Technische Universität Graz

Regularization and finite element error estimates for elliptic distributed optimal control problems with energy regularization and state or control constraints

P. Gangl, R. Löscher, O. Steinbach

Berichte aus dem
Institut für Angewandte Mathematik

Bericht 2023/4

Technische Universität Graz

Regularization and finite element error estimates for elliptic distributed optimal control problems with energy regularization and state or control constraints

P. Gangl, R. Löscher, O. Steinbach

Berichte aus dem Institut für Angewandte Mathematik

Bericht 2023/4

Technische Universität Graz
Institut für Angewandte Mathematik
Steyrergasse 30
A 8010 Graz

WWW: http://www.applied.math.tugraz.at

Regularization and finite element error estimates for elliptic distributed optimal control problems with energy regularization and state or control constraints

Peter Gangl[1], Richard Löscher[2], Olaf Steinbach[2]

[1]Johann Radon Institute for Computational and Applied Mathematics,
Altenberger Straße 69, 4040 Linz, Austria

[2]Institut für Angewandte Mathematik, TU Graz,
Steyrergasse 30, 8010 Graz, Austria

Abstract

In this paper we discuss the numerical solution of elliptic distributed optimal control problems with state or control constraints when the control is considered in the energy norm. As in the unconstrained case we can relate the regularization parameter and the finite element mesh size in order to ensure an optimal order of convergence which only depends on the regularity of the given target, also including discontinuous target functions. While in most cases, state or control constraints are discussed for the more common L^2 regularization, much less is known in the case of energy regularizations. But in this case, and for both control and state constraints, we can formulate first kind variational inequalities to determine the unknown state, from wich we can compute the control in a post processing step. Related variational inequalities also appear in obstacle problems, and are well established both from a mathematical and a numerical analysis point of view. Numerical results confirm the applicability and accuracy of the proposed approach.

1 Introduction

Optimal control problems aim to determine a control of a system that drives the corresponding state as closely as possible to a given desired state under acceptable costs for the control, see [37] for a thorough overview of the mathematical theory. Over the past decades, such problems have been studied for a wide variety of applications. A prominent example is the medical application of cancer treatment by hyperthermia [10] where a heat source should be placed in such a way that the temperature is increased (only) in the cancerous tissue. The control variable can, in general, be defined on the full domain or on

the boundary, and typically the application of the control comes at a certain cost which can be measured in different norms such as the L^2 norm or an energy norm. This cost is typically added to the objective functional with a certain weight and can also be seen as a regularization of the PDE-constrained optimization problem. Finite element error estimates of solutions to optimal control problems have been studied by many authors, see, e.g., [20], for an elliptic boundary control problem or [1, 8, 14, 31] for boundary control with energy regularization. More recently, time-dependent optimal control problems in the context of space-time finite element methods and corresponding error estimates were studied by some of the authors, see, e.g., [25, 26, 27] for parabolic problems or [29] for the wave equation. For many optimal control problems, it is important to pose pointwise constraints for either the state or the control variable, or both. These constraints can be incorporated in different ways, e.g., by augmented Lagrangian methods [21], barrier methods [33] or by reformulating the optimality system as a variational inequality [6, 13, 28]. This is closely related to the treatment of obstacle problems, where constraints can be handled using a penalization technique, see, e.g., [22]. After discretization, variational inequalities can be solved by means of primal-dual active set strategies [2] or semi-smooth Newton methods where the latter two strategies can be shown to be equivalent [16]. In particular the latter class has been used in different physical contexts such as elasticity [23], fluid mechanics [9], wave problems [24], and for different kinds of constraints including mixed control-state constraints [32] or constraints on derivatives of the state [17]. As in the unconstrained case, the control can be measured in different norms, depending on the regularity assumptions on the control, which leads to a different behavior of the solutions in particular in the case of less regular, i.e., discontinuous targets. The, nowaday, common L^2 regularization with state constraints was already studied in [11]. Considering the L^2 norm as energy norm leads to a fourth order elliptic partial differential equation, see [30]. The recent survey paper [3] gives an overview on the numerical analysis incorporating state constraints in this case. While in most cases, state or control constraints are discussed in the context of L^2 regularizations, much less is known in the case of energy regularizations. The very recent work [15] examines state constraints in the case of energy regularization for the Laplace equation, which is also a starting point for our analysis, see [36].

In this paper, we consider the problem to find the minimizer $(u_\varrho, z_\varrho) \in H_0^1(\Omega) \times H^{-1}(\Omega)$ of the functional

$$\mathcal{J}(u_\varrho, z_\varrho) := \frac{1}{2} \|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} + \frac{1}{2} \varrho \|z_\varrho\|^2_{H^{-1}(\Omega)} \tag{1.1}$$

subject to the Dirichlet boundary value problem of the Poisson equation,

$$-\Delta u_\varrho = z_\varrho \quad \text{in } \Omega, \quad u_\varrho = 0 \quad \text{on } \partial\Omega. \tag{1.2}$$

Here, $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$, is some bounded Lipschitz domain, $\overline{u} \in L^2(\Omega)$ is a given target, and $\varrho \in \mathbb{R}_+$ is some regularization parameter on which the minimizer depends on. The variational formulation of the Dirichlet boundary value problem (1.2) is to find $u_\varrho \in H_0^1(\Omega)$ such that

$$\langle \nabla u_\varrho, \nabla v \rangle_{L^2(\Omega)} = \langle z_\varrho, v \rangle_\Omega \quad \text{for all } v \in H_0^1(\Omega), \tag{1.3}$$

where $\langle z_\varrho, v\rangle_\Omega$ denotes the duality pairing for $z_\varrho \in H^{-1}(\Omega) = [H_0^1(\Omega)]^*$ and $v \in H_0^1(\Omega)$ as extension of the inner product in $L^2(\Omega)$. Recall that $\|\nabla v\|_{L^2(\Omega)}$ defines an equivalent norm for $v \in H_0^1(\Omega)$, and that the dual norm in $H^{-1}(\Omega)$ is given by

$$\|z\|_{H^{-1}(\Omega)} = \sup_{0\neq v\in H_0^1(\Omega)} \frac{\langle z, v\rangle_\Omega}{\|\nabla v\|_{L^2(\Omega)}} \quad \text{for all } z \in H^{-1}(\Omega).$$

With this we easily conclude

$$\|z_\varrho\|^2_{H^{-1}(\Omega)} = \|\nabla u_\varrho\|^2_{L^2(\Omega)} = \langle z_\varrho, u_\varrho\rangle_\Omega,$$

where $u_\varrho \in H_0^1(\Omega)$ is the unique solution of the variational formulation (1.3). Hence we can write the cost functional (1.1) as reduced cost functional

$$\widetilde{\mathcal{J}}(u_\varrho) = \frac{1}{2}\,\|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} + \frac{1}{2}\,\varrho\,\|\nabla u_\varrho\|^2_{L^2(\Omega)}. \tag{1.4}$$

In the case of neither state nor control constraints, the minimizer of the reduced cost functional (1.4) is given as the unique solution $u_\varrho \in H_0^1(\Omega)$ of the variational formulation

$$\varrho\,\langle\nabla u_\varrho, \nabla v\rangle_{L^2(\Omega)} + \langle u_\varrho, v\rangle_{L^2(\Omega)} = \langle\overline{u}, v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega). \tag{1.5}$$

Depending on the regularity of the target $\overline{u}$ we can prove the following regularization error estimates:

Lemma 1.1 ([30, Theorem 3.2]) *For $\varrho > 0$, let $u_\varrho \in H_0^1(\Omega)$ be the unique solution of the variational formulation (1.5). Assume $\overline{u} \in H_0^s(\Omega) := [L^2(\Omega), H_0^1(\Omega)]_s$ for some $s \in [0,1]$ or $\overline{u} \in H_0^1(\Omega) \cap H^s(\Omega)$ for some $s \in (1,2]$. Then there holds the regularization error estimate*

$$\|u_\varrho - \overline{u}\|_{L^2(\Omega)} \leq c\,\varrho^{s/2}\,\|\overline{u}\|_{H^s(\Omega)}. \tag{1.6}$$

Let $V_h \subset H_0^1(\Omega)$ be some finite element space of piecewise linear and continuous basis functions which are defined with respect to some admissible decomposition of the domain Ω into simplicial shape regular finite elements τ_ℓ of local mesh size h_ℓ, and with a global mesh size $h := \max_\ell h_\ell$. For simplicity we may assume that Ω is a polygonally ($n = 2$) or polyhedrally ($n = 3$) bounded domain. The finite element approximation of (1.5) is to find $u_{\varrho h} \in V_h$ such that

$$\varrho\,\langle\nabla u_{\varrho h}, \nabla v_h\rangle_{L^2(\Omega)} + \langle u_{\varrho h}, v_h\rangle_{L^2(\Omega)} = \langle\overline{u}, v_h\rangle_{L^2(\Omega)} \quad \text{for all } v_h \in V_h. \tag{1.7}$$

The numerical analysis of this variational formulation as well as the construction of robust iterative solution methods was considered in [27].

Lemma 1.2 ([27, Theorem 1]) *Let us assume, for simplicity, that $\Omega \subset \mathbb{R}^n$ is convex, and that the target function satisfies either $\overline{u} \in H_0^s(\Omega)$ for $s \in [0,1]$ or $\overline{u} \in H_0^1(\Omega) \cap H^s(\Omega)$ for $s \in (1,2]$. For the choice $\varrho = h^2$ there holds the error estimate*

$$\|u_{\varrho h} - \overline{u}\|_{L^2(\Omega)} \leq c\,h^s\,\|\overline{u}\|_{H^s(\Omega)}. \tag{1.8}$$

The aim of this paper is to provide related estimates for both the regularization and the finite element error when considering the minimization of (1.4) with either state or control constraints. Note that related work, considering state constraints, but not with respect to the regularization parameter ϱ, was recently presented in [15].

The remainder of this paper is structured as follows: In Section 2 we provide estimates for the error $\|u_\varrho - \overline{u}\|_{L^2(\Omega)}$ with respect to the regularization parameter ϱ for both state and control constraints. These results follow similar as in the unconstrained case, due to the structure of the variational inequality to be solved. In order to include state or control constraints we have to solve a first kind variational inequality to find the unknown state. Their numerical approximation using finite element methods is well established, and we can transfer the general results to the particular application of constrained optimal control problems with energy regularization. The finite element discretization and the related a priori error estimates are given in Section 3. In a post processing step we finally compute the control, when the state is known. The resulting discrete variational inequality can be reformulated as a nonlinear system of algebraic equations, which can be solved by applying a semi-smooth Newton method which turns out to be an active set strategy, and which is described in Section 4. Several numerical results are given in Section 5 in order to demonstrate the applicability and accuracy of the proposed approach. Finally, we summarize and comment on ongoing work.

2 Regularization error estimates

In this section we will discuss regularization error estimates for the minimization of (1.4) with additional constraints on either the state u_ϱ or the control z_ϱ.

2.1 State constraints

We consider the reduced functional $\widetilde{\mathcal{J}}(u_\varrho)$ as defined in (1.4), but now we minimize over the convex subset

$$K_s := \Big\{ v \in H_0^1(\Omega) : g_-(x) \leq v(x) \leq g_+(x) \text{ for almost all } x \in \Omega \Big\},$$

where $g_\pm \in H^1_\Delta(\Omega) := \{v \in H_0^1(\Omega) : \Delta v \in L^2(\Omega)\}$ are given barrier functions, and where we assume $g_- < g_+$ and $0 \in K_s$ to be satisfied. From this it follows that $g_- \leq 0$, and $g_+ \geq 0$. The minimizer $u_\varrho \in K_s$ satisfying

$$\widetilde{\mathcal{J}}(u_\varrho) = \min_{v \in K_s} \widetilde{\mathcal{J}}(v)$$

is determined as the unique solution $u_\varrho \in K_s$ of the first kind variational inequality

$$\varrho \, \langle \nabla u_\varrho, \nabla(v - u_\varrho) \rangle_{L^2(\Omega)} + \langle u_\varrho, v - u_\varrho \rangle_{L^2(\Omega)} \geq \langle \overline{u}, v - u_\varrho \rangle_{L^2(\Omega)} \quad \text{for all } v \in K_s. \tag{2.1}$$

As in the unconstrained case [30] we can prove the following regularization error estimates.

Lemma 2.1 *For $\varrho > 0$, let $u_\varrho \in K_s$ be the unique solution of the variational inequality (2.1). For $\overline{u} \in L^2(\Omega)$ there holds*

$$\|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le \|\overline{u}\|_{L^2(\Omega)}, \tag{2.2}$$

while for $\overline{u} \in K_s$ we have

$$\|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le \sqrt{\varrho}\, \|\nabla \overline{u}\|_{L^2(\Omega)}, \tag{2.3}$$

and

$$\|\nabla(u_\varrho - \overline{u})\|_{L^2(\Omega)} \le \|\nabla \overline{u}\|_{L^2(\Omega)}. \tag{2.4}$$

If in addition $\Delta \overline{u} \in L^2(\Omega)$ is satisfied for $\overline{u} \in K_s$,

$$\|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le \varrho\, \|\Delta \overline{u}\|_{L^2(\Omega)} \tag{2.5}$$

as well as

$$\|\nabla(u_\varrho - \overline{u})\|_{L^2(\Omega)} \le \sqrt{\varrho}\, \|\Delta \overline{u}\|_{L^2(\Omega)} \tag{2.6}$$

follow.

Proof. From the variational inequality (2.1) we obviously have

$$\varrho\, \langle \nabla u_\varrho, \nabla(v - u_\varrho)\rangle_{L^2(\Omega)} \ge \langle \overline{u} - u_\varrho, v - u_\varrho\rangle_{L^2(\Omega)} \quad \text{for all } v \in K_s.$$

In particular for $v = 0 \in K_s$ this gives

$$\|\overline{u} - u_\varrho\|^2_{L^2(\Omega)} + \varrho\, \langle \nabla u_\varrho, \nabla u_\varrho\rangle_{L^2(\Omega)} \le \langle u_\varrho - \overline{u}, \overline{u}\rangle_{L^2(\Omega)} \le \|\overline{u} - u_\varrho\|_{L^2(\Omega)} \|\overline{u}\|_{L^2(\Omega)},$$

i.e., (2.2) follows. When assuming $\overline{u} \in K_s$ we can consider $v = \overline{u}$ to obtain

$$\begin{aligned} \varrho\, \|\nabla(u_\varrho - \overline{u})\|^2_{L^2(\Omega)} + \|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} &\le \varrho\, \langle \nabla \overline{u}, \nabla(\overline{u} - u_\varrho)\rangle_{L^2(\Omega)} \\ &\le \varrho\, \|\nabla \overline{u}\|_{L^2(\Omega)} \|\nabla(u_\varrho - \overline{u})\|_{L^2(\Omega)}, \end{aligned}$$

i.e.,

$$\|\nabla(u_\varrho - \overline{u})\|_{L^2(\Omega)} \le \|\nabla \overline{u}\|_{L^2(\Omega)},$$

that is (2.4), and (2.3),

$$\|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} \le \varrho\, \|\nabla \overline{u}\|^2_{L^2(\Omega)}.$$

If $\overline{u} \in K_s$ is such that $\Delta \overline{u} \in L^2(\Omega)$ is satisfied, then we conclude

$$\begin{aligned} \varrho\, \|\nabla(u_\varrho - \overline{u})\|^2_{L^2(\Omega)} + \|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} &\le \varrho\, \langle \nabla \overline{u}, \nabla(\overline{u} - u_\varrho)\rangle_{L^2(\Omega)} \\ &= \varrho\, \langle (-\Delta \overline{u}), \overline{u} - u_\varrho\rangle_{L^2(\Omega)} \le \varrho\, \|\Delta \overline{u}\|_{L^2(\Omega)} \|u_\varrho - \overline{u}\|_{L^2(\Omega)}, \end{aligned}$$

and hence

$$\|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le \varrho\, \|\Delta \overline{u}\|_{L^2(\Omega)},$$

i.e., (2.5), follows. Finally,

$$\varrho\, \|\nabla(u_\varrho - \overline{u})\|^2_{L^2(\Omega)} \le \varrho\, \|\Delta \overline{u}\|_{L^2(\Omega)} \|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le \varrho^2\, \|\Delta \overline{u}\|^2_{L^2(\Omega)}$$

implies (2.6). □

For the solution u_ϱ of (2.1) we introduce the active sets $\Omega_{s,\pm} := \{x \in \Omega : u_\varrho(x) = g_\pm(x)\}$.

Lemma 2.2 *For $u_\varrho \in K_s$ being the unique solution of the variational inequality* (2.1)*, let* $\lambda := -\varrho\Delta u_\varrho + u_\varrho - \overline{u} \in H^{-1}(\Omega)$*. Assume* $\overline{u} \in K_s \cap H^1_\Delta(\Omega)$ *and* $g_\pm \in H^1_\Delta(\Omega)$*. Then we conclude* $z_\varrho = -\Delta u_\varrho \in L^2(\Omega)$ *and hence,* $\lambda \in L^2(\Omega)$*.*

Proof. When using integration by parts we can write (2.1) as

$$\langle -\varrho\Delta u_\varrho + u_\varrho - \overline{u}, v - u_\varrho\rangle_\Omega \geq 0 \quad \text{for all } v \in K_s,$$

i.e.,

$$\langle \lambda, v - u_\varrho\rangle_\Omega \geq 0 \quad \text{for all } v \in K_s. \tag{2.7}$$

The definition of λ implies

$$\lambda + \varrho\Delta u_\varrho = u_\varrho - \overline{u} \in L^2(\Omega).$$

For $x \in \Omega_\pm$ we have $u_\varrho(x) = g_\pm(x)$, and hence $\Delta u_\varrho = \Delta g_\pm \in L^2(\Omega_{s,\pm})$ as well as $\lambda \in L^2(\Omega_{s,\pm})$. Let $w \in H^1_0(\Omega)$ satisfying

$$0 \leq w(x) \leq \min\Big\{g_+(x) - u_\varrho(x), u_\varrho(x) - g_-(x)\Big\} \quad \text{for } x \in \Omega,$$

i.e., $w(x) = 0$ for $x \in \Omega_{s,\pm}$. For $v = u_\varrho + w \in K_s$ we then obtain from (2.7) $\langle \lambda, w\rangle_\Omega \geq 0$, while for $v = u_\varrho - w \in K_s$ we conclude $\langle \lambda, w\rangle_\Omega \leq 0$. Hence we have $\langle \lambda, w\rangle_\Omega = 0$ for all $w \in H^1_0(\Omega\backslash\Omega_{s,\pm})$, i.e., $\lambda = 0$ in $H^{-1}(\Omega\backslash\Omega_{s,\pm})$, which remains true in $L^2(\Omega\backslash\Omega_{s,\pm})$. This already gives $\lambda \in L^2(\Omega)$. Moreover, by $0 = \lambda = -\varrho\Delta u_\varrho + u_\varrho - \overline{u}$ in $\Omega \setminus \Omega_{s,\pm}$ and Lemma 2.1 we obtain

$$\|\varrho\Delta u_\varrho\|_{L^2(\Omega\backslash\Omega_{s,\pm})} = \|u_\varrho - \overline{u}\|_{L^2(\Omega\backslash\Omega_{s,\pm})} \leq \|u_\varrho - \overline{u}\|_{L^2(\Omega)} \leq \varrho\,\|\Delta\overline{u}\|_{L^2(\Omega)},$$

which implies

$$\|\Delta u_\varrho\|_{L^2(\Omega\backslash\Omega_\pm)} \leq \|\Delta\overline{u}\|_{L^2(\Omega)},$$

i.e., $\Delta u_\varrho \in L^2(\Omega \setminus \Omega_\pm)$ and together with $\Delta u_\varrho = \Delta g_\pm$ in $\Omega_{s,\pm}$, $\Delta u_\varrho \in L^2(\Omega)$, independent of ϱ. □

Due to $\lambda \in L^2(\Omega)$ we can write (2.7) as

$$\int_\Omega \lambda(x)\,[v(x) - u_\varrho(x)]\,dx \geq 0 \quad \text{for all } v \in K_s.$$

For arbitrary $w_+ \in H^1_0(\Omega)$ satisfying $0 \leq w_+(x) \leq g_+(x) - u_\varrho(x)$ for almost all $x \in \Omega$ we have $v = u_\varrho + w_+ \in K_s$, and we conclude

$$\int_{\Omega\backslash\Omega_+} \lambda(x)\,w_+(x)\,ds_x \geq 0 \quad \text{for all } w_+ \in H^1_0(\Omega\backslash\Omega_{s,+}), \quad w_+ \geq 0.$$

Hence we obtain $\lambda(x) \geq 0$ for almost all $x \in \Omega\backslash\Omega_{s,+}$. In the same way we choose $w_- \in H^1_0(\Omega)$ satisfying $g_-(x) - u_\varrho(x) \leq w_-(x) \leq 0$ for almost all $x \in \Omega$. Hence we have $v = u_\varrho + w_- \in K_s$, and we conclude

$$\int_{\Omega\backslash\Omega_-} \lambda(x)\,w_-(x)\,dx \geq 0 \quad \text{for all } w_- \in H^1_0(\Omega\backslash\Omega_{s,-}), \quad w_- \leq 0,$$

i.e., $\lambda(x) \le 0$ for almost all $x \in \Omega \backslash \Omega_{s,-}$. With this we finally obtain the complementarity conditions which hold for almost all $x \in \Omega$:

$$g_-(x) < u_\varrho(x) < g_+(x) \ : \ \lambda(x) = 0; \ u_\varrho(x) = g_-(x) \ : \ \lambda(x) \ge 0; \ u_\varrho(x) = g_+(x) \ : \ \lambda(x) \le 0.$$

Remark 2.1 *The variational inequality* (2.1) *corresponds to the two obstacle problem as considered, e.g., in* [5]*, where also a more general discussion on the regularity of solutions is given, i.e.,* [5, Theorémè I.1, Remarque I.4, Remarque I.5]*, which also fits our application.*

2.2 Control constraints

Since $-\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)$ defines an isomorphism, we can also write $u_\varrho = S z_\varrho$, where $S : H^{-1}(\Omega) \to H_0^1(\Omega)$ is the solution operator of the Dirichlet boundary value problem (1.2). Instead of (1.1) and (1.4) we now consider the reduced cost functional

$$\widehat{\mathcal{J}}(z_\varrho) = \frac{1}{2} \, \| S z_\varrho - \overline{u} \|^2_{L^2(\Omega)} + \frac{1}{2} \, \varrho \, \langle S z_\varrho, z_\varrho \rangle_\Omega \quad \text{for } z_\varrho \in H^{-1}(\Omega). \tag{2.8}$$

Box constraints in $H^{-1}(\Omega)$ are defined in weak form, i.e., for given $f_\pm \in L^2(\Omega)$ we define

$$Z_c := \Big\{ z \in H^{-1}(\Omega) : \langle f_-, v \rangle_{L^2(\Omega)} \le \langle z_\varrho, v \rangle_\Omega \le \langle f_+, v \rangle_{L^2(\Omega)} \, \forall \, v \in H_0^1(\Omega), \ v(x) \ge 0 \Big\}. \tag{2.9}$$

Hence we find the minimizer $z_\varrho \in Z_c$ of the reduced cost functional (2.8) as the unique solution of the variational inequality

$$\langle S^* S z_\varrho + \varrho \, S z_\varrho, \psi - z_\varrho \rangle_\Omega \ge \langle S^* \overline{u}, \psi - z_\varrho \rangle_\Omega \quad \text{for all } \psi \in Z_c. \tag{2.10}$$

When using $u_\varrho = S z_\varrho$ and the fact that S is self-adjoint, this can be written as

$$\langle u_\varrho + \varrho \, z_\varrho, v - u_\varrho \rangle_\Omega \ge \langle \overline{u}, v - u_\varrho \rangle_\Omega \quad \text{for all } v = S\psi, \ \psi \in Z_c.$$

When introducing

$$K_c := \Big\{ u \in H_0^1(\Omega) : \langle f_-, v \rangle_{L^2(\Omega)} \le \langle \nabla u, \nabla v \rangle_{L^2(\Omega)} \le \langle f_+, v \rangle_{L^2(\Omega)} \, \forall \, v \in H_0^1(\Omega), \, v(x) \ge 0 \Big\}, \tag{2.11}$$

and using $z_\varrho = -\Delta u_\varrho$, we finally end up with a variational inequality to find $u_\varrho \in K_c$ such that

$$\langle u_\varrho, v - u_\varrho \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla u_\varrho, \nabla (v - u_\varrho) \rangle_{L^2(\Omega)} \ge \langle \overline{u}, v - u_\varrho \rangle_{L^2(\Omega)} \quad \text{for all } v \in K_c. \tag{2.12}$$

Since the variational inequality (2.12) coincides with (2.1), all the regularization error estimates as given in Lemma 2.1 remain valid, but we have to assume $\overline{u} \in K_c$ instead of $\overline{u} \in K_s$, when required.

For the unique solution $u_\varrho \in K_c$ and for the target $\overline{u} \in L^2(\Omega)$ we define $w \in H_0^1(\Omega)$ as the unique weak solution of the Dirichlet boundary value problem

$$-\Delta w = -\varrho \Delta u_\varrho + u_\varrho - \overline{u} \quad \text{in } \Omega, \quad w = 0 \quad \text{on } \partial\Omega, \tag{2.13}$$

satisfying

$$\langle \nabla w, \nabla v\rangle_{L^2(\Omega)} = \varrho\,\langle \nabla u_\varrho, \nabla v\rangle_{L^2(\Omega)} + \langle u_\varrho - \overline{u}, v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

When using integration by parts we can write this as

$$\langle -\Delta w + \varrho \Delta u_\varrho, v\rangle_\Omega = \langle u_\varrho - \overline{u}, v\rangle_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

For $f_\pm \in L^2(\Omega)$ we define $g_{c,\pm} \in H_0^1(\Omega)$ as unique solutions of the Dirichlet boundary value problems

$$-\Delta g_{c,\pm} = f_\pm \quad \text{in } \Omega, \quad g_{c,\pm} = 0 \quad \text{on } \partial\Omega,$$

and we introduce $\Omega_{c,\pm} := \{x \in \Omega : u_\varrho(x) = g_{c,\pm}(x)\}$. Due to $f_\pm \in L^2(\Omega)$ we therefore have $f_\pm(x) = -\Delta g_{c,\pm}(x) = -\Delta u_\varrho(x)$ for almost all $x \in \Omega_{c,\pm}$.

Lemma 2.3 *For $u_\varrho \in K_c$ being the unique solution of the variational inequality* (2.12)*, let $w \in H_0^1(\Omega)$ be the weak solution of the Dirichlet boundary value problem* (2.13)*. Assume $\overline{u} \in K_c \cap H^1_\Delta(\Omega)$ and $f_\pm \in L^2(\Omega)$. Then we conclude $z_\varrho = -\Delta u_\varrho \in L^2(\Omega)$, and hence, $\Delta w \in L^2(\Omega)$.*

Proof. The definition of $w \in H_0^1(\Omega)$ as weak solution of the Poisson equation in (2.13) implies

$$-\Delta w(x) = \varrho\, f_\pm(x) + u_\varrho(x) - \overline{u}(x) \quad \text{for almost all } x \in \Omega_{c,\pm},$$

and hence, $\Delta w \in L^2(\Omega_{c,\pm})$ follows. Since $u_\varrho \in K_c$ is the unique solution of the variational inequality (2.12), and using the definition of $w \in H_0^1(\Omega)$, this gives

$$\langle \nabla w, \nabla(v - u_\varrho)\rangle_{L^2(\Omega)} \geq 0 \quad \text{for all } v \in K_c,$$

or equivalently,

$$\langle w, -\Delta v + \Delta u_\varrho\rangle_\Omega \geq 0 \quad \text{for all } v \in K_c.$$

Let $v_+ \in H_0^1(\Omega)$ be the unique solution of the Dirichlet boundary value problem

$$-\Delta v_+ = -\Delta u_\varrho + \psi \quad \text{in } \Omega, \quad v_+ = 0 \quad \text{on } \partial\Omega,$$

where $\psi \in L^2(\Omega)$, $\psi(x) \geq 0$ for almost all $x \in \Omega$, is given. To ensure $v_+ \in K_c$ we need to assume

$$\langle \psi, v\rangle_{L^2(\Omega)} \leq \langle f_+ + \Delta u_\varrho, v\rangle_\Omega \quad \text{for all } v \in H_0^1(\Omega),\ v \geq 0.$$

From this we conclude, when considering $v \in H_0^1(\Omega_{c,+})$, $\psi(x) = 0$ for almost all $x \in \Omega_{c,+}$, and

$$\langle w, \psi\rangle_{L^2(\Omega)} = \langle w, \psi\rangle_{L^2(\Omega\setminus\Omega_{c,+})} \geq 0\,. \tag{2.14}$$

Next, and using the same ψ as above, let $v_- \in H_0^1(\Omega)$ be the unique solution of the Dirichlet boundary value problem

$$-\Delta v_- = -\Delta u_\varrho - \psi \quad \text{in } \Omega, \quad v_- = 0 \quad \text{on } \partial\Omega.$$

To ensure $v_- \in K_c$ we now have to satisfy

$$\langle \psi, v\rangle_\Omega \le -\langle f_- + \Delta u_\varrho, v\rangle_\Omega \quad \text{for all } v \in H_0^1(\Omega),\ v \ge 0.$$

This gives $\psi(x) = 0$ for almost all $x \in \Omega_{c,-}$, and we have $\langle w, \psi\rangle_{L^2(\Omega\setminus\Omega_{c,-})} \le 0$. Hence we conclude $\langle w, \psi\rangle_{L^2(\Omega\setminus\Omega_{c,\pm})} = 0$ for all $\psi \in L^2(\Omega \setminus \Omega_{c,\pm})$ satisfying

$$\langle \psi, v\rangle_\Omega \le \min\Big\{\langle f_+ + \Delta u_\varrho, v\rangle_\Omega, -\langle f_- + \Delta u_\varrho, v\rangle_\Omega\Big\} \quad \text{for all } v \in H_0^1(\Omega),\ v \ge 0,$$

and therefore $w(x) = 0$ for almost all $x \in \Omega\setminus\Omega_{c,\pm}$ follows. Using Lemma 2.1, this implies

$$\|\varrho\,\Delta u_\varrho\|_{L^2(\Omega\setminus\Omega_{c,\pm})} = \|u_\varrho - \overline{u}\|_{L^2(\Omega\setminus\Omega_{c,\pm})} \le \|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le \varrho\,\|\Delta\overline{u}\|_{L^2(\Omega)},$$

i.e.,

$$\|\Delta u_\varrho\|_{L^2(\Omega\setminus\Omega_{c,\pm})} \le \|\Delta\overline{u}\|_{L^2(\Omega)}.$$

Together with $-\Delta u_\varrho = f_\pm \in L^2(\Omega_{c,\pm})$ this gives $-\Delta u_\varrho \in L^2(\Omega)$, and $-\Delta w \in L^2(\Omega)$. □

From the proof of Lemma 2.3 we already have $f_-(x) < -\Delta u_\varrho(x) < f_+(x)$ and $w(x) = 0$ for $x \in \Omega\setminus\Omega_{c,\pm}$. Moreover, (2.14) gives $\langle w, \psi\rangle_{L^2(\Omega_{c,-})} \ge 0$ for all $\psi \in L^2(\Omega_{c,-})$ with $\psi \ge 0$, and hence we obtain $-\Delta u_\varrho(x) = f_-(x)$ and $w(x) \ge 0$ for $x \in \Omega_{c,-}$. In the same way we conclude $-\Delta u_\varrho(x) = f_+(x)$ and $w(x) \le 0$ for $x \in \Omega_{c,+}$. Note that these relations are the complimentarity conditions of the variational inequality (2.12).

3 Finite element discretization

Let us consider the variational inequality to find $u_\varrho \in K$ such that

$$\varrho\,\langle \nabla u_\varrho, \nabla(v - u_\varrho)\rangle_{L^2(\Omega)} + \langle u_\varrho, v - u_\varrho\rangle_{L^2(\Omega)} \ge \langle \overline{u}, v - u_\varrho\rangle_{L^2(\Omega)} \quad \text{for all } v \in K, \tag{3.1}$$

which corresponds to (2.1) with $K = K_s$ in the case of state constraints, and to (2.12) with $K = K_c$ for control constraints. We now assume that Ω is either convex or sufficiently regular such that $\|\Delta u\|_{L^2(\Omega)}$ defines an equivalent norm in $H_0^1(\Omega) \cap H^2(\Omega) = H_\Delta^1(\Omega)$.

As in the unconstrained case, let $V_h = S_h^1(\Omega) \cap H_0^1(\Omega) = \mathrm{span}\{\varphi_k\}_{k=1}^M$ be a conforming finite element space, e.g., of piecewise linear and continuous basis functions φ_k which are defined with respect to some admissible decomposition of Ω into simplicial shape regular finite elements τ_ℓ of local mesh size h_ℓ.

Let $K_h \subset V_h$ be some appropriate approximation of K to be specified later. Then we consider the Galerkin variational inequality of (3.1) to find $u_{\varrho h} \in K_h$ such that

$$\varrho\,\langle \nabla u_{\varrho h}, \nabla(v_h - u_{\varrho h})\rangle_{L^2(\Omega)} + \langle u_{\varrho h}, v_h - u_{\varrho h}\rangle_{L^2(\Omega)} \ge \langle \overline{u}, v_h - u_{\varrho h}\rangle_{L^2(\Omega)} \tag{3.2}$$

is satisfied for all $v_h \in K_h$, which is obviously equivalent to

$$\langle \overline{u} - u_{\varrho h}, v_h - u_{\varrho h}\rangle_{L^2(\Omega)} - \varrho\,\langle \nabla u_{\varrho h}, \nabla(v_h - u_{\varrho h})\rangle_{L^2(\Omega)} \le 0 \quad \text{for all } v_h \in K_h. \tag{3.3}$$

Following [12] we can prove the following a priori error estimate for the solution $u_{\varrho h} \in K_h$ of the variational inequality (3.3).

Lemma 3.1 *For $u_\varrho \in K$ and $u_{\varrho h} \in K_h$ being the unique solutions of the variational inequalities* (3.1) *and* (3.2)*, respectively, there holds the error estimate*

$$\begin{aligned} &\|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \varrho \, \|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)} \qquad (3.4)\\ &\leq 8 \Big[\inf_{v_h \in K_h} \Big(\|u_\varrho - v_h\|^2_{L^2(\Omega)} + \varrho \, \|\nabla(u_\varrho - v_h)\|^2_{L^2(\Omega)} \Big) + \varrho^2 \, \|\Delta u_\varrho\|^2_{L^2(\Omega)} + \|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} \Big]. \end{aligned}$$

Proof. For arbitrary $v_h \in K_h$, using (3.3) and integration by parts, we can write

$$\begin{aligned} &\|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \varrho \, \|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)}\\ &\quad = \langle u_\varrho - u_{\varrho h}, u_\varrho - u_{\varrho h} \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla(u_\varrho - u_{\varrho h}), \nabla(u_\varrho - u_{\varrho h}) \rangle_{L^2(\Omega)}\\ &\quad = \langle u_\varrho - u_{\varrho h}, u_\varrho - v_h \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla(u_\varrho - u_{\varrho h}), \nabla(u_\varrho - v_h) \rangle_{L^2(\Omega)}\\ &\qquad + \langle u_\varrho - u_{\varrho h}, v_h - u_{\varrho h} \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla(u_\varrho - u_{\varrho h}), \nabla(v_h - u_{\varrho h}) \rangle_{L^2(\Omega)}\\ &\quad = \langle u_\varrho - u_{\varrho h}, u_\varrho - v_h \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla(u_\varrho - u_{\varrho h}), \nabla(u_\varrho - v_h) \rangle_{L^2(\Omega)}\\ &\qquad + \langle u_\varrho - \overline{u}, v_h - u_{\varrho h} \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla u_\varrho, \nabla(v_h - u_{\varrho h}) \rangle_{L^2(\Omega)}\\ &\qquad\quad + \langle \overline{u} - u_{\varrho h}, v_h - u_{\varrho h} \rangle_{L^2(\Omega)} - \varrho \, \langle \nabla u_{\varrho h}, \nabla(v_h - u_{\varrho h}) \rangle_{L^2(\Omega)}\\ &\quad \leq \langle u_\varrho - u_{\varrho h}, u_\varrho - v_h \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla(u_\varrho - u_{\varrho h}), \nabla(u_\varrho - v_h) \rangle_{L^2(\Omega)}\\ &\qquad + \langle u_\varrho - \overline{u}, v_h - u_{\varrho h} \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla u_\varrho, \nabla(v_h - u_{\varrho h}) \rangle_{L^2(\Omega)}\\ &\quad = \langle u_\varrho - u_{\varrho h}, u_\varrho - v_h \rangle_{L^2(\Omega)} + \varrho \, \langle \nabla(u_\varrho - u_{\varrho h}), \nabla(u_\varrho - v_h) \rangle_{L^2(\Omega)}\\ &\qquad + \langle -\varrho \Delta u_\varrho + u_\varrho - \overline{u}, v_h - u_{\varrho h} \rangle_{L^2(\Omega)}\\ &\quad \leq \|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)} \|u_\varrho - v_h\|_{L^2(\Omega)} + \varrho \, \|\nabla(u_\varrho - u_{\varrho h})\|_{L^2(\Omega)} \|\nabla(u_\varrho - v_h)\|_{L^2(\Omega)}\\ &\qquad + \| - \varrho \Delta u_\varrho + u_\varrho - \overline{u}\|_{L^2(\Omega)} \|v_h - u_{\varrho h}\|_{L^2(\Omega)} \, . \end{aligned}$$

When using Young's inequality we further have

$$\begin{aligned} \|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \varrho \, \|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)} \;\leq\; & \frac{1}{4} \, \|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \|u_\varrho - v_h\|^2_{L^2(\Omega)}\\ & + \frac{1}{2} \varrho \, \|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)} + \frac{1}{2} \varrho \|\nabla(u_\varrho - v_h)\|^2_{L^2(\Omega)}\\ & + \| - \varrho \Delta u_\varrho + u_\varrho - \overline{u}\|^2_{L^2(\Omega)} + \frac{1}{4} \, \|v_h - u_{\varrho h}\|^2_{L^2(\Omega)} \, , \end{aligned}$$

i.e.,

$$\begin{aligned} \frac{3}{4} \, \|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \frac{1}{2} \, \varrho \, \|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)} \;\leq\; & \|u_\varrho - v_h\|^2_{L^2(\Omega)} + \frac{1}{2} \, \varrho \, \|\nabla(u_\varrho - v_h)\|^2_{L^2(\Omega)}\\ & + \Big(\varrho \, \|\Delta u_\varrho\|_{L^2(\Omega)} + \|u_\varrho - \overline{u}\|_{L^2(\Omega)} \Big)^2 + \frac{1}{4} \, \Big(\|v_h - u_\varrho\|_{L^2(\Omega)} + \|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)} \Big)^2\\ \leq\; & \|u_\varrho - v_h\|^2_{L^2(\Omega)} + \frac{1}{2} \, \varrho \, \|\nabla(u_\varrho - v_h)\|^2_{L^2(\Omega)}\\ & + 2 \, \varrho^2 \, \|\Delta u_\varrho\|^2_{L^2(\Omega)} + 2 \, \|u_\varrho - \overline{u}\|^2_{L^2(\Omega)} + \frac{1}{2} \, \|v_h - u_\varrho\|^2_{L^2(\Omega)} + \frac{1}{2} \, \|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)}, \end{aligned}$$

and hence,

$$\frac{1}{4}\,\|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \frac{1}{2}\,\varrho\,\|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)}$$
$$\leq\ \frac{3}{2}\,\|u_\varrho - v_h\|^2_{L^2(\Omega)} + \frac{1}{2}\,\varrho\,\|\nabla(u_\varrho - v_h)\|^2_{L^2(\Omega)} + 2\,\varrho^2\,\|\Delta u_\varrho\|^2_{L^2(\Omega)} + 2\,\|u_\varrho - \overline{u}\|^2_{L^2(\Omega)},$$

and the assumption follows. □

3.1 State constraints

Let $I_h : C(\overline{\Omega}) \to S^1_h(\Omega)$ be the nodal interpolation operator. When assuming $g_\pm \in H^1_\Delta(\Omega) = H^1_0(\Omega) \cap H^2(\Omega)$ we then define

$$K_{sh} := \Big\{v_h \in V_h : I_h g_- \leq v_h \leq I_h g_+ \text{ in } \Omega\Big\},$$

and we consider the variational inequality (3.2) for $K_h = K_{sh}$.

Theorem 3.2 *Let $u_\varrho \in K_s$ and $u_{\varrho h} \in K_{sh}$ be the unique solutions of the variational inequalities (2.1) and (3.2), respectively. Assume $\overline{u} \in K_s \cap H^1_\Delta(\Omega)$ and $g_\pm \in H^1_\Delta(\Omega)$. When choosing $\varrho = h^2$, then there holds the error estimate*

$$\|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + h^2\,\|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)} \leq c\,h^4\,\Big[\|\Delta\overline{u}\|^2_{L^2(\Omega)} + \|\Delta g_\pm\|^2_{L^2(\Omega)}\Big]. \tag{3.5}$$

Proof. Due to Lemma 2.2 we have

$$\|\Delta u_\varrho\|^2_{L^2(\Omega)} = \|\Delta g_\pm\|^2_{L^2(\Omega_\pm)} + \|\Delta u_\varrho\|^2_{L^2(\Omega\setminus\Omega_\pm)},$$

and since $\|\Delta u\|_{L^2(\Omega)}$ defines an equivalent norm in $H^1_0(\Omega) \cap H^2(\Omega)$,

$$|u_\varrho|^2_{H^2(\Omega)} \leq c\,\|\Delta u_\varrho\|^2_{L^2(\Omega)} \leq c\,\Big[\|\Delta\overline{u}\|_{L^2(\Omega)} + \|\Delta g_\pm\|^2_{L^2(\Omega)}\Big]$$

follows. Hence we can consider the nodal interpolation $I_h u_\varrho \in K_{sh}$ and we can use standard interpolation error estimates to conclude

$$\|u_\varrho - I_h u_\varrho\|^2_{L^2(\Omega)} + \varrho\,\|\nabla(u_\varrho - I_h u_\varrho)\|^2_{L^2(\Omega)} \leq c\,\Big(h^4 + \varrho\,h^2\Big)\,|u_\varrho|^2_{H^2(\Omega)}.$$

With this and using (2.5) we can write the general error estimate (3.4) as

$$\|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + \varrho\,\|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)}$$
$$\leq\ 8\,\Big[c\,\Big(h^4 + \varrho\,h^2\Big)\,|u_\varrho|^2_{H^2(\Omega)} + 2\,\varrho^2\,\Big(\|\Delta\overline{u}\|^2_{L^2(\Omega)} + \|\Delta g_\pm\|^2_{L^2(\Omega)}\Big)\Big].$$

The assertion finally follows when choosing $\varrho = h^2$. □

Corollary 3.3 *Assume $\overline{u} \in K_s \cap H^r(\Omega)$ for some $r \in (1,2]$, or $\overline{u} \in H_0^r(\Omega)$ for some $r \in [0,1]$. In the latter case we also assume $g_-(x) \le \overline{u}(x) \le g_+(x)$ for almost all $x \in \Omega$, where $g_\pm \in H^r(\Omega)$, $r \in [0,2]$. Then there holds the error estimate*

$$\|u_{\varrho h} - \overline{u}\|_{L^2(\Omega)} \le c\, h^r \sqrt{\|\overline{u}\|^2_{H^r(\Omega)} + \|g_\pm\|^2_{H^r(\Omega)}}. \tag{3.6}$$

Proof. When considering the variational inequality (3.2) for $v_h = 0$, this gives

$$\varrho\, \langle \nabla u_{\varrho h}, \nabla u_{\varrho h} \rangle_{L^2(\Omega)} + \langle u_{\varrho h} - \overline{u}, u_{\varrho h} - \overline{u} \rangle_{L^2(\Omega)} \le \langle \overline{u} - u_{\varrho h}, \overline{u} \rangle_{L^2(\Omega)},$$

from which we derive the trivial error estimate

$$\|u_{\varrho h} - \overline{u}\|_{L^2(\Omega)} \le \|\overline{u}\|_{L^2(\Omega)} \le \sqrt{\|\overline{u}\|^2_{L^2(\Omega)} + \|g_\pm\|^2_{L^2(\Omega)}}.$$

On the other hand, (3.5) and (2.5) imply, recall $\varrho = h^2$,

$$\|u_{\varrho h} - \overline{u}\|_{L^2(\Omega)} \le \|u_{\varrho h} - u_\varrho\|_{L^2(\Omega)} + \|u_\varrho - \overline{u}\|_{L^2(\Omega)} \le c\, h^2 \sqrt{\|\overline{u}\|^2_{H^2(\Omega)} + \|g_\pm\|^2_{H^2(\Omega)}}.$$

Now the assertion follows from a space interpolation argument. □

Using the isomorphism $v_h \in K_{sh} \leftrightarrow \underline{v} \in \mathbb{R}^M$ we can write the variational inequality (3.2) as: Find $\underline{u} \in \mathbb{R}^M \leftrightarrow u_{\varrho h} \in K_{sh}$ such that

$$\varrho\, (K_h \underline{u}, \underline{v} - \underline{u}) + (M_h \underline{u}, \underline{v} - \underline{u}) \ge (\underline{\overline{u}}, \underline{v} - \underline{u}) \tag{3.7}$$

is satisfied for all $\underline{v} \in \mathbb{R}^M \leftrightarrow v_h \in K_{sh}$. Here, M_h and K_h are the standard finite element mass and stiffness matrices, defined by

$$K_h[j,k] = \int_\Omega \nabla \varphi_k(x) \cdot \nabla \varphi_j(x)\, dx, \quad M_h[j,k] = \int_\Omega \varphi_k(x) \varphi_j(x)\, dx, \quad j,k = 1, \ldots, M,$$

and $\underline{\overline{u}}$ is the load vector with the entries

$$\overline{u}_j = \int_\Omega \overline{u}(x)\, \varphi_j(x)\, dx \quad \text{for } j = 1, \ldots, M.$$

As in the continuous case we define $\underline{\lambda} := M_h \underline{u} + \varrho K_h \underline{u} - \underline{\overline{u}} \in \mathbb{R}^M$. Further, let the index set of the active nodes be denoted by $D_\pm := \{k = 1, \ldots, M : u_k = g_{\pm,k}\}$. Then we conclude the discrete complementarity conditions

$$\lambda_k = 0,\ g_{-,k} < u_k < g_{+,k} \text{ for } k \notin D_\pm,\ \lambda_k \le 0 \text{ for } k \in D_+,\ \lambda_k \ge 0 \text{ for } k \in D_-, \tag{3.8}$$

which are equivalent to

$$\lambda_k = \min\{0, \lambda_k + c(g_{+,k} - u_k)\} + \max\{0, \lambda_k + c(g_{-,k} - u_k)\}, \quad c > 0.$$

Hence we have to solve a system $\underline{F}(\underline{u}, \underline{\lambda}) = \underline{0}$ of (non)linear equations

$$F_1(\underline{u}, \underline{\lambda}) = M_h \underline{u} + \varrho K_h \underline{u} - \underline{\lambda} - \underline{\overline{u}} = \underline{0}, \tag{3.9}$$

$$F_2(\underline{u}, \underline{\lambda}) = \underline{\lambda} - \min\{0, \underline{\lambda} + c(\underline{g}_+ - \underline{u})\} - \max\{0, \underline{\lambda} + c(\underline{g}_- - \underline{u})\}, \tag{3.10}$$

where the latter have to be considered componentwise. This will be discussed in Section 4.

3.2 Control constraints

In the case of control constraints we consider the variational inequality (3.2) for $K_h = K_{ch}$ where

$$K_{ch} := \left\{ u_h \in V_h : \langle f_-, v_h \rangle_{L^2(\Omega)} \leq \langle \nabla u_h, \nabla v_h \rangle_{L^2(\Omega)} \leq \langle f_+, v_h \rangle_{L^2(\Omega)} \; \forall v_h \in V_h, \; v_h \geq 0 \right\}.$$

Theorem 3.4 *Let $u_\varrho \in K_c$ and $u_{\varrho h} \in K_{ch}$ be the unique solutions of the variational inequalities (2.12) and (3.2), respectively. Assume $\overline{u} \in K_c \cap H^1_\Delta(\Omega)$ and $f_\pm \in L^2(\Omega)$. When choosing $\varrho = h^2$, then there holds the error estimate*

$$\|u_\varrho - u_{\varrho h}\|^2_{L^2(\Omega)} + h^2 \, \|\nabla(u_\varrho - u_{\varrho h})\|^2_{L^2(\Omega)} \leq c \, h^4 \left[\|\Delta \overline{u}\|^2_{L^2(\Omega)} + \|f_\pm\|^2_{L^2(\Omega)} \right]. \tag{3.11}$$

Proof. Due to Lemma 2.3 we have

$$\|\Delta u_\varrho\|^2_{L^2(\Omega)} = \|f_\pm\|^2_{L^2(\Omega_\pm)} + \|\Delta u_\varrho\|^2_{L^2(\Omega \setminus \Omega_\pm)} \leq \|\Delta \overline{u}\|^2_{L^2(\Omega)} + \|f_\pm\|^2_{L^2(\Omega)},$$

and

$$|u_\varrho|^2_{H^2(\Omega)} \leq c \, \|\Delta u_\varrho\|^2_{L^2(\Omega)} \leq c \left[\|\Delta \overline{u}\|^2_{L^2(\Omega)} + \|f_\pm\|^2_{L^2(\Omega)} \right]$$

follows. For $u_\varrho \in K_c$ we define $P_h u_\varrho \in V_h$ as the unique solution of the variational formulation

$$\langle \nabla P_h u_\varrho, \nabla v_h \rangle_{L^2(\Omega)} = \langle \nabla u_\varrho, \nabla v_h \rangle_{L^2(\Omega)} \quad \text{for all } v_h \in V_h.$$

By construction we have $P_h u_\varrho \in K_{ch}$, and using standard finite element error estimates including the Nitsche trick we conclude

$$\|u_\varrho - P_h u_\varrho\|^2_{L^2(\Omega)} + \varrho \, \|\nabla(u_\varrho - P_h u_\varrho)\|^2_{L^2(\Omega)} \leq c \left(h^4 + \varrho \, h^2 \right) |u_\varrho|^2_{H^2(\Omega)}.$$

The assertion now follows as in the case of state constraints, we skip the details. □

As in the case of state constraints we can also derive error estimates for less regular target functions.

Using the isomorphism $v_h \in K_{ch} \leftrightarrow \underline{v} \in \mathbb{R}^M$ we can write the variational inequality (3.2) as: Find $\underline{u} \in \mathbb{R}^M \leftrightarrow u_{\varrho h} \in K_{ch}$ such that

$$((M_h + \varrho K_h)\underline{u} - \overline{\underline{u}}, \underline{v} - \underline{u}) \geq 0 \quad \text{for all } \underline{v} \in \mathbb{R}^M \leftrightarrow v_h \in K_{ch} \, .$$

The control constraints $v_h \in K_{ch}$ are equivalent to

$$f_{-,i} \leq (K_h \underline{v})_i \leq f_{+,i} \quad \text{for all } i = 1, \ldots, M \, .$$

On the other hand, the discrete variational inequality can be written as

$$(\underline{w}, K_h \underline{v} - K_h \underline{u}) \geq 0, \quad \text{where} \quad \underline{w} := (K_h^{-1} M_h + \varrho I)\underline{u} - K_h^{-1} \overline{\underline{u}}.$$

We introduce the discrete active sets $I_\pm := \{i \in \mathbb{R}^M : (K_h\underline{u})_i = f_{\pm,i}\}$ and conclude

$$\sum_{i\in I_+} w_i \underbrace{[(K_h\underline{v})_i - f_{+,i}]}_{\le 0} + \sum_{i\in I_-} w_i \underbrace{[(K_h\underline{v})_i - f_{-,i}]}_{\ge 0} + \sum_{i\in I\setminus I_\pm} w_i[(K_h\underline{v})_i - (K_h\underline{u})_i] \ge 0\,.$$

Let us define $g_i = f_{+,i}$ for $i \in I_+$, $g_i = f_{-,i}$ for $i \in I_-$, and $g_i = (K_h\underline{u})_i$ for $i \in I\setminus I_\pm$. For some $j \in I_+$ we set $g_j = f_{+,j} - \alpha$ with $0 < \alpha < f_{+,j} - f_{-,j}$, and we solve $K_h\underline{v} = \underline{g}$. By construction we obtain $-w_j\alpha \ge 0$, i.e., $w_j \le 0$ for $j \in I_+$, and in the same way we conclude $w_j \ge 0$ for $j \in I_-$ as well as $w_j = 0$ for $j \in I\setminus I_\pm$. Hence we have the discrete complementarity conditions

$$w_j = 0 \ : \ f_{-,j} < (K_h\underline{u})_j < f_{+,j},\ j \notin I_\pm,\ w_j \le 0 \ : \ (K_h\underline{u})_j = f_{+,j},\ w_j \ge 0 \ : \ (K_h\underline{u})_j = f_{-,j}.$$

As in the case of state constraints we have to solve a system of (non)linear equations,

$$F_1(\underline{u},\underline{w}) = M_h\underline{u} + \varrho K_h\underline{u} - K_h\underline{w} - \overline{\underline{u}} = \underline{0}, \tag{3.12}$$

$$F_2(\underline{u},\underline{w}) = \underline{w} - \min\{0, \underline{w} + c(\underline{f}_+ - K_h\underline{u})\} - \max\{0, \underline{w} + c(\underline{f}_- - K_h\underline{u})\}, \quad c > 0. \tag{3.13}$$

3.3 Finite element approximation of the control

When the state u_ϱ, i.e., its finite element approximation $u_{\varrho h}$, is known it remains to find the related control $z_\varrho = -\Delta u_\varrho \in H^{-1}(\Omega)$, i.e., an appropriate finite element approximation of $\widetilde{z}_\varrho = -\Delta u_{\varrho h} \in H^{-1}(\Omega)$. To do so, we define $A : H^1_0(\Omega) \to H^{-1}(\Omega)$ satisfying

$$\langle Au, v\rangle_\Omega = \langle \nabla u, \nabla v\rangle_{L^2(\Omega)} \quad \text{for all } u, v \in H^1_0(\Omega),$$

and we can compute $\widetilde{z}_\varrho \in H^{-1}(\Omega)$ as unique solution of the variational formulation

$$\langle A^{-1}\widetilde{z}_\varrho, \psi\rangle_\Omega = \langle u_{\varrho h}, \psi\rangle_{L^2(\Omega)} \quad \text{for all } \psi \in H^{-1}(\Omega).$$

In addition to the finite element space $V_h \subset H^1_0(\Omega)$ of piecewise linear and continuous basis functions φ_k we now define the ansatz space $Z_H = S^0_H(\Omega) = \text{span}\{\psi_\ell\}_{\ell=1}^N$ of piecewise constant basis functions ψ_ℓ which are defined with respect to some mesh of mesh size $H \simeq h$. Then we can find the finite element approximation $\widetilde{z}_{\varrho H} \in Z_H$ as unique solution of the variational formulation

$$\langle A^{-1}\widetilde{z}_{\varrho H}, \psi_H\rangle_\Omega = \langle u_{\varrho h}, \psi_H\rangle_{L^2(\Omega)} \quad \text{for all } \psi_H \in Z_H. \tag{3.14}$$

In addition, let $z_{\varrho H} \in Z_H$ be the solution of the variational formulation

$$\langle A^{-1}z_{\varrho H}, \psi_H\rangle_\Omega = \langle u_\varrho, \psi_H\rangle_{L^2(\Omega)} = \langle A^{-1}z_\varrho, \psi_H\rangle_\Omega \quad \text{for all } \psi_H \in Z_H.$$

When using standard arguments we conclude Cea's lemma

$$\|z_\varrho - z_{\varrho H}\|_{H^{-1}(\Omega)} \le \inf_{\psi_H \in Z_H} \|z_\varrho - \psi_H\|_{H^{-1}(\Omega)},$$

and the error estimate

$$\|z_\varrho - z_{\varrho H}\|_{H^{-1}(\Omega)} \leq c\, H\, \|z_\varrho\|_{L^2(\Omega)} = c\, H\, \|\Delta u_\varrho\|_{L^2(\Omega)}.$$

In the case of state constraints we further have, as in the proof of Lemma 2.2,

$$\begin{aligned}
\|\Delta u_\varrho\|^2_{L^2(\Omega)} &= \|\Delta g_\pm\|^2_{L^2(\Omega_\pm)} + \|\Delta u_\varrho\|^2_{L^2(\Omega\setminus\Omega_\pm)} = \|\Delta g_\pm\|^2_{L^2(\Omega_\pm)} + \frac{1}{\varrho^2}\,\|\overline{u} - u_\varrho\|^2_{L^2(\Omega\setminus\Omega_\pm)} \\
&\leq \|\Delta g_\pm\|^2_{L^2(\Omega)} + \varrho^{s-2}\,\|\overline{u}\|^2_{H^s(\Omega)} \leq \Big(\|\Delta g_\pm\|_{L^2(\Omega)} + \varrho^{s/2-1}\,\|\overline{u}\|_{H^s(\Omega)}\Big)^2,
\end{aligned}$$

i.e., recall $H \simeq h$ and $\varrho = h^2$,

$$\|z_\varrho - z_{\varrho H}\|_{H^{-1}(\Omega)} \leq c\, h^{s-1}\,\Big(\|g_\pm\|_{H^2(\Omega)} + \|\overline{u}\|_{H^s(\Omega)}\Big).$$

Note that in the case of control constraints we obtain a similar result, when assuming $f_\pm \in L^2(\Omega)$ instead of $g_\pm \in H^1_\Delta(\Omega)$. In any case, we have the perturbed Galerkin orthogonality

$$\langle A^{-1}(z_{\varrho H} - \widetilde{z}_{\varrho H}), \psi_H\rangle_\Omega = \langle u_\varrho - u_{\varrho h}, \psi_H\rangle_{L^2(\Omega)} \quad \text{for all } \psi_H \in Z_H,$$

from which we conclude

$$\begin{aligned}
&\|z_{\varrho H} - \widetilde{z}_{\varrho H}\|^2_{H^{-1}(\Omega)} = \langle A^{-1}(z_{\varrho H} - \widetilde{z}_{\varrho H}), z_{\varrho H} - \widetilde{z}_{\varrho H}\rangle_\Omega = \langle u_\varrho - u_{\varrho h}, z_{\varrho H} - \widetilde{z}_{\varrho H}\rangle_{L^2(\Omega)} \\
&\leq \|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)}\|z_{\varrho H} - \widetilde{z}_{\varrho H}\|_{L^2(\Omega)} \leq c\, h^s\,\sqrt{\|\overline{u}\|^2_{H^s(\Omega)} + \|g_\pm\|^2_{H^s(\Omega)}}\, H^{-1}\,\|z_{\varrho H} - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)},
\end{aligned}$$

when using an inverse inequality in Z_H, and the related error estimates for the approximate state. This gives

$$\|z_{\varrho H} - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)} \leq c\, h^{s-1}\,\Big(\|\overline{u}\|^2_{H^s(\Omega)} + \|g_\pm\|^2_{H^2(\Omega)}\Big),$$

and, therefore,

$$\|z_\varrho - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)} \leq \|z_\varrho - z_{\varrho H}\|_{H^{-1}(\Omega)} + \|z_{\varrho H} - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)} \leq c\, h^{s-1}\,\Big(\|\overline{u}\|^2_{H^s(\Omega)} + \|g_\pm\|^2_{H^2(\Omega)}\Big)$$

follows. Note that we cannot expect any order of convergence for the approximate control in $H^{-1}(\Omega)$ when we have $\overline{u} \in H^s(\Omega)$ for $s < 1$ only. In this case we have to measure the

error in a weaker norm. For this we consider, using the L^2 projection $Q_H : L^2(\Omega) \to Z_H$,

$$\begin{aligned}
\|z_\varrho - \widetilde{z}_{\varrho H}\|_{H^{-2}(\Omega)} &= \sup_{0\neq v\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{\langle z_\varrho - \widetilde{z}_{\varrho H}, v\rangle_\Omega}{\|v\|_{H^2(\Omega)}} \\
&= \sup_{0\neq v=A^{-1}\psi\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{\langle z_\varrho - \widetilde{z}_{\varrho H}, A^{-1}\psi\rangle_\Omega}{\|v\|_{H^2(\Omega)}} \\
&= \sup_{0\neq v=A^{-1}\psi\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{\langle A^{-1}(z_\varrho - \widetilde{z}_{\varrho H}), \psi\rangle_\Omega}{\|v\|_{H^2(\Omega)}} \\
&= \sup_{0\neq v=A^{-1}\psi\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{\langle A^{-1}(z_\varrho - \widetilde{z}_{\varrho H}), \psi - Q_H\psi\rangle_\Omega + \langle A^{-1}(z_\varrho - \widetilde{z}_{\varrho H}), Q_H\psi\rangle_\Omega}{\|v\|_{H^2(\Omega)}} \\
&= \sup_{0\neq v=A^{-1}\psi\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{\langle A^{-1}(z_\varrho - \widetilde{z}_{\varrho H}), \psi - Q_H\psi\rangle_\Omega + \langle u_\varrho - u_{\varrho h}, Q_H\psi\rangle_\Omega}{\|v\|_{H^2(\Omega)}} \\
&= \sup_{0\neq v=A^{-1}\psi\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{\|z_\varrho - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)}\|\psi - Q_H\psi\|_{H^{-1}(\Omega)} + \|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)}\|Q_H\psi\|_{L^2(\Omega)}}{\|v\|_{H^2(\Omega)}} \\
&\leq \sup_{0\neq v=A^{-1}\psi\in H_0^1(\Omega)\cap H^2(\Omega)} \frac{c\,H\,\|z_\varrho - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)}\|\psi\|_{L^2(\Omega)} + \|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)}}{\|v\|_{H^2(\Omega)}} \\
&\leq c\,H\,\|z_\varrho - \widetilde{z}_{\varrho H}\|_{H^{-1}(\Omega)} + \|u_\varrho - u_{\varrho h}\|_{L^2(\Omega)} = c\,h^s\Big(\|\overline{u}\|_{H^s(\Omega)} + \|g_\pm\|_{H^2(\Omega)}\Big),
\end{aligned}$$

when assuming $\overline{u} \in H_0^s(\Omega)$ for $s \in [0,1]$, or $\overline{u} \in H_0^1(\Omega) \cap H^s(\Omega)$ for $s \in (1,2]$. In particular, this error estimate also allows the consideration of discontinuous target functions $\overline{u} \in H_0^s(\Omega)$, $s < 1/2$. However, the application of the inverse A^{-1} does not allow a direct evaluation, and hence we need to introduce a suitable approximation as follows: For any $z \in H^{-1}(\Omega)$ we define $p_z \in H_0^1(\Omega)$ as the unique solution of the variational formulation

$$\langle Ap_z, q\rangle_\Omega = \langle \nabla p_z, \nabla q\rangle_{L^2(\Omega)} = \langle z, q\rangle_\Omega \quad \text{for all } q \in H_0^1(\Omega).$$

Moreover, we compute an approximate solution $p_{zh} \in V_h$, for simplicity we consider the finite element space V_h as already used for the approximation of the state, such that

$$\langle \nabla p_{zh}, \nabla q_h\rangle_{L^2(\Omega)} = \langle z, q_h\rangle_\Omega \quad \text{for all } q_h \in V_h$$

which defines an approximation $\widetilde{A}^{-1}z := p_{zh}$ of $p_z = A^{-1}z$. From

$$\|\nabla p_{zh}\|_{L^2(\Omega)}^2 = \langle \nabla p_{zh}, \nabla p_{zh}\rangle_{L^2(\Omega)} = \langle z, p_{zh}\rangle_\Omega \leq \|z\|_{H^{-1}(\Omega)}\|\nabla p_{zh}\|_{L^2(\Omega)}$$

we immediately conclude boundedness, i.e.,

$$\|\widetilde{A}^{-1}z\|_{H_0^1(\Omega)} = \|p_{zh}\|_{H_0^1(\Omega)} = \|\nabla p_{zh}\|_{L^2(\Omega)} \leq \|z\|_{H^{-1}(\Omega)} \quad \text{for all } z \in H^{-1}(\Omega).$$

For $z_H \in Z_H$ let $p_{z_H h} = \widetilde{A}^{-1}z_H \in V_h$ be the unique solution of the variational formulation

$$\langle \nabla p_{z_H h}, \nabla q_h\rangle_{L^2(\Omega)} = \langle z_H, q_h\rangle_{L^2(\Omega)} \quad \text{for all } q_h \in V_h,$$

while $p_{z_H} = A^{-1} z_H$ solves

$$\langle \nabla p_{z_H}, \nabla q \rangle_{L^2(\Omega)} = \langle z_H, q \rangle_{L^2(\Omega)} \quad \text{for all } q \in H_0^1(\Omega),$$

which is the weak formulation of the Poisson equation $-\Delta p_{z_H} = z_H$ with homogeneous Dirichlet boundary conditions. When using standard arguments, i.e., Cea's lemma and the approximation property of V_h, we conclude the error error estimate

$$\begin{aligned} \|\nabla(p_{z_H} - p_{z_H h})\|_{L^2(\Omega)} &\leq \inf_{q_h \in V_h} \|\nabla(p_{z_H} - q_h)\|_{L^2(\Omega)} \leq c_A\, h\, |p_{z_H}|_{H^2(\Omega)} \\ &\leq \widetilde{c}_A\, h\, \|\Delta p_{z_H}\|_{L^2(\Omega)} = \widetilde{c}_A\, h\, \|z_H\|_{L^2(\Omega)} \leq \widetilde{c}_A c_I\, \frac{h}{H}\, \|z_H\|_{H^{-1}(\Omega)}, \end{aligned}$$

and when using an inverse inequality, e.g., [34]. In the case

$$h < \frac{1}{2\widetilde{c}_A c_I}\, H \tag{3.15}$$

we therefore have

$$\|\nabla(p_{z_H} - p_{z_H h})\|_{L^2(\Omega)} \leq \frac{1}{2}\, \|z_H\|_{H^{-1}(\Omega)}.$$

Hence we can write

$$\begin{aligned} \langle \widetilde{A}^{-1} z_H, z_H \rangle_\Omega &= \langle p_{z_H h}, z_H \rangle_{L^2(\Omega)} = \langle p_{z_H}, z_H \rangle_\Omega - \langle p_{z_H} - p_{z_H h}, z_H \rangle_\Omega \\ &\geq \langle A^{-1} z_H, z_H \rangle_\Omega - \|\nabla(p_{z_H} - p_{z_H h})\|_{L^2(\Omega)} \|z_H\|_{H^{-1}(\Omega)} \geq \frac{1}{2}\, \|z_H\|^2_{H^{-1}(\Omega)}. \end{aligned}$$

The approximate operator $\widetilde{A}$ is therefore discrete elliptic, if the mesh condition (3.15) is satisfied. Although the constants $\widetilde{c}_A$ and c_I are in general unknown, in our numerical experiments we have used $h = \frac{1}{4} H$.

Instead of (3.14) we now consider the variational formulation to find $\widehat{z}_{\varrho H} \in Z_H$ such that

$$\langle \widetilde{A}^{-1} \widehat{z}_{\varrho H}, \psi_H \rangle_\Omega = \langle u_{\varrho h}, \psi_H \rangle_{L^2(\Omega)} \quad \text{for all } \psi_H \in Z_H. \tag{3.16}$$

Unique solvability of (3.16) follows since $\widetilde{A}^{-1}$ is discrete elliptic. Note that (3.16) can be written as a mixed variational formulation to find $(\widehat{z}_{\varrho H}, p_{\widehat{z}_{\varrho H} h}) \in Z_H \times V_h$ such that

$$\langle p_{\widehat{z}_{\varrho H} h}, \psi_H \rangle_{L^2(\Omega)} = \langle u_{\varrho h}, \psi_H \rangle_{L^2(\Omega)}, \quad \langle \nabla p_{\widehat{z}_{\varrho H} h}, \nabla q_h \rangle_{L^2(\Omega)} = \langle \widehat{z}_{\varrho H}, q_h \rangle_{L^2(\Omega)}$$

is satisfied for all $(\psi_H, q_h) \in Z_H \times V_h$. This formulation is equivalent to the linear system of algebraic equations,

$$\hat{M}_h^\top \underline{p} = \hat{M}_h^\top \underline{u}, \quad K_h \underline{p} = \hat{M}_h \underline{z},$$

where in addition to the standard finite element stiffness matrix K_h we have used the mass matrix $\hat{M}_h$ defined by

$$\hat{M}_h[j, \ell] = \langle \psi_\ell, \varphi_j \rangle_{L^2(\Omega)}, \quad \ell = 1, \ldots, N;\ j = 1, \ldots, M.$$

Since the finite element stiffness matrix K_h is invertible, we can eliminate $\underline{p} = K_h^{-1}\hat{M}_h\underline{z}$ to end up with the Schur complement system to be solved,

$$\hat{M}_h^\top K_h^{-1}\hat{M}_h\underline{z} = \hat{M}_h^\top \underline{u}. \tag{3.17}$$

The mesh condition (3.15) not only implies unique solvability of (3.17), but the discrete ellipticity of $\widetilde{A}^{-1}$ also provides related error estimates, when applying the Strang lemma, e.g., [4, 34]. With this we conclude the final error estimate

$$\|z_\varrho - \widehat{z}_{\varrho H}\|_{H^{-2}(\Omega)} \le c\, h^s \left(\|\overline{u}\|_{H^s(\Omega)} + \|g_\pm\|_{H^2(\Omega)} \right). \tag{3.18}$$

Note that this estimate remains true when considering control constraints $f_\pm \in L^2(\Omega)$.

4 Semi-smooth Newton method

In this section we discuss the iterative solution of the discrete variational inequality (3.7). For the solution of (3.9) and (3.10) we can apply a semi-smooth Newton method which is equivalent to an active set strategy as given in Algorithm 1, see [7, 16, 18, 19], and [35]. The generalization to the iterative solution of (3.12) and (3.13) is straightforward and will not be discussed here.

Algorithm 1 Active set algorithm [16]

Require: Initial values $\underline{u}^0, \underline{\lambda}^0$

(a) $m = 0$

(b) Set

$$y_{+,k}^m = \lambda_k^m + c\,[g_+(x_k) - u_k^m] \quad \text{and} \quad y_{-,k}^m = \lambda_k^m + c\,[g_-(x_k) - u_k^m]$$

while stop criterion is not fulfilled **do**

(i) Set

$$\mathcal{I}^m = \{k :\, y_{+,k}^m \ge 0,\, y_{-,k}^m \le 0\}, \quad \mathcal{A}_-^m = \{k :\, y_{-,k}^m > 0\}, \quad \mathcal{A}_+^m = \{k :\, y_{+,k}^m < 0\}$$

(ii) Solve

$$(M_h + \varrho K_h)\underline{u}^{m+1} - \underline{\lambda}^{m+1} = \underline{\overline{u}},\ u_k^{m+1} = g_\pm(x_k),\ k \in \mathcal{A}_\pm^m,\ \lambda_k^{m+1} = 0,\ k \in \mathcal{I}^m.$$

(iii) $m = m + 1$

end while

The semi-smooth Newton method successively computes the roots of $\underline{F}(\underline{u}, \underline{\lambda}) = \underline{0}$ by

$$\begin{pmatrix} \underline{u}^{m+1} \\ \underline{\lambda}^{m+1} \end{pmatrix} = \begin{pmatrix} \underline{u}^m \\ \underline{\lambda}^m \end{pmatrix} - \left(D\underline{F}(\underline{u}^m, \underline{\lambda}^m) \right)^{-1} \underline{F}(\underline{u}^m, \underline{\lambda}^m), \tag{4.1}$$

with the Jacobian $D\underline{F}$ given by

$$D\underline{F}(\underline{u},\underline{\lambda}) = \begin{pmatrix} M_h + \varrho K_h & -I \\ c(G'_{min}(\underline{u},\underline{\lambda}) + G'_{max}(\underline{u},\underline{\lambda})) & I - (G'_{min}(\underline{u},\underline{\lambda}) + G'_{max}(\underline{u},\underline{\lambda})) \end{pmatrix}.$$

The diagonal entries of

$$\begin{aligned} G'_{min}(\underline{u},\underline{\lambda}) &= \operatorname{diag}\Big(g'_{min}\big(\lambda_k + c\,[g_+(x_k) - u_k]\big)\Big), \\ G'_{max}(\underline{u},\underline{\lambda}) &= \operatorname{diag}\Big(g'_{max}\big(\lambda_k + c\,[g_-(x_k) - u_k]\big)\Big) \end{aligned}$$

are the slant derivatives of the functions $g_{\min}(y) = \min\{0, y\}$ and $g_{\max}(y) = \max\{0, y\}$ defined by

$$g'_{\min}(y) = \begin{cases} 1, & y < 0, \\ 0, & y \geq 0, \end{cases} \quad \text{and} \quad g'_{\max}(y) = \begin{cases} 0, & y \leq 0, \\ 1, & y > 0. \end{cases}$$

Rewriting the system (4.1) gives

$$\begin{aligned} &\begin{pmatrix} M_h + \varrho K_h & -I \\ c(G'_{\min}(\underline{u}^m,\underline{\lambda}^m) + G'_{\max}(\underline{u}^m,\underline{\lambda}^m)) & I - (G'_{\min}(\underline{u}^m,\underline{\lambda}^m) + G'_{\max}(\underline{u}^m,\underline{\lambda}^m)) \end{pmatrix} \begin{pmatrix} \underline{u}^m - \underline{u}^{m+1} \\ \underline{\lambda}^m - \underline{\lambda}^{m+1} \end{pmatrix} \\ &\quad = F(\underline{u}^m,\underline{\lambda}^m). \end{aligned} \tag{4.2}$$

From the first line we get

$$(M_h + \varrho K_h)(\underline{u}^m - \underline{u}^{m+1}) - \underline{\lambda}^m + \underline{\lambda}^{m+1} = (M_h + \varrho K_h)\underline{u}^m - \underline{\lambda}^m - \overline{\underline{u}},$$

from which we conclude

$$(M_h + \varrho K_h)\underline{u}^{m+1} - \underline{\lambda}^{m+1} = \overline{\underline{u}}.$$

With

$$y^m_{+,k} := \lambda^m_k + c\,[g_+(x_k) - u^m_k] \quad \text{and} \quad y^m_{-,k} := \lambda^m_k + c\,[g_-(x_k) - u^m_k],$$

the second line reads, componentwise,

$$\begin{aligned} &c[g'_{\min}(y^m_{+,k}) + g'_{\max}(y^m_{-,k})](u^m_k - u^{m+1}_k) + \lambda^m_k - \lambda^{m+1}_k \\ &\quad -[g'_{\min}(y^m_{+,k})) + g'_{\max}(y^m_{-,k})](\lambda^m_k - \lambda^{m+1}_k) = \lambda^m_k - \min\{0, y^m_{+,k}\} - \max\{0, y^m_{-,k}\}. \end{aligned}$$

We distinguish the following three cases.

1. $y^m_{+,k} \geq 0$ and $y^m_{-,k} \leq 0$: Then, $g'_{min}(y^m_{+,k}) = g'_{max}(y^m_{-,k}) = 0$, and we compute

$$\lambda^{m+1}_k = 0.$$

2. $y^m_{-,k} > 0$: From this we get $\lambda^m_k > c\,[u^m_k - g_-(x_k)]$ and we compute

$$\lambda^m_k + c\,[g_+(x_k) - u^m_k] > c\,[u^m_k - g_-(x_k) + g_+(x_k) - u^m_k] = c\,[g_+(x_k) - g_-(x_k)] > 0,$$

i.e., $y^m_{+,k} > 0$. Therefore, $g'_{\min}(y^m_{+,k}) = 0$ and $g'_{\max}(y^m_{-,k}) = 1$, and we get

$$u^{m+1}_k = g_-(x_k).$$

3. $y^m_{+,k} < 0$: Then, as in the second case, we compute $y^m_{-,k} < 0$ to get $g'_{\min}(y^m_{+,k}) = 1$, $g'_{\max}(y^m_{-,k}) = 0$, and thus
$$u^{m+1}_k = g_+(x_k).$$

Therefore we see, that the iterates of the semi-smooth Newton method (4.1) fulfill the active set strategy as given in Algorithm 1.

5 Numerical results

For our numerical tests we consider the domain $\Omega = (0,1)^2$ and the following target functions $\overline{u}_i \in \mathcal{C}^\infty(\Omega) \cap H^1_0(\Omega)$, $i = 1, 2$,

$$\overline{u}_1(x, y) = \sin(\pi x)\sin(\pi y),$$

and

$$\overline{u}_2(x, y) := \overline{u}_2(x, y; k) = H_k(x)H_k(y), \quad \text{where } H_k(s) = \frac{1}{1 + e^{-k(s-0.25)}} - \frac{1}{1 + e^{-k(s-0.75)}},$$

for $k = 40$, see Fig. 1, for which we can compute $z_i = -\Delta\overline{u}_i$ analytically. We also consider the discontinuous target $\overline{u}_3 := \lim_{k\to\infty} \overline{u}_2(\cdot, \cdot; k) \in H^{1/2-\varepsilon}(\Omega)$, $\varepsilon > 0$, for which we *cannot* compute the control analytically, given as

$$\overline{u}_3(x, y) = \begin{cases} 1, & (x, y) \in [0.25, 0.75]^2, \\ 0, & \text{else.} \end{cases}$$

5.1 State constraints

In order to incorporate constraints on the state, we apply the semi-smooth Newton algorithm, where we set $\varrho = h^2$, and the initial values

$$\underline{u}^0 = (h^2 K_h + M_h)^{-1}\underline{\overline{u}} \in \mathbb{R}^M \quad \text{and} \quad \underline{\lambda}^0 = \underline{0}.$$

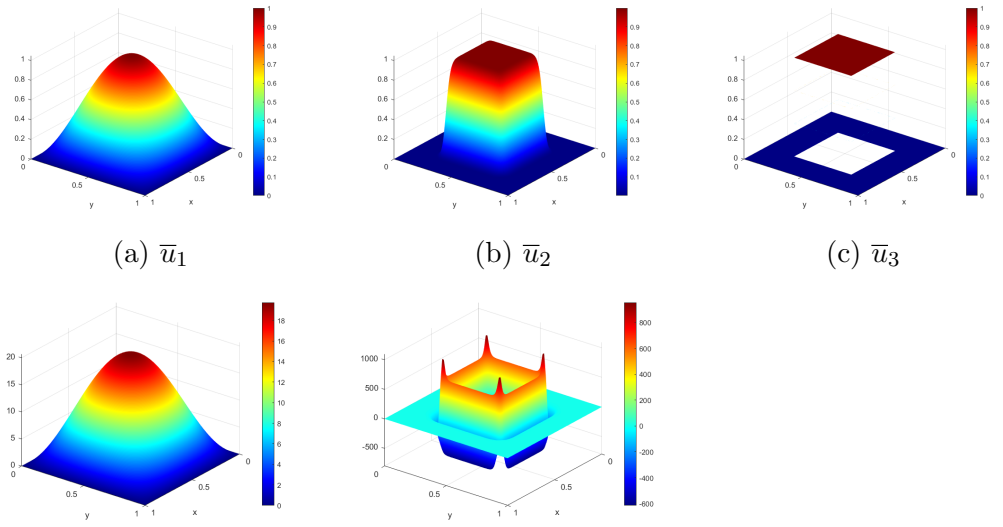
A stopping criterion is then defined using a maximal absolute error in each node, i.e., we stop if

$$\text{tol} := \max\{\text{tol}_+, \text{tol}_-\} < 10^{-5}, \tag{5.1}$$

where

$$\text{tol}_+ := \max_{\{k:u_k > g_+(x_k)\}} |u_k - g_+(x_k)| \quad \text{and} \quad \text{tol}_- := \max_{\{k:u_k < g_-(x_k)\}} |u_k - g_-(x_k)|.$$

After computing the state $\underline{u} \leftrightarrow u_{\varrho h} \in K_{sh}$, we can reconstruct the control $\underline{z} \leftrightarrow z_{\varrho H} \in Z_H$ by solving the Schur complement system (3.17). In order to ensure stability of the discrete



(a) $\overline{u}_1$ (b) $\overline{u}_2$ (c) $\overline{u}_3$

(d) $z_1 = -\Delta\overline{u}_1$ (e) $z_2 = -\Delta\overline{u}_2$

Figure 1: Target functions $\overline{u}_i$ and $z_j = -\Delta\overline{u}_j$, $i = 1, 2, 3$, $j = 1, 2$.

system, we choose $h = H/4$. For the targets $\overline{u}_i$ we consider the upper and lower constraints $g_{\pm}^{(j)}$ given by

$$g_{-}^{(1)}(x) \equiv 0,\ g_{+}^{(1)}(x) = 0.5 \cdot \overline{u}_1(x), \quad g_{-}^{(2)}(x) \equiv 0,\ g_{+}^{(2)}(x) = 0.5 \cdot \overline{u}_2(x).$$

The results are depicted in Fig. 2 and Fig. 4. Note that $g_{+}^{(2)}(x, y) \leq \overline{u}_i(x, y)$ for $i = 1, 2$ and all $(x, y) \in \Omega$. Thus, the constrained solutions as shown in Fig. 2 (c) and (i) as well as the controls in (f) and (l) coincide.

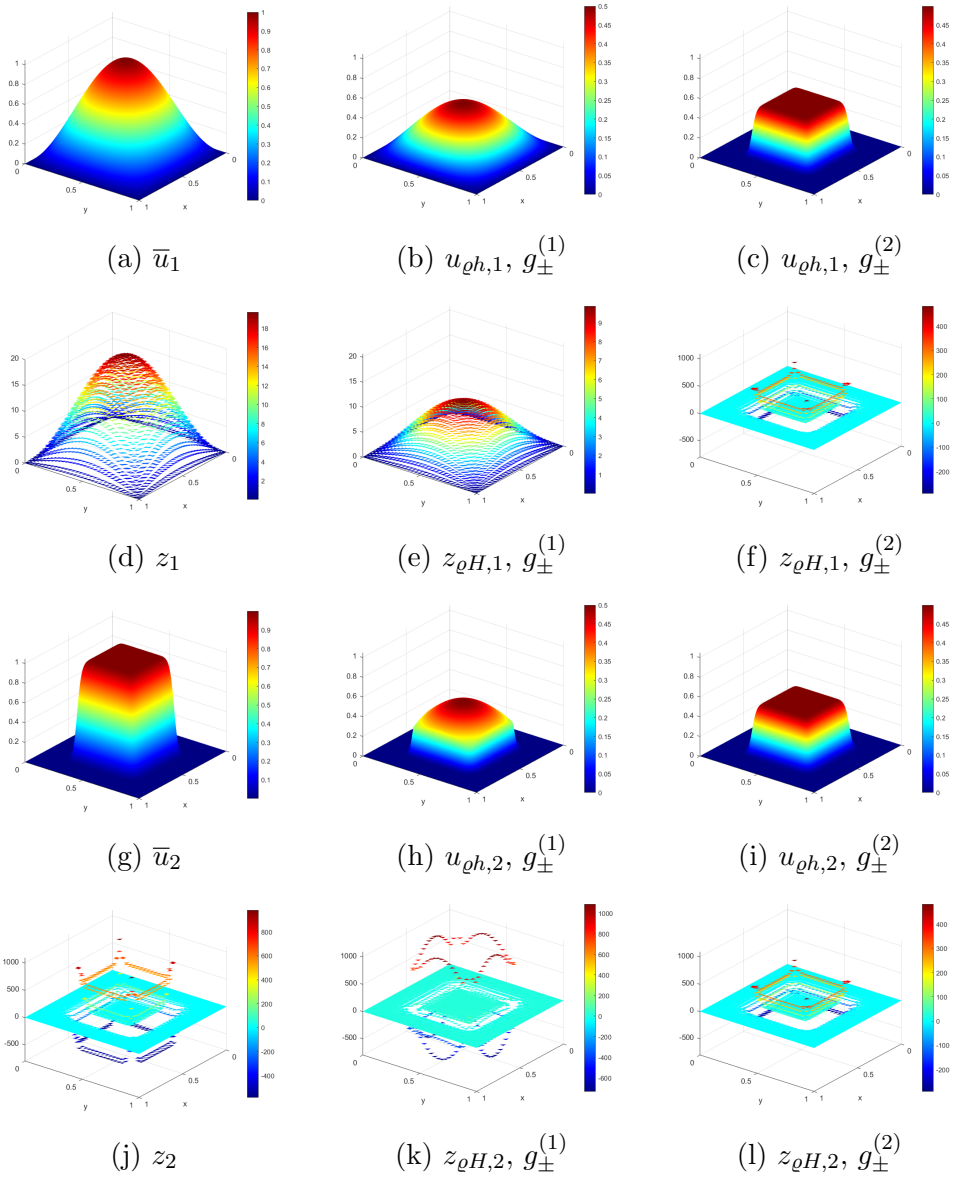
5.2 Control constraints

In order to incorporate contraints on the control, we apply the semi-smooth Newton algorithm, where we set $\varrho = h^2$, and the initial values

$$\underline{u}^0 = (h^2 K_h + M_h)^{-1} \underline{\overline{u}} \in \mathbb{R}^M \quad \text{and} \quad \underline{w}^0 = \underline{0}.$$

Again we apply the stopping criteria (5.1) but now we use

$$\mathrm{tol}_+ := \max_{\{k:(K_h \underline{u})_k > f_{+,k}\}} |(K_h \underline{u})_k - f_{+,k}| \quad \text{and} \quad \mathrm{tol}_- := \max_{\{k:(K_h \underline{u})_k < f_{-,k}\}} |(K_h \underline{u})_k - f_{-,k}|.$$



(a) $\overline{u}_1$ (b) $u_{\varrho h,1}$, $g_{\pm}^{(1)}$ (c) $u_{\varrho h,1}$, $g_{\pm}^{(2)}$

(d) z_1 (e) $z_{\varrho H,1}$, $g_{\pm}^{(1)}$ (f) $z_{\varrho H,1}$, $g_{\pm}^{(2)}$

(g) $\overline{u}_2$ (h) $u_{\varrho h,2}$, $g_{\pm}^{(1)}$ (i) $u_{\varrho h,2}$, $g_{\pm}^{(2)}$

(j) z_2 (k) $z_{\varrho H,2}$, $g_{\pm}^{(1)}$ (l) $z_{\varrho H,2}$, $g_{\pm}^{(2)}$

Figure 2: Targets $\overline{u}_i$ and unconstrained controls z_i, $i = 1, 2$, computed constrained states $u_{\varrho h,i}$ on a mesh with $N = 32768$ elements and $M = 16129$ DoFs with constraints $g_{\pm}^{(j)}$, and reconstruction of the controls $z_{\varrho H,i}$ on a mesh with $N_H = 2048$ elements.

For the upper and lower constraints on the control we consider the functions $f_{\pm}^{(j)}$ given by

$$\begin{aligned} f_-^{(1)}(x) &\equiv 0 \quad \text{and} \quad f_+^{(1)}(x) = 0.5 \cdot z_1(x) \\ f_-^{(2)}(x) &\equiv 0 \quad \text{and} \quad f_+^{(2)}(x) = \min\{z_1(x), 10\} \\ f_-^{(3)}(x) = \max\{\min\{z_2(x), 0\}, -500\} &\quad \text{and} \quad f_+^{(3)}(x) = \min\{\max\{z_2(x), 0\}, 500\} \\ f_-^{(4)}(x) &\equiv 0 \quad \text{and} \quad f_+^{(4)}(x) = \min\{\max\{z_2(x), 0\}, 1000\} \\ f_-^{(5)}(x) &\equiv 0 \quad \text{and} \quad f_+^{(5)}(x) = 4 \cdot \min\{\max\{z_2(x), 0\}, 250\}. \end{aligned}$$

The results are depicted in Fig. 3 and Fig. 4. Note, that for $\varrho = 0$ the control for $\overline{u}_3$ is given by $z_3 = -\Delta \overline{u}_3 \in H^{-3/2-\varepsilon}(\Omega)$ and thus can only be seen in a distributional sense. If we compute the control on a sufficiently fine mesh, this behaviour is resembled and the control explodes only in some points. Thus, we also give the reconstruction on a coarser mesh in Fig. 4 and with suitable constraints, which still gives a meaningful control.

6 Conclusions

In this paper, we have described and analyzed state and control constraints when considering elliptic distributed optimal control problems with energy regularization. We have proven optimal error estimates with respect to both the regularization parameter ϱ, and the finite element mesh width h. While for the solution of the nonlinear system we have used a semi-smooth Newton method, the design of an overall efficient iterative solution method including preconditioning was not within the scope of this paper. This approach can be extended to optimal control problems in three space dimensions, and to more involved applications. Moreover, following existing work for unconstrained optimal control problems subject to time dependent problems such as the heat and the wave equation, we can include state and control constraints also in these cases.

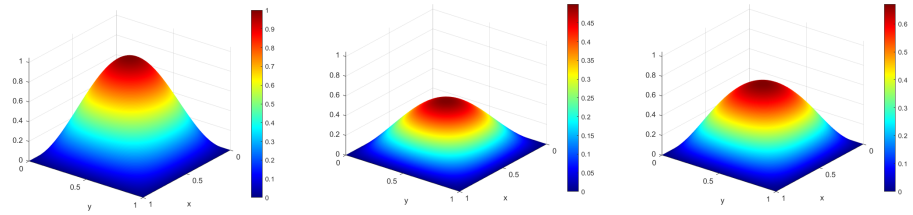
Acknowledgments

This work has been partially supported by the Austrian Science Fund (FWF) under the Grant Collaborative Research Center TRR361/F90: CREATOR Computational Electric Machine Laboratory.

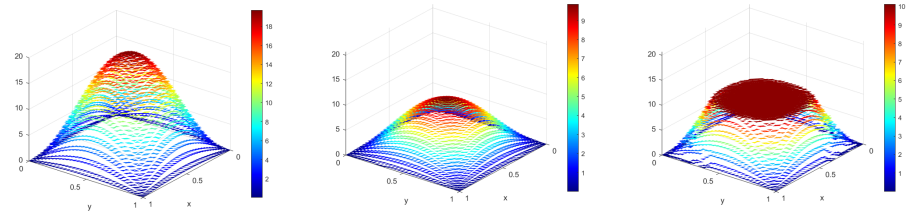
References

[1] T. Apel, O. Steinbach, and M. Winkler. Error estimates for neumann boundary control problems with energy regularization. *J. Numer. Math.*, 24:207–233, 2016.

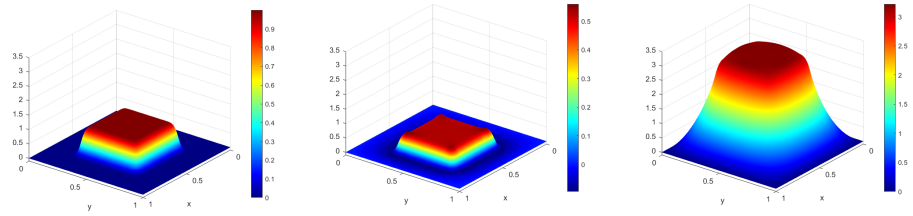
[2] M. Bergounioux and K. Kunisch. Primal-dual strategy for state-constrained optimal control problems. *Comp. Opt. Appl.*, 22:193–224, 2002.



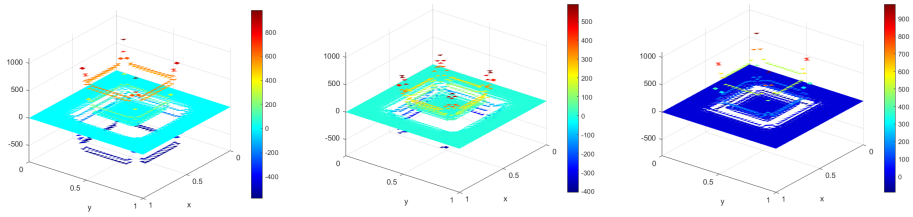
(a) $\overline{u}_1$ (b) $u_{\varrho h,1}$, $f_{\pm}^{(1)}$ (c) $u_{\varrho h,1}$, $f_{\pm}^{(2)}$



(d) z_1 (e) $z_{\varrho H,1}$, $f_{\pm}^{(1)}$ (f) $z_{\varrho H,1}$, $f_{\pm}^{(2)}$

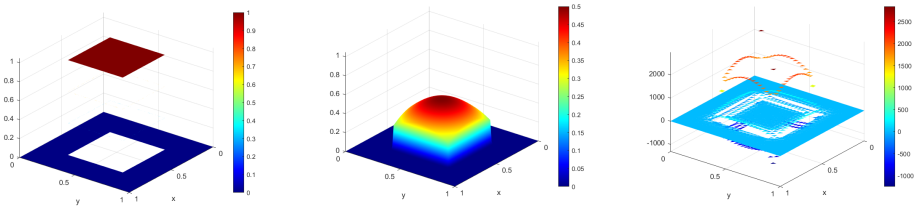


(g) $\overline{u}_2$ (h) $u_{\varrho h,2}$, $f_{\pm}^{(3)}$ (i) $u_{\varrho h,2}$, $f_{\pm}^{(4)}$

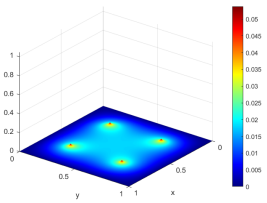


(j) z_2 (k) $z_{\varrho H,2}$, $f_{\pm}^{(3)}$ (l) $z_{\varrho H,2}$, $f_{\pm}^{(4)}$

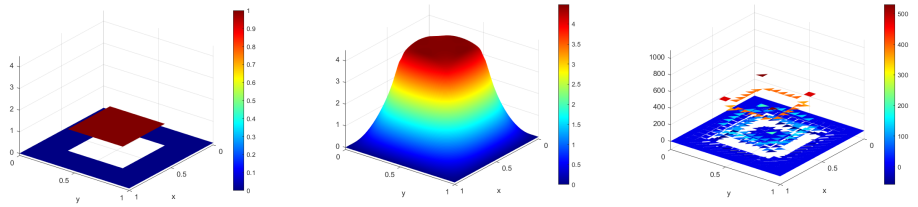
Figure 3: Targets $\overline{u}_i$ and unconstrained controls z_i, $i = 1, 2$, computed states $u_{\varrho h,i}$ on a mesh with $N = 32768$ elements and $M = 16129$ DoFs and reconstruction of the constrained controls $z_{\varrho H,i}$, with constraints $f_{\pm}^{(j)}$, on a mesh with $N_H = 2048$ elements.



(a) $\overline{u}_3$ (b) $u_{\varrho h,3}$, $g_{\pm}^{(1)}$ (c) $z_{\varrho H,3}$, $g_{\pm}^{(1)}$



(d) $u_{\varrho h,3}$, $f_{\pm}^{(4)}$ (e) $z_{\varrho H,3}$, $f_{\pm}^{(4)}$



(f) $\overline{u}_3$ (g) $u_{\varrho h,3}$, $f_{\pm}^{(5)}$ (h) $z_{\varrho H,3}$, $f_{\pm}^{(5)}$

Figure 4: Computed state $u_{\varrho h,3}$ for state constaints $g_{\pm}^{(1)}$, control constaints $f_{\pm}^{(4)}$ and $f_{\pm}^{(5)}$ on a mesh with $N = 32768$ elements and $M = 16129$ DoFs adn reconstruction of the control $z_{\varrho H,3}$ on a mesh with $N_H = 2048$ elements (1^{st} and 2^{nd} row) and with $N_H = 512$ (3^{rd} row).

[3] S. C. Brenner. Finite element methods for elliptic distributed optimal control problems with pointwise state constraints (survey). *Advances in Mathematical Sciences*, 21:3–16, 2020.

[4] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, 1994.

[5] H. R. Brezis and G. Stampacchia. Sur la régularité de la solution d'inéquations elliptiques. *Bull. Soc. Math. France*, 96:153–180, 1968.

[6] F. Brezzi, W. W. Hager, and P. A. Raviart. Error estimates for the finite element solution of variational inequalities. Part I. Primal theory. *Numer. Math.*, 28:431–443, 1977.

[7] X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38:1200–1216, 2001.

[8] S. Chowdhury, T. Gudi, and A. K. Nandakumaran. Error bounds for a Dirichlet boundary control problem based on energy spaces. *Math. Comp.*, 86:1103–1126, 2017.

[9] J. C. de los Reyes and K. Kunisch. A semi-smooth Newton method for control constrained boundary optimal control of the Navier–Stokes equations. *Nonlinear Analysis: Theory, Methods & Applications*, 62:1289–1316, 2005.

[10] P. Deuflhard, A. Schiela, and M. Weiser. Mathematical cancer therapy planning in deep regional hyperthermia. *Acta Numerica*, 21:307–378, 2012.

[11] R. S. Falk. Approximation of a class of optimal control problems with order of convergence estimates. *J. Math. Anal. Appl.*, 44:28–47, 1973.

[12] R. S. Falk. Error estimates for the approximation of a class of variational inequalities. *Math. Comp.*, 28:963–971, 1974.

[13] R. Glowinski. *Lectures on numerical methods for nonlinear variational problems*, volume 65 of *Lectures on Mathematics and Physics*. Springer, Berlin, New York, 1980.

[14] W. Gong, M. Mateos, J. Singler, and Y. Zhang. Analysis and approximations of Dirichlet boundary control of Stokes flows in the energy space. *SIAM J. Numer. Anal.*, 60:450–474, 2022.

[15] W. Gong and Z. Tan. A new finite element method for elliptic optimal control problems with pointwise state constraints in energy spaces. arXiv:2306.03246v1, 2023.

[16] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semi-smooth Newton method. *SIAM J. Optim.*, 13:865–888, 2002.

[17] M. Hintermüller and K. Kunisch. PDE-constrained optimization subject to pointwise constraints on the control, the state, and its derivative. *SIAM J. Optim.*, 20:1133–1156, 2010.

[18] M. Hintermüller and M. Ulbrich. A mesh-independence result for semi-smooth Newton methods. *Math. Program. Ser. B*, 101:151–184, 2004.

[19] K. Ito and K. Kunisch. Semi-smooth Newton methods for the Signorini problem. *Appl. Math.*, 53:455–468, 2008.

[20] M. Karkulik. A finite element method for elliptic Dirichlet boundary control problems. *Comput. Meth. Appl. Math.*, 20:827–843, 2020.

[21] V. Karl and D. Wachsmuth. An augmented Lagrange method for elliptic state constrained optimal control problems. *Comput. Optim. Appl.*, pages 857–880, 2018.

[22] N. Kikuchi. Convergence of a penalty-finite element approximation for an obstacle problem. *Numer. Math.*, 37:105–120, 1981.

[23] A. Kröner. Semi-smooth Newton methods for optimal control of the dynamical Lamé system with control constraints. *Num. Funct. Ana. Optim.*, 34:741–769, 2013.

[24] A. Kröner, K. Kunisch, and B. Vexler. Semismooth Newton methods for optimal control of the wave equation with control constraints. *SIAM J. Contr. Optim.*, 49:830–858, 2011.

[25] U. Langer, O. Steinbach, F. Tröltzsch, and H. Yang. Space-time finite element discretization of parabolic optimal control problems with energy regularization. *SIAM J. Numer. Anal.*, 59:675–695, 2021.

[26] U. Langer, O. Steinbach, and H. Yang. Robust space-time finite element error estimates for parabolic distributed optimal control problems with energy regularization. arXiv:2206.06455v1, 2022.

[27] U. Langer, O. Steinbach, and H. Yang. Robust discretization and solvers for elliptic optimal control problems with energy regularization. *Comput. Meth. Appl. Math.*, 22:97–111, 2022.

[28] J. L. Lions and G. Stampacchia. Variational inequalities. *Comm. Pure Appl. Math.*, 20:493–519, 1967.

[29] R. Löscher and O. Steinbach. Space-time finite element methods for distributed optimal control of the wave equation. arXiv:2211.02562v1, 2022.

[30] M. Neumüller and O. Steinbach. Regularization error estimates for distributed control problems in energy spaces. *Math. Methods Appl. Sci.*, 44:4176–4191, 2021.

[31] G. Of, T. X. Phan, and O. Steinbach. An energy space finite element approach for elliptic Dirichlet boundary control problems. *Numer. Math.*, 129:723–748, 2015.

[32] A. Rösch and D. Wachsmuth. Semi-smooth Newton method for an optimal control problem with control and mixed control-state constraints. *Optim. Meth. Softw.*, 26:169–186, 2011.

[33] A. Schiela and M. Weiser. Barrier methods for a control problem from hyperthermia treatment planning. In M. Diehl, F. Glineur, E. Jarlebring, and W. Michiels, editors, *Recent advances in optimization and its applications in engineering*, pages 419–428. Springer, Berlin, Heidelberg, 2010.

[34] O. Steinbach. *Numerical approximation methods for elliptic boundary value problems. Finite and boundary elements*. Springer, New York, 2008.

[35] O. Steinbach. Boundary element methods for variational inequalities. *Numer. Math.*, 126:173–197, 2014.

[36] O. Steinbach. Space-time finite element methods for optimal control problems. In H. Harbrecht, A. Kunoth, V. Simoncini, and K. Urban, editors, *Optimization problems for PDEs in weak space-time form*, Oberwolfach Rep. EMS Press, 2023. Abstracts from the Oberwolfach workshop held March 5–10, 2023, joint with U. Langer, R. Löscher, F. Tröltzsch, H. Yang, M. Zank.

[37] F. Tröltzsch. *Optimal control of partial differential equations: Theory, methods and applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2010.