Optimal Control Problems with Control Complementarity Constraints

Christian Clason, Yu Deng, Patrick Mehlitz, Uwe Prüfert

SPP 1962 | Non-smooth and Complementarity-based
Distributed Parameter Systems:
Simulation and Hierarchical Optimization

Optimal control problems with control complementarity constraints
Existence results, optimality conditions, and a penality method

Christian Clason[*]   Yu Deng[†]   Patrick Mehlitz[‡]   Uwe Prüfert[§]

September 25, 2018

A special class of optimal control problems with complementarity constraints on the control functions is studied. It is shown that such problems possess optimal solutions whenever the underlying control space is a first-order Sobolev space. After deriving necessary optimality conditions of strong stationarity-type, a penalty method based on the Fischer–Burmeister function is suggested and its theoretical properties are analyzed. Finally, the numerical treatment of the problem is discussed and results of computational experiments are presented.

Keywords: Fischer–Burmeister function, Mathematical problems with complementarity constraints, Optimal control, Optimality conditions, Penalty method

MSC: 49J20, 49K20, 49M05, 49M25, 90C48

1. Introduction

Complementarity conditions appear in many mathematical optimization problems arising from real-world applications, and this phenomenon is not restricted to the finite-dimensional setting, see Luo et al. [1996], Ulbrich [2011], Wachsmuth [2015] and the references therein. A prominent example for a complementarity problem in function spaces is the optimal control of the obstacle problem, see Harder and Wachsmuth [2018a] for an overview of existing literature. Mathematical problems with complementarity constraints (MPCCs) suffer from an inherent lack of regularity,

---

[*]Universität Duisburg-Essen, Faculty of Mathematics, 45117 Essen, Germany, christian.clason@uni-due.de, https://udue.de/clason

[†]Technische Universität Bergakademie Freiberg, Faculty of Mathematics and Computer Science, 09596 Freiberg, Germany, yu.deng@math.tu-freiberg.de, http://www.mathe.tu-freiberg.de/nmo/mitarbeiter/yu-deng

[‡]Brandenburgische Technische Universität Cottbus-Senftenberg, Institute of Mathematics, Chair of Optimal Control, 03046 Cottbus, Germany, mehlitz@b-tu.de, https://www.b-tu.de/fg-optimale-steuerung/team/dr-patrick-mehlitz

[§]Technische Universität Bergakademie Freiberg, Faculty of Mathematics and Computer Science, 09596 Freiberg, Germany, uwe.pruefert@math.tu-freiberg.de, http://www.mathe.tu-freiberg.de/nmo/mitarbeiter/uwe-pruefert

1

see [Ye et al., 1997, Proposition 1.1] and [Mehlitz and Wachsmuth, 2016, Lemma 3.1] for the finite- and infinite-dimensional situation, respectively, which is why the construction of suitable optimality conditions, constraint qualifications, and numerical methods is a challenging task. Using so-called *NCP functions*, complementarity constraints can be transformed into possibly nonsmooth equality constraints that can be handled by, e.g., Newton-type methods, see De Luca et al. [2000], Facchinei et al. [1998], Ulbrich [2011] and the references therein. A satisfying overview of NCP functions can be found in Sun and Qi [1999]. One of the most popular NCP-functions is the so-called *Fischer–Burmeister function* $\phi \colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$\forall a, b \in \mathbb{R}\colon \quad \phi(a, b) := \sqrt{a^2 + b^2} - a - b, \tag{1.1}$$

see Fischer [1992]. Obviously, one has

$$\forall a, b \in \mathbb{R}\colon \quad \phi(a, b) = 0 \iff a \geq 0 \ \wedge \ b \geq 0 \ \wedge \ ab = 0,$$

which (by definition) holds for all NCP functions. Thus, NCP-functions allow the replacement of a complementarity condition by a single equality constraint. In Ulbrich [2011], it is shown that NCP-functions can be applied to solve complementarity problems in function space settings as well.

In this paper, an optimal control problem with complementarity constraints on the control functions is studied. Control complementarity constraints have been the subject of several recent papers including Clarke and de Pinho [2010], Guo and Ye [2016], Mehlitz and Wachsmuth [2018], Pang and Stewart [2008]. Classically, such constraints arise from reformulating a bilevel optimal control problem with lower level control constraints as a single-level problem using lower level first-order optimality conditions, see [Mehlitz and Wachsmuth, 2016, Section 5]. On the other hand, control complementarity constraints are closely related to switching conditions on the control functions, see Clason et al. [2016a,b, 2017] and the references therein. Here, it will be shown that such problems possess an optimal solution if the control space is taken as $H^1(\Omega)$. Recently, optimal control problems with control constraints in first-order Sobolev spaces were studied in Deng et al. [2018a,b].

It will also be demonstrated that the Fischer–Burmeister function can be used to design penalty methods that can be exploited to find minimizers of the corresponding optimal control problem. One major advantage of this procedure is that the resulting penalized problems are unconstrained. In contrast, simply penalizing the equilibrium condition and leaving the non-negativity conditions in the constraints would lead to the appearance of Lagrange multipliers from $H^1(\Omega)^\star$ in the necessary optimality conditions of the penalized problems, which would cause some theoretical and numerical difficulties due to the presumed high regularity of the control space, see Deng et al. [2018b].

The paper is organized as follows: In the remainder of this section, the basic notation is introduced. Afterwards, the optimal control problem is formally stated and the existence of solutions is discussed in Section 2. Necessary optimality conditions of strong stationarity-type are derived in Section 3. Section 4 is dedicated to the theoretical investigation of a penalization procedure for the H^1-setting. The practical implementation of the proposed numerical method and some corresponding examples are discussed in Section 5 and Section 6, respectively.

Basic notation For a Banach space $\mathcal{X}$, $\|\cdot\|_{\mathcal{X}}$ denotes its norm. The expression $\mathcal{X}^{\star}$ is used to represent the topological dual space of $\mathcal{X}$. Let $\langle\cdot,\cdot\rangle_{\mathcal{X}} : \mathcal{X}^{\star} \times \mathcal{X} \to \mathbb{R}$ be the associated dual pairing. For another Banach space $\mathcal{Y}$, $\mathbb{L}[\mathcal{X},\mathcal{Y}]$ represents the Banach space of all bounded, linear operators which map from $\mathcal{X}$ to $\mathcal{Y}$. For $F \in \mathbb{L}[\mathcal{X},\mathcal{Y}]$, $F^{\star} \in \mathbb{L}[\mathcal{Y}^{\star},\mathcal{X}^{\star}]$ denotes its adjoint. If $\mathcal{X} \subset \mathcal{Y}$ holds true while the associated identity mapping from $\mathcal{X}$ into $\mathcal{Y}$ is continuous, then $\mathcal{X}$ is said to be continuously embedded into $\mathcal{Y}$, denoted by $\mathcal{X} \hookrightarrow \mathcal{Y}$.

Recall that a set $A \subset \mathcal{X}$ is said to be weakly sequentially closed if all the limit points of weakly convergent sequences contained in A belong to A as well, and that any closed, convex set is weakly sequentially closed by Mazur's lemma. For any $A \subset \mathcal{X}$, define the polar cone

$$A^{\circ} := \left\{ x^{\star} \in \mathcal{X}^{\star} \,\middle|\, \forall x \in A\colon \ \langle x^{\star}, x\rangle_{\mathcal{X}} \leq 0 \right\},$$

as well as the annihilator

$$A^{\perp} := \left\{ x^{\star} \in \mathcal{X}^{\star} \,\middle|\, \forall x \in A\colon \ \langle x^{\star}, x\rangle_{\mathcal{X}} = 0 \right\}.$$

By definition, $A^{\perp} = A^{\circ} \cap (-A)^{\circ}$ holds true. It is well known that A° is a nonempty, closed, convex cone while $A^{\perp}$ is a closed subspace of $\mathcal{X}^{\star}$. For an arbitrary vector $x \in \mathcal{X}$, set $x^{\perp} := \{x\}^{\perp}$ for the sake of brevity.

Finally, if a function $F\colon \mathcal{X} \to \mathcal{Y}$ is Fréchet differentiable at $\bar{x} \in \mathcal{X}$, then the bounded, linear operator $F'(\bar{x}) \in \mathbb{L}[\mathcal{X},\mathcal{Y}]$ denotes its Fréchet derivative at $\bar{x}$.

Function spaces For an arbitrary bounded domain $\Omega \subset \mathbb{R}^{d}$ and $p \in [1,\infty]$, $L^{p}(\Omega)$ denotes the usual Lebesgue space of (equivalence classes of) Lebesgue measurable functions mapping from Ω to $\mathbb{R}$, which is equipped with the usual norm. It is well known that for $p \in [1,\infty)$, the space $L^{p}(\Omega)^{\star}$ is isometric to $L^{p'}(\Omega)$ for $p' \in (1,\infty]$ such that $1/p + 1/p' = 1$. The associated dual pairing is given by

$$\forall u \in L^{p}(\Omega)\,\forall v \in L^{p'}(\Omega)\colon \quad \langle v, u\rangle_{L^{p}(\Omega)} := \int_{\Omega} u(x)v(x)\mathrm{d}x.$$

Recall that $L^{2}(\Omega)$ is a Hilbert space which will be identified with $L^{2}(\Omega)$ by means of Riesz' representation theorem. For an arbitrary function $u \in L^{1}(\Omega)$, $\operatorname{supp} u := \{x \in \Omega \,|\, u(x) \neq 0\}$ denotes the support of u. Supposing that $A \subset \Omega$ is a Lebesgue measurable set, $\chi_{A}\colon \Omega \to \mathbb{R}$ represents the characteristic function of A which is 1 for all $x \in A$ and 0 else. Clearly, for a bounded domain Ω and $p \in [1,\infty)$, the relation $\|\chi_{A}\|_{L^{p}(\Omega)} = |A|^{1/p}$ is obtained where $|A|$ denotes the Lebesgue measure of A.

The Banach space of all weakly differentiable functions from $L^{2}(\Omega)$ whose weak derivatives belong to $L^{2}(\Omega)$ is denoted by $H^{1}(\Omega)$. It is equipped with the usual norm

$$\forall y \in H^{1}(\Omega)\colon \quad \|y\|_{H^{1}(\Omega)} := \left(\|y\|_{L^{2}(\Omega)}^{2} + \sum_{i=1}^{d} \|\partial_{x_i} y\|_{L^{2}(\Omega)}^{2} \right)^{1/2}.$$

Clearly, $H^{1}(\Omega)$ is a Hilbert space. However, its dual $H^{1}(\Omega)^{\star}$ will not be identified with $H^{1}(\Omega)$ so that $H^{1}(\Omega)$, $L^{2}(\Omega)$, and $H^{1}(\Omega)^{\star}$ form a so-called Gelfand triple, i.e., they satisfy the relations $H^{1}(\Omega) \hookrightarrow L^{2}(\Omega) \hookrightarrow H^{1}(\Omega)^{\star}$. A detailed study of duality in Sobolev spaces can be found in [Adams and Fournier, 2003, Section 3].

Whenever Ω satisfies the so-called cone condition, see [Adams and Fournier, 2003, Section 4], then the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact, see [Adams and Fournier, 2003, Theorem 6.3]. In this paper, $E \in \mathbb{L}\left[H^1(\Omega), L^2(\Omega)\right]$ is used to denote the latter.

For later use, let $L_+^2(\Omega) \subset L^2(\Omega)$ and $H_+^1(\Omega) \subset H^1(\Omega)$ denote the nonempty, closed, and convex cones of almost everywhere nonnegative functions in $L^2(\Omega)$ and $H^1(\Omega)$, respectively.

2. Problem setting and existence of optimal solutions

In this work, the model complementarity problem

$$\left.\begin{array}{r} \frac{1}{2}\left\|D[y] - y_d\right\|_{\mathcal{D}}^2 + J(u, v) \;\rightarrow\; \min_{y, u, v} \\[4pt] A[y] - B[u] - C[v] \;=\; 0 \\[4pt] (u, v) \;\in\; \mathbb{C} \end{array}\right\} \tag{OC^4}$$

is studied, where for some $\alpha_1, \alpha_2 \geq 0$ and $\varepsilon \geq 0$,

$$\forall u, v \in H^1(\Omega): \quad J(u, v) := \frac{\alpha_1}{2}\|u\|_{L^2(\Omega)}^2 + \frac{\alpha_2}{2}\|v\|_{L^2(\Omega)}^2 + \frac{\varepsilon}{2}\left(\|u\|_{H^1(\Omega)}^2 + \|v\|_{H^1(\Omega)}^2\right),$$

and $\mathbb{C}$ denotes the complementarity set

$$\mathbb{C} := \left\{(w, z) \in H^1(\Omega)^2 \,\middle|\, 0 \leq w(x) \perp z(x) \geq 0 \text{ a.e. on } \Omega\right\}.$$

Observing that A can represent a differential operator, one can interpret (OC^4) as an optimal control problem with complementarity constraints on the control functions that can be used to model switching requirements on the controls. In the context of ordinary differential equations, optimal control problems with mixed control-state complementarity constraints have been studied in Clarke and de Pinho [2010], Guo and Ye [2016], Pang and Stewart [2008] recently. In Harder and Wachsmuth [2018a,b], Mehlitz and Wachsmuth [2018], the interested reader can find some theoretical investigations of optimization problems with complementarity constraints with respect to $L_+^2(\Omega)$ and $H_+^1(\Omega)$. Recently, optimal control problems with switching constraints related to (OC^4) have been studied in Clason et al. [2016b, 2017].

For the remainder of this work, the following standing assumptions on the problem (OC^4) are postulated.

Assumption 2.1. *The domain $\Omega \subset \mathbb{R}^d$ is nonempty, bounded, and satisfies the cone condition. Let the observation space $\mathcal{D}$ as well as the state space $\mathcal{Y}$ be Hilbert spaces. The target $y_d \in \mathcal{D}$ will be fixed. The operator $A \in \mathbb{L}\left[\mathcal{Y}, \mathcal{Y}^\star\right]$ is an isomorphism while $B, C \in \mathbb{L}\left[H^1(\Omega), \mathcal{Y}^\star\right]$ and $D \in \mathbb{L}[\mathcal{Y}, \mathcal{D}]$ are arbitrarily chosen. Finally, $\varepsilon > 0$ holds.*

Let $S \in \mathbb{L}\left[H^1(\Omega)^2, \mathcal{D}\right]$ be the control-to-observation operator which maps any pair of controls $(u, v) \in H^1(\Omega)^2$ to $D[y]$, where $y \in \mathcal{Y}$ is the associated uniquely determined solution of the state equation

$$A[y] - B[u] - C[v] = 0.$$

Then, S is a well-defined continuous linear operator since A is assumed to be an isomorphism.

In the following, the existence of optimal solutions to (OC^4) is discussed. First, the overall H^1-setting needed for the further theoretical treatment of (OC^4) is analyzed in Section 2.1. Some comments on the setting where controls come from $L^2(\Omega)$ are presented in Section 2.2.

2.1. First-order Sobolev spaces

Since the objective function of (OC^4) is continuously Fréchet differentiable, convex, and bounded from below, the only critical point for existence is the weak sequential closedness of the complementarity set $\mathbb{C}$.

Lemma 2.2. *The set $\mathbb{C}$ is closed.*

Proof. Let $\{(u_k, v_k)\}_{k \in \mathbb{N}} \subset \mathbb{C}$ be a sequence converging to $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$. Due to the continuity of the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$, the strong convergences $u_k \to \bar{u}$ and $v_k \to \bar{v}$ hold in $L^2(\Omega)$. Noting that the cone $L^2_+(\Omega)$ is closed in $L^2(\Omega)$, the relations $\{u_k\}_{k \in \mathbb{N}}$, $\{v_k\}_{k \in \mathbb{N}} \subset L^2_+(\Omega)$ yield $\bar{u}, \bar{v} \in L^2_+(\Omega)$.

Hence, by supposing $(\bar{u}, \bar{v}) \notin \mathbb{C}$, there is a measurable set $\Omega' \subset \Omega$ of positive measure and some $\epsilon > 0$ such that $\bar{u}(x)\bar{v}(x) \geq \epsilon$ holds for almost all $x \in \Omega'$. Noting that $\bar{u}\bar{v} \in L^1(\Omega)$ holds true, $\|\bar{u}\bar{v}\|_{L^1(\Omega)} \geq \epsilon |\Omega'| > 0$ follows. On the other hand, Hölder's inequality and the boundedness of $\{u_k\}_{k \in \mathbb{N}}$ in $L^2(\Omega)$ yield that $\{u_k v_k\}_{k \in \mathbb{N}} \subset L^1(\Omega)$ converges in $L^1(\Omega)$ to $\bar{u}\bar{v}$ since

$$\|u_k v_k - \bar{u}\bar{v}\|_{L^1(\Omega)} \leq \|u_k(v_k - \bar{v})\|_{L^1(\Omega)} + \|(u_k - \bar{u})\bar{v}\|_{L^1(\Omega)}$$

$$\leq \|u_k\|_{L^2(\Omega)} \|v_k - \bar{v}\|_{L^2(\Omega)} + \|u_k - \bar{u}\|_{L^2(\Omega)} \|\bar{v}\|_{L^2(\Omega)} \xrightarrow{k \to \infty} 0.$$

However, due to $\|u_k v_k\|_{L^1(\Omega)} = 0$ for all $k \in \mathbb{N}$, the assumption $\|\bar{u}\bar{v}\|_{L^1(\Omega)} > 0$ cannot be satisfied. Thus, $(\bar{u}, \bar{v}) \in \mathbb{C}$, and, consequently, $\mathbb{C}$ is closed. $\qquad\square$

Although $\mathbb{C}$ is a nonconvex set, the compactness of the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ can be used in order to show that $\mathbb{C}$ is weakly sequentially closed.

Lemma 2.3. *The set $\mathbb{C}$ is weakly sequentially closed.*

Proof. First, a similar proof as for Lemma 2.2 shows that the complementarity set in $L^2(\Omega)$ given by

$$\widetilde{\mathbb{C}} := \left\{ (w, z) \in L^2(\Omega)^2 \,\middle|\, 0 \leq w(x) \perp z(x) \geq 0 \text{ a.e. on } \Omega \right\}$$
$$= \left\{ (w, z) \in L^2_+(\Omega)^2 \,\middle|\, \langle w, z \rangle_{L^2(\Omega)} = 0 \right\} \tag{2.1}$$

is closed as well.

Next, choose a sequence $\{(u_k, v_k)\}_{k \in \mathbb{N}} \subset \mathbb{C}$ converging weakly to $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$. Exploiting $u_k \rightharpoonup \bar{u}$ and $v_k \rightharpoonup \bar{v}$ as well as the compactness of the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$, there is a subsequence of $\{(u_k, v_k)\}_{k \in \mathbb{N}}$ that converges strongly to $(\bar{u}, \bar{v})$ in $L^2(\Omega)^2$. Due to the closedness of $\widetilde{\mathbb{C}}$ in $L^2(\Omega)^2$, $(\bar{u}, \bar{v}) \in \widetilde{\mathbb{C}} \cap H^1(\Omega)^2$ holds, and, consequently, $(\bar{u}, \bar{v})$ is already an element of $\mathbb{C}$. Thus, $\mathbb{C}$ is weakly sequentially closed. $\qquad\square$

As a corollary, the existence of optimal solutions to (OC^4) is obtained.

Corollary 2.4. *The problem (OC^4) possesses an optimal solution.*

Proof. The objective functional of (OC^4) is continuously Fréchet differentiable, convex, and (due to $\varepsilon > 0$) coercive. Furthermore, by Lemma 2.3, the complementarity set $\mathbb{C}$ is weakly sequentially closed, and so is the feasible set induced by the PDE constraint. Hence, the claim follows by application of Tonelli's direct method. $\qquad\square$

2.2. Lebesgue spaces

In the remainder of this section, the existence of optimal controls in $L^2(\Omega)$ is investigated. In this case, the corresponding model problem is given by

$$\left.\begin{aligned}
\tfrac{1}{2}\,\|\mathrm{D}[y] - y_\mathrm{d}\|_{\mathcal{D}}^2 + \tfrac{\alpha_1}{2}\,\|u\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2}\,\|v\|_{L^2(\Omega)}^2 \;\rightarrow\; \min_{y,u,v} \\
\mathrm{A}[y] - \tilde{\mathrm{B}}[u] - \tilde{\mathrm{C}}[v] \;=\; 0 \\
(u,v) \in \widetilde{\mathbb{C}}
\end{aligned}\right\} \tag{OC_{L^2}}$$

where the complementarity set $\widetilde{\mathbb{C}}$ has been defined in (2.1). Furthermore, $\tilde{\mathrm{B}}, \tilde{\mathrm{C}} \in \mathbb{L}\left[L^2(\Omega), \mathcal{Y}^\star\right]$ need to be chosen. As already shown in the proof of Lemma 2.3, $\widetilde{\mathbb{C}}$ is closed. However, $\widetilde{\mathbb{C}}$ is in general *not* weakly sequentially closed, as the following example shows.

Example 2.5. *For any $k \in \mathbb{N}$, define the two open sets*

$$P_k := \left\{ x \in \mathbb{R}^d \,\middle|\, \textstyle\prod_{j=1}^d \sin(k\pi x_j) > 0 \right\},$$

$$Q_k := \left\{ x \in \mathbb{R}^d \,\middle|\, \textstyle\prod_{j=1}^d \sin(k\pi x_j) < 0 \right\}.$$

Now, set $u_k := \chi_{\Omega \cap P_k}$ and $v_k := \chi_{\Omega \cap Q_k}$. Obviously, $(u_k, v_k) \in \widetilde{\mathbb{C}}$ holds true for all $k \in \mathbb{N}$. Furthermore, the sequence $\{(u_k, v_k)\}_{k \in \mathbb{N}} \subset L^2(\Omega)^2$ converges weakly to the point $(\tfrac{1}{2}\chi_\Omega, \tfrac{1}{2}\chi_\Omega)$, which does not belong to $\widetilde{\mathbb{C}}$. Thus, $\widetilde{\mathbb{C}}$ is not weakly sequentially closed.

It may still happen that there exists an optimal solution of the complementarity-constrained problem (OC_{L^2}), as illustrated by the following example. For $\mathcal{D} := L^2(\Omega)$ and $\mathcal{Y} := H^1(\Omega)$, consider the elliptic optimal control problem

$$\left.\begin{aligned}
\tfrac{1}{2}\,\|\mathrm{E}[y] - y_\mathrm{d}\|_{L^2(\Omega)}^2 + \tfrac{\alpha_1}{2}\,\|u\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2}\,\|v\|_{L^2(\Omega)}^2 \;\rightarrow\; \min_{y,u,v} \\
-\nabla \cdot (\mathbf{C}\nabla y) + \mathbf{a}y \;=\; \chi_{\Omega_u} u + \chi_{\Omega_v} v \qquad \text{a.e. on } \Omega \\
\vec{\mathbf{n}} \cdot (\mathbf{C}\nabla y) + \mathbf{q}y \;=\; 0 \qquad \text{a.e. on } \mathrm{bd}\,\Omega \\
(u,v) \in \widetilde{\mathbb{C}}
\end{aligned}\right\} \tag{2.2}$$

where $\alpha_1, \alpha_2 > 0$ are constants, $\mathbf{C} \in L^\infty(\Omega; S^d(\mathbb{R}))$ (where $S^d(\mathbb{R})$ denotes the set of real symmetric $d \times d$ matrices) satisfies the condition of uniform ellipticity, i.e.,

$$\exists c_0 > 0 \;\forall x \in \Omega \;\forall \xi \in \mathbb{R}^d : \quad \xi^\top \mathbf{C}(x)\xi \geq c_0 |\xi|_2^2. \tag{2.3}$$

Moreover, $\mathbf{a} \in L^\infty(\Omega)$ and $\mathbf{q} \in L^\infty(\mathrm{bd}\,\Omega)$ are nonnegative and satisfy $\|\mathbf{a}\|_{L^\infty(\Omega)} + \|\mathbf{q}\|_{L^\infty(\mathrm{bd}\,\Omega)} > 0$, and $\Omega_u, \Omega_v \subset \Omega$ are measurable sets of positive measure satisfying $\Omega_u \cup \Omega_v = \Omega$. Here, the PDE constraint is interpreted in the weak sense. It is well known that the associated differential operator A is elliptic, see [Evans, 2010, Section 6], and, thus, an isomorphism.

Proposition 2.6. *The problem (2.2) possesses an optimal solution.*

Proof. Assume without loss of generality that $\alpha_1 \leq \alpha_2$; the other case can be handled analogously. Consider then the surrogate optimal control problem

$$
\left.
\begin{aligned}
\tfrac{1}{2} \left\| E[y] - y_d \right\|_{L^2(\Omega)}^2 + \tfrac{1}{2} \left\| \left(\sqrt{\alpha_1} \chi_{\Omega_u} + \sqrt{\alpha_2} \chi_{\Omega_v \setminus \Omega_u} \right) z \right\|_{L^2(\Omega)}^2 \;\rightarrow\; \min_{y, z} \\
-\nabla \cdot (\mathbf{C}\nabla y) + \mathbf{a}y \;=\; z \qquad \text{a.e. on } \Omega \\
\vec{\mathbf{n}} \cdot (\mathbf{C}\nabla y) + \mathbf{q}y \;=\; 0 \qquad \text{a.e. on bd } \Omega \\
z \;\in\; L_+^2(\Omega).
\end{aligned}
\right\} \quad (2.4)
$$

Note that its objective is equivalent to

$$
H^1(\Omega) \times L^2(\Omega) \ni (y, z) \mapsto \tfrac{1}{2} \left\| E[y] - y_d \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_1}{2} \left\| \chi_{\Omega_u} z \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| \chi_{\Omega_v \setminus \Omega_u} z \right\|_{L^2(\Omega)}^2 \in \mathbb{R}.
$$

The ellipticity of the underlying PDE in (2.4) implies that the associated control-to-observation operator $\check{S} \colon L^2(\Omega) \rightarrow L^2(\Omega)$ is linear and continuous, see [Evans, 2010, Section 6.2]. Observing that $\Omega_u \cup \Omega_v = \Omega$ holds by assumption, the reduced objective functional

$$
L^2(\Omega) \ni z \mapsto \tfrac{1}{2} \left\| \check{S}[z] - y_d \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_1}{2} \left\| \chi_{\Omega_u} z \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| \chi_{\Omega_v \setminus \Omega_u} z \right\|_{L^2(\Omega)}^2 \in \mathbb{R}
$$

is convex, continuous, and coercive. This shows that the optimal control problem (2.4) possesses an optimal solution $(\bar{y}, \bar{z}) \in H^1(\Omega) \times L^2(\Omega)$ with objective value $\bar{m} \in \mathbb{R}$.

Let $(y, u, v) \in H^1(\Omega) \times L^2(\Omega) \times L^2(\Omega)$ be feasible to (2.2). Defining $z := \chi_{\Omega_u} u + \chi_{\Omega_v \setminus \Omega_u} v$, (y, z) is feasible for (2.4). Then, the estimate

$$
\begin{aligned}
\tfrac{1}{2} \left\| E[y] - y_d \right\|_{L^2(\Omega)}^2 &+ \tfrac{\alpha_1}{2} \left\| u \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| v \right\|_{L^2(\Omega)}^2 \\
&\geq \tfrac{1}{2} \left\| E[y] - y_d \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_1}{2} \left\| \chi_{\Omega_u} u \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| \chi_{\Omega_v \setminus \Omega_u} v \right\|_{L^2(\Omega)}^2 \\
&= \tfrac{1}{2} \left\| E[y] - y_d \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_1}{2} \left\| \chi_{\Omega_u} z \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| \chi_{\Omega_v \setminus \Omega_u} z \right\|_{L^2(\Omega)}^2 \geq \bar{m}
\end{aligned}
$$

is obtained. In particular, the objective value of (2.2) is bounded from below by $\bar{m}$.

Define $\bar{u} := \chi_{\Omega_u} \bar{z}$ and $\bar{v} := \chi_{\Omega_v \setminus \Omega_u} \bar{z}$. Then, $(\bar{y}, \bar{u}, \bar{v})$ is feasible to (2.2) since $\bar{y}$ is the state associated with $\bar{z}$ and $\chi_{\Omega_u} \bar{u} + \chi_{\Omega_v} \bar{v} = \bar{z}$ holds true. Moreover, the relation

$$
\begin{aligned}
\tfrac{1}{2} \left\| E[\bar{y}] - y_d \right\|_{L^2(\Omega)}^2 &+ \tfrac{\alpha_1}{2} \left\| \bar{u} \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| \bar{v} \right\|_{L^2(\Omega)}^2 \\
&= \tfrac{1}{2} \left\| E[\bar{y}] - y_d \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_1}{2} \left\| \chi_{\Omega_u} \bar{z} \right\|_{L^2(\Omega)}^2 + \tfrac{\alpha_2}{2} \left\| \chi_{\Omega_v \setminus \Omega_u} \bar{z} \right\|_{L^2(\Omega)}^2 = \bar{m}
\end{aligned}
$$

follows. Thus, $(\bar{y}, \bar{u}, \bar{v})$ is an optimal solution of (2.2). $\qquad\square$

Note that the proof of Proposition 2.6 yields a strategy for the solution of (2.2) by means of standard arguments from optimal control by solving the surrogate problem (2.4).

3. Optimality conditions

Consider the so-called state-*reduced* problem

$$
\left.
\begin{aligned}
\tfrac{1}{2} \left\| S[u, v] - y_d \right\|_{\mathcal{D}}^2 + J(u, v) \;\rightarrow\; \min_{u, v} \\
(u, v) \;\in\; \mathbb{C}
\end{aligned}
\right\} \quad (3.1)
$$

which is equivalent to (OC[4]) by definition of the control-to-observation operator S. Using the embedding operator $E \colon H^1(\Omega) \to L^2(\Omega)$, (3.1) can be stated equivalently as

$$
\left.
\begin{aligned}
\tfrac{1}{2} \, \|S[u,v] - y_{\mathrm{d}}\|_{\mathcal{D}}^2 + J(u,v) \; &\to \; \min_{u,v} \\
E[u] \; &\in \; L_+^2(\Omega) \\
E[v] \; &\in \; L_+^2(\Omega) \\
\langle E[u], E[v] \rangle_{L^2(\Omega)} \; &= \; 0
\end{aligned}
\right\}
\tag{3.2}
$$

which is a generalized MPCC in the Banach space $L^2(\Omega)$. It was shown in [Mehlitz and Wachsmuth, 2016, Lemma 3.1] that Robinson's constraint qualification does not hold at the feasible points of this problem. Moreover, since E is not surjective, the constraint qualifications needed to show that local optimal solutions of this problem satisfy MPCC-tailored stationarity conditions (e.g., the weak or strong stationarity conditions) are not satisfied, see Mehlitz and Wachsmuth [2016], Wachsmuth [2015] for details.

On the other hand, it is still possible to derive necessary optimality conditions for (3.1) using a standard trick from finite-dimensional MPCC theory: Define appropriate surrogate problems which do not contain a complementarity constraint anymore and thus can be handled with the classical KKT conditions in Banach spaces.

In order to formulate an appropriate surrogate problem, let $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$ be a feasible point of (3.1) and define the measurable sets

$$
I^{+0}(\bar{u}, \bar{v}) := \{x \in \Omega \,|\, \bar{u}(x) > 0 \,\wedge\, \bar{v}(x) = 0\},
\tag{3.3}
$$

$$
I^{0+}(\bar{u}, \bar{v}) := \{x \in \Omega \,|\, \bar{u}(x) = 0 \,\wedge\, \bar{v}(x) > 0\},
\tag{3.4}
$$

$$
I^{00}(\bar{u}, \bar{v}) := \{x \in \Omega \,|\, \bar{u}(x) = 0 \,\wedge\, \bar{v}(x) = 0\}.
\tag{3.5}
$$

If $(\bar{u}, \bar{v})$ is a locally optimal solution of (3.1), then it is also a locally optimal solution of the auxiliary problems

$$
\left.
\begin{aligned}
\tfrac{1}{2} \, \|S[u,v] - y_{\mathrm{d}}\|_{\mathcal{D}}^2 + J(u,v) \; &\to \; \min_{u,v} \\
u \; &\geq \; 0 && \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \\
u \; &= \; 0 && \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\
v \; &\geq \; 0 && \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\
v \; &= \; 0 && \text{a.e. on } I^{+0}(\bar{u}, \bar{v})
\end{aligned}
\right\}
\tag{rNLP$_{\bar{u}}$}
$$

and

$$
\left.
\begin{aligned}
\tfrac{1}{2} \, \|S[u,v] - y_{\mathrm{d}}\|_{\mathcal{D}}^2 + J(u,v) \; &\to \; \min_{u,v} \\
u \; &\geq \; 0 && \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\
u \; &= \; 0 && \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \\
v \; &\geq \; 0 && \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \\
v \; &= \; 0 && \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v})
\end{aligned}
\right\}
\tag{rNLP$_{\bar{v}}$}
$$

since their respective feasible sets are smaller than $\mathbb{C}$ but contain $(\bar{u}, \bar{v})$. By standard notion, see Pang and Fukushima [1999], Scheel and Scholtes [2000], Wachsmuth [2015], (rNLP$_{\bar{u}}$) and (rNLP$_{\bar{v}}$) are referred to as *restricted nonlinear problems*. Furthermore, the corresponding *relaxed nonlinear problem* is introduced by means of

$$
\left.
\begin{aligned}
\tfrac{1}{2}\,\|S[u,v] - y_\mathrm{d}\|_{\mathcal{D}}^2 + J(u,v) \;\to\; &\min_{u,v} \\
u \;\geq\; 0 \qquad &\text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\
u \;=\; 0 \qquad &\text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \\
v \;\geq\; 0 \qquad &\text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\
v \;=\; 0 \qquad &\text{a.e. on } I^{+0}(\bar{u}, \bar{v}).
\end{aligned}
\right\}
\tag{RNLP}
$$

Note that the feasible points $(u, v) \in H^1(\Omega)^2$ of (RNLP) do not necessarily satisfy the complementarity condition $(u, v) \in \mathbb{C}$. Combining standard techniques from finite-dimensional MPCC theory and optimization in Banach spaces, the following result is obtained, see also [Wachsmuth, 2015, Theorems 3.1 and 5.2].

Theorem 3.1. *Let $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$ be a locally optimal solution of* (3.1)*. Then, there exist multipliers $\mu, \nu \in H^1(\Omega)^\star$ satisfying*

$$
0 = S^\star\left[S[\bar{u}, \bar{v}] - y_\mathrm{d}\right] + J'(\bar{u}, \bar{v}) + (\mu, \nu),
\tag{3.6a}
$$

$$
\mu \in \left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} &z \geq 0 \quad \textit{a.e. on } I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\ &z = 0 \quad \textit{a.e. on } I^{0+}(\bar{u}, \bar{v}) \end{aligned} \right\}^\circ,
\tag{3.6b}
$$

$$
\langle \mu, \bar{u} \rangle_{H^1(\Omega)} = 0,
\tag{3.6c}
$$

$$
\nu \in \left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} &z \geq 0 \quad \textit{a.e. on } I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\ &z = 0 \quad \textit{a.e. on } I^{+0}(\bar{u}, \bar{v}) \end{aligned} \right\}^\circ,
\tag{3.6d}
$$

$$
\langle \nu, \bar{v} \rangle_{H^1(\Omega)} = 0.
\tag{3.6e}
$$

Proof. Introducing the cones

$$
\mathcal{K}_{+0} := \left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} &z \geq 0 \quad \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \\ &z = 0 \quad \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \end{aligned} \right\},
$$

$$
\mathcal{K}_{0+, 00} := \left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} &z \geq 0 \quad \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\ &z = 0 \quad \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \end{aligned} \right\},
$$

(rNLP$_{\bar{u}}$) is equivalent to

$$
\left.
\begin{aligned}
\tfrac{1}{2}\,\|S[u,v] - y_\mathrm{d}\|_{\mathcal{D}}^2 + J(u,v) \;\to\; &\min_{u,v} \\
u \;\in\; &\mathcal{K}_{+0} \\
v \;\in\; &\mathcal{K}_{0+, 00}.
\end{aligned}
\right\}
$$

Since $(\bar{u}, \bar{v})$ is a locally optimal solution of (rNLP$_{\bar{u}}$), there exist multipliers $\mu^1, v^1 \in H^1(\Omega)^\star$ which satisfy the corresponding KKT conditions

$$\left.\begin{aligned}
0 &= \mathrm{S}^\star\big[\mathrm{S}[\bar{u}, \bar{v}] - y_{\mathrm{d}}\big] + J'(\bar{u}, \bar{v}) + (\mu^1, v^1), \\
\mu^1 &\in \mathcal{K}_{+0}^\circ \cap \bar{u}^\perp, \\
v^1 &\in \mathcal{K}_{0+,00}^\circ \cap \bar{v}^\perp,
\end{aligned}\right\} \tag{3.7}$$

see [Bonnans and Shapiro, 2000, Theorem 3.9]. Considering (rNLP$_{\bar{v}}$) in a similar way, there exist $\mu^2, v^2 \in H^1(\Omega)^\star$ which satisfy

$$\left.\begin{aligned}
0 &= \mathrm{S}^\star\big[\mathrm{S}[\bar{u}, \bar{v}] - y_{\mathrm{d}}\big] + J'(\bar{u}, \bar{v}) + (\mu^2, v^2), \\
\mu^2 &\in \mathcal{K}_{+0,00}^\circ \cap \bar{u}^\perp, \\
v^2 &\in \mathcal{K}_{0+}^\circ \cap \bar{v}^\perp,
\end{aligned}\right\} \tag{3.8}$$

where

$$\mathcal{K}_{+0,00} := \left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} z &\geq 0 \quad \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\ z &= 0 \quad \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \end{aligned} \right\},$$

$$\mathcal{K}_{0+} := \left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} z &\geq 0 \quad \text{a.e. on } I^{0+}(\bar{u}, \bar{v}) \\ z &= 0 \quad \text{a.e. on } I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \end{aligned} \right\}.$$

Combining the respective first condition in (3.7) and (3.8) yields $\mu^1 = \mu^2$ and $v^1 = v^2$. Since $\mathcal{K}_{+0,00}^\circ \cap \bar{u}^\perp$ is a subset of $\mathcal{K}_{+0}^\circ \cap \bar{u}^\perp$ while $\mathcal{K}_{0+,00}^\circ \cap \bar{v}^\perp$ is a subset of $\mathcal{K}_{0+}^\circ \cap \bar{v}^\perp$, the desired result is obtained by setting $\mu := \mu^2$ and $v := v^1$. $\qquad\square$

Note that the system (3.6) coincides with the KKT conditions of (RNLP). In this regard, it is reasonable to call the conditions (3.6) a *strong stationarity-type system*.

Remark 3.2. *It is difficult to give an explicit characterization of the multipliers $\mu, v \in H^1(\Omega)^\star$. Assume that Ω has a Lipschitz boundary. Introducing $\mathcal{H}_{\mathcal{A}} := \{z \in H^1(\Omega) \mid z = 0 \text{ a.e. on } \mathcal{A}\}$ for a fixed measurable set $\mathcal{A} \subset \Omega$ and using the relation $H_+^1(\Omega)^\circ = H^1(\Omega)^\star \cap \mathcal{M}_-(\overline{\Omega})$, see [Bonnans and Shapiro, 2000, Section 6.4.3], it holds that*

$$\mu \in \big(H_+^1(\Omega) \cap \mathcal{H}_{I^{0+}(\bar{u}, \bar{v})}\big)^\circ = \mathrm{cl}\left(H^1(\Omega)^\star \cap \mathcal{M}_-(\overline{\Omega}) + \mathcal{H}_{I^{0+}(\bar{u}, \bar{v})}^\perp\right)$$

where $\mathcal{M}_-(\overline{\Omega})$ denotes the set of all finite, nonpositive Borel measures on $\overline{\Omega}$. A similar result can be obtained to characterize v. However, due to the appearance of the closure as well as the annihilated subspace associated with $\mathcal{H}_{I^{0+}(\bar{u}, \bar{v})}$, this characterization is of limited practical use; in particular, it cannot be deduced that μ and v are measures. Applying the machinery of capacity theory, see Attouch et al. [2006], Bonnans and Shapiro [2000], a more advanced approach to the characterization of μ and v can be attempted. For this purpose, one could strengthen the constraints in (rNLP$_{\bar{u}}$), (rNLP$_{\bar{v}}$), and (RNLP) to hold quasi-everywhere on the respective subdomains, i.e., the respective conditions hold up to sets of H^1-capacity zero. Then, one needs to find explicit expressions for the polar cone associated with sets of type

$$\left\{ z \in H^1(\Omega) \,\middle|\, \begin{aligned} z &\geq 0 \quad \text{quasi-everywhere on } \mathcal{A} \\ z &= 0 \quad \text{quasi-everywhere on } \Omega \setminus \mathcal{A} \end{aligned} \right\}$$

where $\mathcal{A} \subset \Omega$ is measurable.

In order to state necessary optimality conditions of strong stationarity-type that avoid the appearance of multipliers and allow a numerical implementation, one can exploit the definition of the polar cone in the system (3.6).

Corollary 3.3. *Let $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$ be a locally optimal solution of (3.1). Then, the condition*

$$0 = \langle S[\bar{u}, \bar{v}] - y_d, S[\bar{u}, \bar{v}] \rangle_{\mathcal{D}^2} + J'(\bar{u}, \bar{v})[\bar{u}, \bar{v}]$$

holds, and for any pair $(z_u, z_v) \in H^1_+(\Omega) \times H^1_+(\Omega)$,

$$\left.\begin{array}{l} \operatorname{supp} z_u \subset I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \\ \operatorname{supp} z_v \subset I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \end{array}\right\} \implies \left\langle S^\star \left[S[\bar{u}, \bar{v}] - y_d \right] + J'(\bar{u}, \bar{v}), (z_u, z_v) \right\rangle_{H^1(\Omega)^2} \geq 0.$$

Proof. Due to Theorem 3.1, there exist $\mu, \nu \in H^1(\Omega)^\star$ satisfying (3.6). Testing (3.6a) with $(\bar{u}, \bar{v})$ while exploiting (3.6c), (3.6e), and the definition of the adjoint operator, the first statement of the corollary follows.

The second statement is a consequence of (3.6a), (3.6b), and (3.6d). $\qquad\square$

Remark 3.4. *According to standard terminology for MPCCs, the necessary optimality conditions (3.6) are of strong stationarity-type, see, e.g., [Wachsmuth, 2015, Definition 5.1] and [Ye, 2005, Definition 2.7]. Recall that a feasible point $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$ of (3.1) and thus of (3.2) is a strongly stationary point of (3.2) in the sense of [Wachsmuth, 2015, Definition 5.1] if and only if there are multipliers $(\mu, \nu) \in L^2(\Omega)^2$ satisfying*

$$0 = S^\star \left[S[\bar{u}, \bar{v}] - y_d \right] + J'(\bar{u}, \bar{v}) + (E, E)^\star[\mu, \nu], \tag{3.9a}$$

$$\mu = 0 \quad \textit{a.e. on } I^{+0}(\bar{u}, \bar{v}), \tag{3.9b}$$

$$\nu = 0 \quad \textit{a.e. on } I^{0+}(\bar{u}, \bar{v}), \tag{3.9c}$$

$$\mu \leq 0 \wedge \nu \leq 0 \quad \textit{a.e. on } I^{00}(\bar{u}, \bar{v}), \tag{3.9d}$$

see also [Mehlitz and Wachsmuth, 2018, Definition 4.1]. If $\mathbb{C}$ is replaced by $\widetilde{\mathbb{C}}$ and $\varepsilon = 0$ (in this case, E is the identity mapping), the systems (3.6) and (3.9) are equivalent. However, for $\mathbb{C}$ and $\varepsilon > 0$, the necessary optimality conditions (3.6) are weaker than (3.9), which can be seen as follows: It is clear that whenever $(\tilde{\mu}, \tilde{\nu}) \in L^2(\Omega)^2$ satisfy the classical strong stationarity conditions (3.9), then the multipliers $\mu := E^\star[\tilde{\mu}]$ and $\nu := E^\star[\tilde{\nu}]$ satisfy (3.6). On the other hand, by means of Theorem 3.1, the multipliers appearing in the system (3.6) may come from $H^1(\Omega)^\star \setminus L^2(\Omega)$ in general.

Remark 3.5. *In this section, only the property of S to be a bounded, linear operator has been exploited. Thus, the optimality conditions obtained in Theorem 3.1 and Corollary 3.3 are applicable in many different situations, e.g., in case where S is the control-to-observation operator associated with a linear elliptic equation where u and v only operate on some subdomain, or for a linear parabolic equation where the controls u and v only depend on time. The latter problems are closely related to the switching-constrained problems examined in Clason et al. [2016a,b, 2017].*

It should be noted that similar necessary optimality conditions can be derived if $S: H^1(\Omega)^2 \to \mathcal{D}$ is Fréchet differentiable but not necessarily linear.

4. Penalization of complementarity constraints

In order to find optimal solutions of (OC^4), an obvious idea would be to penalize the violation of the equilibrium condition

$$u(x)v(x) = 0 \quad \text{a.e. on } \Omega \tag{4.1}$$

in (OC^4). This is related to the approaches used in Clason et al. [2016a,b, 2017] for the treatment of switching-constrained optimal control problems. However, the resulting penalized problem would still involve inequality constraints for the controls in $H^1(\Omega)$, and thus the associated KKT conditions would involve Lagrange multipliers from $H^1(\Omega)^\star$, see [Deng et al., 2018b, Section 5] for details. This, however, may provoke theoretical and numerical difficulties that should be avoided here.

To get around these issues, the penalization of the overall complementarity constraint using the Fischer–Burmeister function is proposed here, which leads to penalized problems in which the only constraint is the state equation.

4.1. Penalty term

Let $\phi \colon \mathbb{R}^2 \to \mathbb{R}$ denote the Fischer–Burmeister function introduced in (1.1) and let the mapping $\Phi \colon L^2(\Omega)^2 \to L^2(\Omega)$ be the associated Nemytskii operator defined by

$$\forall (w, z) \in L^2(\Omega)^2 \, \forall x \in \Omega \colon \quad \Phi(w, z)(x) := \phi(w(x), z(x)).$$

This operator is well-defined since

$$\begin{aligned}
\|\Phi(w, z)\|_{L^2(\Omega)} &\leq \left(\int_\Omega \left(w^2(x) + z^2(x) \right) \mathrm{d}x \right)^{1/2} + \|w\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)} \\
&\leq \left(\int_\Omega \left(|w(x)| + |z(x)| \right)^2 \mathrm{d}x \right)^{1/2} + \|w\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)} \\
&\leq 2 \left(\|w\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)} \right) < +\infty,
\end{aligned}$$

i.e., Φ maps from $L^2(\Omega)^2$ to $L^2(\Omega)$, see also [Ulbrich, 2011, Section 3.3].

The violation of the complementarity constraint $(u, v) \in \mathbb{C}$ can then be penalized using the functional $F \colon H^1(\Omega)^2 \to \mathbb{R}_0^+$ defined by

$$\forall (u, v) \in H^1(\Omega)^2 \colon \quad F(u, v) := \tfrac{1}{2} \int_\Omega \phi^2(u(x), v(x)) \mathrm{d}x = \tfrac{1}{2} \|\Phi(\mathrm{E}[u], \mathrm{E}[v])\|_{L^2(\Omega)}^2. \tag{4.2}$$

Recall that $\mathrm{E} \in \mathbb{L}\left[H^1(\Omega), L^2(\Omega) \right]$ represents the natural embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$.

It is obvious that Φ cannot be Fréchet differentiable since ϕ is not smooth. In contrast, F is a continuously Fréchet differentiable mapping.

Lemma 4.1. *Let $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$ be arbitrary. Then, F is continuously Fréchet differentiable at $(\bar{u}, \bar{v})$. The associated Fréchet derivative is given by*

$$\forall (\delta^u, \delta^v) \in H^1(\Omega)^2 \colon \quad F'(\bar{u}, \bar{v})[\delta^u, \delta^v] = \int_\Omega \phi(\bar{u}(x), \bar{v}(x)) \left(\eta_{\bar{u}}(x) \delta^u(x) + \eta_{\bar{v}}(x) \delta^v(x) \right) \mathrm{d}x,$$

where $\eta_{\bar{u}}, \eta_{\bar{v}} \in L^\infty(\Omega)$ are defined by

$$\forall x \in \Omega: \quad \eta_{\bar{u}}(x) = \begin{cases} \dfrac{\bar{u}(x)}{\sqrt{\bar{u}(x)^2 + \bar{v}(x)^2}} - 1 & \text{if } x \notin I^{00}(\bar{u}, \bar{v}), \\ 0 & \text{if } x \in I^{00}(\bar{u}, \bar{v}), \end{cases} \tag{4.3a}$$

$$\forall x \in \Omega: \quad \eta_{\bar{v}}(x) = \begin{cases} \dfrac{\bar{v}(x)}{\sqrt{\bar{u}(x)^2 + \bar{v}(x)^2}} - 1 & \text{if } x \notin I^{00}(\bar{u}, \bar{v}), \\ 0 & \text{if } x \in I^{00}(\bar{u}, \bar{v}), \end{cases} \tag{4.3b}$$

and $I^{00}(\bar{u}, \bar{v})$ is defined by (3.5).

Proof. Let $f \colon \mathbb{R}^2 \to \mathbb{R}$ be given by

$$\forall (a, b) \in \mathbb{R}^2: \quad f(a, b) := \tfrac{1}{2}\phi(a, b)^2.$$

One can check that f is continuously differentiable with gradient

$$\forall (a, b) \in \mathbb{R}^2: \quad \nabla f(a, b) = \begin{cases} \phi(a, b) \begin{pmatrix} \frac{a}{\sqrt{a^2 + b^2}} - 1 \\ \frac{b}{\sqrt{a^2 + b^2}} - 1 \end{pmatrix} & \text{if } (a, b) \neq (0, 0), \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{if } (a, b) = (0, 0). \end{cases}$$

Clearly, the Nemytskii-operator $\mathcal{F}$ associated with f maps from $L^2(\Omega)^2$ to $L^1(\Omega)$, since Φ maps $L^2(\Omega)^2$ to $L^2(\Omega)$. Noting that $a/\sqrt{a^2 + b^2} \in [-1, 1]$ and $b/\sqrt{a^2 + b^2} \in [-1, 1]$ hold for all $(a, b) \in \mathbb{R}^2 \setminus \{(0, 0)\}$, the Nemytskii operator associated with ∇f maps from $L^2(\Omega)^2$ to $L^2(\Omega)$. Applying [Goldberg et al., 1992, Theorems 4 and 7], $\mathcal{F} \colon L^2(\Omega)^2 \to L^1(\Omega)$ is continuously Fréchet differentiable. Furthermore,

$$\forall x \in \Omega: \quad \mathcal{F}'(w, z)[\delta_w, \delta_z](x) = \nabla_a f(w(x), z(x))\delta_w(x) + \nabla_b f(w(x), z(x))\delta_w(x)$$

for any $(w, z), (\delta_w, \delta_z) \in L^2(\Omega)^2$.

Define $\mathrm{L} \in \mathbb{L}\left[L^1(\Omega), \mathbb{R}\right]$ by $\mathrm{L}[w] := \int_\Omega w(x)\mathrm{d}x$. Then, $F = \mathrm{L} \circ \mathcal{F} \circ (\mathrm{E}, \mathrm{E})$. Since all involved mappings are continuously Fréchet differentiable, the assertion of the lemma follows by exploiting the chain rule for Fréchet differentiable functions, see [Tröltzsch, 2010, Theorem 2.20]. □

Remark 4.2. *As the penalty functional F is smooth, it cannot lead to exact penalization of the complementarity constraints. Although* Section 6 *demonstrates that a penalty method using F behaves well in numerical practice, in principle any other NCP-function, see* Sun and Qi [1999] *for an overview, can be used to construct similar penalty methods.*

One possible alternative would be to use $F_1 \colon H^1(\Omega)^2 \to \mathbb{R}_0^+$ given by

$$\forall (u, v) \in H^1(\Omega)^2: \quad F_1(u, v) := \int_\Omega |\phi(u(x), v(x))|\mathrm{d}x = \left\| \tilde{\Phi}(\mathrm{E}[u], \mathrm{E}[v]) \right\|_{L^1(\Omega)},$$

where $\tilde{\Phi} \colon L^2(\Omega)^2 \to L^1(\Omega)$ is the mapping $\mathrm{E}_{L^2 \to L^1} \circ \Phi$ where $\mathrm{E}_{L^2 \to L^1}$ represents the continuous embedding $L^2(\Omega) \hookrightarrow L^1(\Omega)$. This leads to a nonsmooth but Lipschitz continuous mapping.

Another approach would be to exploit the so-called smoothed *Fischer–Burmeister function* $\phi_\theta \colon \mathbb{R}^2 \to \mathbb{R}$ *given by*

$$\forall (a, b) \in \mathbb{R}^2 \colon \quad \phi_\theta(a, b) := \sqrt{a^2 + b^2 + 2\theta} - a - b,$$

which is continuously differentiable for any $\theta > 0$, *see* Kanzow [1996]. *Using [Goldberg et al., 1992, Theorems 4 and 7], one can check that the associated Nemytskii operator* $\tilde{\Phi}_\theta \colon L^2(\Omega)^2 \to L^1(\Omega)$ *is continuously Fréchet differentiable. Define* $F_{1,\theta} \colon H^1(\Omega)^2 \to \mathbb{R}_0^+$ *by means of*

$$\forall (u, v) \in H^1(\Omega)^2 \colon \quad F_{1,\theta}(u, v) := \int_\Omega |\phi_\theta(u(x), v(x))| \mathrm{d}x = \left\| \tilde{\Phi}_\theta(\mathrm{E}[u], \mathrm{E}[v]) \right\|_{L^1(\Omega)}.$$

Clearly, $F_{1,0}$ *corresponds to* F_1. *For* $\theta > 0$ *this approach can be seen as a mixture of a penalty and a smoothing method. However, it needs to be noted that* $F_{1,\theta}$ *is nonsmooth even for positive values of* θ.

4.2. Existence, convergence results, and optimality conditions

Using the penalty functional F defined in (4.2) to penalize the complementarity constraints in (3.1) leads to the family of penalized problems

$$\tfrac{1}{2} \|\mathrm{S}[u, v] - y_\mathrm{d}\|_\mathcal{D}^2 + J(u, v) + \sigma_k F(u, v) \; \to \; \min_{u, v}, \tag{P_k}$$

where $\{\sigma_k\}_{k \in \mathbb{N}} \subset \mathbb{R}^+$ is a sequence of positive real numbers tending to infinity as $k \to \infty$. The first question is about the existence of solutions of (P_k).

Proposition 4.3. *For any* $\sigma_k > 0$, *the penalized problem* (P_k) *possesses an optimal solution.*

Proof. Let $\{(u_l, v_l)\}_{l \in \mathbb{N}} \subset H^1(\Omega)^2$ be a minimizing sequence for (P_k) and let $\bar{m} \in \overline{\mathbb{R}}$ be the corresponding infimal value. Since J is, due to $\varepsilon > 0$, coercive and bounded from below, this sequence is bounded in $H^1(\Omega)^2$ and, thus, possesses a weakly convergent subsequence (without relabeling) with weak limit $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$. Due to the compactness of $H^1(\Omega) \hookrightarrow L^2(\Omega)$, the strong convergences $u_l \to \bar{u}$ and $v_l \to \bar{v}$ hold in $L^2(\Omega)$. Noting that the operator Φ is continuous on $L^2(\Omega)^2$, see [Goldberg et al., 1992, Theorem 4], it follows that

$$\lim_{l \to \infty} F(u_l, v_l) = F(\bar{u}, \bar{v}).$$

Thus, the continuity of S and the weak lower semicontinuity of norms imply that

$$\tfrac{1}{2} \|\mathrm{S}[\bar{u}, \bar{v}] - y_\mathrm{d}\|_\mathcal{D}^2 + J(\bar{u}, \bar{v}) + \sigma_k F(\bar{u}, \bar{v})$$
$$\leq \liminf_{l \to \infty} \left(\tfrac{1}{2} \|\mathrm{S}[u_l, v_l] - y_\mathrm{d}\|_\mathcal{D}^2 + J(u_l, v_l) \right) + \sigma_k \lim_{l \to \infty} F(u_l, v_l)$$
$$= \liminf_{l \to \infty} \left(\tfrac{1}{2} \|\mathrm{S}[u_l, v_l] - y_\mathrm{d}\|_\mathcal{D}^2 + J(u_l, v_l) + \sigma_k F(u_l, v_l) \right) = \bar{m},$$

i.e., $(\bar{u}, \bar{v})$ is a global minimizer of (P_k). $\qquad \square$

Next, the convergence of solutions of (P_k) as $\sigma_k \to \infty$ is addressed.

Proposition 4.4. *Fix a sequence $\{\sigma_k\}_{k \in \mathbb{N}} \subset \mathbb{R}^+$ tending to infinity as $k \to \infty$. For any $k \in \mathbb{N}$, let $(u_k, v_k) \in H^1(\Omega)^2$ be a global minimizer of* (P$_k$). *Then, $\{(u_k, v_k)\}_{k \in \mathbb{N}}$ contains a subsequence converging strongly in $H^1(\Omega)^2$ to a point $(\bar{u}, \bar{v}) \in \mathbb{C}$ such that $(\bar{y}, \bar{u}, \bar{v})$, where $\bar{y} \in \mathcal{Y}$ is the state associated with $(\bar{u}, \bar{v})$, is an optimal solution of* (OC^4).*

Moreover, any subsequence of $\{(u_k, v_k)\}_{k \in \mathbb{N}}$ converging weakly to $(\bar{u}, \bar{v})$ in $H^1(\Omega)^2$ produces a global minimizer of (OC^4) *in the above sense.*

Proof. For any $k \in \mathbb{N}$, the estimate

$$\tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k) + \sigma_k F(u_k, v_k) \leq \tfrac{1}{2} \left\| y_d \right\|_{\mathcal{D}}^2$$

follows from the feasibility of $(0,0) \in H^1(\Omega)^2$ for (P$_k$). Thus, since J is coercive and bounded from below while F only takes nonnegative values, $\{(u_k, v_k)\}_{k \in \mathbb{N}}$ is bounded and therefore contains a weakly convergent subsequence (which, as all further subsequences, will not be relabeled). Recalling the compactness of $H^1(\Omega) \hookrightarrow L^2(\Omega)$, the sequence $\{(u_k, v_k)\}_{k \in \mathbb{N}}$ converges strongly to $(\bar{u}, \bar{v})$ in $L^2(\Omega)^2$ and thus pointwise almost everywhere at least along a subsequence. Furthermore, the relation

$$0 \leq \left\| \Phi(\mathrm{E}[u_k], \mathrm{E}[v_k]) \right\|_{L^2(\Omega)} \leq \sqrt{\tfrac{1}{\sigma_k}} \left\| y_d \right\|_{\mathcal{D}} \to 0$$

is obtained as $k \to \infty$. Consequently, at least along a subsequence, $\{\Phi(\mathrm{E}[u_k], \mathrm{E}[v_k])\}_{k \in \mathbb{N}}$ converges pointwise a.e. to 0. By definition of Φ, $(\bar{u}, \bar{v}) \in \mathbb{C}$ follows.

Now choose $(u, v) \in \mathbb{C}$ arbitrarily. Since this point is feasible to (P$_k$), it follows for any $k \in \mathbb{N}$ that

$$\begin{aligned}
\tfrac{1}{2} \left\| S[u, v] - y_d \right\|_{\mathcal{D}}^2 + J(u, v) &\geq \tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k) + \sigma_k F(u_k, v_k) \\
&\geq \tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k).
\end{aligned}$$

Thus, using the weak lower semicontinuity of the functionals, one obtains

$$\begin{aligned}
\tfrac{1}{2} \left\| S[\bar{u}, \bar{v}] - y_d \right\|_{\mathcal{D}}^2 + J(\bar{u}, \bar{v}) &\leq \liminf_{k \to \infty} \left(\tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k) \right) \\
&\leq \limsup_{k \to \infty} \left(\tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k) \right) \\
&\leq \limsup_{k \to \infty} \left(\tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k) + \sigma_k F(u_k, v_k) \right) \\
&\leq \tfrac{1}{2} \left\| S[u, v] - y_d \right\|_{\mathcal{D}}^2 + J(u, v)
\end{aligned}$$

for all $(u, v) \in \mathbb{C}$. Consequently, $(\bar{u}, \bar{v})$ is a global minimizer of the state-reduced problem (3.1). Choosing $u := \bar{u}$ and $v := \bar{v}$ in the above estimate, one obtains

$$\tfrac{1}{2} \left\| S[u_k, v_k] - y_d \right\|_{\mathcal{D}}^2 + J(u_k, v_k) \to \tfrac{1}{2} \left\| S[\bar{u}, \bar{v}] - y_d \right\|_{\mathcal{D}}^2 + J(\bar{u}, \bar{v}),$$

and $J(u_k, v_k) \to J(\bar{u}, \bar{v})$ follows by Lemma A.1. Since $u_k \to \bar{u}$ and $v_k \to \bar{v}$ in $L^2(\Omega)$, the definition of J and $\varepsilon > 0$ imply that

$$\left\| u_k \right\|_{H^1(\Omega)}^2 + \left\| v_k \right\|_{H^1(\Omega)}^2 \to \left\| \bar{u} \right\|_{H^1(\Omega)}^2 + \left\| \bar{v} \right\|_{H^1(\Omega)}^2 .$$

Now, applying Lemma A.1 once more yields

$$\|u_k\|^2_{H^1(\Omega)} \to \|\bar{u}\|^2_{H^1(\Omega)}, \qquad \|v_k\|^2_{H^1(\Omega)} \to \|\bar{v}\|^2_{H^1(\Omega)}.$$

Combining this with the weak convergences $u_k \rightharpoonup \bar{u}$ and $v_k \rightharpoonup \bar{v}$ in $H^1(\Omega)$, the convergences $u_k \to \bar{u}$ and $v_k \to \bar{v}$ in $H^1(\Omega)$ follow since the latter is a Hilbert space. This yields the first assertion.

If $\{(u_k, v_k)\}_{k\in\mathbb{N}}$ contains a subsequence converging weakly to $(\bar{u}, \bar{v}) \in H^1(\Omega)^2$ in $H^1(\Omega)$, then the above arguments can be partially repeated to show that $(\bar{u}, \bar{v})$ is a global minimizer of (3.1). This completes the proof. □

An obvious advantage of (P_k) is that it is a smooth and unconstrained problem, allowing the straightforward derivation of necessary optimality conditions. Hence, the following result is a direct consequence of Fermat's rule and Lemma 4.1.

Proposition 4.5. *For fixed $\sigma_k > 0$, let $(u_k, v_k) \in H^1(\Omega)^2$ be a locally optimal solution of (P_k). Then, the corresponding functions $\eta_{u_k}, \eta_{v_k} \in L^\infty(\Omega)$ defined as in (4.3) satisfy*

$$0 = \mathrm{S}^\star\big[\mathrm{S}[u_k, v_k] - y_\mathrm{d}\big] + J'(u_k, v_k) + \sigma(\mathrm{E}, \mathrm{E})^\star\big[\Phi(\mathrm{E}[u_k], \mathrm{E}[v_k])\eta_{u_k}, \Phi(\mathrm{E}[u_k], \mathrm{E}[v_k])\eta_{v_k}\big].$$

Remark 4.6. *Similar results as in this section can be shown for the penalty terms induced by the nonsmooth functionals F_1 and F_{1,θ_k} given in Remark 4.2 using the continuity of the associated Nemytskii operators $\tilde{\Phi}$ and $\tilde{\Phi}_{\theta_k}$ as well as calculus rules for Clarke's generalized derivative, see Clarke [1990]. Obtaining a convergence result as in Proposition 4.4 for F_{1,θ_k} additionally requires to choose σ_k and θ_k such that $\sigma_k\sqrt{\theta_k} \to 0$ as $k \to \infty$.*

5. Numerical treatment

This section deals with the numerical implementation of the penalization technique described in Section 4 following a "first-discretize-then-optimize approach" based on a finite element discretization. In order to concentrate on the complementarity constraint, the state equation is chosen as the elliptic model problem

$$\left.\begin{aligned} -\nabla \cdot (\mathbf{C}\nabla y) + \mathbf{a}y &= \mathbf{b}u + \mathbf{c}v & \text{a.e. on } \Omega \\ \vec{\mathbf{n}} \cdot (\mathbf{C}\nabla y) &= 0 & \text{a.e. on } \mathrm{bd}\,\Omega. \end{aligned}\right\} \tag{PDE}$$

Here, $\Omega \subset \mathbb{R}^d$, $\mathbf{C} \in L^\infty(\Omega; S^d(\mathbb{R}))$ satisfies the condition of uniform ellipticity (2.3), and the functions $\mathbf{a}, \mathbf{b}, \mathbf{c} \in L^\infty(\Omega)$ do not vanish while $\mathbf{a}$ is additionally nonnegative, see also Section 2.2. Set $\mathcal{D} := L^2(\Omega)$ and $\mathcal{Y} := H^1(\Omega)$. The operator $\mathrm{D} := \mathrm{E}$ represents the natural embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$. Note that the weak formulation of the associated state equation can be written in the abstract form $\mathrm{A}[y] - \mathrm{B}[u] - \mathrm{C}[v] = 0$, where the bounded, linear operators

A, B, C ∈ $\mathbb{L}[H^1(\Omega), H^1(\Omega)^\star]$ are given for all $y, u, v, w \in H^1(\Omega)$ as

$$\langle A[y], w\rangle_{H^1(\Omega)} := \int_\Omega (C(x)\nabla y(x)) \cdot \nabla w(x)\mathrm{d}x + \int_\Omega a(x)y(x)w(x)\mathrm{d}x,$$

$$\langle B[u], w\rangle_{H^1(\Omega)} := \int_\Omega b(x)u(x)w(x)\mathrm{d}x,$$

$$\langle C[v], w\rangle_{H^1(\Omega)} := \int_\Omega c(x)v(x)w(x)\mathrm{d}x.$$

It can be checked that the operator A is elliptic and self-adjoint under the postulated assumptions, see, e.g., [Evans, 2010, Section 6]. The operators B and C are self-adjoint as well.

5.1. Finite element discretization

The domain Ω is discretized by a suitable tessellation Ω_Δ. The finite element space $\mathcal{P}^1(\Omega_\Delta)$ of piecewise affine elements is considered to replace the Sobolev space $H^1(\Omega)$. For later use, let n_p denote the number of vertices in Ω_Δ and n_e the number of elements. Accordingly, since the state y, the controls u and v, as well as the adjoint state p associated with the state equation (i.e., the associated Lagrange multiplier) all come from the space $H^1(\Omega)$, they are represented via linear combinations of the basis functions of the finite element space $\mathcal{P}^1(\Omega_\Delta)$. The corresponding weight vectors will be denoted by $\vec{y}, \vec{u}, \vec{v}$, and $\vec{p}$, which can be interpreted as the discrete counterparts of y, u, v, and p, respectively. The set of test functions $H^1(\Omega)$ is represented by the same basis functions.

The coefficient functions C, a, b, and c as well as the (interpolant of the) desired state y_d are assumed to be essentially bounded. Thus, they will be discretized analogously using the finite element space $\mathcal{P}^0(\Omega_\Delta)$ of piecewise constant functions. Their discrete approximations are denoted by $C, \vec{a}, \vec{b}, \vec{c}$, and $\vec{y}_\mathrm{d}$, respectively. The matrix $E_{10} \in \mathbb{R}^{n_e \times n_p}$ transforms the H^1-approximations into L^2-elements and corresponds to the natural embedding operator $E\colon H^1(\Omega) \to L^2(\Omega)$. The mass matrices $M_0(1)$ and $M_1(1)$ represent the finite element spaces $\mathcal{P}^0(\Omega_\Delta)$ and $\mathcal{P}^1(\Omega_\Delta)$, respectively, while the mass matrix M_{01} is used for the mixed finite element systems and represents the relationship between the different elements evaluated at grid nodes and barycenters of the elements. The stiffness matrix associated with the (constant) identity matrix is denoted by $K(1)$. A detailed description of this discretization and the specific forms of these matrices can be found in [Deng et al., 2018a].

This discretization leads to the finite-dimensional problem associated with (P_k) given by

$$\left.\begin{aligned}
&\tfrac{1}{2}(E_{10}\vec{y} - \vec{y}_\mathrm{d})^\top M_0(1)(E_{10}\vec{y} - \vec{y}_\mathrm{d}) + \tfrac{\alpha_1}{2}\vec{u}^\top M_1(1)\vec{u} + \tfrac{\alpha_2}{2}\vec{v}^\top M_1(1)\vec{v} \\
&\quad + \tfrac{\varepsilon}{2}\vec{u}^\top(M_1(1) + K(1))\vec{u} + \tfrac{\varepsilon}{2}\vec{v}^\top(M_1(1) + K(1))\vec{v} + \sigma_k\tilde{F}(\vec{u}, \vec{v}) \;\to\; \min_{\vec{y},\vec{u},\vec{v}} \\
&\qquad\qquad (M_{01}(\vec{a}) + K(C))\vec{y} - M_{01}(\vec{b})\vec{u} - M_{01}(\vec{c})\vec{v} \;=\; 0,
\end{aligned}\right\} \tag{5.1}$$

where $\tilde{F}(\vec{u}, \vec{v})$ denotes the discretized counterpart of the penalty term $F(u, v)$.

The chief difficulty in deriving useful optimality conditions for (5.1) lies in the discretization of the penalty term $F(u, v)$ and its derivatives using a discrete penalty term $\tilde{F}(\vec{u}, \vec{v})$ for the discretized controls $\vec{u}$ and $\vec{v}$. Since the Fischer–Burmeister function is penalized with respect

17

to the space $L_2(\Omega)$, the mass matrix $M_0(1)$ can be used to evaluate integrals over all elements. Interpreting powers and square roots of a vector in a componentwise fashion, a suitable choice for $\tilde{F}$ is given by

$$\tilde{F}(\vec{u}, \vec{v}) = \tfrac{1}{2} \left(\sqrt{(E_{10}\vec{u})^2 + (E_{10}\vec{v})^2} - E_{10}\vec{u} - E_{10}\vec{v} \right)^{\top} M_0(1) \left(\sqrt{(E_{10}\vec{u})^2 + (E_{10}\vec{v})^2} - E_{10}\vec{u} - E_{10}\vec{v} \right)$$

for all $\vec{u}, \vec{v} \in \mathbb{R}^{n_p}$. The appearance of E_{10} is motivated by the proof of Lemma 4.1, where the penalty functional F has been represented as the composition of three differentiable mapping: the natural embedding E, the Nemytskii-operator associated with the squared Fischer–Burmeister function (as a mapping from $L^2(\Omega)^2$ to $L^1(\Omega)$), and some linear integral operator. The quadratic function $\tilde{F}$ is first-order differentiable everywhere and its derivative at $(\vec{u}, \vec{v})$ is given by

$$\tilde{F}'(\vec{u}, \vec{v}) = \begin{pmatrix} E_{10}^{\top} \mathrm{diag}\left(T_u(\vec{u}, \vec{v})\right) M_0(1) \left(\sqrt{(E_{10}\vec{u})^2 + (E_{10}\vec{v})^2} - E_{10}\vec{u} - E_{10}\vec{v} \right) \\ E_{10}^{\top} \mathrm{diag}\left(T_v(\vec{u}, \vec{v})\right) M_0(1) \left(\sqrt{(E_{10}\vec{u})^2 + (E_{10}\vec{v})^2} - E_{10}\vec{u} - E_{10}\vec{v} \right) \end{pmatrix}, \tag{5.2}$$

where the vectors $T_u(\vec{u}, \vec{v}), T_v(\vec{u}, \vec{v}) \in \mathbb{R}^{n_e}$ are defined for all $i \in \{1, \ldots, n_e\}$ as

$$T_u(\vec{u}, \vec{v})_i := \begin{cases} \dfrac{(E_{10}\vec{u})_i}{\sqrt{(E_{10}\vec{u})_i^2 + (E_{10}\vec{v})_i^2}} - 1 & \text{if } (E_{10}\vec{u})_i \neq 0 \text{ or } (E_{10}\vec{v})_i \neq 0, \\ 0 & \text{if } (E_{10}\vec{u})_i = (E_{10}\vec{v})_i = 0, \end{cases}$$

$$T_v(\vec{u}, \vec{v})_i := \begin{cases} \dfrac{(E_{10}\vec{v})_i}{\sqrt{(E_{10}\vec{u})_i^2 + (E_{10}\vec{v})_i^2}} - 1 & \text{if } (E_{10}\vec{u})_i \neq 0 \text{ or } (E_{10}\vec{v})_i \neq 0, \\ 0 & \text{if } (E_{10}\vec{u})_i = (E_{10}\vec{v})_i = 0. \end{cases}$$

Note that the case $(E_{10}\vec{u})_i = (E_{10}\vec{v})_i = 0$ corresponds to the *biactive* case, i.e., where the discretized controls $\vec{u}$ and $\vec{v}$ (interpreted in the discretized counterpart of $L^2(\Omega)$, i.e., elementwise) are zero at the same time.

Combining (5.1) and (5.2), it is now possible to obtain for the problem (5.1) the KKT system

$$E_{10}^{\top} M_0(1) E_{10} \vec{y} - E_{10}^{\top} M_0(1) \vec{y}_\mathrm{d} - (M_{01}(\vec{a}) + K(C))\vec{p} = 0 \tag{5.3a}$$

$$[\alpha_1 M_1(1) + \varepsilon\,(M_1(1) + K(1))]\,\vec{u} + \sigma_k \tilde{F}'_{\vec{u}}(\vec{u}, \vec{v}) + M_{01}(\vec{b})\vec{p} = 0 \tag{5.3b}$$

$$[\alpha_2 M_1(1) + \varepsilon\,(M_1(1) + K(1))]\,\vec{v} + \sigma_k \tilde{F}'_{\vec{v}}(\vec{u}, \vec{v}) + M_{01}(\vec{c})\vec{p} = 0 \tag{5.3c}$$

$$-(M_{01}(\vec{a}) + K(C))\vec{y} + M_{01}(\vec{b})\vec{u} + M_{01}(\vec{c})\vec{v} = 0. \tag{5.3d}$$

Recall that $\vec{p}$ represents the discretized adjoint state and can also be considered as multiplier related to the discretized state equation. Since the function $\tilde{F}'$ is nonsmooth but Lipschitz continuous, the nonlinear system (5.3) can be solved using a damped semismooth Newton-type method, see Qi and Sun [1999]. Note that the domain of nonsmoothness associated with the mapping $\tilde{F}' \colon \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_p} \times \mathbb{R}^{n_p}$ is given by

$$\{(\vec{u}, \vec{v}) \in \mathbb{R}^{n_p} \times \mathbb{R}^{n_p} \mid \exists i \in \{1, \ldots, n_e\} \colon (E_{10}\vec{u})_i = (E_{10}\vec{v})_i = 0\}.$$

A particular Newton derivative can then be chosen as an element of Clarke's generalized Jacobian, see Clarke [1990], associated with $\tilde{F}'$ at $(\vec{u}, \vec{v})$ that is zero at indices corresponding to biactive components of $(\vec{u}, \vec{v})$. This choice will be used in the proposed method.

Next, due to the well-known local convergence behavior of Newton's method, the initialization of $\vec{u}$ and $\vec{v}$ for the numerical solution of (5.3) has to be taken into consideration. For that purpose, consider the (infinite-dimensional) problem

$$
\left.
\begin{aligned}
\tfrac{1}{2}\left\|\mathrm{E}[y] - y_{\mathrm{d}}\right\|^2_{L^2(\Omega)} + J(u,v) \;&\rightarrow\; \min_{y,u,v} \\
-\nabla\cdot(\mathbf{C}\nabla y) + \mathbf{a}y \;&=\; \mathbf{b}u + \mathbf{c}v \quad &\text{a.e. on } \Omega \\
\vec{\mathbf{n}}\cdot(\mathbf{C}\nabla y) \;&=\; 0 \quad &\text{a.e. on bd } \Omega \\
u, v \;&\geq\; 0 \quad &\text{a.e. on } \Omega
\end{aligned}
\right\}
\qquad \text{(OCNC)}
$$

which results from (OC^4) by omitting the equilibrium condition (4.1) and merely imposing nonnegativity constraints. Note that (OCNC) is convex and can be solved globally by combining a penalty algorithm and a semismooth Newton method, see Deng et al. [2018b]. The associated global minimizer is uniquely determined. If its solution already satisfies the equilibrium condition (4.1), then a global minimizer of (OC^4) has already been detected. The discretized counterpart of (OCNC) can be derived similarly as stated above. The associated (discrete) optimal solution $(\vec{y}_0, \vec{u}_0, \vec{v}_0)$ will be used as the starting vector of the semismooth Newton-type method. An abstract description of the proposed numerical method for the computational solution of (OC^4) is presented in Algorithm 1. In step **S2** of this algorithm, $\|\cdot\|_M$ denotes a weighted Euclidean norm which represents the discretized H^1-norm, see Deng et al. [2018b] for details.

Algorithm 1 Abstract algorithm

S0 Let $\{\sigma_k\}_{k\in\mathbb{N}}$ be a sequence of positive penalty parameters with $\sigma_k \to \infty$ as $k \to \infty$. Let a tolerance eps > 0 be given. Let $(\vec{y}_0, \vec{u}_0, \vec{v}_0)$ be the (discrete) optimal solution associated with (OCNC). Compute $\vec{p}_0$ as a solution of the discretized adjoint equation with source $E_{10}\vec{y}_0 - \vec{y}_{\mathrm{d}}$. Set $k := 1$.

S1 Solve the discretized KKT system (5.3) for fixed σ_k by a damped, semismooth Newton-type method with starting point $(\vec{y}_{k-1}, \vec{u}_{k-1}, \vec{v}_{k-1}, \vec{p}_{k-1})$. Let $(\vec{y}_k, \vec{u}_k, \vec{v}_k, \vec{p}_k)$ be the associated solution.

S2 If $\|(\vec{u}_k, \vec{v}_k) - (\vec{u}_{k-1}, \vec{v}_{k-1})\|_M <$ eps holds true, then accept $(\vec{u}_k, \vec{v}_k)$ as the discrete optimal control. Otherwise, set $k := k + 1$ and go to **S1**.

5.2. Checking strong stationarity

It has to be noted that in step **S1** of Algorithm 1 one generally only computes critical points to (5.1). Since the penalty functional F defined in (4.2) is not convex, these cannot be guaranteed to be global minimizers of (5.1) and therefore the convergence result of Proposition 4.4 does not apply. It is therefore sensible to verify whether the output is at least a strongly stationary point of (3.1) in the sense of Corollary 3.3, which is the best possible output of a method for the numerical solution of complementarity problems, see, e.g., Hoheisel et al. [2013] for a discussion of the finite-dimensional situation. A possible approach for such a verification is described in the following.

Let $(y, u, v) \in H^1(\Omega)^3$ be feasible to (OC^4). If this point is a local minimizer, then Corollary 3.3 implies that

$$\langle y - y_{\mathrm{d}}, y \rangle_{L^2(\Omega)} + \alpha_1 \langle u, u \rangle_{L^2(\Omega)} + \alpha_2 \langle v, v \rangle_{L^2(\Omega)} + \varepsilon \langle u, u \rangle_{H^1(\Omega)} + \varepsilon \langle v, v \rangle_{H^1(\Omega)} = 0 \qquad (5.4)$$

and that

$$\left\langle \bar{y} - y_{\mathrm{d}}, z_y \right\rangle_{L^2(\Omega)} + \alpha_1 \langle u, z_u \rangle_{L^2(\Omega)} + \alpha_2 \langle v, z_v \rangle_{L^2(\Omega)} + \varepsilon \langle u, z_u \rangle_{H^1(\Omega)} + \varepsilon \langle v, z_v \rangle_{H^1(\Omega)} \geq 0 \quad (5.5)$$

for any pair $(z_u, z_v) \in H^1_+(\Omega)^2$ with

$$\operatorname{supp} z_u \subset I^{+0}(u, v) \cup I^{00}(u, v), \qquad \operatorname{supp} z_v \subset I^{0+}(u, v) \cup I^{00}(u, v),$$

where $z_y \in H^1(\Omega)$ is the solution of the state equation $\mathrm{A}[z_y] - \mathrm{B}[z_u] - \mathrm{C}[z_v] = 0$.

Using the same discretization technique as described in Section 5.1, a discrete counterpart to (5.4) is

$$\Theta := \vec{y}^\top M_1(1) \vec{y} - \vec{y}^\top M_{01} \vec{y}_{\mathrm{d}} + \alpha_1 \vec{u}^\top M_1(1) \vec{u} + \alpha_2 \vec{v}^\top M_1(1) \vec{v}$$
$$+ \varepsilon \vec{u}^\top (K(1) + M_1(1)) \vec{u} + \varepsilon \vec{v}^\top (K(1) + M_1(1)) \vec{v} = 0. \qquad (5.6)$$

Clearly, a certain tolerance for the violation of (5.6) needs to be imposed in practice.

The numerical verification of condition (5.5) requires an appropriate choice of discrete test functions $\vec{z}_u, \vec{z}_v$ for given discretized controls $(\vec{u}, \vec{v})$ in the finite element space $\mathcal{P}^1(\Omega_\Delta)$. Considering the employed finite element discretization of (3.1), one particular choice is from the set of basis functions associated with $\mathcal{P}^1(\Omega_\Delta)$. Since the support of each of these "hat functions" covers all elements adjoining a single vertex, a corresponding elementwise approximation of the set $I^{+0}(u, v)$, $I^{00}(u, v)$, and $I^{0+}(u, v)$ is required as well. This can be defined using the projection of $\vec{u}, \vec{v}$ from $\mathcal{P}^1(\Omega_\Delta)$ to $\mathcal{P}^0(\Omega_\Delta)$ using the matrix E_{10}, which will be denoted by $\vec{u}^0 := E_{10} \vec{u}$ and $\vec{v}^0 := E_{10} \vec{v}$, respectively. This leads to the corresponding discrete sets

$$I^{+0}(\vec{u}, \vec{v}) := \left\{ i \in \{1, \ldots, n_e\} \,\middle|\, \vec{u}_i^0 > 0 \text{ and } \vec{v}_i^0 = 0 \right\},$$
$$I^{00}(\vec{u}, \vec{v}) := \left\{ i \in \{1, \ldots, n_e\} \,\middle|\, \vec{u}_i^0 = 0 \text{ and } \vec{v}_i^0 = 0 \right\},$$
$$I^{0+}(\vec{u}, \vec{v}) := \left\{ i \in \{1, \ldots, n_e\} \,\middle|\, \vec{u}_i^0 = 0 \text{ and } \vec{v}_i^0 > 0 \right\}.$$

For any pair of basis vectors $(\vec{z}_u, \vec{z}_v)$ whose support is contained in $I^{+0}(\vec{u}, \vec{v}) \cup I^{00}(\vec{u}, \vec{v})$ and $I^{0+}(\vec{u}, \vec{v}) \cup I^{00}(\vec{u}, \vec{v})$, respectively, one can then check whether

$$\Sigma(\vec{z}_u, \vec{z}_v) := \vec{y}^\top M_1(1) \vec{z}_y - \vec{z}_y^\top M_{01} \vec{y}_{\mathrm{d}} + \alpha_1 \vec{u}^\top M_1(1) \vec{z}_u + \alpha_2 \vec{v}^\top M_1(1) \vec{z}_v$$
$$+ \varepsilon \vec{u}^\top (K(1) + M_1(1)) \vec{z}_u + \varepsilon \vec{v}^\top (K(1) + M_1(1)) \vec{z}_v \geq 0, \quad (5.7)$$

where the state $\vec{z}_y$ associated with $(\vec{z}_u, \vec{z}_v)$ is obtained via

$$(M_1(1) + K(1)) \vec{z}_y = M_{01}(\vec{b}) \vec{z}_u + M_{01}(\vec{c}) \vec{z}_v.$$

In numerical practice, a certain tolerance with respect to negative values of $(\vec{u}^0, \vec{v}^0)$ is necessary since Algorithm 1 involves a penalty procedure and hence yields, in general, only *almost* feasible points for (OC^4). Rather than testing for nonnegativity, it is thus checked whether $\Sigma(\vec{z}_u, \vec{z}_v)$ is larger than a given negative tolerance.

6. Numerical examples

The proposed numerical method from Section 5 is illustrated by means of several examples. The implementation is carried out using the object oriented finite element MATLAB class library OOPDE, see Prüfert [2015].

Let $\Omega = (0,1)^2 \subset \mathbb{R}^2$. For all examples in this section, let C be the identity matrix in $\mathbb{R}^{2\times2}$ and let $a \equiv 1$, $b = \chi_{\Omega_u}$, as well as $c = \chi_{\Omega_v}$ hold where $\Omega_u := \{(x_1, x_2) \in \Omega \mid x_2 < 0.25\}$ and $\Omega_v := \{(x_1, x_2) \in \Omega \mid x_2 > 0.75\}$ are fixed subdomains of Ω. The values $\alpha_1 = \alpha_2 = 0$ are fixed for this section.

In order to construct examples where the controls are independent of x_2, cf. [Clason et al., 2016a, Section 6] where parabolic problems were considered and the controls only depend on time, the problem (OC^4) will be equipped with the additional restrictions

$$\partial_{x_2} u = \partial_{x_2} v = 0 \qquad \text{a.e. on } \Omega. \tag{6.1}$$

These constraints realize controls depending only on x_1 and being constant with respect to x_2 while allowing to use the same finite element space for the discretization of u, v, and y. Note that the additional constraints do not influence the complementarity constraints (which are now imposed on Ω rather than $(0,1)$). A detailed discussion of optimal control problems with gradient constraints can be found in Deng et al. [2018a].

To compare results, the solutions of the control problem (OCNC) without complementarity constraints (equipped with the additional constraints (6.1)) will be considered. Recall that optimal controls $(u, v) \in H^1(\Omega)^2$ of (OCNC) additionally fulfilling the equilibrium condition (4.1) solve (OC^4) as well, and that these controls are used as starting points for solving (OC^4). Since the computed controls are nearly constant with respect to x_2, only $u(x_1, 0)$ and $v(x_1, 0)$ are plotted for the sake of easier comparison. To evaluate the satisfaction of the complementarity conditions, the maximal absolute value of the Fischer–Burmeister function applied componentwise to $(\vec{u}^0, \vec{v}^0)$ is reported. Furthermore, $\Sigma(\vec{z}_u, \vec{z}_v)$ from (5.7) is checked with a tolerance

$$\text{tol} := 0.01 \left| \min_{(\vec{z}_u, \vec{z}_v) \text{ feasible test pair}} \Sigma(\vec{z}_u, \vec{z}_v) \right|, \tag{6.2}$$

and the number as well as distribution of pairs $(\vec{z}_u, \vec{z}_u)$ for which $\Sigma(\vec{z}_u, \vec{z}_v) > \text{tol}$ ("numerically positive"), $|\Sigma(\vec{z}_u, \vec{z}_v)| \le \text{tol}$ ("numerically zero"), or $\Sigma(\vec{z}_u, \vec{z}_v) < -\text{tol}$ ("numerically negative") holds is given.

Example 1 Here, the desired state is chosen to be the (weak) solution of the elliptic boundary value problem

$$\left. \begin{aligned} -\Delta y(x) &= 0 & \text{a.e. on } \Omega \\ y(x) &= 2\max\{0; x_1 \cos(0.75\pi x_1)\} & \text{a.e. on } \Gamma_1 \\ y(x) &= 0.25 & \text{a.e. on } \Gamma_2 \\ \vec{n}(x) \cdot \nabla y(x) &= 0 & \text{a.e. on } \Gamma_3 \end{aligned} \right\}$$

where $\Gamma_1 := [0,1] \times \{0\}$, $\Gamma_2 := [0,1] \times \{1\}$, and $\Gamma_3 := \{0,1\} \times [0,1]$ are fixed. Furthermore, $\varepsilon := 10^{-8}$ is used. The optimal controls of the associated problem (OCNC) do not fulfill the complementarity
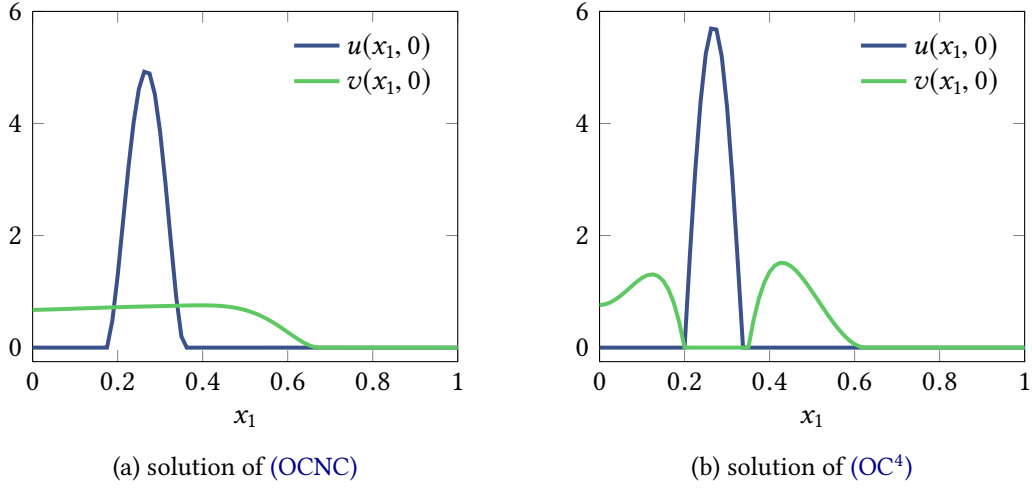
21

(a) solution of (OCNC)

(b) solution of (OC^4)

Figure 1: Example 1: computed controls

(a) $\Sigma(\vec{z}_u, \vec{z}_v)$

(b) pairs marked numerically positive (white), numerically zero (gray), numerically negative (black)
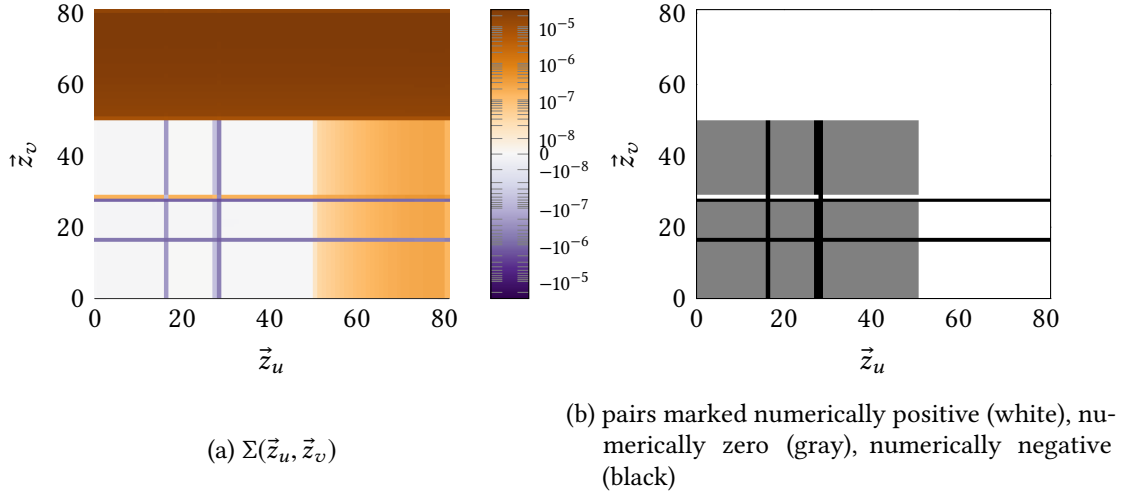
Figure 2: Example 1: values of stationarity test and distribution of failed pairs

condition but already provide a biactive set, see Figure 1a. On the other hand, the computed solution for (OC^4) approximately satisfies the complementarity condition, see Figure 1b, with a maximal absolute value of the Fischer–Burmeister function of approximately $3.58 \cdot 10^{-6}$.

The minimal value of $\Sigma(\vec{z}_u, \vec{z}_v)$ was approximately $-1.62 \cdot 10^{-6}$, cf. Figure 2a. Accordingly, the tolerance for the stationarity test was chosen as tol $= 1.617 \cdot 10^{-8}$. This leads to 4000 pairs $(\vec{z}_u, \vec{z}_v)$ marked as "numerically positive", 2256 as "numerically zero", and 305 as "numerically negative" and thus failing the strong stationarity test (5.7), see Figure 2b. These amount to approximately 4.7% of the total number 6561 of pairs. Note that pairs where the stationarity test fails correlate with those basis functions associated with nodes where the subdomains $I^{+0}(\vec{u}, \vec{v})$ and $I^{0+}(\vec{u}, \vec{v})$ meet. Finally, $\Theta = -2.01 \cdot 10^{-9}$ holds for the constant defined in (5.6).

(a) solution of (OCNC)

(b) solution of (OC^4)
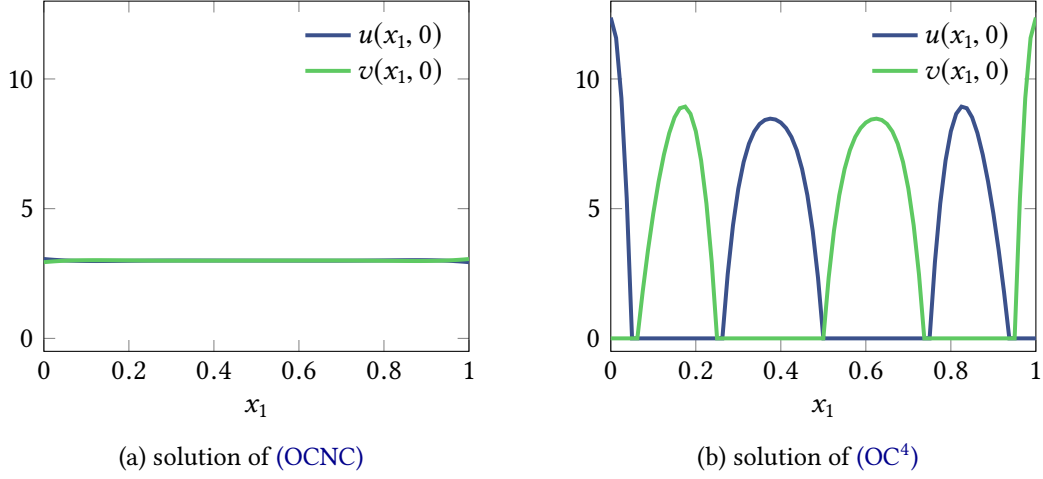
Figure 3: Example 2: computed controls

(a) $\Sigma(\vec{z}_u, \vec{z}_v)$

(b) pairs marked numerically positive (white), numerically zero (gray), numerically negative (black)
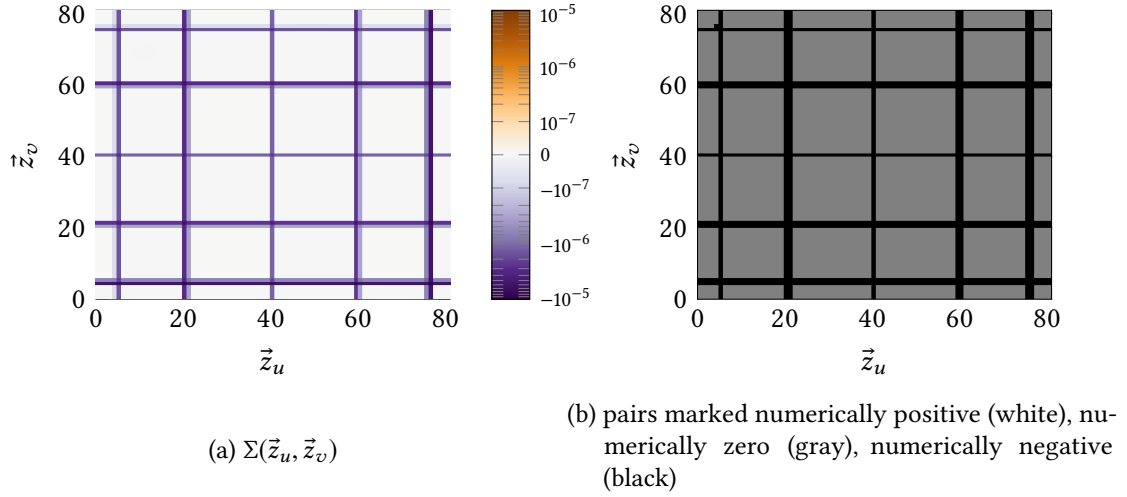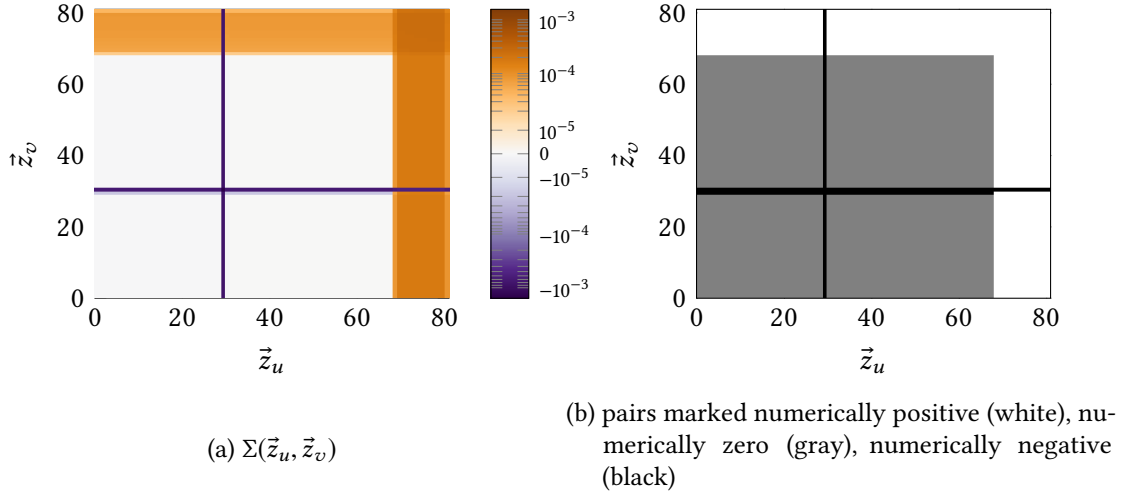
Figure 4: Example 2: values of stationarity test and distribution of failed pairs

Example 2   Here, the desired state is given by $y_d \equiv 1.5$, while the parameter ε is the same as in Example 1. The optimal controls for the problem (OCNC) are nearly constant functions, see Figure 3a. The controls for the problem (OC^4) computed via Algorithm 1 are complementary, see Figure 3b. The maximal absolute value of the Fischer–Burmeister function is $2.02 \cdot 10^{-6}$. Using the tolerance tol $= 1.11 \cdot 10^{-7}$ leads to 0 numerically positive, 5328 numerically zero, and 1233 numerically negative pairs, see Figure 4b. These are more than 18.5% of all tested pairs. In this example, $\Theta = -3.35 \cdot 10^{-10}$ holds true.

(a) solution of (OCNC)

(b) solution of (OC^4)
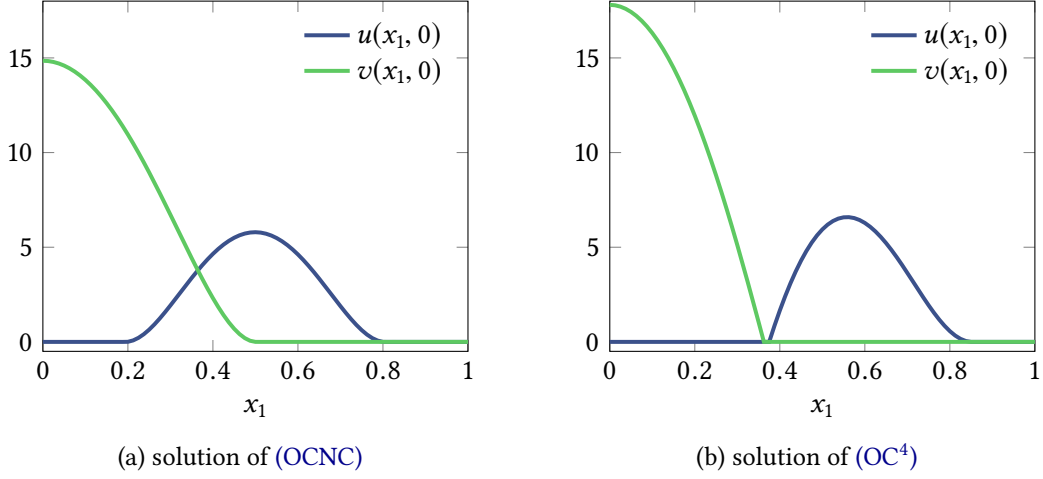
Figure 5: Example 3: computed controls

(a) $\Sigma(\vec{z}_u, \vec{z}_v)$

(b) pairs marked numerically positive (white), numerically zero (gray), numerically negative (black)

Figure 6: Example 3: values of stationarity test and distribution of failed pairs

Example 3 In this example, the desired state is given by the discontinuous function

$$y_d(x) := \begin{cases} 3 & \text{for } x \in [(0.25, 0.75) \times (0, 0.25)] \cup [(0, 0.5) \times (0.75, 1)] \\ 1 & \text{otherwise}, \end{cases}$$

and the Tikhonov parameter is chosen to be $\varepsilon = 10^{-5}$. In Figure 5, the computed solutions of (OCNC) and (OC^4) are visualized. The optimal control for (OC^4) is again approximately complementary with a maximal absolute value of the Fischer–Burmeister function of $3.18 \cdot 10^{-4}$. Here, the minimal value of $\Sigma(\vec{z}_u, \vec{z}_v)$ leads to the tolerance tol = $1.57 \cdot 10^{-5}$ leading to 1911 numerically positive, 4422 numerically zero, and 228 numerically negative pairs, see Figure 6b. The latter represent 3.5% of all tested pairs. Furthermore, $\Theta = -2.65 \cdot 10^{-6}$ is obtained.

24

(a) solution of (OCNC)

(b) solution of (OC⁴)
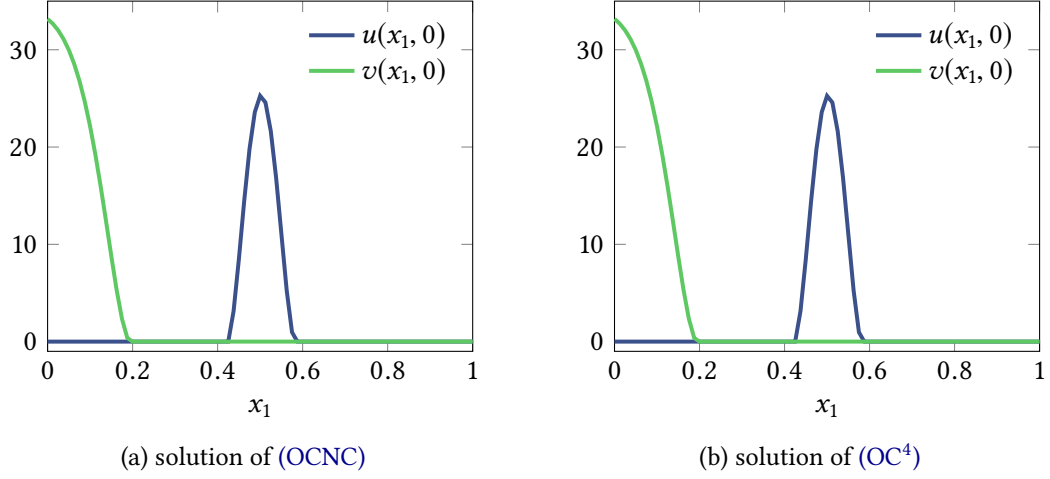
Figure 7: Example 4: computed controls

(a) $\Sigma(\vec{z}_u, \vec{z}_v)$

(b) pairs marked numerically positive (white), numerically zero (gray), numerically negative (black)
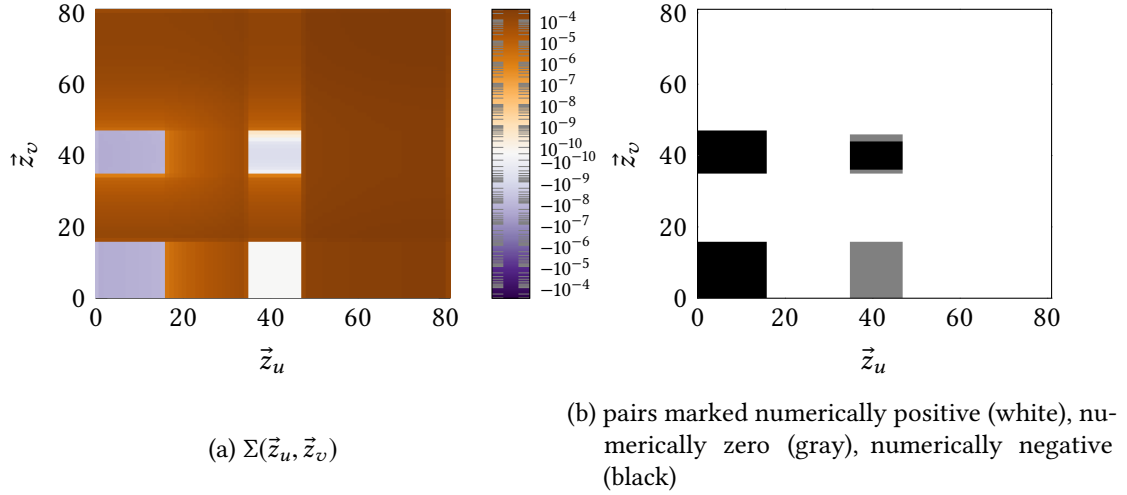
Figure 8: Example 4: values of stationarity test and distribution of failed pairs

Example 4 Here, the setting of Example 3 is used except the Tikhonov parameter is chosen to be $\varepsilon = 10^{-8}$. In this case, the optimal controls of problem (OCNC) are already (numerically) complementary, see Figure 7a, they provide a globally optimal solution of (OC⁴). Correspondingly, they coincide with the controls computed for (OC⁴), see Figure 7b, for which the maximal absolute value of the Fischer–Burmeister function is $2.08 \cdot 10^{-5}$. With the tolerance chosen as tol $= 2.27 \cdot 10^{-10}$, 5789 pairs are labeled as numerically positive, 228 as numerically zero, and 544 as numerically negative, see Figure 6b. Thus, only 8.3% of all tested pairs belong to the latter category. Note that $\Theta = -1.65 \cdot 10^{-7}$ holds.

Observing that Algorithm 1 computes the globally optimal solution of (OC⁴) in this example, the above data represent an *approximately* passed stationarity test.

	Example 1	Example 2	Example 3	Example 4
y_d	in $H^1(\Omega)$	constant	in $L^2(\Omega)$	in $L^2(\Omega)$
ε	10^{-8}	10^{-8}	10^{-5}	10^{-8}
complementarity	$3.58 \cdot 10^{-6}$	$2.02 \cdot 10^{-6}$	$3.18 \cdot 10^{-4}$	$2.08 \cdot 10^{-5}$
tol	$1.62 \cdot 10^{-8}$	$1.11 \cdot 10^{-7}$	$1.57 \cdot 10^{-5}$	$2.27 \cdot 10^{-10}$
Θ	$-2.01 \cdot 10^{-9}$	$-3.35 \cdot 10^{-10}$	$-2.65 \cdot 10^{-6}$	$-1.65 \cdot 10^{-7}$
num. neg. pairs	4.7%	18.5%	3.5%	8.3%
stationarity test	passed	failed	passed	passed

Table 1: summary of experiments

Summary The results of the numerical experiments are summarized in Table 1, where "complementarity" refers to the maximal absolute value of the elementwise Fischer–Burmeister function. Noting that Experiment 4 provides a benchmark for a passed stationarity test, a computed solution of (OC^4) is considered as *approximately passing* the strong stationarity test if $|\Theta| \leq \sqrt{\text{tol}}$ holds for Θ defined in (5.6) and the tolerance defined in (6.2), while the number of numerically negative tested pairs is at most 10% of the total number of tested pairs.

7. Conclusions

Optimal control problems with complementarity constraints on the controls admit solutions if the controls are chosen from a first-order Sobolev space. Although necessary optimality conditions of strong stationarity-type can be derived in this case, the explicit characterization of the associated Lagrange multipliers is difficult and remains the topic of further research. However, a penalty method based on the Fischer–Burmeister function can be formulated that ensures convergence to a global minimizers of the original complementarity-constrained problem. In theory, this requires computing global minimizers of the penalized problems, and it has to be investigated whether an adapted method based on KKT points is theoretically possible. Nevertheless, numerical examples illustrate that combined with a computable check for a discrete strong stationarity-type condition, this approach leads to a numerical procedure that in many cases results in nearly strongly stationary points. In light of prominent literature which deals with the numerical treatment of finite-dimensional complementarity problems, see Hoheisel et al. [2013] and the references therein, this seems to be the best to be hoped for.

A. A helpful lemma

In the proof of Proposition 4.4, the following lemma is used twice.

Lemma A.1. *Let $\{\alpha_k\}_{k\in\mathbb{N}}, \{\beta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ be sequences such that $\alpha_k + \beta_k \to \alpha + \beta$ holds where $\alpha, \beta \in \mathbb{R}$ satisfy*

$$\alpha \leq \liminf_{k\to\infty} \alpha_k, \qquad \beta \leq \liminf_{k\to\infty} \beta_k.$$

Then, the convergences $\alpha_k \to \alpha$ and $\beta_k \to \beta$ are valid.

Proof. The assumptions imply that

$$\alpha \leq \liminf_{k \to \infty} \alpha_k \leq \limsup_{k \to \infty} \alpha_k = \limsup_{k \to \infty} (\alpha_k + \beta_k - \beta_k)$$
$$= \lim_{k \to \infty} (\alpha_k + \beta_k) + \limsup_{k \to \infty} (-\beta_k) \leq \alpha + \beta - \beta = \alpha,$$

which implies that $\alpha_k \to \alpha$. Now, $\beta_k \to \beta$ follows from $\alpha_k + \beta_k \to \alpha + \beta$. □

Acknowledgments

References

R. A. Adams and J. J. F. Fournier. *Sobolev Spaces.* Elsevier Science, 2003.

H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces*, volume 6 of *MPS/SIAM Series on Optimization.* Society for Industrial and Applied Mathematics (SIAM), 2006. doi: 10.1137/1.9781611973488.

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems.* Springer, 2000. doi: 10.1007/978-1-4612-1394-9.

F. Clarke and M. R. de Pinho. Optimal control problems with mixed constraints. *SIAM Journal on Control and Optimization*, 48(7):4500–4524, 2010. doi: 10.1137/090757642.

F. H. Clarke. *Optimization and Nonsmooth Analysis.* Classics Appl. Math. 5. SIAM, 2nd edition, 1990. doi: 10.1137/1.9781611971309.

C. Clason, K. Ito, and K. Kunisch. A convex analysis approach to optimal controls with switching structure for partial differential equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 22(2):581–609, 2016a. doi: 10.1051/cocv/2015017.

C. Clason, A. Rund, K. Kunisch, and R. C. Barnard. A convex penalty for switching control of partial differential equations. *Systems & Control Letters*, 89:66–73, 2016b. doi: 10.1016/j.sysconle.2015.12.013.

C. Clason, A. Rund, and K. Kunisch. Nonconvex penalization of switching control of partial differential equations. *Systems & Control Letters*, 106:1–8, 2017. doi: 10.1016/j.sysconle.2017.05.006.

T. De Luca, F. Facchinei, and C. Kanzow. A theoretical and numerical comparison of some semismooth algorithms for complementarity problems. *Computational Optimization and Applications*, 16(2):173–205, 2000. doi: 10.1023/A:1008705425484.

Y. Deng, P. Mehlitz, and U. Prüfert. On an optimal control problem with gradient constraints. Technical Report SPP-1962-050, DFG Priority Programme 1962, 2018a. URL https://spp1962.wias-berlin.de/preprints/050.pdf.

Y. Deng, P. Mehlitz, and U. Prüfert. Optimal control in first-order Sobolev spaces with inequality constraints. Technical Report SPP-1962-053, DFG Priority Programme 1962, 2018b. URL https://spp1962.wias-berlin.de/preprints/053.pdf.

L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, 2010. doi: 10.1112/blms/20.4.375.

F. Facchinei, A. Fischer, and C. Kanzow. Regularity properties of a semismooth reformulation of variational inequalities. *SIAM Journal on Optimization*, 8(3):850–869, 1998. doi: 10.1137/S1052623496298194.

A. Fischer. A special Newton-type optimization method. *Optimization*, 24(3-4):269–284, 1992. doi: 10.1080/02331939208843795.

H. Goldberg, W. Kampowsky, and F. Tröltzsch. On Nemytskij operators in L_p-spaces of abstract functions. *Mathematische Nachrichten*, 155:127–140, 1992. doi: 10.1002/mana.19921550110.

L. Guo and J. J. Ye. Necessary optimality conditions for optimal control problems with equilibrium constraints. *SIAM Journal on Control and Optimization*, 54(5):2710–2733, 2016. doi: 10.1137/15M1013493.

F. Harder and G. Wachsmuth. Comparison of optimality systems for the optimal control of the obstacle problem. *GAMM-Mitteilungen*, 40(4):312–338, 2018a. doi: 10.1002/gamm.201740004.

F. Harder and G. Wachsmuth. The limiting normal cone of a complementarity set in Sobolev spaces. *Optimization*, pages 1–25, 2018b. doi: 10.1080/02331934.2018.1484467.

T. Hoheisel, C. Kanzow, and A. Schwartz. Theoretical and numerical comparison of relaxation methods for mathematical programs with complementarity constraints. *Mathematical Programming*, 137(1):257–288, 2013. doi: 10.1007/s10107-011-0488-5.

C. Kanzow. Some noninterior continuation methods for linear complementarity problems. *SIAM Journal on Matrix Analysis and Applications*, 17(4):851–868, 1996. doi: 10.1137/S0895479894273134.

Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511983658.

P. Mehlitz and G. Wachsmuth. Weak and strong stationarity in generalized bilevel programming and bilevel optimal control. *Optimization*, 65(5):907–935, 2016. doi: 10.1080/02331934.2015.1122007.

P. Mehlitz and G. Wachsmuth. The limiting normal cone to pointwise defined sets in Lebesgue spaces. *Set-Valued and Variational Analysis*, 26(3):449–467, 2018. doi: 10.1007/s11228-016-0393-4.

J.-S. Pang and M. Fukushima. Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with equilibrium constraints. *Computational Optimization and Applications*, 13(1):111–136, 1999. doi: 10.1023/A:1008656806889.

J.-S. Pang and D. E. Stewart. Differential variational inequalities. *Mathematical Programming A*, 113(2):345–424, Jun 2008. doi: 10.1007/s10107-006-0052-x.

U. Prüfert. *OOPDE: An object oriented toolbox for finite elements in Matlab*. TU Bergakademie Freiberg, 2015. URL http://www.mathe.tu-freiberg.de/files/personal/255/oopde-quickstart-guide-2015.pdf.

L. Qi and D. Sun. A survey of some nonsmooth equations and smoothing Newton methods. In *Progress in Optimization: Contributions from Australasia*, pages 121–146. Springer US, Boston, 1999. doi: 10.1007/978-1-4613-3285-5_7.

S. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000. doi: 10.1287/moor.25.1.1.15213.

D. Sun and L. Qi. On NCP-functions. *Computational Optimization and Applications*, 13(1):201–220, 1999. doi: 10.1023/A:1008669226453.

F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. American Mathematical Society, 2010. doi: 10.1090/gsm/112.

M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM, 2011. doi: 10.1137/1.9781611970692.

G. Wachsmuth. Mathematical programs with complementarity constraints in Banach spaces. *Journal on Optimization Theory and Applications*, 166:480–507, 2015. doi: 10.1007/s10957-014-0695-3.

J. J. Ye. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *J. Math. Anal. Appl.*, 307:350–369, 2005. doi: 10.1016/j.jmaa.2004.10.032.

J. J. Ye, D. L. Zhu, and Q. J. Zhu. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM Journal on Optimization*, 7(2):481–507, 1997. doi: 10.1137/S1052623493257344.