LEAST-SQUARES FINITE ELEMENT METHODS
FOR OPTIMAL DESIGN AND CONTROL PROBLEMS

Max Gunzburger
School of Computational Science
Florida State University
gunzburg@scs.fsu.edu

in collaboration with
Pavel Bochev
Computational Mathematics and Algorithms Department
Sandia National Laboratories
pbboche@sandia.gov

Last updated in 2006
CONTROL AND OPTIMIZATION PROBLEMS

- The optimal control or optimization problems we consider consist of the following ingredients
  - state variables – variables that describe the system being modeled
  - control variables or design parameters – variables at our disposal that can be used to affect the state variables
  - a state system – partial differential equations relating the state and control variables
  - a functional of the state and control variables

- We consider the constrained optimization problem
  find state and control variables that minimize the given functional subject to the state system being satisfied
• We restrict attention to (generalizations are entirely possible)
  – linear, elliptic state systems
  – quadratic functionals

• The **Lagrange multiplier rule** is a standard approach for solving constrained optimization problems
  – applying the rule results in an **optimality system** whose solutions also solve the optimization problem
  – the optimality system consists of
    - the **state system** – the given partial differential equations that relate the unknown state and control variables
    - an **adjoint** or **co-state system** – partial differential equations involving the adjoint operator of the state system
    - an **optimality condition** – reflects the fact that the gradient of the functional vanishes for optimal values of the state and control variables
in the current context, the optimality system is a symmetric and \textit{weakly coercive linear system} in the state, adjoint-state (Lagrange multiplier), and control variables.

in the context of finite element methods,
- optimality systems are usually discretized using \textit{Galerkin methods}
- typically resulting in saddle-point type matrix problems that are symmetric and \textit{indefinite}

we will discretize the optimality system via a \textit{least-squares finite element method}
- resulting in matrix problems that are symmetric and \textit{positive definite}
- other advantages over the Galerkin approach will also accrue
Another popular approach for solving constrained optimization problems are penalty methods.

- we will briefly consider this approach in conjunction with least-square finite element methods for the solution of control and optimization problems.
We begin with four given Hilbert spaces

\[
\begin{align*}
\Theta &= \text{the control space} \\
\Phi &= \text{the state space (solution space for the PDE)} \\
\hat{\Phi} &= \text{a data space} \\
\tilde{\Phi} &= \text{a pivot space}
\end{align*}
\]

along with their dual spaces denoted by \((\cdot)^*\)

- we assume that

\[
\Phi \subseteq \hat{\Phi} \subseteq \tilde{\Phi} \subseteq \hat{\Phi}^* \subseteq \Phi^* \quad \text{with continuous embeddings}
\]

and

\[
\langle \psi, \phi \rangle_{\Phi^*, \Phi} = \langle \psi, \phi \rangle_{\hat{\Phi}^*, \hat{\Phi}} = (\psi, \phi)_{\tilde{\Phi}} \quad \forall \psi \in \hat{\Phi}^* \subseteq \Phi^* \quad \text{and} \quad \forall \phi \in \Phi \subseteq \hat{\Phi}
\]
often, the control set $\Theta$ is chosen to be a bounded set in a Hilbert space
- we consider the less general situation of $\Theta$ itself being a Hilbert space

• We define the quadratic functional

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

where

- $\phi = \text{the state variable}$
- $\theta = \text{the control variable}$
- $a_1(\cdot, \cdot)$ is a symmetric bilinear form on $\hat{\Phi} \times \hat{\Phi}$
- $a_2(\cdot, \cdot)$ is a symmetric bilinear form on $\Theta \times \Theta$
- $\hat{\phi} \in \hat{\Phi}$ is a given function

- often, the control space $\Theta$ is finite dimensional in which case $\theta$ is usually referred to as the vector of design variables
– the second term in the functional can be interpreted as a penalty term which limits the size of the control $\theta$

– we make the following assumptions about the bilinear forms

\[
\begin{align*}
  a_1(\phi, \mu) &\leq C_1 \|\phi\|_{\hat{\Phi}} \|\mu\|_{\hat{\Phi}} & \forall \phi, \mu \in \hat{\Phi} & \quad \text{← continuity} \\
  a_2(\theta, \nu) &\leq C_2 \|\theta\|_{\Theta} \|\nu\|_{\Theta} & \forall \theta, \nu \in \Theta & \quad \text{← continuity} \\
  a_1(\phi, \phi) &\geq 0 & \forall \phi \in \hat{\Phi} & \quad \text{← non-negativity} \\
  a_2(\theta, \theta) &\geq K_2 \|\theta\|_{\Theta}^2 & \forall \theta \in \Theta & \quad \text{← coercivity}
\end{align*}
\]
• Given

another Hilbert space $\Lambda$

a bilinear form $b_1(\cdot, \cdot)$ on $\Phi \times \Lambda$

a bilinear form $b_2(\cdot, \cdot)$ on $\Theta \times \Lambda$

a function $g \in \Lambda^*$

we define the **linear constraint equation**

$$b_1(\phi, \psi) + b_2(\theta, \psi) = \langle g, \psi \rangle_{\Lambda^*, \Lambda} \quad \forall \psi \in \Lambda$$
we make the following assumptions about the bilinear forms

\[
\begin{align*}
  b_1(\phi, \psi) & \leq c_1 \|\phi\|_\Phi \|\psi\|_\Lambda & \forall \phi \in \Phi, \psi \in \Lambda & \leftarrow \text{continuity} \\
  b_2(\theta, \psi) & \leq c_2 \|\psi\|_\Phi \|\theta\|_\Theta & \forall \theta \in \Theta, \psi \in \Lambda & \leftarrow \text{continuity} \\
  \sup_{\psi \in \Lambda, \psi \neq 0} \frac{b_1(\phi, \psi)}{\|\psi\|_\Lambda} & \geq k_1 \|\phi\|_\Phi & \forall \phi \in \Phi & \leftarrow \text{weak coercivity} \\
  \sup_{\phi \in \Phi, \phi \neq 0} \frac{b_1(\phi, \psi)}{\|\phi\|_\Phi} & > 0 & \forall \psi \in \Lambda & \leftarrow \text{weak coercivity}
\end{align*}
\]

these assumptions suffice to guarantee that,

given any control \(\theta \in \Theta\),

the constraint equation is uniquely solvable for \(\phi \in \Phi\)
• We consider the optimal control problem

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

– we will transform this problem into a more familiar constrained optimization problem

- we will then be able to apply the well-known Brezzi theory for the approximation of saddle point problems

• In defining least-squares finite element methods, it is useful to also view our optimization problem in operator notation

- so we recast our problem to fit that view
• The bilinear forms serve to define operators

\[ A_1 : \hat{\Phi} \rightarrow \hat{\Phi}^*, \quad A_2 : \Theta \rightarrow \Theta^*, \quad B_1 : \Phi \rightarrow \Lambda^*, \]
\[ B_1^* : \Lambda \rightarrow \Phi^*, \quad B_2 : \Theta \rightarrow \Lambda^*, \quad B_2^* : \Lambda \rightarrow \Theta^* \]

through the relations

\[ a_1(\phi, \mu) = \langle A_1 \phi, \mu \rangle_{\hat{\Phi}^*, \hat{\Phi}} \quad \forall \phi, \mu \in \hat{\Phi} \]
\[ a_2(\theta, \nu) = \langle A_2 \theta, \nu \rangle_{\Theta^*, \Theta} \quad \forall \theta, \nu \in \Theta \]
\[ b_1(\phi, \psi) = \langle B_1 \phi, \psi \rangle_{\Lambda^*, \Lambda} = \langle B_1^* \psi, \phi \rangle_{\Phi^*, \Phi} \quad \forall \phi \in \Phi, \psi \in \Lambda \]
\[ b_2(\psi, \theta) = \langle B_2 \theta, \psi \rangle_{\Lambda^*, \Lambda} = \langle B_2^* \psi, \theta \rangle_{\Theta^*, \Theta} \quad \forall \theta \in \Theta, \psi \in \Lambda \]

– the assumptions we have made imply that the operators are bounded with

\[ \| A_1 \|_{\hat{\Phi} \rightarrow \hat{\Phi}^*} \leq C_1, \quad \| A_2 \|_{\Theta \rightarrow \Theta^*} \leq C_2, \quad \| B_1 \|_{\Phi \rightarrow \Phi^*} \leq c_1, \]
\[ \| B_1^* \|_{\Phi^* \rightarrow \Phi} \leq c_1, \quad \| B_2 \|_{\Phi \rightarrow \Theta^*} \leq c_2, \quad \| B_2^* \|_{\Theta \rightarrow \Phi^*} \leq c_2 \]

and that the operator \( B_1 \) is invertible with

\[ \| B_1^{-1} \|_{\Lambda^* \rightarrow \Phi} \leq 1/k_1 \]
we also have that the functional can be expressed as

\[ J(\phi, \theta) = \frac{1}{2} \langle A_1(\phi - \hat{\phi}), (\phi - \hat{\phi}) \rangle_{\Phi^*, \hat{\Phi}} + \frac{1}{2} \langle A_2 \theta, \theta \rangle_{\Theta^*, \Theta} \quad \forall \phi \in \Phi, \theta \in \Theta \]

the constraint equation can be expressed as

\[ B_1 \phi + B_2 \theta = g \quad \text{in} \ \Lambda^* \]

and the optimal control problem takes the form

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

– note that for any given control \( \theta \in \Theta \), there corresponds the unique state

\[ \phi = B_1^{-1}(g - B_2 \theta) \in \Phi \]
Recasting the abstract optimization problem as a classical saddle point problem

Let
\[
V \equiv \Phi \times \Theta \quad \text{(the state and control spaces)}
\]
\[
S \equiv \Lambda \quad \text{(the Lagrange multiplier space)}
\]
and let
\[
\|\{\phi, \theta\}\|_V = \sqrt{\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2} \quad \forall \{\phi, \theta\} \in V
\]

Let
\[
a(\{\phi, \theta\}, \{\mu, \nu\}) \equiv a_1(\phi, \mu) + a_2(\theta, \nu) \quad \forall \phi, \mu \in \Phi, \ \theta, \nu \in \Theta
\]
\[
b(\{\phi, \theta\}, \{\psi\}) \equiv b_1(\phi, \psi) + b_2(\theta, \psi) \quad \forall \phi \in \Phi, \ \theta \in \Theta, \ \psi \in \Lambda
\]
\[
\langle f, \{\mu, \nu\}\rangle_{V^*, V} \equiv a_1(\mu, \hat{\phi}) \quad \forall \mu \in \Phi, \ \nu \in \Theta
\]
\[
t = \frac{1}{2}a_1(\hat{\phi}, \hat{\phi})
\]
• Let

\[ u = \{\phi, \theta\} \in V \]

\[ q = \{\psi\} \in S \]

• It is simple to see that our functional can be expressed as

\[ \mathcal{J}(u) = \frac{1}{2} a(u, u) - \langle f, u \rangle_{V^*, V} + t \quad \forall u \in V \]

the linear constraint equation can be expressed as

\[ b(u, q) = \langle g, q \rangle_{S^*, S} \quad \forall q \in S \]

and the constrained minimization problem can be recast in the form

\[ \min_{u \in V} \mathcal{J}(u) \quad \text{subject to} \quad b(u, q) = \langle g, q \rangle_{S^*, S} \quad \forall q \in S \]

• This is exactly a classical saddle point setting to which we can apply the Brezzi theory

---

1It is easy to show that \( t \leq (C_1/2) \) and \( \|f\|_{V^*} \leq C_1 \|\hat{\phi}\|_{\hat{\Phi}} \) so that \( f \in V^* \)
Applying the Brezzi theory

• Let the subspace $Z \subset V$ be defined by by

$$Z = \{ v \in V : \ b(v, q) = 0 \ \forall q \in S \}$$

– the subspace $Z \subset V = \Phi \times \Theta$ has the equivalent definition

$$Z = \{ \{ \phi, \theta \} \in \Phi \times \Theta : \ \phi = -B_1^{-1}B_2\theta \ \forall \theta \in \Theta \}$$

• Then, it can be shown that the assumptions we have made imply that

$$\begin{align*}
    a(u, v) &\leq C_a \| u \|_V \| v \|_V & \forall u, v \in V \\
    b(u, q) &\leq C_b \| u \|_V \| q \|_S & \forall u \in V, \ q \in S \\
    a(u, u) &\geq 0 & \forall u \in V \\
    a(u, u) &\geq K_a \| u \|_V^2 & \forall u \in Z \\
    \sup_{v \in V, v \neq 0} \frac{b(v, q)}{\| v \|_V} &\geq K_b \| q \|_S & \forall q \in S
\end{align*}$$
These are exactly the assumptions of the Brezzi theory; they guarantee that

– the constrained optimization problem

\[
\min_{u \in V} J(u) \quad \text{subject to} \quad b(u, q) = \langle g, q \rangle_{S^*, S} \quad \forall q \in S
\]

has a unique solution \( u \in V \)

– that solution can be determined by solving the optimality system

\[
\begin{cases}
    a(u, v) + b(v, p) = \langle f, v \rangle_{V^*, V} & \forall v \in V \\
    b(u, q) = \langle g, q \rangle_{S^*, S} & \forall q \in S
\end{cases}
\]

where \( p \in S \) is a Lagrange multiplier introduced to enforce the constraint equation

– the optimality system has a unique solution \((u, p) \in V \times S\) and, moreover,

\[
\|u\|_V + \|p\|_S \leq C(\|f\|_{V^*} + \|g\|_{S^*})
\]
• Due to the equivalence of the classical saddle point problem just considered and our abstract optimal control problem, we can easily verify that

– the optimal control problem

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

has a unique solution \((\phi, \theta) \in \Phi \times \Theta\)

– that solution can be determined by solving the optimality system

\[
\begin{aligned}
& a_1(\phi, \mu) + b_1(\mu, \lambda) = a_1(\hat{\phi}, \mu) \quad \forall \mu \in \Phi \\
& a_2(\theta, \nu) + b_2(\nu, \lambda) = 0 \quad \forall \nu \in \Theta \\
& b_1(\phi, \psi) + b_2(\theta, \psi) = \langle g, \psi \rangle_{\Lambda^*, \Lambda} \quad \forall \psi \in \Lambda
\end{aligned}
\]

where \(\lambda \in \Lambda\) is the Lagrange multiplier introduced to enforce the constraint equation

– the optimality system has a unique solution \((\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda\) and, moreover,

\[
\|\phi\|_\Phi + \|\theta\|_\Theta + \|\lambda\|_\Lambda \leq C(\|g\|_{\Lambda^*} + \|\hat{\phi}\|_{\hat{\Phi}})
\]
• The partitioning

\[
\begin{pmatrix}
  a(u, v) + b(v, p) \\
  b(u, q)
\end{pmatrix} = \begin{pmatrix}
  a_1(\phi, \mu) & + & b_1(\mu, \lambda) \\
  a_2(\theta, \nu) & + & b_2(\nu, \lambda) \\
  b_1(\phi, \psi) & + & b_2(\theta, \psi)
\end{pmatrix}
\]

makes the correspondence between the classical Brezzi setting and our optimal control setting clear

• Using the operators previously introduced, the optimality system takes the equivalent operator form

\[
\begin{cases}
  A_1\phi + B_1^*\lambda = A_1\hat{\phi} & \text{in } \Phi^* \\
  A_2\theta + B_2^*\lambda = 0 & \text{in } \Theta^* \\
  B_1\phi + B_2\theta = g & \text{in } \Lambda^*
\end{cases}
\]
We choose (conforming) finite dimensional subspaces

$$\Phi^h \subset \Phi \quad \Theta^h \subset \Theta \quad \Lambda^h \subset \Lambda$$

and then restrict the optimality system to the subspaces, i.e.,

- we seek $$(\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$$ that satisfies

$$\begin{cases}
  a_1(\phi^h, \mu^h) + b_1(\mu^h, \lambda^h) = a_1(\hat{\phi}, \mu^h) & \forall \mu^h \in \Phi^h \\
  a_2(\theta^h, \nu^h) + b_2(\nu^h, \lambda^h) = 0 & \forall \nu^h \in \Theta^h \\
  b_1(\phi^h, \psi^h) + b_2(\theta^h, \psi^h) = \langle g, \psi^h \rangle_{\Lambda^*, \Lambda} & \forall \psi^h \in \Lambda^h
\end{cases}$$
– this discrete optimality system is also the optimality system for the minimization of the functional

\[
\min_{\phi^h \in \Phi^h, \theta^h \in \Theta^h} \mathcal{J}(\phi^h, \theta^h)
\]

subject to the constraint

\[
b_1(\phi^h, \psi^h) + b_2(\psi^h, \theta^h) = \langle g, \psi^h \rangle_{\Lambda^*, \Lambda} \quad \forall \psi^h \in \Lambda^h
\]
• We make the following additional assumptions about the bilinear form $b_1(\cdot, \cdot)$ and the spaces $\Phi^h$ and $\Lambda^h$

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

- these discrete weak coercivity assumptions are not in general inherited from the corresponding assumptions previously made on the bilinear form $b_1(\cdot, \cdot)$ with respect to the spaces $\Phi$ and $\Lambda$

  - of course, an exception is the special case for which the bilinear form $b_1(\cdot, \cdot)$ is strongly coercive

- these assumptions are needed in order to guarantee that, given any $\theta^h \in \Theta^h$, the discrete state system (the third equation in the discrete optimality system) is uniquely solvable for $\phi^h \in \Phi^h$. 
We next define the subspace \( Z^h \subset \Phi^h \times \Theta^h \) by
\[
Z^h = \left\{ \{ \phi^h, \theta^h \} \in \Phi^h \times \Theta^h : b_1(\phi^h, \psi^h) + b_2(\theta^h, \psi^h) = 0 \quad \forall \psi^h \in \Lambda^h \right\}.
\]

Note that, in general, \( Z^h \nsubseteq Z \) even though \( \Phi^h \subset \Phi \), \( \Theta^h \subset \Theta \), and \( \Lambda^h \subset \Lambda \).

Now we can connect our discretized control problem to the saddle point problem that fits into the Brezzi theory

- let \( V^h = \Phi^h \times \Theta^h \) and \( S^h = \Lambda^h \)
  - we then have that \( V^h \subset V \) and \( \Lambda^h \subset \Lambda \)

- the discrete optimality system is then equivalent to the discrete saddle point system
\[
\begin{align*}
\begin{cases}
a(u^h, v^h) + b(v^h, p^h) &= \langle f, v^h \rangle_{V^*, V} \quad \forall v^h \in V^h \\
b(u^h, q^h) &= \langle g, q^h \rangle_{S^*, S} \quad \forall q^h \in S^h
\end{cases}
\end{align*}
\]
with the additional assumptions we have made on the bilinear form $b_1(\cdot, \cdot)$ and the spaces $\Phi^h$ and $\Lambda^h$, we can show that

$$a(u^h, u^h) \geq K^h_a \|u^h\|_V^2 \quad \forall u^h \in Z^h$$

and

$$\sup_{v^h \in V^h, v^h \neq 0} \frac{b(v^h, q^h)}{\|v^h\|_V} \geq K^h_b \|q^h\|_S \quad \forall q^h \in S^h$$

these are exactly the assumptions needed to apply the Brezzi theory to the discretized saddle point problem

from that theory, we conclude that

- the discretized saddle point problem
  has a unique solution $(u^h, p^h) \in V^h \times S^h$
  - moreover, we have that
    $$\|u^h\|_V + \|p^h\|_S \leq C \left( \|f\|_{V^*} + \|g\|_{S^*} \right).$$
  - furthermore, the following optimal error estimate holds
    $$\|u - u^h\|_V + \|p - p^h\|_S \leq C \left( \inf_{v^h \in V^h} \|u - v^h\|_V + \inf_{q^h \in S^h} \|p - q^h\|_S \right)$$
using these results and the equivalence of the discretized optimality and saddle point systems, we can easily obtain that

- the discrete optimality system has a unique solution
  \((\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h\)

- moreover
  \[\|\phi^h\|_\Phi + \|\theta^h\|_\Theta + \|\lambda^h\|_\Lambda \leq C(\|g\|_{\Lambda^*} + \|\hat{\phi}\|_{\hat{\Phi}})\]

- furthermore, the following optimal error estimate holds
  \[\|\phi - \phi^h\|_\Phi + \|\theta - \theta^h\|_\Theta + \|\lambda - \lambda^h\|_\Lambda \leq C\left(\inf_{\mu^h \in \Phi^h} \|\phi - \mu^h\|_\Phi + \inf_{\xi^h \in \Theta^h} \|\theta - \xi^h\|_\Theta + \inf_{\psi^h \in \Lambda^h} \|\lambda - \psi^h\|_\Lambda\right)\]
The discrete optimality system is equivalent to the linear system

\[
\begin{pmatrix}
  A_1 & 0 & B_1^T \\
  0 & A_2 & B_2^T \\
  B_1 & B_2 & 0
\end{pmatrix}
\begin{pmatrix}
  \vec{\phi} \\
  \vec{\theta} \\
  \vec{\lambda}
\end{pmatrix}
= 
\begin{pmatrix}
  \vec{f} \\
  \vec{0} \\
  \vec{g}
\end{pmatrix}
\]

where, once bases for the finite element spaces are chosen, the various vectors and matrices are defined from the corresponding linear functionals and bilinear forms in a standard manner.

- the assumptions we have made imply that
  - the matrix $A_2$ is symmetric and positive definite
  - the matrix $B_1$ is invertible

- these are sufficient to guarantee that the above optimality system coefficient matrix is itself invertible
  - note that that matrix is symmetric
• The saddle point nature of optimality system cannot be avoided
  – it occurs even when the state system is strongly coercive, i.e.,
    - even if the bilinear form $b_1(\cdot, \cdot)$ is coercive
    - even if the matrix $B_1$ is positive definite

• Thus, the Galerkin approach to discretization of the optimality system will always yield **indefinite matrix problems**
• Note that there are two layers of indefiniteness in the problem

  – the outer layer corresponds to the matrix
    \[
    \begin{pmatrix}
    A_1 & 0 & B_1^T \\
    0 & A_2 & B_2^T \\
    B_1 & B_2 & 0
    \end{pmatrix}
    \] that is indefinite regardless of the nature of the matrix $B_1$

  – the inner layer corresponds to the matrix $B_1$ itself that in general is merely invertible but is not necessarily symmetric or definite

  – for both layers, inf-sup conditions are needed to guarantee stable invertibility

  – it turns out that the inf-sup conditions on the bilinear form $b_1(\cdot, \cdot)$ (along with the coercivity of the bilinear form $a_2(\cdot, \cdot)$) are enough to insure the stable invertibility of both $B_1$ and the optimality system matrix
• The discretized optimality system is formidable
  
  – it is at least twice the size of the state system $\mathbb{B}_1$

  – solving it in one shot (as a fully coupled system) is often not practical

  – many strategies for “uncoupling” the block equations in the optimality
    system have been proposed
      
      - this, of course, necessitates an block-iterative approach
        to solving the discrete optimality system

  – alternately, iterative solution methods may be applied to the whole coupled
    system
      
      - such strategies are made “more difficult” by the indefiniteness of
        the whole coefficient matrix and by the indefiniteness of $\mathbb{B}_1$
• We start with the optimality system written in operator form

\[
\begin{align*}
A_{1}\phi + B_{1}^*\lambda &= A_{1}\hat{\phi} \quad \text{in } \Phi^* \\
A_{2}\theta + B_{2}^*\lambda &= 0 \quad \text{in } \Theta^* \\
B_{1}\phi + B_{2}\theta &= g \quad \text{in } \Lambda^*
\end{align*}
\]

• PDE theory tells that this system is well posed with respect to the

  – data space \( \Phi^* \times \Theta^* \times \Lambda^* \)
  – the solution space \( \Phi \times \Theta \times \Lambda \)
Thus, we define a least-squares functional by summing the squares of the data space norms of the residuals of the PDE’s in the optimality system

\[ K(\phi, \theta, \lambda; \hat{\phi}, g) = \| A_1 \phi + B_1^* \lambda - A_1 \hat{\phi} \|_{\Phi^*}^2 + \| A_2 \theta + B_2^* \lambda \|_{\Theta^*}^2 + \| B_1 \phi + B_2 \theta - g \|_{\Lambda^*}^2 \]

We then pose the least-squares minimization problem

\[ \min_{(\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda} K(\phi, \theta, \lambda; \hat{\phi}, g) \]

The solution of the least-squares principle is a solution of the optimality system and is therefore also a solution of the optimal control problem.
The first-order necessary conditions corresponding to the minimization problem are given by

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

where

$$B((\phi, \theta, \lambda), (\mu, \nu, \psi)) = (A_1\mu + B^*_1\psi, A_1\phi + B^*_1\lambda)_{\Phi^*}$$

$$+ (A_2\nu + B^*_2\psi, A_2\theta + B^*_2\lambda)_{\Theta^*} + (B_1\mu + B_2\nu, B_1\phi + B_2\theta)_{\Lambda^*}$$

$$\forall (\phi, \theta, \lambda), (\mu, \nu, \psi) \in \Phi \times \Theta \times \Lambda$$

and

$$F((\mu, \nu, \psi); (A_1\hat{\phi}, g)) = (A_1\mu + B^*_1\psi, A_1\hat{\phi})_{\Phi^*}$$

$$+ (B_1\mu + B_2\nu, g)_{\Lambda^*} \quad \forall (\mu, \nu, \psi) \in \Phi \times \Theta \times \Lambda$$
• The assumptions previously made can be used to show that

  – the bilinear form \( B(\cdot, \cdot) \) is symmetric and continuous on 
    \((\Phi \times \Theta \times \Lambda) \times (\Phi \times \Theta \times \Lambda)\)

  – the linear functional \( F(\cdot) \) is continuous on \((\Phi \times \Theta \times \Lambda)\)

• Moreover, the bilinear form \( B(\cdot, \cdot) \) is coercive on \((\Phi \times \Theta \times \Lambda)\)

\[
B((\phi, \theta, \lambda), (\phi, \theta, \lambda)) \geq C(\|\phi\|_{\Phi}^2 + \|\theta\|_{\Theta}^2 + \|\lambda\|_{\Lambda}^2) \quad \forall (\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda
\]
Note that since

\[ K(\phi, \theta, \lambda; 0, 0) = \| A_1 \phi + B^*_1 \lambda \|^2_{\Phi^*} + \| A_2 \theta + B^*_2 \lambda \|^2_{\Theta^*} + \| B_1 \phi + B_2 \theta \|^2_{\Lambda^*} \]

\[ = B((\phi, \theta, \lambda), (\phi, \theta, \lambda)) \]

the coercivity and continuity of the bilinear form \( B(\cdot, \cdot) \) are equivalent to stating that the functional \( K(\phi, \theta, \lambda; 0, 0) \) is norm-equivalent

- there exist constants \( C_1 > 0 \) and \( C_2 > 0 \) such that

\[ C_1 \left( \| \phi \|^2_{\Phi} + \| \theta \|^2_{\Theta} + \| \lambda \|^2_{\Lambda} \right) \leq K(\phi, \theta, \lambda; 0, 0) \leq C_2 \left( \| \phi \|^2_{\Phi} + \| \theta \|^2_{\Theta} + \| \lambda \|^2_{\Lambda} \right) \]

for all \( (\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda \)
We define a finite element discretization of the first-order necessary conditions by

- choosing conforming finite element subspaces
  \[ \Phi^h \subset \Phi, \Theta^h \subset \Theta, \text{ and } \Lambda^h \subset \Lambda \]

- then requiring that \((\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h\) satisfy
  \[ B((\phi^h, \theta^h, \lambda^h), (\mu^h, \nu^h, \psi^h)) = F((\mu^h, \nu^h, \psi^h); (A_1 \hat{\phi}, 0, g)) \]
  \[ \forall (\mu^h, \nu^h, \psi^h) \in \Phi^h \times \Theta^h \times \Lambda^h \]

- equivalently, this system can be obtained by minimizing the least-squares functional over the finite element space

Using the assumptions previously made, one can show that for any \((\hat{\phi}, g) \in \hat{\Phi}^* \times \Lambda^*\),

the discretized least-squares problem 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
\[
\|\phi - \phi^h\|_\Phi + \|\theta - \theta^h\|_\Theta + \|\lambda - \lambda^h\|_\Lambda \\
\leq C \left( \inf_{\tilde{\phi}^h \in \Phi^h} \|\phi - \tilde{\phi}^h\|_\Phi + \inf_{\tilde{\theta}^h \in \Theta^h} \|\theta - \tilde{\theta}^h\|_\Theta + \inf_{\tilde{\lambda}^h \in \Lambda^h} \|\lambda - \tilde{\lambda}^h\|_\Lambda \right)
\]

• The discrete problem is equivalent to a linear algebraic system of the form
\[
\begin{pmatrix}
K_1 & C_1^T & C_2^T \\
C_1 & K_2 & C_3^T \\
C_2 & C_3 & K_3
\end{pmatrix}
\begin{pmatrix}
\vec{\phi} \\
\vec{\theta} \\
\vec{\lambda}
\end{pmatrix}
= 
\begin{pmatrix}
\vec{f} \\
\vec{h} \\
\vec{g}
\end{pmatrix}
\]

• The coefficient matrix of this linear system is symmetric and positive definite
  – this should be compared to the linear system that results from a Galerkin finite element discretization of the optimality system for which the coefficient matrix is symmetric and indefinite
• The results about approximations of the least-squares principle do not require the inf-sup conditions

\[
\begin{align*}
\sup_{\psi^h \in \Lambda^h, \psi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\psi^h\|_{\Lambda}} & \geq k_h \|\phi^h\|_{\Phi} \quad \forall \phi^h \in \Phi^h \\
\sup_{\phi^h \in \Phi^h, \phi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\phi^h\|_{V}} & > 0 \quad \forall \psi^h \in \Lambda^h
\end{align*}
\]

that are required for the stability of Galerkin approximations

– this is one of the advantages of least-square finite element methods
• The discrete least-squares system can be viewed as a Galerkin discretization of the system

\[
\begin{align*}
(A_1^*A_1 + B_1^*B_1)\phi + (B_1^*B_2)\theta + (A_1^*B_1^*)\lambda &= (A_1^*A_1)\hat{\phi} + (B_1^*)g & \text{in } \Phi \\
(A_2^*A_2 + B_2^*B_2)\theta + (A_2^*B_2^*)\lambda + (B_2^*B_1^*)\phi &= (B_2^*)g & \text{in } \Theta \\
(B_1B_1^* + B_2B_2^*)\lambda + (B_1A_1^*)\phi + (B_2A_2^*)\theta &= (B_1A_1^*)\hat{\phi} & \text{in } \Lambda 
\end{align*}
\]

– the equations in this system are linear combinations of derivatives of the equations in the optimality system

– it is clear then that the discrete least-squares system 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 (the PDEs) in optimization or control problem
  - in fact, it leads one to
    cast the constraint PDEs into first-order systems
• Obviously, since we have the symmetry and positive definiteness of the matrix

\[
\begin{pmatrix}
K_1 & C_1^T & C_2^T \\
C_1 & K_2 & C_3^T \\
C_2 & C_3 & K_3
\end{pmatrix}
\]

so are the matrices \(K_1, K_2,\) and \(K_3\)

• In addition, the matrices \(K_1\) and \(K_3\) are differentially dominant over the matrices \(C_1, C_2,\) and \(C_3\)

• The matrix \(K_2\) often can
  – also be differentially dominant
    - in which case the whole coefficient matrix is block differentially diagonally dominant
  or
  – be small due to the fact that the control vector \(\tilde{\theta}\) involves only a few design parameters
    - in which case it can be explicitly inverted
• The discrete equations emanating from the least-squares finite element methods also form a formidable system so that uncoupling or iterative solution strategies are called for here as well.

• For the Galerkin discrete system

\[
\begin{pmatrix}
A_1 & 0 & B_1^T \\
0 & A_2 & B_2^T \\
B_1 & B_2 & 0
\end{pmatrix}
\begin{pmatrix}
\phi \\
\theta \\
\lambda
\end{pmatrix}
= 
\begin{pmatrix}
f \\
0 \\
g
\end{pmatrix}
\]

uncoupling approaches rely on the invertibility of the matrices \( B_1 \) and \( A_2 \) - the first of these is, in general, non-symmetric and indefinite.

• For the least-squares discrete system

\[
\begin{pmatrix}
K_1 & C_1^T & C_2^T \\
C_1 & K_2 & C_3^T \\
C_2 & C_3 & K_3
\end{pmatrix}
\begin{pmatrix}
\phi \\
\theta \\
\lambda
\end{pmatrix}
= 
\begin{pmatrix}
f \\
h \\
g
\end{pmatrix}
\]

uncoupling strategies rely on the invertibility of the matrices \( K_1, K_2, \) and \( K_3 \).
- all three of these matrices are symmetric and positive definite even when the discrete inf-sup condition on the bilinear form $b_1(\cdot, \cdot)$ is not satisfied

- An example of a simple uncoupling strategy is to apply a block-Gauss Seidel method to the discrete equations

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

  $\mathbb{K}_3 \vec{\lambda}^{(k+1)} = \vec{g} - \mathbb{C}_2 \vec{\phi}^{(k)} - \mathbb{C}_3 \vec{\theta}^{(k)}$

  $\mathbb{K}_1 \vec{\phi}^{(k+1)} = \vec{f} - \mathbb{C}_1^\top \vec{\theta}^{(k)} - \mathbb{C}_2^\top \vec{\lambda}^{(k+1)}$

  $\mathbb{K}_2 \vec{\theta}^{(k+1)} = \vec{h} - \mathbb{C}_1 \vec{\phi}^{(k+1)} - \mathbb{C}_3^\top \vec{\lambda}^{(k+1)}$

  until satisfactory convergence is achieved

  – the matrices $\mathbb{K}_3$, $\mathbb{K}_1$, and $\mathbb{K}_2$ are all positive definite so that efficient solution strategies are easy to devise

  – of course, more sophisticated uncoupling strategies can also be defined
In the least-squares case, the fact that the matrix
\[
\begin{pmatrix}
K_1 & C_1^T & C_2^T \\
C_1 & K_2 & C_3^T \\
C_2 & C_3 & K_3
\end{pmatrix}
\begin{pmatrix}
\vec{\phi} \\
\vec{\theta} \\
\vec{\lambda}
\end{pmatrix} =
\begin{pmatrix}
\vec{f} \\
\vec{h} \\
\vec{g}
\end{pmatrix}
\]
is positive definite and block-differentially diagonally dominant may have useful implications for iterative solution strategies.
EXAMPLE: OPTIMIZATION AND CONTROL PROBLEMS FOR THE STOKES EQUATIONS

• The Stokes system

\[
\begin{aligned}
-\Delta u + \nabla p + \theta &= g \\
\nabla \cdot u &= 0 \\
\end{aligned}
\]

in \(\Omega\), \quad u = 0 \quad \text{on} \quad \Gamma, \quad \int_\Omega p \, d\Omega = 0

where \(g\) is given function

• The functionals

Case I: \( J_1(u, \theta) = \frac{1}{2} \int_\Omega |\nabla \times u|^2 \, d\Omega + \frac{\delta}{2} \int_\Omega |\theta|^2 \, d\Omega \)

Case II: \( J_2(u, \theta; \hat{u}) = \frac{1}{2} \int_\Omega |u - \hat{u}|^2 \, d\Omega + \frac{\delta}{2} \int_\Omega |\theta|^2 \, d\Omega \)

where \(\hat{u}\) is given function and \(\delta\) a given parameter
We study the two problems of finding states \((u, p)\) and controls \(\theta\) that minimize either of the functionals subject to the Stokes system being satisfied.

Since least-squares finite element methods involve “squaring” operators, we recast the Stokes equations into a first-order system:

- we choose the velocity-vorticity-pressure formulation

\[
\begin{aligned}
\nabla \times \omega + \nabla p + \theta &= g \\
\nabla \cdot u &= 0 \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \Gamma, \\
\n\int_{\Omega} p \, d\Omega &= 0 \\
\n\nabla \times u - \omega &= 0
\end{aligned}
\]

- note that the first functional can now be written as

Case I: \(J_1(\omega, \theta) = \frac{1}{2} \int_{\Omega} |\omega|^2 \, d\Omega + \frac{\delta}{2} \int_{\Omega} |\theta|^2 \, d\Omega\)
Precise statement of the optimal control problems

- **Spaces**
  
  state and co-state space: \( \Phi = \Lambda = H^1_0(\Omega) \times L^2(\Omega) \times L^2_0(\Omega) \)
  
  control space: \( \Theta = L^2(\Omega) \)
  
  dual spaces: \( \Phi^* = \Lambda^* = H^{-1}(\Omega) \times L^2(\Omega) \times L^2_0(\Omega) \) and \( \Theta^* = L^2(\Omega) \)
  
  other spaces: \( \hat{\Phi} = \tilde{\Phi} = L^2(\Omega) \times L^2(\Omega) \times L^2_0(\Omega) \)

- **Functions**
  
  trial functions: \( \phi = \{u, \omega, p\} \in \Phi \quad \theta = \{\theta\} \in \Theta \quad \lambda = \{v, \sigma, q\} \in \Lambda \)
  
  test functions: \( \mu = \{\tilde{u}, \tilde{\omega}, \tilde{p}\} \in \Phi \quad \nu = \{\tilde{\theta}\} \in \Theta \quad \psi = \{\tilde{v}, \tilde{\sigma}, \tilde{r}\} \in \Lambda \)
  
  data functions: \( g = \{g, 0, 0\} \in \Lambda^* \quad \hat{\phi} = \{\tilde{u}, 0, 0\} \in \hat{\Phi} \)
• Bilinear forms

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

\[ a_2(\theta, \nu) = \delta(\theta, \tilde{\theta}) \quad \forall \theta = \{\theta\} \in \Theta, \quad \nu = \{\tilde{\theta}\} \in \Theta \]

\[ b_1(\phi, \psi) = (\omega, \nabla \times \tilde{\nu}) - (p, \nabla \cdot \tilde{v}) + (\nabla \times u - \omega, \tilde{\sigma}) - (\nabla \cdot u, \tilde{r}) \quad \forall \phi = \{u, \omega, p\} \in \Phi, \quad \psi = \{\tilde{v}, \tilde{\sigma}, \tilde{r}\} \in \Lambda \]

\[ b_2(\theta, \psi) = (\theta, \tilde{v}) \quad \forall \theta = \{\theta\} \in \Theta, \quad \psi = \{\tilde{v}, \tilde{\sigma}, \tilde{r}\} \in \Lambda \]

• Linear functional

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

for \( g \in H^{-1}(\Omega) \)
• Operators

\[
A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for Case I} \quad A_1 = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for Case II}
\]

\[
A_2 = \delta I \quad B_1 = \begin{pmatrix} 0 & \nabla \times & \nabla \\ \nabla \times & -I & 0 \\ -\nabla \cdot & 0 & 0 \end{pmatrix} \quad B_2 = \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix}
\]

• Least-squares functional

\[
\mathcal{K}(\{\mathbf{u}, \mathbf{\omega}, p\}, \theta, \{\mathbf{v}, \mathbf{\sigma}, q\}; \hat{\mathbf{u}}, g)
= \|\nabla \times \mathbf{\sigma} + \nabla q + \delta_2(\mathbf{u} - \hat{\mathbf{u}})\|_{-1}^2 + \|\nabla \times \mathbf{v} - \mathbf{\sigma} + \delta_1 \mathbf{\omega}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2
+ \|\delta \theta + \mathbf{v}\|_0^2
+ \|\nabla \times \mathbf{\omega} + \nabla p + \theta - g\|_{-1}^2 + \|\nabla \times \mathbf{u} - \mathbf{\omega}\|_0^2 + \|\nabla \cdot \mathbf{u}\|_0^2
\]

where

\[
\delta_1 = \begin{cases} 1 & \text{for Case I} \\ 0 & \text{for Case II} \end{cases} \quad \text{and} \quad \delta_2 = \begin{cases} 0 & \text{for Case I} \\ 1 & \text{for Case II} \end{cases}
\]
• Bilinear form for least-squares problem

\[ B \left( \{ u, \omega, p \}, \theta, \{ v, \sigma, q \}; \{ \tilde{u}, \tilde{\omega}, \tilde{p} \}, \tilde{\theta}, \{ \tilde{v}, \tilde{\sigma}, \tilde{q} \} \right) \]

\[ = \left( \nabla \times \sigma + \nabla q + \delta_2 u, \nabla \times \tilde{\sigma} + \nabla \tilde{q} + \delta_2 \tilde{u} \right)^{-1} \]

\[ + \left( \nabla \times v - \sigma + \delta_1 \omega, \nabla \times \tilde{v} - \tilde{\sigma} + \delta_1 \tilde{\omega} \right) + \left( \nabla \cdot v, \nabla \cdot \tilde{v} \right) \]

\[ + \left( \delta \theta + v, \tilde{\delta} \tilde{\theta} + \tilde{v} \right) \]

\[ + \left( \nabla \times \omega + \nabla p + \theta, \nabla \times \tilde{\omega} + \nabla \tilde{p} + \tilde{\theta} \right)^{-1} \]

\[ + \left( \nabla \times u - \omega, \nabla \times \tilde{u} - \tilde{\omega} \right) + \left( \nabla \cdot u, \nabla \cdot \tilde{u} \right) \]

• Linear functional

\[ F \left( \{ \tilde{u}, \tilde{\omega}, \tilde{p} \}, \tilde{\theta}, \{ \tilde{v}, \tilde{\sigma}, \tilde{q} \}; \hat{u}, g \right) \]

\[ = \left( \delta_2 \hat{u}, \nabla \times \tilde{\sigma} + \nabla \tilde{q} + \delta_2 \tilde{u} \right)^{-1} + \left( g, \nabla \times \tilde{\omega} + \nabla \tilde{p} + \tilde{\theta} \right)^{-1} \]
• With all this notation, the optimization problems for the Stokes system are exactly of the type we have considered.

• All the assumptions that have been made in the abstract setting - and therefore all the results that were obtained in that setting can be verified in the setting of the Stokes system.

• An example of the type of result that one obtains
  – suppose one chooses continuous, piecewise polynomial finite element spaces of degree $r$ for the approximation of all variables - this is permissible for least-squares finite element methods.
  – suppose also that the solution of the optimality system satisfies
    \[
    \begin{align*}
    u & \in H^{r+1}(\Omega) \cap H^1_0(\Omega) & \omega & \in H^r(\Omega) & p & \in H^r(\Omega) \cap L^2_0(\Omega) \\
    \theta & \in H^r(\Omega) \\
    v & \in H^{r+1}(\Omega) \cap H^1_0(\Omega) & \sigma & \in H^r(\Omega) & q & \in H^r(\Omega) \cap L^2_0(\Omega)
    \end{align*}
    \]
– then, we have the error estimate 
\[ \|u - u^h\|_1 + \|\omega - \omega^h\|_0 + \|p - p^h\|_0 + \|\theta - \theta^h\|_0 \]
\[ + \|\nu - \nu^h\|_1 + \|\sigma - \sigma^h\|_0 + \|q - q^h\|_0 = O(h^r) \]

– Of course, the least-squares method we have just described is not practical due to the appearance of the $H^{-1}(\Omega)$ norm
– ways to get around using that norm are available
• We briefly consider
  – as an alternative to discretizing the optimality system that results from applying the Lagrange multiplier rule
  penalty methods for obtaining solution of our constrained optimization of control problem
Least-squares formulation of the constraint equation

- Recall the constraint equation written in operator form
  \[ B_1 \phi + B_2 \theta = g \quad \text{in } \Lambda^* \]
  where \( g \in \Lambda^* \) is a given function

- For the moment, let us consider the control \( \theta \in \Theta \) as also being a given function and consider the problem of finding \( \phi \in \Phi \) such that
  \[ B_1 \phi = g - B_2 \theta \quad \text{in } \Lambda^* \]
  – this equation is well posed with respect to the data space \( \Lambda^* \) and the solution space \( \Phi \)

- For any \( \theta \in \Theta \) and \( g \in \Lambda^* \), we define the least-squares functional
  \[ \mathcal{H}(\phi; \theta, g) = \| B_1 \phi + B_2 \theta - g \|_{\Lambda^*}^2 \quad \forall \phi \in \Phi \]

- We then consider the problem: given \( \theta \in \Theta \) and \( g \in \Lambda^* \)
  \[ \min_{\phi \in \Phi} \mathcal{H}(\phi; \theta, g) \]
• The Euler-Lagrange equation corresponding to this minimization problem is given, in variational form, by

\[ \tilde{b}_1(\phi, \mu) = (g, B_1 \mu)_{\Lambda^*} - \tilde{b}_2(\theta, \mu) \quad \forall \mu \in \Phi \]

where

\[ \tilde{b}_1(\phi, \mu) = (B_1 \mu, B_1 \phi)_{\Lambda^*} \quad \forall \phi, \mu \in \Phi \]

\[ \tilde{b}_2(\theta, \nu) = (B_1 \nu, B_2 \theta)_{\Lambda^*} \quad \forall \theta \in \Theta, \nu \in \Phi \]

• It can be shown that the two bilinear forms are continuous and that the bilinear form \( \tilde{b}_1(\cdot, \cdot) \) is coercive

  – as a result, it can be shown that

\[ C_1 \|\phi\|_{\Phi}^2 \leq \mathcal{H}(\phi; 0, 0) \leq C_2 \|\phi\|_{\Phi}^2 \]

  so that the functional \( \mathcal{H}(\phi; 0, 0) \) is norm equivalent
• In operator form, the least-squares principle yields the first-order necessary condition

\[ B_1^*B_1\phi = B_1^*g - B_1^*B_2\theta \quad \text{in } \Phi^* \]

which are simply the normal equations for the constraint equation

\[ B_1\phi = g - B_2\theta \]

 – the operator \( B_1^*B_1 \) is symmetric and positive definite even when the operator \( B_1 \) is indefinite

• We can define a least-squares finite element method which is of norm-equivalent type

 – so that the analysis is simple
 – optimal accuracy is achieved
 – the coefficient matrix of the resulting algebraic equation is symmetric and positive definite
Multiobjective form of the optimal control problem

- Instead of considering the optimal control problem

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

we consider the equivalent problem

\[
\min_{(\phi, \theta) \in \Phi \times \Theta} J(\phi, \theta) \quad \text{subject to} \quad \min_{\phi \in \Phi} \mathcal{H}(\phi; \theta, g)
\]

- this is a multiobjective optimization problem
- there are, of course, many ways to address such problems
One method for treating the multiobjective optimization problem is to form the functional

$$J_\epsilon(\phi, \theta) = J(\phi, \theta) + \frac{1}{\epsilon} \mathcal{H}(\phi; \theta, g)$$

so that

$$J_\epsilon(\phi, \theta) = \frac{1}{2} \langle A_1(\phi - \hat{\phi}), (\phi - \hat{\phi}) \rangle_{\Phi*},_{\Phi} + \frac{1}{2} \langle A_2 \theta, \theta \rangle_{\Theta*},_{\Theta}$$

$$+ \frac{1}{2\epsilon} \langle B_1 \phi + B_2 \theta - g, D^{-1}(B_1 \phi + B_2 \theta - g) \rangle_{\Lambda*},_{\Lambda}$$

$$= \frac{1}{2} a_1(\phi - \hat{\phi}, \phi - \hat{\phi}) + \frac{1}{2} a_2(\theta, \theta)$$

$$+ \frac{1}{2\epsilon} \tilde{b}_1(\phi, \phi) + \tilde{b}_2(\theta, \phi) + c(\theta, \theta)$$

$$- \frac{1}{2\epsilon} (2\langle \tilde{g}_1, \phi \rangle_{\Phi*},_{\Phi} + 2\langle \tilde{g}_2, \theta \rangle_{\Theta*},_{\Theta} - \langle g, D^{-1}g \rangle_{\Lambda*},_{\Lambda})$$
• Unfortunately, this approach does not take full advantage of the least-squares formulation of the constraint equation

  – in particular, the discrete inf-sup condition on the bilinear form \( b_1(\cdot, \cdot) \) cannot be circumvented

  – one also has to worry about choosing a “good” value for the penalty parameter \( \epsilon \)
Methods based on constraining by the least-squares first-order necessary conditions

• Another equivalent reformulation of our optimal control problem is

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

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

$$\begin{cases} A_1 \phi + B_1^* B_1 \lambda = A_1 \hat{\phi} & \text{in } \Phi^* \\ A_2 \theta + B_2^* B_1 \lambda = 0 & \text{in } \Theta^* \\ B_1^* B_1 \phi + B_1^* B_2 \theta = B_1^* g & \text{in } \Phi^* , \end{cases}$$

where $\lambda \in \Phi$ is the Lagrange multiplier introduced to enforce the constraint

– this is a saddle point type system, so that the resulting matrix problems are symmetric but indefinite

– the only advantage over the standard Galerkin approach is that the operator $B_1^* B_1$ is positive definite, even when $B_1$ is not so
There are several effective means for discretizing the new optimality system:

- penalty methods are one approach

- one also has the choice of discretize-then-optimize and optimize-then-discretize approaches
Comparisons of the Different Approaches To Solving the Optimal Control Problem

• We consider seven computational methods for solving our optimal control problem
  – the first is the standard Lagrange multiplier followed by a Galerkin discretization of the resulting optimality system
  – the second is a penalty approach to solving the optimality system
  – the other five approaches involve least-squares principles in one way or another

• The methods
  1. Lagrange multiplier rule applied to the optimization problem followed by a mixed-Galerkin finite element discretization of the resulting optimality system
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

3. 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

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

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 to obtain an optimality system followed by a finite element discretization
6. 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.

7. 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.
• We use the following criteria to compare the methods

  – **discrete inf-sup not required** – are the finite element spaces required to satisfy a discrete inf-sup condition 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 $\epsilon$ 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>
<thead>
<tr>
<th></th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>discrete inf-sup not required</td>
<td>× × √ √ √ √ √</td>
</tr>
<tr>
<td>locking impossible</td>
<td>√ √ √ × √ √ ×</td>
</tr>
<tr>
<td>optimal error estimate</td>
<td>√ √ √ × √ √ ×</td>
</tr>
<tr>
<td>symmetric matrix system</td>
<td>√ √ √ √ √ √ √</td>
</tr>
<tr>
<td>reduced number of unknowns</td>
<td>× √ × √ × √ √</td>
</tr>
<tr>
<td>positive definite matrix system</td>
<td>× √ √ × √ √</td>
</tr>
</tbody>
</table>

- From the table, we see that only method 6 has all its boxes checked
  - so it seems that it is the preferred method
• However, there are additional issues that arise in practice that must also be considered
  – use standard finite element spaces
  – ease of assembly of the discrete system
  – manageable conditioning of the discrete systems

• When these additional criteria are added to the mix, it seems that method three wins out
• However, there are additional issues that arise in practice that must also be considered
  – use standard finite element spaces
  – ease of assembly of the discrete system
  – manageable conditioning of the discrete systems

• When these additional criteria are added to the mix, it seems that method three wins out
  – which is why I spent the most time on it!