Least-squares finite elements for distributed optimal control problems

Thomas Führer1 · Michael Karkulik2

Received: 28 October 2022 / Revised: 29 July 2023 / Accepted: 29 July 2023 / Published online: 10 August 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
We provide a framework for the numerical approximation of distributed optimal control problems, based on least-squares finite element methods. Our proposed method simultaneously solves the state and adjoint equations and is inf–sup stable for any choice of conforming discretization spaces. A reliable and efficient a posteriori error estimator is derived for problems where box constraints are imposed on the control. It can be localized and therefore used to steer an adaptive algorithm. For unconstrained optimal control problems, i.e., the set of controls being a Hilbert space, we obtain a coercive least-squares method and, in particular, quasi-optimality for any choice of discrete approximation space. For constrained problems we derive and analyze a variational inequality where the PDE part is tackled by least-squares finite element methods. We show that the abstract framework can be applied to a wide range of problems, including scalar second-order PDEs, the Stokes problem, and parabolic problems on space-time domains. Numerical examples for some selected problems are presented.

Mathematics Subject Classification 65N30 · 65N12 · 35F35 · 65M50 · 49M41

1 Introduction

Optimal control problems subject to PDEs form an important class of problems in practice, see, e.g., [24] for various applications. Nowadays, the theory of distributed optimal control problems subject to linear PDEs is well understood, and the most
common approach to solve such problems is by deriving first-order optimality conditions and introducing either the adjoint state or a Lagrangian multiplier. This approach leads, in general, to a symmetric but indefinite system. Quite naturally, finite element techniques (FEM) have been used to discretize the resulting formulations, see [2, Ch.11]. Standard references on different finite element methods used in optimal control are [3, 5, 10]. In the present work, we particularly consider least-squares finite element methods, which are treated thoroughly in [2]. We also refer to [2, Ch.11] for an overview on references and different techniques to treat optimal control problems with FEM resp. least-squares FEM (LSFEM).

In the present work we consider a different approach, inspired by the work [1]. In [1, Sec.4] the optimality system is written in operator form and, then, a quadratic functional is defined by summing the squared norms of the residuals. Minimizing this functional is equivalent to solving the optimal control problem, and minimizing over finite element spaces gives a LSFEM. The method from [1, Sec.4] is defined when the control variable is sought in a Hilbert space. However, when the control is restricted to a convex subset only—a situation encountered in many practical applications—there is no simple, i.e., linear, relation between the adjoint state and the control. Here, we consider the general case of optimal control problems where the set of admissible controls is a convex closed subset of the Lebesgue space of square-integrable functions. Starting from the optimality system consisting of the state equation, the adjoint state equation and a variational inequality relating the adjoint state and the control, see [24, Ch. 2], our proposed method is based on a least-squares functional (for the state and adjoint state equation) and a duality term corresponding to the variational inequality of the optimality system. The whole method results in a variational inequality and we prove that the bilinear form, though not symmetric, is coercive, thus, the discrete systems are invertible for any choice of discretization space. The proposed method can also be interpreted as a Nitsche-type coupling of a LSFEM and a Galerkin method. Nitsche-type methods are defined by adding terms to a bilinear form to include, e.g., essential boundary conditions in a weak sense, see, e.g., [28]. In the past, we have used such a coupling technique for other minimum residual methods like the discontinuous Petrov–Galerkin method with optimal test functions (DPG) to define a coupling to boundary element methods [13] or for the analysis of the DPG method for Signorini problems [14]. If the set of admissible controls is a closed subspace, then our method simplifies to a LSFEM method that is essentially a reduced version of the method from [1, Sec.4] (we eliminated the optimality condition from [1, Eq.(4.1)]).

In this article, we restrict the presentation to a framework where we consider LSFEMs which minimize first-order system residuals in $L^2$ norms only. The practical implications of this framework are that standard finite element spaces (continuous and piecewise polynomial) are conforming, and that the system matrices are easily computable up to numerical quadrature. A disadvantage is an increased number of degrees of freedom due to the introduction of a new variable. We show that the framework is quite general in the sense that it is applicable to a wide range of problems, including distributed optimal control problems subject to reaction–diffusion–convection problems, the Stokes problem, the heat equation and many more. The heat equation, in particular, will be formulated in space-time following methods developed in the recent works [15, 17]. The advantage of this approach is that it is robust in standard energy
norms even on locally refined space-time meshes and that standard finite element spaces are conforming. These space-time discretizations have been used recently for the optimal control of parabolic problems in [16, 18]. While the latter two works use a Lagrangian multiplier in order to obtain a complete optimality system, in the present work we directly include the adjoint state equation. An advantage of this approach is that we immediately obtain an approximation of the adjoint state without any post-processing. We also stress that robustness of the employed space-time discretization allows us to use final-time desired states, cf. [16]. On the downside, using first-order formulations for the state and the adjoint-state equations implies an increased number of degrees of freedom. In general, this is also a critical point in space-time discretizations. However, in transient problems such as optimal control for parabolic equations, the entire history of the state needs to be stored anyway, even if time-stepping methods are employed. For unconstrained optimal control problems our proposed method yields a symmetric and positive definite system of equations, whereas the approaches from [16, 18] yield symmetric saddle-point systems. Furthermore, in the present work the control is eliminated (for unconstrained problems) which leads to slightly smaller dimensions of the discrete spaces. Regarding convergence we note that all three methods use the same Hilbert spaces defined over the space-time cylinder and are quasi-optimal with respect to canonical norms. For constrained optimal control problems our proposed method requires us to choose sufficiently large parameters to ensure invertibility where the optimal choice of these parameters is not explicitly known. For unconstrained problems, however, our proposed method as well as the works [16, 18] do not require sufficiently large parameters. Other works on space-time methods for the optimal control of parabolic problems include [19, 20, 22, 23, 26]. These works use different space-time discretization methods such as residual minimization for second order PDEs, wavelet methods, mixed finite element methods, or discontinuous Galerkin methods, which do not share all the advantages as the space-time discretization employed in the present work.

Furthermore, we derive an a posteriori error estimator that is fully localizable as well as reliable and efficient in the case that box constraints are imposed on the control. The estimator consists of the norms of the residuals of the state and adjoint state equations plus the error between the discrete computed control and the projection of a similar quantity defined by the discrete adjoint state. In contrast to the estimator from [16] the a posteriori error estimator from the work at hand does not rely on solving an auxiliary problem.

The remainder of this article is organized as follows: In Sect. 2 we introduce the underlying assumptions of the LSFEMs under consideration, derive the optimality system in the notation of LSFEMs and define a general LSFEM based method to solve distributed optimal control problems. A well-posedness analysis and quasi-best approximation estimates are given together with the derivation of reliable and efficient error estimators. Section 3 gives various examples where the abstract framework is applicable and Sect. 4 presents numerical experiments for some selected problems.
2 LSFEM based framework

In this section we present a framework using least-squares finite elements for optimal control problems. In Sect. 2.2 we describe the optimal control problem and assumptions tailored to the use of LSFEMs. A LSFEM based method for the constrained optimal control problem is defined in Sect. 2.3 and a priori as well as a posteriori analysis is provided. For unconstrained optimal control problems we show in Sect. 2.4 that our proposed method can be simplified to a pure LSFEM problem.

2.1 Notation

For a Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$, we use common notations for Lebesgue and Sobolev spaces, e.g., $L^2(\Omega; \mathbb{R})$, $L^2(\Omega; \mathbb{R}^d)$, $L^2(\Omega; \mathbb{R}^{d \times d})$, $H^1(\Omega)$, $H^1_0(\Omega)$. We use $(\cdot, \cdot)_{L^2(\Omega)}$ to denote the canonical inner product in $L^2(\Omega; \mathbb{R})$, $L^2(\Omega; \mathbb{R}^d)$ or $L^2(\Omega; \mathbb{R}^{d \times d})$. The induced norm is denoted by $\| \cdot \|_{L^2(\Omega)}$. Throughout, we identify Lebesgue-like spaces with their duals. The Sobolev space $H^1_0(\Omega)$ is equipped with the norm $\| \nabla (\cdot) \|_{L^2(\Omega)}$. We also need $H(\text{div}; \Omega) = \{ \sigma \in L^2(\Omega; \mathbb{R}^d) : \text{div} \sigma \in L^2(\Omega; \mathbb{R}) \}$ with norm $\| \sigma \|^2_{H(\text{div}; \Omega)} = \| \sigma \|^2_{L^2(\Omega)} + \| \text{div} \sigma \|^2_{L^2(\Omega)}$. For a generic Hilbert space $H$ we denote by $(\cdot, \cdot)_H$ its inner product and by $\| \cdot \|_H$ the induced norm.

If $X_1, \ldots, X_n$ are Hilbert spaces, then we equip the product space $X = X_1 \times \cdots \times X_n$ with the norm

$$
\| (x_1, \ldots, x_n) \|^2_X = \| x_1 \|^2_{X_1} + \cdots + \| x_n \|^2_{X_n}.
$$

For the functional analytic setting of the heat equation we use Bochner spaces. Let $X$ be a Hilbert space and $I = (0, T)$ a time interval. The elements of the Bochner space $L^2(I; X)$ are functions $x : I \to X$ which are strongly measurable with respect to the Lebesgue measure $dt$ with

$$
\int_I \| x(t) \|^2_X dt < \infty.
$$

Similarly, the Bochner space $H^1(I; X)$ is defined as the space of functions $x \in L^2(X)$ such that $\partial_t x \in L^2(X)$, where $\partial_t (\cdot)$ denotes the weak derivative with respect to time. We refer to, e.g., [11, Sec.5.9.2], for a short introduction to Bochner spaces.

2.2 Optimal control problem

In this work we consider the following optimal control problem. Let $f \in Z$ be given and let the desired state $z_d \in H$ be given. With the cost parameter $0 < \lambda \leq 1$ define the cost functional $J : X \to \mathbb{R}$,
where the state $y = y(u) \in Y$ is the unique solution of the state equation

$$\mathcal{L}y + Bu = f.$$ 

The next subsection collects all the assumptions on the spaces and operators for problem (1).

### 2.2.1 Assumptions

Let $\Omega_1, \ldots, \Omega_n$ denote Lipschitz domains. We consider the following spaces resp. sets,

- $H$, $X$, $Y$, and $Y^*$ Hilbert spaces,
- $X_{ad} \subseteq X$ a non-empty, closed and convex subset,
- $Z = L^2(\Omega_1) \times \cdots \times L^2(\Omega_n)$.

Here, $L^2(\Omega_j)$ stands for a Lebesgue space, e.g., $L^2(\Omega_j, \mathbb{R}) = L^2(\Omega_j)$ (with $d$ depending on $\Omega_j$). We also consider bounded linear operators,

- $A : Z \to H$, with adjoint $A^* : H \to Z$,
- $B : X \to Z$, with adjoint $B^* : Z \to X$,
- $C : X \to X$ self-adjoint and positive definite, i.e.,

$$C = C^* \text{ and there exists } \kappa > 0 \text{ such that } (C u, u)_X \geq \kappa \|u\|_X^2 \forall u \in X. \quad (A1)$$

Moreover, we consider linear operators,

$$\mathcal{L} : Y \to Z, \quad \mathcal{L}^* : Y^* \to Z,$$

and assume that

$$\mathcal{L} \text{ and } \mathcal{L}^* \text{ are bounded, } \ker(\mathcal{L}) = \{0\} = \ker(\mathcal{L}^*), \quad \text{ran}(\mathcal{L}) = Z = \text{ran}(\mathcal{L}^*), \quad (A2)$$

i.e., $\mathcal{L}$ and $\mathcal{L}^*$ are boundedly invertible. Furthermore, we assume that there exist bounded linear operators

$$\mathcal{I} : Y \to Z \quad \text{and} \quad \mathcal{I}^* : Y^* \to Z$$

such that $\mathcal{L}$ and $\mathcal{L}^*$ are adjoint in the sense that

$$(\mathcal{L}y, \mathcal{I}^* p)_Z = (\mathcal{I}y, \mathcal{L}^* p)_Z \forall y \in Y, p \in Y^*. \quad (A3)$$

We anticipate that $\mathcal{L}$ resp. $\mathcal{L}^*$ will correspond to least-squares formulations of PDEs. Various examples will be given in Sect. 3.
Remark 1 As stated in the introduction we consider LSFEMs which minimize residuals in $L^2$ norms. It is possible to replace $Z$ by a general Hilbert space. To solve the operator equation $Ly = f$ in $Z$, a least-squares formulation would read: Find the minimizer of

$$\min_{y_h \in Y_h} \|Ly_h - f\|_Z^2.$$  

This requires to implement the inner product in $Z$ which may require the inversion of certain operators, e.g., if $Z$ is the dual space of $H^1_0(\Omega)$. For some problems it is possible to replace the inner product by a discrete version, see, e.g., [4].

Remark 2 In most applications $I$ and $I^*$ are canonical embedding operators, e.g., $H^1_0(\Omega) \times H(\text{div}; \Omega) \hookrightarrow L^2(\Omega) \times L^2(\Omega; \mathbb{R}^d)$. For the formulation of the heat equation considered below (Sect. 3.4) these operators also include a trace operation (restriction to initial resp. end time).

2.2.2 Auxiliary results

By following the proof of [24, Ch. 2, Theorem 1.4] we obtain the next result.

**Proposition 3** The minimization problem to find $u \in X_{ad}$ such that

$$J(u) = \min_{v \in X_{ad}} J(v)$$

is equivalent to

$$Ly = f - Bu, \quad (2a)$$
$$L^* p = A^*(AI y - z_d), \quad (2b)$$
$$(-B^* I^* p + \lambda Cu, v - u)_X \geq 0 \quad (2c)$$

for all $v \in X_{ad}$.

**Proof** Note that the functional $J$ is differentiable and convex. The minimization problem therefore is equivalent, see, e.g. [29, Lemma 2.21], to the variational inequality

$$(AI y(u) - z_d, AI (y(v) - y(u)))_H + (\lambda Cu, v - u)_X \geq 0 \quad \forall v \in X_{ad}.$$  

Define $p \in Y^*$ as the unique solution of

$$L^* p = A^*(AI y(u) - z_d)$$

which is possible due to (A2). Then, using (A3) gives

$$\begin{align*}
(AI y(u) - z_d, AI (y(v) - y(u)))_H &= (A^*(AI y(u) - z_d), I (y(v) - y(u)))_Z \\
&= (L^* p, I (y(v) - y(u)))_Z \\
&= (I^* p, L (y(v) - y(u)))_Z.
\end{align*}$$
Employing that $L y(u) = f - B u$ and $L y(v) = f - B v$ we end up with

$$(A I y(u) - z_d, A I (y(v) - y(u)))_H = -(I^* p, B (v - u))_Z = (-B^* I^* p, v - u)_X.$$  

This leads to the variational inequality

$$(-B^* I^* p + \lambda C u, v - u)_X \geq 0 \quad \forall v \in X_{ad},$$

which finishes the proof. \hfill \Box

For the analysis below we make use of a parameter-dependent norm on the product space $X \times Y \times Y^*$ given for all $(u, y, p) \in X \times Y \times Y^*$ by

$$\| (u, y, p) \|_\lambda^2 := \lambda \| u \|_X^2 + \| y - y_u \|_{Y^*}^2 + \lambda^{-1} \| p - p_y \|_{Y^*}^2 + \| A I y_u \|_H^2,$$

where for each $u \in X$, $y \in Y$ the functions $y_u \in Y$ and $p_y \in Y^*$ are defined as solutions of

$$L y_u = -B u, \quad L^* p_y = A^* A I y.$$  

From this point on we write $A \lesssim B$ for $A, B \geq 0$ if there exists a generic constant $C > 0$ independent of the cost parameter $\lambda$ such that $A \leq C \cdot B$. If both inequalities hold, i.e., $A \lesssim B \lesssim A$ then we write $A \approx B$.

**Lemma 4** The mapping $\| \cdot \|_\lambda : X \times Y \times Y^* \to \mathbb{R}_{\geq 0}$, $(u, y, p) \mapsto \| (u, y, p) \|_\lambda$ defines a norm on the product space $X \times Y \times Y^*$ with

$$\lambda^{1/2} \| (u, y, p) \|_{X \times Y \times Y^*} \lesssim \| (u, y, p) \|_\lambda \lesssim \lambda^{-1/2} \| (u, y, p) \|_{X \times Y \times Y^*}, \quad \forall (u, y, p) \in X \times Y \times Y^*.$$  

**Proof** By linearity of the definitions $y_u$, $p_y$ the triangle inequality and homogeneity of $\| \cdot \|_\lambda$ follow. Definiteness can be seen from the equivalence estimate. Let $(u, y, p) \in X \times Y \times Y^*$ be given. By the definition of $y_u$ and $p_y$ and the assumptions of Sect. 2.2.1 we have $\| y_u \|_Y \lesssim \| u \|_X$ and $\| p_y \|_{Y^*} \lesssim \| y \|_Y$. Using the latter estimates together with the triangle inequality proves

$$\lambda \| (u, y, p) \|_{X \times Y \times Y^*}^2 = \lambda \| u \|_X^2 + \lambda \| y \|_{Y^*}^2 + \lambda \| p \|_{Y^*}^2 \leq \lambda \| u \|_X^2 + \lambda \| y \|_{Y^*}^2 + \lambda \| p - p_y \|_{Y^*}^2 \leq \lambda \| u \|_X^2 + \lambda \| y - y_u \|_{Y^*}^2 + \lambda \| p - p_y \|_{Y^*}^2 \leq \lambda \| u \|_X^2 + \lambda \| y - y_u \|_{Y^*}^2 + \lambda^{-1} \| p - p_y \|_{Y^*}^2 + \| A I y_u \|_H^2 = \| (u, y, p) \|_\lambda^2,$$
where we used $0 < \lambda \leq 1$. For the upper bound the very same arguments show that

\[
\| (u, y, p) \|_2^2 = \lambda \| u \|_X^2 + \| y - y_u \|_Y^2 + \lambda^{-1} \| p - p_y \|_Y^* + \| A^* y_u \|_H^2
\]

\[
\lesssim \lambda \| u \|_X^2 + \lambda^{-1} \| y \|_Y^2 + \lambda^{-1} \| p \|_Y^* + \| y_u \|_Y^2 + \lambda^{-1} \| p_y \|_Y^* + \| u \|_X^2
\]

\[
\lesssim \| u \|_X^2 + \lambda^{-1} \| y \|_Y^2 + \lambda^{-1} \| p \|_Y^* \leq \lambda^{-1} \|(u, y, p)\|_{X \times Y \times Y^*}^2.
\]

This finishes the proof. \(\square\)

### 2.3 LSFEM for constrained optimal control problem

Let us define the bilinear form

\[
a(u, y, p; v, z, q) = \alpha (\mathcal{L} y + \mathcal{B} u, \mathcal{L} z + \mathcal{B} v)_Z
\]

\[
+ \beta (\mathcal{L}^* p - A^* A^* y, \mathcal{L}^* q - A^* A^* z)_Z
\]

\[
+ (-B^* T^* p + \lambda Cu, v)_X
\]

for $u, v \in X$, $y, z \in Y$, $p, q \in Y^*$. Here $\alpha, \beta$ denote positive constants. We also consider the load functional $\ell: X \times Y \times Y^* \to \mathbb{R}$, for given $f \in Z, z_d \in H$, defined by

\[
\ell(v, z, q) = \alpha (f, \mathcal{L} z + \mathcal{B} v)_Z + \beta (-A^* z_d, \mathcal{L}^* q - A^* A^* z)_Z.
\]

We now show that (2) is equivalent to the following variational inequality: Find $(u, y, p) \in X_{\text{ad}} \times Y \times Y^*$ such that

\[
a(u, y, p; v - u, z - y, q - p) \geq \ell(v - u, z - y, q - p)
\]

for all $(v, z, q) \in X_{\text{ad}} \times Y \times Y^*$. (3)

**Proposition 5** Problems (2) and (3) are equivalent.

**Proof** If $u, y, p$ solve (2), then they also satisfy (3).

Let $u, y, p$ be a solution of (3). By testing (3) with $v = u$, $z = y$, $q = p \pm w$, $w \in Y^*$, we see that

\[
\pm \beta (\mathcal{L}^* p - A^* A^* y, \mathcal{L}^* w)_Z \geq \pm \beta (-A^* z_d, \mathcal{L}^* w)_Z,
\]

or equivalently

\[
(\mathcal{L}^* p - A^* A^* y, \mathcal{L}^* w)_Z = (-A^* z_d, \mathcal{L}^* w)_Z
\]

for all $w \in Y^*$. With (A2), this yields

\[
\mathcal{L}^* p = A^* (A^* y - z_d).
\]
which is (2b). By testing (3) with \( v = u, z = y \pm w, w \in Y, q = p \), and using the latter identity we further conclude with a similar argumentation as before that

\[
(\mathcal{L} y + B u, L w)_Z = (f, L w)_Z \quad \forall w \in Y,
\]

which means that, using again (A2),

\[
\mathcal{L} y + B u = f,
\]

i.e., (2a). Using the identities (2a)–(2b), we see that the variational inequality (3) reduces to

\[
0 \leq a(u, y, p; v - u, z - y, q - p) - \ell(v - u, z - y, q - p)
\]

\[
= (-B^* I^* p + \lambda C u, v - u)_X
\]

for all \( v \in X_{ad} \). This is (2c) and, thus, finishes the proof. \( \square \)

The next result shows that the bilinear form \( a \) is coercive and bounded, given that the constants \( \alpha, \beta \) are sufficiently large. In other words, \( a \) induces an equivalent norm on the underlying energy space, and thus Problem (3) admits a unique solution.

**Theorem 6** There exists \( \alpha_0, \beta_0 \) with \( \alpha_0 \approx 1, \beta_0 \approx \lambda^{-1} \) such that for all \( \alpha \geq \alpha_0, \beta \geq \beta_0 \) and all \( (u, y, p), (v, z, q) \in X \times Y \times Y^* \),

\[
\|(u, y, p)\|_\lambda^2 \lesssim a(u, y, p; u, y, p),
\]

\[
|a(u, y, p; v, z, q)| \lesssim \max\{\alpha, \beta\} \|(u, y, p)\|_\lambda \|(v, z, q)\|_\lambda,
\]

i.e., the bilinear form \( a(\cdot, \cdot) \) is coercive and bounded with respect to the norm \( \|\cdot\|_\lambda \).
For any non-empty closed convex subset \( W \subseteq X \times Y \times Y^* \), the variational inequality to find \( (u, y, p) \in W \) such that

\[
a(u, y, p; v - u, z - y, q - p) \geq \ell(v - u, z - y, q - p) \quad \text{for all } (v, z, q) \in W
\]

admits a unique solution. In particular, Problem (2) respectively (3) admits a unique solution.

**Proof** It suffices to prove that the bilinear form is coercive and bounded and that the right-hand side is linear and bounded. Existence and uniqueness of the given variational inequality then follow from the Lions–Stampacchia theorem, cf. [25]. Let \( (u, y, p), (v, z, q) \in X \times Y \times Y^* \) be given. Recall that \( y_u \) and \( p_y \) are given as solutions of

\[
\mathcal{L} y_u = -B u, \quad L^* p_y = A^* A I y.
\]
By the assumptions of Sect. 2.2.1 we have $\|y_u\|_Y \lesssim \|u\|_X$ and $\|p_y\|_{Y^*} \lesssim \|y\|_Y$. Moreover,

$$\|y - y_u\|_Y \approx \|\mathcal{L}(y - y_u)\|_Z = \|\mathcal{L}y + \mathcal{B}u\|_Z$$

and

$$\|p - p_y\|_{Y^*} \approx \|\mathcal{L}^*(p - p_y)\|_Z = \|\mathcal{L}^*p - \mathcal{A}^*\mathcal{A}^T y\|_Z.$$ 

Putting the last two estimates together proves

$$\|\mathcal{L}y + \mathcal{B}u\|_Z + \lambda^{-1/2}\|\mathcal{L}^*p - \mathcal{A}^*\mathcal{A}^T y\|_Z \approx \|y - y_u\|_Y + \lambda^{-1/2}\|p - p_y\|_{Y^*} \lesssim \|(u, y, p)\|_{\lambda}, \quad (5)$$

which is used in the remainder of the proof.

To prove boundedness we consider the individual terms in the bilinear form. First, by the Cauchy–Schwarz inequality and (5) we see that

$$\alpha |(\mathcal{L}y + \mathcal{B}u, \mathcal{L}z + \mathcal{B}v)_Z| \lesssim \alpha \|\mathcal{L}y + \mathcal{B}u\|_Z \|\mathcal{L}z + \mathcal{B}v\|_Z \lesssim \alpha \|(u, y, p)\|_{\lambda} \|(v, z, q)\|_{\lambda}.$$ 

Second, the same arguments yield

$$\beta |(\mathcal{L}^*p - \mathcal{A}^*\mathcal{A}^T y, \mathcal{L}^*q - \mathcal{A}^*\mathcal{A}^T z)_Z| \leq \lambda \beta \lambda^{-1/2}\|\mathcal{L}^*p - \mathcal{A}^*\mathcal{A}^T y\|_Z \lambda^{-1/2}\|\mathcal{L}^*q - \mathcal{A}^*\mathcal{A}^T z\|_Z \lesssim \lambda \beta \|(u, y, p)\|_{\lambda} \|(v, z, q)\|_{\lambda}.$$ 

For the estimate of the final term in the bilinear form we note that

$$(\mathcal{B}^*\mathcal{I}^*p_y, v)_X = (\mathcal{I}^*p_y, \mathcal{B}v)_Z = (\mathcal{I}^*p_y, -\mathcal{L}y_v)_Z = -(\mathcal{L}^*p_y, \mathcal{I}y_v)_Z = -(\mathcal{A}^*\mathcal{A}^T y, \mathcal{I}y_v)_Z = -(\mathcal{A}^*\mathcal{A}^T y, \mathcal{A}^T y_v)_H.$$ 

(6)

This yields

$$|(\mathcal{B}^*\mathcal{I}^*p_y, v)_X| \leq |(\mathcal{A}^*y - y_u, \mathcal{A}^*y_v)_H| + |(\mathcal{A}^*y_u, \mathcal{A}^*y_v)_H| \lesssim (\|y - y_u\|_Y + \|\mathcal{A}^*\mathcal{A}^T y_u\|_H) \|\mathcal{A}^*\mathcal{A}^T y_v\|_H \lesssim \|(u, y, p)\|_{\lambda} \|(v, z, q)\|_{\lambda}.$$ 

Then,

$$|(-\mathcal{B}^*\mathcal{I}^*p + \lambda \mathcal{C}u, v)_X| \lesssim |(\mathcal{B}^*\mathcal{I}^*p - p_y, v)_X| + |(\mathcal{B}^*\mathcal{I}^*p_y, v)_X| + \lambda |(\mathcal{C}u, v)_X| \lesssim \lambda^{-1/2}\|p - p_y\|_{Y^*} \lambda^{1/2}\|v\|_X + \|(u, y, p)\|_{\lambda} \|(v, z, q)\|_{\lambda} \lesssim \|(u, y, p)\|_{\lambda} \|(v, z, q)\|_{\lambda}.$$
This shows boundedness of the bilinear form. The very same arguments can be used to prove that the right-hand side functional $\ell$ is bounded. The details are omitted.

It remains to prove coercivity. By the definition of the norm $\| \cdot \|_\lambda$ and (5) we get

$$\|(u, y, p)\|_\lambda^2 = \|y - y_u\|_Z^2 + \lambda^{-1}\|p - p_y\|_Z^2 + \lambda\|u\|_X^2 + \|\Lambda\Lambda y_u\|_H^2$$

$$\approx \|Ly + Bu\|_Z^2 + \lambda^{-1}\|\Lambda^* p - A^*\Lambda\Lambda y\|_Z^2 + \lambda \langle Cu, u \rangle + \|\Lambda\Lambda y_u\|_H^2.$$

In the last estimate we have used Young’s inequality with $\delta > 0$ and a constant $C_1 > 0$. Employing Young’s inequality again we further obtain

$$|\langle B^*\Lambda^* (p - p_y), u \rangle| \leq C_2 \delta^{-1} / 2 \lambda^{-1} \|\Lambda^* p - A^*\Lambda\Lambda y\|_Z^2 + \delta / 2 \|u\|_X^2$$

where $C_2 > 0$ denotes another positive constant. Combining the last two estimates and using $\lambda\|u\|_X^2 + \|\Lambda\Lambda y_u\|_H^2 \leq \|(u, y, p)\|_\lambda^2$ we conclude that

$$\|(u, y, p)\|_\lambda^2 \lesssim \|Ly + Bu\|_Z^2 + \lambda^{-1}\|\Lambda^* p - A^*\Lambda\Lambda y\|_Z^2 + (B^*\Lambda^* (p - p_y), u) + \|\Lambda\Lambda y_u\|_H^2$$

$$\leq (1 + C_1 \delta^{-1}/2)\|Ly + Bu\|_Z^2 + (1 + C_2 \delta^{-1}/2)\lambda^{-1}$

Subtracting the last term on the right-hand side for sufficiently small $\delta > 0$ and setting $\alpha_0 := (1 + C_1 \delta^{-1}/2), \beta_0 := \lambda^{-1}(1 + C_2 \delta^{-1})$ finishes the proof. \hfill \qed

### 2.3.1 A priori error analysis

In this section we derive quasi-optimality results using well-known techniques for variational inequalities, see, e.g. [12]. In particular, following the proof of [12, Theorem 1] shows the next result.

**Theorem 7** Suppose that $1 \approx \alpha \geq \alpha_0$ and $\lambda^{-1} \approx \beta \geq \beta_0$, such that the results of Theorem 6 hold true. Let $(u, y, p) \in X_{ad} \times Y \times Y^*$ denote the unique solution of (3). Let $X_{h, ad} \subset X_{ad}$ denote a non-empty closed convex subset and $Y_h \subset Y, Y^*_h \subset Y^*$ closed subspaces. Set $W_h = X_{h, ad} \times Y_h \times Y^*_h$. If $(u_h, y_h, p_h) \in W_h$ denotes the
solution of (4), then

\[
\| (u - u_h, y - y_h, p - p_h) \|^2_\lambda \\
\lesssim \inf_{(v_h, z_h, q_h) \in W_h} \left( \| (u - v_h, y - z_h, p - q_h) \|^2_\lambda + |(-B^* T^* p + \lambda C u, v_h - u)_X | \right). 
\]

**Proof** For the convenience of the reader we include the proof. First, from (2a)–(2b) we see that the solution \((u, y, p)\) satisfies

\[
a(u, y, p; v, w, q) = \ell(v, w, q) + (-B^* T^* p + \lambda C u, v)_X \quad \forall (v, w, q) \in X \times Y \times Y^*. 
\]

(7)

We use the short notation \(x = (u, y, p), x_h = (u_h, y_h, p_h)\) and \(\bar{x}_h = (v_h, z_h, q_h) \in W_h\). Second, Theorem 6 shows that the bilinear form is coercive. Together with (7), this leads to

\[
\| x - x_h \|^2_\lambda \lesssim a(x - x_h; x - x_h) \\
= a(x; x - x_h) - a(x_h; x - \bar{x}_h) - a(x_h; \bar{x}_h - x_h) \\
\leq \ell(x - x_h) - a(x_h; x - \bar{x}_h) - \ell(\bar{x}_h - x_h) \\
= \ell(x - \bar{x}_h) - a(x_h; x - \bar{x}_h). 
\]

Using (7), boundedness of the bilinear form and Young’s inequality with parameter \(\delta > 0\) we see that

\[
\ell(x - \bar{x}_h) - a(x_h; x - \bar{x}_h) \\
\leq a(x; x - \bar{x}_h) - (-B^* T^* p + \lambda C u, u - v_h)_X - a(x_h; x - \bar{x}_h) \\
\leq a(x - x_h; x - \bar{x}_h) + (-B^* T^* p + \lambda C u, v_h - u)_X \\
\lesssim \delta^{-1} \| x - \bar{x}_h \|^2_\lambda + (-B^* T^* p + \lambda C u, v_h - u)_X + \delta \| x - x_h \|^2_\lambda. 
\]

Putting all the estimates together and subtracting the last term for \(\delta\) sufficiently small proves the asserted quasi-optimality result. \(\square\)

An immediate consequence of Theorem 7 and norm equivalence from Lemma 4 is the following result.

**Corollary 8** With the assumptions and notations of Theorem 7, the estimate

\[
\lambda \| (u - u_h, y - y_h, p - p_h) \|^2_{X \times Y \times Y^*} \\
\lesssim \inf_{(v_h, z_h, q_h) \in W_h} \left( \lambda^{-1} \| (u - v_h, y - z_h, p - q_h) \|^2_{X \times Y \times Y^*} + |(-B^* T^* p + \lambda C u, v_h - u)_X | \right) 
\]

holds.
Remark 9 If $X_{h, \text{ad}}$ is not a subset of $X_{\text{ad}}$, i.e., if we consider a non-conforming discretization of the convex set $X_{\text{ad}}$, then the additional term
\[ \inf_{v \in X_{\text{ad}}} |(-B^*I^*p_h + \lambda Cu_h, v - u_h)_X|, \]
enters on the right-hand side of the a priori estimate in Theorem 7.

2.3.2 A posteriori error analysis

In this section we derive reliable and efficient a posteriori error estimators. Let us recall the variational inequality from (2),
\[ (-B^*I^*p + \lambda Cu, v - u)_X \geq 0 \quad \forall v \in X_{\text{ad}}. \]

By our assumptions on the operator $C$ we have that $(\cdot, \cdot)_C := (C(\cdot), (\cdot))_X$ defines an inner product on $X$ which induces the norm $\| \cdot \|_C$ that is equivalent to the norm on $X$. Thus, the variational inequality can be put as
\[ (-\lambda^{-1} C^{-1} B^*I^*p + u, v - u)_C \geq 0 \quad \forall v \in X_{\text{ad}}. \]

This means that $u = \Pi_{X_{\text{ad}}} \lambda^{-1} C^{-1} B^*I^*p$ where $\Pi_{X_{\text{ad}}}: X \to X_{\text{ad}}$ denotes the projection with respect to the inner product $(\cdot, \cdot)_C$, see, e.g. [6, Theorem 5.2]. Note that $\Pi_{X_{\text{ad}}}$ is a non linear operator unless $X_{\text{ad}}$ is a closed subspace. In the latter case it is the orthogonal projection. In either case, $\Pi_{X_{\text{ad}}}$ is non-expansive, i.e.,
\[ \| \Pi_{X_{\text{ad}}}(v) - \Pi_{X_{\text{ad}}}(w) \|_C \leq \| v - w \|_C \quad \forall v, w \in X. \]

Hence, by norm equivalence we also have that
\[ \| \Pi_{X_{\text{ad}}}(v) - \Pi_{X_{\text{ad}}}(w) \|_X \leq \| v - w \|_X \quad \forall v, w \in X. \]

Theorem 10 Suppose that $1 \sim \alpha \geq \alpha_0$, $\lambda^{-1} \sim \beta \geq \beta_0$, such that the results of Theorem 6 hold true. Let $(u, y, p) \in X_{\text{ad}} \times Y \times Y^*$ denote the unique solution of (3). Let $X_{h, \text{ad}} \subset X_{\text{ad}}$ denote a non-empty closed convex subset and $Y_{h} \subset Y$, $Y_{h}^* \subset Y^*$ closed subspaces. Let $(u_h, y_h, p_h) \in X_{h, \text{ad}} \times Y_h \times Y_h^*$ be some arbitrary element and define
\[ \tilde{u}_h = \tilde{u}_h(p_h) = \Pi_{X_{\text{ad}}} \lambda^{-1} C^{-1} B^*I^*p_h. \]

Then, the estimator given by
\[ \eta^2 = \alpha \| L \gamma_h + Bu_h - f \|_Z^2 + \beta \| L^*p_h - A^*(AIy_h - z_d) \|_Z^2 + \lambda \| \tilde{u}_h - u_h \|_X^2 \]
is reliable and efficient in the sense that
\[ \lambda \|(u - u_h, y - y_h, p - p_h)\|_{X \times Y \times Y^*} \lesssim \eta \lesssim \lambda^{-1/2} \|(u - u_h, y - y_h, p - p_h)\|_{X \times Y \times Y^*}. \]
Proof We start by noting that \( \tilde{u}_h \in X_{\text{ad}} \) satisfies

\[
(\lambda C \tilde{u}_h - B^* \mathcal{I}^* p_h, v - \tilde{u}_h)_{X} \geq 0 \quad \forall v \in X_{\text{ad}}.
\]

First, we show the efficiency estimate. By the triangle inequality, \( L y + B u = f \), \( L^* p = A^*(\mathcal{A} I y - z_d) \), estimate (5), Lipschitz continuity of the projection \( \Pi_{X_{\text{ad}}} \) and Lemma 4, we have that

\[
\eta^2 = \alpha \| L y_h + B u_h - f \|_Z^2 + \beta \| L^* p_h - A^*(\mathcal{A} I y_h - z_d) \|_Z^2 + \lambda \| \tilde{u}_h - u_h \|_X^2
\]

\[
= \alpha \| L (y_h - y) + B(u_h - u) \|_Z^2 + \beta \| L^* (p_h - p) - A^* A I (y_h - y) \|_Z^2 + \lambda \| \tilde{u}_h - u_h \|_X^2
\]

\[
\leq (u - u_h, y - y_h, \eta^2 \leq \lambda^{-1} \| (u - u_h, y - y_h, p - p_h) \|_Z^2 + \lambda \| \tilde{u}_h - u_h \|_X^2
\]

For the lower bound, we apply Theorem 6 to see that

\[
\lambda \| (u - \tilde{u}_h, y - y_h, p - p_h) \|_Z^2 \leq \lambda \| L(y_h - y) + B(\tilde{u}_h - u) \|_Z^2 + \| L^* (p_h - p) - A^* A I (y_h - y) \|_Z^2 + \lambda (-B^* \mathcal{I}^*) (p_h - p) + \lambda C(\tilde{u}_h - u), \tilde{u}_h - u)_X
\]

Note that \((-B^* \mathcal{I}^*) (p + \lambda C u, \tilde{u}_h - u)_X \geq 0\) since \( \tilde{u}_h \in X_{\text{ad}} \), as well as \((-B^* \mathcal{I}^*) (p + \lambda C \tilde{u}_h, u - \tilde{u}_h)_X \geq 0\) since \( u \in X_{\text{ad}} \). Thus,

\[
(-B^* \mathcal{I}^*) (p_h - p) + \lambda C(\tilde{u}_h - u), \tilde{u}_h - u)_X \leq (-B^* \mathcal{I}^*) (p_h + \lambda C \tilde{u}_h, \tilde{u}_h - u)_X \leq 0.
\]

We conclude that

\[
\lambda \| (u - \tilde{u}_h, y - y_h, p - p_h) \|_X^2 \leq \lambda \| L y_h + B \tilde{u}_h - f \|_Z^2 + \| L^* p_h - A^*(\mathcal{A} I y - z_d) \|_Z^2 \leq \eta^2.
\]

The triangle inequality \( \| u - u_h \|_X \leq \| \tilde{u}_h - u_h \|_X + \| u - \tilde{u}_h \|_X \) and Lemma 4 then show

\[
\lambda^2 \| (u - u_h, y - y_h, p - p_h) \|_{X \times Y \times Y^*}^2 \leq \lambda \| (u - \tilde{u}_h, y - y_h, p - p_h) \|_X^2 \leq \eta^2,
\]

which finishes the proof. \( \square \)
For general space $X$ and non-empty convex closed sets $X_{ad}$ the evaluation of $\Pi_{X_{ad}}$ may not be easily accessible. However, for the widely considered case of $X = \prod_{j=1}^{m} L^2(\Omega_j')$ ($\Omega_1', \ldots, \Omega_m'$ Lipschitz domains) together with box constraints there is an explicit representation of $\Pi_{X_{ad}}$: Let $a, b$ be given vector-valued and essentially bounded functions, $a < b$ (understood componentwise a.e.) and consider the box constraint control set

$$X_{ad}^{\text{box}} = \{ v \in X : a \leq v \leq b \ \text{a.e.} \}.$$ 

Suppose that $C$ is the identity operator. The projection on $X_{ad}^{\text{box}}$, characterized by $(\Pi_{X_{ad}}^{\text{box}} v - v, v - w) \geq 0$ for all $w \in X_{ad}^{\text{box}}$, has the explicit representation, see, e.g., [29, Theorem 2.28],

$$\Pi_{X_{ad}}^{\text{box}} v = \min \{ b, \max \{ v, a \} \} \ \text{a.e.}$$

The latter definition is understood componentwise.

### 2.4 LSFEM for unconstrained optimal control problem

In this section we consider the unconstrained optimal control problem, i.e., $X_{ad} \subseteq X$ is a closed subspace. This allows us to eliminate the control variable: The variational inequality (2c) is

$$(-B^* T^* p + \lambda C u , v - u)_X \geq 0 \ \forall v \in X_{ad}.$$ 

Taking $v = w + u$ and $v = u - w$ with $w \in X_{ad}$ implies

$$(-B^* T^* p + \lambda C u , w)_X = 0 \ \forall w \in X_{ad}.$$ 

Therefore, $u = \Pi_{X_{ad}}^{\lambda^{-1}C^{-1}B^*T^*} p$ where $\Pi_{X_{ad}}^{\lambda^{-1}C^{-1}B^*T^*} p$ is the orthogonal projection with respect to the inner product $(\cdot, \cdot)_C$. Problem (2) simplifies to

$$\mathcal{L} y = f - B \Pi_{X_{ad}}^{\lambda^{-1}C^{-1}B^*T^*} p,$$

$$\mathcal{L}^* p = A^* (AI y - z_d).$$

Define the least-squares functional

$$G(y, p; f, z_d) = \alpha \| \mathcal{L} y + B \Pi_{X_{ad}}^{\lambda^{-1}C^{-1}B^*T^*} p - f \|^2_Z + \beta \| \mathcal{L}^* p - A^* (AI y - z_d) \|^2_Z,$$

where $\alpha, \beta > 0$, and consider the minimization problem

$$\min_{(y, p) \in Y \times Y^*} G(y, p; f, z_d). \quad (8)$$
This problem has a unique solution if we can prove that the functional \((y, p) \mapsto G(y, p; 0, 0)\) is equivalent to a norm on \(Y \times Y^*\), see, e.g., [2, Section 2.2.1 and Theorem 2.5]. We use the norm

\[
\| (y, p) \|_\lambda^2 := \| (\Pi_{X_{ad}^\lambda} \lambda^{-1} C^{-1} B^* \mathcal{I}^* p, y, p) \|_\lambda^2 \quad \forall (y, p) \in Y \times Y^*.
\]

By following the proof of Lemma 4 one verifies the next result.

**Lemma 11** The estimates

\[
\lambda^{1/2} \| (y, p) \|_{Y \times Y^*} \lesssim \| (y, p) \|_\lambda \lesssim \lambda^{-1/2} \| y \|_Y + \lambda^{-1} \| p \|_{Y^*}
\]

hold for all \((y, p) \in Y \times Y^*\).

**Proof** The lower bound directly follows from the definitions and Lemma 4. For the upper bound we use \(u = \Pi_{X_{ad}^\lambda} \lambda^{-1} C^{-1} B^* \mathcal{I}^* p\). In the proof of Lemma 4 it was shown that

\[
\| (u, y, p) \|_\lambda^2 \lesssim \| u \|_X^2 + \lambda^{-1} \| y \|_Y^2 + \lambda^{-1} \| p \|_{Y^*}^2.
\]

This together with \(\| u \|_X \lesssim \lambda^{-1} \| p \|_{Y^*}\) shows the upper bound. \(\square\)

**Theorem 12** If \(\alpha \approx 1, \beta \approx \lambda^{-1}\), then

\[
\| (y, p) \|_\lambda^2 \approx G(y, p; 0, 0) \quad \forall (y, p) \in Y \times Y^*.
\]

**Proof** Let \(\alpha_0 \approx 1, \beta_0 \approx \lambda^{-1}\) denote the constants from Theorem 6. Note that we assume \(\alpha \approx 1\) and \(\beta \approx \lambda^{-1}\). Therefore, \(\alpha \approx \alpha_0, \beta \approx \beta_0\). Let \(y \in Y\) and \(p \in Y^*\) be given. Define \(u = \Pi_{X_{ad}^\lambda} \lambda^{-1} C^{-1} B^* \mathcal{I}^* p\). Note that then \((-B^* \mathcal{I}^* p + \lambda C u, v)_X = 0\) for any \(v \in X_{ad}\). We have that

\[
G(y, p; 0, 0) = \alpha \| \mathcal{L} y + B \Pi_{X_{ad}^\lambda} \lambda^{-1} C^{-1} B^* \mathcal{I}^* p \|_Z^2 + \beta \| \mathcal{L}^* p - A^* A \mathcal{I} y \|_Z^2
\]

\[
= \alpha_0 \| \mathcal{L} y + B u \|_Z^2 + \beta_0 \| \mathcal{L}^* p - A^* A \mathcal{I} y \|_Z^2
\]

\[
\approx \alpha_0 \| \mathcal{L} y + B u \|_Z^2 + \beta_0 \| \mathcal{L}^* p - A^* A \mathcal{I} y \|_Z^2 + (-B^* \mathcal{I}^* p + \lambda C u, v)_X
\]

\[
\approx \| (u, y, p) \|_\lambda^2 = \| (y, p) \|_\lambda^2.
\]

The last equivalence “\(\approx\)” follows from the proof of Theorem 6. \(\square\)

The variational formulation of the convex minimization problem (8) reads: Find \((y, p) \in Y \times Y^*\) such that

\[
b(y, p; w, q) = m(w, q) \quad \forall (w, q) \in Y \times Y^*,
\]

(9)
where the bilinear form $b : (Y \times Y^*)^2 \to \mathbb{R}$ and the functional $m : Y \times Y^* \to \mathbb{R}$ are given by

$$
\begin{align*}
    b(y, p; w, q) &= \alpha (\mathcal{L} y + B \Pi_{X_{ad}} \lambda^{-1} C^1 B^* T^* p, \mathcal{L} w + B \Pi_{X_{ad}} \lambda^{-1} C^1 B^* T^* q)_Z \\
    &\quad + \beta (L^* p - A^* A T y, L^* q - A^* A T w)_Z \\
    m(w, q) &= \alpha (f, \mathcal{L} w + B \Pi_{X_{ad}} \lambda^{-1} C^1 B^* T^* q)_Z + \beta (-A^* z_d, L^* q - A^* A T w)_Z.
\end{align*}
$$

Given a closed subspace $Y_h \times Y_h^* \subseteq Y \times Y^*$ we consider the discrete formulation: Find $(y_h, p_h) \in Y_h \times Y_h^*$ such that

$$
    b(y_h, p_h; w, q) = m(w, q) \quad \forall (w, q) \in Y_h \times Y_h^*. 
$$

(10)

Thanks to the norm equivalence established in Theorem 12 the next result follows immediately from the theory of LSFEMs, see, e.g., [2, Chapter 3 and Theorem 2.5].

**Theorem 13** Let $f \in Z$, $z_d \in Z$ be given. The problems (9) and (10) admit unique solutions $(y, p) \in Y \times Y^*$ and $(y_h, p_h) \in Y_h \times Y_h^*$. They satisfy the quasi-optimality estimate

$$
    \| (y - y_h, p - p_h) \|_\lambda \lesssim \inf_{(z_h, q_h) \in Y_h \times Y_h^*} \| (y - z_h, p - q_h) \|_\lambda.
$$

Another immediate result of the norm equivalence is the following a posteriori error estimate.

**Corollary 14** Let $(y, p) \in Y \times Y^*$ denote the unique solution of (9) and let $Y_h \times Y_h^* \subset Y \times Y^*$ denote a closed subspace. The least-squares functional defines an efficient and reliable a posteriori error estimator, i.e.,

$$
    \| (y - y_h, p - p_h) \|_\lambda^2 \approx G(y_h, p_h; f, z_d) \quad \text{for any } (y_h, p_h) \in Y_h \times Y_h^*.
$$

**Remark 15** To obtain an approximation of the control $u$ given that $(y_h, p_h) \in Y_h \times Y_h^*$ is the solution of (10), we define

$$
    u_h = \Pi_{X_{ad}} \lambda^{-1} C^1 B^* T^* p_h.
$$

Since $u = \Pi_{X_{ad}} \lambda^{-1} C^1 B^* T^* p$ we get that

$$
    \lambda^{1/2} \| u - u_h \|_X \leq \| (y - y_h, p - p_h) \|_\lambda \lesssim \inf_{(z_h, q_h) \in Y_h \times Y_h^*} \| (y - z_h, p - q_h) \|_\lambda
$$

by the boundedness of the involved operators and Theorem 13.

### 3 Examples

In this section we present various model problems where the framework developed in Sect. 2 can be applied.
3.1 Second-order PDE

Let \( \Omega \subseteq \mathbb{R}^d, d \geq 2 \) denote a bounded Lipschitz domain. Given \( f \in L^2(\Omega; \mathbb{R}), f_k \in L^2(\Omega; \mathbb{R}^d) \) we consider the first-order formulation of a general second-order PDE,

\[
\begin{align*}
\text{div} \, \sigma + cy &= f, \quad (11a) \\
\nabla y - by + A^{-1} \sigma &= f, \quad (11b) \\
y|_{\partial \Omega} &= 0, \quad (11c)
\end{align*}
\]

where \( A \in L^\infty(\Omega; \mathbb{R}^{d \times d}) \) is a symmetric, uniformly positive matrix, \( b \in L^\infty(\Omega; \mathbb{R}^d) \), \( c \in L^\infty(\Omega; \mathbb{R}) \) such that

\[
L^\infty(\Omega) \ni \frac{1}{2} \text{div} (Ab) + c \geq 0.
\]

We can also include the Helmholtz problem, where \( A \) is the identity, \( b = 0 \), and \( c < 0 \) is a constant such that it is not an eigenvalue of

\[
-\Delta y = \mu y, \quad y|_{\partial \Omega} = 0.
\]

We note that the problem admits a unique solution \((y, \sigma) \in H_0^1(\Omega) \times H(\text{div} ; \Omega)\). This can be seen by solving the second equation in (11) for \( \sigma \) and then replacing \( \sigma \) in the first equation. The resulting second-order PDE is

\[
-\text{div} (A \nabla y) + \text{div} (Ab y) + cy = f - \text{div} (Af), \quad y|_{\partial \Omega} = 0.
\]

Unique solvability can be shown by standard arguments and we omit the details here.

The choices

- \( X = H = L^2(\Omega; \mathbb{R}), X_{ad} = \{ v \in X : a \leq v \leq b \text{ a.e. in } \Omega \} \) with \( a < b \),

- \( Y = Y^* = H_0^1(\Omega) \times H(\text{div} ; \Omega) \),

- \( Z = L^2(\Omega; \mathbb{R}) \times L^2(\Omega; \mathbb{R}^d) \)

- \( \mathcal{I}(y, \sigma) = (y, \sigma), \mathcal{I}^*(p, \xi) = (p, \xi) \),

- \( A(y, \sigma) = y \),

- \( Bu = -(u, 0) \),

- \( Cu = \lambda u \) with \( \lambda > 0 \),

give rise to the following optimal control problem: Given \( z_d \in L^2(\Omega; \mathbb{R}) \) and \((f, f_k) \in Z \) find the minimizer of

\[
\min \| y - z_d \|^2_{L^2(\Omega)} + \lambda \| u \|^2_{L^2(\Omega)} \quad \text{subject to}
\]

\[
\begin{align*}
\text{div} \, \sigma + cy &= f + u, \\
\nabla y - by + A^{-1} \sigma &= f, \\
y|_{\partial \Omega} &= 0.
\end{align*}
\]
The corresponding least-squares operators are given by

\begin{align*}
\mathcal{L} y &= \left( \text{div } \sigma + cy, \nabla y - by + A^{-1}\sigma \right), \\
\mathcal{L}^* p &= \left( -\text{div } \xi - b \cdot \xi + cp, -\nabla p + A^{-1}\xi \right)
\end{align*}

for \( y = (y, \sigma) \in Y, \ p = (p, \xi) \in Y^* \). In particular, integration by parts proves that

\[ (\mathcal{L} y, \mathcal{I}^* p) = (\mathcal{I} y, \mathcal{L}^* p) \quad \forall \ y \in Y, \ p \in Y^*. \]

We note that \( \text{ran}(\mathcal{L}) = Z = \text{ran}(\mathcal{L}^*), \ ker(\mathcal{L}) = \{0\} = ker(\mathcal{L}^*) \) which can be seen by following the analysis of, e.g., [7], so that this example fits the abstract framework from Sect. 2.

### 3.2 Stokes problem

Let \( \Omega \subseteq \mathbb{R}^d, d = 2, 3 \) denote a bounded Lipschitz domain. We consider the Stokes equations

\begin{align*}
-\Delta y + \nabla p &= f, \\
\text{div } y &= 0, \\
y|_{\partial\Omega} &= 0.
\end{align*}

For a suitable least-squares formulation we use the pseudostress tensor \( M = \nabla y - p I \), see [8, Section 3.2] for the definition and analysis of a similar formulation, where \( I \) denotes the identity tensor. Here, the vector gradient \( \nabla v \) is the tensor where the \( j \)-th row is the gradient of the \( j \)-th component of \( v \) and \( \Delta v \) is the vector where the \( j \)-th row is the Laplacian of the \( j \)-th component of \( v \). In particular, by noting that \( \text{tr } M = \text{div } y - dp = -dp \) we can eliminate the pressure variable \( p \) from the system. Let \( \text{Dev } M = M - \frac{1}{d} \Pi_O \text{tr } M \) denote the deviatoric part of \( M \in L^2(\Omega; \mathbb{R}^{d \times d}), \ \Pi_O \) the \( L^2(\Omega; \mathbb{R}) \) orthogonal projection onto constants, and \( \text{div} \) the row-wise divergence operator. Using the spaces

\begin{align*}
Y &= Y^* = H^1_0(\Omega) \times \mathcal{H}(\text{div } ; \Omega) \\
&= H^1_0(\Omega) \times \left\{ N \in L^2(\Omega; \mathbb{R}^{d \times d}) : \text{div } N \in L^2(\Omega; \mathbb{R}^d) \right\}
\end{align*}

we define the least-squares operators

\[ \mathcal{L}(y, M) = \mathcal{L}^*(y, M) = (-\text{div } M, \nabla y - \text{Dev } M - \frac{1}{d} \Pi_O \text{tr } M), \quad (12) \]

and \( \mathcal{I}(y, M) = \mathcal{I}^*(y, M) = (y, M) \). One verifies that these operators are linear and bounded in the canonical norms. We note that \( (M, N)_{L^2(\Omega)} = \int_\Omega M : N \, dx \), where \( M : N \) denotes the Frobenius inner product of two tensors \( M, N \in \mathcal{H}(\text{div } ; \Omega) \).
Then, \((\text{Dev } M, I)_{L^2(\Omega)} = 0\) and \((\text{Dev } M, \text{N})_{L^2(\Omega)} = (M, \text{Dev } N)_{L^2(\Omega)}\) by a straightforward calculation. To see (A3) set \(Z = L^2(\Omega; \mathbb{R}^d) \times L^2(\Omega; \mathbb{R}^{d \times d})\) and let \((y, M), (p, \text{N}) \in Y = Y^*\) be given. Integration by parts yields

\[
(L(y, M), I^*(p, \text{N}))_Z = (-\text{div } M, p)_{L^2(\Omega)} + (\nabla y - \text{Dev } M - \frac{1}{d} \Pi_\Omega \text{tr } M, N)_{L^2(\Omega)} = (y, -\text{div } N)_{L^2(\Omega)} + (M, -\text{Dev } M)_{L^2(\Omega)} - \frac{1}{d} (\Pi_\Omega \text{tr } M, \text{tr } N)_{L^2(\Omega)} = (y, -\text{div } N)_{L^2(\Omega)} + (M, \nabla p - \text{Dev } N - \frac{1}{d} I \Pi_\Omega \text{tr } N)_{L^2(\Omega)} = (I(y, M), L^*(p, \text{N}))_Z.
\]

Assumption (A2) is shown in Theorem 16 below. For the optimal control problem of interest we further have

- \(X = H = L^2(\Omega; \mathbb{R}^d), X_{\text{ad}} = \{v \in X : a \leq v \leq b \text{ a.e.}\}\) with \(a, b \in \mathbb{R}^d\),
- \(A(y, M) = y, Eu = -(u, 0), Cu = u\).

Thus, we have the optimal control problem: Given \((f, 0) \in Z, z_d \in H\) find the solution \(u \in X_{\text{ad}}\) of

\[
\begin{align*}
\min \|y - z_d\|^2_{L^2(\Omega)} + \lambda \|u\|^2_{L^2(\Omega)} & \text{ subject to } \\
-\text{div } M &= f + u, \\
\nabla y - \text{Dev } M - \frac{1}{d} I \Pi_\Omega \text{tr } M &= 0, \\
y|_{\partial \Omega} &= 0.
\end{align*}
\]

We stress that the equations above are a formulation of the Stokes problem

\[-\Delta y + \nabla p = f + u, \quad \nabla y = 0, \quad \text{and } \Pi_\Omega p = 0.\]

The proof of the next theorem uses unique solvability of the weak formulation of the Stokes problem, see, e.g., [3, Theorem 8.2.1].

**Theorem 16** Let \(L\) denote the operator from (12). We have that \(\ker L = \{0\}\) and \(\text{ran } L = Z\).

**Proof** Let \((f, F) \in Z\) be given. We prove surjectivity of \(L\). To that end, let \((y, p) \in H^1_0(\Omega) \times L^2(\Omega; \mathbb{R})\) with \(\Pi_\Omega p = 0\) denote the (unique) weak solution of the Stokes equations

\[-\Delta y + \nabla p = f - \text{div } F, \quad \text{div } y = (1 - \Pi_\Omega) \text{tr } F.\]
Then, define $M := \nabla y - pI - F \in L^2(\Omega; \mathbb{R}^{d \times d})$. Taking the trace gives

$$\text{tr} \, M = \text{div} \, y - dp - \text{tr} \, F = (1 - \Pi_{\Omega}) \text{tr} \, F - dp - \text{tr} \, F = -\Pi_{\Omega} \text{tr} \, F - dp.$$ 

Applying $\Pi_{\Omega}$ we get $\Pi_{\Omega} \text{tr} \, M = -\Pi_{\Omega} \text{tr} \, F$. Putting the last identities together we see that

$$\text{Dev} \, M = \nabla y - pI - F - \frac{1}{d} I \Pi_{\Omega} \text{tr} \, M + pI,$$

or equivalently, $\nabla y - \text{Dev} \, M - \frac{1}{d} I \Pi_{\Omega} \text{tr} \, M = F$. We have to show that $\text{div} \, M \in L^2(\Omega; \mathbb{R}^d)$ and $\text{div} \, M = -f$. Recall that from the weak formulation of the Stokes problem we have that

$$(\nabla y, \nabla v)_{L^2(\Omega)} - (p, \text{div} \, v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} + (F, \nabla v)_{L^2(\Omega)} \quad \forall v \in H^1_0(\Omega).$$

Using this identity we get for any $v \in D(\Omega) \subseteq H^1_0(\Omega)$ (smooth test functions with vanishing traces)

$$\langle \text{div} \, M, v \rangle = -(M, \nabla v)_{L^2(\Omega)} = -(\nabla y, \nabla v)_{L^2(\Omega)} + (p, \text{div} \, v)_{L^2(\Omega)} + (F, \nabla v)_{L^2(\Omega)}$$

$$= (-f, v)_{L^2(\Omega)}.$$

Due to density of $D(\Omega)$ in $L^2(\Omega; \mathbb{R}^d)$ this proves that $\text{div} \, M = -f$. Thus, we have proven that for given $(f, F) \in Z$ there exists $(y, M) \in Y$ with $L(y, M) = (f, F)$.

To show injectivity of $L$ let $(y, M) \in Y$ with $L(y, M) = 0$ be given, i.e.,

$$-\text{div} \, M = 0,$$

$$\nabla y - \text{Dev} \, M - \frac{1}{d} I \Pi_{\Omega} \text{tr} \, M = 0.$$ 

Taking the trace and then applying $\Pi_{\Omega}$ to the second equation shows $\Pi_{\Omega} \text{tr} \, M = 0$. We then have

$$\nabla y - \text{Dev} \, M = 0.$$ 

Taking the trace of the latter identity we see that $\text{div} \, y = 0$. Setting $p := -\frac{1}{d} \text{tr} \, M$ and eliminating $M$ in the equation $\text{div} \, M = 0$ we see that $(y, p)$ satisfy the Stokes equations

$$-\Delta y + \nabla p = 0, \quad \text{div} \, y = 0, \quad \Pi_{\Omega} p = 0.$$ 

By unique solvability we conclude that $y = 0$. Furthermore, we conclude $\text{Dev} \, M = 0$, that is $M = \frac{1}{d} I \text{tr} \, M = -pI = 0$. This finishes the proof. \hfill \square
3.3 Maxwell equation

We consider the real 3D Maxwell problem: Let $\Omega$ denote a simply connected Lipschitz domain. Given $f_1, f_2 \in L^2(\Omega; \mathbb{R}^3)$, $c \in L^\infty(\Omega; \mathbb{R})$, find $(y, \sigma) \in H_0(\text{curl}; \Omega) \times H(\text{curl}; \Omega)$ such that

\[
\begin{align*}
\text{curl } \sigma + cy &= f_1, \\
\text{curl } y - \sigma &= f_2.
\end{align*}
\]  

(14)

This is a first-order reformulation of the PDE

\[
\text{curl curl } y + cy = f_1 + \text{curl } f_2, \quad y \times n|_{\partial \Omega} = 0.
\]  

(15)

Here, the spaces are given by

\[
H(\text{curl}; \Omega) := \{ v \in L^2(\Omega; \mathbb{R}^3) : \text{curl } v \in L^2(\Omega; \mathbb{R}^3) \},
\]

\[
H_0(\text{curl}; \Omega) = \{ v \in H(\text{curl}; \Omega) : v \times n|_{\partial \Omega} = 0 \},
\]

where $n$ denotes the normal vector field on $\partial \Omega$. Moreover, we assume that $c$ is strictly positive, or if $c < 0$, that $c$ is a constant and is not an eigenvalue of the cavity problem, see, e.g. [9, Section 2.4] and [27, Section 4.7]. We note that $\mathcal{I}(y, \sigma) = (y, \sigma), \mathcal{L}^* = \mathcal{L}$,

\[
\mathcal{L}(y, \sigma) = (\text{curl } \sigma + cy, \text{curl } y - \sigma) \quad \forall (y, \sigma) \in Y = Y^* := H_0(\text{curl}; \Omega) \times H(\text{curl}; \Omega),
\]

$\mathcal{L}^* = \mathcal{L}$, and $Z = L^2(\Omega; \mathbb{R}^3) \times L^2(\Omega; \mathbb{R}^3)$. To see (A3) integration by parts yields for all $(y, \sigma), (z, \tau) \in Y$,

\[
(\mathcal{L}(y, \sigma), \mathcal{L}^*(z, \tau))_Z = (\text{curl } \sigma + cy, z)_{L^2(\Omega)} + (\text{curl } y - \sigma, \tau)_{L^2(\Omega)}
\]

\[
= (\sigma, \text{curl } z)_{L^2(\Omega)} + (y, cz)_{L^2(\Omega)}
\]

\[
+ (y, \text{curl } \tau)_{L^2(\Omega)} - (\sigma, \tau)_{L^2(\Omega)}
\]

\[
= (y, \text{curl } \tau + cz)_{L^2(\Omega)} + (\sigma, \text{curl } z - \tau)_{L^2(\Omega)}
\]

\[
= (\mathcal{I}(y, \sigma), \mathcal{L}^*(z, \tau))_Z.
\]

Surjectivity resp. injectivity of $\mathcal{L}$ can be seen by analyzing problem (14) or, equivalently, (15). In the case $f_2 = 0$ problem (15) has been studied in full length in [27, Ch.4]. There it is also mentioned that general right-hand side ($f_2 \neq 0$ in our setting) can be included. We provide a proof for completeness.

**Proposition 17** We have that $\text{ran}(\mathcal{L}) = Z$ and $\ker(\mathcal{L}) = \{0\}$.

**Proof** It is straightforward to verify that (15) and (14) are equivalent. Thus, it suffices to analyze the weak form of (15). The variational formulation of (15) is: Find $y \in H_0(\text{curl}; \Omega)$ such that

\[
(\text{curl } y, \text{curl } z)_{L^2(\Omega)} + (cy, z)_{L^2(\Omega)} = (f_1, z)_{L^2(\Omega)} + (f_2, \text{curl } z)_{L^2(\Omega)}
\]  

(16)
for all \( z \in H_0(\text{curl}; \Omega) \). If \( c > 0 \), then the left-hand side defines a symmetric bilinear form that induces an inner product that is equivalent to the canonic inner product in \( H_0(\text{curl}; \Omega) \). We may thus assume that \( c < 0 \) is a constant and not an eigenvalue of the cavity problem. The case \( f_2 = 0 \) is found in [27, Corollary 4.19]. Note that [27, Corollary 4.19] also implies that \( \ker(L) \) is trivial. By linearity we may thus assume that \( f_1 = 0 \) and \( f_2 \neq 0 \). Consider the Helmholtz decomposition

\[
y = y_0 + \nabla y_0 \quad \text{where } y_0 \in H_0^1(\Omega) \text{ solves } \Delta y_0 = \text{div } y.
\]

By construction, \( \text{div } y_0 = \text{div } y - \Delta y_0 = 0 \), and \( \nabla y_0 \times n|_{\partial \Omega} = 0 \) since \( y_0|_{\partial \Omega} = 0 \). Then, \( y_0 = y - \nabla y_0 \in \tilde{H} = \{ v \in H_0(\text{curl}; \Omega) : \text{div } v = 0 \} \). Using \( z = \nabla q \) for any \( q \in H_0^1(\Omega) \) in (16) one finds that \( \text{div } y = 0 \), thus, a solution to (16) satisfies \( y = y_0 \). Consequently, (16) is equivalent to: Find \( y_0 \in \tilde{H} \) such that

\[
(\text{curl } y_0, \text{curl } z_0)_{L^2(\Omega)} + c(y_0, z_0)_{L^2(\Omega)} = (f_2, \text{curl } z_0)_{L^2(\Omega)}
\]

for all \( z_0 \in \tilde{H} \). This variational formulation has a unique solution because the first term on the left-hand side defines a bounded and coercive bilinear form [27, Corollary 3.51] and the embedding from \( \tilde{H} \) to \( L^2(\Omega, \mathbb{R}^3) \) is compact [27, Corollary 3.49]. □

The optimal control problem under consideration is: Given \( f \in L^2(\Omega; \mathbb{R}^3) \) and \( z_d \in L^2(\Omega; \mathbb{R}^3) \), solve the minimization problem

\[
\min \| y - z_d \|_{L^2(\Omega)}^2 + \lambda \| u \|_{L^2(\Omega)}^2 \quad \text{subject to}
\]

\[
\text{curl } \sigma + cy = f + u, \\
\text{curl } y - \sigma = 0, \\
y \times n|_{\partial \Omega} = 0.
\]

The remaining operators and spaces to fit the abstract framework are given by

- \( X = H = L^2(\Omega; \mathbb{R}^3), X_{ad} = \{ v \in X : a \leq v \leq b \text{ a.e. } \Omega \} \), \( a < b \).
- \( A(y, \sigma) = y, B u = -(u, 0), C u = u \).

### 3.4 Heat equation

Let \( \Omega \subseteq \mathbb{R}^d, d \geq 1 \) denote a bounded Lipschitz domain, \( J = (0, T) \) a time interval, and \( Q = J \times \Omega \) the time-space domain. Given the data \( f \in L^2(Q; \mathbb{R}), y_0 \in L^2(\Omega; \mathbb{R}) \) consider the heat equation

\[
\begin{align*}
\partial_t y - \Delta y &= f, \\
y|_{J \times \partial \Omega} &= 0, \\
y(0) &= y_0.
\end{align*}
\]
and its first-order reformulation

\[ \partial_t y - \text{div} \sigma = f, \]
\[ \nabla y - \sigma = 0, \]
\[ y|_{\partial \Omega} = 0, \]
\[ y(0) = y_0. \]

We analyze the following optimal control problem. Given \( f \in L^2(Q; \mathbb{R}), y_0 \in L^2(\Omega; \mathbb{R}) \) and the desired states \( z_d \in L^2(Q; \mathbb{R}), z_{d,T} \in L^2(\Omega; \mathbb{R}) \), we consider

\[
\min \| y - z_d \|_{L^2(Q)}^2 + \| y(T) - z_{d,T} \|_{L^2(\Omega)}^2 + \lambda \| u \|_{L^2(Q)}^2 + \lambda \| u_0 \|_{L^2(\Omega)}^2 \]
subject to
\[
\partial_t y - \text{div} \sigma = f + u, \]
\[ \nabla y - \sigma = 0, \]
\[ y|_{\partial \Omega} = 0, \]
\[ y(0) = y_0 + u_0. \]

The spaces and operators that fit the abstract framework are:

- \( X = H = L^2(Q; \mathbb{R}) \times L^2(\Omega; \mathbb{R}) \),
- \( X_{\text{ad}} = \{ (v, v_0) \in X : a \leq v \leq b \text{ a.e. in } Q, a_0 \leq v_0 \leq b_0 \text{ a.e. in } \Omega \} \),
- \( Y = \{ (y, \sigma) \in L^2(H^1_0(\Omega)) \times L^2(Q; \mathbb{R}^d) : \partial_t y - \text{div} \sigma \in L^2(Q; \mathbb{R}) \} \),
- \( Y^* = \{ (y, \sigma) \in L^2(H^1_0(\Omega)) \times L^2(Q; \mathbb{R}^d) : \partial_t y + \text{div} \sigma \in L^2(Q; \mathbb{R}) \} \),
- \( Z = L^2(Q; \mathbb{R}) \times L^2(Q; \mathbb{R}) \times L^2(\Omega; \mathbb{R}) \),
- \( \mathcal{T}(y, \sigma) = (y, \sigma, y(T)), \mathcal{T}^*(p, \xi) = (p, \xi, p(0)) \),
- \( \mathcal{A}(y, \sigma, z) = (y, z) \text{ for } (y, \sigma, z) \in Z \),
- \( \mathcal{B}(u, u_0) = -(u, 0, u_0) \),
- \( \mathcal{C}(u, u_0) = (u, u_0) \).

From the analysis of [17, Section 2] we note that \( Y \) and \( Y^* \) are Hilbert spaces with norms

\[
\| (y, \sigma) \|_Y^2 = \| \nabla y \|_{L^2(Q)}^2 + \| \sigma \|_{L^2(Q)}^2 + \| \partial_t y - \text{div} \sigma \|_{L^2(Q)}^2, \\
\| (y, \sigma) \|_{Y^*}^2 = \| \nabla y \|_{L^2(Q)}^2 + \| \sigma \|_{L^2(Q)}^2 + \| \partial_t y + \text{div} \sigma \|_{L^2(Q)}^2. 
\]

The least-squares operators are given by

\[
\mathcal{L} y = (\partial_t y - \text{div} \sigma, \nabla y - \sigma, y(0)), \quad \mathcal{L}^* p = (-\partial_t p - \text{div} \xi, \nabla p - \xi, p(T)).
\]
In [17, Theorem 2.3 and Remark 2.6] it is shown that \( L : Y \rightarrow Z \) is an isomorphism. The same argumentation proves that \( L^* \) is an isomorphism. To see (A3) we employ integration by parts in space and in time to get for all \((y, \sigma) \in Y, (p, \xi) \in Y^*\)

\[
(L(y, \sigma), I^*(p, \xi))_Z = (\partial_t y - \text{div} \sigma, p)_{L^2(Q)} + (\nabla y - \sigma, \xi)_{L^2(Q)} + (y(0), p(0))_{L^2(\Omega)}
\]

\[
= (\sigma, \nabla p - \xi)_{L^2(Q)} + (y, -\partial_t p - \text{div} \xi)_{L^2(Q)} + (y(T), p(T))_{L^2(\Omega)}
\]

\[
= (I(y, \sigma), L^*(p, \xi))_Z.
\]

### 4 Numerical experiments

In this section we present numerical results for some of the examples described in Sect. 3. We consider optimal control problems subject to the Poisson equation (Sect. 4.2 for an unconstrained problem and Sects. 4.3, 4.4 for a constrained problem), the Stokes equations (Sect. 4.5) and the heat equation (Sect. 4.6).

All programs have been implemented in MATLAB version 2022a on a LINUX machine with AMD RYZEN 1700+ processor and 32GB RAM.

#### 4.1 Discretization

For all problems except the heat equation we consider a two-dimensional domain \( \Omega \) with polygonal boundary. For the heat equation we consider the time interval \( J = (0, T) \) and a one-dimensional spatial domain \( \omega \) (i.e., interval), hence, the space-time domain \( Q = J \times \omega \) is a two-dimensional domain with polygonal boundary. For simplicity we write \( \Omega \) for spatial and space-time domains in the definitions below.

We consider a regular triangulation \( T \) of \( \Omega \) into open, non-empty triangles,

\[
\overline{\Omega} = \bigcup_{T \in T} \overline{T}.
\]

The maximum element diameter is denoted by \( h \). The following finite element spaces are used:

\[
\mathcal{P}^p(T) := \{ v \in L^2(\Omega) : v|_T \text{ is a polynomial of degree } \leq p, \text{ for all } T \in T \},
\]

\[
S^1(T) := \mathcal{P}^1(T) \cap H^1(\Omega), \quad S^1_0(T) := \mathcal{P}^1(T) \cap H^1_0(\Omega),
\]

\[
\mathcal{R}T^0(T) := \{ \tau \in H(\text{div}; \Omega) : \tau|_T(x, y) = (\alpha, \beta)^\top + \gamma(x, y)^\top, \alpha, \beta, \gamma \in \mathbb{R}, \text{ for all } T \in T \}.
\]

We consider box constraints, i.e., \( X_{\text{ad}} = \{ v \in L^2(\Omega; \mathbb{R}) : a \leq v \leq b \text{ a.e. in } \Omega \} \) with \( a < b \) and use the discrete space

\[
X_{h, \text{ad}} := X_{\text{ad}} \cap \mathcal{P}^0(T).
\]
By equipping $P^0(T)$ with the canonical basis of characteristic functions, the constraints in $X_{h, \text{ad}}$ are pointwise constraints and we use the active set strategy proposed in [21, Algorithm A2] to solve the discretized variational inequalities.

We employ a standard adaptive finite element loop consisting of the four basic steps: SOLVE, ESTIMATE, MARK, and REFINE. For the estimation step we use the a posteriori error estimators proposed in Sect. 2.3 for constrained problems and in Sect. 2.4 for unconstrained problems. Note that both estimators can be written as a sum of local contributions, i.e.,

$$\xi^2 = \sum_{T \in T} \xi(T)^2,$$

where $\xi$ denotes one of the proposed estimators. As marking strategy in the example from Sect. 4.4 we use the bulk criterion: Find a (minimal) set $M \subseteq T$ such that

$$\frac{1}{4} \xi^2 \leq \sum_{T \in M} \xi(T)^2.$$

The refinement step is realized employing the newest-vertex bisection algorithm. Uniform refinements correspond to the case where all elements in the mesh are marked for refinement.

Let us note that the bilinear form $a(\cdot; \cdot)$ is coercive provided that the scaling parameters $\alpha, \beta$ are chosen sufficiently large, see Theorem 6. In the numerical experiments we have chosen $\alpha = 1, \beta = \lambda^{-1}$ and found that the results do not significantly change for different choices of this parameter provided that $\alpha \approx 1, \beta \approx \lambda^{-1}$.

### 4.2 Unconstrained problem subject to Poisson equation

Let $\Omega = (0, 1)^2$. We consider the optimal control problem

$$\min_{u \in L^2(\Omega)} \|y - z_d\|_{L^2(\Omega)}^2 + \lambda \|u\|_{L^2(\Omega)}^2 \quad \text{subject to}$$

$$-\Delta y = u + f \quad \text{in } \Omega,$$

$$y|_{\partial \Omega} = 0.$$  \hspace{1cm} (18)

In view of the abstract framework we have that $X = X_{\text{ad}} = L^2(\Omega; \mathbb{R})$ and thus an unconstrained optimal control problem. For this problem we consider the manufactured solution

$$y(x_1, x_2) = \sin(\pi x_1) \sin(\pi x_2),$$

$$p(x_1, x_2) = x_1(1 - x_1)x_2(1 - x_2).$$  \hspace{1cm} (19)
and \( u \) resp. \( f, z_d \) are then computed by

\[
\begin{align*}
u &= -\frac{1}{\lambda} p, \\
f &= -\Delta y - u, \\
z_d &= \Delta p + y.
\end{align*}
\]

Since this is an unconstrained control problem we use the method described in Sect. 2.4 with the lowest-order finite element space

\[Y_h \times Y_h^* = (S_0^1(T) \times RT^0(T)) \times (S_0^1(T) \times RT^0(T)),\]

i.e., we compute minimizer \((y_h, p_h)\) of the functional \(G(\cdot; f, z_d)\) over the space \(Y_h \times Y_h^*\). Since the solution is smooth we expect that

\[
\|y - y_h\|_Y + \|p - p_h\|_{Y^*} = O(h).
\]

This is observed in the left plot of Fig. 1 where we visualize the errors \(\|y - y_h\|_Y\), \(\|p - p_h\|_Y\) and the a posteriori error estimator \(G(y_h, p_h; f, z_d)\) for \(\lambda = 10^{-2}\). The bottom row of Fig. 1 shows the effectivity index

\[
\text{eff}_{\text{ue}} := \frac{\sqrt{G(y_h, p_h; f, z_d)}}{\|(y - y_h, p - p_h)\|_{Y \times Y^*}}
\]

for different values of \(\lambda\). Here, we use uniform mesh refinement.

### 4.3 Constrained problem subject to Poisson equation

In this section we consider the control problem (18) and the same setting again but replace the control space \(X_{ad} = L^2(\Omega; \mathbb{R})\) by

\[X_{ad} = \{v \in L^2(\Omega; \mathbb{R}) : -1 \leq v \leq 0 \text{ a.e. in } \Omega\}.\]

We prescribe the solutions (19) and define

\[u = \Pi_{X_{ad}} (-\lambda^{-1} p),\]

and compute the data \(f, z_d\) by

\[
\begin{align*}
f &= -\Delta y - u, \\
z_d &= \Delta p + y.
\end{align*}
\]

We use the proposed method described in Sect. 2.3, i.e., we solve the variational inequality (4) with the lowest-order finite element space

\[H = X_{ad} \cap P^0(T) \times (S_0^1(T) \times RT^0(T)) \times (S_0^1(T) \times RT^0(T)).\]
The solution pair \((y, p)\) is smooth and \(u \in H^1(\Omega)\). Let \(\Pi_h^0\) denote the \(L^2(\Omega)\) orthogonal projection onto piecewise constants. Using Corollary 8 we infer that

\[
\|u - u_h\|_{L^2(\Omega)}^2 + \|y - y_h\|_Y^2 + \|p - p_h\|_Y^2 \\
\lesssim \|u - \Pi_h^0 u\|_{L^2(\Omega)}^2 + \|y - z_h\|_Y^2 + \|p - q_h\|_Y^2 + \|(p + \lambda u, \Pi_h^0 u - u)\|_{L^2(\Omega)}
\]

for \((z_h, q_h) \in Y_h \times Y_h\), which is possible since \(\Pi_h^0 u \in [a, b]\) if \(a \leq u \leq b\) a.e., thus, \(\Pi_h^0 u \in X_{h, \text{ad}}\). Employing \(\|(1 - \Pi_h^0)u\|_{L^2(\Omega)} \lesssim h\|\nabla u\|_{L^2(\Omega)}\) we infer that

\[
\|(1 - \Pi_h^0)u\|_{L^2(\Omega)}^2 + \|(p + \lambda u, (1 - \Pi_h^0)u)\|_{L^2(\Omega)} \\
= \|(1 - \Pi_h^0)u\|_{L^2(\Omega)}^2 + \|(1 - \Pi_h^0)(p + \lambda u), (1 - \Pi_h^0)u\|_{L^2(\Omega)} \\
\lesssim h^2\|\nabla u\|_{L^2(\Omega)}^2 + h\|\nabla (p + \lambda u)\|_{L^2(\Omega)}\|\nabla u\|_{L^2(\Omega)} = O(h^2).
\]
Putting all together we obtain with standard approximation results that

$$
\| u - u_h \|_{L^2(\Omega)} + \| y - y_h \|_Y + \| p - p_h \|_Y = O(h).
$$

This rate is also observed in our numerical experiment for the moderate value $\lambda = 10^{-2}$, see the top row of Fig. 1 (right plot). In the bottom row we also show the effectivity index

$$
\text{eff}_c := \frac{\eta}{\| (u - u_h, y - y_h, p - p_h) \|_{X \times Y \times Y^*}}
$$

for different values of $\lambda$. Here, we use uniform mesh refinement.

### 4.4 Constrained problem with singular solution

We consider the control problem (18) with $\lambda = 1$ and the L-shaped domain $\Omega = (-1, 1)^2 \setminus [-1, 0]^2$. The set of admissible controls is given by

$$
X_{ad} = \{ v \in L^2(\Omega) : 0.1 \leq v \leq 0.12 \text{ a.e. in } \Omega \}
$$

and as data we use

$$
f(x_1, x_2) = 0, \ z_d(x_1, x_2) = 1, \ (x_1, x_2) \in \Omega.
$$

For this setup we do not know an explicit representation of the solution $(u, y, p)$ but stress that reduced regularity is expected due to the non-convexity of the domain. Figure 2 shows that the error estimator $\eta$ seems to asymptotically converge at $O(h^{2/3})$ for a sequence of uniform meshes. Using adaptively refined meshes we observe that optimal rates are reestablished, i.e., the error estimator converges at $O(N^{-1/2})$ with $N = \dim(X_h \times Y_h \times Y_h)$. The right plot in top row of Fig. 2 visualizes the solution $u_h \in X_{h, ad}$ on a mesh with $\#T = 720$ triangles obtained from the adaptive loop. Strong mesh refinements towards the re-entrant corner are observed (Fig. 2 bottom row), indicating that the adaptive algorithm detects singularities.

### 4.5 Unconstrained problem subject to Stokes equation

Let $\Omega = (0, 1)^2$. We consider the optimal control problem (13) with $X = X_{ad}$, i.e., the unconstrained control problem. Our manufactured solution is given by

$$
y(x, y) = p(x, y) = \text{curl} \left( x(1 - x)y(1 - y) \right)^2
$$

and the data $f = -\Delta y + \lambda^{-1} p, \ z_d = \Delta p + y$ is computed thereof. Since this is an unconstrained control problem we employ the method described in Sect. 2.4, see also Sect. 4.2 for a similar numerical experiment for the Poisson problem. The solution is smooth so that we expect optimal convergence of the lowest-order scheme which is
Fig. 2 Top row: The left plot shows the error estimator $\eta$ on a sequence of uniformly refined meshes and a sequence of locally refined meshes generated by the adaptive algorithm. The black dashed lines indicate $O(N^{-1/3})$ resp. $O(N^{-1/2})$ where $N = \dim(X_{h,\text{ad}} \times Y_h \times Y_h)$. The right plot shows the discrete control $u_h$ on a mesh with $#T = 720$ triangles. Bottom row: Two meshes generated by the adaptive algorithm with $#T = 513$ resp. $#T = 720$ elements

also observed in Fig. 3 for different values of $\lambda$. Here, we use uniform mesh refinement. Figure 3 also shows the effectivity index $\text{eff}_{uc}$. One observes that as $\lambda$ gets smaller the error in $y$ has a pre-asymptotic range where it converges at a higher rate. This explains that for $\lambda = 10^{-3}$ the effectivity index grows.

4.6 Constrained problem subject to heat equation

Let $J = (0, 1) = \Omega$, $Q = J \times \Omega$. We consider the optimal control problem (17) with

$$X_{\text{ad}} = \{(v, v_0) \in L^2(Q) \times L^2(\Omega) : -1 \leq v \leq 0, -1 \leq v_0 \leq 0\}.$$

We use the manufactured solution

$$y(x, t) = t \sin(\pi x),$$

$$p(x, t) = (1 - t)x(1 - x).$$
Then, \((u, u_0) = \Pi_{X_{ad}} (-\lambda^{-1} p, -\lambda^{-1} p_0)\) and the data \(f, z_d, z_d, T\) is computed by

\[
\begin{align*}
    f &= \partial_t y - \partial_{xx} y - u, \\
    z_d &= \partial_t p + \partial_{xx} p + y, \\
    z_d, T &= -p(T) + y(T).
\end{align*}
\]

Since the solution \((y, p)\) is smooth and \(u \in H^1(Q)\), \(u_0 \in H^1(\Omega)\) one expects that

\[
\|u - u_h\|_{L^2(Q)} + \|u_0 - u_h(0)\|_{L^2(\Omega)} + \|y - y_h\|_Y + \|p - p_h\|_{Y^*} = O(h),
\]

see Theorem 7 by using a similar argumentation as in Sect. 4.3. Figure 4 shows the a posteriori estimators and errors for this problem for different values of \(\lambda\). We observe the optimal rate \(O(h) = O(N^{-1/2})\) where \(N = \dim(X_h \times Y_h \times Y^*_h)\) on a sequence of uniformly refined meshes. Furthermore, Fig. 4 also shows the effectivity index \(\text{eff}_{\text{uc}}\). We note that for \(\lambda \leq 10^{-3}\) the employed solver (active set strategy) did not converge.
Fig. 4 Error estimator and errors for the optimal control problem (17) considered in Sect. 4.6. The dashed black line indicates $O(h)$. The right plot in the bottom row shows the effectivity index $\text{eff}_\text{c}$.

within reasonable time. This requires some further analysis and possible fine-tuning of parameters.

Acknowledgements This work was supported by ANID through FONDECYT Projects and 1210391 (TF), and 1210579 (MK).

References

1. Bochev, P., Gunzburger, M.D.: Least-squares finite element methods for optimality systems arising in optimization and control problems. SIAM J. Numer. Anal. 43(6), 2517–2543 (2006)
2. Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Applied Mathematical Sciences, vol. 166. Springer, New York (2009)
3. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications. Springer Series in Computational Mathematics, vol. 44. Springer, Heidelberg (2013)
4. Bramble, J.H., Lazarov, R.D., Pasciak, J.E.: A least-squares approach based on a discrete minus one inner product for first order systems. Math. Comput. 66(219), 935–955 (1997)
Least-squares finite elements for distributed optimal…

5. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, vol. 15, 3rd edn. Springer, New York (2008)
6. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext. Springer, New York (2011)
7. Cai, Z., Lazarov, R., Manteuffel, T.A., McCormick, S.F.: First-order system least squares for second-order partial differential equations. I. SIAM J. Numer. Anal. 31(6), 1785–1799 (1994)
8. Cai, Z., Lee, B., Wang, P.: Least-squares methods for incompressible Newtonian fluid flow: linear stationary problems. SIAM J. Numer. Anal. 42(2), 843–859 (2004)
9. Carstensen, C., Storn, J.: Asymptotic exactness of the least-squares finite element residual. SIAM J. Numer. Anal. 56(4), 2008–2028 (2018)
10. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements. Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
11. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence, RI (1998)
12. Falk, R.S.: Error estimates for the approximation of a class of variational inequalities. Math. Comput. 28, 963–971 (1974)
13. Führer, T., Heuer, N., Karkulik, M.: On the coupling of DPG and BEM. Math. Comput. 86(307), 2261–2284 (2017)
14. Führer, T., Heuer, N., Stephan, E.P.: On the DPG method for Signorini problems. IMA J. Numer. Anal. 38(4), 1893–1926 (2018)
15. Führer, T., Karkulik, M.: Space-time least-squares finite elements for parabolic equations. Comput. Math. Appl. 92, 27–36 (2021)
16. Führer, T., Karkulik, M.: Space-time finite element methods for parabolic distributed optimal control problems. arXiv arXiv:2208.09879 (2022)
17. Gantner, G., Stevenson, R.: Further results on a space-time FOSLS formulation of parabolic PDEs. ESAIM Math. Model. Numer. Anal. 55(1), 283–299 (2021)
18. Gantner, G., Stevenson, R.: Applications of a space-time FOSLS formulation for parabolic PDEs. IMA J. Numer. Anal. (2023)
19. Gong, W., Hinze, M., Zhou, Z.J.: Space-time finite element approximation of parabolic optimal control problems. J. Numer. Math. 20(2), 111–145 (2012)
20. Gunzburger, M.D., Kunoth, A.: Space-time adaptive wavelet methods for optimal control problems constrained by parabolic evolution equations. SIAM J. Control Optim. 49(3), 1150–1170 (2011)
21. Kärkkäinen, T., Kunisch, K., Tarvainen, P.: Augmented Lagrangian active set methods for obstacle problems. J. Optim. Theory Appl. 119(3), 499–533 (2003)
22. Langer, U., Steinbach, O., Tröltzsch, F., Yang, H.: Space-time finite element discretization of parabolic optimal control problems with energy regularization. SIAM J. Numer. Anal. 50(2), 675–695 (2021)
23. Langer, U., Steinbach, O., Tröltzsch, F., Yang, H.: Unstructured space-time finite element methods for optimal control of parabolic equations. SIAM J. Sci. Comput. 43(2), A744–A771 (2021)
24. Lions, J.-L.: Optimal control of systems governed by partial differential equations. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer, New York-Berlin (1971). Translated from the French by S. K. Mitter
25. Lions, J.-L., Stampacchia, G.: Variational inequalities. Commun. Pure Appl. Math. 20, 493–519 (1967)
26. Meidner, D., Vexler, B.: Adaptive space-time finite element methods for parabolic optimization problems. SIAM J. Control Optim. 46(1), 116–142 (2007)
27. Monk, P.: Finite Element Methods for Maxwell’s Equations. Numerical Mathematics and Scientific Computation. Oxford University Press, New York (2003)
28. Stenberg, R.: On some techniques for approximating boundary conditions in the finite element method, vol. 63, pp. 139–148 (1995). International Symposium on Mathematical Modelling and Computational Methods Modelling 94 (Prague, 1994)
29. Tröltzsch, F.: Optimal control of partial differential equations, volume 112 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI (2010). Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.