Least-squares finite element methods

Pavel Bochev*and Max Gunzburger†

Abstract. Least-squares finite element methods are an attractive class of methods for the numerical solution of partial differential equations. They are motivated by the desire to recover, in general settings, the advantageous features of Rayleigh–Ritz methods such as the avoidance of discrete compatibility conditions and the production of symmetric and positive definite discrete systems. The methods are based on the minimization of convex functionals that are constructed from equation residuals. This paper focuses on theoretical and practical aspects of least-square finite element methods and includes discussions of what issues enter into their construction, analysis, and performance. It also includes a discussion of some open problems.

1. Introduction

Finite element methods (FEMs) for the approximate numerical solution of partial differential equations (PDEs) were first developed and analyzed for problems in linear elasticity and other settings for which solutions can be characterized as (unconstrained) minimizers of convex, quadratic functionals on infinite-dimensional Hilbert spaces [46]. A Rayleigh–Ritz approximation of such solutions is defined by minimizing the functional over a family of finite-dimensional subspaces. An FEM results when these spaces consist of piecewise polynomial functions defined with respect to a family of grids. When applied to problems such as linear elasticity or the Poisson equation, the Rayleigh–Ritz setting gives rise to FEMs with several advantageous features that led to their great success and popularity:

1. general regions and boundary conditions are relatively easy to treat in a systematic manner;

2. the conformity[1] of the finite element spaces suffices to guarantee the stability

and optimal accuracy[2] of the approximate solutions;

3. all variables can be approximated using a single type of finite element space, e.g., the same degree piecewise polynomials defined with respect to a same grid;

4. the resulting linear systems are
   a) sparse;   b) symmetric;   c) positive definite.

The success of FEMs in the Rayleigh–Ritz setting quickly led both engineers and mathematicians to apply and analyze FEMs in other settings, motivated by the fact that properties 1 and 4a are retained for all FEMs.[3] For example, *mixed* FEMs arose from minimization problems constrained by PDEs such as the Stokes problem; the Lagrange multiplier rule was applied to enforce the constraints, resulting in saddle-point problems [19]. In this setting, the only other property retained from the Rayleigh–Ritz setting is 4b. More generally, *Galerkin* FEMs can, in principle, be defined for any PDE by forcing the residual of the PDE (posed in a weak, variational formulation) to be orthogonal to the finite element subspace [3]. In this general setting, one usually loses all the features of the Rayleigh–Ritz setting other than 1 and 4a. Using the same formalisms, Galerkin FEMs were even applied to *nonlinear* problems such as the Navier–Stokes equations [34]. It is a testament to the importance of advantage 1 that despite the loss of other advantages, mixed and Galerkin FEMs are in widespread use and have also been extensively analyzed.[4]

Not surprisingly, despite the success of mixed and Galerkin FEMs in general settings, there has been substantial interest and effort devoted to developing finite element approaches that recover at least some of the advantages of the Rayleigh–Ritz setting. Notable among these efforts have been penalty and stabilized FEMs, e.g., for the Stokes problem, stabilized FEMs [4]–[5], [6], [12], [15], [29], [30], [36], [37], [44] recover advantages 2 and 3 but fail to recover advantage 4c and often lose advantage 4b.

Least-squares finite element methods (LSFEMs) can be viewed as another attempt at retaining the advantages of the Rayleigh–Ritz setting even for much more general problems. In fact, they offer the possibility of, in principle, retaining *all* of the advantages of that setting for practically any PDE problem. In §2, we show how this is possible. However, this is not the whole story. Any FEM, including an LSFEM, must also meet additional practicality criteria:

A. bases for conforming subspaces are easily constructed;

B. linear systems are easily assembled;

C. linear systems are relatively well conditioned.

---

[2]An approximate solution is referred to as being *optimally accurate* if the corresponding error is bounded by a constant times the error of the best approximation.

[3]These properties follow from the way finite element spaces are constructed, e.g., based on grids and choosing basis functions of compact support.

[4]It should be noted that in the general settings for which FEMs lose many of the advantages they possess in the Rayleigh–Ritz setting, they do not suffer from any disadvantages compared to other discretization methods such as finite difference, finite volume, and spectral methods.

In judging whether or not an LSFEM meets theses criteria, we will measure them up against Galerkin FEMs for the Poisson equation; in particular, we will ask the questions: can we use standard, piecewise polynomial spaces that are merely continuous and for which bases are easily constructed? can the assembly of the linear systems be accomplished by merely applying quadrature rules to integrals? and, are the condition numbers of the linear systems of[5] $O(h^{-2})$? Unfortunately, naively defined LSFEMs often fail to meet one or more of the practicality criteria.

LSFEMs possess two additional advantageous features that other FEMs, even in the Rayleigh–Ritz setting, do not possess. First, least-square functionals provide an easily computable residual error indicator that can be used for adapting grids. Second, the treatment of general boundary conditions, including nonhomogeneous ones, is greatly facilitated because boundary condition residuals can be incorporated into the least-square functional.

2. The most straightforward LSFEM

Let Ω denote a bounded domain in $\Re^d$, $d = 2$ or 3, with boundary Γ. Consider the problem

$$\mathcal{L}u = f \text{ in } \Omega \quad \text{and} \quad \mathcal{R}u = g \text{ on } \Gamma, \tag{1}$$

where $\mathcal{L}$ is a linear differential operator and $\mathcal{R}$ is a linear boundary operator. We assume that the problem (1) is well posed so that there exists a solution Hilbert space S, data Hilbert spaces H_Ω and H_Γ, and positive constants α_1 and α_2 such that

$$\alpha_1 \|u\|_S^2 \leq \|\mathcal{L}u\|_{H_\Omega}^2 + \|\mathcal{R}u\|_{H_\Gamma}^2 \leq \alpha_2 \|u\|_S^2 \quad \text{for all } u \in S. \tag{2}$$

Then consider the least-squares functional[6]

$$J(u; f, g) = \|\mathcal{L}u - f\|_{H_\Omega}^2 + \|\mathcal{R}u - g\|_{H_\Gamma}^2 \tag{3}$$

and the unconstrained minimization problem

$$\min_{u \in S} J(u; f, g). \tag{4}$$

Note that the functional (3) measures the residuals of the components of the system (1) using the data space norms H_Ω and H_Γ and the minimization problem (4) seeks a solution in the solution space S for which (2) is satisfied. It is clear that the problems (1) and (4) are equivalent in the sense that $u \in S$ is a solution of (4) if and only if it is also a solution, perhaps in a generalized sense, of (1).

An LSFEM can be defined by choosing a family of finite element subspaces $S^h \subset S$ parameterized by h tending to zero and then restricting the minimization problem (4)

---

[5] Usually, h is a measure of the size of the grid used in the construction of the finite element space.

[6] A least-squares functional may be viewed as an "artificial" energy that plays the same role for LSFEMs as a bona fide physically energy plays for Rayleigh–Ritz FEMs.

to the subspaces. Thus, the LSFEM approximation $u^h \in S^h$ to the solution $u \in S$ of (1) or (4) is the solution of the problem

$$\min_{u^h \in S^h} J(u^h; f, g). \tag{5}$$

The Euler–Lagrange equations corresponding to the minimization problems (4) and (5) are given by

$$\textit{seek } u \in S \quad \textit{such that} \quad B(u, v) = F(v) \quad \textit{for all } v \in S, \tag{6}$$

$$\textit{seek } u^h \in S^h \quad \textit{such that} \quad B(u^h, v^h) = F(v^h) \quad \textit{for all } v^h \in S^h, \tag{7}$$

respectively, where for all $u, v \in S$,

$$B(u, v) = (\mathcal{L}v, \mathcal{L}u)_{H_\Omega} + (\mathcal{R}v, \mathcal{R}u)_{H_\Gamma} \quad \text{and} \quad F(v) = (\mathcal{L}v, f)_{H_\Omega} + (\mathcal{R}v, g)_{H_\Gamma}. \tag{8}$$

If we choose a basis $\{U_j\}_{j=1}^J$, where $J = \dim(S^h)$, then we have that $u^h = \sum_{j=1}^J c_j U_j$ for some constants $\{c_j\}_{j=1}^J$ and then the discretized problem (7) is equivalent to the linear system

$$\mathbb{K}\mathbf{c} = \mathbf{f}, \tag{9}$$

where the elements of the matrix $\mathbb{K} \in \Re^{J \times J}$ and the vectors $\mathbf{f} \in \Re^J$ and $\mathbf{c} \in \Re^J$ are given, for $i, j = 1, \ldots, J$, by $\mathbf{c}_j = c_j$,

$$\mathbb{K}_{ij} = (\mathcal{L}U_i, \mathcal{L}U_j)_{H_\Omega} + (\mathcal{R}U_i, \mathcal{R}U_j)_{H_\Gamma}, \quad \text{and} \quad \mathbf{f}_i = (\mathcal{L}U_i, f)_{H_\Omega} + (\mathcal{R}U_i, g)_{H_\Gamma}.$$

The results of the following theorem follow directly from (2).

Theorem 2.1. *Assume that* (2) *holds and that* $S^h \subset S$. *Then,*
– the bilinear form $B(\cdot, \cdot)$ *defined in* (8) *is continuous, symmetric, and coercive;*
– the linear functional $F(\cdot)$ *defined in* (8) *is continuous;*
– the problem (6) *has a unique solution* $u \in S$ *that is also the unique solution of the minimization problem* (4);
– the problem (7) *has a unique solution* $u^h \in S^h$ *that is also the unique solution of the minimization problem* (5);
– for some constant $C > 0$, *we have that* $\|u\|_S \leq C(\|f\|_{H_\Omega} + \|g\|_{H_\Gamma})$ *and* $\|u^h\|_S \leq C(\|f\|_{H_\Omega} + \|g\|_{H_\Gamma})$;
– for some constant $C > 0$, u *and* u^h *satisfy the error estimate*

$$\|u - u^h\|_S \leq C \inf_{v^h \in S^h} \|u - v^h\|_S; \tag{10}$$

– the matrix $\mathbb{K}$ *of* (9) *is symmetric and positive definite.*

Note that it is not assumed that the system (1) is self-adjoint or positive as it would have to be in the Rayleigh–Ritz setting; it is only assumed that it is well posed. Despite the generality of the system (1), the LSFEM based on (5) recovers *all* desirable

features of FEMs in the Rayleigh–Ritz setting. Note that (10) shows that least-squares finite element approximations are optimally accurate with respect to solution norm $\| \cdot \|_S$ for which the system (1) is well posed.

In defining the least-squares principle (4), we have not restricted the spaces S and S^h to satisfy the boundary conditions. Instead, we have included the residual $\mathcal{R}u - g$ of the boundary condition in the functional $J(\cdot\,; \cdot, \cdot)$ defined in (3). Thus, we see that LSFEMs possess a desirable feature that is absent even from standard FEMs in the Rayleigh–Ritz setting: the imposition of boundary conditions can be effected through the functional and need not be imposed on the finite element spaces.[7] Notwithstanding this advantage, one can impose essential boundary conditions on the space S in which case all terms in (2)–(8) involving the boundary condition are omitted and we also set $H_\Omega = H$. Note also that since

$$J(u_h; f, g) = \|\mathcal{L}u_h - f\|^2_{H_\Omega} + \|\mathcal{R}u_h - g\|^2_{H_\Gamma}$$
$$= B(u_h, u_h) - 2F(u_h) + (f, f)_{H_\Omega} + (g, g)_{H_\Gamma},$$

the least-square functional $J(u_h; f, g)$ provides a *computable* indicator for the residual error in the LSFEM approximation u^h. Such indicators are in widespread used for grid adaption.

The problems (6) and (7) display the *normal equation* form typical of least-squares systems; see (8). It is important to note that since $\mathcal{L}$ is a differential operator, (6) involves a higher-order differential operator. We shall see that this observation has a profound effect on how practical LSFEMs are defined.

2.1. The practicality of the straightforward LSFEM. The complete recovery, in general settings, of all desirable features of the Rayleigh–Ritz setting is what makes LSFEMs intriguing and attractive. But, what about the practicality of the method defined by (5)? We explore this issue using examples.

2.1.1. An impractical application of the straightforward LSFEM. Consider the problem

$$-\Delta u = f \text{ in } \Omega \quad \text{and} \quad u = 0 \text{ on } \Gamma, \tag{11}$$

where we assume that Ω is either a convex, Lipschitz domain or that it has a smooth boundary. Of course, this is a problem which fits into the Rayleigh–Ritz framework so that there is no apparent need[8] to use any other type of FEM. However, let us proceed and use the LSFEM method anyway, and see what happens. Here we have that (2) holds with[9] $S = H^2(\Omega) \cap H^1_0(\Omega)$, $H = L^2(\Omega)$, and $\mathcal{L} = -\Delta$. We then have

---

[7]This advantage of LSFEM can be useful for imposing inhomogeneous boundary conditions, essential boundary conditions such as Dirichlet boundary conditions for second-order elliptic PDEs, and boundary conditions involving a particular component, e.g., the normal component, of a vector variable.

[8]Inhomogeneous Dirichlet boundary conditions provide a situation in which one might want to use LSFEMs even for the Poisson problem.

[9]We use standard Sobolev space notation throughout the paper. Also, in this and most of our examples, we will be imposing the boundary condition on the solutions space S.

that, for all $u, v \in H^2(\Omega) \cap H^1_0(\Omega)$,

$$J(u; f) = \|\Delta u + f\|_0^2, \quad F(v) = \int_\Omega f \Delta v \, d\Omega, \quad \text{and} \quad B(u, v) = \int_\Omega \Delta v \, \Delta u \, d\Omega.$$

Note that minimizing the least-squares functional has turned the second-order Poisson problem into a fourth-order problem.

An LSFEM is defined by choosing a subspace $S^h \subset S = H^2(\Omega) \cap H^1_0(\Omega)$ and then posing the problem (7). It is well known that in this case, the finite element space S^h has to consist of continuously differentiable functions; this requirement greatly complicates the construction of bases and the assembly of the matrix problem. Furthermore, it is also well known that the condition number of the matrix problem is $O(h^{-4})$ which should be contrasted with the $O(h^{-2})$ condition number obtained through a Rayleigh–Ritz discretization of the Poisson equation. Thus, for this problem, the straightforward LSFEM fails all three practicality tests.

Since it is also true that (2) holds with $S = H^1_0(\Omega)$ and $H = H^{-1}(\Omega)$, one could develop an LSFEM based on the functional $J(u; f) = \|\Delta u + f\|_{-1}$ and the solution space $S = H^1_0(\Omega)$. This approach would allow one to use a finite element space S^h consisting of merely continuous functions so that bases may be easily constructed. Moreover, it can be shown that because of the use of the $H^{-1}(\Omega)$ inner product, the condition number of the resulting matrix system is $O(h^{-2})$ which is the same as for a Rayleigh–Ritz discretization. However, the $H^{-1}(\Omega)$ inner product is computed by inverting the Laplacian operator which leads to the loss of property 4a and also makes the assembly of the matrix problem more difficult. So, as it stands, the straightforward LSFEM remains impractical for the second-order Poisson problem.

2.1.2. A practical application of the straightforward LSFEM.

Consider now the problem

$$-\nabla \cdot \boldsymbol{u} = f \text{ and } \nabla \times \boldsymbol{u} = \boldsymbol{g} \text{ in } \Omega \quad \text{and} \quad \boldsymbol{n} \cdot \boldsymbol{u} = 0 \text{ on } \Gamma. \tag{12}$$

Here $\boldsymbol{u} \in S = \boldsymbol{H}^1_n(\Omega) = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega) \mid \boldsymbol{n} \cdot \boldsymbol{v} = 0 \text{ on } \Gamma\}$ and $\{f, \boldsymbol{g}\} \in H = L^2_0(\Omega) \times \boldsymbol{L}^2_s(\Omega)$, where $L^2_0(\Omega) = \{f \in L^2(\Omega) \mid \int_\Omega f \, d\Omega = 0\}$, and $\boldsymbol{L}^2_s(\Omega) = \{\boldsymbol{g} \in \boldsymbol{L}^2(\Omega) \mid \nabla \cdot \boldsymbol{g} = 0 \text{ in } \Omega\}$. We then have that (2) holds so that we may define the least-squares functional

$$J(\boldsymbol{u}; f, \boldsymbol{g}) = \|\nabla \cdot \boldsymbol{u} + f\|_0^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{g}\|_0^2 \quad \text{for all } \boldsymbol{u} \in S = \boldsymbol{H}^1_n(\Omega) \tag{13}$$

that results in

$$B(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \left((\nabla \cdot \boldsymbol{u})(\nabla \cdot \boldsymbol{v}) + (\nabla \times \boldsymbol{u}) \cdot (\nabla \times \boldsymbol{v})\right) d\Omega \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in S = \boldsymbol{H}^1_n(\Omega)$$

and

$$F(\boldsymbol{v}) = \int_\Omega \left(-f \nabla \cdot \boldsymbol{v} + \boldsymbol{g} \cdot \nabla \times \boldsymbol{v}\right) d\Omega \quad \text{for all } \boldsymbol{v} \in S = \boldsymbol{H}^1_n(\Omega).$$

An LSFEM is defined by choosing a subspace $S^h \subset S = \boldsymbol{H}_n^1(\Omega)$ and then solving the problem (7).

The LSFEM based on the functional (13) not only recovers all the good properties of the Rayleigh–Ritz setting for the problem (12), but also satisfies all three practicality criteria. Since we merely require that $S^h \subset \boldsymbol{H}_n^1(\Omega)$, we can choose standard finite element spaces for which bases are easily constructed. Furthermore, since the functional (13) only involves $L^2(\Omega)$ inner products, the assembly of the matrix system is accomplished in a standard manner. Finally, it can be shown that the condition number of the matrix system is $O(h^{-2})$.

2.2. Norm-equivalence vs. practicality.
Since (2) and (3) imply that

$$\alpha_1 \|u\|_S^2 \leq J(u; 0, 0) \leq \alpha_2 \|u\|_S^2, \tag{14}$$

we refer to the functional $J(\cdot; \cdot, \cdot)$ as being *norm equivalent.* This property of the functional causes the LSFEM defined by (5) to recover all the desirable properties of the Rayleigh–Ritz setting. However, the norms that enter the definition of the functional $J(\cdot; \cdot, \cdot)$ as well as the form of the PDE system (1) can render the resulting LSFEM impractical. Thus, in order to define a practical LSFEM, one may have to define a least-squares functional that is not norm equivalent in the sense of (14). We take up this issue in §3. Here, we examine the examples of §2.1 to see what guidance they give us about what makes an LSFEM practical.

2.2.1. First-order system form of the PDEs.
Perhaps the most important observation that can be made from the examples of §2.1 is that the example of §2.1.2 involved a first-order system of PDEs and an LSFEM that allowed for the easy construction of finite element bases (because one could work with merely continuous finite element spaces) and resulted in matrix systems with relative good conditioning. As a result, all modern LSFEMs are based on first-order formulations of PDE systems. Of course, many if not most PDEs of practical interest are not usually posed as first-order systems. Thus, *the first step in defining an LSFEM should be recasting a given PDE system into a first-order system.*

Unfortunately, there is no unique way to do this. For example, the three problems

$$\left\{ \begin{array}{ll} \boldsymbol{u} + \nabla\phi = \boldsymbol{0} & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = f & \text{in } \Omega \\ \phi = 0 & \text{on } \Gamma \end{array} \right\}, \quad \left\{ \begin{array}{ll} \boldsymbol{u} + \nabla\phi = \boldsymbol{0} & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = f & \text{in } \Omega \\ \nabla \times \boldsymbol{u} = \boldsymbol{0} & \text{in } \Omega \\ \phi = 0 & \text{on } \Gamma \end{array} \right\}, \quad \left\{ \begin{array}{ll} \nabla \cdot \boldsymbol{u} = f & \text{in } \Omega \\ \nabla \times \boldsymbol{u} = \boldsymbol{0} & \text{in } \Omega \\ \boldsymbol{n} \times \boldsymbol{u} = \boldsymbol{0} & \text{on } \Gamma \end{array} \right\}$$

are all first-order systems that are equivalent to the Poisson problem (11). Each happens to be norm equivalent, but with respect to different norms. If we assume that in each case the boundary condition is imposed on the solutions space, we have that, for the three problems, the space S in (2) is respectively given by $H_0^1(\Omega) \times H(\Omega, \text{div})$, $H_0^1(\Omega) \times \boldsymbol{H}^1(\Omega)$, and $\boldsymbol{H}_\tau^1(\Omega)$, where $H(\Omega, \text{div}) = \{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega) \mid \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\}$ and $\boldsymbol{H}_\tau^1(\Omega) = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega) \mid \boldsymbol{n} \times \boldsymbol{v} = \boldsymbol{0} \text{ on } \Gamma\}$.

2.2.2. Functionals formed using L^2 norms of equation residuals. Another observation that can be gleaned from the examples of §2.1 is that if one wants to be able to assemble the matrix system using standard finite element techniques, *then one should use L^2 norms of equation residuals in the definition of the least-squares functional.* Unfortunately, it is not always the case that the resulting least-squares functional is norm equivalent. Let us explore this issue in more detail.

Consider the Stokes problem

$$-\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \quad \nabla \cdot \boldsymbol{u} = 0 \text{ in } \Omega \quad \text{and} \quad \boldsymbol{u} = \boldsymbol{0} \text{ on } \Gamma. \tag{15}$$

The most popular LSFEM for this problem is based on the first-order system

$$\nabla \times \boldsymbol{\omega} + \nabla p = \boldsymbol{f}, \quad \boldsymbol{\omega} = \nabla \times \boldsymbol{u}, \quad \nabla \cdot \boldsymbol{u} = 0 \text{ in } \Omega \quad \text{and} \quad \boldsymbol{u} = \boldsymbol{0} \text{ on } \Gamma \tag{16}$$

that is known for obvious reasons as the *velocity–vorticity–pressure formulation.* One would then be tempted to use the functional

$$J_0(\boldsymbol{u}, \boldsymbol{\omega}, p; \boldsymbol{f}) = \|\nabla \times \boldsymbol{\omega} + \nabla p - \boldsymbol{f}\|_0^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \boldsymbol{u}\|_0^2 \tag{17}$$

that involves only $L^2(\Omega)$ norms of equation residuals. Indeed, this is the most popular approach for defining LSFEM for the Stokes equations. Unfortunately, the functional (17) is not norm equivalent [13]. On the other hand, the functional

$$J_{-1}(\boldsymbol{u}, \boldsymbol{\omega}, p; \boldsymbol{f}) = \|\nabla \times \boldsymbol{\omega} + \nabla p - \boldsymbol{f}\|_{-1}^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \boldsymbol{u}\|_0^2$$

is equivalent to $\|\boldsymbol{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2$ [13]. So, on the one hand, the lack of norm equivalence for the functional $J_0(\,\cdot\,, \,\cdot\,, \,\cdot\,; \,\cdot\,)$ results in a loss of accuracy of the LSFEM approximations based on that functional. On the other hand, the appearance of the $\boldsymbol{H}^{-1}(\Omega)$ norm in the functional $J_{-1}(\,\cdot\,, \,\cdot\,, \,\cdot\,; \,\cdot\,)$ results in an impractical LSFEM because the matrix systems are not easily assembled.[10]

3. More sophisticated LSFEMs

To define the least-squares principle (4), one had to choose the pair $\{S, J(\,\cdot\,; f, g)\}$, where S denotes a solution Hilbert space and $J(\,\cdot\,; f, g)$ a functional defined over S that satisfies the norm-equivalence relation (14). We refer to the variational principle (4) as the *continuous* least-squares principle.[11] The straightforward LSFEM was defined by choosing a finite element subspace $S^h \subset S$ and then minimizing the functional $J(\,\cdot\,; f, g)$ over S^h; see (5). We refer to the straightforward LSFEM as the *conforming* LSFEM. For such LSFEMs, we obtain the error estimate (10).

---

[10]A similar dilemma arises when one imposes boundary conditions through the least-squares functional.

[11]Here, "continuous" refers to the fact that solutions of (4) are also solutions of the PDE system (1). Recall also that (14) follows from the well-posedness relation (2) for the PDE system.

Conforming LSFEMs can be generalized so that their applicability and practicality are enhanced. Here, we briefly discuss some of these generalizations. We still have in mind approximating solutions of the continuous least-squares principle (4) or what is equivalent, solutions of the PDE system (1). We again choose a finite element space S^h and a convex, quadratic functional $J_h(\cdot; f, g)$ defined over S^h. The pair $\{S^h, J_h(\cdot; f, g)\}$ gives rise to the *discrete* least-squares principle

$$\min_{u^h \in S^h} J_h(u^h; f, g). \tag{18}$$

Since we only require that the functional $J_h(\cdot; f, g)$ be defined for functions in S^h, we refer to LSFEMs constructed in this manner as *discrete* LSFEMs.

The functional $J_h(\cdot; f, g)$ is required to satisfy the following non-restrictive assumptions.

H1. There exists a *discrete energy inner product* $(\cdot, \cdot)_h : S^h \times S^h \mapsto \Re$ and a *discrete energy norm* $\| \cdot \|_h = (\cdot, \cdot)_h^{1/2}$ such that $J_h(u^h; 0, 0) = (u^h, u^h)_h = \|u^h\|_h^2$ for all $u^h \in S^h$.

H2. There exist bilinear forms $E(\cdot, \cdot)$ and $T(\cdot, \cdot)$ such that for all smooth functions $u \in S$ and all $u^h \in S^h$

$$J_h(u^h; \mathcal{L}u, \mathcal{R}u) = \|u - u^h\|_h^2 + E(u, u^h) + T(u, u). \tag{19}$$

The two assumptions are sufficient to prove the following results about solutions of (18).

Theorem 3.1. *Assume that hypotheses* H1 *and* H2 *hold for the discrete principle* $\{S^h, J_h(\cdot; f, g)\}$ *and let u denote a sufficiently smooth solution of* (1). *Then the problem* (18) *has a unique solution $u^h \in S^h$. Moreover, u^h satisfies*

$$\|u - u^h\|_h \leq \inf_{v^h \in S^h} \|u - v^h\|_h + \sup_{v^h \in S^h} \frac{E(u, v^h)}{\|v^h\|_h}. \tag{20}$$

A discrete least-squares functional $J_h(\cdot; f, g)$ will be referred to being *order r-consistent* if there exists a positive number r such that for all sufficiently smooth functions $u \in S$, the second term on the right-hand side of (20) can be bounded from above by $C(u)h^r$, where $C(u)$ is a positive number whose value may depend on u but not on h. If $J_h(\cdot; f, g)$ is order r-consistent, then (20) implies that

$$\|u - u^h\|_h \leq \inf_{v^h \in S^h} \|u - v^h\|_h + C(u)h^r. \tag{21}$$

Theorem 3.1 shows that discrete LSFEMs can work under a minimal set of assumptions. It also explains why LSFEMs tend to be much more robust than their mixed FEM counterparts; unlike the inf-sup conditions that are required for the latter type of method, defining pairs $\{S^h, J_h(\cdot; f, g)\}$ such that the assumptions H1 and H2 are satisfied is not a difficult task.

Constructing discrete least-squares functionals. Theorem 3.1 provides estimates for the error with respect to the discrete norm $\|\cdot\|_h$. Of greater interest is estimating errors using the (mesh-independent) solution norm $\|\cdot\|_S$ associated with the PDE problem (1). Since $S^h \subset S$, it is certainly true that $\|\cdot\|_S$ acts as another norm on S^h, in addition to $\|\cdot\|_h$. Thus, since S^h is finite dimensional, these two norms are comparable. However, the comparability constants may depend on h; if they do, then error estimates analogous to (20) and (21) but in terms of the norm $\|\cdot\|_S$ will surely involve constants that depend on inverse powers of h and, at the least, accuracy may be compromised. We conclude that hypotheses H1 and H2 do not sufficiently connect $J_h(\cdot\,; f, g)$ to the problem (1) for us to determine much about the properties of the error in the discrete LSFEM solution with respect to $\|\cdot\|_S$ norm. Thus, we now discuss how to construct discrete least-squares functionals so that we can get a handle on these properties.

We assume that (2) and (14) hold for the problem (1), the least-squares functional $J(\cdot\,; f, g)$, the solution space S, and the data spaces H_Ω and H_Γ. Let $\mathcal{D}_S$, $\mathcal{D}_\Omega$, and $\mathcal{D}_\Gamma$ denote *norm-generating* operators that allow us to relate the norms on S, H_Ω, and H_Γ, respectively, to[12] $L^2(\Omega)$ norms, i.e., such that, for all $u \in S$, $f \in H_\Omega$, and $g \in H_\Gamma$, $\|u\|_S = \|\mathcal{D}_S u\|_0$, $\|f\|_{H_\Omega} = \|\mathcal{D}_\Omega f\|_0$, and $\|g\|_{H_\Gamma} = \|\mathcal{D}_\Gamma g\|_{0,\Gamma}$. We then let

$$J_h(u^h; f, g) = \|\mathcal{D}_\Omega^h(\mathcal{L}^h u^h - \mathcal{Q}_\Omega^h f)\|_0^2 + \|\mathcal{D}_\Gamma^h(\mathcal{R}^h u^h - \mathcal{Q}_\Gamma^h g)\|_0^2,$$

where $\mathcal{D}_\Omega^h$, $\mathcal{D}_\Gamma^h$, $\mathcal{L}^h$, and $\mathcal{R}^h$ are approximations of the operators $\mathcal{D}_\Omega$, $\mathcal{D}_\Gamma$, $\mathcal{L}$, and $\mathcal{R}$, respectively, and $\mathcal{Q}_\Omega^h : H_\Omega \mapsto L^2(\Omega)$ and $\mathcal{Q}_\Gamma^h : H_\Gamma \mapsto L^2(\Gamma)$ are projections. It can be shown that $J_h(u^h; f, g)$ satisfies (19) with a specific form for $E(u, v^h)$.

The operators $\mathcal{L}$ and $\mathcal{R}$ define the problem (1) that is being solved so that the main objective in choosing $\mathcal{L}^h$ and $\mathcal{R}^h$ is to make $J_h(u; f, g)$ as small as possible for the exact solutions u. An appropriate choice is to use operators that will lead to truncation errors of order r in (19), i.e., $\mathcal{L}^h$ and $\mathcal{R}^h$ should be such that (21) holds. On the other hand, $\mathcal{D}_\Omega$ and $\mathcal{D}_\Gamma$ define the energy balance of (1), i.e., the proper scaling between data and solution spaces. As a result, the main objective in the choice of $\mathcal{D}_\Omega^h$ and $\mathcal{D}_\Gamma^h$ is to ensure that the scaling induced by $J_h(\cdot\,; f, g)$ is as close as possible to (2), i.e., to "bind" $\{S^h, J_h(\cdot\,; f, g)\}$ to the energy balance of $\{S, J(\cdot\,; f, g)\}$.

For *norm-equivalent* discrete least-squares principles, $J_h(\cdot, f, g)$ satisfies

$$\hat{\alpha}_1 \|u^h\|_S \leq J_h(\cdot\,; 0, 0) \leq \hat{\alpha}_2 \|u^h\|_S \quad \text{for all } u^h \in S^h.$$

If the finite element space satisfies standard inverse assumptions, minimizers of this functional satisfy the error estimate

$$\|u - u^h\|_S \leq C \left\{ \inf_{v^h \in S^h} \|u - v^h\|_S + \left(\inf_{v^h \in S^h} \|u - v^h\|_h + \sup_{v^h \in S^h} \frac{E(u, v^h)}{\|v^h\|_h} \right) \right\}.$$

---

[12] Recall from §2.2 that the use of $L^2(\Omega)$ norms in the definition of the least-squares functional is a key factor to making an LSFEM practical.

For *quasi norm-equivalent* discrete least-squares principles, $J_h(\,\cdot\,; f, g)$ satisfies

$$\hat{\alpha}_1^h \|u^h\|_S \le J_h(\,\cdot\,; 0, 0) \le \hat{\alpha}_2^h \|u^h\|_S,$$

where $\hat{\alpha}_1^h > 0$ and $\hat{\alpha}_2^h > 0$ for all $h > 0$ but may depend on h. Under additional assumptions, error estimates can also be derived in this case.

4. Compatible LSFEMs

Stable mixed finite element methods (MFEMs) for the Poisson equation[13] based on first-order formulations involving a scalar variable ϕ and a vector (or flux) variable $\boldsymbol{u}$ require the use of finite element spaces that satisfy an appropriate inf-sup condition [19], [20]. It is well known that pairs of standard, nodal-based, continuous finite element spaces fail the inf-sup condition and lead to unstable mixed methods. It is also well known that the inf-sup condition is circumvented if one uses such simple element pairs in LSFEMs based on L^2 least-squares functionals. Ever since such LSFEMs for first-order formulations of the Poisson equation were first considered in [38], this fact has been deemed as an important advantage of those methods over MFEMs. On the other hand, such LSFEMs suffer from two deficiencies. Computationally-based observations indicate that nodal-based LSFEMs do a poor job, compared to stable MFEMs, of conserving mass, i.e., of locally satisfying $\nabla \cdot \boldsymbol{u} = 0$. In addition, excepting in one special case, such methods produce suboptimally accurate (with respect to $L^2(\Omega)$ norms) flux approximations.[14]

Already in [38], optimal L^2 error estimates for LSFEMs were established for the scalar variable; however, there and in all subsequent analyses, optimal L^2 error estimates for the flux could not be obtained[15] without the addition of a "redundant" curl equation; see, e.g., [23], [24], [26], [39], [43]. Moreover, computational studies in [32] strongly suggested that optimal L^2 convergence for flux approximations may in fact be nearly impossible to obtain if one uses pairs of standard, nodal-based, continuous finite element spaces. A notable exception was a case studied in [32] for which optimal L^2 error estimates for both the scalar variable and the flux were obtained when these variables were approximated by continuous nodal spaces corresponding to a criss-cross grid. The key to proving these results was the validity of a grid decomposition property (GDP) which was established for the criss-cross grid in [33]. So far, the criss-cross grid remains the only known case of a continuous, nodal-based

---

[13]Although we consider only the Poisson problem, much of what we discuss can be easily extended to other systems of elliptic PDEs.

[14]The least-squares functionals in question are norm equivalent so that optimally accurate approximations are obtained with respect to the norms for which the equivalences hold. Here, we are interested in error estimates in weaker $L^2(\Omega)$ norms for which the norm equivalence of the least-square functional does not by itself guarantee optimal accuracy.

[15]A somewhat different situation exists for negative-norm-based LSFEMs for which it is known that the L^2 accuracy of the flux is optimal with respect to the spaces used; however, for such methods, no error bound for the divergence of the flux could be established; see [18].

finite element space for which the GDP can be verified. More importantly, it was shown in [33] (see also [17]) that the GDP is necessary and sufficient for the stability of MFEMs.

The correlation between the stability of MFEMs and the optimal accuracy of LSFEMs, established in [32], opens up the intriguing possibility that optimal L^2 accuracy for the flux may be obtainable for an LSFEM, provided that this variable is approximated using finite element spaces that are stable for an appropriate MFEM. Today, the stability of MFEMs is well understood, and many examples of stable finite element pairs are known. We will show that the use of some of these spaces in an LSFEM indeed can help improve the L^2 accuracy of flux approximations.

What we conclude is that if one gives up the use of nodal-based, continuous finite element spaces for the approximation of the flux, one can obtain optimally accurate approximations of the flux with respect to $L^2(\Omega)$ norms. While this conclusion may disappoint the adherents of equal-order implementations,[16] our results do not void LSFEMs as a viable or even preferable computational alternative to MFEMs. To the contrary, they demonstrate that *an LSFEM can be designed that combines the best computational properties of two dual MFEMs* and at the same time manages to avoid the inf-sup conditions and indefinite linear systems that make the latter more difficult to solve. Although we reach this conclusion in the specific context of MFEMs and LSFEMs for the Poisson problem, the idea of defining the latter type of method so that it inherits the best characteristics of a pair of mixed methods that are related through duality may have considerably wider application.

In the rest of this section, we focus the Poisson equation

$$-\Delta\phi = f \text{ in } \Omega, \quad \phi = 0 \text{ on } \Gamma_d, \quad \text{and} \quad \partial\phi/\partial n = 0 \text{ on } \Gamma_n, \qquad (22)$$

where Ω denotes a bounded region in $\Re^d$, $d = 2, 3$, with a Lipschitz continuous boundary Γ that consists of two disjoint parts denoted by Γ_d and Γ_n.

4.1. MFEMs for the Poisson problem. So as to provide a background for subsequent discussions concerning LSFEMs, we first consider two[17] (dual) MFEMs for the Poisson problem (22) written in the first-order form

$$\nabla \cdot \boldsymbol{u} = f, \quad \boldsymbol{u} + \nabla\phi = \boldsymbol{0} \text{ in } \Omega, \quad \phi = 0 \text{ on } \Gamma_d, \quad \boldsymbol{u} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma_n. \qquad (23)$$

4.1.1. Stable MFEMs for the Dirichlet principle. Continuous, nodal finite element spaces built from mth degree polynomials, $m \geq 1$, and whose elements satisfy the boundary condition $\phi = 0$ on Γ_n are denoted by $\mathcal{S}_m^0$. Note that $\mathcal{S}_m^0 \subset \{\psi \in H^1(\Omega) \mid \psi = 0 \text{ on } \Gamma_d\}$. We denote by $\mathcal{S}_m^1$ the space $\nabla(\mathcal{S}_m^0)$.[18]

---

[16]Recall that the ability to approximate all variables using simple nodal finite element spaces was one or the advantages of the FEMs in the Rayleigh–Ritz setting that we set out to recover using LSFEMs.

[17]Because they can be derived from two classical optimization problems, we will refer to the two methods as the discretized Dirichlet and Kelvin principles, respectively.

[18]Except for $m = 1$, $\mathcal{S}_m^1$ is not a complete $(m-1)$st degree polynomial space. However, characterizing $\mathcal{S}_m^1$ is not difficult and turns out to be unnecessary in practice.

A stable MFEM based on the Dirichlet principle is defined as follows: seek $\psi_h \in \mathcal{S}_m^0$ and $\boldsymbol{u}_h \in \mathcal{S}_m^1 = \nabla(\mathcal{S}_m^0)$ such that

$$
\begin{cases}
\displaystyle \int_\Omega \boldsymbol{u}_h \cdot \boldsymbol{v}_h \, d\Omega + \int_\Omega \nabla \phi_h \cdot \boldsymbol{v}_h \, d\Omega = 0 & \text{for all } \boldsymbol{v}_h \in \mathcal{S}_m^1, \\[2mm]
\displaystyle \int_\Omega \nabla \psi_h \cdot \boldsymbol{u}_h \, d\Omega = - \int_\Omega f \psi_h \, d\Omega & \text{for all } \psi \in \mathcal{S}_m^0.
\end{cases}
\tag{24}
$$

Note that since $\mathcal{S}_m^1 = \nabla(\mathcal{S}_m^0)$, even at the discrete level, we may eliminate the flux approximation to obtain the equivalent discrete problem for $\phi_h \in \mathcal{S}_m^0$

$$
\int_\Omega \nabla \phi_h \cdot \nabla \psi_h \, d\Omega = \int_\Omega f \psi_h \, d\Omega \quad \text{for all } \psi \in \mathcal{S}_m^0
\tag{25}
$$

that we recognize as the standard Galerkin discretization of (22). In fact, (24) and (25) are equivalent in that whenever ϕ_h is a solution of (25), then ϕ_h and $\boldsymbol{u}_h = \nabla \phi_h$ are a solution pair for (24) and conversely. In this way we see that for (24), i.e., the Dirichlet principle, the required inf-sup condition is completely benign in the sense that it can be avoided by eliminating the flux approximation $\boldsymbol{u}_h$ from (24) and solving (25) instead. The required inf-sup condition is implicitly satisfied by the pair of spaces $\mathcal{S}_m^0$ and $\mathcal{S}_m^1 = \nabla(\mathcal{S}_m^0)$. If one insists on solving (24), then one needs to explicitly produce a basis for $\mathcal{S}_m^1$; this is easily accomplished.

From either (24) or (25) one obtains, for the Dirichlet principle, that if $\phi \in H^{m+1}(\Omega) \cap H_d^1(\Omega)$, then

$$
\|\phi - \phi_h\|_0 \le h^{m+1} \|\phi\|_{m+1} \quad \text{and} \quad \|\boldsymbol{u} - \boldsymbol{u}_h\|_0 = \|\nabla(\phi - \phi_h)\|_0 \le h^m \|\phi\|_{m+1}.
\tag{26}
$$

4.1.2. Stable MFEMs for the Kelvin principle. The BDM_k and RT_k spaces on Ω are built from the individual element spaces defined with respect to a finite element $\mathcal{K}$ in a partition $\mathcal{T}_h$ of Ω

$$
\mathrm{BDM}_k(\mathcal{K}) = (P_k(\mathcal{K}))^n \quad \text{and} \quad \mathrm{RT}_k(\mathcal{K}) = (P_k(\mathcal{K}))^n + \boldsymbol{x} P_k(\mathcal{K})
$$

in a manner that ensures the continuity of the normal component across element boundaries; see [20] for details and definitions of the corresponding element degrees of freedom. Since BDM_k and RT_k both contain complete polynomials of degree k, their approximation properties in L^2 are the same. Since RT_k also contains the higher-degree polynomial component $\boldsymbol{x} P_k(\mathcal{K})$, it approximates the divergence of the flux with better accuracy than does BDM_k. Note, however, that this additional component does not help to improve the L^2 accuracy of RT_k spaces because it does not increase to $k + 1$ the order of the complete polynomials contained in RT_k.

In what follows, we will denote by $\mathcal{S}_k^2$ the RT and BDM spaces having *equal approximation orders with respect to the divergence operator*, i.e., we set $\mathcal{S}_k^2 = \{\boldsymbol{v} \in H_n(\Omega, \mathrm{div}) \mid \boldsymbol{v}|_{\mathcal{K}} \in \mathcal{S}_k^2(\mathcal{K})\}$, where $\mathcal{S}_k^2(\mathcal{K})$ is one of the finite element spaces

$RT_{k-1}(\mathcal{K})$ or $BDM_k(\mathcal{K})$ and $H_n(\Omega, \text{div}) = \{v \in H_n(\Omega, \text{div}) \mid v \cdot n = 0 \text{ on } \Gamma_n\}$. We denote by $\mathcal{S}_k^3$ the space $\nabla \cdot (\mathcal{S}_k^2)$. For characterizations of these spaces, see [20].

A stable MFEM based on the Kelvin principle is defined as follows: we seek $u_h \in \mathcal{S}_k^2$ and $\phi_h \in \mathcal{S}_k^3 = \nabla \cdot (\mathcal{S}_k^2)$ such that

$$\begin{cases} \int_\Omega u_h \cdot v_h \, d\Omega - \int_\Omega \phi_h \nabla \cdot v_h \, d\Omega = 0 & \text{for all } v_h \in \mathcal{S}_k^2, \\ \int_\Omega \psi_h \nabla \cdot u_h \, d\Omega = \int_\Omega f \psi_h \, d\Omega & \text{for all } \psi_h \in \mathcal{S}_k^3. \end{cases} \tag{27}$$

For (27), the required inf-sup condition is much more onerous than for (24) in the sense that defining a pair of stable finite element spaces for the scalar variable and the flux is not so straightforward a matter. We refer to [20] for a proof that $(\mathcal{S}_k^3, \mathcal{S}_k^2)$ is a stable pair for the mixed finite element problem (27). Moreover, one can show [20] that for any sufficiently regular exact solution of (23), one has

$$\|u - u_h\|_0 \le Ch^r \|u\|_r \quad \begin{cases} \text{for } 1 \le r \le k & \text{if } \mathcal{S}_k^2(\mathcal{K}) = RT_{k-1}, \\ \text{for } 1 \le r \le k+1 & \text{if } \mathcal{S}_k^2(\mathcal{K}) = BDM_k, \end{cases} \tag{28}$$

$$\|\nabla \cdot (u - u_h)\|_0 \le Ch^r \|\nabla \cdot u\|_r \quad \text{for } 1 \le r \le k \tag{29}$$

and

$$\|\phi - \phi_h\|_0 \le Ch^r (\|\phi\|_r + \|u\|_r) \quad \text{for } 1 \le r \le k. \tag{30}$$

It is important to note that if one uses continuous, nodal based finite element spaces for both the scalar variable and the flux, then (24) and (27) are identical discrete systems. It is well known that this leads to unstable approximations, so that one cannot use such pairs of finite element spaces in the MFEMs (24) or (27).

4.1.3. The grid decomposition property.
The following result establishes the GDP for the spaces $\mathcal{S}_k^2$ used for the discretized Kelvin principle (27);[19] for a proof, see [14].

Theorem 4.1. *For every $u_h \in \mathcal{S}_k^2$, there exist w_h, z_h in $\mathcal{S}_k^2$ such that*

$$u_h = w_h + z_h, \quad \nabla \cdot z_h = 0, \quad \int_\Omega w_h \cdot z_h \, d\Omega = 0, \quad and$$
$$\|w_h\|_0 \le C(\|\nabla \cdot u_h\|_{-1} + h\|\nabla \cdot u_h\|_0). \tag{31}$$

It was shown in [33] that the GDP, i.e., (31), along with the relation $\mathcal{S}_k^3 = \nabla \cdot (\mathcal{S}_k^2)$, are necessary and sufficient for the stability of the discretized Kelvin principle (27).

---

[19]An analogous GDP can be defined in the context of the finite element spaces $\mathcal{S}_m^0$ used for the discretized Dirichlet principle (24) but it is trivially satisfied.

4.2. LSFEMs for the Poisson problem. An LSFEM for the Poisson problem (22) can be defined based on the quadratic functional

$$J(\phi, \boldsymbol{u}; f) = \|\nabla \cdot \boldsymbol{u} - f\|_0^2 + \|\nabla\phi + \boldsymbol{u}\|_0^2 \tag{32}$$

and the least-squares principle

$$\min_{(\phi,\boldsymbol{u})\in H_d^1(\Omega)\times H_n(\Omega,\mathrm{div})} J(\phi, \boldsymbol{u}; f). \tag{33}$$

Note that we have used the first-order form (23) of the Poisson problem and that we use $L^2(\Omega)$ norms to measure the equation residuals. Also, we require the functions in the spaces $H_d^1(\Omega)$ and $H_n(\Omega, \mathrm{div})$ to satisfy the boundary conditions $\phi = 0$ on Γ_d and $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ on Γ_n, respectively. The Euler–Lagrange equations corresponding to (33) are given by: seek $\{\phi, \boldsymbol{u}\} \in H_d^1(\Omega) \times H_n(\Omega, \mathrm{div})$ such that

$$B(\{\phi, \boldsymbol{u}\}, \{\psi, \boldsymbol{v}\}) = F(\{\psi, \boldsymbol{v}\}) \quad \text{for all } \{\psi, \boldsymbol{v}\} \in H_d^1(\Omega) \times H_n(\Omega, \mathrm{div}), \tag{34}$$

where

$$B(\{\phi, \boldsymbol{u}\}, \{\psi, \boldsymbol{v}\}) = \int_\Omega (\nabla \cdot \boldsymbol{u})(\nabla \cdot \boldsymbol{v})\, d\Omega + \int_\Omega (\nabla\phi + \boldsymbol{u}) \cdot (\nabla\psi + \boldsymbol{v})\, d\Omega \tag{35}$$

and

$$F(\{\psi, \boldsymbol{v}\}) = \int_\Omega f(\nabla \cdot \boldsymbol{v})\, d\Omega. \tag{36}$$

To define an LSFEM, we restrict (33) to the conforming subspace $\mathcal{S}_m^0 \times \mathcal{S}_k^2 \subset H_d^1(\Omega) \times H_n(\Omega, \mathrm{div})$ or, equivalently, restrict (34) to those subspaces to obtain the discrete problem: seek $\{\phi_h, \boldsymbol{u}_h\} \in \mathcal{S}_m^0 \times \mathcal{S}_k^2$ such that

$$B(\{\phi_h, \boldsymbol{u}_h\}, \{\psi_h, \boldsymbol{v}_h\}) = F(\{\psi_h, \boldsymbol{v}_h\}) \quad \text{for all } \{\psi_h, \boldsymbol{v}_h\} \in \mathcal{S}_m^0 \times \mathcal{S}_k^2. \tag{37}$$

The next theorem states that the functional (32) is norm equivalent.[20] For a proof, see any of [21], [23], [24], [43].

Theorem 4.2. *There exist positive constants α_1 and α_2 such that for any $\{\phi, \boldsymbol{u}\} \in H_d^1(\Omega) \times H_n(\Omega, \mathrm{div})$,*

$$\alpha_1\big(\|\phi\|_1^2 + \|\boldsymbol{u}\|_{H(\Omega,\mathrm{div})}^2\big) \le J(\phi, \boldsymbol{u}; 0) \le \alpha_2\big(\|\phi\|_1^2 + \|\boldsymbol{u}\|_{H(\Omega,\mathrm{div})}^2\big). \tag{38}$$

Thus, the LSFEM defined through (37) is an example of an LSFEM that recovers all the desirable properties of the Rayleigh–Ritz setting, except that by using the finite element spaces $\mathcal{S}_m^0$ and $\mathcal{S}_k^2$, we have forced ourselves to not use continuous, nodal-based finite element spaces for the flux approximation.[21] Because we are using finite element spaces that are compatible for the MFEMs (24) and (27), we refer to the LSFEM defined by (37) as a *compatible* LSFEM.

---

[20]In the theorem, we have that $\|\boldsymbol{u}\|_{H(\Omega,\mathrm{div})} = (\|\boldsymbol{u}\|_0^2 + \|\nabla \cdot \boldsymbol{u}\|_0^2)^{1/2}$.

[21]We could, of course, use such spaces for the flux approximation, but, as indicated previously, we would then not be able to obtain optimal error estimates with respect to $L^2(\Omega)$ norms.

4.2.1. Error estimates in $H^1(\Omega) \times H(\Omega, \text{div})$. We now review the convergence properties of LSFEMs for the Poisson equation with respect to the $H^1(\Omega) \times H(\Omega, \text{div})$ norm. For a proof of the following theorem, see [14].

Theorem 4.3. *Assume that the solution $\{\phi, \boldsymbol{u}\}$ of (34) satisfies $\{\phi, \boldsymbol{u}\} \in H_d^1(\Omega) \cap H^{m+1}(\Omega) \times H_n(\Omega, \text{div}) \cap \boldsymbol{H}^{k+1}(\Omega)$ for some integers $k, m \geq 1$. Let $\{\phi_h, \boldsymbol{u}_h\} \in \mathcal{S}_m^0 \times \mathcal{S}_k^2$ be the solution of the least-squares finite element problem (37). Then there exists a constant $C > 0$ such that*

$$\|\phi - \phi_h\|_1 + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{H(\Omega, \text{div})} \leq C\big(h^k \|\boldsymbol{u}\|_{k+1} + h^m \|\phi\|_{m+1}\big). \qquad (39)$$

The error estimate (39) remains valid if $\boldsymbol{u}$ is approximated in the continuous, nodal-based finite element space $(P_k(\Omega))^n$.

Theorem 4.3 shows that the errors in $\boldsymbol{u}_h$ and ϕ_h are equilibrated when $k = m$ and that $(\mathcal{S}_k^0, \mathcal{S}_k^2)$ has the same asymptotic accuracy in the norm of $H^1(\Omega) \times H(\Omega, \text{div})$ as the C^0 pair $(\mathcal{S}_k^0, (P_k)^n)$. For this reason, in the implementation of the LSFEM, one usually chooses the nodal-based pair $(\mathcal{S}_k^0, (P_k)^n)$ because it is easier to implement. Indeed, the ability to use equal-order interpolation has been often cited as a primary reason for choosing to use LSFEMs. Nevertheless, the pair is not flawless because optimal L^2 norm errors for the flux approximation have proven impossible to obtain without using the very restrictive criss-cross grid or augmenting (23) with an additional redundant curl constraint equation.[22] Also, as we have already mentioned, numerical studies in [32] indicate that the L^2 convergence of the flux is indeed suboptimal with such finite element spaces.

We will see that if the nodal approximation of the flux is replaced by an approximation in $\mathcal{S}_k^2$, it may be possible to recover optimal L^2 convergence rates without adding the curl constraint. As in [32], the key to this is the GDP.

4.2.2. Error estimates in L^2. We assume that the solution of the problem

$$-\Delta \psi = \eta \text{ in } \Omega, \quad \psi = 0 \text{ on } \Gamma_d, \quad \frac{\partial \psi}{\partial n} = 0 \text{ on } \Gamma_n$$

satisfies the regularity estimate $\|\psi\|_{s+2} \leq C\|\eta\|_s$ for $s = 0, 1$ and for all $\eta \in H^s(\Omega)$. This is needed since L^2 error estimates are based on duality arguments.

L^2 error estimates for the scalar variable.

Theorem 4.4. *Assume that the regularity assumption is satisfied, and assume that the solution $(\phi, \boldsymbol{u})$ of (34) satisfies $(\phi, \boldsymbol{u}) \in H_d^1(\Omega) \cap H^{m+1}(\Omega) \times H_n(\Omega, \text{div}) \cap \boldsymbol{H}^{k+1}(\Omega)$*

---

[22]The redundant curl constraint $\nabla \times \boldsymbol{u} = \boldsymbol{0}$, first introduced in the least-squares finite element setting in [26] and subsequently utilized by many others (see, e.g., [21], [23], [24], [39]), renders the least-squares functional norm-equivalent with respect to the $H^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ norm but, in some situations, may unduly restrict the range of the data and should be avoided.

for some integers $k, m \geq 1$. Let $(\phi_h, \boldsymbol{u}_h) \in \mathcal{S}_m^0 \times \mathcal{S}_k^2$ be the solution of the least-squares finite element problem (37). *Then there exists a constant $C > 0$ such that*
$$\|\phi - \phi_h\|_0 \leq C(h^{k+1}\|\boldsymbol{u}\|_{k+1} + h^{m+1}\|\phi\|_{m+1}).$$

For a proof of this theorem, see [14]. The optimal L^2 error bound of Theorem 4.4 for the scalar variable does not require that the finite element space for flux approximations satisfy (31), i.e., the GDP. Thus, it remains valid even when continuous, nodal-based finite element spaces are used for the flux approximations, a result first shown in [38]. On the other hand, we will see that the GDP is needed if one wants to improve the L^2 accuracy of the flux.

L^2 error estimate for the flux.   The L^2 error estimates for approximations to the flux depend on whether $\mathcal{S}_k^2$ represents the RT_{k-1} or the BDM_k family. To this end, we have the following result whose proof may be found in [14].

Theorem 4.5. *Assume that the hypotheses of Theorem* 4.4 *hold with $k = m = r$. Then there exists a constant $C > 0$ such that*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \leq C \begin{cases} h^r(\|\boldsymbol{u}\|_{r+1} + \|\phi\|_{r+1}) & \textit{if } \mathcal{S}_r^2(\Omega) = \mathrm{RT}_{r-1}, \\ h^{r+1}(\|\boldsymbol{u}\|_{r+1} + \|\phi\|_{r+1}) & \textit{if } \mathcal{S}_r^2(\Omega) = \mathrm{BDM}_r. \end{cases} \qquad (40)$$

Consider, for example, the lowest-order case for which $r = 1$, $\mathcal{S}_1^0(\Omega) = P_1$, and $\mathcal{S}_1^2(\Omega)$ is either RT_0 or BDM_1. If the least-squares finite element method is implemented with RT_0 elements, (40) specializes to

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \leq h(\|\boldsymbol{u}\|_2 + \|\phi\|_2).$$

If instead we use BDM_1 elements, we then obtain the improved error bound

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_0 \leq h^2(\|\boldsymbol{u}\|_2 + \|\phi\|_2).$$

4.3. Interpretation of results and mass conservation. We have seen that an LS-FEM method implemented using equal-order, continuous, nodal-based finite element spaces approximates the scalar variable with the same accuracy (with respect to both $H^1(\Omega)$ and $L^2(\Omega)$ norms) as the Galerkin method (25) (or, equivalently, the mixed method (24) for the Dirichlet principle. However, the approximation properties of the Kelvin principle (27) are only partially inherited in the sense that the accuracy in the approximation to the divergence of the flux is recovered, but the accuracy in the flux approximation itself may be of one order less. This should not be too much of a surprise because continuous, nodal-based finite elements provide stable discretization only for the Dirichlet principle (with the exception of the criss-cross grid; see [32]). While least-squares minimization is stable enough to allow for the approximation of scalar variables and the flux by equal-order, continuous, nodal-based finite element

spaces, it cannot completely recover from the fact that such spaces are unstable for the Kelvin principle.

The key observation from §4.2.2 is that an LSFEM can inherit the best properties of *both* the discretized Dirichlet principle (24) and Kelvin principle (27), provided the scalar variable and the flux are approximated by finite element spaces that are stable with respect to these two principles, respectively. Then least-squares finite element solutions recover the accuracy of the Dirichlet principle for the scalar variable and the accuracy of the Kelvin principle for the flux. In a way, we see that, implemented in this particular manner, the LSFEM represents a balanced mixture of the two principles. In [16], an explanation of this observation using the apparatus of differential form calculus is provided as are the results of several illustrative computational experiments.

Unlike LSFEMs based on the use of continuous, nodal-based finite element spaces for all variables, it can be shown that through a simple local post-processing procedure, the compatible LSFEM inherits the local mass conservation properties of the discretized Kelvin principle (27); see [16] for details.

5. Alternative LSFEMs

The LSFEMs considered so far follow variants of the template established in §2: first, spaces S, H_Ω, and H_Γ that verify (2) are determined, then a least-squares functional (3) is defined by measuring equation residuals in the norms of H_Ω and H_Γ and, finally, an LSFEM is obtained by minimizing (3) over a finite-dimensional subspace S^h of S. Here, we provide examples of methods that, while still relying on least-squares notions, deviate in more significant ways from that template.

5.1. Collocation LSFEMs. The least-squares optimization steps (3) and (4) precede the discretization step (5). In the broadest sense, *collocation* LSFEM (CLSFEM) are methods [25], [31], [41] that reverse the order of these two steps. They are also known as *point least-squares* or *overdetermined collocation* methods.

Let $\{U_j(\boldsymbol{x})\}_{j=1}^J$ denote a basis for a finite element space. We seek an approximate solution of (1) of the form $u(\boldsymbol{x}) \approx \hat{u}_h(\boldsymbol{x}) = \sum_{j=1}^J c_j U_j(\boldsymbol{x})$, where $\boldsymbol{c} = (c_1, c_2, \ldots, c_J)$ is a vector of unknown coefficients. Then *collocation points* $\{\boldsymbol{x}_i\}_{i=1}^{M_1} \subset \Omega$ and $\{\boldsymbol{x}_i\}_{i=M_1+1}^M \subset \Gamma$ are chosen in such a way that the corresponding point residuals $\mathcal{L}\hat{u}_h(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)$ and $\mathcal{R}\hat{u}_h(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)$ are well defined. Then a CLSFEM is defined by minimizing, over $\boldsymbol{c} \in \Re^J$, the discrete functional

$$J_c(\boldsymbol{c}; f, g) = \sum_{i=1}^{M_1} \alpha_i \big(\mathcal{L}\hat{u}_h(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\big)^2 + \sum_{i=M_1+1}^M \beta_i \big(\mathcal{R}\hat{u}_h(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\big)^2.$$

The weights α_i and β_i can be used to adjust the relative importance of the terms in the functional. The necessary condition for the minimization of $J_c(\,\cdot\,; f, g)$ gives rise

to an $M \times J$ linear system $\mathbb{A}c = b$. If $M = J$, then the method reduces to a standard collocation method. If $M > J$, the solution c is obtained in a least-squares sense by solving the normal equations $\mathbb{A}^T \mathbb{A}c = \mathbb{A}^T b$. If the collocation points and weights correspond to a quadrature rule, then the CLSFEM is equivalent to an LSFEM in which integrals are approximated by a quadrature rule.

Since only a finite set of collocation points belonging to the domain $\overline{\Omega}$ need be specified, collocation LSFEMs are attractive for problems posed on irregularly shaped domains; see [41]. On the other hand, since the normal equations tend to become ill-conditioned, such methods require additional techniques such as scaling or orthonormalization in order to obtain a reliable solution; see [31].

5.2. Discontinuous LSFEMs. The LSFEMs of §2, 3, and 4 are defined using a conforming finite element subspace S^h of the solution space S. Discontinuous LSFEMs (DLSFEMs) are an alternative approach that use finite element subspaces of $L^2(\Omega)$ that consist of piecewise polynomial functions that are not constrained by inter-element continuity requirements. The degrees of freedom on each element can be chosen independently of each other and the elements can have hanging nodes. These features offer great flexibility in implementing adaptive methods because first, resolution on each element can be adjusted as needed and second, new elements can be added by simple subdivisions of existing elements.

In general, the least-squares problem (4) cannot be restricted to a discontinuous space S^h because it is not a proper subspace of S. To take advantage of discontinuous spaces, it is necessary to modify (3) so that it is well defined on the "broken" (with respect to a partition $\mathcal{T}_h$ of the domain Ω) data space $S = \{u \in L^2(\Omega) \,|\, u \in S(\mathcal{K}) \text{ for all } \mathcal{K} \in \mathcal{T}_h\}$. The first DLSFEMs appeared in [2], [22] as least-squares formulations for interface and transmission problems for the Poisson equation. We follow [22], where a DLSFEM is developed for the problem

$$\begin{cases} \nabla \cdot (a_i u_i) = f_i & \text{and} \quad u_i + \nabla \phi_i = 0 \quad \text{in } \Omega_i, \quad i = 1, 2, \\ \phi_i = 0 \text{ on } \Gamma_{i,d} & \text{and} \quad u_i \cdot n_i = 0 \quad \text{on } \Gamma_{i,n}, \quad i = 1, 2, \\ \phi_1 = \phi_2 & \text{and} \quad a_1 u_1 \cdot n_1 + a_2 u_2 \cdot n_2 = 0 \quad \text{on } \Gamma_{12} \end{cases} \tag{41}$$

that is a first-order formulation of a transmission problem for the Poisson equation.[23] Here, Ω_1 and Ω_2 are two[24] open subsets of Ω such that $\overline{\Omega}_1 \cup \overline{\Omega}_2 = \overline{\Omega}$ and $\Omega_1 \cap \Omega_2 = \emptyset$. The set $\Gamma_{12} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ is the *interface* between the two subdomains and $\Gamma_{i,d} = \Gamma_d \cap \overline{\Omega}_i$ and $\Gamma_{i,n} = \Gamma_n \cap \overline{\Omega}_i$, $i = 1, 2$.

In the conforming case, an LSFEM for the Poisson equation was defined by using the functional (32) and conforming subspaces of the solution space $S = H^1_d(\Omega) \times H_n(\Omega, \text{div})$. For the problem (41), we instead use the "broken" (with respect to the

---

[23]The functions a_1 and a_2 denote a "media property" that is discontinuous across Γ_{12}.
[24]The generalization to more than two subdomains is straightforward.

partition $\{\Omega_1, \Omega_2\}$) solution space $\boldsymbol{S} = \boldsymbol{H}_d^1(\Omega) \times \boldsymbol{H}_n(\Omega, \mathrm{div})$, where

$\boldsymbol{H}_d^1(\Omega) = \{\widetilde{\phi} = \{\phi_1, \phi_2\} \mid \phi_i \in H_d^1(\Omega_i),\ i = 1, 2\}$ for the scalar variable,
$\boldsymbol{H}_n(\Omega, \mathrm{div}) = \{\widetilde{\boldsymbol{u}} = \{\boldsymbol{u}_1, \boldsymbol{u}_2\} \mid \boldsymbol{u}_i \in H_n(\Omega_i, \mathrm{div}),\ i = 1, 2\}$ for the flux.

To define a DLSFEM, we also need to replace (32) by a least-squares functional that can be minimized over $\boldsymbol{S}$. Of course, we also want a functional whose minimizer is a solution of (41). A functional with the desired properties is given by (see [22])

$$J_{12}(\widetilde{\phi}, \widetilde{\boldsymbol{u}}; f_1, f_2) = \sum_{i=1}^{2} \left(\|\nabla \cdot (a_i \boldsymbol{u}_i) - f_i\|_{0,\Omega_i^h}^2 + \|\boldsymbol{u}_i + \nabla \phi_i\|_{0,\Omega_i^h}^2 \right) \tag{42}$$
$$+ |\phi_1 - \phi_2\|_{1/2,\Gamma_{12}}^2 + \|a_1 \boldsymbol{u}_1 \cdot \boldsymbol{n}_1 + a_2 \boldsymbol{u}_2 \cdot \boldsymbol{n}_2\|_{-1/2,\Gamma_{12}}^2.$$

Interface terms in (42) are treated in exactly the same way as one would impose weak Dirichlet and Neumann conditions, respectively. To obtain a practical method, they are replaced by weighted L^2 norms on Γ_{12}. Choosing $\boldsymbol{S}^h \subset \boldsymbol{S}$ completes the formulation of the DLSFEM; see [22] for further details.

     The Treffetz element least-squares method [42], [45] can be viewed as a variant of the DLSFEM. The term "Treffetz elements" usually refers to methods that use approximation spaces consisting of piecewise analytic solutions of the PDE. Such spaces provide highly accurate approximations of the broken solution space $\boldsymbol{S}$ so that they also require functionals that are well-posed with respect to that space. Given a Treffetz element space, it is a trivial matter to use (42) to define a DLSFEM; see [42], [45] for further details.

6. Open problems in LSFEM

We close with a brief discussion of some of the open problems that exist in the theory and application of LSFEMs.

6.1. Hyperbolic PDEs. Recovery of the Rayleigh–Ritz properties by LSFEMs relies on the existence of Hilbert spaces that validate the bounds (2) for (1). Such bounds are natural for elliptic PDEs and can be derived for any such PDE by using the Agmon–Douglis–Nirenberg theory [1]. On the other hand, for *hyperbolic* PDEs such bounds are not so natural, partly because they admit data in L^p spaces and their solutions may have contact discontinuities and shock waves.

     Recall that (7) can be viewed as a Galerkin method applied to a higher-order PDE. As a result, LSFEMs for hyperbolic equations designed using a Hilbert space setting are equivalent to a Galerkin discretization of a *degenerate* elliptic PDE. The result is an LSFEM that will have excellent stability properties but which will smear shocks and discontinuities; see [11] for numerical examples.

To illustrate some of the pitfalls that can be encountered with hyperbolic PDEs, it suffices to consider the simple linear convection-reaction problem

$$\nabla \cdot (\boldsymbol{b}u) + cu = f \text{ in } \Omega \quad \text{and} \quad u = g \text{ on } \Gamma_-, \tag{43}$$

where $\boldsymbol{b}$ is a given convection vector, $c(\boldsymbol{x})$ is a bounded measurable function on Ω, and $\Gamma_- = \{\boldsymbol{x} \in \Gamma \mid \boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{b}(\boldsymbol{x}) < 0\}$ is the inflow part of the boundary Γ. A straightforward $L^2(\Omega)$ norm-based least-squares principle for (43) is defined by minimizing the functional

$$J(v; f, g) = \|\nabla \cdot (\boldsymbol{b}v) + cv - f\|_0^2 + \|v - g\|_{0,\Gamma_-}^2 \tag{44}$$

over the Hilbert space $S = \{u \in L^2(\Omega) \mid \mathcal{L}u = \nabla \cdot (\boldsymbol{b}v) + cv \in L^2(\Omega)\}$. Then the following theorem can be obtained [10].

Theorem 6.1. *Assume that Γ_- is non-characteristic and $c + \frac{1}{2}\nabla \cdot \boldsymbol{b} \geq \sigma > 0$ for some constant σ. Then $J(v; 0, 0) = \|\nabla \cdot (\boldsymbol{b}v) + cv\|_0^2 + \|v\|_{0,\Gamma_-}^2$ is equivalent to the graph norm $\|v\|_S^2 = \|v\|_0^2 + \|\mathcal{L}v\|_0^2$. For every $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_-)$, (44) has a unique minimizer $u \in S$ and for that u we have that $J(u; 0, 0) \leq \|f\|_0 + \|g\|_{0,\Gamma_-}^2$.*

This theorem shows that if the data belongs to L^2, all the prerequisites needed to define an LSFEM are fulfilled. We can proceed as in §2 and define a method in the most straightforward way by restricting the Euler–Lagrange equation corresponding to the minimization of (44) to a finite dimensional subspace $S^h \subset S$.

However, the convection-reaction problem (43) is meaningful even if the data[25] f belongs only to the Banach space $L^1(\Omega)$. In this case, proper solution and data spaces for (43) are given by $S = \{v \in L^1(\Omega) \mid \nabla \cdot (\boldsymbol{b}v) \in L^1(\Omega)\}$ and $H = L^1(\Omega)$, respectively. One can show [35] that $\mathcal{L}$ is an isomorphism $S \mapsto H$ and so, instead of (2), we have a similar bound but in Banach spaces: $\alpha_1 \|u\|_S \leq \|\mathcal{L}u\|_H$ for all $u \in S$.

Now, consider the unconstrained minimization problem associated with the spaces S and H:

$$\min_{u \in S} J_1(u; f), \quad \text{where } J_1(u; f) = \|\mathcal{L}u - f\|_{L^1(\Omega)} = \int_{\Omega} |\mathcal{L}u - f| \, d\Omega. \tag{45}$$

For our model equation (43), this is the "correct" minimization problem that, restricted to $S^h \subset S$, will have solutions that do not smear discontinuities. This fact has been recognized independently in [40] and more recently in [35]. In [35], it is also shown that under some reasonable assumptions on S^h, the discrete problem

$$\min_{u^h \in S^h} J_1(u^h; f) \tag{46}$$

has at least one global minimizer, no local minimizers, and a solution that satisfies the stability bound $\|u^h\|_S \leq C\|f\|_H$.

---

[25]We assume now that $g = 0$.

We can view (45) as yet another example of the conflict between practicality and optimality. In this case, however, the practicality issue is much more severe because (45) is not differentiable, we cannot write a first-order optimality condition, and the discrete problem (46) does not give rise to a matrix problem. This is the chief reason that so far there are only two examples [35], [40] of FEMs for (43) based on the L^1 optimization problem (45). In [35], the minimizer of (46) is approximated by solving a sequence of regularized L^1 optimization problems that are differentiable. The method of [40] uses a sequence of more conventional L^2 least-squares approaches, but defined using an adaptively weighted L^2 inner product. The weights are used to weaken contributions to the least-squares functional from elements that contain solution discontinuities.

At this point, there is very limited experience with solving hyperbolic PDEs by minimizing functionals over Banach spaces. For problems with non-smooth data, computational experiments with the methods of [35] and [40] show that they are superior to LSFEMs defined through the minimization of (44); most notable is their ability to provide sharp discontinuity profiles without over- and under-shooting. A series of experiments in [35] also points strongly towards a possibility that the numerical solutions actually obey a maximum principle on general unstructured grids and that the L^1-based algorithm seems to be able to select viscosity solutions. However, at present, there are no mathematical confirmations of these facts, nor is it known whether such algorithms for hyperbolic conservation laws are able to provide accurate shock positions and speeds.

Despite the promise of L^1 optimization techniques, the state of LSFEMs for hyperbolic problems is far from satisfactory. Straightforward L^2 norm-based LSFEMs are clearly not the most appropriate as they are based on the "wrong" stability estimate for the problem. L^1 norm-based techniques give far better results but are more complex and, in the case of [35], require the solution of nonlinear optimization problems. Thus, the jury is still out on whether or not it is possible to define a simple, robust, and efficient LSFEMs for hyperbolic problems that will be competitive with specially designed, upwind schemes employing flux limiters.

6.2. Mass conservation. In §4 it was shown that LSFEMs for the Poisson equation can be implemented in a way that allows them to inherit the best computational properties of MFEMs for the same problem. In particular, it is possible to define an LSFEM for (23) so that the approximation *locally conserves mass*.

Currently, the methods in §4 are the only such example. Achieving local mass conservation in LSFEMs for incompressible, viscous flows remains an important open problem. All existing LSFEMs for incompressible, viscous flows conserve mass only approximately so that $\|\nabla \cdot \boldsymbol{u}^h\|_0 = O(h^r)$, where r is the approximation order of the finite element space. For low-order elements, which are among the most popular and easy to use elements, LSFEMs have experienced severe problems with mass conservation. For LSFEMs based on the velocity–vorticity–pressure system (16), these problems were first identified in [27] where also a solution was proposed that

combines least-squares principles *and* Lagrange multipliers to achieve element-wise mass conservation. Then the resulting *restricted* LSFEM treats the continuity equation $\nabla \cdot \boldsymbol{u} = 0$ as an additional constraint that is enforced on each element by a Lagrange multiplier. The method achieves remarkable local conservation but compromises the main motivation underlying LSFEMs: to recover a Rayleigh–Ritz setting for the PDE. In particular, property 4c does not hold.

An alternative to exact local conservation is an LSFEM with *enhanced* total mass conservation. This can be effected by increasing the importance of the continuity residual by using weights. A weighted LSFEM for (16) using the functional

$$J_W(\boldsymbol{\omega}, p, \boldsymbol{u}) = \|\nabla \times \boldsymbol{\omega} + \nabla p - \boldsymbol{f}\|_0^2 + \sum_{\mathcal{K} \in \mathcal{T}_h} h_{\mathcal{K}}^2 \left(W \|\nabla \cdot \boldsymbol{u}\|_{0,\mathcal{K}}^2 + \|\nabla \times \boldsymbol{u} - \boldsymbol{\omega}\|_{0,\mathcal{K}}^2 \right)$$

was studied in [28] where numerical studies showed that fairly a small weight, e.g., $W = 10$, helps to significantly improve total mass conservation.

Thus, for the Stokes problem, at present there are methods that either recover local mass conservation but forfeit some important advantages of the Rayleigh–Ritz settings or retain all those advantages but can at best provide improved global conservation. It is of interest to explore whether or not the ideas of §4 can be extended to develop compatible LSFEMs for viscous flows that retain all the Rayleigh–Ritz advantages and at the same time locally conserve mass.

6.3. LSFEMs for nonlinear problems.

Consider the nonlinear version of (1)

$$\mathcal{L}u + \mathcal{G}(u) = f \ \text{ in } \Omega \quad \text{and} \quad \mathcal{R}u = g \ \text{ on } \Gamma, \tag{47}$$

where $\mathcal{G}(u)$ is a nonlinear term. Formally, a least-squares principle for (1) can be easily extended to handle (47) by modifying (4) and (3) to

$$\min_{u \in S} J_{\mathcal{G}}(u; f, g), \quad \text{where} \quad J_{\mathcal{G}}(u; f, g) = \|\mathcal{L}u + \mathcal{G}(u) - f\|_{H_\Omega}^2 + \|\mathcal{R}u - g\|_{H_\Gamma}^2 \tag{48}$$

and then define an LSFEM by restricting (48) to a family $S^h \subset S$. While the extension of LSFEMs to (47) is trivial, its analysis is not and remains one of the open problems in LSFEMs. Compared with the well-developed mathematical theory for linear elliptic problems [2], [13], [18], [21], [23], [24], [26], [32], [38], analyses of LSFEMs for nonlinear problems are mostly confined to the Navier–Stokes equations [7]–[9].

It can be shown that the Euler–Lagrange equation associated with the least-squares principle (48) for the Navier–Stokes equations has the abstract form

$$F(\lambda, U) \equiv U + T \cdot G(\lambda, U) = 0, \tag{49}$$

where λ is the Reynolds number, T is a least-squares solution operator for the associated Stokes problem, and G is a nonlinear operator. As a result, the corresponding discrete nonlinear problem has the same abstract form

$$F^h(\lambda, U^h) \equiv U^h + T^h \cdot G(\lambda, U^h) = 0, \tag{50}$$

where T^h is an approximation of T. The importance of (50) is signified by the fact that discretization in (50) is introduced solely by means of an approximation to the *linear* operator T in (49). As a result, under some assumptions, one can show that the error in the nonlinear approximation defined by (50) is of the same order as the error in the least-squares solution of the linear Stokes problem.

One of the obstacles in extending this approach to a broader class of nonlinear problems is that after the application of a least-squares principle, the (differentiation) order of the nonlinear term may change.

References

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

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

[3] Babuška, I., Aziz, K., Survey lectures on the mathematical foundations of the finite element method. In *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (ed. by K. Aziz and I. Babuška), Academic Press, New York 1972.

[4] Barth, T., Bochev, P., Gunzburger, M., Shadid, J., A Taxonomy of consistently stabilized finite element methods for the Stokes problem. *SIAM J. Sci. Comput.* **25** (2004), 1585–1607.

[5] Becker, R., Braack, M., A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo* **38** (2001), 173–199.

[6] Blasco, J., Codina, R., Stabilized finite element method for the transient Navier-Stokes equations based on a pressure gradient projection. *Comput. Methods Appl. Mech. Engrg.* **182** (2000), 277–300.

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

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

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

[10] Bochev, P., Choi, J., Improved least-squares error estimates for scalar hyperbolic problems. *Comput. Meth. Appl. Math.* **1** (2001), 115–124.

[11] Bochev, P., Choi, J., A comparative numerical study of least-squares, SUPG and Galerkin methods for convection problems. *Int. J. Comput. Fluid Dyn.* **15** (2001), 127–146.

[12] Bochev, P., Dohrmann, C., Gunzburger, M., Stabilization of low-order mixed finite elements for the Stokes equations. *SIAM J. Numer. Anal.* **44** (1) (2006), 82–101.

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

[14] Bochev, P., Gunzburger, M., On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles. *SIAM J. Numer. Anal.* **43** (2005), 340–362.

[15] Bochev, P., Gunzburger, M., An absolutely stable pressure-Poisson stabilized method for the Stokes equations. *SIAM J. Numer. Anal.* **42** (2005), 1189–1207.

[16] Bochev, P., Gunzburger, M., Compatible least-squares finite element methods. *SIAM J. Numer. Anal.*, to appear.

[17] Boffi, D., Brezzi, F., Gastaldi, L., On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form. *Math. Comp.* **69** (2000), 121–140.

[18] Bramble, J., Lazarov, R., Pasciak, J., A least squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.* **66** (1997), 935–955.

[19] Brezzi, F., On existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers. *RAIRO Model. Math. Anal. Numer.* **21** (1974) 129–151.

[20] Brezzi, F., Fortin, M., *Mixed and Hybrid Finite Element Methods*. Springer Ser. Comput. Math. 15, Springer-Verlag, New York 1991.

[21] Cai, Z., Lazarov, R., Manteuffel, T., McCormick, S., First-order system least squares for second-order partial differential equations: Part I. *SIAM J. Numer. Anal.* **31** (1994), 1785–1799.

[22] Cao, Y., Gunzburger, M., Least-squares finite element approximations to solutions of interface problems. *SIAM J. Numer. Anal.* **35** (1998), 393-405.

[23] Carey, G., Pehlivanov, A., Error estimates for least-squares mixed finite elements. *Math. Model Numer. Anal.* **28** (1994), 499–516.

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

[25] Chang, C.-L., Gunzburger, M., A subdomain Galerkin/least squares method for first order elliptic systems in the plane. *SIAM J. Numer. Anal.* **27** (1990), 1197–1211.

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

[27] Chang, C.-L., Nelson, J., Least squares finite element method for the Stokes problem with zero residual of mass conservation. *SIAM J. Numer. Anal.* **34** (1997), 480–489.

[28] Deang, J., Gunzburger, M., Issues related to least-squares finite element methods for the Stokes equations. *SIAM J. Sci. Comput.* **20** (1998), 878–906.

[29] Dohrmann, C., Bochev, P., A stabilized finite element method for the Stokes problem based on polynomial pressure projections. *Internat. J. Numer. Methods Fluids* **46** (2004), 183–201.

[30] Douglas, J., Wang, J., An absolutely stabilized finite element method for the Stokes problem. *Math. Comp.* **52** (1989) 495–508.

[31] Eason, E., A review of least-squares methods for solving partial differential equations. *Internat. J. Numer. Methods Engrg.* **10** (1976), 1021–1046

[32] Fix, G., Gunzburger, M., Nicolaides, R., On finite element methods of the least-squares type. *Comput. Math. Appl.* **5** (1979), 87–98.

[33] Fix, G., Gunzburger, M., Nicolaides, R., On mixed finite element methods for first-order elliptic systems. *Numer. Math.* **37** (1981), 29–48.

[34] Girault, V., Raviart, P.-A., *Finite Element Methods for Navier-Stokes Equations*. Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin 1986.

[35] Guermond, J.-L., A finite element technique for solving first-order PDEs in L^p. *SIAM J. Numer. Anal.* **42** (2004), 714–737.

[36] Hughes, T., Franca, L., A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity pressure spaces. *Comput. Methods Appl. Mech. Engrg.* **65** (1987), 85–96.

[37] Hughes, T., Franca, L., Balestra, M., A new finite element formulation for computational fluid dynamics: Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Comput. Methods Appl. Mech. Engrg.* **59** (1986), 85–99.

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

[39] Jiang, B.-N., Povinelli, L., Optimal least-squares finite element methods for elliptic problems. *Comput. Methods Appl. Mech. Engrg.* **102** (1993), 199–212.

[40] Jiang, B.-N., Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method. *J. Comput. Phys.* **105** (1993), 108–121.

[41] Liable, J., Pinder, G., Least-squares collocation solution of differential equations on irregularly shaped domains using orthogonal meshes. *Numer. Methods Partial Differential Equations* **5** (1989), 347–361.

[42] Monk, P., Wang, D.-Q., A least-squares method for the Helmholtz equation. *Comput. Meth. Appl. Mech. Engrg.* **175** (1999), 121–136.

[43] Pehlivanov, A., Carey, G., Lazarov, R., Least-squares mixed finite elements for second-order elliptic problems. *SIAM J. Numer. Anal.* **31** (1994), 1368–1377.

[44] Silvester, D., Optimal low order finite element methods for incompressible flow. *Comput. Meth. Appl. Mech. Engrg.* **111** (1994), 357–368.

[45] Stojek, M., Least-squares Treffetz-type elements for the Helmholtz equation. *Internat. J. Numer. Methods Engrg.* **41** (1998), 831–849.

[46] Strang, G., Fix, G., *An Analysis of the Finite Element Method*. Prentice Hall, Englewood Cliffs, NJ, 1973.

Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1110, U.S.A.

E-mail: pbboche@sandia.gov

School of Computational Science, Florida State University, Tallahassee, FL 32306-4120, U.S.A.

E-mail: gunzburg@scs.fsu.edu