

A least-squares finite element method for optimization and control problems

*Pavel B. Bochev** and *Max D. Gunzburger†*

1 Introduction

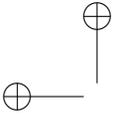
Optimization and control problems for systems governed by partial differential equations arise in many applications. Experimental studies of such problems go back 100 years [20]. Computational approaches have been applied since the advent of the computer age. Most of the efforts in the latter direction have employed elementary optimization strategies but, more recently, there has been considerable practical and theoretical interest in the application of sophisticated local and global optimization strategies, e.g., Lagrange multiplier methods, sensitivity or adjoint-based gradient methods, quasi-Newton methods, evolutionary algorithms, etc.

The optimal control or optimization problems we consider consist of

- *state variables*, i.e., variables that describe the system being modeled;
- *control variables* or *design parameters*, i.e., variables at our disposal that can be used to affect the state variables;
- a *state system*, i.e., partial differential equations relating the state and control variables; and

*Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque NM 87185-1110 (pboche@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

†School of Computational Science and Information Technology, Florida State University, Tallahassee FL 32306-4120 (gunzburg@csit.fsu.edu). Supported in part by CSRI, Sandia National Laboratories under contract 18407.



- a *functional* of the state and control variables whose minimization is the goal.

Then, the problems we consider consist of finding state and control variables that minimize the given functional subject to the state system being satisfied. Here, we restrict attention to linear, elliptic state systems and to quadratic functionals.

The Lagrange multiplier rule is a standard approach for solving finite-dimensional optimization problems. It is not surprising then that several popular approaches to solving optimization and control problems constrained by partial differential equations are based on solving optimality systems deduced from the application of the Lagrange multiplier rule. In [8], least-squares finite element methods were used to develop methods for the approximate solution of the optimality systems; see also [18].

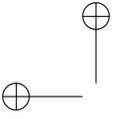
Penalty methods, another popular method for solving optimization problems, have also been applied, albeit to a lesser extent, to problems such as those considered here. In particular, these algorithms enforce the partial differential equations constraints by using well-posed least-squares functionals as penalty terms that are added to the original cost functional. This type of penalty methods offers certain efficiency-related advantages compared to methods based on the solution of the Lagrange multiplier optimality system either by Galerkin or least-squares finite element methods. Such least-squares/penalty methods have been used for an optimal shape design problem [1], for controlling Stokes equations [7], for the Dirichlet control of the Navier-Stokes equations [2, 4], and for optimal control problems constrained by model first-order elliptic systems [19]. An extensive discussion of least-squares/penalty methods is given in [9].

In this paper we discuss a new approach in which the cost functional is *constrained* by the least-squares functional. This approach is more effective than are least squares/penalty methods. For the latter, one has methods that either require the satisfaction of discrete stability conditions or are prone to locking; see, e.g., [10]. Using the new approach, one can define a method that avoids both of these undesirable features. A more detailed discussion comparing these two approaches to incorporating least-square principles into PDE constrained optimization problems as well as the approach of [8] is given in [9].

The paper is organized as follows. In §2, we define an abstract quadratic optimization and control problem constrained by linear, elliptic partial differential equations. Then, in §3, we review results about least-squares finite element methods for the approximate solution of the constraint equations. In §4, we present and analyze the new approach that involves constraining the cost functional by the least-squares functional. Finally, in §5, we provide a concrete example of the abstract theory.

2 Quadratic optimization and control problems in Hilbert spaces with linear constraints

We begin with four given Hilbert spaces Θ , Φ , $\widehat{\Phi}$, and $\widetilde{\Phi}$ along with their dual spaces denoted by $(\cdot)^*$. We assume that $\Phi \subseteq \widehat{\Phi} \subseteq \widetilde{\Phi}$ with continuous embeddings and that



$\tilde{\Phi}$ acts as the pivot space for both the pair $\{\Phi^*, \Phi\}$ and the pair $\{\widehat{\Phi}^*, \widehat{\Phi}\}$ so that we not only have that $\Phi \subseteq \widehat{\Phi} \subseteq \tilde{\Phi} \subseteq \widehat{\Phi}^* \subseteq \Phi^*$, but also

$$\langle \psi, \phi \rangle_{\Phi^*, \Phi} = \langle \psi, \phi \rangle_{\widehat{\Phi}^*, \widehat{\Phi}} = (\psi, \phi)_{\tilde{\Phi}} \quad \forall \psi \in \widehat{\Phi}^* \subseteq \Phi^* \quad \text{and} \quad \forall \phi \in \Phi \subseteq \widehat{\Phi}, \quad (1)$$

where $(\cdot, \cdot)_{\tilde{\Phi}}$ denotes the inner product on $\tilde{\Phi}$. Next, we define the *functional*

$$\mathcal{J}(\phi, \theta) = \frac{1}{2}a_1(\phi - \widehat{\phi}, \phi - \widehat{\phi}) + \frac{1}{2}a_2(\theta, \theta) \quad \forall \phi \in \Phi, \theta \in \Theta, \quad (2)$$

where $a_1(\cdot, \cdot)$ and $a_2(\cdot, \cdot)$ are symmetric bilinear forms on $\widehat{\Phi} \times \widehat{\Phi}$ and $\Theta \times \Theta$, respectively, and $\widehat{\phi} \in \widehat{\Phi}$ is a given function. In the language of control theory, Φ is called the *state space*, ϕ the *state variable*, Θ the *control space*, and θ the *control variable*. In many applications, the control space is finite dimensional in which case θ is often referred to as the vector of *design variables*. We note that often Θ is chosen to be a bounded set in a Hilbert space but, for our purposes, we can consider the less general situation of Θ itself being a Hilbert space. The second term in the functional (2) can be interpreted as a penalty term¹ which limits the size of the control θ .

We make the following assumptions about the bilinear forms $a_1(\cdot, \cdot)$ and $a_2(\cdot, \cdot)$:

$$\begin{cases} a_1(\phi, \mu) \leq C_1 \|\phi\|_{\widehat{\Phi}} \|\mu\|_{\widehat{\Phi}} & \forall \phi, \mu \in \widehat{\Phi} \\ a_2(\theta, \nu) \leq C_2 \|\theta\|_{\Theta} \|\nu\|_{\Theta} & \forall \theta, \nu \in \Theta \\ a_1(\phi, \phi) \geq 0 & \forall \phi \in \widehat{\Phi} \\ a_2(\theta, \theta) \geq K_2 \|\theta\|_{\Theta}^2 & \forall \theta \in \Theta, \end{cases} \quad (3)$$

where C_1 , C_2 , and K_2 are all positive constants.

Given another Hilbert space Λ , the additional bilinear forms $b_1(\cdot, \cdot)$ on $\Phi \times \Lambda$ and $b_2(\cdot, \cdot)$ on $\Theta \times \Lambda$, and the function $g \in \Lambda^*$, we define the *constraint equation*

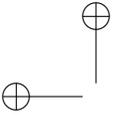
$$b_1(\phi, \psi) + b_2(\theta, \psi) = \langle g, \psi \rangle_{\Lambda^*, \Lambda} \quad \forall \psi \in \Lambda. \quad (4)$$

We make the following assumptions about the bilinear forms $b_1(\cdot, \cdot)$ and $b_2(\cdot, \cdot)$:

$$\begin{cases} b_1(\phi, \psi) \leq c_1 \|\phi\|_{\Phi} \|\psi\|_{\Lambda} & \forall \phi \in \Phi, \psi \in \Lambda \\ b_2(\theta, \psi) \leq c_2 \|\theta\|_{\Theta} \|\psi\|_{\Lambda} & \forall \theta \in \Theta, \psi \in \Lambda \\ \sup_{\psi \in \Lambda, \psi \neq 0} \frac{b_1(\phi, \psi)}{\|\psi\|_{\Lambda}} \geq k_1 \|\phi\|_{\Phi} & \forall \phi \in \Phi \\ \sup_{\phi \in \Phi, \phi \neq 0} \frac{b_1(\phi, \psi)}{\|\phi\|_{\Phi}} > 0 & \forall \psi \in \Lambda, \end{cases} \quad (5)$$

where c_1 , c_2 , and k_1 are all positive constants.

¹The usage of the terminology ‘‘penalty term’’ in conjunction with the second term in (2) should not be confused with the usage of that terminology in §1. In particular, the second term in (2) has no connection with the terminology ‘‘least-squares/penalty’’ previously used.



We consider the *optimal control problem*

$$\min_{(\phi, \theta) \in \Phi \times \Theta} \mathcal{J}(\phi, \theta) \quad \text{subject to} \quad b_1(\phi, \psi) + b_2(\theta, \psi) = \langle g, \psi \rangle_{\Lambda^*, \Lambda} \quad \forall \psi \in \Lambda. \quad (6)$$

The following result is proved in, e.g., [8].

Theorem 1. *Let the assumptions (3) and (5) hold. Then, the optimal control problem (6) has a unique solution $(\phi, \theta) \in \Phi \times \Theta$.*

It is instructive to rewrite the functional (2), the constraint (4), and the optimal control problem (6) in operator notation. To this end, we note that the bilinear forms serve to define operators

$$\begin{aligned} A_1 : \widehat{\Phi} &\rightarrow \widehat{\Phi}^*, & A_2 : \Theta &\rightarrow \Theta^*, & B_1 : \Phi &\rightarrow \Lambda^*, \\ B_1^* : \Lambda &\rightarrow \Phi^*, & B_2 : \Theta &\rightarrow \Lambda^*, & B_2^* : \Lambda &\rightarrow \Theta^* \end{aligned}$$

through the following relations:

$$\begin{aligned} a_1(\phi, \mu) &= \langle A_1 \phi, \mu \rangle_{\widehat{\Phi}^*, \widehat{\Phi}} & \forall \phi, \mu \in \widehat{\Phi} \\ a_2(\theta, \nu) &= \langle A_2 \theta, \nu \rangle_{\Theta^*, \Theta} & \forall \theta, \nu \in \Theta \\ b_1(\phi, \psi) &= \langle B_1 \phi, \psi \rangle_{\Lambda^*, \Lambda} = \langle B_1^* \psi, \phi \rangle_{\Phi^*, \Phi} & \forall \phi \in \Phi, \psi \in \Lambda \\ b_2(\psi, \theta) &= \langle B_2 \theta, \psi \rangle_{\Lambda^*, \Lambda} = \langle B_2^* \psi, \theta \rangle_{\Theta^*, \Theta} & \forall \theta \in \Theta, \psi \in \Lambda. \end{aligned} \quad (7)$$

Then, the functional (2) and the constraint (4) respectively take the forms

$$\mathcal{J}(\phi, \theta) = \frac{1}{2} \langle A_1(\phi - \widehat{\phi}), (\phi - \widehat{\phi}) \rangle_{\widehat{\Phi}^*, \widehat{\Phi}} + \frac{1}{2} \langle A_2 \theta, \theta \rangle_{\Theta^*, \Theta} \quad \forall \phi \in \Phi, \theta \in \Theta \quad (8)$$

and

$$B_1 \phi + B_2 \theta = g \quad \text{in } \Lambda^* \quad (9)$$

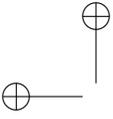
and the optimal control problem (6) takes the form

$$\min_{(\phi, \theta) \in \Phi \times \Theta} \mathcal{J}(\phi, \theta) \quad \text{subject to} \quad B_1 \phi + B_2 \theta = g \quad \text{in } \Lambda^*. \quad (10)$$

Assumptions (3) and (5) imply that A_1 , A_2 , B_1 , B_2 , B_1^* , and B_2^* are bounded with

$$\begin{aligned} \|A_1\|_{\widehat{\Phi} \rightarrow \widehat{\Phi}^*} &\leq C_1, & \|A_2\|_{\Theta \rightarrow \Theta^*} &\leq C_2, & \|B_1\|_{\Phi \rightarrow \Lambda^*} &\leq c_1, \\ \|B_1^*\|_{\Lambda \rightarrow \Phi^*} &\leq c_1, & \|B_2\|_{\Theta \rightarrow \Lambda^*} &\leq c_2, & \|B_2^*\|_{\Lambda \rightarrow \Theta^*} &\leq c_2 \end{aligned}$$

and that the operator B_1 is invertible with $\|B_1^{-1}\|_{\Lambda^* \rightarrow \Phi} \leq 1/k_1$. See [8] for details. Note that, given $\theta \in \Theta$ and $g \in \Lambda^*$, the assumptions in (5) about the bilinear form $b_1(\cdot, \cdot)$ imply that the constraint equation (9) may be solved for $\phi \in \Phi$ to yield $\phi = B^{-1}(g - B_2 \theta)$.



3 Least-squares formulation of the constraint equations

The constraint equations are given in variational form in (4) and in equivalent operator form in (9). They may also be defined through a least-squares minimization problem. Let $D : \Lambda \rightarrow \Lambda^*$ be a self-adjoint, strongly coercive operator,² i.e., there exist constants $c_d > 0$ and $k_d > 0$ such that

$$\langle D\lambda, \psi \rangle_{\Lambda^*, \Lambda} \leq c_d \|\lambda\|_{\Lambda} \|\psi\|_{\Lambda} \quad \text{and} \quad \langle D\lambda, \lambda \rangle_{\Lambda^*, \Lambda} \geq k_d \|\lambda\|_{\Lambda}^2 \quad \forall \lambda, \psi \in \Lambda. \quad (13)$$

Note that then $k_d \leq \|D\|_{\Lambda \rightarrow \Lambda^*} \leq c_d$ and $1/c_d \leq \|D^{-1}\|_{\Lambda^* \rightarrow \Lambda} \leq 1/k_d$. Let³

$$\begin{aligned} \mathcal{K}(\phi; \theta, g) = \\ \langle B_1\phi + B_2\theta - g, D^{-1}(B_1\phi + B_2\theta - g) \rangle_{\Lambda^*, \Lambda} \quad \forall \phi \in \Phi, \theta \in \Theta, g \in \Lambda^*. \end{aligned} \quad (14)$$

Given $\theta \in \Theta$ and $g \in \Lambda^*$, consider the problem

$$\min_{\phi \in \Phi} \mathcal{K}(\phi; \theta, g). \quad (15)$$

Clearly, this problem is equivalent to (4) and (9), i.e., solutions of (15) are solutions of (4) or (9) and conversely. The Euler-Lagrange equation corresponding to the problem (15) is given, in variational form, by

$$\tilde{b}_1(\phi, \mu) = \langle \tilde{g}_1, \mu \rangle_{\Phi^*, \Phi} - \tilde{b}_2(\theta, \mu) \quad \forall \mu \in \Phi, \quad (16)$$

where

$$\tilde{b}_1(\phi, \mu) = \langle B_1\mu, D^{-1}B_1\phi \rangle_{\Lambda^*, \Lambda} = \langle B_1^*D^{-1}B_1\phi, \mu \rangle_{\Phi^*, \Phi} \quad \forall \phi, \mu \in \Phi \quad (17)$$

$$\tilde{b}_2(\theta, \mu) = \langle B_1\mu, D^{-1}B_2\theta \rangle_{\Lambda^*, \Lambda} = \langle B_1^*D^{-1}B_2\theta, \mu \rangle_{\Phi^*, \Phi} \quad \forall \theta \in \Theta, \mu \in \Phi \quad (18)$$

and

$$\tilde{g}_1 = B_1^*D^{-1}g \in \Phi^*. \quad (19)$$

²In the sequel, we will also use the induced bilinear form

$$d(\lambda, \psi) = \langle D\lambda, \psi \rangle_{\Lambda^*, \Lambda} \quad \forall \lambda, \psi \in \Lambda. \quad (11)$$

The following results are immediate.

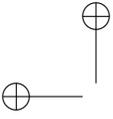
Proposition 2. *Assume that the operator D is symmetric and that (13) holds. Then, the bilinear form $d(\cdot, \cdot)$ is symmetric and*

$$d(\lambda, \psi) \leq c_d \|\lambda\|_{\Lambda} \|\psi\|_{\Lambda} \quad \forall \lambda, \psi \in \Lambda \quad \text{and} \quad d(\lambda, \lambda) \geq k_d \|\lambda\|_{\Lambda}^2 \quad \forall \lambda \in \Lambda. \quad (12)$$

³Let $R : \Lambda \rightarrow \Lambda^*$ denote the Riesz operator, i.e., we have that if $v = R\lambda$ and $\chi = R\psi$ for $\lambda, \psi \in \Lambda$ and $v, \chi \in \Lambda^*$, then $\|\lambda\|_{\Lambda} = \|v\|_{\Lambda^*}$, $\|\psi\|_{\Lambda} = \|\chi\|_{\Lambda^*}$, and

$$(\psi, \lambda)_{\Lambda} = \langle R\psi, \lambda \rangle_{\Lambda^*, \Lambda} = \langle \chi, R^{-1}v \rangle_{\Lambda^*, \Lambda} = (v, \chi)_{\Lambda^*}.$$

Then, if one chooses $D = R$, the functional (14) reduces to $\mathcal{K}(\phi; \theta, g) = (B_1\phi + B_2\theta - g, B_1\phi + B_2\theta - g)_{\Lambda^*} = \|B_1\phi + B_2\theta - g\|_{\Lambda^*}^2$. Note that, in general, (14) can also be written as an inner product, i.e., $\mathcal{K}(\phi; \theta, g) = (B_1\phi + B_2\theta - g, RD^{-1}(B_1\phi + B_2\theta - g))_{\Lambda^*}$.



The following proposition shows that the bilinear forms $\tilde{b}_1(\cdot, \cdot)$ and $\tilde{b}_2(\cdot, \cdot)$ are continuous and the former is strongly coercive.

Proposition 3. *Assume that (5) and (13) hold. Then, the bilinear form $\tilde{b}_1(\cdot, \cdot)$ is symmetric and there exist positive constants \tilde{c}_1 , \tilde{c}_2 , and \tilde{k}_1 such that*

$$\begin{cases} \tilde{b}_1(\phi, \mu) \leq \tilde{c}_1 \|\phi\|_{\Phi} \|\mu\|_{\Phi} & \forall \phi, \mu \in \Phi \\ \tilde{b}_2(\theta, \mu) \leq \tilde{c}_2 \|\mu\|_{\Phi} \|\theta\|_{\Theta} & \forall \theta \in \Theta, \mu \in \Phi \\ \tilde{b}_1(\phi, \phi) \geq \tilde{k}_1 \|\phi\|_{\Phi}^2 & \forall \phi \in \Phi. \end{cases} \quad (20)$$

Moreover, $\|\tilde{g}_1\|_{\Phi^*} \leq \frac{c_1}{k_d} \|g\|_{\Lambda^*}$ and the problem (16), or equivalently (15), has a unique solution.

Proof. The symmetry of the bilinear form $\tilde{b}_1(\cdot, \cdot)$ follows immediately from the symmetry of the operator D . Since $b_1(\phi, \psi) \leq c_1 \|\phi\|_{\Phi} \|\psi\|_{\Lambda}$ for all $\phi \in \Phi$ and $\psi \in \Lambda$, we have that

$$\langle B_1 \phi, \psi \rangle_{\Lambda^*, \Lambda} = b_1(\phi, \psi) \leq c_1 \|\phi\|_{\Phi} \|\psi\|_{\Lambda} \quad \forall \phi \in \Phi, \psi \in \Lambda$$

from which it easily follows that $\|B_1 \phi\|_{\Lambda^*} \leq c_1 \|\phi\|_{\Phi}$ for all $\phi \in \Phi$. We then have that

$$\tilde{b}_1(\phi, \mu) = \langle B_1 \mu, D^{-1} B_1 \phi \rangle_{\Lambda^*, \Lambda} \leq \|D^{-1}\|_{\Lambda \rightarrow \Lambda^*} \|B_1 \phi\|_{\Lambda^*} \|B_1 \mu\|_{\Lambda^*} \leq \frac{c_1^2}{k_d} \|\phi\|_{\Phi} \|\mu\|_{\Phi}.$$

In a similar way, one shows that $\|B_2 \theta\|_{\Lambda^*} \leq c_2 \|\theta\|_{\Theta}$ for all $\theta \in \Theta$ and

$$\tilde{b}_2(\theta, \mu) = \langle B_1 \mu, D^{-1} B_2 \theta \rangle_{\Lambda^*, \Lambda} \leq \frac{c_1 c_2}{k_d} \|\theta\|_{\Theta} \|\mu\|_{\Phi}.$$

Also in a similar way, the bound for $\|\tilde{g}_1\|_{\Phi^*}$ can be obtained.

Next, since $\sup_{\psi \in \Lambda, \psi \neq 0} (b_1(\phi, \psi) / \|\psi\|_{\Lambda}) \geq k_1 \|\phi\|_{\Phi}$ for all $\phi \in \Phi$, we have that

$$\|B_1 \phi\|_{\Lambda^*} = \sup_{\psi \in \Lambda, \psi \neq 0} \frac{\langle B_1 \phi, \psi \rangle_{\Lambda^*, \Lambda}}{\|\psi\|_{\Lambda}} = \sup_{\psi \in \Lambda, \psi \neq 0} \frac{b_1(\phi, \psi)}{\|\psi\|_{\Lambda}} \geq k_1 \|\phi\|_{\Phi} \quad \forall \phi \in \Phi.$$

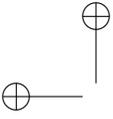
We then have that, with $\lambda = D^{-1} B_1 \phi$ so that $\|B_1 \phi\|_{\Lambda^*} \leq \|D\|_{\Lambda \rightarrow \Lambda^*} \|\lambda\|_{\Lambda}$,

$$\begin{aligned} \tilde{b}_1(\phi, \phi) &= \langle B_1 \phi, D^{-1} B_1 \phi \rangle_{\Lambda^*, \Lambda} = \langle D \lambda, \lambda \rangle_{\Lambda^*, \Lambda} \geq k_d \|\lambda\|_{\Lambda}^2 \\ &\geq \frac{k_d}{\|D\|_{\Lambda \rightarrow \Lambda^*}^2} \|B_1 \phi\|_{\Lambda^*}^2 \geq \frac{k_1^2 k_d}{c_d^2} \|\phi\|_{\Phi}^2. \end{aligned}$$

Thus, (20) holds with $\tilde{c}_1 = c_1^2 / k_d$, $\tilde{c}_2 = c_1 c_2 / k_d$, and $\tilde{k}_1 = k_1^2 k_d / c_d^2$.

The unique solvability of (16) then follows from the Lax-Milgram lemma. \square

As an immediate consequence of Proposition 3, we have that the least-squares functional (14) is norm equivalent in the following sense.



Corollary 4. *Assume that (13) and the conditions on the bilinear form $b_1(\cdot, \cdot)$ in (5) hold. Then,*

$$\tilde{k}_1 \|\phi\|_{\Phi}^2 \leq \mathcal{K}(\phi; 0, 0) = \tilde{b}_1(\phi, \phi) = \langle B_1 \phi, D^{-1} B_1 \phi \rangle_{\Lambda^*, \Lambda} \leq \tilde{c}_1 \|\phi\|_{\Phi}^2 \quad \forall \phi \in \Phi. \quad (21)$$

For all $\mu \in \Phi$, we can rewrite (16) as $\langle B_1 \mu, D^{-1}(B_1 \phi + B_2 \theta - g) \rangle_{\Lambda^*, \Lambda} = 0$ or $\langle B_1^* D^{-1}(B_1 \phi + B_2 \theta - g), \mu \rangle_{\Phi^*, \Phi} = 0$ so that, in operator form, we have that (16) is equivalent to

$$\tilde{B}_1 \phi + \tilde{B}_2 \theta = \tilde{g}_1 \quad \text{in } \Phi^*, \quad (22)$$

where

$$\tilde{B}_1 = B_1^* D^{-1} B_1 : \Phi \rightarrow \Phi^*, \quad \text{and} \quad \tilde{B}_2 = B_1^* D^{-1} B_2 : \Theta \rightarrow \Phi^*. \quad (23)$$

Note that (21) implies that the operator $\tilde{B}_1 = B_1^* D^{-1} B_1$ in (22) is symmetric and coercive even when the operator B_1 in (9) is indefinite and/or non-symmetric; these observations, of course, follow from the fact that the bilinear form $b_1(\cdot, \cdot)$ is weakly coercive (see (5)) while the bilinear form $\tilde{b}_1(\cdot, \cdot)$ is strongly coercive (see (20)). It is also easy to see that (22) has the same solutions as (9).

Discretization of (16), or equivalently of (22), is accomplished in the standard manner. One chooses a subspace $\Phi^h \subset \Phi$ and then, given $\theta \in \Theta$ and $\tilde{g} \in \Phi^*$, one solves the problem

$$\tilde{b}_1(\phi^h, \mu^h) = \langle \tilde{g}_1, \mu^h \rangle_{\Phi^*, \Phi} - \tilde{b}_2(\theta, \mu^h) \quad \forall \mu^h \in \Phi^h. \quad (24)$$

Then, (21) and the Lax-Milgram and Cea lemmas immediately imply the following results.

Proposition 5. *Assume that (5) and (13) hold. Then, the problem (24) has a unique solution and, if ϕ denotes the solution of the problem (16), or equivalently, of (22), there exists a constant $C > 0$ whose value is independent of h , ϕ , and ϕ^h such that*

$$\|\phi - \phi^h\|_{\Phi} \leq C \inf_{\phi^h \in \Phi^h} \|\phi - \phi^h\|_{\Phi}.$$

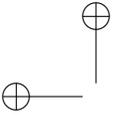
If $\{\phi_j\}_{j=1}^J$ denotes a basis for Φ^h , then the problem (24) is equivalent to the matrix problem

$$\tilde{\mathbb{B}}_1 \vec{\phi} = \tilde{\mathbf{g}}_0, \quad (25)$$

where $\vec{\phi}$ is the vector of coefficients for ϕ^h , $(\tilde{\mathbb{B}}_1)_{ij} = \tilde{b}_1(\phi_i, \phi_j) = \langle \tilde{B}_1 \phi_i, \phi_j \rangle_{\Phi^*, \Phi}$, and $(\tilde{\mathbf{g}}_0)_i = \langle \tilde{g}_1, \phi_i \rangle_{\Phi^*, \Phi} - \tilde{b}_2(\theta, \phi_i) = \langle \tilde{g}_1 - \tilde{B}_2 \theta, \phi_i \rangle_{\Phi^*, \Phi} = \langle B_1^* D^{-1} g - B_1^* D^{-1} B_2 \theta, \phi_i \rangle_{\Phi^*, \Phi}$.

The following result follows easily from Proposition 3 and Corollary 4.

Corollary 6. *Assume that (13) and the conditions on the bilinear form $b_1(\cdot, \cdot)$ in (5) hold. Then, the matrix $\tilde{\mathbb{B}}_1$ is symmetric positive definite and spectrally equivalent to the Gramm matrix \mathbb{G} , $(\mathbb{G})_{i,j} = (\phi_i, \phi_j)_{\Phi}$.*



The main advantages of using a least-squares finite element method to solve the constraint equation (9) are that the matrix $\widetilde{\mathbb{B}}_1$ in (25) is symmetric and positive definite even when the operator B_1 in (9) is indefinite and/or non-symmetric, and that the conforming finite element subspace $\Phi^h \subset \Phi$ is not subject to any additional discrete stability conditions.⁴ In incorporating the least-squares formalism into the optimization setting of §2, we want to preserve these advantages.

In the sequel, we will also use the bilinear form

$$c(\theta, \nu) = \langle B_2 \nu, D^{-1} B_2 \theta \rangle_{\Lambda^*, \Lambda} = \langle B_2^* D^{-1} B_2 \theta, \nu \rangle_{\Theta^*, \Theta} \quad \forall \theta, \nu \in \Theta \quad (27)$$

and the function

$$\widetilde{g}_2 = B_2^* D^{-1} g \in \Theta^*. \quad (28)$$

The following results are immediate.

Proposition 7. *Assume that the operator D is symmetric and that (13) and the condition on the bilinear form $b_2(\cdot, \cdot)$ in (5) hold. Then, the bilinear form $c(\cdot, \cdot)$ is symmetric and, for some constant $C_c > 0$,*

$$c(\theta, \nu) \leq C_c \|\theta\|_{\Theta} \|\nu\|_{\Theta} \quad \forall \theta, \nu \in \Theta \quad \text{and} \quad c(\theta, \theta) \geq 0 \quad \forall \theta \in \Theta. \quad (29)$$

Moreover, $\|\widetilde{g}_2\|_{\Theta^*} \leq \frac{c_2}{k_d} \|g\|_{\Lambda^*}$.

Associated with the bilinear form $c(\cdot, \cdot)$ we have the operator $C = B_2^* D^{-1} B_2 : \Theta \rightarrow \Theta^*$, i.e., $c(\theta, \nu) = \langle C\theta, \nu \rangle_{\Theta^*, \Theta}$ for all $\theta, \nu \in \Theta$.

4 Methods based on constraining by the least-squares functional

A means for incorporating least-squares notions into a solution method for the constrained optimization problem of §2 is to solve, instead of (6) or its equivalent form (10), the bilevel minimization problem

$$\min_{(\phi, \theta) \in \Phi \times \Theta} \mathcal{J}(\phi, \theta) \quad \text{subject to} \quad \min_{\phi \in \Phi} \mathcal{K}(\phi; \theta, g). \quad (30)$$

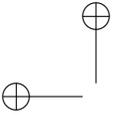
From (22), one sees that this is equivalent to the problem

$$\min_{(\phi, \theta) \in \Phi \times \Theta} \mathcal{J}(\phi, \theta) \quad \text{subject to} \quad \widetilde{B}_1 \phi + \widetilde{B}_2 \theta = \widetilde{g} \quad \text{in } \Phi^*. \quad (31)$$

⁴A direct, conforming Galerkin finite element discretization of (9) requires that the discrete stability conditions

$$\begin{cases} \sup_{\psi^h \in \Lambda, \psi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\psi^h\|_{\Lambda}} \geq k_1^h \|\phi^h\|_{\Phi} & \forall \phi^h \in \Phi^h \\ \sup_{\phi^h \in \Phi^h, \phi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\phi^h\|_{\Phi}} > 0 & \forall \psi^h \in \Lambda^h, \end{cases} \quad (26)$$

be satisfied.



Using a Lagrange multiplier $\mu \in \Phi$ to enforce the constraint in (31) we obtain the Euler-Lagrange equations

$$\begin{cases} A_1\phi & + \tilde{B}_1\mu & = A_1\hat{\phi} & \text{in } \Phi^* \\ & A_2\theta & + \tilde{B}_2^*\mu & = 0 & \text{in } \Theta^* \\ \tilde{B}_1\phi & + \tilde{B}_2\theta & & = \tilde{g}_1 & \text{in } \Phi^*, \end{cases} \quad (32)$$

for the saddle point $\{(\phi, \theta), \mu\}$.

The problem (31) should be contrasted with the problem (10). Both (10) and (31) involve the same functional $\mathcal{J}(\cdot, \cdot)$, but are constrained differently. As a result, the former leads to the optimality system (see [8])

$$\begin{cases} A_1\phi & + B_1\mu & = A_1\hat{\phi} & \text{in } \Phi^* \\ & A_2\theta & + B_2^*\mu & = 0 & \text{in } \Theta^* \\ B_1\phi & + B_2\theta & & = g & \text{in } \Lambda^*, \end{cases} \quad (33)$$

while the latter leads to the optimality system (32). Although both optimality systems are of saddle point type, their internal structures are significantly different. For example, the operator B_1 that plays a central role in (33) may be non-symmetric and indefinite; on the other hand, the operator $\tilde{B}_1 = B_1^*D^{-1}B_1$ that plays the analogous role in (32) is always symmetric and positive definite whenever the assumptions (5) and (13) hold.

Penalization can be used to facilitate the solution of the system (32). To this end, we let $\tilde{D} : \Phi \rightarrow \Phi^*$ be a self-adjoint, strongly coercive operator, i.e., there exist constants $\tilde{c}_d > 0$ and $\tilde{k}_d > 0$ such that

$$\langle \tilde{D}\mu, \phi \rangle_{\Phi^*, \Phi} \leq \tilde{c}_d \|\mu\|_{\Phi} \|\phi\|_{\Phi} \quad \text{and} \quad \langle \tilde{D}\mu, \mu \rangle_{\Phi^*, \Phi} \geq \tilde{k}_d \|\mu\|_{\Phi}^2. \quad (34)$$

for all $\phi, \mu \in \Phi$. Corresponding to the operator \tilde{D} , we have the symmetric, coercive bilinear form

$$\tilde{d}(\phi, \mu) = \langle \tilde{D}\mu, \phi \rangle_{\Phi^*, \Phi} \quad \forall \phi, \mu \in \Phi.$$

We then consider the penalized functional

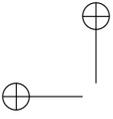
$$\tilde{\mathcal{J}}_{\epsilon}(\phi, \theta) = \mathcal{J}(\phi, \theta) + \frac{1}{\epsilon} \langle \tilde{B}_1\phi + \tilde{B}_2\theta - \tilde{g}_1, \tilde{D}^{-1}(\tilde{B}_1\phi + \tilde{B}_2\theta - \tilde{g}_1) \rangle_{\Phi^*, \Phi}$$

and the unconstrained optimization problem

$$\min_{\phi \in \Phi, \theta \in \Theta} \tilde{\mathcal{J}}_{\epsilon}(\phi, \theta). \quad (35)$$

The Euler-Lagrange equations corresponding to this problem are given by

$$\begin{cases} \left(A_1 + \frac{1}{\epsilon} \tilde{B}_1 \tilde{D}^{-1} \tilde{B}_1 \right) \phi_{\epsilon} + \frac{1}{\epsilon} \tilde{B}_1 \tilde{D}^{-1} \tilde{B}_2 \theta_{\epsilon} = A_1 \hat{\phi} + \frac{1}{\epsilon} \tilde{B}_1 \tilde{D}^{-1} \tilde{g}_1 & \text{in } \Phi^* \\ \left(A_2 + \frac{1}{\epsilon} \tilde{B}_2^* \tilde{D}^{-1} \tilde{B}_2 \right) \theta_{\epsilon} + \frac{1}{\epsilon} \tilde{B}_2^* \tilde{D}^{-1} \tilde{B}_1 \phi_{\epsilon} = \frac{1}{\epsilon} \tilde{B}_2^* \tilde{D}^{-1} \tilde{g}_1 & \text{in } \Theta^* \end{cases} \quad (36)$$



or

$$\begin{cases} \left(A_1 + \frac{1}{\epsilon} B_1^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} B_1 \right) \phi_\epsilon \\ + \frac{1}{\epsilon} B_1^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} B_2 \theta_\epsilon = A_1 \hat{\phi} + \frac{1}{\epsilon} B_1^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} g & \text{in } \Phi^* \\ \left(A_2 + \frac{1}{\epsilon} B_2^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} B_2 \right) \theta_\epsilon \\ + \frac{1}{\epsilon} B_2^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} B_1 \phi_\epsilon = \frac{1}{\epsilon} B_2^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} g & \text{in } \Theta^*. \end{cases} \quad (37)$$

Letting $\mu_\epsilon = \tilde{D}^{-1}(\tilde{B}_1 \phi_\epsilon + \tilde{B}_2 \theta_\epsilon - \tilde{g}_1)$, it is easy to see that (36) is equivalent to the following regular perturbation of (32):

$$\begin{cases} A_1 \phi_\epsilon & + \tilde{B}_1 \mu_\epsilon & = A_1 \hat{\phi} & \text{in } \Phi^* \\ & A_2 \theta_\epsilon + \tilde{B}_2^T \mu_\epsilon & = 0 & \text{in } \Theta^* \\ \tilde{B}_1 \phi_\epsilon + \tilde{B}_2 \theta_\epsilon - \epsilon \tilde{D} \mu_\epsilon & = \tilde{g}_1 & \text{in } \Phi^*. \end{cases} \quad (38)$$

The systems (36) and (38) are equivalent, but their discretizations are not, even if we use the same subspaces $\Phi^h \subset \Phi$ and $\Theta^h \subset \Theta$ to discretize both systems. However, unlike the situation that occurs when one directly penalizes the cost functional with a least squares functional (see [9]), the discretization of either (36) or (38) will result in matrix systems (after elimination in the second case) that are uniformly (with respect to h) positive definite without regard to (26).

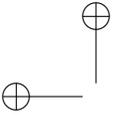
4.1 Discretize-then-eliminate

Let $\{\phi_j\}_{j=1}^J$ and $\{\theta_k\}_{k=1}^K$, where $J = \dim(\Phi^h)$ and $K = \dim(\Theta^h)$, denote the chosen basis sets for Φ^h and Θ^h , respectively. In addition to the matrix $\tilde{\mathbb{B}}_1$ defined previously in (25), we define the matrices

$$\begin{cases} (\mathbb{A}_1)_{ij} = a_1(\phi_i, \phi_j) & \text{for } i, j = 1, \dots, J \\ (\mathbb{A}_2)_{k\ell} = a_2(\theta_k, \theta_\ell) & \text{for } k, \ell = 1, \dots, K \\ (\tilde{\mathbb{B}}_2)_{jk} = \tilde{b}_2(\theta_k, \phi_j) = \langle B_2 \theta_k, D^{-1} B_1 \phi_j \rangle_{\Lambda^*, \Lambda} & \text{for } k = 1, \dots, K, j = 1, \dots, J \\ (\tilde{\mathbb{D}})_{ij} = \tilde{d}(\phi_i, \phi_j) = \langle \tilde{D} \phi_i, \phi_j \rangle_{\Phi^*, \Phi} & \text{for } i, j = 1, \dots, J \end{cases}$$

and the vectors

$$\begin{cases} (\vec{\mathbf{f}})_j = a_1(\hat{\phi}, \phi_j) & \text{for } j = 1, \dots, J \\ (\vec{\mathbf{g}}_1)_i = \langle \tilde{g}_1, \phi_i \rangle_{\Phi^*, \Phi} = \langle B_1 \phi_i, D^{-1} g \rangle_{\Lambda^*, \Lambda} & \text{for } k = 1, \dots, K. \end{cases}$$



Then, discretizing the equivalent weak formulation corresponding to (38) results in the matrix problem⁵

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \tilde{\mathbb{B}}_1 \\ 0 & \mathbb{A}_2 & \tilde{\mathbb{B}}_2^T \\ \tilde{\mathbb{B}}_1 & \tilde{\mathbb{B}}_2 & -\epsilon\tilde{\mathbb{D}} \end{pmatrix} \begin{pmatrix} \vec{\phi}_\epsilon \\ \vec{\theta}_\epsilon \\ \vec{\mu}_\epsilon \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{0} \\ \vec{g}_1 \end{pmatrix}. \quad (40)$$

The system (40) is symmetric and indefinite, but it is uniformly (with respect to h) invertible without regard to (26). Indeed, we have that the matrices $\tilde{\mathbb{B}}_1$ and \mathbb{A}_2 are symmetric and positive definite whenever (3), (5), and (13) hold.

The vector of coefficients $\vec{\mu}_\epsilon$ may be eliminated from (40) to yield

$$\begin{cases} \left(\mathbb{A}_1 + \frac{1}{\epsilon}\tilde{\mathbb{B}}_1\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{B}}_1 \right) \vec{\phi}_\epsilon + \frac{1}{\epsilon}\tilde{\mathbb{B}}_1\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{B}}_2\vec{\theta}_\epsilon = \vec{f} + \frac{1}{\epsilon}\tilde{\mathbb{B}}_1\tilde{\mathbb{D}}^{-1}\vec{g}_1 \\ \left(\mathbb{A}_2 + \frac{1}{\epsilon}\tilde{\mathbb{B}}_2^T\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{B}}_2 \right) \vec{\theta}_\epsilon + \frac{1}{\epsilon}\tilde{\mathbb{B}}_2^T\tilde{\mathbb{D}}^{-1}\tilde{\mathbb{B}}_1\vec{\phi}_\epsilon = \frac{1}{\epsilon}\tilde{\mathbb{B}}_2^T\tilde{\mathbb{D}}^{-1}\vec{g}_1. \end{cases} \quad (41)$$

Theorem 8. *Let (3), (5), and (13) hold. Then, (40) has a unique solution $\phi_\epsilon^h \in \Phi^h$, $\theta_\epsilon^h \in \Theta^h$, and $\mu_\epsilon^h \in \Phi^h$. Moreover, if $\phi \in \Phi$, $\theta \in \Theta$, and $\mu \in \Phi$ denotes the unique solution of (30), or equivalently, of the optimization problem (6), then there exist a constant $C > 0$ whose value is independent of ϵ and h such that*

$$\begin{aligned} \|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta + \|\mu - \mu_\epsilon^h\|_\Phi &\leq C\epsilon \left(\|g\|_{\Lambda^*} + \|\hat{\phi}\|_{\hat{\Phi}} \right) \\ &+ C \left(\inf_{\tilde{\phi}^h \in \Phi^h} \|\phi_\epsilon - \tilde{\phi}^h\|_\Phi + \inf_{\tilde{\theta}^h \in \Theta^h} \|\theta_\epsilon - \tilde{\theta}^h\|_\Theta + \inf_{\tilde{\mu}^h \in \Phi^h} \|\mu_\epsilon - \tilde{\mu}^h\|_\Phi \right). \end{aligned} \quad (42)$$

It is usually the case that the approximation-theoretic terms on the right-hand side of (42) satisfy inequalities of the type

$$\inf_{\tilde{\phi}^h \in \Phi^h} \|\phi_\epsilon - \tilde{\phi}^h\|_\Phi \leq Ch^\alpha, \quad \inf_{\tilde{\mu}^h \in \Phi^h} \|\mu_\epsilon - \tilde{\mu}^h\|_\Phi \leq Ch^\alpha,$$

and

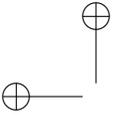
$$\inf_{\tilde{\theta}^h \in \Theta^h} \|\theta_\epsilon - \tilde{\theta}^h\|_\Theta \leq Ch^\beta,$$

where $\alpha > 0$ and $\beta > 0$ depend on the degree of the polynomials used for the spaces Φ^h and Θ^h and the regularity of the solution ϕ_ϵ , θ_ϵ , and μ_ϵ of (30). Combining with (42) we have that

$$\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta + \|\mu - \mu_\epsilon^h\|_\Phi \leq C(\epsilon + h^\alpha + h^\beta) \quad (43)$$

⁵Discretization of the unperturbed system (32) yields the related discrete system

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \tilde{\mathbb{B}}_1 \\ 0 & \mathbb{A}_2 & \tilde{\mathbb{B}}_2^T \\ \tilde{\mathbb{B}}_1 & \tilde{\mathbb{B}}_2 & 0 \end{pmatrix} \begin{pmatrix} \vec{\phi} \\ \vec{\theta} \\ \vec{\mu} \end{pmatrix} = \begin{pmatrix} \vec{f} \\ \vec{0} \\ \vec{g}_1 \end{pmatrix}. \quad (39)$$



so that if $\beta \geq \alpha$ and one chooses $\epsilon = h^\alpha$, one obtains the optimal error estimate

$$\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta + \|\mu - \mu_\epsilon^h\|_\Phi \leq C\epsilon = Ch^\alpha. \quad (44)$$

It is again important to note that the result (42) does not require that (26) is satisfied and that locking cannot occur.

4.2 Eliminate-then-discretize

Alternately, one could discretize (36) to obtain

$$\begin{pmatrix} \mathbb{A}_1 + \frac{1}{\epsilon}\mathbb{K}_1 & \frac{1}{\epsilon}\mathbb{K}_2 \\ \frac{1}{\epsilon}\mathbb{K}_2^T & \mathbb{A}_2 + \frac{1}{\epsilon}\tilde{\mathbb{C}} \end{pmatrix} \begin{pmatrix} \vec{\phi}_\epsilon \\ \vec{\theta}_\epsilon \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} + \frac{1}{\epsilon}\tilde{\mathbf{g}}_1 \\ \frac{1}{\epsilon}\tilde{\mathbf{g}}_2 \end{pmatrix}. \quad (45)$$

The matrices \mathbb{A}_1 and \mathbb{A}_2 and the vector $\vec{\mathbf{f}}$ are defined as before; we also have, in terms of the basis vectors for Φ^h and Θ^h , that

$$\begin{cases} (\mathbb{K}_1)_{ij} = \langle \tilde{B}_1\phi_i, \tilde{D}^{-1}\tilde{B}_1\phi_j \rangle_{\Phi^*,\Phi} = \langle B_1^*D^{-1}B_1\phi_i, \tilde{D}^{-1}B_1^*D^{-1}B_1\phi_j \rangle_{\Phi^*,\Phi} \\ (\mathbb{K}_2)_{jk} = \langle \tilde{B}_2\theta_k, \tilde{D}^{-1}\tilde{B}_1\phi_j \rangle_{\Phi^*,\Phi} = \langle B_1^*D^{-1}B_2\theta_k, \tilde{D}^{-1}B_1^*D^{-1}B_1\phi_j \rangle_{\Phi^*,\Phi} \\ (\tilde{\mathbb{C}})_{k\ell} = \langle \tilde{B}_2\theta_k, \tilde{D}^{-1}\tilde{B}_2\theta_\ell \rangle_{\Phi^*,\Phi} = \langle B_1^*D^{-1}B_2\theta_k, \tilde{D}^{-1}B_1^*D^{-1}B_2\theta_\ell \rangle_{\Phi^*,\Phi}; \end{cases}$$

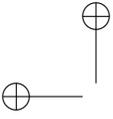
and

$$\begin{cases} (\tilde{\mathbf{g}}_1)_j = \langle \tilde{B}_1\phi_j, \tilde{D}^{-1}\tilde{g}_1 \rangle_{\Phi^*,\Phi} = \langle B_1^*D^{-1}B_1\phi_j, \tilde{D}^{-1}B_1^*D^{-1}g \rangle_{\Phi^*,\Phi} \\ (\tilde{\mathbf{g}}_2)_k = \langle \tilde{B}_2\theta_k, \tilde{D}^{-1}\tilde{g}_2 \rangle_{\Phi^*,\Phi} = \langle B_1^*D^{-1}B_2\theta_k, \tilde{D}^{-1}B_1^*D^{-1}g \rangle_{\Phi^*,\Phi}. \end{cases}$$

Theorem 9. *Let (3), (5), and (13) hold. Then, for $0 < \epsilon \leq 1$, (45) has a unique solution $\phi_\epsilon^h \in \Phi^h$ and $\theta_\epsilon^h \in \Theta^h$. Moreover, if $\phi \in \Phi$ and $\theta \in \Theta$ denotes the unique solution of the optimization problem (6), then there exist a constant $C > 0$ whose value is independent of ϵ and h such that*

$$\begin{aligned} \|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta &\leq C\epsilon \left(\|g\|_{\Lambda^*} + \|\hat{\phi}\|_{\hat{\Phi}} \right) \\ &+ C \left(1 + \frac{1}{\epsilon} \right) \left(\inf_{\tilde{\phi}^h \in \Phi^h} \|\phi_\epsilon - \tilde{\phi}^h\|_\Phi + \inf_{\tilde{\theta}^h \in \Theta^h} \|\theta_\epsilon - \tilde{\theta}^h\|_\Theta \right). \end{aligned} \quad (46)$$

Clearly, (41) and (45) are not the same. However, the coefficient matrices of both systems are symmetric and uniformly (with respect to h) positive definite without regards to (26).



5 Example: Optimization problems for the Stokes system

We use the same concrete examples as in [8]. Consider the velocity-vorticity-pressure formulation of the Stokes system:⁶

$$\left. \begin{aligned} \nabla \times \boldsymbol{\omega} + \nabla p + \boldsymbol{\theta} &= \mathbf{g} \\ \nabla \cdot \mathbf{u} &= 0 \\ \nabla \times \mathbf{u} - \boldsymbol{\omega} &= \mathbf{0} \end{aligned} \right\} \text{ in } \Omega, \quad \mathbf{u} = \mathbf{0} \text{ on } \Gamma, \quad \int_{\Omega} p \, d\Omega = 0 \quad (47)$$

and the functionals

$$\text{Case I: } \mathcal{J}_1(\boldsymbol{\omega}, \boldsymbol{\theta}) = \frac{1}{2} \int_{\Omega} |\boldsymbol{\omega}|^2 \, d\Omega + \frac{\delta}{2} \int_{\Omega} |\boldsymbol{\theta}|^2 \, d\Omega \quad (48)$$

$$\text{Case II: } \mathcal{J}_2(\mathbf{u}, \boldsymbol{\theta}; \hat{\mathbf{u}}) = \frac{1}{2} \int_{\Omega} |\mathbf{u} - \hat{\mathbf{u}}|^2 \, d\Omega + \frac{\delta}{2} \int_{\Omega} |\boldsymbol{\theta}|^2 \, d\Omega, \quad (49)$$

where Ω denotes an open, bounded domain in \mathcal{R}^s , $s = 2$ or 3 , with boundary Γ , \mathbf{u} , $\boldsymbol{\omega}$ and p denote the velocity, vorticity, and pressure fields, respectively, $\boldsymbol{\theta}$ denotes a distributed control, and \mathbf{g} and $\hat{\mathbf{u}}$ are given functions. The optimization problems we study are to find $(\mathbf{u}, \boldsymbol{\omega}, p, \boldsymbol{\theta})$ that minimizes either the functional in (48) or (49), subject to the Stokes system in the form (47) being satisfied.

5.1 Precise statement of optimization problems

We recall the space $L^2(\Omega)$ of all square integrable functions with norm $\|\cdot\|_0$ and inner product (\cdot, \cdot) , the space $L_0^2(\Omega) \equiv \{q \in L^2(\Omega) : \int_{\Omega} p \, d\Omega = 0\}$, the space $H^1(\Omega) \equiv \{v \in L^2(\Omega) : \nabla v \in [L^2(\Omega)]^s\}$, and the space $H_0^1(\Omega) \equiv \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\}$. A norm for functions $v \in H^1(\Omega)$ is given by $\|v\|_1 \equiv (\|\nabla v\|^2 + \|v\|_0^2)^{1/2}$. The dual space of $H_0^1(\Omega)$ is denoted by $H^{-1}(\Omega)$. The inner product in $H^{-1}(\Omega)$ is denoted by $(\cdot, \cdot)_{-1}$. Note that we may define $(\cdot, \cdot)_{-1} = \langle \cdot, (-\Delta)^{-1}(\cdot) \rangle_{\mathbf{H}^{-1}(\Omega), \mathbf{H}^1(\Omega)} = (\cdot, (-\Delta)^{-1}(\cdot))$, where $\Delta : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$ denotes the Laplace operator with respect to Ω with zero Dirichlet boundary conditions along Γ .

The corresponding spaces of vector-valued functions are denoted in bold face, e.g., $\mathbf{H}^1(\Omega) = [H^1(\Omega)]^s$ is the space of vector-valued functions each of whose components belongs to $H^1(\Omega)$. We note the following equivalence of norms [15]:

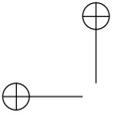
$$\tilde{C}_1 \|\mathbf{v}\|_1^2 \leq \|\nabla \times \mathbf{v}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2 \leq \tilde{C}_2 \|\mathbf{v}\|_1^2 \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega) \quad (50)$$

for some constants $\tilde{C}_1 > 0$ and $\tilde{C}_2 > 0$.

Let $\Phi = \Lambda = \mathbf{H}_0^1(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega)$ and $\Theta = \mathbf{L}^2(\Omega)$ so that $\Phi^* = \Lambda^* = \mathbf{H}^{-1}(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega)$ and $\Theta^* = \mathbf{L}^2(\Omega)$. Let $\hat{\Phi} = \tilde{\Phi} = \mathbf{L}^2(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega)$. Then, $\Phi \subset \hat{\Phi} = \tilde{\Phi} = \hat{\Phi}^* \subset \Phi^*$. For $\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi$, we define the norm

$$\|\phi\|_{\Phi} = (\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2)^{1/2}$$

⁶The reasons for using velocity-vorticity-pressure formulation of the Stokes equations instead of, say, the standard primitive variable formulation, are discussed in [8].



and likewise for the other product spaces.

We make the associations of

$$\begin{array}{lll}
 \text{trial functions:} & \phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi, & \theta = \{\boldsymbol{\theta}\} \in \Theta, & \lambda = \{\mathbf{v}, \boldsymbol{\sigma}, q\} \in \Lambda, \\
 \text{test functions:} & \mu = \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\} \in \hat{\Phi}, & \nu = \{\tilde{\boldsymbol{\theta}}\} \in \hat{\Theta}, & \psi = \{\tilde{\mathbf{v}}, \tilde{\boldsymbol{\sigma}}, \tilde{r}\} \in \Lambda, \\
 \text{data:} & g = \{\mathbf{g}, \mathbf{0}, 0\} \in \Lambda^*, & \hat{\phi} = \left\{ \left(\begin{array}{c} \mathbf{0} \\ \hat{\mathbf{u}} \end{array} \right), \mathbf{0}, 0 \right\} \in \hat{\Phi} & \begin{array}{l} \text{for Case I} \\ \text{for Case II.} \end{array}
 \end{array}$$

We next define the bilinear forms

$$a_1(\phi, \mu) = \begin{cases} \langle \tilde{\boldsymbol{\omega}}, \boldsymbol{\omega} \rangle & \text{for Case I} \\ \langle \tilde{\mathbf{u}}, \mathbf{u} \rangle & \text{for Case II} \end{cases} \quad \forall \phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \hat{\Phi}, \quad \mu = \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\} \in \hat{\Phi},$$

$$a_2(\theta, \nu) = \delta(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \quad \forall \theta = \{\boldsymbol{\theta}\} \in \Theta, \quad \nu = \{\tilde{\boldsymbol{\theta}}\} \in \hat{\Theta},$$

$$b_1(\phi, \psi) = \langle \boldsymbol{\omega}, \nabla \times \tilde{\mathbf{v}} \rangle - \langle p, \nabla \cdot \tilde{\mathbf{v}} \rangle + \langle \nabla \times \mathbf{u} - \boldsymbol{\omega}, \tilde{\boldsymbol{\sigma}} \rangle - \langle \nabla \cdot \mathbf{u}, \tilde{r} \rangle \\ \forall \phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \hat{\Phi}, \quad \psi = \{\tilde{\mathbf{v}}, \tilde{\boldsymbol{\sigma}}, \tilde{r}\} \in \Lambda,$$

$$b_2(\theta, \psi) = \langle \boldsymbol{\theta}, \tilde{\mathbf{v}} \rangle \quad \forall \theta = \{\boldsymbol{\theta}\} \in \Theta, \quad \psi = \{\tilde{\mathbf{v}}, \tilde{\boldsymbol{\sigma}}, \tilde{r}\} \in \Lambda.$$

For $\mathbf{g} \in \mathbf{H}^{-1}(\Omega)$, we also define the linear functional

$$\langle g, \psi \rangle_{\Lambda^*, \Lambda} = \langle \mathbf{g}, \tilde{\mathbf{v}} \rangle_{\mathbf{H}^{-1}(\Omega), \mathbf{H}_0^1(\Omega)} \quad \forall \psi = \{\tilde{\mathbf{v}}, \tilde{\boldsymbol{\sigma}}, \tilde{r}\} \in \Lambda.$$

The operators associated with the bilinear forms are then

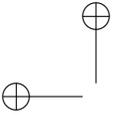
$$\begin{aligned}
 A_1 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for Case I,} & A_1 &= \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for Case II,} \\
 A_2 &= \delta I, & B_1 &= \begin{pmatrix} 0 & \nabla \times & \nabla \\ \nabla \times & -I & 0 \\ -\nabla \cdot & 0 & 0 \end{pmatrix}, & \text{and} & B_2 &= \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix}.
 \end{aligned} \tag{51}$$

Note that $B_1^* = B_1$.

It is now easily seen that the functionals $\mathcal{J}_1(\cdot, \cdot)$ and $\mathcal{J}_2(\cdot, \cdot; \cdot)$ defined in (48) and (49), respectively, can be written in the form (2). Likewise, the Stokes system (47) can be written in the form (4). Thus, the two optimization problems for the Stokes system can both be written in the form (6), with $\mathcal{J}(\cdot, \cdot)$ being either $\mathcal{J}_1(\cdot, \cdot)$ or $\mathcal{J}_2(\cdot, \cdot)$ as appropriate.

Additional notation relevant to example problems

In §§3 and 4, additional bilinear forms, linear functionals, operators, and functions are introduced. In the context of the optimization problems considered in this



section, the bilinear forms (17), (18), (27), (11), and (34) are respectively given by

$$\begin{aligned}
\tilde{b}_1(\{\mathbf{u}, \boldsymbol{\omega}, p\}, \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\}) &= \left(\nabla \times \boldsymbol{\omega} + \nabla p, \nabla \times \tilde{\boldsymbol{\omega}} + \nabla \tilde{p} \right)_{-1} \\
&\quad + \left(\nabla \times \mathbf{u} - \boldsymbol{\omega}, \nabla \times \tilde{\mathbf{u}} - \tilde{\boldsymbol{\omega}} \right) + \left(\nabla \cdot \mathbf{u}, \nabla \cdot \tilde{\mathbf{u}} \right) \\
&= \left(\nabla \times \boldsymbol{\omega} + \nabla p, (-\Delta)^{-1}(\nabla \times \tilde{\boldsymbol{\omega}} + \nabla \tilde{p}) \right) \\
&\quad + \left(\nabla \times \mathbf{u} - \boldsymbol{\omega}, \nabla \times \tilde{\mathbf{u}} - \tilde{\boldsymbol{\omega}} \right) + \left(\nabla \cdot \mathbf{u}, \nabla \cdot \tilde{\mathbf{u}} \right), \\
\tilde{b}_2(\{\boldsymbol{\theta}, \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\}\}) &= \left(\boldsymbol{\theta}, \nabla \times \tilde{\boldsymbol{\omega}} + \nabla \tilde{p} \right)_{-1} = \left(\boldsymbol{\theta}, (-\Delta)^{-1}(\nabla \times \tilde{\boldsymbol{\omega}} + \nabla \tilde{p}) \right), \\
c(\{\boldsymbol{\theta}\}, \{\tilde{\boldsymbol{\theta}}\}) &= \left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \right)_{-1} = \left(\tilde{\boldsymbol{\theta}}, (-\Delta)^{-1}\boldsymbol{\theta} \right), \\
d(\{\mathbf{u}, \boldsymbol{\omega}, p\}, \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\}) &= (\mathbf{u}, \tilde{\mathbf{u}})_{-1} + (\boldsymbol{\omega}, \tilde{\boldsymbol{\omega}}) + (p, \tilde{p}) \\
&= ((-\Delta)^{-1}\tilde{\mathbf{u}}, \mathbf{u}) + (\boldsymbol{\omega}, \tilde{\boldsymbol{\omega}}) + (p, \tilde{p}), \\
\tilde{d}(\{\mathbf{u}, \boldsymbol{\omega}, p\}, \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\}) &= d(\{\mathbf{u}, \boldsymbol{\omega}, p\}, \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\}).
\end{aligned}$$

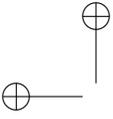
We also have the linear functionals

$$\begin{aligned}
\langle \tilde{g}_1, \{\tilde{\mathbf{u}}, \tilde{\boldsymbol{\omega}}, \tilde{p}\} \rangle_{\Phi^*, \Phi} &= \left(\mathbf{g}, \nabla \times \tilde{\boldsymbol{\omega}} + \nabla \tilde{p} \right)_{-1} = \left(\mathbf{g}, (-\Delta)^{-1}(\nabla \times \tilde{\boldsymbol{\omega}} + \nabla \tilde{p}) \right) \\
&= \left(\nabla \times (-\Delta)^{-1}\mathbf{g}, \tilde{\boldsymbol{\omega}} \right) - \left(\nabla \cdot (-\Delta)^{-1}\mathbf{g}, \tilde{p} \right) \\
\langle \tilde{g}_2, \{\tilde{\boldsymbol{\theta}}\} \rangle_{\Theta^*, \Theta} &= \left(\mathbf{g}, \tilde{\boldsymbol{\theta}} \right)_{-1} = \left(\mathbf{g}, (-\Delta)^{-1}\tilde{\boldsymbol{\theta}} \right) = \left((-\Delta)^{-1}\mathbf{g}, \tilde{\boldsymbol{\theta}} \right)
\end{aligned}$$

and the operators and functions

$$\begin{aligned}
\tilde{B}_1 &= \begin{pmatrix} \nabla \times \nabla \times -\nabla \nabla \cdot & -\nabla \times & 0 \\ -\nabla \times & I + \nabla \times (-\Delta)^{-1}\nabla \times & \nabla \times (-\Delta)^{-1}\nabla \\ 0 & -\nabla \cdot (-\Delta)^{-1}\nabla \times & -\nabla \cdot (-\Delta)^{-1}\nabla \end{pmatrix}, \\
\tilde{B}_2 &= \begin{pmatrix} 0 \\ \nabla \times (-\Delta)^{-1} \\ -\nabla \cdot (-\Delta)^{-1} \end{pmatrix}, \quad D = \tilde{D} = \begin{pmatrix} -\Delta & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, \\
\tilde{g}_1 &= \begin{pmatrix} \mathbf{0} \\ \nabla \times (-\Delta)^{-1}\mathbf{g} \\ -\nabla \cdot (-\Delta)^{-1}\mathbf{g} \end{pmatrix}, \quad \text{and} \quad \tilde{g}_2 = (-\Delta)^{-1}\mathbf{g}.
\end{aligned}$$

With these definitions, all of the methods discussed in the preceding sections can be defined for the optimization problems considered in this section.



Verification of hypotheses

In [8], the following results are proved.

Proposition 10. *Let the spaces Φ , $\widehat{\Phi}$, Θ , and Λ and the bilinear forms $a_1(\cdot, \cdot)$, $a_2(\cdot, \cdot)$, $b_1(\cdot, \cdot)$, and $b_2(\cdot, \cdot)$ be defined as above. Then, the assumptions (3) and (5) are satisfied with $C_1 = 1$, $C_2 = \delta$, $K_2 = \delta$, $c_1 = 3$, $c_2 = 1$, and $k_1 = \min\{1, \widetilde{C}_1\} / \left(2\sqrt{\max\left\{\frac{1}{\widetilde{C}_1}, \widetilde{C}_2\right\}}\right)$.*

We also easily have the following result.

Proposition 11. *Let the spaces Φ and Λ and the bilinear forms $d(\cdot, \cdot)$ and $\widetilde{d}(\cdot, \cdot)$ be defined as above. Then, (13) and (34) are satisfied with $c_d = \widetilde{c}_d = 1$ and $k_d = \widetilde{k}_d = \frac{1}{2} \min\{1, \sigma\}$, where σ denotes the smallest eigenvalue of $-\Delta$ with respect to Ω with zero Dirichlet boundary conditions along Γ .*

Having verified the assumptions (3), (5), (13), and (34) in the context of the two optimization problems for the Stokes system, we have that all the results of §§2–4 will apply to those systems.

5.2 Least-squares formulation of the constraint equations

Using the associations of spaces and variables defined in §5.1 as well as the operators defined there, it is easy to see that the least-squares functional (14) is given by, for the example problems we are considering,

$$\begin{aligned} \mathcal{K}(\{\mathbf{u}, \boldsymbol{\omega}, p\}; \boldsymbol{\theta}, \mathbf{g}) &= \|\nabla \times \boldsymbol{\omega} + \nabla p + \boldsymbol{\theta} - \mathbf{g}\|_{-1}^2 + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2 \\ &= (\nabla \times \boldsymbol{\omega} + \nabla p + \boldsymbol{\theta} - \mathbf{g}, (-\Delta)^{-1}(\nabla \times \boldsymbol{\omega} + \nabla p + \boldsymbol{\theta} - \mathbf{g})) \\ &\quad + \|\nabla \times \mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2. \end{aligned} \quad (52)$$

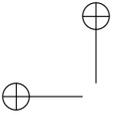
The Euler-Lagrange equation corresponding to the minimization of the least-squares functional (52) is given by (16), which now takes the form: for given $\boldsymbol{\theta} \in \mathbf{L}^2(\Omega)$ and $\mathbf{g} \in \mathbf{H}^{-1}(\Omega)$, find $\{\mathbf{u}, \boldsymbol{\omega}, p\} \in \mathbf{H}_0^1(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega)$ such that

$$\begin{aligned} \widetilde{b}_1(\{\mathbf{u}, \boldsymbol{\omega}, p\}, \{\widetilde{\mathbf{u}}, \widetilde{\boldsymbol{\omega}}, \widetilde{p}\}) &= \langle \widetilde{g}_1, \{\widetilde{\mathbf{u}}, \widetilde{\boldsymbol{\omega}}, \widetilde{p}\} \rangle_{\Phi^*, \Phi} - \widetilde{b}_2(\{\boldsymbol{\theta}, \{\widetilde{\mathbf{u}}, \widetilde{\boldsymbol{\omega}}, \widetilde{p}\}\}) \\ \forall \{\widetilde{\mathbf{u}}, \widetilde{\boldsymbol{\omega}}, \widetilde{p}\} &\in \mathbf{H}_0^1(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega). \end{aligned} \quad (53)$$

We then have the following results; see [5, 6].

Proposition 12. *Let the spaces Φ , $\widehat{\Phi}$, Θ , and Λ and the bilinear forms $a_1(\cdot, \cdot)$, $a_2(\cdot, \cdot)$, $b_1(\cdot, \cdot)$, and $b_2(\cdot, \cdot)$ be defined as in §5.1. Let $\mathcal{K}(\cdot, \cdot, \cdot; \cdot, \cdot)$ be defined by (52). Then, we have the norm equivalence result*

$$\gamma_1(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2) \leq \mathcal{K}(\mathbf{u}, \boldsymbol{\omega}, p; \mathbf{0}, \mathbf{0}) \leq \gamma_2(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2) \quad (54)$$



for some constants $\gamma_1, \gamma_2 > 0$. Moreover, the bilinear form $\tilde{b}_1(\{\cdot, \cdot, \cdot\}, \{\cdot, \cdot, \cdot\})$ defined in §5.1 is symmetric, continuous, and coercive and the problem (53) has a unique solution; that solution is the unique minimizer of the least-squares functional (52).

Proof. The results follow immediately from Lemma 4 and Propositions 3 and 10. \square

To define least-squares finite element approximations of the constraint equations, we first choose conforming finite element subspaces $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, $\mathbf{W}^h \subset \mathbf{L}^2(\Omega)$, and $S^h \subset L_0^2(\Omega)$. We then minimize the functional in (52) over the subspaces, or equivalently, solve the problem: for given $\boldsymbol{\theta} \in \mathbf{L}^2(\Omega)$ and $\mathbf{g} \in \mathbf{H}^{-1}(\Omega)$, find $\{\mathbf{u}^h, \boldsymbol{\omega}^h, p^h\} \in \mathbf{V}^h \times \mathbf{W}^h \times S^h$ such that

$$\begin{aligned} & \tilde{b}_1(\{\mathbf{u}^h, \boldsymbol{\omega}^h, p^h\}, \{\tilde{\mathbf{u}}^h, \tilde{\boldsymbol{\omega}}^h, \tilde{p}^h\}) \\ &= \langle \tilde{g}_1, \{\tilde{\mathbf{u}}^h, \tilde{\boldsymbol{\omega}}^h, \tilde{p}^h\} \rangle_{\Phi^*, \Phi} - \tilde{b}_2(\{\boldsymbol{\theta}, \{\tilde{\mathbf{u}}^h, \tilde{\boldsymbol{\omega}}^h, \tilde{p}^h\}\}) \\ & \quad \forall \{\tilde{\mathbf{u}}^h, \tilde{\boldsymbol{\omega}}^h, \tilde{p}^h\} \in \mathbf{V}^h \times \mathbf{W}^h \times S^h. \end{aligned} \quad (55)$$

We then have the following results.

Proposition 13. *Let $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, $\mathbf{W}^h \subset \mathbf{L}^2(\Omega)$, and $S^h \subset L_0^2(\Omega)$. Then, the discrete problem (55) has a unique solution $\{\mathbf{u}^h, \boldsymbol{\omega}^h, p^h\} \in \mathbf{V}^h \times \mathbf{W}^h \times S^h$. Let $\{\mathbf{u}, \boldsymbol{\omega}, p\} \in \mathbf{H}_0^1(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega)$ denote the unique solution of (53). Then,*

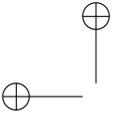
$$\begin{aligned} & \|\mathbf{u} - \mathbf{u}^h\|_1 + \|\boldsymbol{\omega} - \boldsymbol{\omega}^h\|_0 + \|p - p^h\|_0 \\ & \leq C \left(\inf_{\tilde{\mathbf{u}}^h \in \mathbf{V}^h} \|\mathbf{u} - \tilde{\mathbf{u}}^h\|_1 + \inf_{\tilde{\boldsymbol{\omega}}^h \in \mathbf{W}^h} \|\boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}^h\|_0 + \inf_{\tilde{p}^h \in S^h} \|p - \tilde{p}^h\|_0 \right). \end{aligned} \quad (56)$$

Proof. The results follow in a straightforward manner from Proposition 12. \square

5.3 Discrete systems for the Stokes control problem

One can also, in a straightforward manner, use the notation introduced in §5.1 to define the concrete discrete systems that correspond to the methods introduced in §4. Here, we consider, in the context of the Stokes equations, the discrete system defined by (41).

Let $\{\mathbf{u}_j\}_{j=1}^{J_1}$, $\{\boldsymbol{\omega}_j\}_{j=1}^{J_2}$, $\{p_j\}_{j=1}^{J_3}$, and $\{\boldsymbol{\theta}_k\}_{k=1}^K$ respectively denote basis sets for the finite element spaces $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, $\mathbf{W}^h \subset \mathbf{L}^2(\Omega)$, $S^h \subset L_0^2(\Omega)$, and $\boldsymbol{\Theta}^h \subset \mathbf{L}^2(\Omega)$, where $\boldsymbol{\Theta}^h$ is the finite element space introduced to approximate the control $\boldsymbol{\theta}$. Then, using the definitions given in §5.1 for the associated bilinear forms and linear



functionals, the matrices and vectors appearing in (41) are given by:

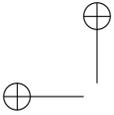
$$\mathbb{A}_1 = \begin{pmatrix} \mathbb{A}_{1,11} & 0 & 0 \\ 0 & \mathbb{A}_{1,22} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \tilde{\mathbb{B}}_1 = \begin{pmatrix} \tilde{\mathbb{B}}_{1,11} & \tilde{\mathbb{B}}_{1,12} & 0 \\ \tilde{\mathbb{B}}_{1,12}^T & \tilde{\mathbb{B}}_{1,22} & \tilde{\mathbb{B}}_{1,23} \\ 0 & \tilde{\mathbb{B}}_{1,23}^T & \tilde{\mathbb{B}}_{1,33} \end{pmatrix}, \quad \tilde{\mathbb{B}}_2 = \begin{pmatrix} 0 \\ \tilde{\mathbb{B}}_{2,21} \\ \tilde{\mathbb{B}}_{2,31} \end{pmatrix},$$

$$\tilde{\mathbb{D}} = \begin{pmatrix} \tilde{\mathbb{D}}_{,11} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathbb{I} \end{pmatrix}, \quad \vec{\mathbf{f}} = \begin{pmatrix} \vec{\mathbf{f}}_{,1} \\ \vec{\mathbf{0}} \\ \vec{\mathbf{0}} \end{pmatrix}, \quad \text{and} \quad \vec{\mathbf{g}}_1 = \begin{pmatrix} \vec{\mathbf{0}} \\ \vec{\mathbf{g}}_{1,2} \\ \vec{\mathbf{g}}_{1,3} \end{pmatrix},$$

where

$$\begin{aligned} (\mathbb{A}_{1,11})_{ij} &= \begin{cases} 0 & \text{(Case I)} \\ (\mathbf{u}_j, \mathbf{u}_i) & \text{(Case II)} \end{cases} && \text{for } i, j = 1, \dots, J_1, \\ (\mathbb{A}_{1,22})_{ij} &= \begin{cases} (\boldsymbol{\omega}_j, \boldsymbol{\omega}_i) & \text{(Case I)} \\ 0 & \text{(Case II)} \end{cases} && \text{for } i, j = 1, \dots, J_2, \\ (\mathbb{A}_2)_{k\ell} &= \delta(\boldsymbol{\theta}_\ell, \boldsymbol{\theta}_k) && \text{for } k, \ell = 1, \dots, K, \\ (\tilde{\mathbb{B}}_{1,11})_{ij} &= (\nabla \times \mathbf{u}_j, \nabla \times \mathbf{u}_i) + (\nabla \cdot \mathbf{u}_j, \nabla \cdot \mathbf{u}_i) && \text{for } i, j = 1, \dots, J_1, \\ (\tilde{\mathbb{B}}_{1,12})_{ij} &= -(\boldsymbol{\omega}_j, \nabla \times \mathbf{u}_i) && \text{for } i = 1, \dots, J_1, j = 1, \dots, J_2, \\ (\tilde{\mathbb{B}}_{1,22})_{ij} &= (\nabla \times \boldsymbol{\omega}_j, \nabla \times \boldsymbol{\omega}_i)_{-1} + (\boldsymbol{\omega}_j, \boldsymbol{\omega}_i) && \text{for } i, j = 1, \dots, J_2, \\ (\tilde{\mathbb{B}}_{1,23})_{ij} &= (\nabla p_j, \nabla \times \boldsymbol{\omega}_i) && \text{for } i = 1, \dots, J_2, j = 1, \dots, J_3, \\ (\tilde{\mathbb{B}}_{1,33})_{ij} &= (\nabla p_j, \nabla p_i)_{-1} && \text{for } i, j = 1, \dots, J_3, \\ (\tilde{\mathbb{B}}_{2,21})_{jk} &= (\boldsymbol{\theta}_k, \nabla \times \boldsymbol{\omega}_j) && \text{for } j = 1, \dots, J_2, k = 1, \dots, K, \\ (\tilde{\mathbb{B}}_{2,31})_{jk} &= (\boldsymbol{\theta}_k, \nabla p_j) && \text{for } j = 1, \dots, J_3, k = 1, \dots, K, \\ (\tilde{\mathbb{D}}_{,11})_{ij} &= (\nabla \mathbf{u}_j, \nabla \mathbf{u}_i) && \text{for } i, j = 1, \dots, J_1, \\ (\vec{\mathbf{f}}_{,1})_i &= \begin{cases} \vec{\mathbf{0}} & \text{(Case I)} \\ (\hat{\mathbf{u}}, \mathbf{u}_i) & \text{(Case II)} \end{cases} && \text{for } i = 1, \dots, J_1, \\ (\vec{\mathbf{g}}_{1,2})_i &= (\mathbf{g}, \nabla \times \boldsymbol{\omega}_i)_{-1} && \text{for } i = 1, \dots, J_2, \\ (\vec{\mathbf{g}}_{1,3})_i &= (\mathbf{g}, \nabla p_i)_{-1} && \text{for } i = 1, \dots, J_3. \end{aligned}$$

In this way, the matrix system (41) is completely defined.



Theorem 8 implies that, if one uses continuous finite element spaces of degree r for all variables and if one chooses $\epsilon = h^r$, then

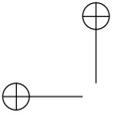
$$\|\mathbf{u} - \mathbf{u}_\epsilon^h\|_1 + \|\boldsymbol{\omega} - \boldsymbol{\omega}_\epsilon^h\|_0 + \|p - p_\epsilon^h\|_0 + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\epsilon^h\|_0 = O(h^r)$$

provided that the solution of the optimization problems we are considering in this section are sufficiently smooth. One could use finite element spaces of one degree lower for the approximations of the vorticity $\boldsymbol{\omega}$, the pressure p , and the control $\boldsymbol{\theta}$ than that used for the velocity \mathbf{u} and still obtain the same error estimate. However, one of the strengths of using least-squares finite element methods is that one can use any conforming finite element spaces and, in particular, one can use the same degree finite element spaces for all variables; we see that the methods introduced here inherit this strength since they do not require the satisfaction of the discrete stability conditions in (26).

5.4 Some practical issues arising in implementations

One difficulty that arises in the implementation discussed in §5.3 is caused by the appearance of the $\mathbf{H}^{-1}(\Omega)$ norm in the least-squares functional (52). For example, this leads to the appearance of the $\mathbf{H}^{-1}(\Omega)$ inner product in the definition of the matrices and vectors that form the discrete system; see §5.3. The equivalence relation $(\cdot, \cdot)_{-1} = (\cdot, (-\Delta)^{-1}\cdot)$ is not of much help since, in general, one cannot exactly invert the Laplace operator, even in the case of zero Dirichlet boundary conditions. Fortunately, there are several approaches available for ameliorating this difficulty; these are discussed in [8]; see also [3, 11, 12]. All the approaches discussed in [8] can be applied to the methods introduced in this paper, with similar comparative effectiveness; thus, here, we do not consider this issue any further.

A second issue that needs to be discussed is the conditioning of the discrete systems. Actually, there are two issues here, i.e., the conditioning with respect to either h as $h \rightarrow 0$ or with respect to ϵ as $\epsilon \rightarrow 0$. First, let's discuss the $h \rightarrow 0$ issue. Least-squares finite element methods typically result in a “squaring” of operators, e.g., the normal equations in the linear algebra context. This is clearly indicated in (22) and (23) where one sees that the operator \tilde{B}_1 that results from applying the least-squares principle (15) to the constraint equations involves the product of the operators B_1^* and B_1 . It is well known that “squaring” operators can result in the squaring of the condition number of the corresponding matrices one obtains after discretization. This is the principal reason for using first-order formulations of the constraint equations, as was done for the Stokes equations in §5.1. The idea here is that after “squaring” first-order operators, one obtains a second-order operator so that the h -condition number of the resulting squared system is hopefully similar to that for Galerkin formulations of second-order equations. However, penalty formulations of optimal control problems can result in a second “squaring” of operators. For example, look at (37); we see there operators such as $B_1^* D^{-1} B_1 \tilde{D}^{-1} B_1^* D^{-1} B_1$ which involves four copies of the operator B_1 . However, that is not the whole story; that operator also involves two copies of the operator D^{-1} and also the operator \tilde{D}^{-1} . Given the nature of all these operators, it is not



at all clear that the h -condition number of the discrete systems of §§4 and 5.3 are similar to those that result from a naive double “squaring” of first-order operators; indeed, norm equivalence relations such as (21) and (54) can sometimes be used to show that h -condition numbers for least-squares-based methods are no worse than those for Galerkin-based methods.

The situation regarding the conditioning of the discrete systems as $\epsilon \rightarrow 0$ is problematic for all penalty methods, even for those for which locking does not occur. Note that to obtain a result such as (44), one chooses $\epsilon = h^\alpha$; with such a choice, ϵ is likely to be small. This situation can be greatly ameliorated by introducing an *iterated* penalty method; see, e.g., [14] and also [13, 16, 17]. To this end, let $\{\vec{\phi}_\epsilon, \vec{\theta}_\epsilon, \vec{\mu}_\epsilon\}$ denote the solution of (40) and set $\vec{\phi}^{(0)} = \vec{\phi}_\epsilon$, $\vec{\theta}^{(0)} = \vec{\theta}_\epsilon$, and $\vec{\mu}^{(0)} = \vec{\mu}_\epsilon$. Then, for $n \geq 1$, we solve the sequence of problems

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \tilde{\mathbb{B}}_1 \\ 0 & \mathbb{A}_2 & \tilde{\mathbb{B}}_2^T \\ \tilde{\mathbb{B}}_1 & \tilde{\mathbb{B}}_2 & -\epsilon\tilde{\mathbb{D}} \end{pmatrix} \begin{pmatrix} \vec{\phi}^{(n)} \\ \vec{\theta}^{(n)} \\ \vec{\mu}^{(n)} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \\ -\epsilon\tilde{\mathbb{D}}\vec{\mu}^{(n-1)} \end{pmatrix}. \quad (57)$$

Then, for any $N > 0$, we let

$$\vec{\phi}_{\epsilon,N} = \sum_{n=0}^N \vec{\phi}^{(n)}, \quad \vec{\theta}_{\epsilon,N} = \sum_{n=0}^N \vec{\theta}^{(n)}, \quad \text{and} \quad \vec{\mu}_{\epsilon,N} = \sum_{n=0}^N \vec{\mu}^{(n)} \quad (58)$$

and we let $\phi_{\epsilon,N}^h \in \Phi^h$, $\theta_{\epsilon,N}^h \in \Theta^h$, and $\mu_{\epsilon,N}^h \in \Phi^h$ be the finite element functions corresponding to the coefficients collected in the respective vectors in (58). Then, instead of the estimate (43), one obtains the estimate (see, e.g., [14] and also [13, 16])

$$\|\phi - \phi_{\epsilon,N}^h\|_\Phi + \|\theta - \theta_{\epsilon,N}^h\|_\Theta + \|\mu - \mu_{\epsilon,N}^h\|_\Phi \leq C(\epsilon^{N+1} + h^\alpha + h^\beta)$$

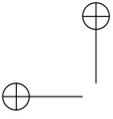
so that if $\beta \geq \alpha$ and one chooses $\epsilon = h^{\alpha/N+1}$, one obtains the optimal error estimate

$$\|\phi - \phi_{\epsilon,N}^h\|_\Phi + \|\theta - \theta_{\epsilon,N}^h\|_\Theta + \|\mu - \mu_{\epsilon,N}^h\|_\Phi \leq C\epsilon^{N+1} = Ch^\alpha.$$

instead of (44). These estimates tell us that we make the error due to penalization as small as we want in two ways: we can choose either ϵ sufficiently small or N sufficiently large. Making the former choice, e.g., choosing $N = 0$ and $\epsilon = h^\alpha$, can lead to conditioning problems for the discrete systems since $\epsilon \ll 1$. Making the latter choice allows us obtain the same effect but with a much larger value for ϵ .

Note that $\vec{\mu}^{(n)}$ may be eliminated from (57) to yield a reduced system with fewer unknowns. Thus, the iteration to compute the pairs $\{\vec{\phi}^{(n)}, \vec{\theta}^{(n)}\}$ for $n = 0, 1, \dots$, using reduced systems proceeds as follows. Let $\vec{\phi}^{(0)} = \vec{\phi}_\epsilon$ and $\vec{\theta}^{(0)} = \vec{\theta}_\epsilon$, where $\vec{\phi}_\epsilon$ and $\vec{\theta}_\epsilon$ denote the solution of (41), and then set

$$\vec{\mathbf{g}}^{(0)} = \tilde{\mathbb{B}}_1 \vec{\phi}^{(0)} + \tilde{\mathbb{B}}_2 \vec{\theta}^{(0)} - \vec{\mathbf{g}}_1.$$



Then, for $n = 1, 2, \dots$, solve the systems

$$\begin{cases} \left(\mathbb{A}_1 + \frac{1}{\epsilon} \tilde{\mathbb{B}}_1 \tilde{\mathbb{D}}^{-1} \tilde{\mathbb{B}}_1 \right) \vec{\phi}^{(n)} + \frac{1}{\epsilon} \tilde{\mathbb{B}}_1 \tilde{\mathbb{D}}^{-1} \tilde{\mathbb{B}}_2 \vec{\theta}^{(n)} = \frac{1}{\epsilon} \tilde{\mathbb{B}}_1 \tilde{\mathbb{D}}^{-1} \vec{\mathbf{g}}^{(n-1)} \\ \left(\mathbb{A}_2 + \frac{1}{\epsilon} \tilde{\mathbb{B}}_2^T \tilde{\mathbb{D}}^{-1} \tilde{\mathbb{B}}_2 \right) \vec{\phi}^{(n)} + \frac{1}{\epsilon} \tilde{\mathbb{B}}_2^T \tilde{\mathbb{D}}^{-1} \tilde{\mathbb{B}}_1 \vec{\theta}^{(n)} = \frac{1}{\epsilon} \tilde{\mathbb{B}}_2^T \tilde{\mathbb{D}}^{-1} \vec{\mathbf{g}}^{(n-1)}. \end{cases}$$

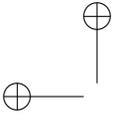
In order to define the next iterate, we set

$$\vec{\mathbf{g}}^{(n)} = \vec{\mathbf{g}}^{(n-1)} + \tilde{\mathbb{B}}_1 \vec{\phi}^{(n)} + \tilde{\mathbb{B}}_2 \vec{\theta}^{(n)}.$$

6 Conclusions

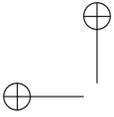
In this paper, a new least-squares method for optimization and control problems was formulated and analyzed. Instead of the more direct approach of penalizing the cost functional by least-squares terms, they are used to reformulate the constraint. This leads to a bilevel minimization problem with a number of attractive computational properties.

Most notably, the new method preserves the desirable features of least-squares methods such as being able to use discretization spaces for the state variables that are not subject to inf-sup stability conditions. The optimality system of the bilevel optimization problem can be solved by penalty methods without locking; this opens up possibilities for the design of more efficient solution methods for optimization and control problems constrained by PDEs.



Bibliography

- [1] D. BEDIVAN AND G. J. FIX, *Least-squares methods for optimal shape design problems*, Computers Math. Applic., 30(2), pp. 17-25, 1995.
- [2] P. BOCHEV, *Least-squares methods for optimal control*, Nonlinear Analysis, Theory, Methods and Applications, 30(3), pp. 1875-1885, 1997.
- [3] P. BOCHEV, Negative norm least-squares methods for the velocity-vorticity-pressure Navier-Stokes equations, *Numerical Methods in PDE's*, 15 (1999) pp. 237-256.
- [4] P. BOCHEV AND D. BEDIVAN, *Least-squares methods for navier-Stokes boundary control problems*, Int. J. Comp. Fluid Dynamics, 9, 1997, pp. 43-58.
- [5] P. BOCHEV AND M. GUNZBURGER, Least-squares finite element methods for elliptic equations, *SIAM Review*, 40/4 (1998) pp. 789-837.
- [6] P. BOCHEV AND M. GUNZBURGER, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479-506.
- [7] P. BOCHEV AND M. GUNZBURGER, *Least-squares finite element methods for optimization and control problems for the Stokes equations*, to appear in Comp. Math. Appl.
- [8] P. BOCHEV AND M. GUNZBURGER, *Least-squares finite element methods for optimality systems arising in optimization and control problems*, to appear in SIAM J. Num. Anal..
- [9] P. BOCHEV AND M. GUNZBURGER, *On least-squares variational principles for the discretization of optimization and control problems*, submitted.
- [10] D. BRAESS, *Finite Elements*, Cambridge, Cambridge, 1997.
- [11] J. BRAMBLE, R. LAZAROV, AND J. PASCIAK, *A least squares approach based on a discrete minus one inner product for first order systems*, Technical Report 94-32, Mathematical Science Institute, Cornell University, 1994.
- [12] J. BRAMBLE AND J. PASCIAK, Least-squares methods for Stokes equations based on a discrete minus one inner product, *J. Comp. App. Math.*, 74 (1996) pp. 155-173.



- [13] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
- [14] M. FORTIN AND A. FORTIN, *A generalization of Uzawa's algorithm for the solution of the Navier-Stokes equations*, Appl. Numer. Meth. 1, 1985, pp. 205–208.
- [15] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
- [16] M. GUNZBURGER, *Iterative penalty methods for the Stokes and Navier-Stokes equations*, in Finite Element Analysis in Fluids, University of Alabama, Huntsville, 1989, pp. 1040–1045.
- [17] M. GUNZBURGER, *Finite Element Methods for Viscous Incompressible Flows*, Academic, Boston, (1989).
- [18] M. GUNZBURGER AND H. C. LEE, *Analysis and approximation of optimal control problems for first-order elliptic systems in three dimensions*, Appl. Math. Comp. **100**, 1999, pp. 49–70.
- [19] M. GUNZBURGER AND H. C. LEE, *A penalty/least-squares method for optimal control problems for first-order elliptic systems*, Appl. Math. Comp. **107**, 2000, pp. 57–75.
- [20] H. SCHLICHTING AND K. GERSTEN, *Boundary Layer Theory*, Springer, Berlin, 2000.