{F}(\delta, v)\|_{-1} \right\}. \qquad (4.2)$$

We introduce a discrete approximation $\mathcal{F}_h$ to the operator $\mathcal{F}$. Define $\mathcal{F}_h :$ $L^2(\Omega) \times W \mapsto W_h$ by

$$(\mathcal{F}_h(\delta, v), \varphi) = -(\delta, \nabla\varphi) + (\mathcal{X}v, \varphi) + \langle \alpha v_t + \beta v, \varphi \rangle_{\Gamma_N}, \quad \text{for all} \ \ \varphi \in W_h.$$

The following lemma will be the basis for the least-squares methods studied in this section.

Lemma 4.1 *Let δ be a function which is in $(H^1(\tau_i))^d$ for each triangle $\tau_i \in \mathcal{T}$ and v be in W. Then,*

$$\|\delta\|_0^2 + \|v\|_1^2 \leq C \left\{ \left\| A^{-1/2}(\delta + A\nabla v) \right\|_0^2 + \|\mathcal{F}_h(\delta, v)\|_{-1,h}^2 + \|\nabla \cdot \delta + \mathcal{X}v\|_h^2 \right.$$

$$\left. + \|[\delta \cdot n]\|_{h,I}^2 + \| - \delta \cdot n + \alpha v_{\mathbf{t}} + \beta v\|_{h,\Gamma_N}^2 \right\}. \tag{4.3}$$

Proof: It follows immediately from (4.2) that

$$\|v\|_1 \leq C \left\{ \left\| A^{-1/2}(\delta + A\nabla v) \right\|_0 + \|\mathcal{F}(\delta, v)\|_{-1} \right\}.$$

Clearly then

$$\|\delta\|_0 \leq C \left\{ \left\| A^{-1/2}(\delta + A\nabla v) \right\|_0 + \|\mathcal{F}(\delta, v)\|_{-1} \right\}.$$

Now $\|\mathcal{F}(\delta, v)\|_{-1}$ may be estimated by the right hand side of (4.3). To this end for $\varphi \in W$ let φ_h be the element in W_h satisfying (3.1). Then by (3.1),

$$\begin{aligned} \|\mathcal{F}(\delta, v)\|_{-1} &= \sup_{\varphi \in W} \frac{(\mathcal{F}_h(\delta, v), \varphi_h) + (\mathcal{F}(\delta, v), \varphi - \varphi_h)}{\|\varphi\|_1} \\ &\leq C \|\mathcal{F}_h(\delta, v)\|_{-1,h} + \sup_{\varphi \in W} \frac{(\mathcal{F}(\delta, v), \varphi - \varphi_h)}{\|\varphi\|_1}. \end{aligned} \tag{4.4}$$

To estimate the last term in (4.4), we integrate by parts to get

$$(\mathcal{F}_h(\delta, v), \varphi - \varphi_h) = \langle \alpha v_{\mathbf{t}} + \beta v, \varphi - \varphi_h \rangle_{\Gamma_N}$$

$$+ \sum_{\tau_i \in \mathcal{T}} \left\{ (\nabla \cdot \delta + \mathcal{X}v, \varphi - \varphi_h)_{\tau_i} - \langle \delta \cdot n, \varphi - \varphi_h \rangle_{\partial \tau_i} \right\}.$$

The same estimates as used in Lemma 3.1 give that

$$(\mathcal{F}_h(\delta, v), \varphi - \varphi_h) \leq C \left\{ \|[\delta \cdot n]\|_{h,I} + \|\nabla \cdot \delta + \mathcal{X}v\|_h \right.$$

$$\left. + \| - \delta \cdot n + \alpha v_{\mathbf{t}} + \beta v\|_{h,\Gamma_N} \right\} \|\varphi\|_1.$$

Combining the above estimates completes the proof of the lemma. $\square$

In order to formulate a finite element least-squares method we need to define an approximation subspace $V_h \subset (L^2(\Omega))^d$. This we choose to consist of piecewise polynomials with respect to the triangulation $\mathcal{T}$. We only need $r - 1$ order approximation. Thus, we can take polynomials of degree $r - 2$.

The piecewise polynomial functions can be discontinuous across triangles and need not satisfy any boundary conditions on Γ_N.

The above lemma suggests a new least-squares method with flux unknowns. Suppose that θ is defined by (1.3) and u solves (1.1). Then

$$\theta + \mathcal{A}\nabla u = 0,$$
$$\mathcal{F}_h(\theta, u) = F_h,$$
$$[\theta \cdot n] = 0 \qquad \text{on interior edges,}$$
$$-\theta \cdot n + \alpha u_{\mathbf{t}} + \beta u = 0 \qquad \text{on } \Gamma_N.$$

Here F_h is the $L^2(\Omega)$ projection of f onto W_h. In order to simplify our notation we introduce the bilinear form $Q_2(\zeta, w; \delta, v)$ defined on $V_h \times W_h$ by

$$Q_2(\zeta, w; \delta, v) = (B_h\left(\mathcal{F}_h(\zeta, w)\right), \mathcal{F}_h(\delta, v)) + (\zeta + A\nabla w, A^{-1}\delta + \nabla v)$$
$$+ \langle [\zeta \cdot n], [\delta \cdot n] \rangle_{h,I} + \langle -\zeta \cdot n + \alpha w_{\mathbf{t}} + \beta w, -\delta \cdot n + \alpha v_{\mathbf{t}} + \beta v \rangle_{h,\Gamma_N}$$
$$+ (\nabla \cdot \zeta + \mathcal{X}w, \nabla \cdot \delta + \mathcal{X}v)_h. \tag{4.5}$$

Lemma 4.1 says that the bilinear form Q_2 is coercive and continuous in $V_h \times W_h$ equipped with the $(L^2)^d$ and H^1 norms. It suggests the following least-squares method.

Least-Squares Method with Flux Unknowns: Find $(\zeta_h, u_h) \in V_h \times W_h$ such that

$$Q_2(\zeta_h, u_h; \delta, v) = (f, \mathcal{F}_h(\delta, v)) + (f, \nabla \cdot \delta + \mathcal{X}v)_h,$$
$$\text{for all } (\delta, v) \in V_h \times W_h. \tag{4.6}$$

Obviously this formulation makes sense only for $f \in L^2(\Omega)$. For f not in $L^2(\Omega)$, one can make a simple modification of the scheme by dropping the second term on the right-hand-side of (4.6). In this case, we get the following least-squares method: Find $(\zeta_h, u_h) \in V_h \times W_h$ such that

$$Q_2(\zeta_h, u_h; \delta, v) = (f, \mathcal{F}_h(\delta, v)), \quad \text{for all } (\delta, v) \in V_h \times W_h. \tag{4.7}$$

Here $(f, \mathcal{F}_h(\delta, v))$ is the value of the linear functional at $\mathcal{F}_h(\delta, v) \in W$. As we show in the next theorem the reduced formulation (4.7) converges with a rate of $O(h^{\gamma-1})$ for $u \in H^{1+\gamma}(\Omega)$ and $1 \leq \gamma \leq 2$.

Remark 4.1 If $V_h \subset V = H(\text{div}; \Omega)$, then $[\zeta_h \cdot n]\mid_{\epsilon_j} = 0$ and the corresponding term in the least-squares method is identically zero. Here ϵ_j denotes an interior edge (face) of the finite element τ_j. If, in addition to $V_h \subset V = H(\text{div}; \Omega)$ $\alpha \equiv \beta \equiv 0$ and the functions in V_h satisfy the boundary condition $\zeta_h \cdot n = 0$ on Γ_N, then this method coinsides with the least-squares method of [7]. However, the method proposed in this paper is more general than that of [7] in two respects:

1. The spaces V_h are allowed to be discontinuous.

2. The Neumann and oblique derivative boundary conditions are naturally incorporated in the bilinear form and the functions in V_h need not satisfy these conditions.

Theorem 4.1 *Assume that the triangulation is quasi-uniform and set $h = \max_i h_{\tau_i}$. If $u \in H^\gamma(\Omega)$ for $2 \leq \gamma \leq r$ then the solution (ζ_h, u_h) of the least-squares method (4.6) satisfies the estimate*

$$\|\zeta_h - \theta\|_0 + \|u_h - u\|_1 \leq Ch^{\gamma-1} \|u\|_\gamma,\tag{4.8}$$

with a constant C independent of u, θ, and h. Furthermore, if $u \in H^\gamma(\Omega)$ for $1 < \gamma \leq 2$ then the reduced method (4.7) satisfies the error estimate (4.8).

5   Numerical Results.

We provide some numerical results in this section which illustrate the convergence behavior of the least-squares methods studied in the earlier sections. Specifically, we will consider the method which does not introduce any additional unknowns (3.19).

We only report the performance of the method on a model problem with oblique derivative boundary condition. Specifically, let Ω be the unit square in R^2. We consider the problem of approximating solutions to

$$\begin{aligned}
-\Delta u &= f &&\text{in } \Omega\\
u &= 0 &&\text{on } \Gamma_D\\
\frac{\partial u}{\partial n} + \frac{\partial u}{\partial \mathbf{t}} &= 0 &&\text{on } \Gamma_N.
\end{aligned}\tag{5.1}$$

Here Γ_N is the right-most edge of the boundary $(x = 1)$ and $\Gamma_D = \partial\Omega/\Gamma_N$.

For approximation, we use a regular grid of triangles and a subspace W_h of continuous piecewise linear functions defined with respect to this triangulation. Specifically, we partion the square into $N \times N$ smaller squares of size $h = 1/N$ and break each of the smaller squares into two triangles by connecting the lower left and upper right hand vertices. Functions in W_h vanish on Γ_D but not on Γ_N.

Instead of introducing a preconditioner for this example, we use $B_h = T_h$. Since the grid is regularly spaced, it is possible to compute the action of T_h by using the discrete Sine Transform. Indeed, let $w_h = T_h f$ and let $\tilde{w}_h$ be the function defined on $\tilde{\Omega} \equiv [0, 2] \times [0, 1]$ resulting from an even extension of w_h with respect to the line $x = 1$. Then, $\tilde{w}_h$ is the solution of

$$\tilde{D}(\tilde{w}_h, \phi) = (\tilde{f}, \phi) \quad \text{for all } \phi \in \tilde{W}_h.\tag{5.2}$$

h	Discrete L^2 error	Maximum norm error
1/8	.013	.024
1/16	.0043	.077
1/32	.0012	.0022
1/64	.00033	.00062
1/128	.000085	.00017

Table 5.1: Convergence behavior for a smooth solution

Here $\tilde{D}$ denotes the inner product in $H^1(\tilde{\Omega})$, $\tilde{W}_h$ is the approximation subspace resulting from reflecting the mesh and $\tilde{f}$ is the even extension of f. Functions in $\tilde{W}_h$ vanish on $\partial\tilde{\Omega}$. Since the mesh is uniform, the discrete Sine Transform provides an algorithm for expanding solution vectors in terms of the discrete eigenvectors corresponding to the stiffness matrix for (5.2). The solution is readily obtained by multipling by the inverse of the discrete eigenvalues and transforming back. The inverse transform is also a discrete Sine Transform. The cost of this evaluation is $O(N^2 \log(N))$ since the Sine Transform can be evaluated in terms of the Fast Fourier Transform.

For the first example, we consider a problem with smooth solution. Specifically, we take as a solution,

$$u = x(y - y^2)$$

along with the corresponding right-hand-side function f and non-homogeneous oblique boundary condition

$$\frac{\partial u}{\partial n} + \frac{\partial u}{\partial \mathbf{t}} = g \qquad \text{on } \Gamma_N.$$

Let U be the solution of (3.20). Table 5.1 gives the error $u - U$ in the discrete L^2 and L^∞ norms as a function of h. Note that the asymptotic convergence appears to be second order in both norms. This is better than predicted by the theory and only holds for smooth solutions.

Smooth solutions of (5.1) are somewhat artificial. In general, the solutions to the above problem fail to be in $H^2(\Omega)$ because of singular behavior at the vertex (1,1). We next illustrate the convergence behavior on a more realistic problem. Specifically, we set up a solution u of (5.1) which illustrates the typical singular behavior while resulting in right-hand-side data f in $L^2(\Omega)$. Let $\eta(r)$ be a C^2 function satisfying

$$\eta(r) = \begin{cases} 1 \text{ if } r \in [0, 1/4], \\ 0 \text{ if } r > 1/2. \end{cases}$$

h	Discrete L^2 error	Maximum norm error
1/8	.033	.11
1/16	.017	.063
1/32	.0078	.037
1/64	.0030	.024
1/128	.0012	.017
1/256	.00048	.012

Table 5.2: Convergence behavior for a singular solution

We then define u by

$$u(x) = r^{1/2} \sin(\phi/2)\, \eta(r)$$

where (r, ϕ) are the polar coordinates with origin at $(1, 1)$. It is easy to see that u satisfies (5.1) with $f = -\Delta u$ in $L^2(\Omega)$.

Table 5.2 gives the error $u - U$ in the discrete L^2 and L^∞ norms as a function of h. As expected, the convergence rate is less than second order but somewhat better than first order for the L^2-norm of the error.

References

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, Inc., New York, 1975.

[2] A. K. AZIZ, R. B. KELLOGG, AND A.B. STEPHENS, *Least-squares methods for elliptic systems*, Math. Comp., 44 (1985), pp. 53–70.

[3] P. B. BOCHEV AND M. D. GUNZBURGER, *Accuracy of least-squares methods for the Navier–Stokes equations*, Comput. Fluids, 22 (1993), pp. 549–563.

[4] ——, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[5] J. BRAMBLE AND J. PASCIAK, *Least-squares methods for Stokes equations based on a discrete minus one inner product*, J. Comp. and App. Math., (1997). (to appear).

[6] J. BRAMBLE AND J. XU, *Some estimates for weighted L^2 projections*, Math. Comp., 56 (1991), pp. 463–476.

[7] J. H. BRAMBLE, R. D. LAZAROV, AND J. E. PASCIAK, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. comp., 66 (1997). (to appear).

[8] 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), pp. 1785–1802.

[9] Z. CAI, T. MANTEUFFEL, AND S. MCCORMICK, *First-order system least-squares for second-order partial differential equations: Part II*, SIAM J. Numer. Anal., 34 (1997). (to appear).

[10] T. F. CHEN AND G. J. FIX, *Least-squares finite element simulation of transonic flows*, Appl. Numer. Math., 2 (1986), pp. 399–408.

[11] P. CIARLET, *Basic error estimates for elliptic problems*, in Finite Element Methods : Handbook of Numerical Analysis,, P. Ciarlet and J. Lions, eds., vol. II, New York, 1991, North-Holland, pp. 18–352.

[12] R. FALK, *A finite element method for the stationary Stokes equations using trial functions which do not satisfy div v = 0*, Math. Comp., 30 (1976), pp. 698–702.

[13] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[14] 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), pp. 85–96.

[15] 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), pp. 85–99.

[16] D. C. JESPERSEN, *A least-square decomposition method for solving elliptic systems*, Math. Comp., 31 (1977), pp. 873–880.

[17] B. N. JIANG AND C. CHANG, *Least-squares finite elements for the Stokes problem*, Comput. Meth. Appl. Mech. Engrg., 81 (1990), pp. 13–37.

[18] P. NEITTAANMÄKI AND J. SARANEN, *On finite element approximation of the gradient for the solution to Poisson equation*, Numer. Math., 37 (1981), pp. 131–148.

[19] A. I. PEHLIVANOV, G. F. CAREY, AND P. S. VASSILEVSKI, *Least-squares mixed finite element methods for non-selfadjoint elliptic problems: I. Error estimates*, Numerische Mathematik, 72 (1996), pp. 502–522.

[20] 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), pp. 1368–1377.

[21] A. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[22] A. Schatz and L. Wahlbin, *Maximum norm estimates in the finite element method on plane polygonal domains. Part 2, refinements*, Math. Comp., 33 (1979), pp. 465–492.

[23] A. Schatz and J. Wang, *Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions*, Math. Comp., 65 (1996), pp. 19–27.

[24] W. L. Wendland, *Elliptic Systems in the Plane*, Pitman, London, 1979.

[25] X. Zhang, *Studies in domain decomposition: multi-level methods and the biharmonic Dirichlet problem*, PhD thesis, NYU Courant Inst. Preprint # TR 1991-583.