Computers and Mathematics with Applications 130 (2023) 98–104

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A simple proof of coerciveness of first-order system least-squares methods for general second-order elliptic PDEs ☆

Shun Zhang

Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong SAR, China

ARTICLE INFO

Keywords:
Least-squares methods
Coerciveness
General second-order elliptic PDEs
LSFEM

ABSTRACT

In this paper, we present a simple proof of the coerciveness of first-order system least-squares methods for general (possibly indefinite) second-order linear elliptic PDEs under a minimal uniqueness assumption. The proof is inspired by Ku's proof [36] based on the a priori estimate of the PDE. For general linear second-order elliptic PDEs, the uniqueness, existence, and well-posedness are equivalent due to the compactness of the operator and Fredholm alternative. Thus only a minimal uniqueness assumption is assumed: the homogeneous equation has a unique zero solution. The proof presented in the paper is a straightforward and short proof using the inf-sup stability of the standard variational formulation. The proof can potentially be applied to other equations or settings once having the standard formulation's stability. As an application, we also discuss least-squares finite element methods for problems with a nonsingular H^{-1} right-hand side.

1. Introduction

The least-squares variational principle and the corresponding least-squares finite element methods based on a first-order system reformulation have been widely used in numerical solutions of partial differential equations, see for example [15,17,6,35,7,28,18,16,36,23,13]. Compared to the standard variational formulation and the related finite element methods, the first-order system least-squares finite element methods have several known advantages: (1) The least-squares continuous problem is well-posed as long as the linear PDE is well-posed. (2) The discrete problem is stable and optimally accurate as long as the discrete spaces are the subspaces of the corresponding abstract solution spaces. (3) The resulting linear systems are symmetric positive definite. Fast solvers like multigrid can be used to solve the discrete problem. (4) The least-squares functional is a good error indicator/estimator for the finite element mesh refinement and the error control. (5) The boundary condition can be easily handled in the least-squares formulation with strong or weak enforcement.

The key to establishing the well-posedness and the a priori and a posteriori error estimates of the first-order system least-squares finite element methods is the coerciveness of the corresponding bilinear forms. The simplest way uses the least-squares graph norm as the norm of choice, then the uniqueness of the PDE can be used to show the well-posedness of the least-squares system. Such analysis is too crude and often only can be used for least-squares bilinear forms we have less understanding of, see [40,39,42].

For problems like linear second-order elliptic PDEs, a more refined analysis is often needed to establish the norm equivalence: the equivalence of the least-squares norm and the standard Sobolev norms of the unknowns. For L^2-based first-order system least-squares methods for second-order elliptic equations with a new variable flux in $H(\mathrm{div})$ space, the norm equivalence of the least-squares norm and the $H(\mathrm{div})\times H^1$-norm is needed. For the H^{-1}-based first-order system least-squares method, the norm equivalence of the least-squares norm and the $L^2(\Omega)^d\times H^1$-norm is needed. The continuity of the least-squares bilinear form is often easy to check, so the main task is to prove the coerciveness of the least-squares bilinear form in the corresponding norms. For elliptic problems with a coercive standard variational formulation, the coerciveness proof is often quite simple with the help of the Poincaré inequality. The proof is subtler for general elliptic equations with possible indefinite weak problems.

☆ This work was supported in part by Research Grants Council of the Hong Kong SAR, China under the GRF Grant Project No. CityU 11302519 and CityU 11305319.

E-mail address: shun.zhang@cityu.edu.hk.

https://doi.org/10.1016/j.camwa.2022.11.021
Received 4 May 2022; Received in revised form 20 October 2022; Accepted 25 November 2022

0898-1221/© 2022 Elsevier Ltd. All rights reserved.

1.1. Review of past coerciveness proofs of least-squares methods for general second-order linear elliptic PDEs

For the general second-order elliptic equation, there are various coerciveness proofs available in the past thirty years for first-order system least-squares methods. In [15], the main assumption is the uniqueness of the solution of the linear elliptic equation, which is a very general setting that includes a real Helmholtz problem (whose standard variational formulation is naturally indefinite). The coerciveness is established for the L^2-based first-order system least-squares method. In the proof of [15], the key tool is a norm equivalence ((3.5) of [15]) based on the bijectivity of the linear elliptic operator. Later in [9], in the same general setting, the uniqueness of the elliptic PDE, the coerciveness of the H^{-1}-based first-order system least-squares method is established based on a compactness argument. In Cai's lecture notes [12], based on the [9]'s arguments, the coerciveness of the L^2-based first-order system least-squares method is established in the same general setting, see also [13]. In [36], the coerciveness is established based on the a priori stability estimate of the general elliptic PDE. For general linear second-order elliptic PDEs, the uniqueness, existence, and well-posedness are equivalent due to the compactness of the operator and Fredholm alternative, so these three proofs are essentially equivalent. One possible shortcoming of [9,12,13] is that it is a proof by contradiction. When applied to problems with mesh-dependent settings, the indirect proof can not show that the coercivity constant is independent of the mesh.

In a more restrictive setting, where some assumptions on the coefficients are assumed to ensure the standard variational formulation is coercive, the coerciveness of the L^2-based first-order system least-squares method was proved in [41]. Such a setting excludes problems like the real Helmholtz problems.

In recent years, several new ideas emerged to prove the coerciveness of first-order system least-squares methods. In [25], based on a priori estimates of two separated problems with the terms of the least-squares functionals as right-hand sides, the coerciveness is established for the Helmholtz equation in a complex setting. The same argument is used in [4] for the self-adjoint diffusion-reaction problems. In [24], a new proof technique based on a lemma introduced in the discontinuous Petrov-Galerkin method [33] is used to prove the coerciveness of specially designed first-order least-squares finite element methods with weakly enforced boundary conditions.

The analysis for finite element approximations to the possible indefinite general second-order elliptic equations is more complicated than the coercive cases. In [43,44], the conforming finite element approximation of the general elliptic equations are discussed. Recently, in [26,21,22], nonconforming and mixed finite element approximations are discussed. All the results require some regularity and the mesh size of the discretization is small enough. Compared to the conforming, nonconforming, and mixed methods, the least-square finite element method for the general elliptic equations has an extra advantage: as long as the finite element spaces are conforming subspaces of the abstract solution spaces, the discrete problem is coercive without any restriction on the mesh size.

1.2. Contribution of this paper

Due to the equivalence of uniqueness, existence, and well-posedness of the second-order elliptic PDE, we only assume a *minimal uniqueness assumption: the homogeneous equation has a unique zero solution*. A simple proof for the coerciveness of two formulations of the first-order system least-squares methods for general second-order elliptic PDEs under the minimal uniqueness assumption is presented. One formulation is based on the diffusion flux reformulation, and the other one is based on a total physical flux reformulation. Both the L^2 and H^{-1} first-order system least-squares methods are discussed.

The central part of our proof is only two to three lines. The main component of the proof is the inf-sup stability of the standard weak formulation, which is guaranteed by the minimal assumption. Besides, this proof has the potential to be applied to other cases and scenarios, where the standard compactness/proof of contradiction tricks in [15,9,12,13,36] may not be applicable, but an inf-sup stable bilinear form is available.

As an application, we discuss least-squares finite element methods for the seconder order elliptic equation with an H^{-1} right-hand side in the paper. The key to setting up a first-order system is to choose the new vector variable carefully so it is in the $H(\mathrm{div})$ space. In [31], least-squares methods with singular righthand sides are discussed. We do not pursue this direction in the paper.

1.3. Outline, notations, and function spaces

The paper is organized as follows. In the remaining part of this subsection, we present notations and function spaces. In Section 2 we introduce mathematical equations for the second-order scalar elliptic partial differential equations and discuss the solution's existence, uniqueness, and well-posedness. Two variants of the first-order systems are presented. Then we set up the corresponding L^2 and H^{-1} least-squares functionals and give sufficient conditions to ensure the coerciveness to avoid repeating. In Section 3, a simple proof based on the inf-sup stability is presented. Section 4 discusses the least-squares finite element method for the elliptic equation with an H^{-1} right-hand side.

Let Ω be a bounded, open, connected subset of $\mathbb{R}^d (d = 2 \text{ or } 3)$ with a Lipschitz continuous boundary $\partial\Omega$. Denote by $\mathbf{n} = (\mathbf{n}_1, ..., \mathbf{n}_d)^T$ the outward unit vector normal to the boundary. We partition the boundary of the domain Ω into two open subsets Γ_D and Γ_N such that $\partial\Omega = \overline{\Gamma_D} \cup \overline{\Gamma_N}$ and $\Gamma_D \cap \Gamma_N = \emptyset$. For simplicity, we assume that Γ_D is not empty (i.e., $\mathrm{meas}(\Gamma_D) \neq 0$).

We use the standard notations and definitions for the Sobolev spaces $H^s(\Omega)^d$ for $s \geq 0$. The standard associated inner product is denoted by $(\cdot,\cdot)_{s,\Omega}$, and their respective norms are denoted by $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,\partial\Omega}$. The notation $|\cdot|_{s,\Omega}$ is used for semi-norms. (We suppress the superscript d because the dependence on dimension will be clear by context. We also omit the subscript Ω from the inner product and norm designation when there is no risk of confusion.) For $s = 0$, $H^s(\Omega)^d$ coincides with $L^2(\Omega)^d$. The symbols $\nabla\cdot$ and ∇ stand for the divergence and gradient operators, respectively. Set $H^1_D(\Omega) := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$ and $H^1_0(\Omega) := \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega\}$. Define $H(\mathrm{div};\Omega) = \{\boldsymbol{\tau} \in L^2(\Omega)^d : \nabla\cdot\boldsymbol{\tau} \in L^2(\Omega)\}$, which is a Hilbert space under the norm $\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} = \left(\|\boldsymbol{\tau}\|_0^2 + \|\nabla\cdot\boldsymbol{\tau}\|_0^2\right)^{1/2}$, and define its subspace $H_N(\mathrm{div};\Omega) = \{\boldsymbol{\tau} \in H(\mathrm{div};\Omega) : \mathbf{n}\cdot\boldsymbol{\tau} = 0 \text{ on } \Gamma_N\}$. Since most of our discussion will be in $H_N(\mathrm{div};\Omega) \times H^1_D(\Omega)$, we introduce the notation for simplicity,

$$\mathbb{X} := H_N(\mathrm{div};\Omega) \times H^1_D(\Omega). \tag{1.1}$$

The space $(H^1_D(\Omega))'$ is defined as the dual space of $H^1_D(\Omega)$ and consists of the functionals v for which the norm

$$\|v\|_{-1} = \sup_{w \in H^1_D(\Omega)} \frac{\langle v, w\rangle_{(H^1_D(\Omega))' \times H^1_D(\Omega)}}{\|\nabla w\|_0} \tag{1.2}$$

is finite. Here, we use the notation $\langle\cdot,\cdot\rangle_{(H^1_D(\Omega))' \times H^1_D(\Omega)}$ to denote the duality pairing between $(H^1_D(\Omega))'$ and $H^1_D(\Omega)$. For simplicity, we will only use $\langle\cdot,\cdot\rangle$ in the rest part of the paper if no confusion is caused.

When $\partial\Omega = \Gamma_D$, then it is the usual $H^{-1}(\Omega)$; see for example, discussions in [30,11]. We use $\|\nabla\cdot\|_0$ instead of the standard $\|\cdot\|_1$ norm due to the Poincaré inequality. The space $(H^1_D(\Omega))'$ is a Hilbert space. Let $S : (H^1_D(\Omega))' \to H^1_D(\Omega)$ to be the solution operator of the following problem:

$$z \in H^1_D(\Omega) \quad \text{with} \quad (\nabla z, \nabla v) = \langle f, v\rangle \quad \forall v \in H^1_D(\Omega). \tag{1.3}$$

That is, for $f \in (H^1_D(\Omega))'$, $Sf = z \in H^1_D(\Omega)$ is the solution of (1.3). Again, we modify the setting in [9] and only use the Poisson problem in its

weak form. In [9], the PDE version of $-\Delta u + u = f$ with boundary conditions is used. It is easy to see that the two settings are equivalent. We have the following result (Lemma 2.1 of [9]): The inner product on $(H^1_D(\Omega))' \times (H^1_D(\Omega))'$ is given by

$$\langle v, Sq\rangle \quad \forall v, q \in (H^1_D(\Omega))'. \tag{1.4}$$

We have $\|q\|^2_{-1} = \langle q, Sq\rangle$, for $q \in (H^1_D(\Omega))'$.

Let $f = S^{-1}z$ and $v = z$ in (1.3), we have

$$\|\nabla z\|^2_0 = \langle S^{-1}z, z\rangle, \quad \forall z \in H^1_D(\Omega). \tag{1.5}$$

Thus, for $v \in H^1_D(\Omega)$ and $w = S^{-1}v \in (H^1_D(\Omega))'$, we have

$$\|w\|^2_{-1} = \langle w, Sw\rangle = \langle S^{-1}v, SS^{-1}v\rangle = \langle S^{-1}v, v\rangle = \|\nabla v\|^2_0. \tag{1.6}$$

By the Poincaré inequality, we also have the following result: Assume $f \in L^2(\Omega)$, then

$$\|f\|_{-1} = \sup_{v\in H^1_D(\Omega)} \frac{(f,v)}{\|\nabla v\|_0} \le C \sup_{v\in H^1_D(\Omega)} \frac{(f,v)}{\|v\|_0} \le C\|f\|_0. \tag{1.7}$$

For $\boldsymbol{\tau} \in H_N(\mathrm{div};\Omega)$, we also have

$$\|\nabla\cdot\boldsymbol{\tau}\|_{-1} = \sup_{v\in H^1_D(\Omega)} \frac{(\nabla\cdot\boldsymbol{\tau}, v)}{\|\nabla v\|_0} = \sup_{v\in H^1_D(\Omega)} \frac{(\boldsymbol{\tau}, \nabla v)}{\|\nabla v\|_0} \le \|\boldsymbol{\tau}\|_0. \tag{1.8}$$

2. General second-order elliptic equations and their least-squares methods

2.1. General second-order elliptic equations

For $f \in L^2(\Omega)$ is a given scalar function, consider the general second-order elliptic equation in divergence form

$$\begin{aligned} -\nabla\cdot(A\nabla u) + \mathbf{b}\cdot\nabla u + cu &= f \text{ in } \Omega,\\ u &= 0 \text{ on } \Gamma_D,\\ A\nabla u\cdot\mathbf{n} &= 0 \text{ on } \Gamma_N, \end{aligned} \tag{2.1}$$

and its adjoint problem (equation in physical form)

$$\begin{aligned} -\nabla\cdot(A\nabla u + \mathbf{b}u) + cu &= f \text{ in } \Omega,\\ u &= 0 \text{ on } \Gamma_D,\\ (A\nabla u + \mathbf{b}u)\cdot\mathbf{n} &= 0 \text{ on } \Gamma_N. \end{aligned} \tag{2.2}$$

We assume the following very mild conditions on the domain and coefficients.

Assumption 2.1. The domain Ω is a bounded, open, connected subset of $\mathbb{R}^d$ ($d = 2$ or 3) with a Lipschitz continuous boundary $\partial\Omega$.

The diffusion coefficient matrix $A \in L^\infty(\Omega)^{d\times d}$ is a given $d\times d$ tensor-valued function; the matrix A is uniformly symmetric positive definite: there exist positive constants $0 < \Lambda_0 \le \Lambda_1$ such that

$$\Lambda_0\mathbf{y}^T\mathbf{y} \le \mathbf{y}^T A\mathbf{y} \le \Lambda_1\mathbf{y}^T\mathbf{y} \tag{2.3}$$

for all $\mathbf{y} \in \mathbb{R}^d$ and almost all $x \in \Omega$. The coefficients $\mathbf{b} \in L^\infty(\Omega)^d$ and $c \in L^\infty(\Omega)$ are given vector- and scalar-valued bounded functions, respectively.

For simplicity, we define

$$Xv = \mathbf{b}\cdot\nabla v + cv. \tag{2.4}$$

Define the following bilinear form:

$$\begin{aligned} &a(w,v) = (A\nabla w, \nabla v) + (Xw, v) = (A\nabla w, \nabla v) + (\nabla w, \mathbf{b}v) + (cw, v)\\ &\quad w, v \in H^1_D(\Omega). \end{aligned} \tag{2.5}$$

With the help of the Poincare inequality (which is true under our assumption of the boundary condition), it is easy to check that the bilinear form is continuous due to Assumption 2.1:

$$a(w,v) \le C\|\nabla v\|_0\|\nabla w\|_0 \quad v \in H^1_D(\Omega) \text{ and } w \in H^1_D(\Omega). \tag{2.6}$$

Remark 2.2. When $\mathbf{b} = 0$ and $c = -k^2$ for some $k > 0$, the equation is a real Helmholtz equation. We have $a(v,v) = \|A^{1/2}\nabla v\|^2_0 - \|kv\|^2_0$, which is indefinite for a large k. However, as long as k^2 square is not an eigenvalue of $(A\nabla v, \nabla w)$, the equation still has a unique solution.

2.2. Existence, uniqueness, and well-posedness of the linear second-order elliptic PDE solution

This subsection discusses the theories on the existence, uniqueness, and well-posedness of the second-order linear elliptic PDE. In order to handle the H^{-1} least-squares formulations, we discuss the two problems with $(H^1_D(\Omega))'$ righthand sides.

For a $g \in (H^1_D(\Omega))'$, assume that $u \in H^1_D(\Omega)$ satisfies the following weak problem of the equation in divergence form:

$$u \in H^1_D(\Omega), \text{ such that } a(u,v) = \langle g, v\rangle_{(H^1_D(\Omega))'\times H^1_D(\Omega)} \quad \forall v \in H^1_D(\Omega), \tag{2.7}$$

or satisfies the adjoint weak problem of the equation in physical form,

$$u \in H^1_D(\Omega), \text{ such that } a(v,u) = \langle g, v\rangle_{(H^1_D(\Omega))'\times H^1_D(\Omega)} \quad \forall v \in H^1_D(\Omega). \tag{2.8}$$

For a possible indefinite linear second-order elliptic equation, the solution's existence and uniqueness theory is based on the Fredholm alternative. Since we assume the ellipticity of the PDEs (conditions (2.3) on A), the operators associated with the divergence form and adjoint physical form problems are Fredholm operators of index zero. The uniqueness, existence, and well-posedness are equivalent. We collect the existence and uniqueness results of linear second-order elliptic equations in the following theorem. These results are standard and are known to the experts. We present here only for completeness and clarifications.

Theorem 2.3. *Assuming that Assumption 2.1 is true, the following assumptions are equivalent:*

(1) The homogeneous equation associated to (2.7), i.e., $a(u,v) = 0$, for all $v \in H^1_D(\Omega)$, has $u = 0$ as its unique solution.

(2) The weak problem (2.7) has the following stability bound:

$$\|\nabla u\|_0 \le C\|g\|_{-1} \quad \forall g \in (H^1_D(\Omega))'. \tag{2.9}$$

(3) The weak problem (2.7) has a unique solution $u \in H^1_D(\Omega)$ for any $g \in (H^1_D(\Omega))'$.

(4) The homogeneous equation associated to (2.8), i.e., $a(v,u) = 0$, for all $v \in H^1_D(\Omega)$, has $u = 0$ as its unique solution.

(5) The adjoint weak problem (2.8) has the following stability bound:

$$\|\nabla u\|_0 \le C\|g\|_{-1} \quad \forall g \in (H^1_D(\Omega))'. \tag{2.10}$$

(6) The adjoint weak problem (2.8) has a unique solution $u \in H^1_D(\Omega)$ for any $g \in (H^1_D(\Omega))'$.

We sketch two proofs.

Proof. (I) By the Fredholm alternative, the solution spaces of the two adjoint homogeneous problems have the same finite number of linearly independent solutions, see page 200 of [5] or Theorem 5.11 of [32]. If one of them has only 0 as its solution, so is the other, thus (1) $\Longleftrightarrow$ (4).

From page 200 of [5], if (4) is assumed, then for any $g \in (H^1_D(\Omega))'$, the weak problem (2.7) has a unique solution in $H^1_D(\Omega)$ since the requirement of the existence of a solution by the Fredholm alternative is that $\langle g, v\rangle = 0$ for any $v \in H^1_D(\Omega)$ being the solution of homogeneous adjoint equation (2.8), which is 0. Then we get (4) $\Longrightarrow$ (3). Similarly, we get (1) $\Longrightarrow$ (6).

Again from the Fredholm alternative, we get (1) $\Longrightarrow$ (2) and (4) $\Longrightarrow$ (5), the a priori estimate.

On the other hand, if (2) is true, we immediately have (1) by letting $g = 0$. Similarly, (5) $\Longrightarrow$ (4).

The conclusions (3) $\Longrightarrow$ (1) and (6) $\Longrightarrow$ (4) are obvious. Thus, we have the equivalence of the assumptions in the theorem. □

Proof. (II) The equivalence of (1) $\Longleftrightarrow$ (2) $\Longleftrightarrow$ (3) can also be proved directly without the help of the adjoint problem. Following the proof of Theorem 8.3 of [32], we get (1) $\Longrightarrow$ (3). (1) $\Longrightarrow$ (2) is a result of the Fredholm alternative. (2) $\Longrightarrow$ (1) and (3) $\Longrightarrow$ (1) are obvious. Similarly, we have (4) $\Longleftrightarrow$ (5) $\Longleftrightarrow$ (6). We only need to show the equivalence of (1-3) and (4-6).

From the Banach-Necăs-Babuška theory [2,8,47], the problem (2.7), which has a continuous bilinear form, is well-posed if and only the following conditions hold:

$$\inf_{v\in H_D^1(\Omega)}\sup_{w\in H_D^1(\Omega)}\frac{a(v,w)}{\|\nabla v\|_0\|\nabla w\|_0}>0,\qquad \inf_{w\in H_D^1(\Omega)}\sup_{v\in H_D^1(\Omega)}\frac{a(v,w)}{\|\nabla v\|_0\|\nabla w\|_0}>0. \tag{2.11}$$

furthermore if (2.11) holds, then

$$\inf_{v\in H_D^1(\Omega)}\sup_{w\in H_D^1(\Omega)}\frac{a(v,w)}{\|\nabla v\|_0\|\nabla w\|_0}=\inf_{w\in H_D^1(\Omega)}\sup_{v\in H_D^1(\Omega)}\frac{a(v,w)}{\|\nabla v\|_0\|\nabla w\|_0}=\beta>0. \tag{2.12}$$

Thus, we immediately get (2.8) is also well-posed and thus (1-3) $\Longleftrightarrow$ (4-6). The proof of the opposite direction is identical. □

Remark 2.4. The inf-sup condition (2.11) or (2.12) is also equivalent to

$$\beta\|\nabla v\|_0\le\sup_{w\in H_D^1(\Omega)}\frac{a(v,w)}{\|\nabla w\|_0}\quad\text{and}\quad \beta\|\nabla v\|_0\le\sup_{w\in H_D^1(\Omega)}\frac{a(w,v)}{\|\nabla w\|_0}\quad \forall v\in H_D^1(\Omega), \tag{2.13}$$

for some $\beta>0$. These conditions are also equivalent to the stability estimates (2.9) and (2.10).

Remark 2.5. The Banach-Necăs-Babuška theory is established for general linear problems in the Banach/Hilbert spaces setting, which states that the problem is well-posed if and only if the bilinear form is continuous and satisfies the inf-sup conditions (2.11). For second-order elliptic equations, due to the compactness and Fredholm alternative, we only need one condition from Theorem 2.3.

Remark 2.6. The stability constant β in (2.12) is mesh-independent, but it can be very close to zero. For example, for the Helmholtz equation with $\mathbf{b}=0$ and $-c$ is very close to the eigenvalue of $(A\nabla v,\nabla w)$.

Remark 2.7. A proof of (1) $\Longrightarrow$ (3) and (4) $\Longrightarrow$ (6) can be found in Lemma 2.2 of [10] for some more general mixed boundary conditions with a compactness argument.

2.3. Least-squares minimization problems

For the elliptic equation in the divergence form (2.1), let the flux $\boldsymbol{\sigma}=-A\nabla u$. We have the first-order system:

$$\begin{cases}\boldsymbol{\sigma}+A\nabla u &= 0 \quad \text{in } \Omega,\\ \nabla\cdot\boldsymbol{\sigma}+Xu &= f \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \Gamma_D,\\ \boldsymbol{\sigma}\cdot\mathbf{n} &= 0 \quad \text{on } \Gamma_N.\end{cases} \tag{2.14}$$

For $u\in H_D^1(\Omega)$, we have $\boldsymbol{\sigma}=-A\nabla u\in L^2(\Omega)^d$ and $\nabla\cdot\boldsymbol{\sigma}=f-Xu\in L^2(\Omega)$, so $(\boldsymbol{\sigma},u)\in\mathbb{X}=H_N(\mathrm{div};\Omega)\times H_D^1(\Omega)$. For $(\boldsymbol{\tau},v)\in\mathbb{X}$, define least-squares functionals for the system (2.14),

$$L(\boldsymbol{\tau},v;f):=\|A^{-1/2}\boldsymbol{\tau}+A^{1/2}\nabla v\|_0^2+\|\nabla\cdot\boldsymbol{\tau}+Xv-f\|_0^2, \tag{2.15}$$

$$L_{-1}(\boldsymbol{\tau},v;f):=\|A^{-1/2}\boldsymbol{\tau}+A^{1/2}\nabla v\|_0^2+\|\nabla\cdot\boldsymbol{\tau}+Xv-f\|_{-1}^2. \tag{2.16}$$

The corresponding least-squares minimization problems are:

$$\text{Find } (\boldsymbol{\sigma},u)\in\mathbb{X} \text{ s.t. } L(\boldsymbol{\sigma},u;f)=\inf_{(\boldsymbol{\tau},v)\in\mathbb{X}}L(\boldsymbol{\tau},v;f) \tag{2.17}$$

$$\text{and find } (\boldsymbol{\sigma},u)\in\mathbb{X} \text{ s.t. } L_{-1}(\boldsymbol{\sigma},u;f)=\inf_{(\boldsymbol{\tau},v)\in\mathbb{X}}L_{-1}(\boldsymbol{\tau},v;f). \tag{2.18}$$

For the elliptic equation in the physical form (2.2), define the total flux as $\boldsymbol{\sigma}=-A\nabla u-\mathbf{b}u$, then

$$\begin{cases}\boldsymbol{\sigma}+A\nabla u+\mathbf{b}u &= 0 \quad \text{in } \Omega,\\ \nabla\cdot\boldsymbol{\sigma}+cu &= f \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \Gamma_D,\\ \boldsymbol{\sigma}\cdot\mathbf{n} &= 0 \quad \text{on } \Gamma_N.\end{cases} \tag{2.19}$$

Similarly, we have $(\boldsymbol{\sigma},u)\in\mathbb{X}$. For $(\boldsymbol{\tau},v)\in\mathbb{X}$, define least-squares functionals for the system (2.19),

$$J(\boldsymbol{\tau},v;f):=\|\boldsymbol{\tau}+A\nabla v+\mathbf{b}v\|_0^2+\|\nabla\cdot\boldsymbol{\tau}+cv-f\|_0^2, \tag{2.20}$$

$$J_{-1}(\boldsymbol{\tau},v;f):=\|\boldsymbol{\tau}+A\nabla v+\mathbf{b}v\|_0^2+\|\nabla\cdot\boldsymbol{\tau}+cv-f\|_{-1}^2. \tag{2.21}$$

Remark 2.8. We discuss the problem (2.2) independently since the total physical flux is different from the flux (2.1) and has many applications. The introduction of the total flux in the least-squares context can be found in [37].

For any $(\boldsymbol{\tau},v)\in L^2(\Omega)^d\times H_D^1(\Omega)$ and $(\boldsymbol{\rho},w)\in\mathbb{X}$, define the following two norms, respectively:

$$|||(\boldsymbol{\tau},v)|||^2:=\|\boldsymbol{\tau}\|_0^2+\|\nabla v\|_0^2\quad\text{and}\quad |||(\boldsymbol{\rho},w)|||^2:=\|\boldsymbol{\rho}\|_0^2+\|\nabla\cdot\boldsymbol{\rho}\|_0^2+\|\nabla w\|_0^2. \tag{2.22}$$

Remark 2.9. By the Poincaré inequality, we can also replace $\|\nabla v\|_0$ by a standard $\|v\|_1$ in the above norms. We use the norm $\|\nabla v\|_0$ for simplicity.

From the Cauchy-Schwarz, triangle, and Poincaré inequalities and the bounds of the coefficients ($A,\mathbf{b}$, and c) of the underlying problems (Assumption 2.1), we immediately have the following upper bounds: There exists $C>0$, such that,

$$L(\boldsymbol{\tau},v;0)\le C|||(\boldsymbol{\tau},v)|||^2\quad\text{and}\quad J(\boldsymbol{\tau},v;0)\le C|||(\boldsymbol{\tau},v)|||^2,\quad\forall(\boldsymbol{\tau},v)\in\mathbb{X}. \tag{2.23}$$

Together with (1.7) and (1.8), we can show that there exists $C>0$, such that,

$$L_{-1}(\boldsymbol{\tau},v;0)\le C\|(\boldsymbol{\tau},v)\|^2\quad\text{and}\quad J_{-1}(\boldsymbol{\tau},v;0)\le C\|(\boldsymbol{\tau},v)\|^2,\quad\forall(\boldsymbol{\tau},v)\in\mathbb{X}. \tag{2.24}$$

In this paper, we want to prove the following coerciveness results: For all $(\boldsymbol{\tau},v)\in\mathbb{X}$, there exists $C>0$, such that,

$$C|||(\boldsymbol{\tau},v)|||^2\le L(\boldsymbol{\tau},v;0),\quad C\|(\boldsymbol{\tau},v)\|^2\le L_{-1}(\boldsymbol{\tau},v;0), \tag{2.25}$$

$$C|||(\boldsymbol{\tau},v)|||^2\le J(\boldsymbol{\tau},v;0),\quad C\|(\boldsymbol{\tau},v)\|^2\le J_{-1}(\boldsymbol{\tau},v;0). \tag{2.26}$$

Remark 2.10. If the coercive results (2.25) to (2.26) hold, then we have the following equivalence:

$$C_1L(\boldsymbol{\tau},v;0)\le J(\boldsymbol{\tau},v;0)\le C_2L(\boldsymbol{\tau},v;0)\quad(\boldsymbol{\tau},v)\in\mathbb{X}; \tag{2.27}$$

$$C_3L_{-1}(\boldsymbol{\tau},v;0)\le J_{-1}(\boldsymbol{\tau},v;0)\le C_3L_{-1}(\boldsymbol{\tau},v;0)\quad(\boldsymbol{\tau},v)\in\mathbb{X}. \tag{2.28}$$

Before discussing the complete coerciveness proofs, we introduce some conditions that ensure coerciveness. These conditions are elementary and well-known. We list them here to avoid repeating them in future proofs.

Theorem 2.11. *For $(\boldsymbol{\tau}, v) \in \mathbb{X}$, we have*

$$C\|\nabla v\|_0^2 \le L(\boldsymbol{\tau}, v; 0) \Longrightarrow C|||(\boldsymbol{\tau}, v)|||^2 \le L(\boldsymbol{\tau}, v; 0), \tag{2.29}$$

$$C\|\nabla v\|_0^2 \le L_{-1}(\boldsymbol{\tau}, v; 0) \Longrightarrow C|||(\boldsymbol{\tau}, v)|||^2 \le L_{-1}(\boldsymbol{\tau}, v; 0), \tag{2.30}$$

$$C\|\nabla v\|_0^2 \le J(\boldsymbol{\tau}, v; 0) \Longrightarrow C|||(\boldsymbol{\tau}, v)|||^2 \le J(\boldsymbol{\tau}, v; 0), \tag{2.31}$$

$$C\|\nabla v\|_0^2 \le J_{-1}(\boldsymbol{\tau}, v; 0) \Longrightarrow C|||(\boldsymbol{\tau}, v)|||_0^2 \le J_{-1}(\boldsymbol{\tau}, v; 0). \tag{2.32}$$

Proof. From the triangle and Poincaré inequalities, we have

$$\|\nabla \cdot \boldsymbol{\tau}\|_0 \le \|\nabla \cdot \boldsymbol{\tau} + Xv\|_0 + \|Xv\|_0 \le \|\nabla \cdot \boldsymbol{\tau} + Xv\|_0 + C\|\nabla v\|_0$$

and $\|\boldsymbol{\tau}\|_0 \le \|\boldsymbol{\tau} + A\nabla v\|_0 + C\|\nabla v\|_0$.

Thus, if the first inequality of (2.29) is true, we have the coerciveness. The rest results in the theorem can be proved similarly. □

3. A simple direct proof from the Inf-Sup stability

We present a simple proof of the coerciveness of the least-squares functionals.

Theorem 3.1. *Assuming one of the conditions of Theorem 2.3 is true, the coercive results in (2.25) and (2.26) hold.*

Proof. By the integration by parts, we have $(\boldsymbol{\tau}, \nabla w) + (\nabla \cdot \boldsymbol{\tau}, w) = 0$ for any $(\boldsymbol{\tau}, w) \in H_N(\mathrm{div}; \Omega) \times H_D^1(\Omega)$. Then for $\boldsymbol{\tau} \in H_N(\mathrm{div}; \Omega)$ and v and w in $H_D^1(\Omega)$, we have

$$a(v, w) = (A\nabla v, \nabla w) + (Xv, w) = (A\nabla v + \boldsymbol{\tau}, \nabla w) + (\nabla \cdot \boldsymbol{\tau} + Xv, w), \tag{3.1}$$

$$\begin{aligned} a(w, v) &= (A\nabla v + \mathbf{b}v, \nabla w) + (cv, w) \\ &= (A\nabla v + \mathbf{b}v + \boldsymbol{\tau}, \nabla w) + (\nabla \cdot \boldsymbol{\tau} + cv, w). \end{aligned} \tag{3.2}$$

It follows from (2.13), (3.1), the Cauchy-Schwarz and Poincaré inequalities, the definition of the minus one norm, the assumption on A, and (1.7), for any $(\boldsymbol{\tau}, v) \in H_N(\mathrm{div}; \Omega) \times H_D^1(\Omega)$,

$$\begin{aligned} \beta\|\nabla v\|_0 &\le \sup_{w \in H_D^1(\Omega)} \frac{a(v, w)}{\|\nabla w\|_0} = \sup_{w \in H_D^1(\Omega)} \frac{(A\nabla v + \boldsymbol{\tau}, \nabla w) + (\nabla \cdot \boldsymbol{\tau} + Xv, w)}{\|\nabla w\|_0} \\ &\le C\|A^{1/2}\nabla v + A^{-1/2}\boldsymbol{\tau}\|_0 + \|\nabla \cdot \boldsymbol{\tau} + Xv\|_{-1} \\ &\le C(\|A^{1/2}\nabla v + A^{-1/2}\boldsymbol{\tau}\|_0 + \|\nabla \cdot \boldsymbol{\tau} + Xv\|_0). \end{aligned}$$

Similarly, using (3.2), for any $(\boldsymbol{\tau}, v) \in H_N(\mathrm{div}; \Omega) \times H_D^1(\Omega)$, we have

$$\begin{aligned} \beta\|\nabla v\|_0 &\le \sup_{w \in H_D^1(\Omega)} \frac{a(w, v)}{\|\nabla w\|_0} = \sup_{w \in H_D^1(\Omega)} \frac{(A\nabla v + \mathbf{b}v + \boldsymbol{\tau}, \nabla w) + (\nabla \cdot \boldsymbol{\tau} + cv, w)}{\|\nabla w\|_0} \\ &\le C\|A\nabla v + \mathbf{b}v + \boldsymbol{\tau}\|_0 + \|\nabla \cdot \boldsymbol{\tau} + cv\|_{-1} \\ &\le C(\|A\nabla v + \mathbf{b}v + \boldsymbol{\tau}\|_0 + \|\nabla \cdot \boldsymbol{\tau} + cv\|_0). \end{aligned}$$

Thus the conditions in Theorem 2.11 are satisfied, we have the coerciveness. □

Remark 3.2. Compared to the proofs in [15,9,12,13,36], the proof is simplified. This proof is a result of studying the existing proofs and finding their connections. We hide the compactness argument used in [15,9,12,13] to the stability of the original bilinear form. For the second-order elliptic equation, the proof of stability relies on the Fredholm alternative, which also relies on compactness; thus, the proofs are essentially equivalent. The advantage of the current proof is that it can be applied to other cases where the stability of the bilinear form is available, and the stability is not necessarily related to compactness. For example, for an L^2 div least-squares method with Crouzeix-Raviart finite element approximation for the general second-order elliptic equation, we have the discrete stability of the bilinear form under the condition the mesh is fine enough [21]. We then can use the above proof method to derive the discrete coerciveness of the L^2 div least-squares method with Crouzeix-Raviart finite element approximation, see [38]. On the other hand, the compactness arguments in [9,12,13,36] do not work in this discrete scenario since the compactness argument is a proof of contradiction and can not be used in mesh-dependent methods to prove that the coercive constant is independent of the mesh size.

As seen from the proof, the proof also guides us in choosing the right weight for different terms.

4. Extension to least-squares finite element methods with H^{-1} right-hand side

This section discusses the least-squares finite element methods for the second-order elliptic equation with an H^{-1} right-hand side. Such a problem appears in many situations, for example, in the goal-oriented, a posteriori error estimate [34] and Darcy flow.

For simplicity, we only consider the (2.1)-type of equation with a pure Dirichlet boundary condition and the L^2-version of the least-squares finite element method. As pointed out in [30,11], any functional in $H^{-1} := (H_0^1(\Omega))'$ can be written as $f - \nabla \cdot \mathbf{g}$ for $f \in L^2(\Omega)$ and $\mathbf{g} \in L^2(\Omega)^d$. Thus, we consider the following problem:

$$-\nabla \cdot (A\nabla u) + \mathbf{b} \cdot \nabla u + cu = f_1 - \nabla \cdot (A\mathbf{f}_2) \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{4.1}$$

Here $f_1 \in L^2(\Omega)$ and $\mathbf{f}_2 \in L^2(\Omega)^d$ are given functions. We have $f_1 + \nabla \cdot (A\mathbf{f}_2) \in H^{-1}(\Omega)$. The divergence of $A\mathbf{f}_2$ should be understood in the distributional sense, i.e.,

$$(\nabla \cdot (A\mathbf{f}_2), v) = -(A\mathbf{f}_2, \nabla v) \quad \forall v \in H_0^1(\Omega).$$

A typical $A\mathbf{f}_2$ can be $A\nabla v_h$ for v_h being a function in the conforming finite element space. Such a righthand side also appears in the recovery-based error estimators, see [20]. The standard variational problem reads: Find $u \in H_0^1(\Omega)$, such that,

$$(A\nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u + cu, v) = (f_1, v) + (A\mathbf{f}_2, \nabla v) \quad \forall v \in H_0^1(\Omega). \tag{4.2}$$

For equation (4.1), the quantity $A\nabla u$ is not in $H(\mathrm{div}; \Omega)$. Since $\nabla \cdot (A\mathbf{f}_2)$ is only in $H^{-1}(\Omega)$, not $L^2(\Omega)$, thus $\nabla \cdot (A\nabla u) = \mathbf{b} \cdot \nabla u + cu - f_1 - \nabla \cdot (A\mathbf{f}_2)$ is only in $H^{-1}(\Omega)$. Taking the simplest example, let $\mathbf{b} = 0$, $c = 0$, $f_1 = 0$, $A = I$, and $\mathbf{f}_2 = \nabla v_h$, where v_h is a continuous piecewise linear finite element function. It is well-known that we usually do not have $[\![\nabla v_h \cdot \mathbf{n}_F]\!]_F = 0$ across an internal edge F of a finite element mesh. This term often appears in the residual type of a posteriori error estimator [1,46], and recovering it in the $H(\mathrm{div})$-conforming space is the foundation of a recovery-based error estimator [19]. Then for this simplest example, $A\nabla u = -A\nabla v_h \notin H(\mathrm{div}; \Omega)$.

On the other hand, let the flux $\boldsymbol{\sigma} = -A\nabla u + A\mathbf{f}_2 \in L^2(\Omega)^d$, then $\nabla \cdot \boldsymbol{\sigma} = f_1 - \mathbf{b} \cdot \nabla u - cu \in L^2(\Omega)$, thus we have $\boldsymbol{\sigma} \in H(\mathrm{div}; \Omega)$. We use the fact $\nabla u = \mathbf{f}_2 - A^{-1}\boldsymbol{\sigma}$ and get a first-order system:

$$\begin{cases} \boldsymbol{\sigma} + A\nabla u = A\mathbf{f}_2 & \text{in } \Omega, \\ \nabla \cdot \boldsymbol{\sigma} + Xu = f_1 & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.3}$$

For $(\boldsymbol{\tau}, v) \in H(\mathrm{div}; \Omega) \times H_0^1(\Omega)$, define the least-squares functional for the system (4.3),

$$M(\boldsymbol{\tau}, v; f_1, \mathbf{f}_2) := \|A^{-1/2}\boldsymbol{\tau} + A^{1/2}\nabla v - A^{1/2}\mathbf{f}_2\|_0^2 + \|\nabla \cdot \boldsymbol{\tau} + Xv - f_1\|_0^2. \tag{4.4}$$

The corresponding least-squares minimization problem is:

Find $(\boldsymbol{\sigma}, u) \in H(\mathrm{div}; \Omega) \times H_0^1(\Omega)$

$$\text{s.t. } M(\boldsymbol{\sigma}, u; f_1, \mathbf{f}_2) = \inf_{(\boldsymbol{\tau}, v) \in H(\mathrm{div}; \Omega) \times H_0^1(\Omega)} M(\boldsymbol{\tau}, v; f_1, \mathbf{f}_2).$$

The Euler-Lagrange weak problem is: Find $(\boldsymbol{\sigma}, u) \in H(\mathrm{div}; \Omega) \times H_0^1(\Omega)$, such that

$$b((\boldsymbol{\sigma},u),(\boldsymbol{\tau},v)) = F(\boldsymbol{\tau},v) \quad \forall (\boldsymbol{\tau},v) \in H(\mathrm{div};\Omega)\times H_0^1(\Omega), \tag{4.5}$$

where the bilinear form a and linear form F are defined for all $(\boldsymbol{\rho},w)$ and $(\boldsymbol{\tau},v)\in H(\mathrm{div};\Omega)\times H_0^1(\Omega)$ as:

$$b((\boldsymbol{\rho},w),(\boldsymbol{\tau},v)) = (\boldsymbol{\rho}+A\nabla w, A^{-1}\boldsymbol{\tau}+\nabla v)+(\nabla\cdot\boldsymbol{\rho}+Xw,\nabla\cdot\boldsymbol{\tau}+Xv),$$

$$F(\boldsymbol{\tau},v) = (A\mathbf{f}_2, A^{-1}\boldsymbol{\tau}+\nabla v)+(f_1,\nabla\cdot\boldsymbol{\tau}+Xv).$$

Let $\mathcal{T}=\{K\}$ be a triangulation of Ω using simplicial elements. The mesh $\mathcal{T}$ is assumed to be regular. For an element $K\in\mathcal{T}$ and an integer $k\ge 0$, let $P_k(K)$ be the space of polynomials with degrees less than or equal to k. Define the finite element spaces RT_k, S_k, and $S_{k,0}$ as follows:

$$RT_k := \{\boldsymbol{\tau}\in H(\mathrm{div};\Omega):\ \boldsymbol{\tau}|_K\in P_k(K)^d + xP_k(K),\ \forall K\in\mathcal{T}\},$$

$$S_k := \{v\in C^0(\Omega):\ v|_K\in P_k(K),\ \forall K\in\mathcal{T}\} \quad\text{and}\quad S_{k,0} := S_k\cap H_0^1(\Omega).$$

Then for $k\ge 0$ being an integer, the corresponding first-order system least-squares minimization problem and finite element problem are: Find $(\boldsymbol{\sigma}_h,u_h)\in RT_k\times S_{k+1,0}$, such that

$$M(\boldsymbol{\sigma}_h,u_h;f_1,\mathbf{f}_2) = \inf_{(\boldsymbol{\tau},v)\in RT_k\times S_{k+1,0}} M(\boldsymbol{\tau},v;f_1,\mathbf{f}_2),$$

and find $(\boldsymbol{\sigma}_h,u_h)\in RT_k\times S_{k+1,0}$, such that

$$b((\boldsymbol{\sigma}_h,u_h),(\boldsymbol{\tau},v)) = F(\boldsymbol{\tau},v) \quad \forall (\boldsymbol{\tau},v)\in RT_k\times S_{k+1,0}, \tag{4.6}$$

respectively.

Since it is obvious that $M(\boldsymbol{\tau},v;0,0)=L(\boldsymbol{\tau},v;0)$, we immediately have

$$b((\boldsymbol{\tau},v),(\boldsymbol{\tau},v))\ge C|||(\boldsymbol{\tau},v)|||^2. \tag{4.7}$$

Thus the existence and uniqueness of both (4.5) and (4.6) are established.

4.1. A priori error estimate

It is easy to derive that the quasi-best approximation property holds:

$$|||(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,u-u_h)||| \le C\inf_{(\boldsymbol{\tau}_h,v_h)\in RT_k\times S_{k+1,0}} |||(\boldsymbol{\sigma}-\boldsymbol{\tau}_h,u-v_h)|||. \tag{4.8}$$

Thus, we only need the approximation properties of the discrete spaces and the regularities of the solutions to finish the a priori error estimate.

First, we discuss the local approximation properties of $S_{k+1,0}$. The space $H^{1+s}(\Omega)$, with $s>0$ for two dimensions and $s>1/2$ for three dimensions, is embedded in $C^0(\Omega)$ by Sobolev's embedding theorem. Hence, we can define the nodal interpolation of the function $v\in H^{1+s}(\Omega)$. It is proved in [29] that if $v\in H^{1+s_K}(K)$ with $s_K>0$ in two dimensions and $s_K>1/2$ in three dimensions, then for $s_K\le k+1$, the following estimate holds for the nodal interpolation I_{no}:

$$\|\nabla(v-I_{no}v)\|_{0,K}\le Ch^{s_K}|v|_{1+s_K,K} \quad \forall K\in\mathcal{T}. \tag{4.9}$$

For solutions with low regularities, the nodal interpolation is not well-defined. We can use the modified Clément interpolation [27,3] or the Scott-Zhang interpolation [45]. For an element $K\in\mathcal{T}$, let Δ_K be the collection of elements in $\mathcal{T}$ that share at least one vertex with K. Assume that $v\in H_0^1(\Omega)$ and $v|_{\Delta_K}\in H^{1+s_{\Delta_K}}(\Delta_K)$ for some $0<s_{\Delta_K}\le k+1$, and let $I_{sz}v$ be the Scott-Zhang interpolation into $S_{k+1,0}$, we have

$$\|\nabla(v-I_{sz}v)\|_{0,K}\le Ch^{s_{\Delta_K}}|v|_{1+s_{\Delta_K},\Delta_K}. \tag{4.10}$$

Define $\mathcal{T}_s$ to be the part of the mesh such that the local element-wise regularity s_K of $H^{1+s_K}(K)$ is big enough to ensure the nodal interpolation:

$$\mathcal{T}_s := \{K\in\mathcal{T}: s_K>0 \text{ for } d=2 \text{ and } s_K>1/2 \text{ for } d=3\}.$$

Combining the approximation properties of (4.9) and (4.10), we have an almost localized approximation result: assume that $u\in H_0^1(\Omega)$, $u|_K\in H^{1+s_K}(K)$ for $K\in\mathcal{T}_s$, and $u|_{\Delta_K}\in H^{1+s_{\Delta_K}}(\Delta_K)$ for $K\in\mathcal{T}\backslash\mathcal{T}_s$, where $\max_{K\in\mathcal{T}_s}\{s_K\}\le k+1$ and $\max_{K\in\mathcal{T}\backslash\mathcal{T}_s}\{s_{\Delta_K}\}\le k+1$,

$$\inf_{v\in S_{k+1,0}}\|\nabla(u-v)\|_0 \le C\Big(\sum_{K\in\mathcal{T}_s} h_K^{s_K}|u|_{1+s_K,K} + \sum_{K\in\mathcal{T}\backslash\mathcal{T}_s} h_K^{s_{\Delta_K}}|u|_{1+s_{\Delta_K},\Delta_K}\Big). \tag{4.11}$$

Assume that $\boldsymbol{\sigma}=A(\mathbf{f}_2-\nabla u)$ and $\nabla\cdot\boldsymbol{\sigma}=f_1-\mathbf{b}\cdot\nabla u-cu$ have the following local regularities, respectively:

$$\boldsymbol{\sigma}|_K\in H^{\ell_K}(K) \quad\text{and}\quad \nabla\cdot\boldsymbol{\sigma}|_K\in H^{t_K}(K) \quad K\in\mathcal{T}.$$

For a fixed $r>0$, denote by $I_{rt}: H(\mathrm{div};\Omega)\cap[H^r(\Omega)]^d\mapsto RT_k$ the standard RT interpolation operator. We have the following local approximation property: for $\boldsymbol{\tau}\in H^{\ell_K}(K)$, $1\le\ell_K\le k+1$,

$$\|\boldsymbol{\tau}-I_{rt}\boldsymbol{\tau}\|_{0,K}\le Ch_K^{\ell_K}|\boldsymbol{\tau}|_{\ell_K,K} \quad \forall K\in\mathcal{T}, \tag{4.12}$$

We also should notice that the interpolations and approximation properties are entirely local.

Denote by $Q_k: L^2(\Omega)\mapsto D_k$ the L^2-projection onto $D_k := \{v\in L^2(\Omega): v|_K\in P_k(K), K\in\mathcal{T}\}$. The following commutativity property is well-known:

$$\nabla\cdot(I_{rt}\boldsymbol{\tau}) = Q_k\nabla\cdot\boldsymbol{\tau} \qquad \forall\,\boldsymbol{\tau}\in H(\mathrm{div};\Omega)\cap H^r(\Omega)^d \ \text{ with } r>0.$$

Thus we have the following local approximation property: for $\nabla\cdot\boldsymbol{\tau}\in H^{t_K}(K)$, $1\le t_K\le k+1$,

$$\|\nabla\cdot(\boldsymbol{\tau}-I_{rt}\boldsymbol{\tau})\|_{0,K}\le Ch_K^{t_K}|\nabla\cdot\boldsymbol{\tau}|_{t_K,K} \quad \forall K\in\mathcal{T}. \tag{4.13}$$

Combining the above approximation properties and (4.8), we have the following a priori error estimate:

Theorem 4.1. ***(A priori error estimate)*** *Assume that $u\in H_0^1(\Omega)$, $u|_K\in H^{1+s_K}(K)$ for $K\in\mathcal{T}_s$, and $u|_{\Delta_K}\in H^{1+s_{\Delta_K}}(\Delta_K)$ for $K\in\mathcal{T}\backslash\mathcal{T}_s$, where $\max_{K\in\mathcal{T}_s}\{s_K\}\le k+1$ and $\max_{K\in\mathcal{T}\backslash\mathcal{T}_s}\{s_{\Delta_K}\}\le k+1$. Assume that $\boldsymbol{\sigma}|_K\in H^{\ell_K}(K)$ and $\nabla\cdot\boldsymbol{\sigma}|_K\in H^{t_K}(K)$, for $1\le\ell_K\le k+1$ and $1\le t_K\le k+1$. Then we have the following a priori error estimate:*

$$\begin{aligned}|||(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,u-u_h)||| \le C\Big(&\sum_{K\in\mathcal{T}_s} h_K^{s_K}|u|_{1+s_K,K} + \sum_{K\in\mathcal{T}\backslash\mathcal{T}_s} h_K^{s_{\Delta_K}}|u|_{1+s_{\Delta_K},\Delta_K}\\ &+\sum_{K\in\mathcal{T}}\big(h_K^{\ell_K}|\boldsymbol{\sigma}|_{\ell_K,K}+h_K^{t_K}|\nabla\cdot\boldsymbol{\sigma}|_{t_K,K}\big)\Big).\end{aligned} \tag{4.14}$$

Remark 4.2. We should notice that the smoothness s_K, ℓ_K, and t_K can be independent. Compared to the standard analysis, the localized a priori does not require global smoothnesses of $\boldsymbol{\sigma}$ and $\nabla\cdot\boldsymbol{\sigma}$, or in terms of the data, global smoothnesses of $A\mathbf{f}_2$ and f_1. We only need the local (element-wise) smoothness of such data.

The discussion of (4.11) can be applied to the conforming C^0-Lagrange approximation to the standard variational formulation too. However, the analysis is not coefficient-robust compared to the results in [14,48] for discontinuous and mixed approximations.

The a priori error estimate with local regularity is the base for adaptive finite element methods to achieve equal discretization error distribution. In a sense, the a priori results show us what an optimal mesh should be. For example, assuming $\mathcal{T}_s=\mathcal{T}$, then an error-equal-distributed mesh should have a similar size of $h_K^{s_K}|u|_{1+s_K,K}+h_K^{\ell_K}|\boldsymbol{\sigma}|_{\ell_K,K}+h_K^{t_K}|\nabla\cdot\boldsymbol{\sigma}|_{t_K,K}$ for all $K\in\mathcal{T}$.

4.2. A posteriori error estimate

Let

$$E=\boldsymbol{\sigma}-\boldsymbol{\sigma}_h \quad\text{and}\quad e=u-u_h. \tag{4.15}$$

Then by the first-order system (4.3), we have

$$M(\boldsymbol{\sigma}_h, u_h; f_1, \mathbf{f}_2) = \|A^{-1/2}(\boldsymbol{\sigma}_h + A\nabla u_h - \boldsymbol{\sigma} - A\nabla u)\|_0^2 + \|\nabla\cdot\boldsymbol{\sigma} + Xu - \nabla\cdot\boldsymbol{\sigma}_h + Xu_h\|_0^2 = \|A^{-1/2}(E + A\nabla e)\|_0^2 + \|\nabla\cdot E + Xe\|_0^2 = M(E, e; 0, \mathbf{0}).$$

Since $E \in H_N(\mathrm{div};\Omega)$ and $e \in H_0^1(\Omega)$, we have

$$C_1 |||(E,e)|||^2 \le M(\boldsymbol{\sigma}_h, u_h; f_1, \mathbf{f}_2) \le C_2 |||(E,e)|||^2. \tag{4.16}$$

Thus we can define a posteriori error estimator and local error indicator as:

$$\eta_K^2 = \|A^{-1/2}\boldsymbol{\tau} + A^{1/2}\nabla v - A^{1/2}\mathbf{f}_2\|_{0,K}^2 + \|\nabla\cdot\boldsymbol{\tau} + Xv - f_1\|_{0,K}^2,$$
$$\eta^2 = \sum_{K\in\mathcal{T}} \eta_K^2 = M(\boldsymbol{\sigma}_h, u_h; f_1, \mathbf{f}_2).$$

The equivalence (4.16) shows the reliability and efficiency of the error estimator.

By the triangle inequality, we also have the local efficiency bound,

$$\eta_K \le C(\|\nabla e\|_{0,K} + \|e\|_{0,K} + \|E\|_{0,K} + \|\nabla\cdot E\|_{0,K}).$$

References

[1] M. Ainsworth, J.T. Oden, A Posteriori Error Estimation in Finite Element Analysis, Wiley, 2000.
[2] I. Babuška, Error bound for finite element method, Numer. Math. 16 (1971) 322–333.
[3] C. Bernardi, V. Girault, A local regularization operator for triangular and quadrilateral finite elements, SIAM J. Numer. Anal. 35 (1998) 1893–1916.
[4] M. Bernkopf, J.M. Melenk, Optimal convergence rates in L^2 for a first order system least squares finite element method. Part I: homogeneous boundary conditions for a first order system least squares finite element method. Part I: homogeneous boundary conditions, arXiv:2012.12919, 2020.
[5] L. Bers, F. John, M. Schechter, Partial Differential Equations, Lecture Notes in Applied Mathematics, vol. 3A, American Mathematical Society, 1964.
[6] P.B. Bochev, M.D. Gunzburger, Finite element methods of least-squares type, SIAM Rev. 40 (1998) 789–837.
[7] P.B. Bochev, M.D. Gunzburger, Least-Squares Finite Element Methods, Applied Mathematical Sciences, vol. 166, Springer, 2009.
[8] D. Braess, Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics, Cambridge University Press, 2007.
[9] J.H. Bramble, R.D. Lazarov, J.E. Pasciak, A least-squares approach based on a discrete minus one inner product for first order systems, Math. Comput. 66 (1997) 935–955.
[10] J.H. Bramble, R.D. Lazarov, J.E. Pasciak, Least squares for second order elliptic problems, Comput. Methods Appl. Mech. Eng. 152 (1998) 195–210.
[11] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Universitext, Springer, 2011.
[12] Z. Cai, Least-Squares Method, Lecture Notes of Purdue University, 2004.
[13] Z. Cai, R. Falgout, S. Zhang, Div first-order system LL* (FOSLL*) least-squares for second-order elliptic partial differential equations, SIAM J. Numer. Anal. 53 (2015) 405–420.
[14] Z. Cai, C. He, S. Zhang, Discontinuous finite element methods for interface problems: robust a priori and a posteriori error estimates, SIAM J. Numer. Anal. 55 (2017) 400–418.
[15] Z. Cai, R. Lazarov, T. Manteuffel, S. McCormick, First order system least-squares for second order partial differential equations: Part I, SIAM J. Numer. Anal. 31 (1994) 1785–1799.
[16] Z. Cai, B. Lee, P. Wang, Least-squares methods for incompressible newtonian fluid flow: linear stationary problems, SIAM J. Numer. Anal. 42 (2004) 843–859.
[17] Z. Cai, T. Manteuffel, S.F. McCormick, First-order system least squares for second-order partial differential equations: Part ii, SIAM J. Numer. Anal. 34 (1997) 425–454.
[18] Z. Cai, G. Starke, Least-squares methods for linear elasticity, SIAM J. Numer. Anal. 42 (2004) 826–842.
[19] Z. Cai, S. Zhang, Recovery-based error estimator for interface problems: conforming linear elements, SIAM J. Numer. Anal. 47 (2009) 2132–2156.
[20] Z. Cai, S. Zhang, Flux recovery and a posteriori error estimators: conforming elements for scalar elliptic equations, SIAM J. Numer. Anal. 48 (2010) 578–602.
[21] C. Carstensen, A.K. Dond, N. Nataraj, A.K. Pani, Error analysis of nonconforming and mixed fems for second-order linear non-selfadjoint and indefinite elliptic problems, Numer. Math. 133 (2016) 557–597.
[22] C. Carstensen, N. Nataraj, A.K. Pani, Stability of mixed fems for non-selfadjoint indefinite second-order linear elliptic pdes, Numer. Math. 150 (2022) 975–992.
[23] J.H. Chaudhry, E.C. Cyr, K. Liu, T.A. Manteuffel, L.N. Olson, L. Tang, Enhancing least-squares finite element methods through a quantity-of-interest, SIAM J. Numer. Anal. 52 (2014) 3085–3105.
[24] H. Chen, G. Fu, J. Li, W. Qiu, First order least squares method with weakly imposed boundary condition for convection dominated diffusion problems, Comput. Math. Appl. 68 (2014) 1635–1652.
[25] H. Chen, W. Qiu, A first order system least squares method for the Helmholtz equation, J. Comput. Appl. Math. 309 (2017) 145–162.
[26] J. Chen, L. Li, Convergence and domain decomposition algorithm for nonconforming and mixed methods for nonselfadjoint and indefinite problems, Comput. Methods Appl. Mech. Eng. 173 (1999).
[27] P. Clément, Approximation by finite element functions using local regularization, RAIRO. Anal. Numér. 9 (1975).
[28] A. Codd, T. Manteuffel, S. McCormick, Multilevel first-order system least squares for nonlinear partial differential equations, SIAM J. Numer. Anal. 41 (2003) 2197–2209.
[29] T. Dupont, R. Scott, Polynomial approximation of functions in Sobolev spaces, Math. Comput. 34 (1980) 441–463.
[30] L.C. Evans, Partial Differential Equations, 2 ed., American Mathematical Society, 2010.
[31] T. Führer, N. Heuer, M. Karkulik, Minres for second-order pdes with singular data, arXiv:2111.00103 [math.NA], 2021.
[32] D. Gilbarg, N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Classics in Mathematics, Springer, 2001.
[33] J. Gopalakrishnan, W. Qiu, An analysis of the practical DPG method, Math. Comput. 83 (2014) 537–552.
[34] M. Innerberger, D. Praetorius, Instance-optimal goal-oriented adaptivity, Comput. Methods Appl. Math. 21 (2021) 109–126.
[35] B.-n. Jiang, The Least-Squares Finite Element Method Theory and Applications in Computational Fluid Dynamics and Electromagnetics, Scientific Computation, Springer, 1998.
[36] J. Ku, A remark on the coercivity for a first-order least-squares method, Numer. Methods Partial Differ. Equ. 23 (2007) 1577–1581.
[37] R. Lazarov, L. Tobiska, P. Vassilevski, Streamline-diffusion least-squares mixed finite element methods for convection-diffusion problems, East-West J. Numer. Math. 5 (1997) 249–264.
[38] Y. Liang, S. Zhang, Least-squares methods with nonconforming finite elements for general second-order elliptic equations, arXiv:2204.07020, 2022.
[39] Q. Liu, S. Zhang, Adaptive flux-only least-squares finite element methods for linear transport equations, J. Sci. Comput. 84 (26) (2020), https://doi.org/10.1007/s10915-020-01269-y.
[40] Q. Liu, S. Zhang, Adaptive least-squares finite element methods for linear transport equations based on an H(div) flux reformulation, Comput. Methods Appl. Mech. Eng. 366 (2020) 113041, https://doi.org/10.1016/j.cma.2020.113041.
[41] A. Pehlivanov, G. Carey, P. Vassilevski, Least-squares mixed finite element methods for non-selfadjoint elliptic problems: I. Error estimates, Numer. Math. 72 (1996).
[42] W. Qiu, S. Zhang, Adaptive first-order system least-squares finite element methods for second order elliptic equations in non-divergence form, SIAM J. Numer. Anal. 58 (2020) 3286–3308.
[43] A.H. Schatz, An observation concerning Ritz-Galerkin methods with indefinite bilinear forms, Math. Comput. 28 (1974) 959–962.
[44] A.H. Schatz, J. Wang, Some new error estimates for Ritz Galerkin methods with minimal regularity assumptions, Math. Comput. 65 (1996) 19–27.
[45] L.R. Scott, S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions, Math. Comput. 54 (1990) 483–493.
[46] R. Verfürth, A Posteriori Error Estimation Techniques for Finite Element Methods, Oxford University Press, 2013.
[47] J. Xu, L. Zikatanov, Some observations on Babuška and Brezzi theories, Numer. Math. 94 (2003) 195–202.
[48] S. Zhang, Robust and local optimal a priori error estimates for interface problems with low regularity: mixed finite element approximations, J. Sci. Comput. 84 (2020), https://doi.org/10.1007/s10915-020-01284-z.