FIRST-ORDER SYSTEM LEAST SQUARES FOR SECOND-ORDER PARTIAL DIFFERENTIAL EQUATIONS: PART I*

Z. CAI†, R. LAZAROV‡, T. A. MANTEUFFEL§, AND S. F. MCCORMICK§

Abstract. This paper develops ellipticity estimates and discretization error bounds for elliptic equations (with lower order terms) that are reformulated as a least–squares problem for an equivalent first–order system. The main result is the proof of ellipticity, which is used in a companion paper to establish optimal convergence of multiplicative and additive solvers of the discrete systems.

Key words. least–squares discretization, second–order elliptic problems, Rayleigh-Ritz, finite elements

1. Introduction. The purpose of this paper is to analyse the least-squares finite element method for second-order convection-diffusion equations written as a first-order system. In general, the standard Galerkin finite element methods applied to non-self-adjoint elliptic equations with significant convection terms exhibit a variety of deficiencies, including oscillations or nonmonotonocity of the solution and poor approximation of its derivatives. A variety of stabilization techniques, such as upwinding, Petrov-Galerkin, and stream-line diffusion approximations, have been introduced to eliminate these and other drawbacks of standard Galerkin methods. Yet, although significant progress has been made, convection-diffusion problems remain among the more difficult problems to solve numerically.

Substantial progress in overcoming the deficiencies of standard Galerkin methods has been made by employing least-squares principles. From a formal point of view, the least-squares method for the linear operator equation $Lp = f$ leads to an approximation of the normal equation $L^*Lp = L^*f$, where L^* is the adjoint of L in the inner product generated by the least-squares norm. Obviously, L^*L is self-adjoint and nonnegative. However, this approach loses its advantages when directly applied to approximating second-order elliptic problems, because of the resulting fourth-order system that requires much more smoothness of the solution, thereby preventing direct use of standard finite element spaces and effectively squaring the condition number of the discrete operator.

In the last thirty years, considerable attention has been directed towards developing alternative methods by introducing physically meaningful new dependent variables (fluxes, velocity, strains and stresses, etc.) that transform the corresponding second-order elliptic problem to a system of equations of first order. The resulting system can then be posed in a weak sense and approximated by finite element methods. In most cases, this procedure leads to a saddle point problem. Due largely to Babuška [2] and Brezzi [5], 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 if one wants

* This work was sponsored by the Air Force Office of Scientific Research under grant number AFOSR–86–0126, by the National Science Foundation under grant number DMS–8704169 and by the Department of Energy under grant number DE-FG03-93ER25165.

† Center for Applied Mathematical Sciences, Department of Mathematics, University of Southern California, 1042 W. 36th Place, DRB 155, Los Angeles, CA 90089–1113 (zcai@wuhan.usc.edu).

‡ Department of Mathematics, Texas A&M University, College Station, TX 77843–3368, and Institute of Mathematics, Bulgarian Academy of Sciences, Sofia (lazarov@math.tamu.edu).

§ Program in Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309–0526.

to have an unconditionally stable scheme of optimal approximation order. In particular, these spaces are usually chosen to satisfy the so-called inf-sup condition of Ladyzhenskaya-Babuška-Brezzi [23, 2, 5].

A general theory of the least-squares method for approximating elliptic boundary value problems of Agmon-Douglas-Nirenberg (ADN) type has been developed by Aziz, Kellog, and Stephens in [1]. The method involves the minimization of a least-squares functional that consists of a weighted sum of the residuals occuring in the equations and the boundary conditions. The weights occuring 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 [20], which is for systems arising from reformulating the Poisson equation by introducing the gradient of the solution as a new unknown, and the method of Wendland [28], which is for elliptic systems of Cauchy-Riemann type. However, this approach leads to an algebraic problem with condition number $O(h^{-4})$, where h is the grid size.

In a recent work by Bochev and Gunzburger [4], the ADN approach was extended to a velocity–vorticity–pressure formulation of Stokes flows with rigorous error analysis and optimal convergence rates. Further extension of these ideas to Navier–Stokes equations with interesting numerical experiments for 2-D problems appears in [3].

The most important applications of the least-squares method are connected with continuum mechanics (cf., [14, 21, 3, 24]). A common feature of all approaches is the introduction of new dependent variables: pressure in elastic problems (see [14]) or vorticity in flow problems (see [3]), for example. This has the effect of transforming the second-order elliptic problem into a system of first order. In most cases, the resulting system will not be of Petrovsky type (cf. [28]). On the other hand, one can add to this system a variety of compatibility conditions to overcome this limitation. For example, the system $-\operatorname{div} \mathbf{u} = f$, $\mathbf{u} = \mathbf{grad}\, p$ that decomposes the Poisson equation $-\Delta p = f$ can be effectively augmented by the compatibility equation $\mathbf{curl}\, \mathbf{u} = \mathbf{0}$ (see [22, 24]). Another possiblity, which was explored by Chen and Fix in [10, 11] for fluid flow, is to eliminate the variable p altogether in favor of a least-squares functional based on $\mathbf{curl}\, \mathbf{u} = \mathbf{0}$ and $\operatorname{div} \mathbf{u} + f = 0$.

The least-squares approach represents a general methodology that can produce a variety of algorithms, depending on such choices as the least-squares norm, the least-squares system, and the boundary treatment, and that can lead to formulations that have substantially different properties. For example, some formulations lead to self-adjoint problems, others yield optimal errors, others have unconditional stability, and others lead to optimal conditioning of the resulting discrete system. Among the deficiencies of some of the incarnations of this approach are non-self-adjointness, conditional stability, $O(h^{-4})$ condition number of the resulting discrete problem, and special essential boundary conditions. The purpose here is to develop a least–squares approach that does not exhibit these limitations. Another deficiency of conventional least-square methodology is the general lack of an efficient solver for the resulting discrete system. This is the subject of a companion paper [8].

Our attention here to least squares was motivated by the numerical experiments of Chen and Fix in [11] and of Carey and Shen in [9], where a second-order equation is solved using least-squares finite elements. In particular, the numerical experiments from [9] show that the finite element approximations of the unknown function p and its gradient $\mathbf{u}$ converge at rates depending only on the approximation properties of the finite element spaces of piecewise polynomials of degree k and r for p and $\mathbf{u}$, respectively. This is an interesting phenomenon since least-squares leads to a coupled system for p and $\mathbf{u}$. In [26], a theory that explains some of these computational results was developed. These ideas are extended in [25] to multidimensional self-adjoint equations of second order split into a system of first order. The main point in [25] is that the least-squares functional generates a bilinear form that is continuous and elliptic in a properly defined subspace of $H(\mathrm{div}) \times H^1$ and, therefore, standard finite element theory can be applied. In particular, one can use Raviart-Thomas mixed finite element spaces from [27] in order to approximate $H(\mathrm{div})$. However, any other finite-dimensional subspace of $H(\mathrm{div})$ can be used since the approximating space need not satisfy the inf-sup condition.

In this paper, our goal is to extend the theory in [25] to systems arising from splitting convection–diffusion and reaction–diffusion equations into a system of equations of first order. The main result here is in §3, where ellipticity with respect to the $H(\mathrm{div}) \times H^1$ norm of the bilinear form corresponding to the least–squares functional is established under very general assumptions on the differential operator. Other possible forms of the least-squares functional are suggested in §4. Approximation by the finite element method and its error analysis is then carried out in §5 in a straightforward manner, showing that the error in the $H(\mathrm{div}) \times H^1$ norm is optimal. Moreover, we show in §6 that the condition number for the resulting linear system is at most $O(h^{-2})$.

In the engineering literature, there is an impressive number of papers on least squares for a variety of applied problems. For a review of these applications up to the mid 70's, see [13].

2. Problem formulation. Let Ω be a bounded domain in $\mathbf{R}^n$, $n = 2$ or 3, with Lipschitz boundary $\partial\Omega$. We consider the boundary value problem

$$-\mathrm{div}\, A\, \mathbf{grad}\, p + Xp = f, \qquad \text{in } \Omega, \tag{2.1}$$

$$p = 0, \qquad \text{on } \Gamma_D, \tag{2.2}$$

$$\mathbf{n} \cdot A\, \mathbf{grad}\, p = 0, \qquad \text{on } \Gamma_N, \tag{2.3}$$

for $f \in L^2(\Omega)$, where $A(x)$ is an $n \times n$ symmetric matrix of functions in $L^2(\Omega)$, X is a linear differential operator of order at most one, $\partial\Omega = \Gamma_D \cup \Gamma_N$ is the boundary of Ω, and $\mathbf{n}$ is the outward unit vector normal to the boundary. We assume that A is uniformly positive definite and scaled appropriately, that is, there exist positive constants

$$0 < \lambda \le 1 \le \Lambda \tag{2.4}$$

such that

$$\lambda \boldsymbol{\xi}^T \boldsymbol{\xi} \le \boldsymbol{\xi}^T A \boldsymbol{\xi} \le \Lambda \boldsymbol{\xi}^T \boldsymbol{\xi} \tag{2.5}$$

for all $\boldsymbol{\xi} \in \mathbf{R}^n$ and almost all $x \in \overline{\Omega}$.

Possible choices for X include:

$$\begin{aligned} &\text{(a)} \quad Xp = \mathrm{div}\,(\mathbf{b}\, p), && \mathbf{b} = (b_1(x), \ldots, b_n(x)) \in (L^2(\Omega))^n\ ; \\ &\text{(b)} \quad Xp = \mathbf{a} \cdot \mathbf{grad}\, p + cp, && \mathbf{a} \in (L^2(\Omega))^n,\ c(x) \in L^2(\Omega). \end{aligned} \tag{2.6}$$

In particular, a real Helmholtz problem arises with $\mathbf{a} \equiv \mathbf{0}$ and $c(x) = -k^2$ in (2.6b).

The classical Sobolev spaces $H^m(\Omega)$, with m-th norm $\|\cdot\|_{m,\Omega}$ and seminorms $|\cdot|_{i,\Omega}$, $0 \le i \le m$, are employed and, as usual, $L^2(\Omega) = H^0(\Omega)$. Denote the corresponding norms on product space $(H^m(\Omega))^n$ by $\|\cdot\|_{m,\Omega,n}$ and $|\cdot|_{i,\Omega,n}$. We also need

$$H(\mathrm{div}) \equiv \left\{ \mathbf{v} \in (L^2(\Omega))^n : \ \mathrm{div}\,\mathbf{v} \in L^2(\Omega) \right\}.$$

We will be interested in the spaces (see (2.2) and (2.3))

$$\begin{aligned} \mathbf{W} &\equiv \{\mathbf{v} \in H(\mathrm{div}) : \mathbf{n} \cdot \mathbf{v} = 0 \text{ on } \Gamma_N\}, \\ V &\equiv \left\{q \in H^1(\Omega) : q = 0 \text{ on } \Gamma_D\right\} \ , \end{aligned} \tag{2.7}$$

with respective norms

$$\begin{aligned} \|\mathbf{v}\|^2_{H(\mathrm{div})} &= \|\mathbf{v}\|^2_{0,\Omega,n} + \|\mathrm{div}\,\mathbf{v}\|^2_{0,\Omega} \equiv \|\mathbf{v}\|^2_{\mathbf{W}}, \\ \|q\|^2_{1,\Omega} &= \|q\|^2_{0,\Omega} + \|\mathbf{grad}\,q\|^2_{0,\Omega,n} \equiv \|q\|^2_V. \end{aligned} \tag{2.8}$$

As usual, the inner product on $L^2(\Omega)$ is denoted by $(\cdot,\,\cdot)$, i.e., for any $p,\ q \in L^2(\Omega)$,

$$(p,\,q) = \int_\Omega p\,q\,dx. \tag{2.9}$$

Likewise, the inner product on $(L^2(\Omega))^n$ is denoted by

$$(\mathbf{u},\,\mathbf{v})_n = \sum_{i=1}^n \int_\Omega u_i v_i\,dx, \quad \forall\, \mathbf{u},\,\mathbf{v} \in (L^2(\Omega))^n. \tag{2.10}$$

Given a linear operator $X : H^1 \to L^2$, denote by $X^* : H^1 \to L^2$ its formal L^2–adjoint defined by

$$(Xp,\,q) = (p,\,X^* q) \tag{2.11}$$

for p and $q \in C_0^\infty(\Omega)$. In other words, we make the definition without regard to boundary conditions. Likewise, given a linear operator $X : H^1 \to (L^2)^n$, denote by $X^* : (H^1)^n \to L^2$ its formal L^2–adjoint defined by

$$(Xq,\,\mathbf{v})_n = (q,\,X^*\mathbf{v}) \tag{2.12}$$

for all $q \in C_0^\infty(\Omega)$ and $\mathbf{v} \in (C_0^\infty(\Omega))^n$. Define the operator $\nabla : H^1 \to (L^2)^n$ by

$$\nabla q \equiv \mathbf{grad}\,q \equiv \left(\frac{\partial q}{\partial x_1}, \frac{\partial q}{\partial x_2}, \ldots, \frac{\partial q}{\partial x_n} \right) . \tag{2.13}$$

Its formal adjoint $\nabla^* : (H^1)^n \to L^2$ is then defined as

$$\nabla^* \mathbf{v} = -\mathrm{div}\,\mathbf{v} = -\left(\frac{\partial \mathbf{v}_1}{\partial x_1} + \ldots + \frac{\partial \mathbf{v}_n}{\partial x_n} \right) . \tag{2.14}$$

Notice that if we restrict ∇ to V and ∇^* to $\mathbf{W}$, then ∇^* is, in fact, the $L^2(\Omega)$ adjoint of ∇.

Consider the least-squares functional

$$G(\mathbf{u},\, p;\, f) = \|A\nabla\, p - \mathbf{u}\|^2_{0,\,\Omega,\,n} + \|\nabla^*\mathbf{u} + Xp - f\|^2_{0,\,\Omega}\,, \tag{2.15}$$

and the associated bilinear form

$$\mathcal{F}(\mathbf{u},\, p;\, \mathbf{v},\, q) = \left(\begin{pmatrix} -I & A\nabla \\ \nabla^* & X \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix},\, \begin{pmatrix} -I & A\nabla \\ \nabla^* & X \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right)_{n+1}. \tag{2.16}$$

Assuming that (2.1)–(2.3) is uniquely solvable in H^1, it is easy to see that the (unique) minimum of $G(\mathbf{u},\, p;\, f)$ over $(\mathbf{u},\, p) \in \mathbf{W} \times V$, for fixed $f \in L^2$, corresponds to the solution of (2.1)–(2.3) in the sense that p solves these equations and $\mathbf{u} = A\nabla\, p$. The main result of this paper will be to show that $G(\mathbf{u},\, p;\, f)$ is elliptic with respect to the $H(\mathrm{div}) \times H^1$ norm, that is, there exist positive constants α and β such that

$$\alpha \left(\|\mathbf{u}\|^2_{H(\mathrm{div})} + \|p\|^2_{1,\,\Omega} \right) \le G(\mathbf{u},\, p;\, 0) \le \beta \left(\|\mathbf{u}\|^2_{H(\mathrm{div})} + \|p\|^2_{1,\,\Omega} \right) \tag{2.17}$$

for every $\mathbf{u} \in \mathbf{W}$ and $p \in V$.

The uniform bounds on A (see (2.5)) imply that the functional in (2.15) is equivalent to the functional

$$\widehat{G}(\mathbf{u},\, p;\, f) = \|A^{1/2}\nabla\, p - A^{-1/2}\mathbf{u}\|^2_{0,\,\Omega,\,n} + \|\nabla^*\mathbf{u} + Xp - f\|^2_{0,\,\Omega} \tag{2.18}$$

in the sense that

$$\lambda \widehat{G}(\mathbf{u},\, p;\, f) \le G(\mathbf{u},\, p;\, f) \le \Lambda \widehat{G}(\mathbf{u},\, p;\, f) \tag{2.19}$$

for every $\mathbf{u} \in \mathbf{W}$, $p \in V$ and $f \in L^2(\Omega)$. In our proof we will be working with $\widehat{G}(\mathbf{u},\, p;\, f)$, which can be written as

$$\begin{aligned} \widehat{G}(\mathbf{u},\, p;\, f) = \Bigg(&\begin{pmatrix} -A^{-1/2} & A^{1/2}\nabla \\ \nabla^* & X \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix}, \\ &\begin{pmatrix} -A^{-1/2} & A^{1/2}\nabla \\ \nabla^* & X \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix} \Bigg)_{n+1}, \end{aligned}$$

and the associated bilinear form

$$\begin{aligned} &\widehat{\mathcal{F}}(\mathbf{u},\, p;\, \mathbf{v},\, q) \\ &\quad = \left(\begin{pmatrix} -A^{-1/2} & A^{1/2}\nabla \\ \nabla^* & X \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix},\, \begin{pmatrix} -A^{-1/2} & A^{1/2}\nabla \\ \nabla^* & X \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right)_{n+1}. \end{aligned} \tag{2.20}$$

If we restrict our attention to functions with the correct properties so that we can, for the moment, ignore the issues of smoothness and boundary conditions, then this form can be written as

$$\widehat{\mathcal{F}}(\mathbf{u},\, p;\, \mathbf{v},\, q) = \left(\begin{pmatrix} A^{-1} + \nabla\nabla^* & \nabla(X - I) \\ (X^* - I)\nabla^* & \nabla^* A \nabla + X^* X \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix},\, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right)_{n+1}. \tag{2.21}$$

The $H(\mathrm{div}) \times H^1$ norm is induced by the inner product

$$(\mathbf{u},\, p;\, \mathbf{v},\, q)_{H(\mathrm{div})\times H^1} = (\mathbf{u},\, \mathbf{v})_n + (\nabla^*\mathbf{u},\, \nabla^*\mathbf{v}) + (p,\, q) + (\nabla p,\, \nabla q)_n. \tag{2.22}$$

Similar in spirit to the definition of $\mathcal{F}$, we define an inner product

(2.23)
$$\mathcal{S}(\mathbf{u},\, p;\, \mathbf{v},\, q) = (A^{-1/2}\mathbf{u},\, A^{-1/2}\mathbf{v})_n + (\nabla^*\mathbf{u},\, \nabla^*\mathbf{v}) + (p,\, q) + (A^{1/2}\nabla p,\, A^{1/2}\nabla q)_n.$$

which satisfies equivalence bounds

$$\frac{1}{C}\mathcal{S}(\mathbf{u},\, p;\, \mathbf{u},\, p) \le (\mathbf{u},\, p;\, \mathbf{u},\, p)_{H(\mathrm{div})\times H^1} \le C\mathcal{S}(\mathbf{u},\, p;\, \mathbf{u},\, p) \tag{2.24}$$

where $C = \max\{1/\lambda,\, \Lambda\}$. (Recall $0 < \lambda \le 1 \le \Lambda$.) Again taking liberties with smoothness and boundary conditions, we can write

$$\mathcal{S}(\mathbf{u},\, p;\, \mathbf{v},\, q) = \left(\begin{pmatrix} A^{-1} + \nabla\nabla^* & 0 \\ 0 & I + \nabla^* A \nabla \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix},\, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right). \tag{2.25}$$

Notice the similarity between the diagonal terms of the matrices defining the bilinear forms in (2.21) and (2.25). The main result of this paper can be viewed as demonstrating the spectral equivalence of these two bilinear forms. Of course, we must pay close attention to the issues of smoothness and boundary conditions to make sure the equivalence holds on all of $\mathbf{W} \times V$.

3. Main result. The main objective of this paper is to establish equivalence of the forms $\mathcal{S}$ in (2.23) and $\mathcal{F}$ in (2.16) on $\mathbf{W}\times V$. If problem (2.1)–(2.3) is invertible and $\partial\Omega$ and the coefficients of A and X are $C^{1,1}$ (first derivative is Lipschitz continuous), then the problem will be $H^2(\Omega)$ regular and the proof flows easily. However, with a little care the result can be proved under more general hypotheses for (2.1)–(2.3). In particlular, we wish to include the important case in which A is discontinuous and $\partial\Omega$ is polygonal. In the next few remarks we discuss certain results that are central to the proof.

Remark 3.1. Our theory assumes that either $\Gamma_D \neq \emptyset$ or an additional constraint is imposed on V, such as $\int_\Omega p\, dx = 0$, so that a Poincaré-Friedrichs inequality holds; that is, there exists a constant $d > 0$ depending only on the domain Ω and the uniform bounds on A (see (2.5)) such that

$$\|p\|^2_{0,\Omega} \le d\|A^{1/2}\nabla p\|^2_{0,\Omega} \tag{3.1}$$

for $p \in V$.

Remark 3.2. We will also assume that (2.1)–(2.3) is invertible in $H^1(\Omega)$; that is, for any $f \in H^{-1}(\Omega)$ there is a unique $p \in V$ such that

$$(A\nabla p,\ \nabla v)_n + (Xp,\ v) = (f, v) \tag{3.2}$$

for every $v \in V$. This is true for the case $X = 0$, $X = I$, or X is given by (2.6a or b) (see [17]). If $f \in L^2(\Omega)$, then the solution p will satisfy $A\nabla p \in \mathbf{W}$ (see [15]). We will make particular use of this result with $X \equiv 0$ and $X \equiv I$. With this in mind, we make the definition

$$D = \{p \in V : A\nabla p \in \mathbf{W}\}, \tag{3.3}$$

and note that invertibility of (2.1)–(2.3) means that it defines a bijection from $L^2(\Omega)$ to D.

Remark 3.3. While it is almost implicit in the assumption that (2.1)–(2.3) is invertible, we will make overt use of the bound

$$\|Xp\|_{0,\Omega} \le \eta\|A^{1/2}\nabla p\|_{0,\Omega} \tag{3.4}$$

for some $\eta > 0$ and every $p \in D$. This bound clearly follows from (3.1) if X has the form (2.6a) and the coefficients are sufficiently smooth, or if X has the form (2.6b).

Remark 3.4. Finally, we will make use of the inequality

$$\|\nabla^* A \nabla p + Xp\|_{0,\Omega}^2 \geq \delta \|\nabla^* A \nabla p\|_{0,\Omega}^2 \tag{3.5}$$

for some $\delta > 0$ and for all $p \in D$. Its validity follows from the results in [16], where it is shown that if two uniformly elliptic operators are invertible in $H^1(\Omega)$ and have the same leading part and boundary conditions, then they are $L^2(\Omega)$-norm equivalent, even in the absence of $H^2(\Omega)$ regularity.

In addition, since $(\nabla^* A \nabla)^{-1}$ is bounded in L^2, there is some constant K such that

$$\|p\|_{0,\Omega} \leq K \|\nabla^* A \nabla p\|_{0,\Omega} \tag{3.6}$$

for every $p \in D$. Thus, by the definition of D and integration by parts, we have

$$\|A^{1/2} \nabla p\|_{0,\Omega}^2 = (\nabla^* A \nabla p,\, p) \leq K(\nabla^* A \nabla p,\, \nabla^* A \nabla p) = K \|\nabla^* A \nabla p\|_{0,\Omega}^2$$

for all $p \in D$. Together with (3.5), this yields

$$\|\nabla^* A \nabla p + X\phi\|_{0,\Omega}^2 \geq \gamma(\|\nabla^* A \nabla p\|_{0,\Omega}^2 + \|A^{1/2} \nabla p\|_{0,\Omega}^2) \tag{3.7}$$

for all $p \in D$, where $\gamma = \delta/(K+1)$.

THEOREM 3.1. *Assume that V implies the Poincaré-Friedrichs inequality (3.1), that X satisfies the bound (3.4), and that (2.1)–(2.3) is invertible in $H^1(\Omega)$. Then there exist positive constants α and β such that*

$$\mathcal{F}(\mathbf{u},\, p;\, \mathbf{v},\, q) \leq \beta \left(\|\mathbf{u}\|_{H(\mathrm{div})}^2 + \|p\|_{1,\Omega}^2\right)^{1/2} \left(\|\mathbf{v}\|_{H(\mathrm{div})}^2 + \|q\|_{1,\Omega}^2\right)^{1/2} \tag{3.8}$$

for every $\mathbf{u}, \mathbf{v} \in \mathbf{W}$ and $p, q \in V$ and

$$\mathcal{F}(\mathbf{u},\, p;\, \mathbf{u},\, p) \geq \alpha \left(\|\mathbf{u}\|_{H(\mathrm{div})}^2 + \|p\|_{1,\Omega}^2\right) \tag{3.9}$$

for every $\mathbf{u} \in \mathbf{W}$ and $p \in V$.

Proof. Continuity of $\mathcal{F}$ (3.8) follows directly from assumption (2.5), definitions (2.8) and (2.16), and (3.4). The proof of the lower bound (3.9) is established in a series of steps.

Formulate an equivalent problem: By (2.19) and (2.24) it is sufficient to find a positive constant α_0 such that

$$\alpha_0 \mathcal{S}(\mathbf{u},\, p;\, \mathbf{u},\, p) \leq \widehat{\mathcal{F}}(\mathbf{u}, p; \mathbf{u}, p) \tag{3.10}$$

for all $\mathbf{u} \in \mathbf{W}$ and $p \in V$, where $\mathcal{S}$ and $\widehat{\mathcal{F}}$ are defined in (2.23) and (2.20), respectively. Then (3.9) would follow with $\alpha = \alpha_0 \lambda \min\{\lambda, 1/\Lambda\}$.

Remove divergence free part: For any $\mathbf{u} \in \mathbf{W}$, we may use a decomposition analogous to Weyl's theorem [15]:

$$\mathbf{u} = A\nabla\phi + \boldsymbol{\psi}, \tag{3.11}$$

where $\phi \in D$ and

$$\nabla^* \boldsymbol{\psi} = 0, \tag{3.12}$$

$$\mathbf{n}\cdot\boldsymbol{\psi} = 0, \qquad \text{on } \Gamma_N. \tag{3.13}$$

This is accomplished by choosing ϕ to be the weak solution of

$$\nabla^* A\nabla\phi = \nabla^*\mathbf{u}, \tag{3.14}$$

$$\phi = 0, \qquad \text{on } \Gamma_D, \tag{3.15}$$

$$\mathbf{n}\cdot A\nabla\,\phi = 0, \qquad \text{on } \Gamma_N. \tag{3.16}$$

From Remark 3.2 it follows that $\phi \in D$, so $A\nabla\phi \in \mathbf{W}$. Setting $\boldsymbol{\psi} = \mathbf{u} - A\nabla\,\phi$, then $\boldsymbol{\psi} \in \mathbf{W}$ and (3.11)–(3.13) follows.

Equations (2.16), (2.22), and (2.23) yield

$$\mathcal{S}(\boldsymbol{\psi},\,0;\,\boldsymbol{\psi},\,0) = \widehat{\mathcal{F}}(\boldsymbol{\psi},\,0;\,\boldsymbol{\psi},\,0) = (A^{-1/2}\boldsymbol{\psi},\,A^{-1/2}\boldsymbol{\psi})_n. \tag{3.17}$$

Consider the cross-product terms. Since $\boldsymbol{\psi} \in \mathbf{W}$ and $\phi, p \in V$, we can integrate by parts (see [15]) to obtain

$$\mathcal{S}(A\nabla\phi,\,p;\,\boldsymbol{\psi},\,0) = (\nabla\phi,\,\boldsymbol{\psi})_n = (\phi,\,\nabla^*\boldsymbol{\psi}) = 0, \tag{3.18}$$

$$\widehat{\mathcal{F}}(A\nabla\phi,\,p;\,\boldsymbol{\psi},\,0) = (\nabla p - \nabla\phi,\,\boldsymbol{\psi})_n = (p-\phi,\,\nabla^*\boldsymbol{\psi}) = 0. \tag{3.19}$$

This yields

$$\begin{aligned}
\mathcal{S}(A\nabla\phi+\boldsymbol{\psi},\,p;\,A\nabla\phi+\boldsymbol{\psi},\,p) &= \mathcal{S}(A\nabla\phi,\,p;\,A\nabla\phi,\,p) + \mathcal{S}(\boldsymbol{\psi},\,0;\,\boldsymbol{\psi},\,0),\\
\widehat{\mathcal{F}}(A\nabla\phi+\boldsymbol{\psi},\,p;\,A\nabla\phi+\boldsymbol{\psi},\,p) &= \widehat{\mathcal{F}}(A\nabla\phi,\,p;\,A\nabla\phi,\,p) + \mathcal{S}(\boldsymbol{\psi},\,0;\,\boldsymbol{\psi},\,0).
\end{aligned}$$

Thus, it only remains to show that there exists a positive constant $\alpha_0 \le 1$ such that

$$\alpha_0\mathcal{S}(A\nabla\phi,\,p;\,A\nabla\phi,\,p) \le \widehat{\mathcal{F}}(A\nabla\phi,\,p;\,A\nabla\phi,\,p) \tag{3.20}$$

for every $\phi \in D,\;\; p \in V$.

Define $\mathcal{S}_0$ and $\mathcal{F}_0$ and show that $c\,\mathcal{S}_0 \le \mathcal{F}_0$: Let

$$\begin{aligned}
&\mathcal{F}_0(\phi,\,p;\,\varphi,\,q)\\
&\quad= \widehat{\mathcal{F}}(A\nabla\phi,\,p;\,A\nabla\varphi,\,q)\\
&\quad= \left(\begin{pmatrix} -A^{1/2}\nabla & A^{1/2}\nabla \\ \nabla^*A\nabla & X \end{pmatrix}\begin{pmatrix}\phi\\ p\end{pmatrix},\,\begin{pmatrix} -A^{1/2}\nabla & A^{1/2}\nabla \\ \nabla^*A\nabla & X \end{pmatrix}\begin{pmatrix}\varphi\\ q\end{pmatrix}\right)_{n+1}
\end{aligned}$$

and

$$\mathcal{S}_0(\phi,\,p;\,\varphi,\,q) = \left(\begin{pmatrix} 0 & A^{1/2}\nabla \\ \nabla^*A\nabla & 0 \end{pmatrix}\begin{pmatrix}\phi\\ p\end{pmatrix},\,\begin{pmatrix} 0 & A^{1/2}\nabla \\ \nabla^*A\nabla & 0 \end{pmatrix}\begin{pmatrix}\varphi\\ q\end{pmatrix}\right)_{n+1}.$$

We now show that there exists a positive constant c such that

$$c\,\mathcal{S}_0(\phi,\,p;\,\phi,\,p) \le \mathcal{F}_0(\phi,\,p;\,\phi,\,p) \tag{3.21}$$

for every $\phi \in D$ and $p \in V$. We start by writing

$$\begin{pmatrix} \phi \\ p \end{pmatrix} = \begin{pmatrix} z \\ z \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \tag{3.22}$$

where $z,\, w_1 \in D$ and $w_2 \in V$ such that

$$\mathcal{S}_0(z,\, z;\, w_1,\, w_2) = 0, \tag{3.23}$$

and

$$\nabla^* A \nabla w_1 = -w_2. \tag{3.24}$$

This is accomplished by choosing $z,\, w_1 \in D$ to satisfy

$$\nabla^* A \nabla z + z = \nabla^* A \nabla \phi + p, \tag{3.25}$$

$$\nabla^* A \nabla w_1 + w_1 = \phi - p \tag{3.26}$$

(see Remark 3.2). To see this, note that adding (3.25) and (3.26) yields

$$\phi = z + w_1. \tag{3.27}$$

Since $D \subset V$, setting

$$w_2 = p - z \tag{3.28}$$

yields $w_2 \in V$ and (3.22). Substituting (3.27) into (3.26) and using (3.28) proves (3.24), which in turn yields (3.23).

Since $z \in D$, the bound (3.7) yields

$$\begin{aligned} \mathcal{F}_0(z,\, z;\, z,\, z) &= (\nabla^* A \nabla z + Xz,\, \nabla^* A \nabla z + Xz) = \|\nabla^* A \nabla z + Xz\|_{0,\,\Omega}^2 \\ &\geq \gamma\left(\|\nabla^* A \nabla z\|_{0,\,\Omega}^2 + \|A^{1/2} \nabla z\|_{0,\,\Omega}^2\right) = \gamma \mathcal{S}_0(z,\, z;\, z,\, z). \end{aligned} \tag{3.29}$$

Note that

$$\mathcal{F}_0(w_1,\, w_2;\, w_1,\, w_2) = \|A^{1/2} \nabla w_2 - A^{1/2} \nabla w_1\|_{0,\,\Omega,\,n}^2 + \|\nabla^* A \nabla w_1 + Xw_2\|_{0,\,\Omega}^2. \tag{3.30}$$

Using (3.24), the first term in (3.30) satisfies

$$\begin{aligned} &\|A^{1/2} \nabla w_2 - A^{1/2} \nabla w_1\|_{0,\,\Omega,\,n}^2 \\ &\quad = (A \nabla w_1,\, \nabla w_1)_n - 2(A \nabla w_1,\, \nabla w_2)_n + (A \nabla w_2,\, \nabla w_2)_n \\ &\quad = (A \nabla w_1,\, \nabla w_1)_n - 2(\nabla^* A \nabla w_1,\, w_2) + (A \nabla w_2,\, \nabla w_2)_n \\ &\quad = (A \nabla w_1,\, \nabla w_1)_n + 2(\nabla^* A \nabla w_1,\, \nabla^* A \nabla w_1) + (A \nabla w_2,\, \nabla w_2)_n \\ &\quad \geq (\nabla^* A \nabla w_1,\, \nabla^* A \nabla w_1) + (A \nabla w_2,\, \nabla w_2)_n \\ &\quad = \mathcal{S}_0(w_1,\, w_2;\, w_1,\, w_2). \end{aligned} \tag{3.31}$$

Note that the integration by parts is justified because $w_1 \in D$ and $w_2 \in V$. Using the bound (3.4) and assuming, as we may, that $\eta \geq 1$, the second term in (3.30) satisfies

$$\begin{aligned} \|\nabla^* A \nabla w_1 + Xw_2\|_{0,\,\Omega}^2 &\leq 2(\|\nabla^* A \nabla w_1\|_{0,\,\Omega}^2 + \|Xw_2\|_{0,\,\Omega}^2) \\ &\leq 2(\|\nabla^* A \nabla w_1\|_{0,\,\Omega}^2 + \eta^2 \|A^{1/2} \nabla w_2\|_{0,\,\Omega,\,n}^2) \\ &\leq 2\eta^2 \mathcal{S}_0(w_1,\, w_2;\, w_1,\, w_2). \end{aligned} \tag{3.32}$$

Using (3.22) and the Cauchy–Schwarz inequality we have

$$
\begin{aligned}
&\mathcal{F}_0(\phi,\, p;\, \phi,\, p) \\
&\quad = \|A^{1/2}\nabla w_2 - A^{1/2}\nabla w_1\|^2_{0,\,\Omega,\,n} + \|\nabla^* A\nabla w_1 + Xw_2\|^2_{0,\,\Omega} \\
&\qquad +2(\nabla^* A\nabla w_1 + Xw_2,\, \nabla^* A\nabla z + Xz) + \|\nabla^* A\nabla z + Xz\|^2_{0,\,\Omega} \\
&\quad \ge \|A^{1/2}\nabla w_2 - A^{1/2}\nabla w_1\|^2_{0,\,\Omega,\,n} + \|\nabla^* A\nabla w_1 + Xw_2\|^2_{0,\,\Omega} \\
&\qquad -2\|\nabla^* A\nabla w_1 + Xw_2\|_{0,\,\Omega}\|\nabla^* A\nabla z + Xz\|_{0,\,\Omega} + \|\nabla^* A\nabla z + Xz\|^2_{0,\,\Omega} \\
&\quad \ge \|A^{1/2}\nabla w_2 - A^{1/2}\nabla w_1\|^2_{0,\,\Omega,\,n} + \left(1 - \frac{1}{\epsilon}\right)\|\nabla^* A\nabla w_1 + Xw_2\|^2_{0,\,\Omega} \\
&\qquad +(1-\epsilon)\|\|\nabla^* A\nabla z + Xz\|^2_{0,\,\Omega}
\end{aligned}
$$

for any $\epsilon > 0$. Considering only the case $\epsilon < 1$ so that $1 - 1/\epsilon < 0$, and using (3.29), (3.31), and (3.32), we have

$$
\mathcal{F}_0(\phi,\, p;\, \phi, p) \ge \left(1 + \left(1 - \frac{1}{\epsilon}\right)2\eta^2\right)\mathcal{S}_0(w_1,\, w_2;\, w_1,\, w_2) + (1-\epsilon)\gamma\mathcal{S}_0(z,\, z;\, z,\, z).
$$

We now make

$$
1 + \left(1 - \frac{1}{\epsilon}\right)2\eta^2 = (1-\epsilon)\gamma
$$

by setting

$$
\epsilon = \frac{\sqrt{(1+2\eta^2-\gamma)^2 + 8\eta^2\gamma} - (1+2\eta^2-\gamma)}{2\gamma}.
$$

Note that $0 < \epsilon < 1$. This and (3.23) yield

$$
\mathcal{F}_0(\phi,\, p;\, \phi,\, p) \ge c(\mathcal{S}_0(w_1,\, w_2;\, w_1,\, w_2) + \mathcal{S}_0(z,\, z;\, z,\, z)) = c\mathcal{S}_0(\phi, p;\, \phi,\, p),
$$

with

$$
c = \frac{(1+2\eta^2+\gamma) - \sqrt{(1+2\eta^2+\gamma)^2 - 4\gamma}}{2} > 0,
$$

which is (3.21).

Finally, show that $\alpha\mathcal{S} \le \widehat{\mathcal{F}}$*:* We now have

$$
\mathcal{S}_0(\phi,\, p;\, \phi,\, p) = \|\nabla^* A\nabla\phi\|^2_{0,\,\Omega} + \|A^{1/2}\nabla p\|^2_{0,\,\Omega,\,n} \le \frac{1}{c}\mathcal{F}_0(\phi,\, p;\, \phi,\, p). \tag{3.33}
$$

By (3.33) and the definition of $\mathcal{F}_0$, we have

$$
\begin{aligned}
\|A^{1/2}\nabla\phi\|^2_{0,\,\Omega,\,n} &\le 2\left(\|A^{1/2}\nabla p\|^2_{0,\,\Omega,\,n} + \|A^{1/2}\nabla p - A^{1/2}\nabla\phi\|^2_{0,\,\Omega,\,n}\right) \\
&\le 2\left(\frac{1}{c} + 1\right)\mathcal{F}_0(\phi,\, p;\, \phi, p).
\end{aligned} \tag{3.34}
$$

The Poincaré-Friedrichs inequality (3.1) and the definition of $\mathcal{F}_0$ again, together with (3.33) and (3.34), now yield

$$
\begin{aligned}
\mathcal{S}(A\nabla\phi,\, p;\, A\nabla\phi,\, p) &= \|A^{1/2}\nabla\phi\|^2_{0,\,\Omega,\,n} + \|\nabla^* A\nabla\phi\|^2_{0,\,\Omega} + \|p\|^2_{0,\,\Omega} + \|A^{1/2}\nabla p\|^2_{0,\,\Omega,\,n} \\
&\le \left(\frac{3+2c+d}{c}\right)\widehat{\mathcal{F}}(A\nabla\phi,\, p;\, A\nabla\phi,\, p)
\end{aligned}
$$

which yields (3.20) with

$$\alpha_0 = \frac{c}{3+2c+d},$$

and the theorem is proved. □

4. Equivalent forms. Problem (2.1)–(2.3) can be written as a first order system in many different ways. In this section, we illustrate that such formulations often lead to least-squares functionals that are equivalent to the form $G(\mathbf{u},\, p;\, f)$ defined in (2.15). For example, if $X = \mathbf{a}^*\nabla + cI$, then one might choose the system

$$\begin{pmatrix} -I & A\nabla \\ \nabla^* + \mathbf{a}^* A^{-1} & cI \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix} \tag{4.1}$$

with $\mathbf{u} \in \mathbf{W}$ and $p \in V$. The functional associated with this system would then be

$$G_\ell(\mathbf{u},\, p;\, f) = \|A\nabla\, p - \mathbf{u}\|^2_{0,\,\Omega,\,n} + \|\nabla^*\mathbf{u} + \mathbf{a}^* A^{-1}\mathbf{u} + cp - f\|^2_{0,\,\Omega}\,. \tag{4.2}$$

System (4.1) can be related to the system implicit in $G(\mathbf{u},\, p;\, f)$ by the simple expression

$$\begin{pmatrix} I & 0 \\ -\mathbf{a}^* A^{-1} & I \end{pmatrix} \begin{pmatrix} -I & A\nabla \\ \nabla^* & \mathbf{a}^*\nabla + cI \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix}. \tag{4.3}$$

Thus, the ratio of $G(\mathbf{u},\, p;\, f)$ and $G_\ell(\mathbf{u},\, p;\, f)$ is bounded above and below, respectively, by the largest and smallest singular values of the transformation matrix. Specifically, if

$$\mathbf{a}^* A^{-2}\mathbf{a} \le C \tag{4.4}$$

for almost every $x \in L^2(\Omega)$, then

$$\alpha_\ell^- G(\mathbf{u},\, p;\, 0) \le G_\ell(\mathbf{u},\, p;\, 0) \le \alpha_\ell^+ G(\mathbf{u},\, p;\, 0), \tag{4.5}$$

for all $\mathbf{u} \in \mathbf{W}$ and $p \in V$, where

$$\alpha_\ell^\pm = \frac{(2+C) \pm \sqrt{C^2+4C}}{2}. \tag{4.6}$$

Note that $0 < \alpha_\ell^- < \alpha_\ell^+ < \infty$. Since equivalence is transitive, then Theorem 3.1 holds with $G(\mathbf{u},\, p;\, 0)$ replaced by $G_\ell(\mathbf{u},\, p;\, 0)$.

If, on the other hand, we have $X = \nabla^*\mathbf{b} + cI$, then one might consider the system

$$\begin{pmatrix} -I & A\nabla + \mathbf{b} \\ \nabla^* & cI \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{u}} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix}. \tag{4.7}$$

The functional associated with this system would then be

$$G_r(\widehat{\mathbf{u}},\, p;\, f) = \|A\nabla\, p + \mathbf{b}p - \widehat{\mathbf{u}}\|^2_{0,\,\Omega,\,n} + \|\nabla^*\widehat{\mathbf{u}} + cp - f\|^2_{0,\,\Omega}\,. \tag{4.8}$$

If $(\widehat{\mathbf{u}},\, p)$ satisfies system (4.7), then $\widehat{\mathbf{u}} = A\nabla p + \mathbf{b}p$. Boundary condition (2.3) now implies

$$\mathbf{n}\cdot\widehat{\mathbf{u}} = (\mathbf{n}\cdot\mathbf{b})p, \quad \text{on } \Gamma_N. \tag{4.9}$$

The choice of spaces on which to pose the minimization of G_r is not immediately clear. However, we can again relate system (4.7) to the system associated with G by writing (4.7) as

$$\begin{pmatrix} -I & A\nabla \\ \nabla^* & \nabla^*\mathbf{b} + cI \end{pmatrix} \begin{pmatrix} I & -\mathbf{b} \\ 0 & I \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{u}} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix}. \tag{4.10}$$

From (4.10) it is clear that

$$G_r(\widehat{\mathbf{u}},\, p;\, f) = G(\widehat{\mathbf{u}} - \mathbf{b}p,\, p;\, f). \tag{4.11}$$

The two functionals become analgous if we pose the minimization of $G_r(\widehat{\mathbf{u}},\, p;\, f)$ over the space

$$\mathbf{Z} \equiv \{(\widehat{\mathbf{u}},\, p) \in H(\mathrm{div}) \times V : \mathbf{n}\cdot(\widehat{\mathbf{u}} - \mathbf{b}p) = 0, \text{ on } \Gamma_N\}. \tag{4.12}$$

The ellipticity of G_r with respect to the $H(\mathrm{div}) \times H^1(\Omega)$ norm can now be established by setting $\mathbf{u} = \widehat{\mathbf{u}} - \mathbf{b}p$ and noting that

$$\frac{1}{K}\left(\|\mathbf{u}\|^2_{H(\mathrm{div})} + \|p\|^2_{1,\Omega}\right) \le \left(\|\widehat{\mathbf{u}}\|^2_{H(\mathrm{div})} + \|p\|^2_{1,\Omega}\right) \le K\left(\|\mathbf{u}\|^2_{H(\mathrm{div})} + \|p\|^2_{1,\Omega}\right), \tag{4.13}$$

where $K = 2 + 4\|\nabla^*\mathbf{b}\|^2_{0,\Omega} + 4\|\mathbf{b}\|^2_{0,\Omega,n}$. Combining (4.11) and (4.13), we see that Theorem 3.1 again holds with G replaced by G_r evaluated on $\mathbf{Z}$.

System (4.7) has the unfortunate property that the space $\mathbf{Z}$ is not a tensor product space. A vector $(\widehat{\mathbf{u}},\, p)$ is admissible only if $\widehat{\mathbf{u}}$ and p satisfy the proper relationship at the boundary. On the other hand, if boundary condition (2.3) were replaced by

$$\mathbf{n}\cdot A\nabla p + bp = 0, \qquad \text{on } \Gamma_N, \tag{4.14}$$

where now b is a scalar, and if a $\mathbf{b} \in H(\mathrm{div})$ could be determined so that

$$\mathbf{n}\cdot\mathbf{b} = b, \qquad \text{on } \Gamma_N, \tag{4.15}$$

then one might consider the system

$$\begin{pmatrix} -I & A\nabla + \mathbf{b} \\ \nabla^* & X - \nabla^*\mathbf{b} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{u}} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ f \end{pmatrix}, \tag{4.16}$$

and its associated functional

$$G_m(\widehat{\mathbf{u}},\, p;\, f) = \|A\nabla\, p + \mathbf{b}p - \widehat{\mathbf{u}}\|^2_{0,\Omega,n} + \|\nabla^*\widehat{\mathbf{u}} + Xp - \nabla^*(\mathbf{b}p) - f\|^2_{0,\Omega}\,. \tag{4.17}$$

The proper subspace is once again $\mathbf{W} \times V$ as in (2.7). A slight modification of the proof of Theorem 3.1 yields that G can be replaced by G_m as well.

5. Finite element approximation. We approximate the minimum of $G(\mathbf{u},\, p;\, f)$ in (2.15) using a Rayleigh-Ritz type finite element method. Let $\mathcal{T}_h$ be a partition of the domain Ω into finite elements, i.e., $\Omega = \cup_{K\in\mathcal{T}_h} K$ with $h = \max\{\mathrm{diam}(K) : K \in \mathcal{T}_h\}$. Assume that the triangulation $\mathcal{T}_h$ is regular (see [12]). Let V_h and $\mathbf{W}_h$ be finite-dimensional subspaces of V and $\mathbf{W}$ with the following properties:

$$\inf_{q_h\in V_h} \|q - q_h\|_{1,\Omega} \le Ch^k\|q\|_{k+1,\Omega}, \tag{5.1}$$

$$\inf_{\mathbf{v}_h \in \mathbf{W}_h} \|\mathbf{v} - \mathbf{v}_h\|_{H(\mathrm{div})} \leq C h^r \|\mathbf{v}\|_{r+1,\Omega}, \tag{5.2}$$

where $k, r > 0$ are integers, $q \in H^{k+1}(\Omega)$, and $\mathbf{v} \in \left(H^{r+1}(\Omega)\right)^n$.

A standard choice for such spaces is piecewise polynomials of degree k and r, respectively, i.e.,

$$V_h = \left\{ q_h \in C^0(\Omega) : q_h|_K \in P_k(K),\ \forall K \in \mathcal{T}_h,\ q_h = 0 \text{ on } \Gamma_D \right\} \tag{5.3}$$

(5.4)

$$\mathbf{W}_h = \{\mathbf{v}_h \in H(\mathrm{div}) : (\mathbf{v}_h)_i|_K \in P_r(K),\ i = 1, \ldots, n,\ \forall K \in \mathcal{T}_h,\ \mathbf{v}_h \cdot \mathbf{n} = 0 \text{ on } \Gamma_N\} \ .$$

Here, $P_s(K)$ is the space of polynomials of degree s on K. Another choice for $\mathbf{W}_h$ is a Raviart-Thomas space (see [27]), where continuity requirements are weaker. In this case, of course, definition (5.4) should be modified since the Raviart-Thomas spaces consist of incomplete polynomials of degree r on each element. Also, we allow the use of standard isoparametric elements.

The finite element approximation to minimizing $G(\mathbf{u}, p; f)$ in (2.15) on $\mathbf{W} \times V$ becomes: find $p_h \in V_h$ and $\mathbf{u}_h \in \mathbf{W}_h$ that satisfy

$$G(\mathbf{u}_h, p_h; f) = \min_{\mathbf{v}_h \in \mathbf{W}_h,\, q_h \in V_h} G(\mathbf{v}_h, q_h; f). \tag{5.5}$$

By Theorem 3.1 and the fact that $V_h \in V$ and $\mathbf{W}_h \in \mathbf{W}$, we conclude that (5.5) has a unique solution and is equivalent to the weak form: find $p_h \in V_h$ and $\mathbf{u}_h \in \mathbf{W}_h$ such that

$$\mathcal{F}(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = (f, \nabla^* \mathbf{v}_h + X q_h) \tag{5.6}$$

for every $q_h \in V_h$ and $\mathbf{v}_h \in \mathbf{W}_h$. Moreover, the error satisfies the orthogonality property

$$\mathcal{F}(\mathbf{u} - \mathbf{u}_h, p - p_h; \mathbf{v}_h, q_h) = 0 \tag{5.7}$$

for every $q_h \in V_h$ and $\mathbf{v}_h \in \mathbf{W}_h$.

THEOREM 5.1. *Let $s = \min(k, r)$ and assume that $p \in H^{s+1}(\Omega)$ and $\mathbf{u} \in \left(H^{s+1}(\Omega)\right)^n$. Then*

$$\|p - p_h\|_{1,\Omega} + \|\mathbf{u} - \mathbf{u}_h\|_{H(\mathrm{div})} \leq C\, h^s \left(\|p\|_{s+1,\Omega} + \|\mathbf{u}\|_{s+1,\Omega}\right),$$

where the constant C does not depend on h, p, or $\mathbf{u}$.

Proof. The proof is a simple consequence of the orthogonality property (5.7) and the approximation properties (5.1) and (5.2) of the finite element spaces V_h and $\mathbf{W}_h$. $\square$

This error estimate is optimal for $k = r$. In many applications, however, it may be useful to have higher order approximations to the fluxes or velocities $\mathbf{u}$, which is why we allow $r > k$. Moreover, numerical experience in [9], [26] and [25] indicates that, for certain values of r and k such that $r \neq k$, the error in $\mathbf{u}$ is $O(h^r)$ while the error in p is $O(h^k)$. It may be possible to develop a theory that bounds the errors in $\mathbf{u}$ and p separately.

6. Condition number. In this section, we bound the condition number of the linear system arising from (5.6). To this end, we additionally assume that the triangulation $\mathcal{T}_h$ satisfies the inverse assumption, i.e., there exists a positive constant θ such that, for all $K \in \mathcal{T}_h$,

$$h \leq \theta \operatorname{diam}(K). \tag{6.1}$$

Under assumption (6.1), many standard finite element spaces satisfy the so-called inverse inequality (see [12]), i.e.,

$$|q|_{1,\Omega} \leq C\, h^{-1} \|q\|_{0,\Omega}, \quad \forall\, q \in V_h, \tag{6.2}$$

$$|\operatorname{div} \mathbf{v}|_{0,\Omega} \leq C\, h^{-1} \|\mathbf{v}\|_{0,\Omega}, \quad \forall\, \mathbf{v} \in \mathbf{W}_h. \tag{6.3}$$

Let ϕ_1, ..., ϕ_N and $\boldsymbol{\psi}_1$, ..., $\boldsymbol{\psi}_M$ be bases for V_h and $\mathbf{W}_h$, respectively. Then, for any $q \in V_h$ and $\mathbf{v} \in \mathbf{W}_h$, we have

$$q = \sum_{i=1}^{N} \eta_i \phi_i \quad \text{and} \quad \mathbf{v} = \sum_{i=1}^{M} \xi_i \boldsymbol{\psi}_i.$$

Denote by $|\boldsymbol{\eta}|$ and $|\boldsymbol{\xi}|$ the ℓ_2-norms of the vectors $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, respectively. Under assumption (6.1), all well-known finite element spaces with the usual choice of bases have the following property: there exist positive constants α_i and β_i ($i = 1, 2$) such that

$$\alpha_1\, h^n\, |\boldsymbol{\eta}| \leq \|q\|_{0,\Omega} \leq \alpha_2\, h^n\, |\boldsymbol{\eta}| \tag{6.4}$$

and

$$\beta_1\, h^n\, |\boldsymbol{\xi}| \leq \|\mathbf{v}\|_{0,\Omega} \leq \beta_2\, h^n\, |\boldsymbol{\xi}|. \tag{6.5}$$

(Recall that n is the spatial dimension of Ω.)

THEOREM 6.1. *Assume that inequalities (6.2)–(6.5) hold. Then the condition number of the linear system resulting from (5.6) is $O(h^{-2})$.*

Proof. It follows from ellipticity and continuity of the bilinear form $\mathcal{F}$ and the inverse inequalities (6.2) and (6.3) that

$$\alpha\, \left(\|\mathbf{v}\|_{0,\Omega}^2 + \|q\|_{0,\Omega}^2\right) \leq \mathcal{F}\,(\mathbf{v},\, q;\, \mathbf{v},\, q) \leq C\, h^{-2} \left(\|\mathbf{v}\|_{0,\Omega}^2 + \|q\|_{0,\Omega}^2\right).$$

Combining inequalities (6.4) and (6.5), we obtain

$$C\, h^n \left(|\boldsymbol{\xi}|^2 + |\boldsymbol{\eta}|^2\right) \leq \mathcal{F}\,(\mathbf{v},\, q;\, \mathbf{v},\, q) \leq C\, h^{n-2} \left(|\boldsymbol{\xi}|^2 + |\boldsymbol{\eta}|^2\right).$$

This completes the proof. ☐

REFERENCES

[1] A. K. AZIZ, R. B. KELLOGG, AND A.B. STEPHENS, *Least-squares methods for elliptic systems*, Math. Comp., 44(169) (1985), pp. 53–70.

[2] I. BABUŠKA, *The finite element method with Lagrange multipliers*, Numer. Math., 20 (1973), pp. 179–192.

[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] P. B. Bochev and M. D. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., to appear.

[5] F. Brezzi, *On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, RAIRO Anal. Numér., 2 (1974), pp. 129–151.

[6] F. Brezzi, J. Douglas, M. Fortin, and D. Matini, *Efficient rectangular mixed finite elements in two and three space variables*, RARIO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

[7] F. Brezzi, J. Douglas, and D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[8] Z. Cai, T. Manteuffel, and S. McCormick, *First-order system least-squares for second-order partial differential equations: Part II*, SIAM J. Numer. Anal., submitted.

[9] G. F. Carey and Y. Shen, *Convergence studies of least-squares finite elements for first order systems*, Comm. Appl. Numer. Meth., 5 (1989), pp. 427–434.

[10] T. F. Chen, *On the least-squares approximations to compressible flow problems*, Numer. Meth. PDE's, 2 (1986), pp. 207–228.

[11] T. F. Chen and G. J. Fix, *Least-squares finite element simulation of transonic flows*, Appl. Numer. Math., 2 (1986), pp. 399–408.

[12] P. G. Ciarlet, *The Finite Element Methods for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[13] E. D. Eason, *A review of least–squares methods for solving partial differential equations*, Int. J. Numer. Math. Engrg., 10 (1976), pp. 1021–1046.

[14] L. Franca and R. Stenberg, *Error analysis of some Galerkin-least-squares method for the elasticity equations*, Report #1054, 1989, INRIA, pp. 1–21.

[15] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, New York, 1986.

[16] C. I. Goldstein, T. A. Manteuffel, and S. V. Parter, *Preconditioning and boundary conditions without H_2 estimates: L_2 condition numbers and the distribution of the singular valves*, SIAM J. of Numer. Anal., 13(1) (1992), pp. 259–288.

[17] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[18] 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.

[19] 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.

[20] D. C. Jespersen, *A least-square decomposition method for solving elliptic systems*, Math. Comp., 31(140) (1977), pp. 873–880.

[21] B. N. Jiang and C. Chang, *Least-squares finite elements for the Stokes problem*, Comput. Meth. Appl. Mech. Engrg., 81 (1990), pp. 13–37.

[22] B. N. Jiang and L. A. Povinelli, *Optimal least-squares finite element method for elliptic problems*, Comput. Meth. Appl. Mech. Engrg., 102 (1993), pp. 199–212.

[23] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flows*, Gordon and Breach, London, 1969.

[24] 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.

[25] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least squares mixed finite elements for second order elliptic problems*, SIAM J. Numer. Anal., to appear.

[26] A. I. Pehlivanov, G. F. Carey, R. D. Lazarov, and Y. Shen, *Convergence of least squares finite elements for first order ODE systems*, Computing, 51 (1993), pp. 111–123.

[27] P. A. Raviart and I. M. Thomas, *A mixed finite element method for second order elliptic problems*, Lect. Notes Math. 606, Springer-Verlag, Berlin and New York, 1977, pp. 292–315.

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