INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 1, Number 1, Pages 49–64

LEAST-SQUARES FINITE ELEMENT METHODS FOR FIRST-ORDER ELLIPTIC SYSTEMS

PAVEL BOCHEV

Abstract. Least-squares principles use artificial "energy" functionals to provide a Rayleigh-Ritz-like setting for the finite element method. These functionals are defined in terms of PDE's residuals and are not unique. We show that viable methods result from reconciliation of a mathematical setting dictated by the *norm-equivalence* of least-squares functionals with *practicality constraints* dictated by the algorithmic design. We identify four universal patterns that arise in this process and develop this paradigm for first-order ADN elliptic systems. Special attention is paid to the effects that each discretization pattern has on the computational and analytic properties of finite element methods, including error estimates, conditioning of the algebraic systems and the existence of efficient preconditioners.

Key Words. finite elements, least-squares, first-order elliptic systems.

1. Introduction

After a somewhat disappointing start in the early seventies[1], the use of least-squares finite elements has been steadily increasing over the last decade. A key factor for the renewed interest in such methods was the idea of their application to equivalent first-order systems rather than to the original PDE problem; see [17], [22], [18], [11] and [13]. This paid off in turning least-squares methods into a viable alternative to Galerkin finite elements, especially in fluid flow computations; see [6]–[12], [18]–[21], [23], and [27]–[29]. From a mathematical viewpoint another notion, namely the concept of *norm-equivalent* least-squares "energy" functionals emerged as a universal prerequisite for recovering fully the Rayleigh-Ritz setting. However, it was soon realized that norm-equivalence is often in conflict with practicality, even for first-order systems (see [6], [11] and [12]); and because practicality is usually the rigid constraint in the algorithmic development, norm equivalence was often neglected.

The main goal of this paper is to establish the reconciliation between practicality, as driven by algorithmic development, and norm-equivalence, as motivated by mathematical analyses, as the defining paradigm of least-squares finite element methods. The key components of this paradigm are a *continuous least-squares principle* (CLSP) which describes a mathematically well-posed, but perhaps impractical,

Received by the editors January 10, 2004.
2000 *Mathematics Subject Classification.* 65F10 (Primary), 65F30 (Secondary).
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

[1]Early examples of least-squares methods suffered from disadvantages that seriously limited their appeal. In many cases discretization required impractical C^1 or better finite element spaces and led to algebraic problems with higher than usual condition numbers; see e.g., [3]–[4], and without efficient preconditioners.

variational setting, and an associated *discrete least-squares principle* (DLSP) which describes an algorithmically feasible setting. The relation between a CLSP and a DLSP follows four universal patterns which lead to four classes of least-squares finite element methods with distinctly different properties.

We develop this paradigm for the important class of first-order systems that are elliptic in the sense of Agmon-Douglis-Nirenberg [1]. In particular, we show that degradation of fundamental properties of least-squares methods such as condition numbers, asymptotic convergence rates, and existence of spectrally equivalent preconditioners occurs when DLSP deviates from the conforming setting induced by a given CLSP.

In what follows Ω will denote a simply connected bounded region in $\mathbb{R}^n$, $n = 2, 3$ with a sufficiently smooth boundary Γ. Throughout the paper we employ the usual notations $H^d(\Omega)$, $\|\cdot\|_d$; $d \geq 0$ for the Sobolev spaces of all functions having square integrable derivatives up to order d on Ω, and the standard Sobolev norm, respectively. As usual, $H_0^d(\Omega)$ will denote the closure of $C^\infty(\Omega)$ with respect to the norm $\|\cdot\|_d$ and $H^{-d}(\Omega)$ will denote the dual of $H_0^d(\Omega)$. The symbol S_d^h will stand for a space of continuous, piecewise polynomial functions defined with respect to a regular triangulation $\mathcal{T}_h$ of the domain Ω. It is assumed that for every $u \in H^{d+1}(\Omega)$ there exists $u^h \in S_d^h$ with

$$\|u - u^h\|_0 + h\|u - u^h\|_1 \leq Ch^{d+1}\|u\|_{d+1}. \tag{1}$$

For regular triangulations the Euclidean norm of the coefficient vector of u^h, denoted by $|\xi|$, and the L^2 norm of u^h are related by the inequality

$$C^{-1}h^M|\xi| \leq \|u^h\|_0 \leq Ch^M|\xi|\,, \tag{2}$$

where M denotes the dimension of S_d^h. We will also need the inverse inequality

$$\|u^h\|_1 \leq Ch^{-1}\|u^h\|_0 \tag{3}$$

which holds for most standard finite element spaces on regular triangulations; see [16].

2. Continuous and discrete least-squares principles

We consider boundary value problems of the form

$$\mathcal{L}(\mathbf{x}, D)\ \mathbf{u} = \mathbf{f} \ \text{ in } \Omega \quad \text{and} \quad \mathcal{R}(\mathbf{x}, D)\ \mathbf{u} = \mathbf{g} \ \text{ on } \Gamma. \tag{4}$$

Here $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ is a vector of dependent variables, $\mathcal{L}(\mathbf{x}, D) = \mathcal{L}_{ij}(\mathbf{x}, D)$, $i, j = 1, \ldots, N$ and $\mathcal{R}(\mathbf{x}, D) = \mathcal{R}_{lj}(\mathbf{x}, D)$, $l = 1, \ldots, L$, $j = 1, \ldots, N$. For simplicity, in what follows we will write $\mathcal{L}\mathbf{u}$ and $\mathcal{R}\mathbf{u}$. Concerning (4), we make the following assumption:

A.: There exist Hilbert spaces $X = X(\Omega)$, $Y = Y(\Omega)$, and $Z = Z(\Gamma)$ such that

$$C_2\|\mathbf{u}\|_X \leq \|\mathcal{L}\mathbf{u}\|_Y + \|\mathcal{R}\mathbf{u}\|_Z \leq C_1\|\mathbf{u}\|_X\,. \tag{5}$$

This relation is fundamental to least-squares methods because it defines the proper "balance" between solution energy as measured by $\|\mathbf{u}\|_X$ and data energy, as measured by $\|\mathcal{L}\mathbf{u}\|_Y + \|\mathcal{R}\mathbf{u}\|_Z$. We note that the setting determined by (5) is not, in general, unique[2].

[2]For example, if $(\mathcal{L}, \mathcal{R})$ has a complete set of homeomorphisms (5) holds on a Hilbert scale; see [24] and [25].

2.1. Continuous least-squares principles. A *continuous least-squares principle*, or CLSP, for (4) is a pair $\{X, J(\cdot)\}$ where the functional

$$J(\mathbf{u};\mathbf{f},\mathbf{g}) = \frac{1}{2}\Big(\|\mathcal{L}\mathbf{u}-\mathbf{f}\|_Y^2 + \|\mathcal{R}\mathbf{u}-\mathbf{g}\|_Z^2\Big) \tag{6}$$

is minimized over the space X, i.e., we solve the optimization problem

$$\textit{seek}\quad \mathbf{u}\in X \quad \textit{such that}\quad J(\mathbf{u};\mathbf{f},\mathbf{g}) \le J(\mathbf{v},\mathbf{f},\mathbf{g}) \quad \forall\,\mathbf{v}\in X\,. \tag{7}$$

For simplicity we will write $J(\mathbf{u})$ instead of $J(\mathbf{u};0,0)$.

Theorem 1. *Assume that* **A.** *holds. Then,*

(1) *the functional* (6) *is norm-equivalent in the sense that*

$$\frac{1}{4}C_2^2\|\mathbf{u}\|_X^2 \le J(\mathbf{u}) \le \frac{1}{2}C_1^2\|\mathbf{u}\|_X^2 \qquad \forall\,\mathbf{u}\in X\,; \tag{8}$$

(2) *problem* (7) *has a unique minimizer* $\mathbf{u}$ *such that*

$$\|\mathbf{u}\|_X \le C\left(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z\right)\,. \tag{9}$$

Moreover, $\mathbf{u}$ *is the unique minimizer of* (6) *if and only if* $\mathbf{u}$ *is the unique solution of* (4).

Proof. To prove 1 it suffices to note that $J(\mathbf{u}) = \frac{1}{2}\Big(\|\mathcal{L}\mathbf{u}\|_X^2 + \|\mathcal{R}\mathbf{u}\|_Z^2\Big)$ so that the norm-equivalence (8) follows from (5). To prove 2, note that minimizers of (6) satisfy the Euler-Lagrange equation

$$\textit{seek}\quad \mathbf{u}\in X \quad \textit{such that}\quad Q(\mathbf{u};\mathbf{v}) = F(\mathbf{v}) \quad \forall\,\mathbf{v}\in X\,. \tag{10}$$

The form $Q(\cdot;\cdot)$ and the functional $F(\cdot)$ in (10) are given by

$$Q(\mathbf{u};\mathbf{v}) = (\mathcal{L}\mathbf{u},\mathcal{L}\mathbf{v})_Y + (\mathcal{R}\mathbf{u},\mathcal{R}\mathbf{v})_Z \tag{11}$$

and

$$F(\mathbf{v}) = (\mathbf{f},\mathcal{L}\mathbf{v})_Y + (\mathbf{g},\mathcal{R}\mathbf{v})_Z\,, \tag{12}$$

respectively. From the lower bound in (5), we obtain

$$Q(\mathbf{u};\mathbf{u}) = 2J(\mathbf{u}) = \|\mathcal{L}\mathbf{u}\|_X^2 + \|\mathcal{R}\mathbf{u}\|_Z^2 \ge \frac{1}{2}C_2^2\|\mathbf{u}\|_X^2\,,$$

while Cauchy's inequality and the upper bound in (5) yield continuity of $Q(\cdot;\cdot)$. Likewise, it is easy to see that $\|\mathcal{F}\| \le C_1(\|\mathbf{f}\|_Y + \|\mathbf{g}\|_Z)$., i.e., $F(\mathbf{v})$ is a bounded linear functional on X. As a result, the existence and uniqueness of a minimizer $\mathbf{u}$ that solves (10) follows from the Lax-Milgram Theorem. Then, (9) easily follows from the coercivity of $Q(\cdot;\cdot)$ and the continuity of $F(\cdot)$. □

Theorem 1 affirms that $\{X, J(\cdot)\}$ is an external Rayleigh-Ritz principle for (4). The "energy" inner product associated with $\{X, J(\cdot)\}$ is $Q(\cdot;\cdot)$, while $|||\mathbf{u}|||^2 = Q(\mathbf{u};\mathbf{u}) = 2J(\mathbf{u})$ is the "energy" norm. The least-squares problem (7) is equivalent to (4) in the sense that their solutions belonging to the space X coincide. However, the variational problem (10) is not a standard, e.g., Galerkin, *weak form* of (4).

3. Discrete least-squares principles

A conforming *discrete least-squares principle* (DLSP) is a pair $\{X^h, \mathcal{J}\}$ where $X^h = \text{span}\{\phi_i^h\}_{i=1}^M$ is a finite element subspace of X and $J(\cdot)$ is given by (6). A conforming DLSP leads to finite element methods for (4) which recover all attractive features of a Rayleigh-Ritz setting: non-restrictive choice of finite element spaces, symmetric and positive definite algebraic systems, and quasioptimal error estimates. Furthermore, if $A^h = Q(\phi_i^h; \phi_j^h)$ and $K^h = (\phi_i^h, \phi_j^h)_X$, the equivalence between $Q(\cdot;\cdot)$ and the standard inner product $(\cdot,\cdot)_X$ implies that

$$C^{-1}\boldsymbol{\xi}^T K^h \boldsymbol{\xi} \le \boldsymbol{\xi}^T A^h \boldsymbol{\xi} \le C\boldsymbol{\xi}^T K^h \boldsymbol{\xi}, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^M, \tag{13}$$

that is, A^h and K^h are spectrally equivalent.

However, depending on $\mathcal{L}$, $\mathcal{R}$ and the choice of X, Y and Z dictated by (5), the pair $\{X^h, J(\cdot)\}$ may be inconvenient for actual implementation. Then, practicality considerations (discussed in §4.1) may force us to abandon the conforming setting described above and consider instead another pair, denoted by $\{X^h, J_h(\cdot)\}$. In this pair X^h is not necessarily a subspace of X and $J_h(\cdot)$ is not necessarily the same as $J(\cdot)$. The pair $\{X^h, J_h(\cdot)\}$ gives rise to another DLSP:

$$\min_{\mathbf{v}^h \in X^h} J_h(\mathbf{v}^h; \mathbf{f}, \mathbf{g}). \tag{14}$$

Let us show that (14) can actually produce meaningful results under the following two very general hypotheses:

D.1: $J_h(\cdot)$ is *consistent* in the sense that $J_h(\mathbf{u}; \mathbf{f}, \mathbf{g}) = 0$ for all smooth data $\mathbf{f}$ and $\mathbf{g}$ and all smooth solutions $\mathbf{u}$ of (4);

D.2: $J_h(\cdot)$ is *positive*: $J_h(\mathbf{v}^h) > 0 \;\; \forall\, 0 \neq \mathbf{v}^h \in X^h$.

From **D.2** we can infer the existence of an inner product

$$((\cdot,\cdot))_h \;:\; X^h \times X^h \mapsto \mathbb{R}, \tag{15}$$

called *discrete energy inner product*. Therefore,

$$((\mathbf{v}^h, \mathbf{v}^h))_h = |||\mathbf{v}^h|||_h^2 \equiv J_h(\mathbf{v}^h) \tag{16}$$

is a norm on X^h, which we refer to as the *discrete energy norm*.

Theorem 2. *Assume that* **D.1** *and* **D.2** *hold for the pair* $\{X^h, J_h(\cdot)\}$ *and let* $\mathbf{u}$ *denote a smooth solution of* (4). *Then, problem* (14) *has a unique minimizer* $\mathbf{u}^h \in X^h$. *This minimizer is the orthogonal projection of* $\mathbf{u}$ *with respect to* (15).

Proof. The minimizer of (14) solves the problem:

$$\textit{seek } \mathbf{u}^h \textit{ in } X^h \textit{ such that} \quad B^h(\mathbf{u}^h; \mathbf{v}^h) = F^h(\mathbf{v}^h) \quad \forall \mathbf{v}^h \in X^h\,, \tag{17}$$

where $B^h(\cdot;\cdot) = ((\cdot,\cdot))_h$ and $F^h(\cdot) = ((\mathbf{u},\cdot))_h$. Let $\mathbf{u}^h = \sum_{i=1}^M \xi_i \phi_i^h$. Problem (17) is a linear system with matrix $A_{ij}^h = ((\phi_j^h, \phi_i^h))_h$ and right hand side $F_i = ((\mathbf{u}, \phi_i^h))_h$. From **D.2** it follows that A^h is symmetric, positive definite and so (17) has a unique solution. To prove the second part note that (17) can be recast as

$$((\mathbf{u}^h - \mathbf{u}, \mathbf{v}^h))_h = 0 \quad \text{for all } \mathbf{v}^h \in X^h.$$

which means that $\mathbf{u}^h$ is orthogonal projection of $\mathbf{u}$ relative to $((\cdot,\cdot))_h$. □

Corollary 1. *The least-squares solution* $\mathbf{u}^h$ *minimizes the discrete energy norm error, that is*

$$|||\mathbf{u} - \mathbf{u}^h|||_h = \inf_{\mathbf{v}^h \in X^h} |||\mathbf{u} - \mathbf{v}^h|||_h. \tag{18}$$

3.0.1. Classes of discrete least-squares principles. Substitution of the conforming DLSP by another principle constitutes a variational crime. Theorem 2 indicates that the penalty for this crime is much less severe than in other methods. It also explains the remarkable robustness of least-squares methods: almost any sensible pair $\{X^h, J_h(\cdot)\}$ will satisfy **D.1**-**D.2**. However, this by no means implies that the CLSP setting established in §2.1 is superficial and may be ignored in the algorithmic development.

Indeed, consider the following situation. Assume that both $|||\cdot|||$ and $\|\cdot\|_X$ are meaningful for $\mathbf{u}^h \in X^h$ so that their restrictions to X^h are well-defined norms. Since $|||\cdot|||_h$ is another norm on this finite-dimensional space, it must be equivalent to the restrictions of $|||\cdot|||$ and $\|\cdot\|_X$. As a result, for every fixed $h > 0$, there are positive numbers $\gamma_1(h)$, $\gamma_2(h)$, $\delta_1(h)$ and $\delta_2(h)$, such that

$$\gamma_1(h)\|\mathbf{u}^h\|_X \le |||\mathbf{u}^h|||_h \le \gamma_2(h)\|\mathbf{u}^h\|_X \quad \forall \mathbf{u}^h \in X^h; \tag{19}$$

$$\delta_1(h)\boldsymbol{\xi}^T K^h \boldsymbol{\xi} \le \boldsymbol{\xi}^T A^h \boldsymbol{\xi} \le \delta_2(h)\boldsymbol{\xi}^T K^h \boldsymbol{\xi} \quad \forall \boldsymbol{\xi} \in \mathbb{R}^M. \tag{20}$$

Thus, a discrete version of (5) and (13) holds for any fixed h. However, the asymptotic behavior of $\gamma_i(h)$ and $\delta_i(h)$ depends entirely on the relation between the two energy norms and neither **D.1** nor **D.2** can control the growth (or decay) of $\gamma_i(h)$ and $\delta_i(h)$.

Conforming DLSP, introduced at the beginning of §3, are restrictions of a given CLSP, i.e, the pair $\{X^h, J_h(\cdot)\}$ represents a subspace X^h of X and $J_h(\cdot) = J(\cdot)$. Thus, (19) is a restriction of (5) to X^h which means that $\gamma_i(h)$ and $\delta_i(h)$ are independent of h; in fact they coincide with the constants C_1 and C_2 from (5).

A deviation from this setting may strengthen the dependence on h and lead to deterioration of algebraic systems and asymptotic convergence of least-squares solutions, as $h \mapsto 0$. There are three possible scenarios according to which this may happen.

Norm-equivalent DLSP are pairs $\{X^h, J_h(\cdot)\}$ for which $X^h \subset X$ and (19) holds with γ_i and δ_i *independent* of h. In general, for such methods $J_h(\cdot) \neq J(\cdot)$, and (19) is not a restriction of (5). Nevertheless, *norm-equivalent* methods do recover all advantages of a Rayleigh-Ritz setting.

Quasi-norm-equivalent DLSP are pairs $\{X^h, J_h(\cdot)\}$ for which $X^h \subset X$ but (19) holds with γ_i and δ_i *dependent* on h. These methods yield optimal convergence rates, however, dependence on h in the equivalence bound leads to higher condition numbers and/or lack of spectral equivalence with the natural inner product on $X \times X$.

And lastly, *non-equivalent DLSP* are pairs $\{X^h, J_h(\cdot)\}$ for which X^h is not necessarily a subspace of X and $J_h(\cdot) \neq J(\cdot)$. As a result, (19) holds with different[3] spaces for the lower and upper bounds. Thus, nothing much can be said about such methods beyond Theorem 2.

4. Application to ADN elliptic systems

To discuss practicality constraints that may force one to abandon the conforming setting we focus on a specific class of PDE problems. For these problems we use Agmon, Douglis and Nirenberg (ADN) theory; see [1], to identify settings which

[3]If X^h satisfies an inverse inequality these bounds may be converted to bounds in terms of the same function space. This necessarily will introduce dependence on h in the lower and/or upper equivalence constants.

verify hypothesis **A.** and lead to well-posed CLS principles. For earlier work based on ADN theory[4] we refer to [2], [11], [12], and [15].

Definition 1. *The system (4) is ADN-elliptic if there exist integer weights* $\{s_i\}$ *and* $\{t_j\}$*, for the equations and the unknowns, respectively, such that* $\deg\mathcal{L}_{ij}(\mathbf{x},\boldsymbol{\xi}) \leq s_i + t_j$*;* $\mathcal{L}_{ij} \equiv 0$ *whenever* $s_i + t_j < 0$*;* $\det\mathcal{L}^P_{ij}(\mathbf{x},\boldsymbol{\xi}) \neq 0$ *for all real* $\boldsymbol{\xi} \neq 0$*; where the* principal part $\mathcal{L}^P$ *of* $\mathcal{L}$ *is defined as all terms* $\mathcal{L}_{ij}$ *for which* $deg\mathcal{L}_{ij}(\mathbf{x},\boldsymbol{\xi}) = s_i + t_j$*.* $\mathcal{L}$ *is called* uniformly elliptic *if there exists a positive constant* C*, such that*

$$C^{-1}|\boldsymbol{\xi}|^{2m} \leq |det\,\mathcal{L}^P(\mathbf{x},\boldsymbol{\xi})| \leq C|\boldsymbol{\xi}|^{2m}. \tag{21}$$

For nondegenerate systems one can always find s_i and t_j so that the principal part $\mathcal{L}^P$ does not vanish identically; see [31]. The orders of $\mathcal{R}_{lj}$ will also depend on two sets of integer weights: the set $\{t_j\}$ already defined for $\mathcal{L}$, and a new set $\{r_l\}$ where each r_l is attached to the lth condition in $\mathcal{R}$. As before, it will be required that $\deg\mathcal{R}_{lj}(\mathbf{x},\boldsymbol{\xi}) \leq r_l + t_j$, with the understanding that $\mathcal{R}_{lj} \equiv 0$ when $r_l + t_j < 0$. The principal part $\mathcal{R}^P$ of the boundary operator will be defined as all terms $\mathcal{R}_{lj}$ such that $\deg\mathcal{R}_{lj}(\mathbf{x},\boldsymbol{\xi}) = r_l + t_j$. The three sets of indices can always be normalized in such a way that $s_i \leq 0$, $r_l \leq 0$ and $t_j \geq 0$. However, even with that normalization the sets of indices may not be unique. This means that several concurrent principal parts may exist. An important subset of ADN elliptic systems is introduced below.

Definition 2. *A system is* elliptic in the sense of Petrovski *if it is elliptic in the sense of ADN and* $s_1 = \ldots = s_N = 0$*. If in addition* $t_1 = \ldots = t_N$*, the system is called* homogeneous elliptic*.*

The boundary value problem (4) will be well-posed only if $\mathcal{R}$ "complements" $\mathcal{L}$ in a proper way. A necessary and sufficient condition for this is given by the *complementing condition* of [1]. We will call (4) *ADN elliptic* when $\mathcal{L}$ is uniformly elliptic and $\mathcal{R}$ satisfies the complementing condition. Let

$$\mathbf{X}_q = \prod_{j=1}^{N} H^{q+t_j}(\Omega); \quad \mathbf{Y}_q = \prod_{i=1}^{N} H^{q-s_i}(\Omega); \quad \mathbf{Z}_q = \prod_{l=1}^{m} H^{q-r_l-1/2}(\Gamma). \tag{22}$$

We now proceed to show that ADN elliptic systems satisfy hypothesis **A.** of §2.1. The first part of the hypothesis follows from a general result due to Agmon, Douglis and Nirenberg [1].

Theorem 3. *Let* $t' = \max t_j$*,* $q \geq d = \max(0, \max r_l + 1)$ *and assume that* Ω *is a bounded domain of class* $C^{q+t'}$*. Furthermore, assume that the coefficients of* $\mathcal{L}$ *are of class* $C^{q-s_i}(\bar{\Omega})$ *and that the coefficients of* $\mathcal{R}$ *are of class* $C^{q-r_l}(\Gamma)$*. If (4) is elliptic and* $\mathbf{f} \in \mathbf{Y}_q$*,* $\mathbf{g} \in \mathbf{Z}_q$ *then*

1. *Every solution* $\mathbf{u} \in \mathbf{X}_d$ *is in fact in* $\mathbf{X}_q$*.*
2. *There exists* $C > 0$*, independent of* $\mathbf{u}$*,* $\mathbf{f}$ *and* $\mathbf{g}$*, such that, for every solution* $\mathbf{u} \in \mathbf{X}_q$

$$\sum_{j=1}^{N} \|\mathbf{u}_j\|_{q+t_j} \leq C\left(\sum_{i=1}^{N} \|\mathbf{f}_i\|_{q-s_i} + \sum_{l=1}^{m} \|\mathbf{g}_l\|_{q-r_l-1/2} + \sum_{j=1}^{N} \|\mathbf{u}_j\|_{0,\Omega}\right). \tag{23}$$

If (4) has a unique solution the L^2*-norm in (23) can be omitted.*

[4]For elliptic problems in the plane a simplified theory exists; see [30], which also has been used in the development of least-squares methods. For examples the reader can consult Wedland's book [30] and the papers [14]–[15].

Since ADN elliptic operators are of Fredholm type; see [24], [25], [30], their range is closed and both the kernel and the co-range are finite dimensional. Therefore, $(\mathcal{L}, \mathcal{R})$ can be augmented by a finite number of constraints so that (4) always has a unique solution. As a result, the L^2 term can be omitted for the modified problem. Lastly, it will be assumed that (23) is valid[5] for $q < 0$. Then (23) can be restated as follows: for all smooth functions $\mathbf{u}$ in Ω and all integers q

$$\|\mathbf{u}\|_{\mathbf{X}_q} \leq C \left(\|\mathcal{L}\mathbf{u}\|_{\mathbf{Y}_q} + \|\mathcal{R}\mathbf{u}\|_{\mathbf{Z}_q} \right). \tag{24}$$

4.1. First-order ADN-elliptic systems. The class of all continuous least-squares principles $\{X, J(\cdot)\}$ for an ADN system follows from Theorem 3 and (24). We identify X with the space $\mathbf{X}_q$, while

$$J(\mathbf{u}; \mathbf{f}, \mathbf{g}) = \frac{1}{2} \Big(\|\mathcal{L}\mathbf{u} - \mathbf{f}\|^2_{\mathbf{Y}_q} + \|\mathcal{R}\mathbf{u} - \mathbf{g}\|^2_{\mathbf{Z}_q} \Big) \tag{25}$$

is the "energy" functional. Thus, $\{\mathbf{X}_q, J(\cdot)\}$ corresponds to the optimization problem

$$\min_{\mathbf{u} \in \mathbf{X}_q} J(\mathbf{u}; \mathbf{f}, \mathbf{g}). \tag{26}$$

The conforming DLSP $\{\mathbf{X}^h, J(\cdot)\}$, where $\mathbf{X}^h \subset \mathbf{X}_q$, provides a Rayleigh-Ritz-like setting for the finite element method. However, this DLSP may be ill-suited for implementation. To deem a DLSP practical we require that at least

- the discrete systems can be obtained with no more difficulty than for a Galerkin method;
- their condition numbers should be comparable to those in the Galerkin method;
- discretization should be accomplished using standard, easy to use finite element spaces.

The first condition will be violated if (25) involves, fractional or negative order Sobolev space norms because such norms are not computable. The second and third conditions will be violated if for some s_i and t_j we have that $s_i + t_j \geq 2$. In this case the term $\|\mathcal{L}_{ij} u_j - f_i\|_{0-s_i}$ will effectively involve second, or higher order derivatives.

For simplicity we consider the case when $\mathbf{X}_q$ is constrained by the boundary condition in (4) and (25) does not involve trace norms[6]. Next, (4) will be transformed to a first-order problem. Here we rely on the important fact that any ADN-elliptic system of order higher than one can be transformed into an equivalent first-order system which is also ADN elliptic[7]. Even though we will see that this by itself is not enough to ensure practical conforming DLSP, first-order systems remain the most convenient setting for least-squares methods. This transformation can be effected through the following process; see [1]. First, all variables are divided into two sets according to their indices: a set $\{u_{k'}\}$ containing all variables for which $t_j > 1$ and a set $\{u_{k''}\}$ of all variables for which $t_j \leq 1$. Then, the new variables are introduced as $u_{k',j} = \partial_j u_{k'}$ and these equations are appended to $\mathcal{L}$. Next, all terms in $\mathcal{L}$ where u'_k is not differentiated remain unchanged. A term in which u'_k is differentiated is substituted according to the rule $D^\alpha(\partial_j u_{k'}) \mapsto D^\alpha(u_{k',j})$. Although rewriting

[5]This amounts to existence of complete sets of homeomorphisms for (4), and is known to hold for self-adjoint ADN and Petrovski systems; see [24]–[25].

[6]For examples of such least-squares methods we refer to [2], [26], and [30].

[7]This is not true if the usual definition of ellipticity, involving only differentiated terms, is used.

of $\mathcal{L}$ is not unique it can be shown; see [1], that the new operator is elliptic and that $\max t_j \leq 2$, $\min s_i \geq -1$. $\mathcal{R}$ is transformed to a boundary operator for the first-order system in a similar manner. While this transformation is not unique too, the relevant fact is that the new operator will satisfy the complementing condition provided it was satisfied by the original $\mathcal{R}$. As a result, we are guaranteed that the first-order system remains ADN elliptic.

4.2. Least-squares functionals for first-order systems. Consider first the case when (4) is homogeneous elliptic. Then $s_i = 0$ for all $i = 1, \ldots, N$ and therefore $t_j = 1$ for all $j = 1, \ldots, N$. Assuming that $\mathbf{X}_q$ is constrained by $\mathcal{R}$ the bound (24) specializes to

$$\|\mathbf{u}\|_{\mathbf{X}_q} = \sum_{j=1}^{N} \|u_j\|_{q+1} \leq C \sum_{i=1}^{N} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j \|_q. \tag{27}$$

If (4) is not homogeneous elliptic, then there will be at least one equation index $s_i = -1$. Since all $\mathcal{L}_{ij}$ are at most of order one, there will be at least one index $t_j = 2$. Without loss of generality we can assume that for some $1 \leq k \leq N$ and $1 \leq l \leq N$

$$s_1 = \ldots = s_k = 0; \quad s_{k+1} = \ldots = s_N = -1; \tag{28}$$

$$t_1 = \ldots = t_l = 1; \quad t_{l+1} = \ldots = t_N = 2, \tag{29}$$

respectively. As a result, now (24) specializes to

$$\sum_{j=1}^{l} \|u_j\|_{q+1} + \sum_{j=l+1}^{N} \|u_j\|_{q+2} \leq C\Big(\sum_{i=1}^{k} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j \|_q + \sum_{i=k+1}^{N} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j \|_{q+1} \Big) \tag{30}$$

To define the norm-equivalent functionals we further restrict the range of q to -1 and 0. The choice $q = 0$ in (27) gives the functional

$$J_P(\mathbf{u}; \mathbf{f}) = \frac{1}{2} \sum_{i=1}^{N} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j - \mathbf{f}_i \|_0^2, \tag{31}$$

while the choices $q = -1$ or $q = 0$ in (30) yield

$$J_{-1}(\mathbf{u}; \mathbf{f}) = \frac{1}{2}\Big(\sum_{i=1}^{k} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j - \mathbf{f}_i \|_{-1}^2 + \sum_{i=k+1}^{N} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j - \mathbf{f}_i \|_0^2 \Big) \tag{32}$$

and

$$J_0(\mathbf{u}; \mathbf{f}) = \frac{1}{2}\Big(\sum_{i=1}^{k} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j - \mathbf{f}_i \|_0^2 + \sum_{i=k+1}^{N} \| \sum_{j=1}^{N} \mathcal{L}_{ij} u_j - \mathbf{f}_i \|_1^2 \Big), \tag{33}$$

respectively.

5. Continuous and discrete least-squares principles

All three functionals (31)-(33) are norm equivalent and lead to well-posed principles $\{\mathbf{X}_q, J(\cdot)\}$; however, only (31) is practical in the sense discussed in §4.1. Functional (32) contains negative order norms while (33) has terms with $t_j - s_i = 2$, i.e., their total order is two. Thus, if the first-order system fails to be homogeneous elliptic then the conforming DLSP is impractical and we must consider the choices of norm equivalent, quasi-norm equivalent, or perhaps even non-equivalent DLSP.

5.1. Homogeneous systems. Consider a first-order homogeneous system and its associated CLS principle $\{\mathbf{X}_0, J_P(\cdot)\}$. The minimization space is $\mathbf{X}_0 = \{\mathbf{u} \,|\, \mathbf{u} \in \prod_{j=1}^N H^1(\Omega);\ \mathcal{R}\mathbf{u} = 0 \text{ on } \Gamma\}$, and the necessary condition is given by the variational equation

$$\text{seek } \mathbf{u} \in \mathbf{X}_0 \text{ such that} \quad Q(\mathbf{u};\mathbf{v}) = F(\mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{X}_0\,, \tag{34}$$

where now

$$Q(\mathbf{u};\mathbf{v}) = \sum_{i=1}^N \Big(\sum_{j=1}^N \mathcal{L}_{ij} u_j, \sum_{j=1}^N \mathcal{L}_{ij} v_j \Big)_0 \quad \text{and} \quad F(\mathbf{v}) = (\mathbf{f}, \sum_{j=1}^N \mathcal{L}_{ij} v_j)_0.$$

The next theorem shows that a conforming DLSP $\{\mathbf{X}^h, J_P(\cdot)\}$ is practical, quasi-optimal, and leads to easy to precondition matrices with condition numbers comparable to those in Galerkin methods.

Theorem 4. *Assume that (4) is homogeneous elliptic and let*

$$\mathbf{X}^h = \{\mathbf{u}^h \,|\, \mathbf{u}^h \in \prod_{j=1}^N S_d^h, \quad \mathcal{R}\mathbf{u}^h = 0 \quad \text{on } \Gamma\}$$

for some integer $d \geq 1$. Assume that $\mathbf{u} \in \mathbf{X}_q$ for some $q \geq 0$. Then,

1. *the least-squares variational problem (34) has a unique solution $\mathbf{u} \in \mathbf{X}_0$ for any $\mathbf{f} \in \mathbf{Y}_0$;*
2. *the discrete least-squares variational problem*

$$\text{seek } \mathbf{u}^h \in \mathbf{X}^h \text{ such that} \quad Q(\mathbf{u}^h;\mathbf{v}^h) = F(\mathbf{v}^h) \quad \forall\, \mathbf{v}^h \in \mathbf{X}^h\,, \tag{35}$$

has a unique solution $\mathbf{u}^h$ such that

$$\|\mathbf{u} - \mathbf{u}^h\|_1 \leq C h^{\tilde{d}} \|\mathbf{u}\|_{\tilde{d}+1}, \quad \tilde{d} = \min\{d, q\}; \tag{36}$$

3. *the least-squares discretization matrix A^h defined by $A_{ij}^h = Q(\boldsymbol{\phi}_i^h; \boldsymbol{\phi}_j^h)$ is spectrally equivalent to the block diagonal matrix $\operatorname{diag}(D, \ldots, D)$ with $D_{ij} = (\phi_i^h, \phi_j^h)_1$. Here $\{\boldsymbol{\phi}_i^h\}$ and $\{\phi_i^h\}$ denote standard nodal bases for $\mathbf{X}^h$ and S_d^h, respectively. Furthermore, cond$(A) = O(h^{-2})$.*

Proof. Since each $\mathcal{L}_{ij}$ is of order at most one,

$$\Big(\mathcal{L}_{ij} u_j, \mathcal{L}_{kl} v_l\Big)_0 \leq \|\mathcal{L}_{ij} u_j\|_0 \|\mathcal{L}_{kl} v_l\|_0 \leq C \|u_j\|_1 \|v_l\|_1.$$

In combination with (27) this verifies hypothesis **A.** for the homogeneous elliptic case and implies that $Q(\cdot;\cdot)$ is continuous and coercive on $\mathbf{X}_0 \times \mathbf{X}_0$. Thus 1. follows from Theorem 1. The second part follows from the fact that $\mathbf{X}^h \subset \mathbf{X}_0$ by standard finite element arguments.

Let $\boldsymbol{\xi}$ and ξ_i denote the nodal coefficients of $\mathbf{u}^h$ and u_i^h. From the identities $\boldsymbol{\xi}^T A^h \boldsymbol{\xi} = Q(\mathbf{u}^h;\mathbf{u}^h)$ and $\xi_i^T D \xi_i = (u_i^h, u_i^h)_1$ and the fact that $Q(\cdot;\cdot)$ is continuous and coercive it follows that

$$C^{-1} \sum_{i=1}^N \xi_i^T D \xi_i \leq \boldsymbol{\xi}^T A^h \boldsymbol{\xi} \leq C \sum_{i=1}^N \xi_i^T D \xi_i,$$

i.e., A^h and $\operatorname{diag}(D, \ldots, D)$ are spectrally equivalent. To find a bound for the condition number of A^h, we use (2) and coercivity of $Q(\cdot;\cdot)$:

$$C^{-1} h^{2M} |\boldsymbol{\xi}|^2 \leq \|\mathbf{u}^h\|_0^2 \leq Q(\mathbf{u}^h;\mathbf{u}^h) \leq C \|\mathbf{u}^h\|_1^2 \leq C h^{2M-2} |\boldsymbol{\xi}|^2$$

The last inequality follows from (3). Thus, cond$(A^h) = O(h^{-2})$. □

5.2. Non-homogeneous systems. For non-homogeneous systems conforming DLSP violates one or more practicality conditions. Here we consider practical DLSP which deviate from the conforming setting. We discuss the impact of such deviations upon the finite element method.

5.2.1. Weighted least-squares principles. Weighted least-squares principles are based on the premise that in finite dimensional spaces all norms are equivalent. Thus, a norm which appears in a least-squares functional and is impractical can be replaced by L^2 norm weighted by the appropriate equivalence constant. For (32) and (33) this substitution formally leads to two different discrete functionals given by

$$J_h(\mathbf{u};\mathbf{f}) = \frac{1}{2}\Big(h^2\sum_{i=1}^{k}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j-\mathbf{f}_i\|_0^2+\sum_{i=k+1}^{N}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j-\mathbf{f}_i\|_0^2\Big); \tag{37}$$

$$J_h(\mathbf{u};\mathbf{f}) = \frac{1}{2}\Big(\sum_{i=1}^{k}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j-\mathbf{f}_i\|_0^2+h^{-2}\sum_{i=k+1}^{N}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j-\mathbf{f}_i\|_0^2\Big), \tag{38}$$

and two different principles $\{\mathbf{X}^h, J_h(\cdot)\}$. However, it is easy to see that both principles belong to an equivalence class of optimization problems

$$\min_{\mathbf{u}^h\in\mathbf{X}^h} J_h(\mathbf{u}^h;\mathbf{f})\,. \tag{39}$$

where $J_h(\cdot)$ is a functional obtained from (37) or (38) by multiplication with a common (and unimportant for the minimization) factor h^α. For reasons that will become clear later we consider (37) to be the generating member of this class. The necessary condition for (37) is

$$\textit{seek } \mathbf{u}^h\in\mathbf{X}^h \textit{ such that}\quad B^h(\mathbf{u}^h;\mathbf{v}^h)=F^h(\mathbf{v}^h)\quad \forall\,\mathbf{v}^h\in\mathbf{X}^h\,, \tag{40}$$

where now

$$B^h(\mathbf{u}^h;\mathbf{v}^h) = h^2\sum_{i=1}^{k}\Big(\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h,\sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h\Big)+\sum_{i=k+1}^{n}\Big(\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h,\sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h\Big)$$

$$F^h(\mathbf{v}) = h^2\sum_{i=1}^{k}(\mathbf{f}_i,\sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h)+\sum_{i=k+1}^{N}(\mathbf{f}_i,\sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h)\,.$$

Theorem 5. *Assume that the indices s_i, t_j are given by (28) and (29), respectively, and let*

$$\mathbf{X}^h=\{\mathbf{u}^h\,|\,\mathbf{u}^h\in\prod_{j=1}^{l}S_d^h\times\prod_{j=l+1}^{N}S_{d+1}^h;\quad \mathcal{R}\mathbf{u}^h=0\quad \textit{on } \Gamma\} \tag{41}$$

where S_d^h and S_{d+1}^h are finite element spaces satisfying (1) for some $d\geq 1$. Also, assume that there exists a positive integer $r\geq d$ such that the exact solution $\mathbf{u}$ of (4) belongs to the space

$$\mathbf{X}_r=\{\mathbf{u}\,|\,\mathbf{u}\in\prod_{j=1}^{l}H^{r+1}(\Omega)\times\prod_{j=l+1}^{N}H^{r+2}(\Omega);\quad \mathcal{R}\mathbf{u}=0\quad \textit{on } \Gamma\}.$$

Then,

(1) *the least-squares problem (40) has a unique solution* $\mathbf{u}^h$ *and*

$$\sum_{j=1}^{l}\|u_j-u_j^h\|_0+\sum_{j=l+1}^{N}\|u_j-u_j^h\|_1\le h^{d+1}\Big(\sum_{j=1}^{l}\|u_j\|_{d+1}+\sum_{j=l+1}^{N}\|u_j\|_{d+2}\Big); \tag{42}$$

(2) *condition number of the matrix in (40) is bounded by* $O(h^{-4})$.

Proof. The first part of this theorem is a modification of a result of Aziz et. al. [2]. To show the second part we use (30) with $q=-1$ and proceed as in Theorem 4 to find that now

$$C^{-1}h^{2M+2}|\boldsymbol{\xi}|^2\le h^2\|\mathbf{u}^h\|_0^2\le B^h(\mathbf{u}^h;\mathbf{u}^h)\le C\|\mathbf{u}^h\|_1^2\le Ch^{2M-2}|\boldsymbol{\xi}|^2.$$

□

The weighted DLSP described here differs substantially from the conforming setting of §5.1. The CLSP related to the generating member of (39) is the pair $\{\mathbf{X}_{-1},J_{-1}(\cdot)\}$ where $\mathbf{X}_{-1}=\prod_{j=1}^{l}L^2(\Omega)\times\prod_{j=l+1}^{n}H^1(\Omega)$ and $J_{-1}(\cdot)$ is (32). In the DLSP this functional is replaced by (37) which is minimized over the subspace $\mathbf{X}^h$ of $\mathbf{X}_{-1}$. As a result, (32) can be restricted to $\mathbf{X}^h$, but (37) is not meaningful for $\mathbf{X}_{-1}$. Furthermore, if (3) holds for $\mathbf{X}^h$ one can show that

$$\frac{h^2}{C}\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big)\le J_h(\mathbf{u}^h)\le C\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big), \tag{43}$$

while in the conforming setting the bounds are independent of h. As a result, now the spectral equivalence between the least-squares matrix and the matrix associated with the standard inner product on $\mathbf{X}_{-1}$ degrades as $h\to 0$. According to §3 we call such DLS principles *quasi norm-equivalent*. The other weighted functional can be formally associated with $\{\mathbf{X}_0,J(\cdot)\}$ where $\mathbf{X}_0=\prod_{j=1}^{l}H^1(\Omega)\times\prod_{j=l+1}^{n}H^2(\Omega)$, and $J(\cdot)$ is (33). In this case $\mathbf{X}^h$ is not a subspace of $\mathbf{X}_0$ and (33) is not meaningful for discrete functions. For this reason we prefer to consider the class (39) as being generated by the first functional. Finally, note that for (38)

$$\frac{1}{C}\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big)\le J_h(\mathbf{u}^h)\le\frac{C}{h^2}\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big). \tag{44}$$

5.2.2. Discrete negative norm least-squares principles. The appearance of the factor h^2 in (43) is caused by the fact that $h\|\phi^h\|_{-1}\le hC\|\phi^h\|_0\le C\|\phi^h\|_{-1}$, which means that the equivalence between $h\|\cdot\|_0$ and $\|\cdot\|_{-1}$ degrades as $h\to 0$. Thus, asymptotically, $h\|\cdot\|_0$ is not a good approximation of the negative norm. To define a DLSP with better equivalence properties we use an approach suggested by Bramble et. al. in [5]. As before, let $D_{ij}=(\phi_i,\phi_j)_1$ and let $\mathbf{B}^h$ denote a symmetric and positive semidefinite operator that is spectrally equivalent to D^{-1} in the sense that

$$C^{-1}(D^{-1}v,v)\le(\mathbf{B}^hv,v)\le C(D^{-1}v,v),\quad\forall v\in L^2(\Omega)\,. \tag{45}$$

We define the discrete negative norm as[8]

$$\|v\|_{-h}=((h^2\mathbf{I}+\mathbf{B}^h)v,v)^{1/2},\quad\forall\mathbf{v}\in L^2(\Omega)\,. \tag{46}$$

[8]Without $\mathbf{B}^h$ norm $\|\cdot\|_{-h}$ reduces to $h\|\cdot\|_0$, i.e., this term is critical.

Lemma 1. *There exists a constant C such that for any $u \in L^2(\Omega)$*

$$C^{-1}\|u\|_{-1} \le \|u\|_{-h} \le C(h\|u\|_0 + \|u\|_{-1})\,. \tag{47}$$

If the inverse inequality (3) holds for S_d^h then

$$C^{-1}\|u^h\|_{-1} \le \|u^h\|_{-h} \le C\|u^h\|_{-1}\,, \tag{48}$$

that is, $\|\cdot\|_{-h}$ is equivalent to $\|\cdot\|_{-1}$ on S_d^h.

For a proof of this lemma we refer to [8]. To define the norm-equivalent DLS principle we replace (32) by the discrete negative norm functional

$$J_{-h}(\mathbf{u};\mathbf{f}) = \frac{1}{2}\Big(\sum_{i=1}^{k}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j - \mathbf{f}_i\|_{-h}^2 + \sum_{i=k+1}^{N}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j - \mathbf{f}_i\|_0^2\Big) \tag{49}$$

and consider the pair $\{\mathbf{X}^h, J_{-h}(\cdot)\}$ where the space $\mathbf{X}^h$ is defined as in (41). The discrete optimization problem is

$$\min_{\mathbf{u}^h\in\mathbf{X}^h} J_{-h}(\mathbf{u}^h;\mathbf{f}). \tag{50}$$

The discrete variational problem is given by

$$\text{seek } \mathbf{u}^h \in \mathbf{X}^h \text{ such that } \quad B^{-h}(\mathbf{u}^h;\mathbf{v}^h) = F^{-h}(\mathbf{v}^h) \quad \forall\,\mathbf{v}^h \in \mathbf{X}^h\,, \tag{51}$$

where

$$B^{-h}(\mathbf{u}^h;\mathbf{v}^h) = \sum_{i=1}^{k}\Big(\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h, \sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h\Big)_{-h} + \sum_{i=k+1}^{N}\Big(\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h, \sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h\Big)$$

$$F^{-h}(\mathbf{v}) = \sum_{i=1}^{k}(\mathbf{f}_i, \sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h)_{-h} + \sum_{i=k+1}^{N}(\mathbf{f}_i, \sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h)\,.$$

Theorem 6. *Assume that $\mathbf{X}^h$ is defined by (41) for some integer $d \ge 1$ and that the exact solution $\mathbf{u}$ of (4) belongs to the space $\mathbf{X}_r$, defined in Theorem 5, for some $r \ge 0$. Then,*

1. *the least-squares variational problem (51) has a unique solution $\mathbf{u}^h$ and*

$$\sum_{j=1}^{l}\|u_j - u_j^h\|_0 + \sum_{j=l+1}^{N}\|u_j - u_j^h\|_1 \le C\,h^{\tilde{d}+1}\Big(\sum_{j=1}^{l}\|u_j\|_{\tilde{d}+1} + \sum_{j=l+1}^{N}\|u_j\|_{\tilde{d}+2}\Big) \tag{52}$$

   where $\tilde{d} = \min\{r, d\}$;

2. *the condition number of the least-squares discretization matrix for (51) is bounded by $O(h^{-2})$ and this matrix is spectrally equivalent to the block-diagonal matrix*

$$M = (\underbrace{G,\ldots,G}_{l},\underbrace{D,\ldots,D}_{N-l}),$$

   where $G = (\phi_i^h, \phi_j^h)_0$ and $D = (\phi_i^h, \phi_j^h)_1$.

Proof. We first show that $B^{-h}(\cdot;\cdot)$ is continuous and coercive on $\mathbf{X}^h \times \mathbf{X}^h$. Since $u_j^h \in S_d^h$ or S_{d+1}^h and the order of each $\mathcal{L}_{ij}$ is at most one, it follows that $\mathcal{L}_{ij}u_j^h \in$

$L^2(\Omega)$ for all $i,j=1,\dots,n$. Then, using the lower bound in (47), the norm-equivalence of (32) and the fact that $\mathbf{X}^h$ is a subspace of $\mathbf{X}_{-1}$ yields

$$\begin{aligned} B^{-h}(\mathbf{u}^h;\mathbf{u}^h) &= \sum_{i=1}^{k}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h\|_{-h}^2+\sum_{i=k+1}^{N}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h\|_0^2 \\ &\geq C\Big(\sum_{i=1}^{k}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h\|_{-1}^2+\sum_{i=k+1}^{N}\|\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h\|_0^2\Big) \\ &= CJ_{-1}(\mathbf{u}^h)\geq C\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big)=\|\mathbf{u}^h\|_{\mathbf{X}_{-1}}^2. \end{aligned}$$

To show continuity we note that all discrete negative norm terms in $B^{-h}(\cdot;\cdot)$ correspond to an equation index $s_i=0$; $i=1,\dots,k$, while all L^2 terms - to an equation index $s_i=-1$; $i=k+1,\dots,n$. Let us fix $1\leq i\leq k$ so that $s_i=0$. Then, using Cauchy's inequality, the fact that the order of each $\mathcal{L}_{ij}$ is at most one, and the inverse inequality (3), the $i-$th term in $B^{-h}(\cdot;\cdot)$ can be bounded as follows:

$$\begin{aligned} \Big(\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h,\sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h\Big)_{-h} &\leq \Big(\sum_{j=1}^{N}\|\mathcal{L}_{ij}u_j^h\|_{-h}\Big)\Big(\sum_{j=1}^{N}\|\mathcal{L}_{ij}v_j^h\|_{-h}\Big) \\ &\leq \sum_{j=1}^{N}\Big(h\|\mathcal{L}_{ij}u_j^h\|_0+\|\mathcal{L}_{ij}u_j^h\|_{-1}\Big)\sum_{j=1}^{N}\Big(h\|\mathcal{L}_{ij}v_j^h\|_0+\|\mathcal{L}_{ij}v_j^h\|_{-1}\Big) \\ &\leq \sum_{j=1}^{N}\Big(h\|u_j^h\|_1+\|u_j^h\|_0\Big)\sum_{j=1}^{N}\Big(h\|v_j^h\|_1+\|v_j^h\|_0\Big) \\ &\leq \sum_{j=1}^{N}\|u_j^h\|_0\sum_{j=1}^{N}\|v_j^h\|_0\,. \end{aligned}$$

Next consider a term with $k+1\leq i\leq n$ so that $s_i=-1$. Since $\deg\mathcal{L}_{ij}\leq s_i+t_j$ and $t_j=1$ for $j=1,\dots,l$ it follows that the first l differential operators have order zero, while the last $N-l$ have orders bounded by 1, that is, $\deg\mathcal{L}_{ij}=0$ for $j=1,\dots,l$ and $\deg\mathcal{L}_{ij}\leq 1$ for $j=l+1,\dots,N$. Then,

$$\begin{aligned} \Big(\sum_{j=1}^{N}\mathcal{L}_{ij}u_j^h,\sum_{j=1}^{N}\mathcal{L}_{ij}v_j^h\Big)_0 &\leq \sum_{j=1}^{N}\|\mathcal{L}_{ij}u_j^h\|_0\sum_{j=1}^{N}\|\mathcal{L}_{ij}v_j^h\|_0 \\ &= \Big(\sum_{j=1}^{l}\|\mathcal{L}_{ij}u_j^h\|_0+\sum_{j=l+1}^{N}\|\mathcal{L}_{ij}u_j^h\|_0\Big)\Big(\sum_{j=1}^{l}\|\mathcal{L}_{ij}v_j^h\|_0+\sum_{j=l+1}^{N}\|\mathcal{L}_{ij}v_j^h\|_0\Big) \\ &\leq C\Big(\sum_{j=1}^{l}\|u_j^h\|_0+\sum_{j=l+1}^{N}\|u_j^h\|_1\Big)\Big(\sum_{j=1}^{l}\|v_j^h\|_0+\sum_{j=l+1}^{N}\|v_j^h\|_1\Big) \end{aligned}$$

Combining both inequalities yields continuity in the norm of $\mathbf{X}_{-1}$:

$$\begin{aligned} B^{-h}(\mathbf{u}^h;\mathbf{v}^h) &\leq \Big(\sum_{j=1}^{l}\|u_j^h\|_0+\sum_{j=l+1}^{N}\|u_j^h\|_1\Big)\Big(\sum_{j=1}^{l}\|v_j^h\|_0+\sum_{j=l+1}^{N}\|v_j^h\|_1\Big) \\ &= \|\mathbf{u}^h\|_{\mathbf{X}_{-1}}\|\mathbf{v}^h\|_{\mathbf{X}_{-1}}. \end{aligned} \tag{53}$$

This establishes existence and uniqueness of the least-squares solution $\mathbf{u}^h$. To prove the error estimate we note that (51) is a consistent scheme and thus, $B^{-h}(\mathbf{u}-$

$\mathbf{u}^h;\mathbf{v}^h) = 0$ for all $\mathbf{v}^h \in \mathbf{X}^h$. However, the error estimate cannot be established using a standard finite element argument because $B^{-h}(\cdot;\cdot)$ is coercive and continuous only on $\mathbf{X}^h \times \mathbf{X}^h$. Thus, we proceed as follows. Let $\mathbf{u}_I^h$ denote the interpolant of the exact solution $\mathbf{u}$ so that from (1) it follows that

$$\|\mathbf{u} - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}} \le h^{\tilde{d}+1}\|\mathbf{u}\|_{\mathbf{X}_{\tilde{d}}}.$$

Since

$$\|\mathbf{u} - \mathbf{u}^h\|_{\mathbf{X}_{-1}} \le \|\mathbf{u} - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}} + \|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}}$$

we only need to bound the last term above, which belongs to $\mathbf{X}^h$:

$$\begin{aligned}
\|\mathbf{u}^h - \mathbf{u}_I^h\|^2_{\mathbf{X}_{-1}} &\le CB^{-h}(\mathbf{u}^h - \mathbf{u}_I^h; \mathbf{u}^h - \mathbf{u}_I^h) = CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}^h - \mathbf{u}_I^h) \\
&\le CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}_I^h - \mathbf{u})^{1/2} B^{-h}(\mathbf{u}^h - \mathbf{u}_I^h; \mathbf{u}^h - \mathbf{u}_I^h)^{1/2} \\
&\le CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}_I^h - \mathbf{u})^{1/2} \|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}}.
\end{aligned}$$

Thus,

$$\|\mathbf{u}^h - \mathbf{u}_I^h\|_{\mathbf{X}_{-1}} \le CB^{-h}(\mathbf{u}_I^h - \mathbf{u}; \mathbf{u}_I^h - \mathbf{u})^{1/2}.$$

To bound the energy norm of $\mathbf{u}_I^h - \mathbf{u} = E$ note that

$$B^{-h}(E;E)^{1/2} \le C\big(\sum_{i=1}^{k} \|\sum_{j=1}^{N} \mathcal{L}_{ij}E_j\|_{-h} + \sum_{i=k+1}^{N} \|\sum_{j=1}^{N} \mathcal{L}_{ij}E_j\|_0\big).$$

Using (47) for $1 \le i \le k$,

$$\begin{aligned}
\|\sum_{j=1}^{N} \mathcal{L}_{ij}E_j\|_{-h} &\le \sum_{j=1}^{N} \big(h\|\mathcal{L}_{ij}E_j\|_0 + \|\mathcal{L}_{ij}E_j\|_{-1}\big) \\
&\le \sum_{j=1}^{N} \big(h\|E_j\|_1 + \|E_j\|_0\big) \\
&\le h^{\tilde{d}+1}\sum_{j=1}^{l} \|u_j\|_{\tilde{d}+1} + h^{\tilde{d}+2}\sum_{j=l+1}^{N} \|u_j\|_{\tilde{d}+2}.
\end{aligned}$$

For $k+1 \le i \le N$, we separate terms of orders zero and one:

$$\begin{aligned}
\|\sum_{j=1}^{N} \mathcal{L}_{ij}E_j\|_0 &\le \sum_{j=1}^{l} \|\mathcal{L}_{ij}E_j\|_0 + \sum_{j=l+1}^{N} \|\mathcal{L}_{ij}E_j\|_0 \\
&\le C\big(\sum_{j=1}^{l} \|E_j\|_0 + \sum_{j=l+1}^{N} \|E_j\|_1\big) \\
&\le Ch^{\tilde{d}+1}\Big(\sum_{j=1}^{l} \|u_j\|_{\tilde{d}+1} + \sum_{j=l+1}^{N} \|u_j\|_{\tilde{d}+2}\Big).
\end{aligned}$$

This establishes (52). Lastly, the spectral equivalence between the least-squares discretization matrix A^h and the matrix M follows from the identities $\boldsymbol{\xi}^T A\boldsymbol{\xi} = B^{-h}(\mathbf{u}^h;\mathbf{u}^h)$, $\xi_j^T D\xi_j = (u_j^h, u_j^h)_1$, $\xi_j^T G\xi_j = (u_j^h, u_j^h)_0$, and continuity and coercivity of the bilinear form. This also implies that $cond(A) = O(h^{-2})$. □

Both the weighted norm DLSP, considered in §5.2.1, and the discrete negative norm DLSP introduced in this section, are associated with the same CLSP given by the pair $\{\mathbf{X}_{-1}, J_{-1}(\cdot)\}$, but do not represent a conforming DLS principle. However,

for the weighted functional (37) the lower equivalence bound in (43) depends on h, while for (49) one can show that

$$\frac{1}{C}\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big)\le J_{-h}(\mathbf{u}^h)\le C\Big(\sum_{j=1}^{l}\|u_j^h\|_0^2+\sum_{j=l+1}^{N}\|u_j^h\|_1^2\Big).$$

According to the terminology of §3 we call such DLSP *norm-equivalent*. Norm equivalent principles give rise to linear systems which are much easier to precondition than systems obtained from quasi-norm equivalent principles. Since the Gramm matrix G is spectrally equivalent to $h^2\,I$, where I is the unit matrix in $\mathbb{R}^M$, the matrix

$$L=\operatorname{diag}(\underbrace{h^2\,I,\dots,h^2\,I}_{l},\underbrace{T,\dots,T}_{N-l}) \tag{54}$$

where T is a preconditioner for the Poisson equation, is a preconditioner for (51). Existence of good preconditioners is critical to the utility of the negative norm DLSP because it leads to dense matrices and must be implemented in an assembly free manner. This rules out the use of direct methods to solve (51).

References

[1] S. Agmon, A. Douglis, and L. Nirenberg; Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II, *Comm. Pure Appl. Math.*, 17, (1964) pp. 35–92.

[2] A. Aziz, R. Kellogg, and A. Stephens; Least-squares methods for elliptic systems, *Math. of Comp.*, 44,169, (1985) pp. 53–70.

[3] J. Bramble and J. Nitsche, *A generalized Ritz-least-squares method for Dirichlet problems,* SIAM J. Numer. Anal., 10 (1973) pp.81–93.

[4] J. Bramble and A. Schatz, *Least-squares methods for 2mth order elliptic boundary value problems,* Math. Comp., 25 (1971) pp. 1–32.

[5] J. Bramble, R. Lazarov, and J. Pasciak, *A least squares approach based on a discrete minus one inner product for first order systems,* Technical Report 94-32, Mathematical Science Institute, Cornell University, 1994.

[6] P. Bochev, *Analysis of least-squares finite element methods for the Navier-Stokes equations,* SIAM J. Num. Anal., 34/5 (1997) pp. 1817–1844.

[7] P. Bochev, Experiences with negative norm least-squares methods for the Navier-Stokes equations, *ETNA*, Vol. 6, (1997) pp. 44–62.

[8] P. Bochev, Negative norm least-squares methods for the velocity-vorticity-pressure Navier-Stokes equations, *Numerical Methods in PDE's*, 15 (1999) pp. 237–256.

[9] P. Bochev, Z. Cai, T. Manteuffel, and S. McCormick, Analysis of velocity-flux least squares methods for the Navier-Stokes equations, Part-I *SIAM. J. Num. Anal.* 35/3 (1998) pp. 990–1009.

[10] P. Bochev, T. Manteuffel, and S. McCormick Analysis of velocity-flux least squares methods for the Navier-Stokes equations, Part-II *SIAM. J. Num. Anal.*, 36/4 (1999) pp. 1125–1144.

[11] P. Bochev and M. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations,* Math. Comp., 63 (1994) pp. 479–506.

[12] P. Bochev and M. Gunzburger, *Least-squares for the velocity-pressure-stress formulation of the Stokes equations,* Comput. Meth. Appl. Mech. Engrg., 126 (1995) pp. 267–287.

[13] Z. Cai, T. Manteuffel, and S. McCormick, *First-order system least-squares for the Stokes equations, with application to linear elasticity,* SIAM J. Num. Anal. 34 (1997) pp. 1727–1741.

[14] C. Chang, *Finite element approximation for grad-div type systems in the plane,* SIAM J. Numer. Anal., 29 (1992) pp. 452–461.

[15] C. Chang and M. Gunzburger, A finite element method for first order elliptic systems in three dimensions, *Appl. Math. Comp.*, 23 (1987) pp. 171–184.

[16] P. Ciarlet; *Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[17] D. Jesperson; A least-squares decomposition method for solving elliptic equations, *Math. Comp.*, 31, (1977) pp. 873–880.

[18] B. Jiang and C. Chang, Least-squares finite elements for the Stokes problem, *Comput. Meth. Appl. Mech. Engrg.*, 78 (1990) pp. 297–311.
[19] B.-N. Jiang and L. Povinelli; Least-squares finite element method for fluid dynamics, *Comput. Meth. Appl. Mech. Engrg.*, 81 (1990) pp. 13–37.
[20] B.-N. Jiang; A least-squares finite element method for incompressible Navier-Stokes problems, *Inter. J. Numer. Meth. Fluids*, 14 (1992) pp. 843–859.
[21] B.-N. Jiang, T. Lin, and L. Povinelli; A least-squares finite element method for 3D incompressible Navier-Stokes equations, AIAA Report 93-0338, (1993).
[22] G. J. Fix, E. Stephan, On the finite element least squares approximation to higher order elliptic systems, *Arch. Rat. Mech. Anal.*, 91/2, (1986) pp. 137–151.
[23] D. Lefebvre, J. Peraire, and K. Morgan; Least-squares finite element solution of compressible and incompressible flows, *Int. J. Num. Meth. Heat Fluid Flow* **2**, (1992) pp. 99–113.
[24] Ya. A. Roitberg; A theorem about the complete set of isomorphisms for systems elliptic in the sense of Douglis and Nirenberg, *Ukrain. Mat. Zh.*, (1975) pp. 447–450.
[25] Ya. A. Roitberg and Z. Seftel; A theorem on homeomorphisms for elliptic systems and its applications, *Math. USSR Sbornik*, **7**, (1969) pp. 439–465.
[26] G. Starke; Multilevel boundary functionals for least-squares mixed finite element methods, *SIAM J. Num. Anal.* 36/4 (1999) pp. 1065–1077.
[27] L. Tang and T. Tsang, A least-squares finite element method for for time dependent incompressible flows with thermal convection, *Int. J. Numer. Methods Fluids*, 17 (1993) pp. 271–289.
[28] L. Tang and T. Tsang, Temporal, spatial and thermal features of 3-D Rayleigh-Benard convection by least-squares finite element method, *Comp. Meth. Appl. Mech. Engrg.*, 140 (1997) pp. 201–219.
[29] X. Ding, T. Tsang, *On first-order formulations of the least-squares finite element method for incompressible flows*, submitted.
[30] W. Wedland, *Elliptic Systems in the Plane*, Pitman, London, 1979.
[31] L. R. Volevich; A problem of linear programming arising in differential equations, *Uspekhi Mat. Nauk*, Vol. 18, No.3, (1963) pp. 155-162.

Pavel B. Bochev, Org. 9214/MS 1110, Sandia National Laboratories, P.O. Box 5800, Albuquerque, New Mexico, 87185

E-mail: pbboche@sandia.gov