METHODS AND APPLICATIONS OF ANALYSIS.
Vol. 12, No. 4, pp. 395–426, December 2005

003

ON LEAST-SQUARES VARIATIONAL PRINCIPLES FOR THE DISCRETIZATION OF OPTIMIZATION AND CONTROL PROBLEMS*

PAVEL B. BOCHEV† AND MAX D. GUNZBURGER‡

Abstract. The approximate solution of optimization and control problems for systems governed by linear, elliptic partial differential equations is considered. Such problems are most often solved using methods based on the application of the Lagrange multiplier rule followed by discretization through, e.g., a Galerkin finite element method. As an alternative, we show how least-squares finite element methods can be used for this purpose. Penalty-based formulations, another approach widely used in other settings, have not enjoyed the same level of popularity in the partial differential equation case perhaps because naively defined penalty-based methods can have practical deficiencies. We use methodologies associated with modern least-squares finite element methods to develop and analyze practical penalty methods for the approximate solution of optimization problems for systems governed by linear, elliptic partial differential equations. We develop an abstract theory for such problems; along the way, we introduce several methods based on least-squares notions, and compare and constrast their properties.

Key words. Optimal control, optimization, least-squares methods, penalty methods, finite element methods

AMS subject classifications. 65N30, 65N22, 49J20, 49K20

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 [23]. 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
- 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, constrained optimization problems. It is not surprising then that several popular approaches to solving optimization and control problems constrained by partial differen-

*Received August 10, 2005; accepted for publication November 7, 2005.

†Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque NM 87185-1110, USA (pbboche@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, Florida State University, Tallahassee FL 32306-4120, USA (gunzburg@csit.fsu.edu). Supported in part by CSRI, Sandia National Laboratories under contract 18407 and by the National Science Foundation under grant number DMS-0308845.

tial equations are based on solving optimality systems deduced from the application of the Lagrange multiplier rule. In these approaches, Galerkin weak forms of the partial differential equation constraints are used. In the finite element method context, these Galerkin variational formulations are usually used as the basis for defining discretizations; see, e.g., [14, 17, 20] for descriptions this approach. Another means for solving the optimality system is to apply least-squares finite element methods; see [8] and also [21].

Instead of constraining the cost functional with a Galerkin weak form of the constraint equations, one can constrain with a least-squares minimization form of the constraints. This leads to a different optimality system that has advantages over using the Galerkin form of the constraints. This approach was considered in [9].

Penalty methods, which are another popular approach for finite-dimensional optimization problems, have not generated much interest for the infinite-dimensional problems which are of interest here. In this paper, we will see why naively defined penalty methods may not be practical and how, using methodologies developed in modern least-squares finite element methods, the penalty approach can be rehabilitated to yield practical and efficient algorithms for optimal control problems. 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. Least-squares/penalty methods have been considered, in concrete settings, in [1, 2, 4, 7, 22].

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 Galerkin and least-squares finite element methods for the approximate solution of the constraint equations. In §4, we consider the use of the Lagrange multiplier rule for deriving an optimality system whose solution is also a solution of the control problem; we also consider Galerkin and least-squares finite element methods for finding approximate solutions of the optimality system. In §§5 and 6, we define and analyze several penalty-based methods for the approximate solution of the abstract control problem of §2. In §5, we begin by directly penalizing the cost functional of the optimal control problem by the least-squares functional; in §6, we begin by constraining the cost functional by the least-squares functional. The two approaches lead to different discrete systems. Methods that result from the approach of §5, which is the more common way to define penalty methods, are not as effective as those resulting from the approach of §6. In the former case, one has methods that either require the satisfaction of discrete stability conditions or are prone to locking. In the latter case, one can define a method that avoids both of these undesirable features. In §7, we critically compare several theoretical properties of the methods; we then briefly discuss some practical issues that also affect the choice of method.

2. The model optimization and control problem. 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 $\widetilde{\Phi}$ acts as the pivot space for both the pairs $\{\Phi^*, \Phi\}$ and $\{\widehat{\Phi}^*, \widehat{\Phi}\}$ so that we not only have that $\Phi \subseteq \widehat{\Phi} \subseteq \widetilde{\Phi} \subseteq \widehat{\Phi}^* \subseteq \Phi^*$, but also

$$\langle \psi, \phi \rangle_{\Phi^*,\Phi} = \langle \psi, \phi \rangle_{\widehat{\Phi}^*,\widehat{\Phi}} = (\psi, \phi)_{\widetilde{\Phi}} \quad \forall\, \psi \in \widehat{\Phi}^* \subseteq \Phi^* \quad \text{and} \quad \forall\, \phi \in \Phi \subseteq \widehat{\Phi}\,, \qquad (2.1)$$

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

$$\mathcal{J}(\phi,\theta) = \frac{1}{2}a_1(\phi-\widehat{\phi},\phi-\widehat{\phi}) + \frac{1}{2}a_2(\theta,\theta) \qquad \forall\,\phi\in\Phi,\ \theta\in\Theta\,, \tag{2.2}$$

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 (2.2) can be interpreted as a penalty term[1] which limits the size of the control θ.

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

$$\begin{cases} a_1(\phi,\mu) \le C_1\|\phi\|_{\widehat{\Phi}}\|\mu\|_{\widehat{\Phi}} & \forall\,\phi,\mu\in\widehat{\Phi} \\ a_2(\theta,\nu) \le C_2\|\theta\|_{\Theta}\|\nu\|_{\Theta} & \forall\,\theta,\nu\in\Theta \\ a_1(\phi,\phi) \ge 0 & \forall\,\phi\in\widehat{\Phi} \\ a_2(\theta,\theta) \ge K_2\|\theta\|_{\Theta}^2 & \forall\,\theta\in\Theta\,, \end{cases} \tag{2.3}$$

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*[2]

$$b_1(\phi,\psi) + b_2(\theta,\psi) = \langle g,\psi\rangle_{\Lambda^*,\Lambda} \quad \forall\,\psi\in\Lambda\,. \tag{2.4}$$

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

$$\begin{cases} b_1(\phi,\psi) \le c_1\|\phi\|_{\Phi}\|\psi\|_{\Lambda} & \forall\,\phi\in\Phi,\ \ \psi\in\Lambda \\ b_2(\theta,\psi) \le c_2\|\theta\|_{\Theta}\|\psi\|_{\Lambda} & \forall\,\theta\in\Theta,\ \ \psi\in\Lambda \\ \displaystyle\sup_{\psi\in\Lambda,\psi\neq 0}\frac{b_1(\phi,\psi)}{\|\psi\|_{\Lambda}} \ge k_1\|\phi\|_{\Phi} & \forall\,\phi\in\Phi \\ \displaystyle\sup_{\phi\in\Phi,\phi\neq 0}\frac{b_1(\phi,\psi)}{\|\phi\|_{\Phi}} > 0 & \forall\,\psi\in\Lambda\,, \end{cases} \tag{2.5}$$

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

We consider the *optimal control problem*

$$\min_{(\phi,\theta)\in\Phi\times\Theta} \mathcal{J}(\phi,\theta) \quad \text{subject to} \quad b_1(\phi,\psi) + b_2(\theta,\psi) = \langle g,\psi\rangle_{\Lambda^*,\Lambda} \quad \forall\,\psi\in\Lambda\,. \tag{2.6}$$

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

THEOREM 2.1. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, the optimal control problem* (2.6) *has a unique solution* $(\phi,\theta)\in\Phi\times\Theta$.

[1]The usage of the terminology "penalty term" in conjunction with the second term in (2.2) should not be confused with the usage of that terminology below.

[2]One should view (2.4) as a Galerkin weak form of the given partial differential equation constraint, i.e., of the operator constraint equation (2.9). In fact, one usually formulates the partial differential equation constraint in the operator form (2.9) and then derives a (Galerkin) weak formulation of the form (2.4).

It is instructive to rewrite the functional (2.2), the constraint (2.4), and the optimal control problem (2.6) in operator notation. To this end, we note that the bilinear forms serve to define operators

$$\begin{array}{lll} A_1 : \widehat{\Phi} \to \widehat{\Phi}^*, & A_2 : \Theta \to \Theta^*, & B_1 : \Phi \to \Lambda^*, \\ B_1^* : \Lambda \to \Phi^*, & B_2 : \Theta \to \Lambda^*, & B_2^* : \Lambda \to \Theta^* \end{array}$$

through the relations

$$\begin{array}{ll} a_1(\phi,\mu) = \langle A_1\phi, \mu\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} & \forall\, \phi, \mu \in \widehat{\Phi} \\ a_2(\theta,\nu) = \langle A_2\theta, \nu\rangle_{\Theta^*,\Theta} & \forall\, \theta, \nu \in \Theta \\ b_1(\phi,\psi) = \langle B_1\phi, \psi\rangle_{\Lambda^*,\Lambda} = \langle B_1^*\psi, \phi\rangle_{\Phi^*,\Phi} & \forall\, \phi \in \Phi,\ \ \psi \in \Lambda \\ b_2(\psi,\theta) = \langle B_2\theta, \psi\rangle_{\Lambda^*,\Lambda} = \langle B_2^*\psi, \theta\rangle_{\Theta^*,\Theta} & \forall\, \theta \in \Theta,\ \ \psi \in \Lambda\,. \end{array} \tag{2.7}$$

Then, the functional (2.2) and the constraint (2.4) respectively take the forms

$$\mathcal{J}(\phi,\theta) = \frac{1}{2}\big\langle A_1(\phi - \widehat{\phi}), (\phi - \widehat{\phi})\big\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} + \frac{1}{2}\langle A_2\theta, \theta\rangle_{\Theta^*,\Theta} \qquad \forall\, \phi \in \Phi,\, \theta \in \Theta \tag{2.8}$$

and

$$B_1\phi + B_2\theta = g \qquad \text{in } \Lambda^* \tag{2.9}$$

and the optimal control problem (2.6) takes the form

$$\min_{(\phi,\theta)\in\Phi\times\Theta} \mathcal{J}(\phi,\theta) \qquad \text{subject to} \qquad B_1\phi + B_2\theta = g \quad \text{in } \Lambda^*. \tag{2.10}$$

Assumptions (2.3) and (2.5) imply that A_1, A_2, B_1, B_2, B_1^*, and B_2^* are bounded with

$$\begin{array}{lll} \|A_1\|_{\widehat{\Phi}\to\widehat{\Phi}^*} \le C_1, & \|A_2\|_{\Theta\to\Theta^*} \le C_2, & \|B_1\|_{\Phi\to\Lambda^*} \le c_1, \\ \|B_1^*\|_{\Lambda\to\Phi^*} \le c_1, & \|B_2\|_{\Theta\to\Lambda^*} \le c_2, & \|B_2^*\|_{\Lambda\to\Theta^*} \le c_2 \end{array}$$

and that the operator B_1 is invertible with $\|B_1^{-1}\|_{\Lambda^*\to\Phi} \le 1/k_1$. See [8] for details.

3. Galerkin and least-squares finite element methods for the constraint equations. The constraint equations are given by (2.9), or in equivalent variational form by (2.4). We consider two finite element approaches for finding approximations of solutions of the constraint equations. The first is a direct discretization of the weak formulation (2.4); the second is based on a reformulation of the constraint equation into a least-square variational principle. Throughout this section, we assume that not only the data function $g \in \Lambda^*$ but also the control function $\theta \in \Theta$ are given and that we wish to determine the corresponding sate $\phi \in \Phi$ satisfying (2.9), or, equivalently, (2.4). In subsequent sections, we will again consider the optimization or control problem problem (2.6) or, equivalently, (2.10), for which the control $\theta \in \Theta$ as well as the state $\phi \in \Phi$ are unknown.

For the constraint equation (2.4), we have the following well-known result; see [13, 14, 17].

PROPOSITION 3.1. *Let the assumptions* (2.5) *hold. Then, given* $\theta \in \Theta$ *and* $g \in \Lambda^*$, (2.4) *has a unique solution* $\phi \in \Phi$. *Moreover, we have that*

$$\|\phi\|_\Phi \le C\big(\|\theta\|_\Theta + \|g\|_{\Lambda^*}\big)\,.$$

Thus, we see that (2.5) are sufficient to guarantee that the constraint equation are solvable for a state $\phi \in \Phi$ for any control $\theta \in \Theta$. Note that, in terms of operators, we have that $\phi = B^{-1}(g - B_2\theta)$. Note also that the operator A_2 is invertible by virtue of (2.3).

3.1. Galerkin finite element methods for the constraint equations. We consider finite element discretizations of the constraint equation (2.4). To this end, we choose (conforming) families of finite-dimensional subspaces $\Phi^h \subset \Phi$ and $\Lambda^h \subset \Lambda$ and then restrict (2.4) to the subspaces, i.e., given $\theta \in \Theta$ and $g \in \Lambda^*$, we seek $\phi^h \in \Phi^h$ that satisfies

$$b_1(\phi^h, \psi^h) + b_2(\theta, \psi^h) = \langle g, \psi^h \rangle_{\Lambda^*,\Lambda} \quad \forall\, \psi^h \in \Lambda^h\,. \tag{3.1}$$

It is well known (see, e.g., [13, 14, 17]) that in order to guarantee that (3.1) is stably solvable, it is not enough to require that (2.5) hold; one must additionally assume that[3]

$$\begin{cases} \displaystyle\sup_{\psi^h \in \Lambda^h, \psi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\psi^h\|_\Lambda} \geq k_1^h \|\phi^h\|_\Phi & \forall\, \phi^h \in \Phi^h \\ \displaystyle\sup_{\phi^h \in \Phi^h, \phi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\phi^h\|_V} > 0 & \forall\, \psi^h \in \Lambda^h\,, \end{cases} \tag{3.2}$$

where k_1^h is a positive constant whose value is independent of h.

PROPOSITION 3.2. *Let the assumptions* (2.5) *and* (3.2) *hold. Then, for any* $\theta \in \Theta$ *and* $g \in \Lambda^*$, (3.1) *has a unique solution* $\phi^h \in \Phi^h$. *Moreover,*

$$\|\phi^h\|_\Phi \leq C\big(\|g\|_{\Lambda^*} + \|\theta\|_\Theta\big)\,.$$

Furthermore, let $\phi \in \Phi$ *denote the unique solution of* (2.4). *Then,*

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

Let $\{\phi_j\}_{j=1}^J$ and $\{\lambda_m\}_{m=1}^M$ denote bases for Φ^h and Λ^h, respectively. Then, the problem (3.1) is equivalent to the matrix problem

$$\mathbb{B}_1 \vec{\boldsymbol{\phi}} = \vec{\mathbf{g}}_0\,,$$

where $\vec{\boldsymbol{\phi}}$ is the vector of coefficients for ϕ^h,

$$(\mathbb{B}_1)_{ij} = b_1(\phi_i, \psi_j) = \langle B_1\phi_i, \psi_j \rangle_{\Lambda^*,\Lambda} \qquad \text{for } i, j = 1, \ldots, J,$$

and

$$\big(\vec{\mathbf{g}}_0\big)_i = \langle g, \psi_i \rangle_{\Lambda^*,\Lambda} - b_2(\theta, \psi_i) = \langle g - B_2\theta, \psi_i \rangle_{\Lambda^*,\Lambda} \qquad \text{for } i = 1, \ldots, J.$$

The assumption (3.2) guarantee that $\mathbb{B}_1$ is a square, invertible matrix.

[3]One of the main motivations for defining least-squares finite element methods for problems of the type (2.4) is to develop discretization methods that do not require the imposition of the discrete stability conditions (3.2).

3.2. Least-squares finite element methods for the constraint equations. The constraint equations are given in variational form in (2.4) and in equivalent operator form in (2.9). The may also be defined through a least-squares minimization problem. Let $D : \Lambda \to \Lambda^*$ be a self-adjoint, strongly coercive operator, i.e., there exist constants $c_d > 0$ and $k_d > 0$ such that

$$\langle D\lambda, \psi\rangle_{\Lambda^*,\Lambda} \le c_d\|\lambda\|_\Lambda\|\psi\|_\Lambda \quad \text{and} \quad \langle D\lambda, \lambda\rangle_{\Lambda^*,\Lambda} \ge k_d\|\lambda\|^2_\Lambda \qquad \forall\, \lambda, \psi \in \Lambda\,. \tag{3.3}$$

Note that then $k_d \le \|D\|_{\Lambda\to\Lambda^*} \le c_d$ and $1/c_d \le \|D^{-1}\|_{\Lambda^*\to\Lambda} \le 1/k_d$. In the sequel, we will also use the induced bilinear form

$$d(\lambda, \psi) = \langle D\lambda, \psi\rangle_{\Lambda^*,\Lambda} \quad \forall\, \lambda, \psi \in \Lambda\,. \tag{3.4}$$

The following results are immediate.

PROPOSITION 3.3. *Assume that the operator D is symmetric and that* (3.3) *holds. Then, the bilinear form $d(\cdot,\cdot)$ is symmetric and*

$$d(\lambda, \psi) \le c_d\|\lambda\|_\Lambda\|\psi\|_\Lambda \quad \forall\, \lambda, \psi \in \Lambda \quad \textit{and} \quad d(\lambda, \lambda) \ge k_d\|\lambda\|^2_\Lambda \quad \forall\, \lambda \in \Lambda\,. \tag{3.5}$$

Let[4,5]

$$\mathcal{K}(\phi; \theta, g) = \left\langle B_1\phi + B_2\theta - g, D^{-1}(B_1\phi + B_2\theta - g)\right\rangle_{\Lambda^*,\Lambda} \quad \forall\, \phi \in \Phi,\ \theta \in \Theta,\ g \in \Lambda^*. \tag{3.6}$$

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

$$\min_{\phi\in\Phi} \mathcal{K}\,(\phi; \theta, g)\,. \tag{3.7}$$

Clearly, this problem is equivalent to (2.4) and (2.9), i.e., solutions of (3.7) are solutions of (2.4) or (2.9) and conversely. The Euler-Lagrange equation corresponding to the problem (3.7) is given, in variational form, by

$$\widetilde{b}_1(\phi, \mu) = \langle \widetilde{g}_1, \mu\rangle_{\Phi^*,\Phi} - \widetilde{b}_2(\theta, \mu) \quad \forall\, \mu \in \Phi\,, \tag{3.8}$$

where

$$\widetilde{b}_1(\phi, \mu) = \left\langle B_1\mu, D^{-1}B_1\phi\right\rangle_{\Lambda^*,\Lambda} = \left\langle B_1^*D^{-1}B_1\phi, \mu\right\rangle_{\Phi^*,\Phi} \qquad \forall\, \phi, \mu \in \Phi \tag{3.9}$$

$$\widetilde{b}_2(\theta, \mu) = \left\langle B_1\mu, D^{-1}B_2\theta\right\rangle_{\Lambda^*,\Lambda} = \left\langle B_1^*D^{-1}B_2\theta, \mu\right\rangle_{\Phi^*,\Phi} \qquad \forall\, \theta \in \Theta,\ \mu \in \Phi \tag{3.10}$$

and

$$\widetilde{g}_1 = B_1^*D^{-1}g \in \Phi^*\,. \tag{3.11}$$

[4]The reason for using D^{-1} and not simply D will become apparent in §5 when we discuss penalty methods.

[5]Let $R : \Lambda \to \Lambda^*$ denote the Reisz operator, i.e., we have that if $\upsilon = R\lambda$ and $\chi = R\psi$ for $\lambda, \psi \in \Lambda$ and $\upsilon, \chi \in \Lambda^*$, then $\|\lambda\|_\Lambda = \|\upsilon\|_{\Lambda^*}$, $\|\psi\|_\Lambda = \|\chi\|_{\Lambda^*}$, and

$$(\psi, \lambda)_\Lambda = < R\psi, \lambda >_{\Lambda^*,\Lambda} = < \chi, R^{-1}\upsilon >_{\Lambda^*,\Lambda} = (\upsilon, \chi)_{\Lambda^*}\,.$$

Then, if one chooses $D = R$, the functional (3.6) reduces to $\mathcal{K}(\phi; \theta, g) = (B_1\phi + B_2\theta - g, B_1\phi + B_2\theta - g)_{\Lambda^*} = \|B_1\phi + B_2\theta - g\|^2_{\Lambda^*}$. Note that, in general, (3.6) can also be written as an inner product, i.e., $\mathcal{K}(\phi; \theta, g) = (B_1\phi + B_2\theta - g, RD^{-1}(B_1\phi + B_2\theta - g))_{\Lambda^*}$.

As is shown in the following proposition, the bilinear forms $\widetilde{b}_1(\cdot,\cdot)$ and $\widetilde{b}_2(\cdot,\cdot)$ are continuous and the former is strongly coercive; see [8, 9] for details.

PROPOSITION 3.4. *Assume that* (2.5) *and* (3.3) *hold. Then, the bilinear form* $\widetilde{b}_1(\cdot,\cdot)$ *is symmetric and there exist positive constants* $\widetilde{c}_1$, $\widetilde{c}_2$, *and* $\widetilde{k}_1$ *such that*

$$\begin{cases} \widetilde{b}_1(\phi,\mu) \le \widetilde{c}_1\|\phi\|_\Phi\|\mu\|_\Phi & \forall\,\phi,\mu\in\Phi \\ \widetilde{b}_2(\theta,\mu) \le \widetilde{c}_2\|\mu\|_\Phi\|\theta\|_\Theta & \forall\,\theta\in\Theta,\ \ \mu\in\Phi \\ \widetilde{b}_1(\phi,\phi) \ge \widetilde{k}_1\|\phi\|_\Phi^2 & \forall\,\phi\in\Phi\,. \end{cases} \tag{3.12}$$

Moreover, $\|\widetilde{g}_1\|_{\Phi^*} \le \frac{c_1}{k_d}\|g\|_{\Lambda^*}$ *and the problem* (3.8), *or equivalently* (3.7), *has a unique solution.*

As an immediate consequence of Proposition 3.4, we have that the least-squares functional (3.6) is norm equivalent in the following sense.

COROLLARY 3.1. *Assume that* (3.3) *and the conditions on the bilinear form* $b_1(\cdot,\cdot)$ *in* (2.5) *hold. Then,*

$$\widetilde{k}_1\|\phi\|_\Phi^2 \le \mathcal{K}(\phi;0,0) = \widetilde{b}_1(\phi,\phi) = \left\langle B_1\phi, D^{-1}B_1\phi\right\rangle_{\Lambda^*,\Lambda} \le \widetilde{c}_1\|\phi\|_\Phi^2 \quad \forall\,\phi\in\Phi\,. \tag{3.13}$$

For all $\mu\in\Phi$, we can the rewrite (3.8) as $\left\langle B_1\mu, D^{-1}(B_1\phi+B_2\theta-g)\right\rangle_{\Lambda^*,\Lambda}=0$ or $\left\langle B_1^*D^{-1}(B_1\phi+B_2\theta-g),\mu\right\rangle_{\Phi^*,\Phi}=0$ so that, in operator form, we have that (3.8) is equivalent to

$$\widetilde{B}_1\phi+\widetilde{B}_2\theta=\widetilde{g}_1 \quad \text{in } \Phi^*, \tag{3.14}$$

where

$$\widetilde{B}_1 = B_1^*D^{-1}B_1 : \Phi\to\Phi^*, \qquad \text{and} \qquad \widetilde{B}_2 = B_1^*D^{-1}B_2 : \Theta\to\Phi^*. \tag{3.15}$$

Note that (3.13) implies that the operator $\widetilde{B}_1 = B_1^*D^{-1}B_1$ in (3.14) is symmetric and coercive even when the operator B_1 in (2.9) is indefinite and/or non-symmetric; these observations, of course, follow from the facts that the bilinear form $b_1(\cdot,\cdot)$ is weakly coercive (see (2.5)) while the bilinear form $\widetilde{b}_1(\cdot,\cdot)$ is strongly coercive (see (3.12)). It is also easy to see that (3.14) has the same solutions as (2.9).

Discretization of (3.8), or equivalently of (3.14), is accomplished in the standard manner. One chooses a subspace $\Phi^h\subset\Phi$ and then, given $\theta\in\Theta$ and $\widetilde{g}\in\Phi^*$, one solves the problem

$$\widetilde{b}_1(\phi^h,\mu^h) = \langle\widetilde{g}_1,\mu^h\rangle_{\Phi^*,\Phi} - \widetilde{b}_2(\theta,\mu^h) \quad \forall\,\mu^h\in\Phi^h\,. \tag{3.16}$$

Then, (3.13) and the Lax-Milgram and Cea lemmas immediately imply the following results.

PROPOSITION 3.5. *Assume that* (2.5) *and* (3.3) *hold. Then, the problem* (3.16) *has a unique solution and, if* ϕ *denotes the solution of the problem* (3.8), *or equivalently, of* (3.14), *there exists a constant* $C>0$ *whose value is independent of* h, ϕ, *and* ϕ^h *such that*

$$\|\phi-\phi^h\|_\Phi \le C\inf_{\widetilde{\phi}^h\in\Phi}\|\phi-\widetilde{\phi}^h\|_\Phi\,.$$

Again, if $\{\phi_j\}_{j=1}^J$ denotes a basis for Φ^h, then the problem (3.16) is equivalent to the matrix problem

$$\widetilde{\mathbb{B}}_1\vec{\boldsymbol{\phi}} = \widetilde{\vec{\mathbf{g}}}_0\,, \tag{3.17}$$

where $\vec{\boldsymbol{\phi}}$ is the vector of coefficients for ϕ^h,

$$(\widetilde{\mathbb{B}}_1)_{ij} = \widetilde{b}_1(\phi_i,\phi_j) = \langle \widetilde{B}_1\phi_i,\phi_j\rangle_{\Phi^*,\Phi} \qquad \text{for } i,j=1,\dots,J,$$

and, for $i=1,\dots,J$,

$$\big(\widetilde{\vec{\mathbf{g}}}_0\big)_i = \langle \widetilde{g}_1,\phi_i\rangle_{\Phi^*,\Phi} - \widetilde{b}_2(\theta,\phi_i) = \langle \widetilde{g}_1 - \widetilde{B}_2\theta,\phi_i\rangle_{\Phi^*,\Phi} = \langle B_1^*D^{-1}g - B_1^*D^{-1}B_2\theta\,,\,\phi_i\rangle_{\Phi^*,\Phi}.$$

The following result follows easily from Proposition 3.4 and Corollary 3.1.

COROLLARY 3.2. *Assume that* (3.3) *and the conditions on the bilinear form* $b_1(\cdot,\cdot)$ *in* (2.5) *hold. Then, the matrix* $\widetilde{\mathbb{B}}_1$ *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 (2.9) are that the matrix $\widetilde{\mathbb{B}}_1$ in (3.17) is symmetric and positive definite even when the operator B_1 in (2.9) 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 such as (3.2).[6] In incorporating the least-squares formalism into the optimization setting of §2, we want to preserve these advantages.

4. Solution of the optimization problem via Lagrange multipliers. For all $\{\mu,\nu,\psi\} \in \Phi\times\Theta\times\Lambda$, we introduce the Lagrangian functional

$$\begin{aligned}\mathcal{L}(\{\mu,\nu\},\{\psi\}) &= \mathcal{J}(\{\mu,\nu\}) + b(\{\mu,\nu\},\{\psi\}) - \langle g,\psi\rangle_{\Lambda^*,\Lambda}\\ &= \frac{1}{2}a_1(\mu-\widehat{\phi},\mu-\widehat{\phi}) + \frac{1}{2}a_2(\nu,\nu) + b_1(\mu,\psi) + b_2(\nu,\psi) - \langle g,\psi\rangle_{\Lambda^*,\Lambda}\,.\end{aligned}$$

Then, (2.6) is equivalent to the unconstrained optimization problem of finding saddle points $\{\phi,\theta,\lambda\} \in \Phi\times\Theta\times\Lambda$ of the Lagrangian functional. These saddle points may be found by solving the *optimality system*, i.e., the first-order necessary conditions

$$\left\{\begin{array}{lcccccll} a_1(\phi,\mu) & & & + & b_1(\mu,\lambda) & = & a_1(\widehat{\phi},\mu) & \forall\,\mu\in\Phi\\ & & a_2(\theta,\nu) & + & b_2(\nu,\lambda) & = & 0 & \forall\,\nu\in\Theta\\ b_1(\phi,\psi) & + & b_2(\theta,\psi) & & & = & \langle g,\psi\rangle_{\Lambda^*,\Lambda} & \forall\,\psi\in\Lambda\,. \end{array}\right. \tag{4.1}$$

The third equation in the optimality system (4.1) is simply the constraint equation. The first equation is commonly referred to as the *adjoint* or *co-state equation* and the Lagrange multiplier λ is referred as the *adjoint* or *co-state* variable. The second equation in (4.1) is referred to as the *optimality condition* since it is merely a statement that the gradient of the functional $\mathcal{J}(\cdot,\cdot)$ defined in (2.2) vanishes at the optimum.

The following result is proved in [8].

THEOREM 4.1. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, the optimality system* (4.1) *has a unique solution* $(\phi,\theta,\lambda) \in \Phi\times\Theta\times\Lambda$. *Moreover*

$$\|\phi\|_\Phi + \|\theta\|_\Theta + \|\lambda\|_\Lambda \le C\big(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\big)$$

[6]The direct, conforming Galerkin finite element discretization considered in §3.1 requires that that discrete stability conditions be satisfied.

and $(\phi, \theta) \in \Phi \times \Theta$ *is the unique solution of the optimal control problem* (2.6).

Using the operators introduced in (2.7), the optimality system (4.1) takes the form

$$\left\{\begin{array}{lllllll} A_1\phi & & & + & B_1^*\lambda & = & A_1\widehat{\phi} \quad \text{in } \Phi^* \\ & & A_2\theta & + & B_2^*\lambda & = & 0 \quad \text{in } \Theta^* \\ B_1\phi & + & B_2\theta & & & = & g \quad \text{in } \Lambda^*\,. \end{array}\right. \tag{4.2}$$

In analogy to the discussion of §3 concerning the discretization of the constraint equation, we consider Galerkin and least-squares finite element methods for finding approximate solutions of the optimality system (4.1). We note that (4.2) is a type of "nested" saddle-point problem that is quite different from that considered in, e.g., [15]

4.1. Galerkin finite element methods for the optimality system. We choose (conforming) finite dimensional subspaces $\Phi^h \subset \Phi$, $\Theta^h \subset \Theta$, and $\Lambda^h \subset \Lambda$ and then restrict (4.1) to the subspaces, i.e., we seek $(\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$ that satisfies

$$\left\{\begin{array}{llll} a_1(\phi^h, \mu^h) & & +b_1(\mu^h, \lambda^h) & = a_1(\widehat{\phi}, \mu^h) \qquad \forall\, \mu^h \in \Phi^h \\ & a_2(\theta^h, \nu^h) & +b_2(\nu^h, \lambda^h) & = 0 \qquad \forall\, \nu^h \in \Theta^h \\ b_1(\phi^h, \psi^h) & +b_2(\theta^h, \psi^h) & & = \langle g, \psi^h\rangle_{\Lambda^*,\Lambda} \qquad \forall\, \psi^h \in \Lambda^h\,. \end{array}\right. \tag{4.3}$$

This is also the optimality system for the minimization of (2.2) over $\Phi^h \times \Theta^h$ subject to the constraint $b_1(\phi^h, \psi^h) + b_2(\psi^h, \theta^h) = \langle g, \psi^h\rangle_{\Lambda^*,\Lambda}$ for all $\psi^h \in \Lambda^h$. The assumptions (2.3) and (2.5) are not sufficient to guarantee that the discrete optimality system (4.3) is solvable. Again, we must assume that the discrete stability conditions (3.2) on the bilinear form $b_1(\cdot,\cdot)$ hold. In this case, we have the following result which is again proved in [8].

THEOREM 4.2. *Let the assumptions* (2.3), (2.5), *and* (3.2) *hold. Then, the discrete optimality system* (4.3) *has a unique solution* $(\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$ *and moreover*

$$\|\phi^h\|_\Phi + \|\theta^h\|_\Theta + \|\lambda^h\|_\Lambda \le C\big(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\big)\,.$$

Furthermore, let $(\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda$ *denote the unique solution of the optimality system* (4.1), *or, equivalently, of the optimal control problem* (2.6). *Then,*

$$\begin{aligned} &\|\phi - \phi^h\|_\Phi + \|\theta - \theta^h\|_\Theta + \|\lambda - \lambda^h\|_\Lambda \\ &\qquad \le C\Big(\inf_{\mu^h\in\Phi^h}\|\phi - \mu^h\|_\Phi + \inf_{\xi^h\in\Theta^h}\|\theta - \xi^h\|_\Theta + \inf_{\psi^h\in\Lambda^h}\|\lambda - \psi^h\|_\Lambda\Big)\,. \end{aligned}$$

In the usual way, the discrete optimality system (4.3) is equivalent to a matrix problem. Let $\{\phi_j\}_{j=1}^J$, $\{\theta_k\}_{k=1}^K$, and $\{\lambda_m\}_{m=1}^M$, where $J = \dim(\Phi^h)$, $K = \dim(\Theta^h)$, and $M = \dim(\Lambda^h)$, denote chosen basis sets for Φ^h, Θ^h, and Λ^h, respectively. We then define the matrices

$$\left\{\begin{array}{ll} (\mathbb{A}_1)_{ij} = a_1(\phi_i, \phi_j) & \text{for } i, j = 1, \ldots, J \\ (\mathbb{A}_2)_{k\ell} = a_2(\theta_k, \theta_\ell) & \text{for } k, \ell = 1, \ldots, K \\ (\mathbb{B}_1)_{mj} = b_1(\phi_j, \lambda_m) & \text{for } j = 1, \ldots, J,\ m = 1, \ldots, M \\ (\mathbb{B}_2)_{km} = b_2(\theta_k, \lambda_m) & \text{for } k = 1, \ldots, K,\ m = 1, \ldots, M \end{array}\right.$$

and the vectors

$$\begin{cases} (\vec{\mathbf{f}})_j = a_1(\widehat{\phi}, \phi_j) & \text{for } j = 1, \ldots, J \\ (\vec{\mathbf{g}})_m = \langle g, \lambda_m \rangle_{\Lambda^*, \Lambda} & \text{for } m = 1, \ldots, M\,. \end{cases}$$

We then have that the problem (4.3) is equivalent to the matrix problem

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \mathbb{B}_1^T \\ 0 & \mathbb{A}_2 & \mathbb{B}_2^T \\ \mathbb{B}_1 & \mathbb{B}_2 & 0 \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}} \\ \vec{\boldsymbol{\theta}} \\ \vec{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \\ \vec{\mathbf{g}} \end{pmatrix}, \tag{4.4}$$

where $\vec{\boldsymbol{\phi}}$, $\vec{\boldsymbol{\theta}}$, and $\vec{\boldsymbol{\lambda}}$ are the vector of coefficients for ϕ^h, θ^h, and λ^h, respectively.

REMARK 4.1. The discrete optimality system (4.3) or its matrix equivalent (4.4), have the typical saddle point structure. This remains true even if the state equations involve a strongly coercive bilinear form $b_1(\cdot, \cdot)$ so that the last two inequalities in (2.5) can be replaced by $b_1(\phi, \phi) \geq k_1 \|\phi\|_\Phi^2$ for all $\phi \in \Phi$. If the assumptions (2.3) and (2.5) hold, then the stability and convergence properties associated with solutions of (4.3) or (4.4) hold merely by assuming that (3.2) holds for the bilinear form $b_1(\cdot, \cdot)$ and the spaces Φ^h and Λ^h. Thus, these properties depend solely on the ability to stably solve, given any discrete control variable, the discrete state equation for a discrete state variable. On the other hand, if (3.2) does not hold, then (4.3) or its matrix equivalent (4.4) may not be solvable, i.e., the coefficient matrix in (4.4) may not be invertible. In fact, the assumptions (3.2) imply that $\mathbb{B}_1$ is uniformly invertible. This, and the facts (which follow from (2.3)) that the symmetric matrices $\mathbb{A}_1$ and $\mathbb{A}_2$ are positive semi-definite and positive definite, respectively, is enough to guarantee that the coefficient matrix in (4.4) is invertible. On the other hand, if (3.2) does not hold so that the matrix $\mathbb{B}_1$ has a nontrivial null space, then, under the other assumptions that have been made, one cannot guarantee the invertibility of the coefficient matrix in (4.4). See [8] for details.

REMARK 4.2. Solving the discrete optimality system (4.3), or equivalently, the linear system (4.4), is often a formidable task. If the constraint equations (2.4) are a system of partial differential equations, then the last (block) row of (4.4) represents a Galerkin finite element discretization of that system. The discrete adjoint equations, i.e., the first row in (4.4), are also a discretization of a system of partial differential equations. Moreover, the dimension of the discrete adjoint vector $\vec{\boldsymbol{\lambda}}$ is essentially the same as that of discrete state vector $\vec{\boldsymbol{\phi}}$. Thus, (4.4) is at least twice the size (we have yet to account for the discrete control variables in $\vec{\boldsymbol{\theta}}$) of the discrete system corresponding to the discretization of the partial differential equation constraints. Thus, if these equations are difficult to approximate, the discrete optimality system will be even more difficult to deal with. For this reason, there have been many approaches suggested for uncoupling the three components of discrete optimality systems such as (4.3), or equivalently, (4.4). See, e.g., [20], for a discussion of several of these approaches. We note that these approaches rely on the invertibility of the matrices $\mathbb{B}_1$ and $\mathbb{A}_2$, properties that follow from (3.2) and (2.3), respectively.

4.2. Least-squares finite element methods for the optimality system. Even if the state equation (2.4) (or (2.9)) involves a symmetric, positive definite operator B_1, i.e., even if the bilinear form $b_1(\cdot, \cdot)$ is symmetric and strongly coercive, the

discrete optimality system (4.3) (or (4.4)) obtained through a Galerkin discretization is indefinite. For example, if $B_1 = -\Delta$ with zero boundary conditions, then $\mathbb{B}_1$ is a symmetric, positive definite matrix, but the coefficient matrix in (4.4) is indefinite. In order to obtain a discrete optimality system that is symmetric and positive definite, we will apply a least-squares finite element discretization. In fact, these desirable properties for the discrete system will remain in place even if the state system bilinear form $b_1(\cdot,\cdot)$ is only weakly coercive, i.e., even if the operator B_1 is merely invertible but not necessarily positive definite.

Given a system of partial differential equations, there are many ways to define least-squares finite element methods for determining approximate solutions. Practicality issues can be used to select the "best" methods from among the many choices available. See, e.g., [5] for a discussion of what factors enter into the choice of a particular least-squares finite element method for a given problem. Here, we will consider the most straightforward means for defining a least-squares finite element method.

4.2.1. A least-squares finite element method for a generalized optimality system. We start with the generalized form of the optimality system (4.2) written in operator form, i.e.,

$$\left\{\begin{array}{lllllll} A_1\phi & & & + & B_1^*\lambda & = & f \quad \text{in } \Phi^* \\ & & A_2\theta & + & B_2^*\lambda & = & s \quad \text{in } \Theta^* \\ B_1\phi & + & B_2\theta & & & = & g \quad \text{in } \Lambda^*, \end{array}\right. \tag{4.5}$$

where $(f,s,g) \in \Phi^* \times \Theta^* \times \Lambda^*$ is a general data triple and $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$ is the corresponding solution triple. In the same way that Theorem 4.1 is proved, we have the following result.

PROPOSITION 4.1. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, for any* $(f,s,g) \in \Phi^* \times \Theta^* \times \Lambda^*$*, the generalized optimality system* (4.5) *has a unique solution* $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$*. Moreover,*

$$\|\phi\|_\Phi + \|\theta\|_\Theta + \|\lambda\|_\Lambda \le C\big(\|f\|_{\Phi^*} + \|s\|_{\Theta^*} + \|g\|_{\Lambda^*}\big)\,. \tag{4.6}$$

A least-squares functional can be defined by summing the squares of the norms of the residuals of the three equations in (4.5) to obtain

$$\mathcal{K}(\phi,\theta,\lambda;f,s,g) = \|A_1\phi + B_1^*\lambda - f\|_{\Phi^*}^2 + \|A_2\theta + B_2^*\lambda - s\|_{\Theta^*}^2 + \|B_1\phi + B_2\theta - g\|_{\Lambda^*}^2\,. \tag{4.7}$$

Clearly, the unique solution of (4.5) is also the solution of the problem

$$\min_{(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda} \mathcal{K}(\phi,\theta,\lambda;f,s,g)\,. \tag{4.8}$$

The first-order necessary conditions corresponding to (4.8) are easily found to be

$$B\big((\phi,\theta,\lambda),(\mu,\nu,\psi)\big) = F\big((\mu,\nu,\psi);(f,s,g)\big) \quad \forall\,(\mu,\nu,\psi) \in \Phi \times \Theta \times \Lambda\,, \tag{4.9}$$

where

$$\begin{aligned} B\big((\phi,\theta,\lambda),(\mu,\nu,\psi)\big) &= (A_1\mu + B_1^*\psi, A_1\phi + B_1^*\lambda)_{\Phi^*} \\ &\quad + (A_2\nu + B_2^*\psi, A_2\theta + B_2^*\lambda)_{\Theta^*} + (B_1\mu + B_2\nu, B_1\phi + B_2\theta)_{\Lambda^*} \\ &\qquad \forall\,(\phi,\theta,\lambda),\ (\mu,\nu,\psi) \in \Phi \times \Theta \times \Lambda \end{aligned} \tag{4.10}$$

and

$$F\big((\mu,\nu,\psi);(f,s,g)\big) = (A_1\mu + B_1^*\psi, f)_{\Phi^*} + (A_2\nu + B_2^*\psi, s)_{\Theta^*} + (B_1\mu + B_2\nu, g)_{\Lambda^*} \quad \forall\,(\mu,\nu,\psi) \in \Phi\times\Theta\times\Lambda\,. \tag{4.11}$$

The following result is proved in [8].

LEMMA 4.1. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, the bilinear form* $B(\cdot,\cdot)$ *is symmetric and continuous on* $(\Phi\times\Theta\times\Lambda)\times(\Phi\times\Theta\times\Lambda)$ *and the linear functional* $F(\cdot)$ *is continuous on* $(\Phi\times\Theta\times\Lambda)$. *Moreover, the bilinear form* $B(\cdot,\cdot)$ *is coercive on* $(\Phi\times\Theta\times\Lambda)$, *i.e.,*

$$B\big((\phi,\theta,\lambda),(\phi,\theta,\lambda)\big) \ge C(\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2) \quad \forall\,(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda\,. \tag{4.12}$$

REMARK 4.3. Since

$$\begin{aligned}\mathcal{K}(\phi,\theta,\lambda;0,0,0) &= \|A_1\phi + B_1^*\lambda\|_{\Phi^*}^2 + \|A_2\theta + B_2^*\lambda\|_{\Theta^*}^2 + \|B_1\phi + B_2\theta\|_{\Lambda^*}^2\\ &= B\big((\phi,\theta,\lambda),(\phi,\theta,\lambda)\big)\,,\end{aligned}$$

the coercivity and continuity of the bilinear form $B(\cdot,\cdot)$ are equivalent to stating that the functional $\mathcal{K}(\phi,\theta,\lambda;0,0,0)$ is norm-equivalent, i.e., that there exist constants $\gamma_1>0$ and $\gamma_2>0$ such that

$$\gamma_1(\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2) \le \mathcal{K}(\phi,\theta,\lambda;0,0,0) \le \gamma_2(\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2) \tag{4.13}$$

for all $(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda$.

The following results follow from Lemma 4.1 and the Lax-Milgram lemma.

PROPOSITION 4.2. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, for any* $(f,s,g)\in\Phi^*\times\Theta^*\times\Lambda^*$, *the problem* (4.9) *has a unique solution* $(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda$. *Moreover, this solution coincides with the solution of the problems* (4.5) *and* (4.8) *and satisfies the estimate* (4.6).

We define a finite element discretization of (4.5) or, equivalently, of (4.9), by choosing conforming finite element subspaces $\Phi^h\subset\Phi$, $\Theta^h\subset\Theta$, and $\Lambda^h\subset\Lambda$ and then requiring that $(\phi^h,\theta^h,\lambda^h)\in\Phi^h\times\Theta^h\times\Lambda^h$ satisfy

$$B\big((\phi^h,\theta^h,\lambda^h),(\mu^h,\nu^h,\psi^h)\big) = F\big((\mu^h,\nu^h,\psi^h);(f,s,g)\big) \quad \forall\,(\mu^h,\nu^h,\psi^h)\in\Phi^h\times\Theta^h\times\Lambda^h\,. \tag{4.14}$$

Note that $(\phi^h,\theta^h,\lambda^h)$ can also be characterized as the solution of the problem

$$\min_{(\phi^h,\theta^h,\lambda^h)\in\Phi^h\times\Theta^h\times\Lambda^h} \mathcal{K}(\phi^h,\theta^h,\lambda^h;f,s,g)\,.$$

The following result follows from Lemma 4.1 and standard finite element analyses.

PROPOSITION 4.3. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, for any* $(f,h,g)\in\Phi^*\times\Theta^*\times\Lambda^*$, *the problem* (4.14) *has a unique solution* $(\phi^h,\theta^h,\lambda^h)\in\Phi^h\times\Theta^h\times\Lambda^h$. *Moreover, we have the optimal error estimate*

$$\begin{aligned}&\|\phi-\phi^h\|_\Phi + \|\theta-\theta^h\|_\Theta + \|\lambda-\lambda^h\|_\Lambda\\ &\quad\le C\Big(\inf_{\widetilde{\phi}^h\in\Phi^h}\|\phi-\widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h\in\Theta^h}\|\theta-\widetilde{\theta}^h\|_\Theta + \inf_{\widetilde{\lambda}^h\in\Lambda^h}\|\lambda-\widetilde{\lambda}^h\|_\Lambda\Big),\end{aligned} \tag{4.15}$$

where $(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda$ *is the unique solution of the problem* (4.9), *or equivalently, of the problems* (4.5) *or* (4.8).

4.2.2. A least-squares finite element method for the optimality system. The results of §4.2.1 easily specialize to the optimality system (4.2). Indeed, letting $f = A_1\widehat{\phi} \in \widehat{\Phi}^* \subset \Phi^*$ and $s = 0$, we have that (4.5) reduces to (4.2). We now have the least-squares functional,

$$\mathcal{K}(\phi,\theta,\lambda;\widehat{\phi},g) = \|A_1\phi+B_1^*\lambda-A_1\widehat{\phi}\|^2_{\Phi^*}+\|A_2\theta+B_2^*\lambda\|^2_{\Theta^*}+\|B_1\phi+B_2\theta-g\|^2_{\Lambda^*}\,, \quad (4.16)$$

the minimization problem

$$\min_{(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda} \mathcal{K}(\phi,\theta,\lambda;\widehat{\phi},g)\,, \quad (4.17)$$

the first-order necessary conditions

$$B\big((\phi,\theta,\lambda),(\mu,\nu,\psi)\big) = F\big((\mu,\nu,\psi);(A_1\widehat{\phi},0,g)\big) \quad \forall\,(\mu,\nu,\psi)\in\Phi\times\Theta\times\Lambda\,, \quad (4.18)$$

where $B(\cdot,\cdot)$ and $F(\cdot)$ are defined as in (4.10) and (4.11), respectively.

We define a finite element discretization of (4.18) by again choosing conforming finite element subspaces $\Phi^h \subset \Phi$, $\Theta^h \subset \Theta$, and $\Lambda^h \subset \Lambda$ and then requiring that $(\phi^h,\theta^h,\lambda^h) \in \Phi^h\times\Theta^h\times\Lambda^h$ satisfy

$$\begin{aligned} B\big((\phi^h,\theta^h,\lambda^h),(\mu^h,\nu^h,\psi^h)\big) &= F\big((\mu^h,\nu^h,\psi^h);(A_1\widehat{\phi},0,g)\big) \\ &\forall\,(\mu^h,\nu^h,\psi^h)\in\Phi^h\times\Theta^h\times\Lambda^h\,. \end{aligned} \quad (4.19)$$

Then, Proposition 4.3 takes the following form.

THEOREM 4.3. *Let the assumptions* (2.3) *and* (2.5) *hold. Then, for any* $(\widehat{\phi},g)\in\widehat{\Phi}^*\times\Lambda^*$, *the problem* (4.19) *has a unique solution* $(\phi^h,\theta^h,\lambda^h)\in\Phi^h\times\Theta^h\times\Lambda^h$. *Moreover, we have the optimal error estimate: there exists a constant* $C>0$ *whose value is independent of* h, *such that*

$$\begin{aligned} &\|\phi-\phi^h\|_\Phi + \|\theta-\theta^h\|_\Theta + \|\lambda-\lambda^h\|_\Lambda \\ &\quad\le C\Big(\inf_{\widetilde{\phi}^h\in\Phi^h}\|\phi-\widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h\in\Theta^h}\|\theta-\widetilde{\theta}^h\|_\Theta + \inf_{\widetilde{\lambda}^h\in\Lambda^h}\|\lambda-\widetilde{\lambda}^h\|_\Lambda\Big), \end{aligned} \quad (4.20)$$

where $(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda$ *is the unique solution of the problem* (4.18) *or, equivalently, of the problems* (4.2) *or* (4.1). *Note also that* $(\phi,\theta)\in\Phi\times\Theta$ *is the unique solution of the problem* (2.6).

REMARK 4.4. The discrete problem (4.19) is equivalent to the linear algebraic system

$$\begin{pmatrix} \mathbb{K}_1 & \mathbb{C}_1^T & \mathbb{C}_2^T \\ \mathbb{C}_1 & \mathbb{K}_2 & \mathbb{C}_3^T \\ \mathbb{C}_2 & \mathbb{C}_3 & \mathbb{K}_3 \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}} \\ \vec{\boldsymbol{\theta}} \\ \vec{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{h}} \\ \vec{\mathbf{g}} \end{pmatrix}. \quad (4.21)$$

Indeed, if one chooses bases $\{\mu_j^h(\mathbf{x})\}_{j=1}^J$, $\{\nu_k^h(\mathbf{x})\}_{k=1}^K$, and $\{\psi_\ell^h(\mathbf{x})\}_{\ell=1}^L$ for Φ_h, Θ_h, and Λ_h, respectively, we then have $\phi^h = \sum_{j=1}^J \phi_j\mu_j^h$, $\theta^h = \sum_{k=1}^K \theta_k\mu_k^h$, and $\lambda^h =$

$\sum_{\ell=1}^{L} \lambda_\ell \psi_\ell^h$ for some sets of coefficients $\{\phi_j\}_{j=1}^{J}$, $\{\theta_k\}_{k=1}^{K}$, and $\{\lambda_\ell\}_{\ell=1}^{L}$ that are determined by solving (4.21). In (4.21), we have that $\vec{\boldsymbol{\phi}} = (\phi_1, \ldots, \phi_J)^T$, $\vec{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_K)^T$, $\vec{\boldsymbol{\lambda}} = (\lambda_1, \ldots, \lambda_L)^T$,

$$
\begin{aligned}
\big(\mathbb{K}_1\big)_{ij} &= (A_1\mu_i, A_1\mu_j)_{\Phi^*} + (B_1\mu_i, B_1\mu_j)_{\Lambda^*} && \text{for } i,j = 1,\ldots,J,\\
\big(\mathbb{K}_2\big)_{ik} &= (A_2\nu_i, A_1\nu_k)_{\Theta^*} + (B_2\nu_i, B_2\nu_k)_{\Lambda^*} && \text{for } i,k = 1,\ldots,K,\\
\big(\mathbb{K}_3\big)_{i\ell} &= (B_1^*\psi_i, B_1^*\psi_\ell)_{\Phi^*} + (B_2\psi_i, B_2\psi_\ell)_{\Theta^*} && \text{for } i,\ell = 1,\ldots,L,\\
\big(\mathbb{C}_1\big)_{ij} &= (B_2\nu_i, B_1\mu_j)_{\Lambda^*} && \text{for } i = 1,\ldots,K,\ j = 1,\ldots,J,\\
\big(\mathbb{C}_2\big)_{ij} &= (B_1^*\psi_i, A_1\nu_j)_{\Phi^*} && \text{for } i = 1,\ldots,L,\ j = 1,\ldots,J,\\
\big(\mathbb{C}_3\big)_{ik} &= (B_2^*\psi_i, A_2\nu_k)_{\Theta^*} && \text{for } i = 1,\ldots,L,\ k = 1,\ldots,K,\\
\big(\vec{\mathbf{f}}\big)_i &= (A_1\mu_i, A_1\widehat{\phi})_{\Phi^*} + (B_1\mu_i, g)_{\Lambda^*} && \text{for } i = 1,\ldots,J,\\
\big(\vec{\mathbf{h}}\big)_i &= (B_2\nu_i, g)_{\Lambda^*} && \text{for } i = 1,\ldots,K,\\
\big(\vec{\mathbf{g}}\big)_i &= (B_1^*\psi_i, A_1\widehat{\phi})_{\Phi^*} && \text{for } i = 1,\ldots,L.
\end{aligned}
$$

REMARK 4.5. It easily follows from Lemma 4.1 that the coefficient matrix of (4.21) is *symmetric and positive definite.* This should be compared to the linear system (4.4) that results from a Galerkin finite element discretization of the optimality system (4.1) for which the coefficient matrix is symmetric and *indefinite.*

REMARK 4.6. The stability of the discrete problem (4.19), the convergence and optimal accuracy of the approximate solution $(\phi^h, \theta^h, \lambda^h)$, and the symmetry and positive definiteness of the discrete system (4.21) obtained by the least-squares finite element method follow from the assumptions (2.3) and (2.5) that guarantee the well posedness of the infinite-dimensional optimization problem (2.6) and its corresponding optimality system (4.1). It is important to note that all of these desirable properties of the least-squares finite element method do not require that the bilinear form $b_1(\cdot,\cdot)$ and the finite element spaces Φ^h and Λ^h satisfy the discrete inf-sup conditions (3.2) that are necessary for the well posedness of the Galerkin finite element discretization (4.3) of the optimality system (4.1). In fact, this is why least-squares finite element methods are often an attractive alternative to Galerkin discretizations; see, e.g., [5].

REMARK 4.7. The observations made in Remark 4.2 about the possible need to uncouple the equations in (4.4) hold as well for the linear system (4.21). Uncoupling approaches for (4.4) rely on the invertibility of the matrices $\mathbb{B}_1$ and $\mathbb{A}_2$; the first of these is, in general, non-symmetric and indefinite, even when the necessary discrete inf-sup conditions in (3.2) are satisfied. For (4.21), uncoupling strategies would rely on the invertibility of the matrices $\mathbb{K}_1$, $\mathbb{K}_2$, and $\mathbb{K}_3$; all three of these matrices are symmetric and positive definite even when (3.2) is not satisfied. An example of a simple uncoupling strategy is to apply a block-Gauss-Seidel method to (4.21), which would proceed as follows.

> Start with initial guesses $\vec{\boldsymbol{\phi}}^{(0)}$ and $\vec{\boldsymbol{\theta}}^{(0)}$ for the discretized state and control; then, for $k = 1, 2, \ldots$, successively solve the linear systems

$$\begin{aligned}
\mathbb{K}_3\vec{\boldsymbol{\lambda}}^{(k+1)} &= \vec{\mathbf{g}} - \mathbb{C}_2\vec{\boldsymbol{\phi}}^{(k)} - \mathbb{C}_3\vec{\boldsymbol{\theta}}^{(k)} \\
\mathbb{K}_1\vec{\boldsymbol{\phi}}^{(k+1)} &= \vec{\mathbf{f}} - \mathbb{C}_1^T\vec{\boldsymbol{\theta}}^{(k)} - \mathbb{C}_2^T\vec{\boldsymbol{\lambda}}^{(k+1)} \\
\mathbb{K}_2\vec{\boldsymbol{\theta}}^{(k+1)} &= \vec{\mathbf{h}} - \mathbb{C}_1\vec{\boldsymbol{\phi}}^{(k+1)} - \mathbb{C}_3^T\vec{\boldsymbol{\lambda}}^{(k+1)}
\end{aligned} \tag{4.22}$$

until satisfactory convergence is achieved, e.g., until some norm of the difference between successive iterates is less than some prescribed tolerance.

Since the coefficient matrix in (4.21) is symmetric and positive definite, this iteration will converge. Moreover, all three coefficient matrices $\mathbb{K}_3$, $\mathbb{K}_1$, and $\mathbb{K}_2$ of the linear systems in (4.22) are themselves symmetric and positive definite so that very efficient solution methodologies, including parallel ones, can be applied for their solution. We also note that, in order to obtain faster convergence rates, better uncoupling iterative methods, e.g., over-relaxation schemes or a conjugate gradient method, can be applied instead of the block Gauss-Seidel iteration of (4.22).

REMARK 4.8. The discrete problem (4.19) (or equivalently, (4.21)) resulting from the least-squares method for the optimality system (4.2) can be viewed as a Galerkin discretization of the system

$$\begin{aligned}
&(A_1^*A_1 + B_1^*B_1)\phi + (B_1^*B_2)\theta + (A_1^*B_1^*)\lambda = (A_1^*A_1)\widehat{\phi} + (B_1^*)g && \text{in } \Phi \\
&(A_2^*A_2 + B_2^*B_2)\theta + (A_2^*B_2^*)\lambda + (B_2^*B_1)\phi = (B_2^*)g && \text{in } \Theta \\
&(B_1B_1^* + B_2B_2^*)\lambda + (B_1A_1)\phi + (B_2A_2)\theta = (B_1A_1)\widehat{\phi} && \text{in } \Lambda\,.
\end{aligned} \tag{4.23}$$

The first equation of this system is the sum of A_1^* applied to the first equation of the optimality system (4.2) and B_1^* applied to the third equation of that system. The other equations of (4.23) are related to the equations of (4.2) in a similar manner. The system (4.23) shows that the discrete system (4.21) essentially involves the discretization of "squares" of operators, e.g., $A_1^*A_1$, $B_1^*B_1$, etc. This observation has a profound effect in how one chooses the form of the constraint equation in (2.6), i.e., the form of (2.9). In particular, practical considerations lead to the need to recast a given partial differential equation system into an equivalent first-order form; see, e.g., [5, 8], for details.

5. Methods based on direct penalization by the least-squares functional. A straightforward way to use least-squares notions in the optimization setting of §2 is to enforce the constraint equations (2.4), or equivalently (2.9), by penalizing the functional (2.2), or its equivalent form (2.8), by the least-squares functional (3.6); see [7, 22] for examples of the use of this approach in concrete settings. Thus, instead of solving the constrained problem (2.6) or its equivalent form (2.10), we solve the unconstrained problem

$$\min_{(\phi,\theta)\in\Phi\times\Theta} \mathcal{J}_\epsilon(\phi,\theta)\,, \tag{5.1}$$

where, for given $\widehat{\phi} \in \widehat{\Phi}$ and $g \in \Lambda^*$,

$$\mathcal{J}_\epsilon(\phi,\theta) = \mathcal{J}(\phi,\theta) + \frac{1}{2\epsilon}\mathcal{K}(\phi;\theta,g) \qquad \forall\,\phi\in\Phi,\ \theta\in\Theta \tag{5.2}$$

so that

$$\begin{aligned}\mathcal{J}_\epsilon(\phi,\theta) &= \frac{1}{2}\big\langle A_1(\phi-\widehat{\phi}),(\phi-\widehat{\phi})\big\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} + \frac{1}{2}\langle A_2\theta,\theta\rangle_{\Theta^*,\Theta} \\ &\qquad + \frac{1}{2\epsilon}\big\langle B_1\phi + B_2\theta - g, D^{-1}(B_1\phi + B_2\theta - g)\big\rangle_{\Lambda^*,\Lambda} \\ &= \frac{1}{2}a_1(\phi-\widehat{\phi},\phi-\widehat{\phi}) + \frac{1}{2}a_2(\theta,\theta) \\ &\qquad + \frac{1}{2\epsilon}\big(\widetilde{b}_1(\phi,\phi) + 2\widetilde{b}_2(\theta,\phi) + c(\theta,\theta)\big) \\ &\qquad - \frac{1}{2\epsilon}\big(2\langle\widetilde{g}_1,\phi\rangle_{\Phi^*,\Phi} + 2\langle\widetilde{g}_2,\theta\rangle_{\Theta^*,\Theta} - \langle g, D^{-1}g\rangle_{\Lambda^*,\Lambda}\big)\,.\end{aligned} \tag{5.3}$$

where

$$c(\theta,\nu) = \big\langle B_2\nu, D^{-1}B_2\theta\big\rangle_{\Lambda^*,\Lambda} = \big\langle B_2^*D^{-1}B_2\theta,\nu\big\rangle_{\Theta^*,\Theta} \qquad \forall\,\theta,\nu\in\Theta \tag{5.4}$$

and

$$\widetilde{g}_2 = B_2^*D^{-1}g \in \Theta^*. \tag{5.5}$$

The following results about the bilinear form $c(\cdot,\cdot)$ and the function $\widetilde{g}_2$ are immediate.

PROPOSITION 5.1. *Assume that the operator D is symmetric and that* (3.3) *and the condition on the bilinear form $b_2(\cdot,\cdot)$ in* (2.5) *hold. Then, the bilinear form $c(\cdot,\cdot)$ is symmetric and, for some constant $C_c>0$,*

$$c(\theta,\nu) \le C_c\|\theta\|_\Theta\|\nu\|_\Theta \quad \forall\,\theta,\nu\in\Theta \quad \textit{and} \quad c(\theta,\theta)\ge 0 \quad \forall\,\theta\in\Theta\,. \tag{5.6}$$

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

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

The Euler-Lagrange equations corresponding to the minimization problem (5.1) are given by

$$\begin{cases} a_1(\phi_\epsilon,\mu) + \dfrac{1}{\epsilon}\widetilde{b}_1(\phi,\mu) + \dfrac{1}{\epsilon}\widetilde{b}_2(\theta,\mu) = a_1(\widehat{\phi},\mu) + \dfrac{1}{\epsilon}\langle\widetilde{g}_1,\mu\rangle_{\Phi^*,\Phi} & \forall\mu\in\Phi \\ a_2(\theta,\nu) + \dfrac{1}{\epsilon}\widetilde{c}(\theta,\nu) + \dfrac{1}{\epsilon}\widetilde{b}_2(\nu,\phi_\epsilon) = \dfrac{1}{\epsilon}\langle\widetilde{g}_2,\nu\rangle_{\Theta^*,\Theta} & \forall\nu\in\Theta \end{cases} \tag{5.7}$$

or equivalently

$$\begin{cases} \langle A_1\phi_\epsilon,\mu\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} + \dfrac{1}{\epsilon}\big\langle B_1\mu, D^{-1}(B_1\phi_\epsilon + B_2\theta_\epsilon)\big\rangle_{\Lambda^*,\Lambda} \\ \qquad = \langle A_1\widehat{\phi},\mu\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} + \dfrac{1}{\epsilon}\big\langle B_1\mu, D^{-1}g\big\rangle_{\Lambda^*,\Lambda} & \forall\,\mu\in\Phi \\ \langle A_2\theta_\epsilon,\nu\rangle_{\Theta^*,\Theta} + \dfrac{1}{\epsilon}\big\langle B_2\nu, D^{-1}(B_1\phi_\epsilon + B_2\theta_\epsilon)\big\rangle_{\Lambda^*,\Lambda} \\ \qquad = \dfrac{1}{\epsilon}\big\langle B_2\nu, D^{-1}g\big\rangle_{\Lambda^*,\Lambda} & \forall\,\nu\in\Theta\,. \end{cases} \tag{5.8}$$

For $\phi_\epsilon\in\Phi$ and $\theta_\epsilon\in\Theta$, (3.3) guarantees that

$$\epsilon D\lambda_\epsilon = B_1\phi_\epsilon + B_2\theta_\epsilon - g \qquad \text{in } \Lambda^* \tag{5.9}$$

has a unique solutions $\lambda_\epsilon \in \Lambda$. Then, one easily sees that (5.7) or (5.8) can be expressed in the equivalent form

$$\left\{\begin{array}{lllllll} a_1(\phi_\epsilon,\mu) & & & + & b_1(\mu,\lambda_\epsilon) & = & a_1(\widehat{\phi},\mu) \qquad \forall \mu \in \Phi \\ & a_2(\theta_\epsilon,\nu) & + & & b_2(\nu,\lambda_\epsilon) & = & 0 \qquad \forall \nu \in \Theta \\ b_1(\phi_\epsilon,\psi) & + \; b_2(\theta_\epsilon,\psi) & - & & \epsilon d(\lambda_\epsilon,\psi) & = & \langle g,\psi\rangle_{\Lambda^*,\Lambda} \qquad \forall \psi \in \Lambda\,. \end{array}\right. \tag{5.10}$$

One recognizes the system (5.10) to be a *regular perturbation* of the system (4.1) that is the Euler-Lagrange equations for the minimization problem (2.6) or its equivalent form (2.10).[7] The following result is proved in, e.g., [8].

Concerning the penalized control problem (2.6), we have the following results.

THEOREM 5.1. *Let the assumptions* (2.3), (2.5), *and* (3.3) *hold. Then, for each* $0 < \epsilon \leq 1$, (5.10) *or, equivalently,* (5.8) *and* (5.9), *or, equivalently, the penalized optimal control problem* (5.1), *has a unique solution* $(\phi_\epsilon,\theta_\epsilon,\lambda_\epsilon) \in \Phi \times \Theta \times \Lambda$. *Let* $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$ *denote the unique solution of the optimality system* (4.2) *or, equivalently, of the optimal control problem* (2.6). *Then, for some constant* $C > 0$ *whose value is independent of* ϵ,

$$\|\phi-\phi_\epsilon\|_\Phi + \|\theta-\theta_\epsilon\|_\Theta + \|\lambda-\lambda_\epsilon\|_\Psi \leq C\epsilon\big(\|g\|_{\Psi^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\big). \tag{5.14}$$

Proof. Define the bilinear forms

$$a(\{\phi,\theta\},\{\mu,\nu\}) = a_1(\phi,\mu) + a_2(\theta,\nu) \qquad \forall\, \{\phi,\theta\},\{\mu,\nu\} \in \Phi \times \Theta$$

and

$$b(\{\phi,\theta\},\{\psi\}) = b_1(\phi,\psi) + b_2(\theta,\psi) \qquad \forall\, \{\phi,\theta\} \in \Phi \times \Theta,\ \psi \in \Lambda\,.$$

Then, (4.2) and (5.10) can be respectively written as

$$\left\{\begin{array}{lllll} a(\{\phi,\theta\},\{\mu,\nu\}) & + \; b(\{\mu,\nu\},\{\lambda\}) & = & a_1(\widehat{\phi},\mu) & \forall\, \{\mu,\nu\} \in \Phi \times \Theta \\ b(\{\phi,\theta\},\{\psi\}) & & = & \langle g,\psi\rangle_{\Lambda^*,\Lambda} & \forall\, \psi \in \Lambda \end{array}\right. \tag{5.15}$$

[7] The systems (5.7), (5.8), and (5.10) can be respectively be expressed in equivalent operator form as

$$\left\{\begin{array}{ll} \left(A_1 + \dfrac{1}{\epsilon}\widetilde{B}_1\right)\phi_\epsilon + \dfrac{1}{\epsilon}\widetilde{B}_2\theta_\epsilon = A_1\widehat{\phi} + \dfrac{1}{\epsilon}\widetilde{g}_1 & \text{in } \Phi^* \\ \left(A_2 + \dfrac{1}{\epsilon}C\right)\theta_\epsilon + \dfrac{1}{\epsilon}\widetilde{B}_2^*\phi_\epsilon = \dfrac{1}{\epsilon}\widetilde{g}_2 & \text{in } \Theta^*\,, \end{array}\right. \tag{5.11}$$

$$\left\{\begin{array}{ll} \left(A_1 + \dfrac{1}{\epsilon}B_1^*D^{-1}B_1\right)\phi_\epsilon + \dfrac{1}{\epsilon}B_1^*D^{-1}B_2\theta_\epsilon = A_1\widehat{\phi} + \dfrac{1}{\epsilon}B_1^*D^{-1}g & \text{in } \Phi^* \\ \left(A_2 + \dfrac{1}{\epsilon}B_2^*D^{-1}B_2\right)\theta_\epsilon + \dfrac{1}{\epsilon}B_2^*D^{-1}B_1\phi_\epsilon = \dfrac{1}{\epsilon}B_2^*D^{-1}g & \text{in } \Theta^*\,, \end{array}\right. \tag{5.12}$$

and

$$\left\{\begin{array}{lllllll} A_1\phi_\epsilon & & & + & B_1^*\lambda_\epsilon & = & A_1\widehat{\phi} \quad \text{in } \Phi^* \\ & A_2\theta_\epsilon & + & & B_2^*\lambda_\epsilon & = & 0 \quad \text{in } \Theta^* \\ B_1\phi_\epsilon & + \; B_2\theta_\epsilon & - & & \epsilon D\lambda_\epsilon & = & g \quad \text{in } \Lambda^*. \end{array}\right. \tag{5.13}$$

Incidentally, we can now see why we use D^{-1} in (3.6), i.e., so that in (5.13) D and not D^{-1} appears.

and

$$\left\{\begin{array}{lllll} a(\{\phi_\epsilon,\theta_\epsilon\},\{\mu,\nu\}) & + \; b(\{\mu,\nu\},\{\lambda_\epsilon\}) & = & a_1(\widehat{\phi},\mu) & \forall\,\{\mu,\nu\}\in\Phi\times\Theta \\ b(\{\phi_\epsilon,\theta_\epsilon\},\{\psi\}) & - \; \epsilon d(\lambda_\epsilon,\psi) & = & \langle g,\psi\rangle_{\Lambda^*,\Lambda} & \forall\,\psi\in\Lambda\,. \end{array}\right. \tag{5.16}$$

Let the subspace Z be defined by

$$Z=\big\{\{\phi,\theta\}\in\Phi\times\Theta \;\mid\; b_1(\phi,\psi)+b_2(\theta,\psi)=0 \quad \forall\,\psi\in\Lambda\big\}\,.$$

In operator notation, the elements $\{\phi,\theta\}\in Z\subset\Phi\times\Theta$ satisfy $B_1\phi+B_2\theta=0$. Note that as a result of (2.5), given any $\theta\in\Theta$, there exists a $\phi_\theta\in\Phi$ satisfying

$$b_1(\phi_\theta,\psi)=-b_2(\theta,\psi) \quad \forall\,\psi\in\Lambda \qquad \text{and} \qquad \|\phi_\theta\|_\Phi\le\frac{c_2}{k_1}\|\theta\|_\Theta \tag{5.17}$$

so that Z can be completely characterized by $(\phi_\theta,\theta)\in\Phi\times\Theta$ where, for arbitrary $\theta\in\Theta$, $\phi_\theta\in\Phi$ satisfies (5.17).

In [8], it is shown that if (2.3) and (2.5) hold, then the subspace Z is closed and

$$\left\{\begin{array}{l} a(\{\phi,\theta\},\{\mu,\nu\})\le C_a\|\{\phi,\theta\}\|_{\Phi\times\Theta}\|\{\mu,\nu\}\|_{\Phi\times\Theta} \;\; \forall\,\{\phi,\theta\},\{\mu,\nu\}\in\Phi\times\Theta \\ b(\{\phi,\theta\},\{\lambda\})\le C_b\|\{\phi,\theta\}\|_{\Phi\times\Theta}\|\{\lambda\}\|_\Lambda \qquad \forall\,\{\phi,\theta\}\in\Phi\times\Theta,\;\{\lambda\}\in\Lambda \\ a(\{\phi,\theta\},\{\phi,\theta\})\ge 0 \qquad \forall\,\{\phi,\theta\}\in\Phi\times\Theta \\ a(\{\phi,\theta\},\{\phi,\theta\})\ge K_a\|\{\phi,\theta\}\|^2_{\Phi\times\Theta} \qquad \forall\,\{\phi,\theta\}\in Z \\ \displaystyle\sup_{\{\mu,\nu\}\in\Phi\times\Theta,\{\mu,\nu\}\ne\{0,0\}}\frac{b(\{\mu,\nu\},\{\lambda\})}{\|\{\mu,\nu\}\|_{\Phi\times\Theta}}\ge K_b\|\{\lambda\}\|_\Lambda \qquad \forall\,\{\lambda\}\in\Lambda\,, \end{array}\right. \tag{5.18}$$

where $C_a=\max\{C_1,C_2\}$, $C_b=\max\{c_1,c_2\}$, $K_a=\frac{1}{2}\min\{1,\frac{k_1^2}{c_2^2}\}$, and $K_b=k_1$.

The results of the theorem then easily follow from well-known results about the systems (5.15) and (5.16) whenever (5.18) holds; see, e.g., [8, 10, 13, 14, 17, 19]. □

Now, let us return to the system (5.8) that can be written in more compact form as

$$\mathcal{A}_\epsilon(\{\phi_\epsilon,\theta_\epsilon\},\{\mu,\nu\})=\mathcal{G}_\epsilon(\{\mu,\nu\}) \qquad \forall\,\{\mu,\nu\}\in\Phi\times\Theta \tag{5.19}$$

where, for all $\{\phi_\epsilon,\theta_\epsilon\},\{\mu,\nu\}\in\Phi\times\Theta$,

$$\mathcal{A}_\epsilon(\{\phi,\theta\},\{\mu,\nu\})=a_1(\phi,\mu)+a_2(\theta,\nu)+\frac{1}{\epsilon}\big\langle B_1\mu+B_2\nu,D^{-1}(B_1\phi+B_2\theta)\big\rangle_{\Lambda^*,\Lambda} \tag{5.20}$$

and

$$\mathcal{G}_\epsilon(\{\mu,\nu\})=a_1(\widehat{\phi},\mu)+\frac{1}{\epsilon}\big\langle B_1\mu+B_2\nu,D^{-1}g\big\rangle_{\Lambda^*,\Lambda}. \tag{5.21}$$

Concerning the bilinear form $\mathcal{A}_\epsilon(\cdot,\cdot)$ and the linear functional $\mathcal{G}_\epsilon(\cdot)$, we have the following results.[8]

[8]The results of Lemma 5.1 provide and alternate means for proving, for any $0<\epsilon\le 1$, that the system (5.8) has a unique solution. Indeed, those results assert that the symmetric bilinear form $\mathcal{A}_\epsilon(\cdot,\cdot)$ is continuous and coercive and that the linear functional $\mathcal{G}_\epsilon(\cdot)$ is continuous so that the existence and uniqueness of the solution of (5.19), or equivalently, of (5.8) follows by the Lax-Milgram lemma. However, due to the ϵ^{-1} in the right-hand side of (5.22), the results of Lemma 5.1 cannot be used to derive the estimate (5.14) for the solution of (5.8); this is done indirectly by using the equivalence of (5.8) and (5.9) with (5.10).

LEMMA 5.1. *Let the bilinear form $\mathcal{A}_\epsilon(\cdot,\cdot)$ be defined by* (5.20) *and let the linear functional $\mathcal{G}_\epsilon(\cdot)$ be defined by* (5.21). *Let the assumptions* (2.3), (2.5), *and* (3.3) *hold and let $0 < \epsilon \le 1$. Then, there exist positive constants c_{a1}, c_{a2}, c_{g1}, c_{g2}, and k_a whose values do not depend on ϵ such that for all $\{\phi,\theta\}$, $\{\mu,\nu\} \in \Phi \times \Theta$,*

$$\mathcal{A}_\epsilon(\{\phi,\theta\},\{\mu,\nu\}) \le \left(c_{a1} + \frac{c_{a2}}{\epsilon}\right) \|\{\phi,\theta\}\|_{\Phi\times\Theta} \|\{\mu,\nu\}\|_{\Phi\times\Theta} \tag{5.22}$$

and

$$\mathcal{A}_\epsilon(\{\phi,\theta\},\{\phi,\theta\}) \ge k_a \|\{\phi,\theta\}\|^2_{\Phi\times\Theta}\,. \tag{5.23}$$

Furthermore,

$$\mathcal{G}_\epsilon(\{\mu,\nu\}) \le \left(c_{g1}\|\widehat{\phi}\|_{\widehat{\Phi}} + \frac{c_{g2}}{\epsilon}\|g\|_{\Lambda^*}\right) \|\{\mu,\nu\}\|_{\Phi\times\Theta}\,. \tag{5.24}$$

Proof. Using (5.18), we have that

$$\begin{aligned}
&\mathcal{A}_\epsilon(\{\phi,\theta\},\{\mu,\nu\}) \\
&\quad \le C_a \|\{\phi,\theta\}\|_{\Phi\times\Theta}\|\{\mu,\nu\}\|_{\Phi\times\Theta} + \frac{1}{\epsilon}\|D^{-1}\|_{\Lambda^*\to\Lambda}\|B_1\phi + B_2\theta\|_{\Lambda^*}\|B_1\mu + B_2\nu\|_{\Lambda^*} \\
&\quad \le \left(C_a + \frac{C_b^2}{\epsilon k_d}\right) \|\{\phi,\theta\}\|_{\Phi\times\Theta}\|\{\mu,\nu\}\|_{\Phi\times\Theta}
\end{aligned}$$

so that (5.22) holds with $c_{a1} = C_a = \max\{C_1, C_2\}$ and $c_{a2} = \frac{C_b^2}{k_d} = \frac{(\max\{c_1,c_2\})^2}{k_d}$.

The proof of (5.24) proceeds in a similar manner; one obtains that $c_{g1} = C_1$ and $c_{g2} = \frac{1}{k_d}\max\{c_1, c_2\}$.

Next, suppose $\{\phi,\theta\} \in Z$ so that $B_1\phi + B_2\theta = 0$ and, by (5.17), $\|\phi\|_\Lambda \le \frac{c_2}{k_1}\|\theta\|_\Theta$. Then,

$$\begin{aligned}
&\mathcal{A}_\epsilon(\{\phi,\theta\},\{\phi,\theta\}) = a_1(\phi,\phi) + a_2(\theta,\theta) \ge a_2(\theta,\theta) \ge K_2\|\theta\|^2_\Theta \\
&\quad \ge \frac{K_2}{2}\left(\|\theta\|^2_\Theta + \frac{k_1^2}{c_2^2}\|\phi\|^2_\Phi\right) \ge \frac{K_2}{2}\min\left\{1, \frac{k_1^2}{c_2^2}\right\}\|\{\phi,\theta\}\|^2 \quad \forall\,\{\phi,\theta\} \in Z\,.
\end{aligned} \tag{5.25}$$

Now, it is well known (see, e.g., [17]) that if (5.18) holds, then

$$\sup_{\{\psi\}\in\Lambda,\{\psi\}\ne\{0\}} \frac{b(\{\phi,\theta\},\{\psi\})}{\|\{\psi\}\|_\Lambda} \ge k_1\|\{\phi,\theta\}\|_{\Phi\times\Theta} \quad \forall\,\{\phi,\theta\} \in Z^\perp\,.$$

Then, since for all $\{\phi,\theta\} \in \Phi\times\Theta$,

$$b(\{\phi,\theta\},\{\psi\}) = b_1(\phi,\psi) + b_2(\theta,\psi) = \langle B_1\phi + B_2\theta, \psi\rangle_{\Lambda^*,\Lambda}\,,$$

we have that

$$\|B_1\phi + B_2\theta\|_{\Lambda^*} \ge k_1\|\{\phi,\theta\}\|_{\Phi\times\Theta} \quad \forall\,\{\phi,\theta\} \in Z^\perp$$

so that, using (5.18) and $0 < \epsilon \le 1$,

$$\begin{aligned}
&\mathcal{A}_\epsilon(\{\phi,\theta\},\{\phi,\theta\}) \\
&\quad = a_1(\phi,\phi) + a_2(\theta,\theta) + \frac{1}{\epsilon}\big\langle B_1\phi + B_2\theta, D^{-1}(B_1\phi + B_2\theta)\big\rangle_{\Lambda^*,\Lambda} \\
&\quad \ge \big\langle B_1\phi + B_2\theta, D^{-1}(B_1\phi + B_2\theta)\big\rangle_{\Lambda^*,\Lambda} \ge \frac{k_d}{c_d^2}\|B_1\phi + B_2\theta\|^2_{\Lambda^*} \\
&\quad \ge \frac{k_d k_1^2}{c_d^2}\|\{\phi,\theta\}\|^2 \qquad \forall\,\{\phi,\theta\} \in Z^\perp.
\end{aligned} \tag{5.26}$$

Since $Z \subset \Phi \times \Theta$ is closed, we obtain (5.23) with $k_a = \min\left\{\frac{K_2}{2}\min\{1, \frac{k_1^2}{c_2^2}\}, \frac{k_d k_1^2}{c_d^2}\right\}$ by combining (5.25) and (5.26). □

As a result of the assumptions in (3.3) for the operator D, we see that (5.8) and (5.10) are completely equivalent. One may then proceed to discretize either of these systems. It is important to note that the two resulting discrete systems are *not equivalent* and can, in fact, have significantly different properties.

5.1. Discretization of the regularized optimality system. We consider obtaining a discretization of (5.8) by first discretizing (5.10) and then eliminating the Lagrange multiplier. Discretization can be effected by choosing conforming finite element spaces $\Phi^h \subset \Phi$, $\Theta^h \subset \Theta$, and $\Lambda^h \subset \Lambda$ and then restricting (5.10) to the subspaces to obtain

$$\begin{array}{rcrcrcll} a_1(\phi_\epsilon^h, \mu^h) & & & + & b_1(\mu^h, \lambda_\epsilon^h) & = & a_1(\widehat{\phi}, \mu^h) & \forall \mu^h \in \Phi^h \\ & & a_2(\theta_\epsilon^h, \nu^h) & + & b_2(\nu^h, \lambda_\epsilon^h) & = & 0 & \forall \nu^h \in \Theta^h \\ b_1(\phi_\epsilon^h, \psi^h) & + & b_2(\theta_\epsilon^h, \psi^h) & - & \epsilon d(\lambda_\epsilon^h, \psi^h) & = & \langle g, \psi^h \rangle_{\Lambda^*,\Lambda} & \forall \psi^h \in \Lambda^h. \end{array} \tag{5.27}$$

In the usual way, the discrete system (5.27) is equivalent to a matrix problem. In addition to the matrices $\mathbb{A}_1$, $\mathbb{A}_2$, $\mathbb{B}_1$, and $\mathbb{B}_2$ and the vectors $\vec{\mathbf{f}}$ and $\vec{\mathbf{g}}$ defined in §4.1, we define the matrix $\mathbb{D}$ by

$$(\mathbb{D})_{mn} = d(\lambda_m, \lambda_n) \quad \text{for } m, n = 1, \ldots, M.$$

Then, the discrete regularized problem (5.27) is equivalent to the linear system

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \mathbb{B}_1^T \\ 0 & \mathbb{A}_2 & \mathbb{B}_2^T \\ \mathbb{B}_1 & \mathbb{B}_2 & -\epsilon\mathbb{D} \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}}_\epsilon \\ \vec{\boldsymbol{\theta}}_\epsilon \\ \vec{\boldsymbol{\lambda}}_\epsilon \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \\ \vec{\mathbf{g}} \end{pmatrix}. \tag{5.28}$$

It is now easy to see how one can eliminate $\vec{\boldsymbol{\lambda}}_\epsilon$ from (5.28), or equivalently, λ_ϵ^h from (5.27). Indeed, (3.5) implies that $\mathbb{D}$ is symmetric and positive definite, and therefore invertible. Then, one easily deduces from (5.28) that

$$\begin{cases} \left(\mathbb{A}_1 + \dfrac{1}{\epsilon}\mathbb{B}_1^T\mathbb{D}^{-1}\mathbb{B}_1\right)\vec{\boldsymbol{\phi}}_\epsilon + \dfrac{1}{\epsilon}\mathbb{B}_1^T\mathbb{D}^{-1}\mathbb{B}_2\vec{\boldsymbol{\theta}}_\epsilon = \vec{\mathbf{f}} + \dfrac{1}{\epsilon}\mathbb{B}_1^T\mathbb{D}^{-1}\vec{\mathbf{g}} \\ \left(\mathbb{A}_2 + \dfrac{1}{\epsilon}\mathbb{B}_2^T\mathbb{D}^{-1}\mathbb{B}_2\right)\vec{\boldsymbol{\theta}}_\epsilon + \dfrac{1}{\epsilon}\mathbb{B}_2^T\mathbb{D}^{-1}\mathbb{B}_1\vec{\boldsymbol{\phi}}_\epsilon = \dfrac{1}{\epsilon}\mathbb{B}_2^T\mathbb{D}^{-1}\vec{\mathbf{g}}. \end{cases} \tag{5.29}$$

Note that (5.29) only involves the approximations $\phi_\epsilon^h \in \Phi^h$ and $\theta_\epsilon^h \in \Theta^h$ of the state variable $\phi \in \Phi$ and the control variable $\theta \in \Theta$, respectively, and does not involve the approximation $\lambda_\epsilon^h \in \Psi^h$ of the adjoint variable $\lambda \in \Psi$. Once (5.29) is used to determine $\vec{\boldsymbol{\phi}}_\epsilon$ and $\vec{\boldsymbol{\theta}}_\epsilon$, $\vec{\boldsymbol{\lambda}}_\epsilon$ may be determined from the last equation in (5.28).

Now, consider what is required to guarantee that the coefficient matrix of the linear system (5.28) or, equivalently, of (5.29) is stably invertible as either or both the grid size h and the penalty parameter ϵ tend to zero. It is not difficult to show, based on the assumptions (2.3), (2.5), and (3.5) that we have made about the bilinear forms appearing in (5.27), that a necessary and sufficient condition for the stable invertibility of (5.28) or (5.29) is that the matrix $\mathbb{B}_1$ be stably invertible. We have already seen in §3.1 that this guarantee can be made if and only if the subspaces Φ^h

and Λ^h satisfy (3.2), i.e., the same requirement needed to insure that the Galerkin discretization (4.3) of the unperturbed optimality system is stably invertible; see §4.1. In other words, despite the fact that

> (5.10) is equivalent to enforcing the constraint (2.9) by penalizing the functional (2.8) by the *well-posed* least-squares functional (3.6)

and despite the fact that

> given a control θ, stable approximations of the state ϕ may be obtained by minimizing the least-squares functional (3.6) *without* having to assume that the discrete spaces Φ^h and Λ^h satisfy (3.2),

the stable solution of (5.27), or equivalently (5.28) or (5.29), requires that (3.2) is satisfied. Thus, one of the main advantages of using least-squares finite element methods, i.e., being able to circumvent (3.2), is lost.[9]

The following error estimate is easily derived using well-known techniques.

THEOREM 5.2. *Let* (2.3), (2.5), (3.3), *and* (3.2) *hold. Then,* (5.27), *or equivalently* (5.29), *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* (2.6) *or equivalently, of* (5.8), *or equivalently, of* (5.10), *then there exist a constant* $C > 0$ *whose value is independent of* ϵ *and* h *such that*

$$\begin{aligned}\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta + \|\lambda - \lambda_\epsilon^h\|_\Lambda \le C\epsilon \left(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\right) \\ + C\Big(\inf_{\widetilde{\phi}^h \in \Phi^h} \|\phi - \widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h \in \Theta^h} \|\theta - \widetilde{\theta}^h\|_\Theta + \inf_{\widetilde{\lambda}^h \in \Lambda^h} \|\lambda - \widetilde{\lambda}^h\|_\Lambda\Big).\end{aligned} \tag{5.30}$$

Proof. Standard finite element analyses [10, 13, 14, 17] yield, for the pair of systems (5.10) and (5.27), that

$$\|\phi_\epsilon - \phi_\epsilon^h\|_\Phi + \|\theta_\epsilon - \theta_\epsilon^h\|_\Theta + \|\lambda_\epsilon - \lambda_\epsilon^h\|_\Lambda \le C\big(\|\phi_\epsilon - \widetilde{\phi}^h\|_\Phi + \|\theta_\epsilon - \widetilde{\theta}^h\|_\Theta + \|\lambda_\epsilon - \widetilde{\lambda}^h\|_\Lambda\big)$$

for all $\widetilde{\phi}^h \in \Phi^h$, $\widetilde{\theta}^h \in \Theta^h$, and $\widetilde{\lambda}^h \in \Lambda^h$. Then, (5.14) and the triangle inequality yields (5.30). ∎

Our discussion serves to point out an important observation about penalty methods, namely that they are not stabilization methods, i.e., penalty methods do not circumvent the discrete conditions (3.2).[10] Penalty methods are properly viewed as being *methods for facilitating the solution of* (4.1) or (4.4). Since here we are primarily interested in retaining the advantage that least-squares finite element methods provide for circumventing conditions such as (3.2), we do not consider discretizations of (5.10) as the best way to incorporate least-square notions into the optimization problems we are considering.

It is usually the case that the approximation-theoretic terms on the right-hand side of (5.30) satisfy inequalities of the type

$$\inf_{\widetilde{\phi}^h \in \Phi^h} \|\phi - \widetilde{\phi}^h\|_\Phi \le Ch^\alpha, \quad \inf_{\widetilde{\lambda}^h \in \Lambda^h} \|\lambda - \widetilde{\lambda}^h\|_\Lambda \le Ch^\alpha, \quad \text{and} \quad \inf_{\widetilde{\theta}^h \in \Theta^h} \|\theta - \widetilde{\theta}^h\|_\Theta \le Ch^\beta, \tag{5.31}$$

[9] Although discretizations of (4.2) and (5.10) both require the imposition of (3.2) on the finite element spaces Φ^h and Λ^h, using the system (5.28) still has some advantages. Foremost among these is that one can reduce the number of variables by eliminating $\vec{\boldsymbol{\lambda}}_\epsilon$ from (5.28) to obtain (5.29). Furthermore, as long as (3.2) is satisfied, the system (5.29) is symmetric and positive definite while (5.28) is symmetric but indefinite as is, of course, (4.4).

[10] The fact that discretizations of (4.2) and (5.10) both require the imposition of (3.2) should not be surprising, given that the latter is a regular perturbation of the former.

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 and θ_ϵ of (5.10), or equivalently, of (5.8). Then, (5.30) implies that

$$\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta + \|\lambda - \lambda_\epsilon^h\|_\Lambda \le C\left(\epsilon + h^\alpha + h^\beta\right) \tag{5.32}$$

with C independent of ϵ and h.

5.2. Discretization of the eliminated system. Instead of discretizing (5.10) and then eliminating the approximation of the Lagrange multiplier to obtain (5.29), one can directly discretize the eliminated system (5.8) or, equivalently, minimize the functional $\mathcal{J}_\epsilon(\cdot,\cdot)$ over $(\phi^h, \theta^h) \in \Phi^h \times \Theta^h$. Choosing approximating subspaces $\Phi^h \subset \Phi$ and $\Theta^h \subset \Theta$, the discrete problem is then given by

$$\begin{cases} a_1(\phi_\epsilon^h, \mu^h) + \dfrac{1}{\epsilon}\big\langle B_1\mu^h, D^{-1}(B_1\phi_\epsilon^h + B_2\theta_\epsilon^h)\big\rangle_{\Lambda^*,\Lambda} \\ \qquad\qquad = a_1(\widehat{\phi}, \mu^h) + \dfrac{1}{\epsilon}\big\langle B_1\mu^h, D^{-1}g\big\rangle_{\Lambda^*,\Lambda} \qquad \forall\, \mu^h \in \Phi^h \\ a_2(\theta_\epsilon^h, \nu^h) + \dfrac{1}{\epsilon}\big\langle B_2\nu^h, D^{-1}(B_1\phi_\epsilon^h + B_2\theta_\epsilon^h)\big\rangle_{\Lambda^*,\Lambda} \\ \qquad\qquad = \dfrac{1}{\epsilon}\big\langle B_2\nu^h, D^{-1}g\big\rangle_{\Lambda^*,\Lambda} \qquad \forall\, \nu^h \in \Theta^h\,. \end{cases} \tag{5.33}$$

This system can be written in the more compact form

$$\mathcal{A}_\epsilon(\{\phi_\epsilon^h, \theta_\epsilon^h\}, \{\mu^h, \nu^h\}) = \mathcal{G}_\epsilon(\{\mu^h, \nu^h\}) \qquad \forall\, \{\mu^h, \nu^h\} \in \Phi^h \times \Theta^h\,, \tag{5.34}$$

where the bilinear form $\mathcal{A}_\epsilon(\cdot,\cdot)$ and linear functional $\mathcal{G}_\epsilon(\cdot)$ are defined in (5.20) and (5.21), respectively.[11]

THEOREM 5.3. *Let* (2.3), (2.5), *and* (3.3) *hold. Then, for* $0 < \epsilon \le 1$, (5.33), *or equivalently,* (5.34) *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* (2.6) *or equivalently, of* (5.8), *or equivalently, of* (5.10), *then there exist a constant* $C > 0$ *whose value is independent of* ϵ *and* h *such that*

$$\begin{aligned} \|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta &\le C\epsilon\left(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\right) \\ &\quad + C\Big(1 + \frac{1}{\epsilon}\Big)\Big(\inf_{\widetilde{\phi}^h \in \Phi^h} \|\phi_\epsilon - \widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h \in \Theta^h} \|\theta_\epsilon - \widetilde{\theta}^h\|_\Theta\Big). \end{aligned} \tag{5.35}$$

Proof. Because of Lemma 5.1, the existence and uniqueness of the solution of (5.34) follow from the Lax-Milgram lemma. Moreover, standard finite element analyses for the problem (5.19) and its discretization (5.34) yield that

$$\|\phi_\epsilon - \phi_\epsilon^h\|_\Phi + \|\theta_\epsilon - \theta_\epsilon^h\|_\Theta \le C\Big(1 + \frac{1}{\epsilon}\Big)\Big(\inf_{\widetilde{\phi}^h \in \Phi^h} \|\phi_\epsilon - \widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h \in \Theta^h} \|\theta_\epsilon - \widetilde{\theta}^h\|_\Theta\Big)\,.$$

Then, (5.14) and the triangle inequality yields (5.35). □

In the usual way, the discrete system (5.33) is equivalent to a matrix problem. 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

[11] The results of Theorem 5.3 do not require that the discrete inf-sup conditions (3.2) holds.

basis sets for Φ^h and Θ^h, respectively. In addition to the matrices $\mathbb{A}_1$, $\mathbb{A}_2$, and $\widetilde{\mathbb{B}}_1$ and the vectors $\vec{\mathbf{f}}$ and $\widetilde{\vec{\mathbf{g}}}$ defined previously, we define the matrices

$$\begin{cases} (\widetilde{\mathbb{B}}_2)_{jk} = \widetilde{b}_2(\theta_k, \phi_j) = \left\langle B_2\theta_k, D^{-1}B_1\phi_j \right\rangle_{\Lambda^*,\Lambda} & \text{for } k = 1, \ldots, K,\ j = 1, \ldots, J \\ (\mathbb{C})_{k\ell} = c(\theta_k, \theta_\ell) = \left\langle B_2\theta_k, D^{-1}B_2\theta_\ell \right\rangle_{\Lambda^*,\Lambda} & \text{for } k, \ell = 1, \ldots, K \end{cases}$$

and the vectors

$$\begin{cases} (\vec{\mathbf{g}}_1)_i = \langle \widetilde{g}_1, \phi_i \rangle_{\Phi^*,\Phi} = \left\langle B_1\phi_i, D^{-1}g \right\rangle_{\Lambda^*,\Lambda} & \text{for } k = 1, \ldots, K \\ (\vec{\mathbf{g}}_2)_k = \langle \widetilde{g}_2, \theta_k \rangle_{\Theta^*,\Theta} = \left\langle B_2\theta_k, D^{-1}g \right\rangle_{\Lambda^*,\Lambda} & \text{for } k = 1, \ldots, K\,. \end{cases}$$

Then, (5.33) is equivalent to the matrix problem

$$\begin{pmatrix} \mathbb{A}_1 + \dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_1 & \dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_2 \\ \dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_2^T & \mathbb{A}_2 + \dfrac{1}{\epsilon}\mathbb{C} \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}}_\epsilon \\ \vec{\boldsymbol{\theta}}_\epsilon \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} + \dfrac{1}{\epsilon}\vec{\mathbf{g}}_1 \\ \dfrac{1}{\epsilon}\vec{\mathbf{g}}_2 \end{pmatrix}, \tag{5.36}$$

where $\vec{\boldsymbol{\phi}}_\epsilon$ and $\vec{\boldsymbol{\theta}}_\epsilon$ are the vectors of coefficients for ϕ_ϵ^h and θ_ϵ^h, respectively.

It is clear that (5.29) and (5.36) are different, i.e., the discretize-then-eliminate approach yields a discrete system that is not equivalent to the system obtained by the eliminate-then-discretize approach, despite the fact that their respective parent continuous systems (5.10) and (5.8) are equivalent. In other words, elimination and discretization steps do not commute!

Note that (5.36) is determined without the need for choosing a subspace Λ^h for the approximation of the Lagrange multiplier. As a result, unlike what is the case for (5.29), for a fixed value of ϵ, the stable invertibility of the system (5.36) does not require the state approximation space Φ^h to satisfy (3.2). In fact, because of (5.22) and (5.23), for a fixed value of ϵ, the coefficient matrix in (5.36) is uniformly (with respect to h) positive definite for any choices for Φ^h and Θ^h.

The approximation-theoretic terms on the right-hand side of (5.35) satisfy inequalities of the type (5.31). Then, (5.35) implies that

$$\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta \leq C\left(\epsilon + \frac{h^\alpha + h^\beta}{\epsilon}\right), \tag{5.37}$$

where the value of $C > 0$ is independent of h and ϵ. The estimate (5.37) shows that nothing bad happens as $h \to 0$ for fixed ϵ. In fact, as $h \to 0$, the error in ϕ_ϵ^h and θ_ϵ^h is of order ϵ which is the best one can hope for for a fixed value of ϵ. However, (5.37) suggests that something bad may[12] happen as $\epsilon \to 0$. In fact, this effect is well known as *locking* and indeed does happen for at least some choices of Φ^h; see, e.g., [10] for a discussion of locking phenomena. Thus, to be safe, (5.37) suggests that as $\epsilon \to 0$, h should be chosen to depend on ϵ in such a way that the right-hand side tends to zero as ϵ and h tend to zero. For example, if $\beta \geq \alpha$, as is often the case, then to equilibrate the two terms in the right-hand side of (5.37), we choose $h = \epsilon^{2/\alpha}$ so that

$$\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta \leq C\epsilon = Ch^{\alpha/2}.$$

[12]Since (5.37) only provides an upper bound for the error, it does not with certainty predict what happens as $\epsilon \to 0$.

In this case, convergence is guaranteed for any choice for Φ^h and Θ^h, but the rate of convergence (with respect to h) may be suboptimal. This should be compared to the results for the discretization of the regularized optimality system (see (5.32)) for which optimal rates of convergence with respect to h are obtained and locking does not occur. Of course, the estimate (5.32) requires that the finite element spaces satisfy the discrete stability conditions in (3.2), while the estimate (5.37) holds without the need to impose those stability conditions.

6. Methods based on constraining by the least-squares functional. Another means of incorporating least-squares notions into a solution method for the constrained optimization problem of §2 is to solve, instead of (2.6) or its equivalent form (2.10), the bilevel minimization problem

$$\min_{(\phi,\theta)\in\Phi\times\Theta} \mathcal{J}(\phi,\theta) \qquad \text{subject to} \qquad \min_{\phi\in\Phi} \mathcal{K}(\phi;\theta,g)\,. \tag{6.1}$$

From (3.14), one sees that this is equivalent to the problem

$$\min_{(\phi,\theta)\in\Phi\times\Theta} \mathcal{J}(\phi,\theta) \qquad \text{subject to} \qquad \widetilde{B}_1\phi + \widetilde{B}_2\theta = \widetilde{g} \quad \text{in } \Phi^*. \tag{6.2}$$

The Euler-Lagrange equations corresponding to the minimization problem (6.2) are given by

$$\left\{\begin{array}{lllllll} A_1\phi & & & + & \widetilde{B}_1\mu & = & A_1\widehat{\phi} \quad \text{in } \Phi^* \\ & & A_2\theta & + & \widetilde{B}_2^*\mu & = & 0 \quad \text{in } \Theta^* \\ \widetilde{B}_1\phi & + & \widetilde{B}_2\theta & & & = & \widetilde{g}_1 \quad \text{in } \Phi^*, \end{array}\right. \tag{6.3}$$

where $\mu \in \Phi$ is the Lagrange multiplier introduced to enforce the constraint in (6.2).

The problem (6.2) should be contrasted with the problem (2.10). Both (2.10) and (6.2) involve the same functional $\mathcal{J}(\cdot,\cdot)$, but are constrained differently. As a result, the former leads to the optimality system (4.2) while the latter leads to the optimality system (6.3). 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 (4.2) may be non-symmetric and indefinte; on the other hand, the operator $\widetilde{B}_1 = B_1^* D^{-1} B_1$ that plays the analogous role in (6.3) is always symmetric and positive definite whenever the assumptions (2.5) and (3.3) hold.

Penalization can be used to facilitate the solution of the system (6.3) in just the same way as (5.10) is related to (4.2). To this end, we let $\widetilde{D} : \Phi \to \Phi^*$ be a self-adjoint, strongly coercive operator, i.e., there exist constants $\widetilde{c}_d > 0$ and $\widetilde{k}_d > 0$ such that

$$\langle \widetilde{D}\mu, \phi\rangle_{\Phi^*,\Phi} \le \widetilde{c}_d \|\mu\|_\Phi \|\phi\|_\Phi \qquad \text{and} \qquad \langle \widetilde{D}\mu, \mu\rangle_{\Phi^*,\Phi} \ge \widetilde{k}_d \|\mu\|_\Phi^2 \tag{6.4}$$

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

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

We then consider the penalized functional

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

and the unconstrained optimization problem

$$\min_{\phi\in\Phi,\,\theta\in\Theta} \widetilde{\mathcal{J}}_\epsilon(\phi,\theta)\,. \tag{6.5}$$

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

$$\begin{cases} \left(A_1+\dfrac{1}{\epsilon}\widetilde{B}_1\widetilde{D}^{-1}\widetilde{B}_1\right)\phi_\epsilon+\dfrac{1}{\epsilon}\widetilde{B}_1\widetilde{D}^{-1}\widetilde{B}_2\theta_\epsilon=A_1\widehat{\phi}+\dfrac{1}{\epsilon}\widetilde{B}_1\widetilde{D}^{-1}\widetilde{g}_1 & \text{in } \Phi^* \\ \left(A_2+\dfrac{1}{\epsilon}\widetilde{B}_2^*\widetilde{D}^{-1}\widetilde{B}_2\right)\theta_\epsilon+\dfrac{1}{\epsilon}\widetilde{B}_2^*\widetilde{D}^{-1}\widetilde{B}_1\phi_\epsilon=\dfrac{1}{\epsilon}\widetilde{B}_2^*\widetilde{D}^{-1}\widetilde{g}_1 & \text{in } \Theta^* \end{cases} \tag{6.6}$$

or

$$\begin{cases} \left(A_1+\dfrac{1}{\epsilon}B_1^*D^{-1}B_1\widetilde{D}^{-1}B_1^*D^{-1}B_1\right)\phi_\epsilon \\ +\dfrac{1}{\epsilon}B_1^*D^{-1}B_1\widetilde{D}^{-1}B_1^*D^{-1}B_2\theta_\epsilon=A_1\widehat{\phi}+\dfrac{1}{\epsilon}B_1^*D^{-1}B_1\widetilde{D}^{-1}B_1^*D^{-1}g & \text{in } \Phi^* \\ \left(A_2+\dfrac{1}{\epsilon}B_2^*D^{-1}B_1\widetilde{D}^{-1}B_1^*D^{-1}B_2\right)\theta_\epsilon \\ +\dfrac{1}{\epsilon}B_2^*D^{-1}B_1\widetilde{D}^{-1}B_1^*D^{-1}B_1\phi_\epsilon=\dfrac{1}{\epsilon}B_2^*D^{-1}B_1\widetilde{D}^{-1}B_1^*D^{-1}g & \text{in } \Theta^*. \end{cases} \tag{6.7}$$

Letting $\mu_\epsilon=\widetilde{D}^{-1}(\widetilde{B}_1\phi_\epsilon+\widetilde{B}_2\theta_\epsilon-\widetilde{g}_1)$, it is easy to see that is (6.6) is equivalent to the following regular perturbation of (6.3):

$$\begin{cases} A_1\phi_\epsilon & & + & \widetilde{B}_1\mu_\epsilon & = & A_1\widehat{\phi} & \text{in } \Phi^* \\ & A_2\theta_\epsilon & + & \widetilde{B}_2^T\mu_\epsilon & = & 0 & \text{in } \Theta^* \\ \widetilde{B}_1\phi_\epsilon & + \widetilde{B}_2\theta_\epsilon & - & \epsilon\widetilde{D}\mu_\epsilon & = & \widetilde{g}_1 & \text{in } \Phi^*. \end{cases} \tag{6.8}$$

The systems (6.6) and (6.8) are equivalent, but once again, 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 for (5.10) and (4.2), now discretization of either (6.6) or (6.8) will result in matrix systems (after elimination in the second case) that are uniformly (with respect to h) positive definite without requiring that (3.2) holds.

6.1. Discretize-then-eliminate. Discretizing the equivalent weak formulation corresponding to (6.8) results in the matrix problem[13]

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \widetilde{\mathbb{B}}_1 \\ 0 & \mathbb{A}_2 & \widetilde{\mathbb{B}}_2^T \\ \widetilde{\mathbb{B}}_1 & \widetilde{\mathbb{B}}_2 & -\epsilon\widetilde{\mathbb{D}} \end{pmatrix}\begin{pmatrix} \vec{\boldsymbol{\phi}}_\epsilon \\ \vec{\boldsymbol{\theta}}_\epsilon \\ \vec{\boldsymbol{\mu}}_\epsilon \end{pmatrix}=\begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \\ \vec{\mathbf{g}}_1 \end{pmatrix}, \tag{6.10}$$

where the matrices $\mathbb{A}_1$, $\mathbb{A}_2$, $\widetilde{\mathbb{B}}_1$, and $\widetilde{\mathbb{B}}_2$ and the vectors $\vec{\mathbf{f}}$ and $\vec{\mathbf{g}}_1$ are as in (5.36) and the matrix $\widetilde{\mathbb{D}}$ corresponds to the bilinear form $\widetilde{d}(\phi,\mu)=\langle\widetilde{D}\mu,\phi\rangle_{\Phi^*,\Phi}$ for $\phi,\mu\in\Phi$.

[13]Discretization of the unperturbed system (6.3) yields the related discrete system

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \widetilde{\mathbb{B}}_1 \\ 0 & \mathbb{A}_2 & \widetilde{\mathbb{B}}_2^T \\ \widetilde{\mathbb{B}}_1 & \widetilde{\mathbb{B}}_2 & 0 \end{pmatrix}\begin{pmatrix} \vec{\boldsymbol{\phi}} \\ \vec{\boldsymbol{\theta}} \\ \vec{\boldsymbol{\mu}} \end{pmatrix}=\begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \\ \vec{\mathbf{g}}_1 \end{pmatrix}. \tag{6.9}$$

The system (6.10) is symmetric and indefinite, but it is uniformly (with respect to h) invertible without regard to (3.2). Indeed, we have that the matrices $\widetilde{\mathbb{B}}_1$ and $\mathbb{A}_2$ are symmetric and positive definite whenever (2.3), (2.5), and (3.3) hold. This should be contrasted with the situation for (5.28) whose uniform invertibility required that the discrete spaces satisfy (3.2).

The vector of coefficients $\vec{\boldsymbol{\mu}}_\epsilon$ may be eliminated from (6.10) to yield

$$\left\{\begin{array}{l} \left(\mathbb{A}_1+\dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_1\widetilde{\mathbb{D}}^{-1}\widetilde{\mathbb{B}}_1\right)\vec{\boldsymbol{\phi}}_\epsilon+\dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_1\widetilde{\mathbb{D}}^{-1}\widetilde{\mathbb{B}}_2\vec{\boldsymbol{\theta}}_\epsilon=\vec{\mathbf{f}}+\dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_1\widetilde{\mathbb{D}}^{-1}\vec{\mathbf{g}}_1 \\ \left(\mathbb{A}_2+\dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_2^T\widetilde{\mathbb{D}}^{-1}\widetilde{\mathbb{B}}_2\right)\vec{\boldsymbol{\theta}}_\epsilon+\dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_2^T\widetilde{\mathbb{D}}^{-1}\widetilde{\mathbb{B}}_1\vec{\boldsymbol{\phi}}_\epsilon=\dfrac{1}{\epsilon}\widetilde{\mathbb{B}}_2^T\widetilde{\mathbb{D}}^{-1}\vec{\mathbf{g}}_1\,. \end{array}\right. \tag{6.11}$$

THEOREM 6.1. *Let* (2.3), (2.5), *and* (3.3) *hold. Then,* (6.10) *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* (2.6) *or equivalently, of* (5.8), *or equivalently, of* (5.10), *then there exist a constant* $C>0$ *whose value is independent of* ϵ *and* h *such that*

$$\begin{aligned}\|\phi-\phi_\epsilon^h\|_\Phi+\|\theta-\theta_\epsilon^h\|_\Theta+\|\mu-\mu_\epsilon^h\|_\Phi&\le C\epsilon\left(\|g\|_{\Lambda^*}+\|\widehat{\phi}\|_{\widehat{\Phi}}\right)\\ &+C\Big(\inf_{\widetilde{\phi}^h\in\Phi^h}\|\phi_\epsilon-\widetilde{\phi}^h\|_\Phi+\inf_{\widetilde{\theta}^h\in\Theta^h}\|\theta_\epsilon-\widetilde{\theta}^h\|_\Theta+\inf_{\widetilde{\mu}^h\in\Phi^h}\|\mu_\epsilon-\mu^h\|_\Phi\Big).\end{aligned} \tag{6.12}$$

Using (5.31), we have from (6.12) that

$$\|\phi-\phi_\epsilon^h\|_\Phi+\|\theta-\theta_\epsilon^h\|_\Theta+\|\mu-\mu_\epsilon^h\|_\Phi\le C(\epsilon+h^\alpha+h^\beta) \tag{6.13}$$

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

$$\|\phi-\phi_\epsilon^h\|_\Phi+\|\theta-\theta_\epsilon^h\|_\Theta+\|\mu-\mu_\epsilon^h\|_\Phi\le C\epsilon=Ch^\alpha\,. \tag{6.14}$$

Note that unlike for Theorem 5.2, the result (6.1) does not require that (3.2) is satisfied. Also, unlike for Theorem 5.3, we get better convergence rates and locking cannot occur.

6.2. Eliminate-then-discretize. Alternately, one could discretize (6.6) to obtain

$$\left(\begin{array}{cc}\mathbb{A}_1+\dfrac{1}{\epsilon}\mathbb{K}_1 & \dfrac{1}{\epsilon}\mathbb{K}_2\\ \dfrac{1}{\epsilon}\mathbb{K}_2^T & \mathbb{A}_2+\dfrac{1}{\epsilon}\widetilde{\mathbb{C}}\end{array}\right)\left(\begin{array}{c}\vec{\boldsymbol{\phi}}_\epsilon\\ \vec{\boldsymbol{\theta}}_\epsilon\end{array}\right)=\left(\begin{array}{c}\vec{\mathbf{f}}+\dfrac{1}{\epsilon}\widetilde{\vec{\mathbf{g}}}_1\\ \dfrac{1}{\epsilon}\widetilde{\vec{\mathbf{g}}}_2\end{array}\right). \tag{6.15}$$

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

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

and

$$\left\{\begin{array}{l}(\widetilde{\vec{\mathbf{g}}}_1)_j=\langle\widetilde{B}_1\phi_j,\widetilde{D}^{-1}\widetilde{g}_1\rangle_{\Phi^*,\Phi}=\langle B_1^*D^{-1}B_1\phi_j,\widetilde{D}^{-1}B_1^*D^{-1}g\rangle_{\Phi^*,\Phi}\\ (\widetilde{\vec{\mathbf{g}}}_2)_k=\langle\widetilde{B}_2\theta_k,\widetilde{D}^{-1}\widetilde{g}_2\rangle_{\Phi^*,\Phi}=\langle B_1^*D^{-1}B_2\theta_k,\widetilde{D}^{-1}B_1^*D^{-1}g\rangle_{\Phi^*,\Phi}\,.\end{array}\right.$$

THEOREM 6.2. *Let* (2.3), (2.5), *and* (3.3) *hold. Then, for* $0 < \epsilon \leq 1$, (6.15) *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* (2.6) *or equivalently, of* (5.8), *or equivalently, of* (5.10), *then there exist a constant* $C > 0$ *whose value is independent of* ϵ *and* h *such that*

$$\begin{aligned}\|\phi - \phi_\epsilon^h\|_\Phi + \|\theta - \theta_\epsilon^h\|_\Theta \leq C\epsilon \left(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\right) \\ + C\Big(1 + \frac{1}{\epsilon}\Big)\Big(\inf_{\widetilde{\phi}^h \in \Phi^h} \|\phi_\epsilon - \widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h \in \Theta^h} \|\theta_\epsilon - \widetilde{\theta}^h\|_\Theta\Big).\end{aligned} \tag{6.16}$$

Clearly, (6.11) and (6.15) are not the same. However, the coefficient matrices of both systems are symmetric and uniformly (with respect to h) positive definite without regards to (3.2).

7. Concluding discussion.

7.1. Preliminary comparison of the different approaches. In the preceding sections, we have discussed several ways to incorporate least-squares finite element notions into optimal control problems. We provide a summary list of the various possibilities. In addition to the various least-squares-related methods, we include the standard approach of applying a Galerkin finite element method to the optimality system obtained after applying the Lagrange multiplier rule to the optimization problem. In the list, equation references that are listed within parentheses correspond to equivalent formulations.

0. *Lagrange multiplier rule applied to the optimization problem followed by a mixed-Galerkin finite element discretization of the resulting optimality system*

$$\begin{aligned}&\{\text{optimization problem } (2.6\,, 2.10)\} \longrightarrow \textit{Lagrange multiplier rule} \longrightarrow \\ &\quad\{\text{optimality system } (4.1\,, 4.2)\} \longrightarrow \textit{Galerkin FE discretization} \longrightarrow \\ &\qquad\{\text{discrete equations } (4.3\,, 4.4)\}\end{aligned}$$

1. *Lagrange multiplier rule applied to the optimization problem followed by a least-squares formulation of the resulting optimality system followed by a finite element discretization*

$$\begin{aligned}&\{\text{optimization problem } (2.6\,, 2.10)\} \longrightarrow \textit{Lagrange multiplier rule} \longrightarrow \\ &\quad\{\text{optimality system } (4.1\,, 4.2)\} \longrightarrow \textit{least-squares formulation} \longrightarrow \\ &\qquad\{\text{least-squares optimality system } (4.17\,, 4.18)\} \longrightarrow \textit{FE discretization} \longrightarrow \\ &\qquad\quad\{\text{discrete system } (4.19\,, 4.21)\}\end{aligned}$$

2. *Lagrange multiplier rule applied to the optimization problem followed by a penalty perturbation of the resulting optimality system followed by a finite element discretization followed by the elimination of the discrete Lagrange multiplier*

$$\begin{aligned}&\{\text{optimization problem } (2.6\,, 2.10)\} \longrightarrow \textit{Lagrange multiplier rule} \longrightarrow \\ &\quad\{\text{optimality system } (4.1\,, 4.2)\} \longrightarrow \textit{penalty perturbation} \longrightarrow \\ &\qquad\{\text{perturbed optimality system } (5.10\,, 5.13)\} \longrightarrow \textit{FE discretization} \longrightarrow \\ &\qquad\quad\{\text{discrete system } (5.27\,, 5.28)\} \longrightarrow \textit{elimination of unknowns} \longrightarrow \\ &\qquad\qquad\{\text{reduced discrete system } (5.29)\}\end{aligned}$$

3. *Penalization of the cost functional by a least-squares functional followed by optimization followed by a finite element discretization of the resulting optimality equations*

$$\{\text{optimization problem } (2.6\,,2.10)\} \longrightarrow \textit{penalization of the cost functional} \longrightarrow$$
$$\{\text{penalized optimization problem } (5.1)\} \longrightarrow \textit{optimization} \longrightarrow$$
$$\{\text{reduced optimality system } (5.7\,,5.8)\} \longrightarrow \textit{FE discretization} \longrightarrow$$
$$\{\text{discrete system } (5.33\,,5.36)\}$$

or, equivalently, *the Lagrange multiplier rule applied to the optimization problem followed by a penalty perturbation of the resulting optimality system followed by the elimination of the Lagrange multiplier followed by a finite element discretization*

$$\{\text{optimization problem } (2.6\,,2.10)\} \longrightarrow \textit{Lagrange multiplier rule} \longrightarrow$$
$$\{\text{optimality system } (4.1\,,4.2)\} \longrightarrow \textit{penalty perturbation} \longrightarrow$$
$$\{\text{perturbed optimality system } (5.10\,,5.13)\} \longrightarrow \textit{elimination of unknowns} \longrightarrow$$
$$\{\text{reduced optimality system } (5.7\,,5.8)\} \longrightarrow \textit{FE discretization} \longrightarrow$$
$$\{\text{reduced discrete system } (5.33\,,5.36)\}$$

4. *Constraining the cost functional by a least-squares formulation of the state equations to obtain a modified optimization problem followed by the Lagrange multiplier rule to obtain an optimality system followed by a finite element discretization*

$$\{\text{modified optimization problem } (6.1\,,6.2)\} \longrightarrow \textit{Lagrange multiplier rule} \longrightarrow$$
$$\{\text{optimality system } (6.3)\} \longrightarrow \textit{FE discretization} \longrightarrow$$
$$\{\text{discrete system } (6.9)\}$$

5. *Constraining the cost functional by a least-squares formulation of the state equations to obtain a modified optimization problem followed by the Lagrange multiplier rule followed by a penalty perturbation of the resulting optimality system followed by a finite element discretization followed by the elimination of the discrete Lagrange multiplier*

$$\{\text{modified optimization problem } (6.1\,,6.2)\} \longrightarrow \textit{Lagrange multiplier rule} \longrightarrow$$
$$\{\text{optimality system } (6.3)\} \longrightarrow \textit{penalty perturbation} \longrightarrow$$
$$\{\text{perturbed optimality system } (6.8)\} \longrightarrow \textit{FE discretization} \longrightarrow$$
$$\{\text{discrete system } (6.10)\} \longrightarrow \textit{elimination of unknowns} \longrightarrow$$
$$\{\text{reduced discrete system } (6.11)\}$$

6. *Constraining the cost functional by a least-squares formulation of the state equations to obtain a modified optimization problem followed by penalization of the cost functional followed by optimization followed by a finite element discretization of the resulting optimality equations*

$$\{\text{modified optimization problem } (6.1\,,6.2)\} \longrightarrow \textit{penalize the cost functional} \longrightarrow$$
$$\{\text{penalized optimization problem } (6.5)\} \longrightarrow \textit{optimization} \longrightarrow$$
$$\{\text{reduced optimality system } (6.6\,,6.7)\} \longrightarrow \textit{FE discretization} \longrightarrow$$
$$\{\text{discrete system } (6.15)\}$$

In Table 1, we compare the seven methods just listed with respect to several desirable properties. The properties are posed in the form of the following questions.

discrete inf-sup not required – are the finite element spaces required to satisfy (3.2) in order that the resulting discrete systems be stably invertible as $h \to 0$?

locking impossible – is it possible to guarantee that the discrete systems are stably invertible as $\epsilon \to 0$ with fixed h?

optimal error estimate – are optimal estimates for the error in the approximate solutions obtainable, possibly after choosing ϵ to depend on h?

symmetric matrix system – are the discrete systems symmetric?

reduced number of unknowns – is it possible to eliminate unknowns to obtain a smaller discrete system?

positive definite matrix system – do the discrete systems, possible after the elimination of unknowns, have a positive definite coefficient matrix?

TABLE 1
Properties of different approaches for the approximate solution of the optimization problem.

	Method						
	0	1	2	3	4	5	6
discrete inf-sup not required	×	√	×	√	√	√	√
locking impossible	√	√	√	×	√	√	×
optimal error estimate	√	√	√	×	√	√	×
symmetric matrix system	√	√	√	√	√	√	√
reduced number of unknowns	×	×	√	√	×	√	√
positive definite matrix system	×	√	√	√	×	√	√

From Table 1, we see that only approach 5 has all its boxes checked, so that as far as the six properties used for comparison purposes in that table, that approach seems preferable. However, there are other issues that arise in the practical implementation of this and other methods that can influence the choice of a "best" method. In §7.2, we discuss some of these issues in the context of concrete examples. When including practical considerations, it seems that Method 1 is also a good candidate. It is probably the case that there is no universal "best" way to incorporate least-square notions into control and optimization problems.

7.2. Some practical issues arising in implementations. One difficulty that arises in the implementation of Method 5 and, indeed, of the other methods we have discussed is that, in concrete practical settings such as the Stokes equations, $H^{-1}(\Omega)$ norms appear in the least-squares functional (4.16). For example, for Method 5, this leads to the appearance of the $H^{-1}(\Omega)$ inner product in the definition of the matrices and vectors that form the discrete system. 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] in the context of Method 1 of §7.1; 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 $h \to 0$ or with respect to ϵ as $\epsilon \to 0$. First, let us discuss the $h \to 0$ 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 (3.14) and (3.15) where one sees that the operator $\widetilde{B}_1$ that results from applying the least-squares principle (3.7) 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. The idea here is that after "squaring" first-order operators, one obtains second-order operators 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 (6.7); we see there operators such as $B_1^* D^{-1} B_1 \widetilde{D}^{-1} B_1^* D^{-1} B_1$ 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 $\widetilde{D}^{-1}$. Given the nature of all these operators, it is not at all clear that the h-condition number of the discrete systems of §§4.2, 5, and 6 are similar to those that result from a naive double "squaring" of first-order operators; indeed, norm equivalence relations such as (3.13) and (4.13) 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 $\epsilon \to 0$ is problematic for all penalty methods, even for those for which locking does not occur. Note that to obtain a result such as (6.14), 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., [16] and also [14, 18, 19]. To this end, let $\{\vec{\boldsymbol{\phi}}_\epsilon, \vec{\boldsymbol{\theta}}_\epsilon, \vec{\boldsymbol{\mu}}_\epsilon\}$ denote the solution of (6.10) and set $\vec{\boldsymbol{\phi}}^{(0)} = \vec{\boldsymbol{\phi}}_\epsilon$, $\vec{\boldsymbol{\theta}}^{(0)} = \vec{\boldsymbol{\theta}}_\epsilon$, and $\vec{\boldsymbol{\mu}}^{(0)} = \vec{\boldsymbol{\mu}}_\epsilon$. Then, for $n \geq 1$, we solve the sequence of problems

$$\begin{pmatrix} \mathbb{A}_1 & 0 & \widetilde{\mathbb{B}}_1 \\ 0 & \mathbb{A}_2 & \widetilde{\mathbb{B}}_2^T \\ \widetilde{\mathbb{B}}_1 & \widetilde{\mathbb{B}}_2 & -\epsilon\widetilde{\mathbb{D}} \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}}^{(n)} \\ \vec{\boldsymbol{\theta}}^{(n)} \\ \vec{\boldsymbol{\mu}}^{(n)} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{0}} \\ \vec{\mathbf{0}} \\ -\epsilon\widetilde{\mathbb{D}}\vec{\boldsymbol{\mu}}^{(n-1)} \end{pmatrix}. \tag{7.1}$$

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

$$\vec{\boldsymbol{\phi}}_{\epsilon,N} = \sum_{n=0}^{N} \vec{\boldsymbol{\phi}}^{(n)}, \qquad \vec{\boldsymbol{\theta}}_{\epsilon,N} = \sum_{n=0}^{N} \vec{\boldsymbol{\theta}}^{(n)}, \qquad \text{and} \qquad \vec{\boldsymbol{\mu}}_{\epsilon,N} = \sum_{n=0}^{N} \vec{\boldsymbol{\mu}}^{(n)} \tag{7.2}$$

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 (7.2). Then, instead of the estimate (6.13), one obtains the estimate (see, e.g., [16] and also [14, 18])

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

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 (6.14). These estimates tell us that we can 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 $\epsilon << 1$. Making the latter choice allows us obtain the same effect but with a much larger value for ϵ.

Note that $\vec{\boldsymbol{\mu}}^{(n)}$ may be eliminated from (7.1) to yield a reduced system with fewer unknowns. Thus, the iteration to compute the pairs $\{\vec{\boldsymbol{\phi}}^{(n)}, \vec{\boldsymbol{\theta}}^{(n)}\}$ for $n = 0, 1, \ldots$, using reduced systems proceeds as follows. Let $\vec{\boldsymbol{\phi}}^{(0)} = \vec{\boldsymbol{\phi}}_\epsilon$ and $\vec{\boldsymbol{\theta}}^{(0)} = \vec{\boldsymbol{\theta}}_\epsilon$, where $\vec{\boldsymbol{\phi}}_\epsilon$ and $\vec{\boldsymbol{\theta}}_\epsilon$ denote the solution of (6.11), and then set

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

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

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

In order to define the next iterate, we set

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

REFERENCES

[1] D. Bedivan and G. J. Fix, *Least-squares methods for optimal shape design problems*, Comp. Math. Appl., 30 (1995), pp. 17–25.

[2] P. Bochev, *Least-squares methods for optimal control*, Non. Anal., Theo, Meth. Appl., 30 (1997), pp. 1875–1885.

[3] P. Bochev, *Negative norm least-squares methods for the velocity-vorticity-pressure Navier-Stokes equations*, Num. Meth. Part. Diff. Eqs., 15 (1999), pp. 237–256.

[4] P. Bochev and D. Bedivan, *Least-squares methods for Navier-Stokes boundary control problems*, Int. J. Comp. Fluid Dyn., 9 (1997), pp. 43–58.

[5] P. Bochev and M. Gunzburger, *Least-squares finite element methods for elliptic equations*, SIAM Rev., 40 (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,* Comp. Math. Appl., 48:7 (2004), pp. 1035–1057.

[8] P. Bochev and M. Gunzburger, *Least-squares finite element methods for optimality systems arising in optimization and control problems*, SIAM J. Num. Anal. 43:6, pp. 2517–2543.

[9] P. Bochev and M. Gunzburger, *A least-squares finite element method for optimization and control problems*, to appear.

[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, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, RAIRO Anal. Numer. R2, 21 (1974), pp. 129–151.

[14] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.

[15] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Num. Anal., 37 (2000), pp. 1542–1570.

[16] 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.

[17] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.

[18] 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.
[19] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic, Boston, 1989.
[20] M. Gunzburger, *Perspectives in Flow Control and Optimization*, SIAM, Philadelphia, 2003.
[21] 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.
[22] 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.
[23] H. Schlichting and K. Gersten, *Boundary Layer Theory*, Springer, Berlin, 2000.