Deutsche Forschungsgemeinschaft
Priority Programme 1962

Coupled versus Decoupled Penalization of Control Complementarity Constraints

Yu Deng, Patrick Mehlitz, Uwe Prüfert

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

Preprint Number SPP1962-125

received on October 25, 2019

Edited by
SPP1962 at Weierstrass Institute for Applied Analysis and Stochastics (WIAS)
Leibniz Institute in the Forschungsverbund Berlin e.V.
Mohrenstraße 39, 10117 Berlin, Germany
E-Mail: spp1962@wias-berlin.de

World Wide Web: http://spp1962.wias-berlin.de/

Coupled versus decoupled penalization of control complementarity constraints

Yu Deng*, Patrick Mehlitz†, and Uwe Prüfert‡

October 25, 2019

Abstract

This paper deals with the numerical solution of optimal control problems with control complementarity constraints. For that purpose, we suggest the use of several penalty methods which differ with respect to the handling of the complementarity constraint which is either penalized as a whole with the aid of NCP-functions or decoupled in such a way that nonnegativity constraints as well as the equilibrium condition are penalized individually. We first present general global and local convergence results which cover several different penalty schemes before two decoupled methods which are based on a classical ℓ_1- and ℓ_2-penalty term, respectively, are investigated in more detail. Afterwards, the numerical implementation of these penalty methods is discussed. Based on some examples, where the optimal boundary control of a parabolic partial differential equation is considered, the qualitative and quantitative properties of the resulting algorithms are compared.

Introduction

Mathematical programs with complementarity constraints (MPCCs) frequently arise when real-world optimization models from e.g. engineering or economics comprising equilibrium conditions are formalized. Furthermore, multi-level optimization problems can be reformulated as MPCCs under suitable conditions. Noting that MPCCs suffer from an inherent lack of regularity due to the special disjunctive structure of their feasible sets, huge effort has been put into the development of problem-tailored optimality conditions, constraint qualifications, and solution algorithms during the last two decades. Exemplary, we refer to [26, 30, 35, 38, 39, 41] for an introduction to complementarity-constrained optimization in the finite- and infinite-dimensional setting. Particularly, optimal control problems with (pointwise) control complementarity constraints were considered in the recent papers [8, 17, 27]. Furthermore, the interested reader is referred to e.g. [18, 19, 30] where complementarity constraints are considered which arise in the context of the optimal control of variational inequalities.

Here, we consider an optimal control problem of parabolic partial differential equations where the available controls, which live only in time, have to satisfy a pointwise complementarity condition. A typical underlying application is given by the optimal boundary control of the non-stationary heat equation where two disjoint heating areas can be controlled but only one is allowed to be active at each time instance due to technical or economical reasons. Noting that parabolic partial differential equations describe a wide variety of time-dependent evolution phenomena from nature, engineering, medicine, and economics, this paper's theory applies to many more practically relevant settings. Following [8], the problem of interest possesses an optimal solution whenever controls are chosen from a first-order Sobolev space. Using a local decomposition approach, its local minimizers can be characterized with the aid of a strong stationarity-type necessary optimality condition. Here, we focus on the numerical solution of the problem with the aid of different penalty methods.

In the literature on finite-dimensional MPCCs, two ways on how to penalize a complementarity condition of the type

$$0 \leq a \perp b \geq 0$$

are suggested. First, one could aim to penalize only the equilibrium condition $ab = 0$ while leaving nonnegativity constraints $a, b \geq 0$ in the feasible set of the surrogate problem, see [20, 24, 34] and the references therein. We refer to this approach as a *decoupled* penalization of the complementarity constraint. Second, it is possible to penalize the overall complementarity constraint at once which will be referred to as a *coupled* approach. This is possible using so-called NCP-functions where NCP abbreviates *nonlinear complementarity program*. A continuous function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ is called NCP-function whenever it satisfies

$$\forall (a,b) \in \mathbb{R}^2\colon \quad \varphi(a,b) = 0 \iff 0 \leq a \perp b \geq 0,$$

*Faculty of Mathematics and Computer Science. Technische Universität Bergakademie Freiberg, Germany. yu.deng@math.tu-freiberg.de
†Institute of Mathematics, Chair of Optimal Control. Brandenburgische Technische Universität Cottbus-Senftenberg, Germany. mehlitz@b-tu.de
‡Faculty of Mathematics and Computer Science. Technische Universität Bergakademie Freiberg, Germany. uwe.pruefert@math.tu-freiberg.de

i.e. the set of roots associated with an NCP-functions precisely recovers the complementarity set in $\mathbb{R}^2$. A satisfying overview of NCP-functions can be found in [14, 22, 36]. For the construction of a solution method associated a complementarity-constrained program, one can focus on penalizing the violation of the equality constraint $\varphi(a,b)=0$ where $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ is an NCP-function. This has been done with the aid of the (squared) Fischer–Burmeister function, see [13], in [21]. In [8], the authors used a similar idea in order to solve optimal control problems with control complementarity constraints. Partial exact penalization in the context of finite-dimensional complementarity programming is discussed in [25].

In this paper, we consider different coupled and decoupled penalty approaches which can be used to tackle optimal control problems with control complementarity constraints. After investigating some convergence properties, we comment in detail on a classical ℓ_1-penalty as well as a quadratic equilibrium penalty. The main focus of the paper lies on the numerical implementation of these penalty methods and their comparison with the already known coupled penalization scheme from [8] which exploits the squared Fischer–Burmeister function. The principle idea is to solve the respective potentially nonsmooth first-order optimality systems associated with the penalized surrogate problems with the aid of a semismooth Newton-type method. In order to avoid dealing with coupled forward-backward systems, we use an all-at-ones approach, i.e. we solve the system for state, controls, and adjoint simultaneously. Generally, we rely on the so-called direct method, which is well known as first-discretize-then-optimize approach.

The paper is organized as follows: In Section 1, the notation and some function spaces used in this manuscript will be presented. Afterwards, the optimal control problem of interest will be introduced in Section 2. Some preliminaries on the state equation will be presented. Furthermore, the existence of solutions as well as necessary and sufficient optimality conditions for the problem of interest are briefly discussed. Section 3 is dedicated to the theoretical study of several penalty approaches which can be used to tackle control complementarity constraints. More precisely, we comment on a coupled penalty approach which exploits an NCP-function, a standard ℓ_1-penalty approach, as well as a decoupled penalty approach where the nonnegativity constraints as well as equilibrium constraint are penalized separately. We first present some quite general convergence results which cover all these penalization schemes. Afterwards, we investigate some more details regarding the ℓ_1-penalty as well the decoupled penalty approach where the equilibrium condition is penalized with the aid of an ℓ_2-penalty term. For the numerical solution of the optimal control problem, the penalized problems are discretized in suitable finite element spaces and the associated first-order optimality systems, which are solved with the aid of a (semismooth) Newton method, are derived in Section 4. Furthermore, we present some comments on the practical implementation of the penalty algorithms. Finally, the different penalty approaches are compared by means of some numerical experiments in Section 5.

1 Notation

Basic notation For two vectors $x, y \in \mathbb{R}^n$, $x \cdot y$ expresses their common Euclidean inner product. Let $\mathcal{X}$ be a Banach space with norm $\|\cdot\|_{\mathcal{X}}$. We denote the topological dual space of $\mathcal{X}$ by $\mathcal{X}^\star$. The associated dual pairing is represented by $\langle\cdot,\cdot\rangle_{\mathcal{X}} : \mathcal{X}^\star \times \mathcal{X} \to \mathbb{R}$. For a sequence $\{x_k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ and some point $\bar{x} \in \mathcal{X}$, strong and weak convergence of $\{x_k\}_{k\in\mathbb{N}}$ to $\bar{x}$ will be denoted by $x_k \to \bar{x}$ and $x_k \rightharpoonup \bar{x}$, respectively. The polar and annihilator of a set $A \subset \mathcal{X}$ are defined by

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

Function spaces For a nonempty, bounded domain $\Xi \subset \mathbb{R}^N$, a Banach space $\mathcal{B}$, and $1 \le p \le \infty$, $L^p(\Xi;\mathcal{B})$ denotes the common Lebesgue space of all (equivalence classes of) abstract measurable functions $u\colon \Xi \to \mathcal{B}$ such that $\Xi \ni \xi \mapsto \|u(\xi)\|_{\mathcal{B}} \in \mathbb{R}$ is p-integrable ($1 \le p < \infty$) or essentially bounded ($p = \infty$). Similarly, $C(\overline{\Xi};\mathcal{B})$ represents the space of all abstract continuous functions $u\colon \overline{\Xi} \to \mathcal{B}$. For brevity, we use $L^p(\Xi) := L^p(\Xi;\mathbb{R})$ for all $p \in [1,\infty]$ and $C(\overline{\Xi}) := C(\overline{\Xi};\mathbb{R})$. For an arbitrary function $u \in L^1(\Xi)$, $\operatorname{supp} u := \{\xi \in \Xi |\ u(\xi) \neq 0\}$ denotes the support of u. Recall that $\mathcal{M}(\overline{\Xi}) := C(\overline{\Xi})^\star$ comprises all finite Borel measures on $\overline{\Xi}$. We use $\mathcal{M}_-(\overline{\Xi})$ to denote the set of all nonpositive measures from $\mathcal{M}(\overline{\Xi})$ in the sense of duality. By $H^1(\Xi)$, we denote the common Sobolev space of first-order weakly differentiable functions from $L^2(\Xi)$ whose weak first-order derivatives belong to $L^2(\Xi)$ as well. Furthermore, we use $H^1_+(\Xi) \subset H^1(\Xi)$ to denote the nonempty, closed and convex cone of almost everywhere nonnegative functions in $H^1(\Xi)$. Further information on these classical function spaces are presented in [1, 37].

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary Γ and let $I := (0,T)$ be a time interval with $T > 0$. The associated space-time cylinder is given by $Q := \Omega \times I$ while its lateral boundary will be denoted by $\Sigma := \Gamma \times I$. The Banach space $W_2^{1,0}(Q)$ contains all functions $y \in L^2(Q)$ that are first-order weakly differentiable w.r.t. the spacial variables from Ω such that the associated spacial gradient ∇y is a function from $L^2(Q;\mathbb{R}^d)$. Similar, the space $W_2^{1,1}(Q)$ comprises all functions $y \in W_2^{1,0}(Q)$ that are first-order weakly differentiable w.r.t. time such that the first order time derivative $\partial_t y$ belongs to $L^2(Q)$. Obviously, the space $W_2^{1,1}(Q)$ may be identified with the Sobolev space $H^1(Q)$. Finally, we use $W(0,T) := \{y \in L^2(I;H^1(\Omega)) \,|\, y' \in L^2(I;H^1(\Omega)^\star)\}$ where y' denotes the distributional derivative of $y \in L^2(I;H^1(\Omega))$. It is well known that we have $W(0,T) \hookrightarrow C(\bar{I};L^2(\Omega))$ and that this embedding is continuous. Further information on function spaces which are suitable for the discussion of parabolic differential equations can be found in [37, 40].

2 Problem statement and preliminaries

For a given time interval $I := (0, T)$ with end time $T > 0$ and a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, with Lipschitz boundary $\Gamma \subset \mathbb{R}^d$, the following optimal control problem with control complementarity constraints will be considered:

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J(u,v) := \tfrac{1}{2}\,\|S(u,v) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\,\|u\|^2_{H^1(I)} + \tfrac{\lambda_2}{2}\,\|v\|^2_{H^1(I)} \\ \text{subject to} \quad & (u,v) \in \mathbb{C} := \left\{(w,z) \in H^1(I)^2 \,\middle|\, 0 \le w(t) \perp z(t) \ge 0 \text{ a.e. on } I\right\}. \end{aligned} \tag{OCP}$$

Here, $S\colon H^1(I)^2 \to L^2(\Omega)$ denotes the control-to-observation-operator which assigns any pair of controls $(u, v) \in H^1(I)^2$ to the terminal state $y(\cdot, T)\colon \Omega \to \mathbb{R}$ where y represents the (weak) solution of the parabolic state equation

$$\begin{aligned} \partial_t y - \nabla\cdot(\mathrm{C}\nabla y) + \mathrm{a}y &= 0 && \text{in } Q \\ \vec{n}\cdot(\mathrm{C}\nabla y) + \mathrm{q}y &= b\,u + c\,v && \text{in } \Sigma \\ y(\cdot,0) &= 0 && \text{in } \Omega, \end{aligned} \tag{PDE}$$

see Section 2.1. Above, $Q := \Omega \times I$ denotes the space-time cylinder while $\Sigma := \Gamma \times I$ represents its lateral boundary. The precise assumptions on (PDE) which guarantee that the linear operator S is continuous will be specified below.

Observe that (PDE) represents the non-stationary heat equation where the appearing coefficient function $\mathrm{C}\colon Q \to \mathbb{R}^{d\times d}$ represents thermal diffusivity while the functions $\mathrm{a}\colon Q \to \mathbb{R}$ and $\mathrm{q}\colon \Sigma \to \mathbb{R}$ describe the heat conduction on the domain Q and its lateral boundary Σ, respectively. Furthermore, $\vec{n}\cdot\nabla y$ stands for the outward normal derivative of y w.r.t. the variables x which address the spacial domain Ω. The coefficient functions $b, c\colon \Gamma \to \mathbb{R}$ are used to model the distribution of the control effort on Γ provided by the controls $u, v\colon I \to \mathbb{R}$ which only live in time. Typically, b and c are characteristic functions of certain subsets of Γ which can be controlled individually. Standard optimal boundary control of parabolic partial differential equations is discussed e.g. in [37, Section 3].

In (OCP), we try to find *complementary* controls $(u, v) \in H^1(I)^2$ such that the resulting terminal state $S(u, v)$ is close to a desired distribution of temperature $y_\mathrm{d}\colon \Omega \to \mathbb{R}$ while the overall control effort is minimal. Clearly, the complementarity constraint $(u, v) \in \mathbb{C}$ is the major difficulty in (OCP). It causes that the optimal control problem of interest is nonconvex and inherently irregular. However, the presence of pointwise complementarity constraints is motivated mainly by practical aspects. Observing that at most one control can be nonzero at each time instance, computed control strategies naturally save power while the model prevents cancellation of control effort. Furthermore, the heating of the body Ω over time via two boundary segments, which can be promoted by an appropriate choice of b and c as characteristic functions, might be technically restricted by the fact that only one boundary segment can be controlled at each time instance. As it is clear from the literature, see, e.g., [8, 10], H^1-regularity of controls is important since this assumption guarantees that $\mathbb{C}$ is weakly sequentially closed. In contrast to the standard setting of optimal control, the control effort is measured in terms of the H^1-norm of u and v. This leads to H^1-coercivity of the objective functional $J\colon H^1(I)^2 \to \mathbb{R}$ and, thus, ensures the existence of a global minimizer. It needs to be noted that $H^1(I)$ is compactly embedded into $C(\bar{I})$ which means that we actually have

$$\mathbb{C} = \left\{(w,z) \in H^1(I)^2 \,\middle|\, 0 \le w(t) \perp z(t) \ge 0 \text{ for all } t \in I\right\}.$$

This observation will be the base for our upcoming analysis.

The precise assumptions on (OCP) are stated below.

Assumption 2.1. *Let the desired state $y_\mathrm{d} \in L^2(\Omega)$ and regularization parameters $\lambda_1, \lambda_2 > 0$ be fixed. Furthermore, let $b, c \in L^\infty(\Gamma)$ be non-vanishing functions. Finally, we fix $\mathrm{C} \in L^\infty(Q; \mathbb{R}^{d\times d})$, $\mathrm{a} \in L^\infty(Q)$, and $\mathrm{q} \in L^\infty(\Sigma)$ such that C satisfies the so-called condition of uniform ellipticity*

$$\exists \gamma > 0\ \forall (x,s) \in Q\ \forall \eta \in \mathbb{R}^d\colon \quad \eta^\top \mathrm{C}(x,s)\eta \ge \gamma\,|\eta|_2^2$$

while q is nonnegative almost everywhere on Σ.

2.1 The state equation

For the discussion of the state equation (PDE), we mainly follow classical arguments provided in [37, Section 3.4]. Using test functions from $\mathcal{W} := \{\xi \in W_2^{1,1}(Q) \mid \forall x \in \Omega\colon\ \xi(x,T) = 0\}$ while exploiting Green's formula as well as the present boundary and initial condition, the associated variational formulation of (PDE) is given by

$$\iint_Q (-y\partial_t\xi + (\mathrm{C}\nabla y)\cdot\nabla\xi + \mathrm{a}y\xi)\,\mathrm{d}x\mathrm{d}t + \iint_\Sigma \mathrm{q}y\xi\,\mathrm{d}s\mathrm{d}t = \iint_\Sigma (bu + cv)\xi\,\mathrm{d}s\mathrm{d}t \qquad \forall \xi \in \mathcal{W}. \tag{2.1}$$

A solution of the variational problem (2.1) will be referred to as a weak solution of (PDE). Noting that the mapping $H^1(I)^2 \ni (u,v) \mapsto bu + cv \in L^2(\Sigma)$ is linear, continuous, and compact due to the postulated L^∞-regularity of b and c as well as the compactness of the embedding $H^1(I) \hookrightarrow L^2(I)$, it follows from [37, Theorem 3.12] that (2.1) possesses a unique solution in $W_2^{1,0}(Q)$ which belongs to $W(0,T)$ after a suitable modification on set of measure zero. Exploiting the continuity of $W(0,T) \hookrightarrow C(\bar{I}, L^2(\Omega))$, the observation operator $W(0,T) \ni y \to y(\cdot,T) \in L^2(\Omega)$ is linear and continuous. Thus, we obtain the following theorem from [37, Theorem 3.13].

Lemma 2.2. *The control-to-observation-operator $S\colon H^1(I)^2 \to L^2(\Omega)$ associated with (PDE) is linear, continuous, and compact.*

Let us briefly note that similar arguments can be applied to the slightly more general parabolic equation

$$\begin{aligned} \partial_t y - \nabla \cdot (\mathrm{C}\nabla y) + \mathrm{a} y &= f && \text{in } Q \\ \vec{n} \cdot (\mathrm{C}\nabla y) + \mathrm{q} y &= bu + cv && \text{in } \Sigma \\ y(\cdot,0) &= y_0 && \text{in } \Omega \end{aligned}$$

where $\mathrm{C} \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ represents thermal diffusivity, $f \in L^2(Q)$ is a fixed heat source, and $y_0 \in L^2(\Omega)$ is an initial distribution of temperature, see [37, Section 7] and [40, Section 4]. The associated control-to-observation-operator is affine and continuous from $H^1(I)^2$ to $L^2(\Omega)$. Clearly, the consideration of $f \equiv 0$ and $y_0 \equiv 0$ is not restrictive observing that one can always shift y_d in order to transfer (OCP) to this setting.

For later purposes, we need to characterize the adjoint operator $S^\star\colon L^2(\Omega) \to (H^1(I)^2)^\star$. In order to do that, the subsequently stated adjoint equation associated with (PDE) has to be considered:

$$\begin{aligned} -\partial_t p - \nabla \cdot (\mathrm{C}\nabla p) + \mathrm{a} p &= 0 && \text{in } Q \\ \vec{n} \cdot (\mathrm{C}\nabla p) + \mathrm{q} p &= 0 && \text{in } \Sigma \\ p(\cdot,T) &= \eta && \text{in } \Omega. \end{aligned} \tag{APDE}$$

For test functions from $\mathcal{W}' := \{\zeta \in W_2^{1,1}(Q) \mid \forall x \in \Omega\colon\ \zeta(x,0) = 0\}$, a suitable variational formulation of (APDE) is given by

$$\iint_Q (p\partial_t \zeta + (\mathrm{C}\nabla p)\cdot\nabla\zeta + \mathrm{a} p\zeta)\,\mathrm{d}x\mathrm{d}t + \iint_\Sigma \mathrm{q} p\zeta\,\mathrm{d}s\mathrm{d}t = \int_\Omega \eta\zeta(\cdot,T)\mathrm{d}x \qquad \forall \zeta \in \mathcal{W}'. \tag{2.2}$$

Again, a solution of (2.2) will be referred to as a weak solution of (APDE). Due to [37, Lemma 3.17], for each $\eta \in L^2(\Omega)$, (APDE) possesses a uniquely determined weak solution in $W_2^{1,0}(Q)$ which belongs to $W(0,T)$ after a modification on a set of measure zero. Furthermore, the linear solution operator associated with (APDE) is continuous as a mapping from $L^2(\Omega)$ to $W(0,T)$.

Lemma 2.3. *For $\eta \in L^2(\Omega)$, let $p \in W(0,T)$ be the weak solution of (APDE). Then, we have*

$$\forall (u,v) \in H^1(I)^2\colon \quad \left\langle S^\star\eta, (u,v)\right\rangle_{H^1(\Omega)^2} = \iint_\Sigma bpu\,\mathrm{d}s\mathrm{d}t + \iint_\Sigma cpv\,\mathrm{d}s\mathrm{d}t.$$

Proof. Fix $(u,v) \in H^1(I)^2$ arbitrarily and let $y \in W(0,T)$ be the associated uniquely determined weak solution of (PDE). Then, [37, Theorem 3.18] yields

$$\left\langle S^\star\eta, (u,v)\right\rangle_{H^1(I)^2} = \langle \eta, S(u,v)\rangle_{L^2(\Omega)} = \int_\Omega \eta y(\cdot,T)\mathrm{d}x = \iint_\Sigma (bu+cv)p\,\mathrm{d}s\mathrm{d}t,$$

and this shows the claim. □

2.2 Existence and optimality conditions

First, we combine Lemma 2.2 and [8, Corollary 2.4] in order to infer the existence of an optimal solution to (OCP).

Proposition 2.4. *Problem (OCP) possesses an optimal solution.*

It has been shown in [8] that (OCP) can be transferred into an MPCC in Banach spaces where the complementarity condition is induced by the cone of all almost everywhere nonnegative functions in $L^2(I)$, see [39] as well. Naturally, reasonable constraint qualification from Banach space programming do not hold at the feasible points of MPCCs. Furthermore, noting that the embedding $H^1(I) \hookrightarrow L^2(I)$ is clearly not surjective, problem-tailored constraint qualifications for MPCCs are violated as well. However, using a local decomposition approach, the authors of [8] were in position to derive necessary optimality

conditions for problem (OCP) directly. Let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a feasible point of (OCP). Due to the continuity of $\bar{u}$ and $\bar{v}$, the sets

$$\begin{aligned} I^{+0}(\bar{u}, \bar{v}) &:= \{t \in I \,|\, \bar{u}(t) > 0 \wedge \bar{v}(t) = 0\}, \\ I^{0+}(\bar{u}, \bar{v}) &:= \{t \in I \,|\, \bar{u}(t) = 0 \wedge \bar{v}(t) > 0\}, \\ I^{00}(\bar{u}, \bar{v}) &:= \{t \in I \,|\, \bar{u}(t) = 0 \wedge \bar{v}(t) = 0\} \end{aligned} \tag{2.3}$$

are well-defined and measurable. In [8, Corollary 3.3], the following multiplier-free necessary optimality condition has been derived.

Proposition 2.5. *Let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a locally optimal solution of (OCP). Furthermore, set*

$$\mathcal{T}(\bar{u}, \bar{v}) := \left\{(z_u, z_v) \in H^1_+(I) \,\middle|\, \operatorname{supp} z_u \subset I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \;\wedge\; \operatorname{supp} z_v \subset I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v})\right\}.$$

Then, the following conditions hold:

$$\langle S(\bar{u}, \bar{v}) - y_\mathrm{d}, S(\bar{u}, \bar{v})\rangle_{L^2(\Omega)} + \lambda_1 \langle \bar{u}, \bar{u}\rangle_{H^1(I)} + \lambda_2 \langle \bar{v}, \bar{v}\rangle_{H^1(I)} = 0, \tag{2.4a}$$

$$\langle S(\bar{u}, \bar{v}) - y_\mathrm{d}, S(z_u, z_v)\rangle_{L^2(\Omega)} + \lambda_1 \langle \bar{u}, z_u\rangle_{H^1(I)} + \lambda_2 \langle \bar{v}, z_v\rangle_{H^1(I)} \geq 0 \qquad \forall (z_u, z_v) \in \mathcal{T}(\bar{u}, \bar{v}). \tag{2.4b}$$

A dual counterpart of the above primal necessary optimality condition can be found in [8]. In the latter paper, is has been justified to call this system a strong stationarity-type necessary optimality condition. For numerical purposes, the system (2.4) from Proposition 2.5 is of essential interest since it allows the implementation of a stationarity test for computed feasible points which is based on the finite element method, see [8, Section 5.2] and Section 4.4.

Below, we present a sufficient optimality condition for (OCP) which is based on a slightly stronger notion of stationarity than provided above.

Proposition 2.6. *Let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a feasible point of (OCP). Furthermore, for $\varepsilon \geq 0$, set*

$$\mathcal{T}_\varepsilon(\bar{u}, \bar{v}) := \left\{(z_u, z_v) \in H^1_+(I)^2 \,\middle|\, \operatorname{supp} z_u \subset \{t \in I \,|\, \bar{v}(t) \leq \varepsilon\} \;\wedge\; \operatorname{supp} z_v \subset \{t \in I \,|\, \bar{u}(t) \leq \varepsilon\}\right\}.$$

Now, assume that (2.4a) is valid while

$$\langle S(\bar{u}, \bar{v}) - y_\mathrm{d}, S(z_u, z_v)\rangle_{L^2(\Omega)} + \lambda_1 \langle \bar{u}, z_u\rangle_{H^1(I)} + \lambda_2 \langle \bar{v}, z_v\rangle_{H^1(I)} \geq 0 \qquad \forall (z_u, z_v) \in \mathcal{T}_\varepsilon(\bar{u}, \bar{v}) \tag{2.5}$$

holds for some $\varepsilon > 0$. Then, there is a neighborhood $U \subset H^1(I)^2$ of $(\bar{u}, \bar{v})$, such that the second-order growth condition

$$\forall (u, v) \in \mathbb{C} \cap U\colon \quad J(u, v) \geq J(\bar{u}, \bar{v}) + \lambda_1 \|u - \bar{u}\|^2_{H^1(I)} + \lambda_2 \|v - \bar{v}\|^2_{H^1(I)} \tag{2.6}$$

holds. Particularly, $(\bar{u}, \bar{v})$ is a strict local minimizer of (OCP).

Proof. For later use, let use define $\bar{\mu}, \bar{\nu} \in H^1(I)^\star$ as stated below:

$$\begin{aligned} \forall z \in H^1(I)\colon \quad \langle \bar{\mu}, z\rangle_{H^1(I)} &:= -\langle S(\bar{u}, \bar{v}) - y_\mathrm{d}, S(z, 0)\rangle_{L^2(\Omega)} - \lambda_1 \langle \bar{u}, z\rangle_{H^1(I)}, \\ \langle \bar{\nu}, z\rangle_{H^1(I)} &:= -\langle S(\bar{u}, \bar{v}) - y_\mathrm{d}, S(0, z)\rangle_{L^2(\Omega)} - \lambda_2 \langle \bar{v}, z\rangle_{H^1(I)}. \end{aligned}$$

By linearity of S, it holds

$$\langle S(\bar{u}, \bar{v}) - y_\mathrm{d}, S(z_u, z_v)\rangle_{L^2(\Omega)} + \lambda_1 \langle \bar{u}, z_u\rangle_{H^1(I)} + \lambda_2 \langle \bar{v}, z_v\rangle_{H^1(I)} + \langle \bar{\mu}, z_u\rangle_{H^1(I)} + \langle \bar{\nu}, z_v\rangle_{H^1(I)} = 0 \tag{2.7}$$

for all $(z_u, z_v) \in H^1(I)^2$. As a consequence, we obtain

$$\langle \bar{\mu}, \bar{u}\rangle_{H^1(I)} + \langle \bar{\nu}, \bar{v}\rangle_{H^1(I)} = 0, \tag{2.8a}$$

$$\langle \bar{\mu}, z_u\rangle_{H^1(I)} + \langle \bar{\nu}, z_v\rangle_{H^1(I)} \leq 0 \qquad \forall (z_u, z_v) \in \mathcal{T}_\varepsilon(\bar{u}, \bar{v}) \tag{2.8b}$$

from (2.4a) and (2.5). Let us define the MPCC-Lagrangian $L\colon H^1(I) \times H^1(I) \times H^1(I)^\star \times H^1(I)^\star \to \mathbb{R}$ of (OCP) by

$$\forall u, v \in H^1(I)\, \forall \mu, \nu \in H^1(I)^\star\colon \quad L(u, v, \mu, \nu) := J(u, v) + \langle \mu, u\rangle_{H^1(I)} + \langle \nu, v\rangle_{H^1(I)}.$$

Performing a second-order Taylor expansion (derivatives are taken only w.r.t. u and v) on L at $(\bar{u}, \bar{v}, \bar{\mu}, \bar{\nu})$ while observing that L is quadratic yields

$$\begin{aligned} L(u, v, \bar{\mu}, \bar{\nu}) = L(\bar{u}, \bar{v}, \bar{\mu}, \bar{\nu}) &+ L'(\bar{u}, \bar{v}, \bar{\mu}, \bar{\nu})(u - \bar{u}, v - \bar{v}) \\ &+ \tfrac{1}{2} L''(\bar{u}, \bar{v}, \bar{\mu}, \bar{\nu})[(u - \bar{u}, v - \bar{v}), (u - \bar{u}, v - \bar{v})] \end{aligned}$$

for each $(u,v) \in H^1(I)^2$. Due to (2.7), the term $L'(\bar{u},\bar{v},\bar{\mu},\bar{\nu})(u-\bar{u},v-\bar{v})$ vanishes. Thus, computing the second-order derivative of L and respecting (2.8a) yield

$$L(u,v,\bar{\mu},\bar{\nu}) = J(\bar{u},\bar{v}) + \tfrac{1}{2}\,\|S(u-\bar{u},v-\bar{v})\|^2_{L^2(\Omega)} + \lambda_1\,\|u-\bar{u}\|^2_{H^1(I)} + \lambda_2\,\|v-\bar{v}\|^2_{H^1(I)}\,. \tag{2.9}$$

Due to compactness of $H^1(I) \hookrightarrow C(\bar{I})$, we can find a neighborhood $U \subset H^1(I)^2$ of $(\bar{u},\bar{v})$ such that

$$\forall (u,v) \in U\ \forall t \in I\colon \quad \bar{u}(t) > \varepsilon \implies u(t) \geq \tfrac{\varepsilon}{2} \qquad \bar{v}(t) > \varepsilon \implies v(t) \geq \tfrac{\varepsilon}{2}.$$

Now, fix $(u,v) \in \mathbb{C} \cap U$. For any $t \in I$ with $\bar{v}(t) > \varepsilon$, it holds $v(t) \geq \frac{\varepsilon}{2}$ and, thus, $u(t) = 0$ by definition of $\mathbb{C}$. This shows $\operatorname{supp} u \subset \{t \in I \,|\, \bar{v}(t) \leq \varepsilon\}$. Similarly, we have $\operatorname{supp} v \subset \{t \in I \,|\, \bar{u}(t) \leq \varepsilon\}$. This shows $(u,v) \in \mathcal{T}_\varepsilon(\bar{u},\bar{v})$, and that is why we deduce $\mathbb{C} \cap U \subset \mathcal{T}_\varepsilon(\bar{u},\bar{v})$. Thus, we can combine (2.8b) and (2.9) in order to obtain

$$\begin{aligned} J(u,v) &\geq L(u,v,\bar{\mu},\bar{\nu}) \\ &= J(\bar{u},\bar{v}) + \tfrac{1}{2}\,\|S(u-\bar{u},v-\bar{v})\|^2_{L^2(\Omega)} + \lambda_1\,\|u-\bar{u}\|^2_{H^1(I)} + \lambda_2\,\|v-\bar{v}\|^2_{H^1(I)} \\ &\geq J(\bar{u},\bar{v}) + \lambda_1\,\|u-\bar{u}\|^2_{H^1(I)} + \lambda_2\,\|v-\bar{v}\|^2_{H^1(I)} \end{aligned}$$

for each $(u,v) \in \mathbb{C} \cap U$, and this shows (2.6). □

Classically, local second-order growth of a function over a set at some reference point is ensured via validity of a second-order optimality condition which demands that the second-order derivative of the Lagrangian function at an underlying stationary point is coercive on a suitable critical cone, see e.g. [3, Section 3.3] for details. Here, the term *stationary* generally refers to Karush–Kuhn–Tucker (KKT) points of the problem of interest. This principle has been extended to finite-dimensional MPCCs in the context of strongly stationary points which are precisely the KKT points in this setting, see [35]. However, it has been remarked in [39] that in infinite dimensions, strong stationarity might be strictly weaker than the KKT conditions of a complementarity-constrained problem, and this gives rise to the idea that strong stationarity of a point may not be enough for the derivation of sufficient optimality conditions. Exemplary, let us mention that second-order sufficient optimality conditions for the optimal control of the obstacle problem indeed turn out to be based on strongly stationary points where the involved multipliers satisfy additional assumptions, see [6, 23]. As discussed in [8, Remark 3.4], the necessary optimality conditions provided in Proposition 2.5 are slightly weaker than the strong stationarity conditions of (OCP) in the sense of [39]. It is, thus, not surprising that our sufficient optimality condition from Proposition 2.6 is *not* based on the strong stationarity-type conditions from Proposition 2.5 but on a stronger concept. Let us note that fixing $\varepsilon := 0$ in the definition of $\mathcal{T}_\varepsilon(\bar{u},\bar{v})$ recovers the definition of $\mathcal{T}(\bar{u},\bar{v})$ which means that the gap between the necessary and sufficient optimality conditions from Propositions 2.5 and 2.6 is somewhat small. On the other hand, one can easily check that the sufficient optimality condition from Proposition 2.6 guarantees that $(\bar{u},\bar{v})$ is already a minimizer of the convex optimization problem

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J(u,v) \\ \text{subject to} \quad & u(t) \geq 0 \qquad \text{on } \{t \in I \,|\, \bar{v}(t) \leq \varepsilon\} \\ & u(t) = 0 \qquad \text{on } \{t \in I \,|\, \bar{v}(t) > \varepsilon\} \\ & v(t) \geq 0 \qquad \text{on } \{t \in I \,|\, \bar{u}(t) \leq \varepsilon\} \\ & v(t) = 0 \qquad \text{on } \{t \in I \,|\, \bar{u}(t) > \varepsilon\}, \end{aligned}$$

whose feasible set is in some sense *far away from* $\mathbb{C}$ even in the setting of strict complementarity, i.e. when $I^{00}(\bar{u},\bar{v}) = \varnothing$ holds, as long as $\varepsilon > 0$ is valid. In this regard, the presented sufficient optimality condition is quite restrictive.

3 How to penalize control complementarity constraints?

In this section, we are going to discuss different penalty approaches which can be used to solve (OCP) numerically. They are motivated by respective ideas from finite-dimensional complementarity programming, see, e.g., [20, 21, 24, 34], and differ w.r.t. the question whether or not the complementarity condition is decomposed (into an equilibrium condition *and* nonnegativity constraints).

3.1 Coupled versus decoupled penalties

Let $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\beta_k\}_{k\in\mathbb{N}}$ be sequences of penalty parameters tending to ∞ as $k \to \infty$. First of all, it is clear that for any NCP-function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$, we have

$$\mathbb{C} = \{(w,z) \in H^1(I)^2 \,|\, \varphi(u(t),v(t)) = 0 \text{ for all } t \in I\}.$$

Thus, it is a nearby idea to study the associated penalized optimization problem

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J_\varphi^k(u,v) := \tfrac{1}{2}\,\|S(u,v)-y_\mathrm{d}\|_{L^2(\Omega)}^2 + \tfrac{\lambda_1}{2}\,\|u\|_{H^1(I)}^2 + \tfrac{\lambda_2}{2}\,\|v\|_{H^1(I)}^2 + \alpha_k \Phi(u,v) \\ \text{subject to} \quad & (u,v) \in H^1(I)^2 \end{aligned} \tag{$\mathrm{P}_\varphi(\alpha_k)$}$$

where $\Phi\colon H^1(I)^2 \to \mathbb{R}$ is the penalty term given by

$$\forall (u,v) \in H^1(I)^2\colon \quad \Phi(u,v) := \tfrac{1}{2}\int_I \varphi^2(u(t),v(t))\,\mathrm{d}t.$$

Due to $H^1(I) \hookrightarrow C(\bar{I})$ and the continuity of φ, the integrand in the definition of Φ is continuous on $\bar{I}$ for each choice of $(u,v) \in H^1(I)^2$ which means that Φ is actually well-defined. Above, squaring the underlying NCP-function does not only ensure that the term under the integral is positive but also enhances the smoothness of the penalty term. However, let us mention that there exist NCP-functions whose domain is already a subset of the nonnegative reals, see [14], and in this case, all the results of Section 3.2 hold even if the integrand φ is not squared. A detailed analysis of the above approach where φ is chosen as the Fischer–Burmeister function can be found in [8, Section 4]. An inherent advantage of this approach is that Φ is continuously Fréchet differentiable in this case since the squared Fischer–Burmeister function is smooth. This observation already has been used in the context of finite-dimensional complementarity programming in [21]. Further information on the use of NCP-functions in the context of optimal control can be found in [38]. Due to the fact that the overall complementarity constraint is penalized within one term, we refer to this approach as coupled penalization.

Next, we observe that (OCP) is equivalent to

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J(u,v) \\ \text{subject to} \quad & (u,v) \in H^1_+(I)^2,\ \int_I u(t)v(t)\,\mathrm{d}t = 0. \end{aligned}$$

Thus, due to the appearing sign conditions on the controls, the consideration of the associated penalized optimization problem

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J_{\ell_1}^k(u,v) := \tfrac{1}{2}\,\|S(u,v)-y_\mathrm{d}\|_{L^2(\Omega)}^2 + \tfrac{\lambda_1}{2}\,\|u\|_{H^1(I)}^2 + \tfrac{\lambda_2}{2}\,\|v\|_{H^1(I)}^2 + \alpha_k \mathrm{P}(u,v) \\ \text{subject to} \quad & (u,v) \in H^1_+(I)^2 \end{aligned} \tag{$\mathrm{P}_{\ell_1}(\alpha_k)$}$$

where $\mathrm{P}\colon H^1(I) \to \mathbb{R}$ is given by

$$\forall (u,v) \in H^1(I)^2\colon \quad \mathrm{P}(u,v) := \int_I u(t)v(t)\,\mathrm{d}t$$

is reasonable as well. Clearly, P is well-defined since we have $uv \in L^1(I)$ for each pair $(u,v) \in H^1(I)^2$. Observing that

$$\forall (u,v) \in H^1_+(I)^2\colon \quad \int_I u(t)v(t)\,\mathrm{d}t = \int_I |u(t)v(t)|\,\mathrm{d}t = \|uv\|_{L^1(I)}$$

holds, P may be interpreted as a classical ℓ_1-penalty function associated with the equilibrium condition

$$u(t)v(t) = 0 \quad \text{for all } t \in I \tag{3.1}$$

on $H^1_+(I)^2$. The restriction of the domain to $H^1_+(I)^2$ is essential here. However, it needs to be noted that due to inequality constraints in H^1, the numerical solution of ($\mathrm{P}_{\ell_1}(\alpha_k)$) is still a challenging issue since the first-order optimality system associated with this program comprises Lagrange multipliers from $\mathcal{M}_-(\bar{I})$, see Proposition 3.7 and [12] for details.

In order to overcome this difficulty, let us fix a continuous function $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ with the property

$$\forall (a,b) \in \mathbb{R}^2\colon \quad \psi(a,b) \geq 0 \ \wedge\ (\psi(a,b) = 0 \iff ab = 0) \tag{NSP}$$

where NSP abbreviates *Nonlinear Switching Penalty*, see Section 3.5. Now, consider the unconstrained penalized program

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J_\psi^k(u,v) := \tfrac{1}{2}\,\|S(u,v)-y_\mathrm{d}\|_{L^2(\Omega)}^2 + \tfrac{\lambda_1}{2}\,\|u\|_{H^1(I)}^2 + \tfrac{\lambda_2}{2}\,\|v\|_{H^1(I)}^2 \\ & \qquad + \alpha_k(\Pi(-u)+\Pi(-v)) + \beta_k \Psi(u,v) \\ \text{subject to} \quad & (u,v) \in H^1(I)^2 \end{aligned} \tag{$\mathrm{P}_\psi(\alpha_k,\beta_k)$}$$

where $\Pi\colon H^1(I) \to \mathbb{R}$ and $\Psi\colon H^1(I)^2 \to \mathbb{R}$ are given as stated below:

$$\forall (u,v) \in H^1(I)^2\colon \quad \Pi(u) := \tfrac{1}{2}\int_I \max{}^2(u(t),0)\,\mathrm{d}t \qquad \Psi(u,v) := \int_I \psi(u(t),v(t))\,\mathrm{d}t.$$

Similar as above one obtains that Π and Ψ are well-defined. Furthermore, we would like to mention that Π is continuously Fréchet differentiable, see [12, Section 4.1]. Noting that the latter approach penalizes the violation of the equilibrium condition (3.1) individually, it is closely related to the considerations in [9, 10, 11] where optimal control problems with switching constraints on the controls which only live in time are investigated. In ($\mathrm{P}_\psi(\alpha_k,\beta_k)$), the equilibrium constraint as well as the nonnegativity constraints on the controls are treated with separate penalty terms and penalty parameters which is why we speak of decoupled penalization. Let us note that the choice $\beta_k := \alpha_k$ is always possible. In numerical practice, however, it might be beneficial to control the growth of the penalty parameters associated with Π and Ψ individually.

3.2 Abstract analysis of the penalty methods

We fix sequences $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\beta_k\}_{k\in\mathbb{N}}$ of nonnegative penalty parameters tending to ∞ as $k \to \infty$. Furthermore, we fix an arbitrary NCP-function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ and a continuous function $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ which satisfies (NSP). The upcoming lemma, which shows that the penalty functionals from Section 3.1 are weakly sequentially continuous, will be important for our analysis.

Lemma 3.1. *The integral functions $\Phi, \mathrm{P}, \Psi\colon H^1(I)^2 \to \mathbb{R}$ and $\Pi\colon H^1(I) \to \mathbb{R}$ are weakly sequentially continuous.*

Proof. We only show the statement for Φ. Noting that the integrands of all the other integral functionals are continuous as well, the same arguments apply to Ψ, P, and Π.
Let $\{(u_k,v_k)\}_{k\in\mathbb{N}} \subset H^1(I)^2$ be a sequence converging weakly to $(\bar{u},\bar{v}) \in H^1(I)^2$. Due to the compactness of $H^1(I) \hookrightarrow C(\bar{I})$, the strong convergences $u_k \to \bar{u}$ and $v_k \to \bar{v}$ hold true in $C(\bar{I})$ and, thus, pointwise on $\bar{I}$. By continuity of φ, we have $\varphi^2(u_k(t),v_k(t)) \to \varphi^2(\bar{u}(t),\bar{v}(t))$ for each $t \in \bar{I}$. Moreover, due to the boundedness of $\{u_k\}_{k\in\mathbb{N}}$ and $\{v_k\}_{k\in\mathbb{N}}$ in $C(\bar{I})$ and the continuity of φ, there is a constant $c > 0$ satisfying

$$\forall k \in \mathbb{N}\,\forall t \in \bar{I}\colon \quad 0 \le \varphi^2(u_k(t),v_k(t)) \le c.$$

Thus, the dominated convergence theorem shows $\Phi(u_k,v_k) \to \Phi(\bar{u},\bar{v})$. □

Based on Lemma 3.1, we obtain the following results which generalize the considerations from [8, Section 4].

Proposition 3.2. *For fixed $k \in \mathbb{N}$, each of the programs ($\mathrm{P}_\varphi(\alpha_k)$), ($\mathrm{P}_{\ell_1}(\alpha_k)$), and ($\mathrm{P}_\psi(\alpha_k,\beta_k)$) possesses a global minimizer.*

Proof. Noting that $S\colon H^1(I)^2 \to L^2(\Omega)$ is linear and continuous while squared norms are weakly sequentially lower semicontinuous, the functional J is weakly sequentially lower semicontinuous. Now, we invoke Lemma 3.1 in order to infer that $J_\varphi^k, J_{\ell_1}^k, J_\psi^k\colon H^1(I)^2 \to \mathbb{R}$ are weakly sequentially lower semicontinuous as well. Furthermore, J_φ^k and J_ψ^k are coercive on $H^1(I)^2$ while $J_{\ell_1}^k$ is coercive on $H^1_+(I)^2$ which means that ($\mathrm{P}_\varphi(\alpha_k)$) and ($\mathrm{P}_\psi(\alpha_k,\beta_k)$) as well as ($\mathrm{P}_{\ell_1}(\alpha_k)$) possess a respective global minimizer. □

Theorem 3.3. *For each $k \in \mathbb{N}$, let $(\bar{u}_k,\bar{v}_k) \in H^1(I)^2$ be a global minimizer of ($\mathrm{P}_\varphi(\alpha_k)$) (($\mathrm{P}_{\ell_1}(\alpha_k)$) or ($\mathrm{P}_\psi(\alpha_k,\beta_k)$), respectively). Then, $\{(\bar{u}_k,\bar{v}_k)\}_{k\in\mathbb{N}}$ possesses a strongly convergent subsequence whose limit point is a global minimizer of (OCP).*

Proof. We first prove the statement for the penalty approach which exploits ($\mathrm{P}_\psi(\alpha_k,\beta_k)$). Noting that the pair of vanishing functions is feasible to ($\mathrm{P}_\psi(\alpha_k,\beta_k)$) for each $k \in \mathbb{N}$, we have $J_\psi^k(\bar{u}_k,\bar{v}_k) \le \frac{1}{2}\|y_\mathrm{d}\|^2_{L^2(\Omega)}$ for each $k \in \mathbb{N}$. This can be used to infer the boundedness of $\{(\bar{u}_k,\bar{v}_k)\}_{k\in\mathbb{N}}$ in $H^1(I)^2$. Consequently, we can extract a weakly convergent subsequence (without relabeling) with weak limit point $(\bar{u},\bar{v}) \in H^1(I)^2$. From above, we particularly have

$$\forall k \in \mathbb{N}\colon \quad \max\left(\Pi(-\bar{u}_k),\Pi(-\bar{v}_k),\Psi(\bar{u}_k,\bar{v}_k)\right) \le \tfrac{1}{2\min(\alpha_k,\beta_k)}\|y_\mathrm{d}\|^2_{L^2(\Omega)},$$

i.e. taking the limit $k \to \infty$ and exploiting the weak continuity of Π and Ψ, see Lemma 3.1, we have $\Pi(-\bar{u}) = 0, \Pi(-\bar{v}) = 0$, and $\Psi(\bar{u},\bar{v}) = 0$. Noting that $\bar{u}$ and $\bar{v}$ are continuous due to $H^1(I) \hookrightarrow C(\bar{I})$, these relations yield $\max(-\bar{u}(t),0) = 0, \max(-\bar{v}(t),0) = 0$, and $\psi(\bar{u}(t),\bar{v}(t)) = 0$ for all $t \in I$ which shows $(\bar{u},\bar{v}) \in H^1_+(I)^2$ and $\int_I \bar{u}(t)\bar{v}(t)\,\mathrm{d}t = 0$. As a result, we have $(\bar{u},\bar{v}) \in \mathbb{C}$. For an arbitrary point $(u,v) \in \mathbb{C}$, it holds $J_\psi^k(\bar{u}_k,\bar{v}_k) \le J_\psi^k(u,v) = J(u,v)$ for all $k \in \mathbb{N}$ by feasibility of (u,v) for each of the programs ($\mathrm{P}_\psi(\alpha_k,\beta_k)$). Taking the limit $k \to \infty$ while observing that J is weakly sequentially lower semicontinuous, we have

$$J(\bar{u},\bar{v}) \le \liminf_{k\to\infty} J(\bar{u}_k,\bar{v}_k) \le \limsup_{k\to\infty} J(\bar{u}_k,\bar{v}_k) \le \limsup_{k\to\infty} J_\psi^k(\bar{u}_k,\bar{v}_k) \le J(u,v),$$

i.e. $(\bar{u},\bar{v})$ is a global minimizer of (OCP). The particular choice $(u,v) := (\bar{u},\bar{v})$ yields $J(\bar{u}_k,\bar{v}_k) \to J(\bar{u},\bar{v})$. Since we have $S(\bar{u}_k,\bar{v}_k) \to S(\bar{u},\bar{v})$ in $L^2(\Omega)$ due to Lemma 2.2, we obtain

$$\lambda_1 \|\bar{u}_k\|^2_{H^1(I)} + \lambda_2 \|\bar{v}_k\|^2_{H^1(I)} \to \lambda_1 \|\bar{u}\|^2_{H^1(I)} + \lambda_2 \|\bar{v}\|^2_{H^1(I)}\,.$$

Exploiting [8, Lemma A.1] as well as $\lambda_1,\lambda_2 > 0$, we have the convergences $\|\bar{u}_k\|_{H^1(I)} \to \|\bar{u}\|_{H^1(I)}$ and $\|\bar{v}_k\|_{H^1(I)} \to \|\bar{v}\|_{H^1(I)}$. Combining this with the weak convergences $\bar{u}_k \rightharpoonup \bar{u}$ and $\bar{v}_k \rightharpoonup \bar{v}$, the desired convergences $\bar{u}_k \to \bar{u}$ and $\bar{v}_k \to \bar{v}$ follow from the fact that $H^1(I)$ is a Hilbert space.

This proof strategy can be easily adapted in order to show the theorem's assertion w.r.t. the programs ($\mathrm{P}_\varphi(\alpha_k)$) and ($\mathrm{P}_{\ell_1}(\alpha_k)$). For the latter approach, one needs to exploit that $\mathrm{P}(u,v) \geq 0$ is valid for all $(u,v) \in H^1_+(I)^2$ and that $H^1_+(I)^2$ is weakly sequentially closed. □

Noting that the appearing surrogate problems ($\mathrm{P}_\varphi(\alpha_k)$), ($\mathrm{P}_{\ell_1}(\alpha_k)$), and ($\mathrm{P}_\psi(\alpha_k,\beta_k)$) are not convex in general, they cannot be solved to global optimality in numerical practice. As a consequence, one should check whether the proposed penalty methods are capable of identifying local minimizers as well. For our respective analysis, we introduce

$$\begin{aligned}
\mathbb{B}_\varepsilon(\bar{u},\bar{v}) &:= \left\{(u,v) \in H^1(I)^2 \,\middle|\, \|u-\bar{u}\|_{H^1(I)} + \|v-\bar{v}\|_{H^1(I)} \leq \varepsilon\right\},\\
\mathbb{S}_\varepsilon(\bar{u},\bar{v}) &:= \left\{(u,v) \in H^1(I)^2 \,\middle|\, \|u-\bar{u}\|_{H^1(I)} + \|v-\bar{v}\|_{H^1(I)} = \varepsilon\right\}.
\end{aligned}$$

where $(\bar{u},\bar{v}) \in H^1(I)^2$ and $\varepsilon > 0$ are fixed. Note that $\mathbb{B}_\varepsilon(\bar{u},\bar{v})$ is weakly sequentially compact while $\mathbb{S}_\varepsilon(\bar{u},\bar{v})$ is closed.

Lemma 3.4. *Let $(\bar{u},\bar{v}) \in \mathbb{C}$ be chosen such that*

$$\forall (u,v) \in \mathbb{C} \cap (\mathbb{B}_\varepsilon(\bar{u},\bar{v})) \setminus \{(\bar{u},\bar{v})\}\colon \quad J(u,v) > J(\bar{u},\bar{v}) \tag{3.2}$$

holds for some $\varepsilon > 0$, i.e. $(\bar{u},\bar{v})$ is a strict local minimizer of (OCP) of radius ε. Then, there are $r > 0$ and $k_0 \in \mathbb{N}$ such that we have $J^k_\varphi(u,v) \geq J(\bar{u},\bar{v}) + r$ ($J^k_{\ell_1}(u,v) \geq J(\bar{u},\bar{v}) + r$ or $J^k_\psi(u,v) \geq J(\bar{u},\bar{v}) + r$, respectively) for all $k \in \mathbb{N}$ with $k \geq k_0$ and all $(u,v) \in \mathbb{S}_\varepsilon(\bar{u},\bar{v})$ which are feasible to ($\mathrm{P}_\varphi(\alpha_k)$) (($\mathrm{P}_{\ell_1}(\alpha_k)$) or ($\mathrm{P}_\psi(\alpha_k,\beta_k)$), respectively).

Proof. First, let us verify the statement for the decoupled penalty approach. We assume on the contrary that there are sequences $\{r_l\}_{l\in\mathbb{N}} \subset \mathbb{R}$ with $r_l \downarrow 0$, $\{k_l\}_{l\in\mathbb{N}}$ with $k_l \to \infty$, and $\{(u_l,v_l)\}_{l\in\mathbb{N}} \subset \mathbb{S}_\varepsilon(\bar{u},\bar{v})$ such that

$$\forall l \in \mathbb{N}\colon \quad J^{k_l}_\psi(u_l,v_l) < J(\bar{u},\bar{v}) + r_l.$$

Noting that $\{r_l\}_{l\in\mathbb{N}}$ is bounded, $\{(u_l,v_l)\}_{l\in\mathbb{N}}$ needs to be bounded as well. Thus, the latter possesses a weakly convergent subsequence (without relabeling) with weak limit point $(\tilde{u},\tilde{v}) \in \mathbb{B}_\varepsilon(\bar{u},\bar{v})$. By definition of $J^{k_l}_\psi$, the above inequality leads to

$$\forall l \in \mathbb{N}\colon \quad \max\left(\Pi(-u_l),\Pi(-v_l),\Psi(u_l,v_l)\right) \leq \frac{1}{\min(\alpha_{k_l},\beta_{k_l})}(J(\bar{u},\bar{v}) + r_l).$$

Similar as in the proof of Theorem 3.3, this can be used to obtain $(\tilde{u},\tilde{v}) \in \mathbb{C}$. Due to $(\tilde{u},\tilde{v}) \in \mathbb{B}_\varepsilon(\bar{u},\bar{v})$, we have $J(\tilde{u},\tilde{v}) \geq J(\bar{u},\bar{v})$ by (3.2). Now, we exploit the weak sequential lower semicontinuity of J in order to infer

$$\begin{aligned}
J(\bar{u},\bar{v}) \leq J(\tilde{u},\tilde{v}) \leq \liminf_{l\to\infty} J(u_l,v_l) &\leq \limsup_{l\to\infty} J(u_l,v_l)\\
&\leq \limsup_{l\to\infty} J^{k_l}_\psi(u_l,v_l) \leq \limsup_{l\to\infty} (J(\bar{u},\bar{v}) + r_l) = J(\bar{u},\bar{v}).
\end{aligned}$$

This already yields $J(\tilde{u},\tilde{v}) = J(\bar{u},\bar{v})$, and thus $(\tilde{u},\tilde{v}) = (\bar{u},\bar{v})$ by (3.2). On the other hand, the above inequalities yield $J(u_l,v_l) \to J(\tilde{u},\tilde{v})$ as $l \to \infty$, and similar arguments as in the proof of Theorem 3.3 guarantee that the strong convergences $u_l \to \tilde{u}$ and $v_l \to \tilde{v}$ hold in $H^1(I)$. Particularly, we infer the relation $(\tilde{u},\tilde{v}) \in \mathbb{S}_\varepsilon(\bar{u},\bar{v})$ since the latter set is closed. Thus, we clearly have $(\tilde{u},\tilde{v}) \neq (\bar{u},\bar{v})$ which is a contradiction.

The proofs for the other two penalty methods can be carried out in similar fashion. □

The above lemma is essential for the proof of the subsequently stated theorem.

Theorem 3.5. *Let $(\bar{u},\bar{v}) \in H^1(I)^2$ be a strict local minimizer of (OCP) satisfying (3.2) for some $\varepsilon > 0$. Then, there is some $k_0 \in \mathbb{N}$ such that ($\mathrm{P}_\varphi(\alpha_k)$) (($\mathrm{P}_{\ell_1}(\alpha_k)$) or ($\mathrm{P}_\psi(\alpha_k,\beta_k)$), respectively) possesses a local minimizer within $\mathbb{B}_\varepsilon(\bar{u},\bar{v})$ for each $k \in \mathbb{N}$ which satisfies $k \geq k_0$.*

Proof. We only proof the theorem for the coupled penalty approach which exploits the surrogate problem ($P_\varphi(\alpha_k)$). These arguments can be transferred directly to the other two approaches. First of all, we note that the optimization problem

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & J_\varphi^k(u,v) \\ \text{subject to} \quad & (u,v) \in \mathbb{B}_\varepsilon(\bar{u},\bar{v}) \end{aligned} \tag{$P_\varphi(\alpha_k,\varepsilon)$}$$

possesses a global minimizer $(\tilde{u}_k,\tilde{v}_k) \in H^1(I)^2$ for each $k \in \mathbb{N}$ since J_φ^k is weakly sequentially lower semicontinuous while $\mathbb{B}_\varepsilon(\bar{u},\bar{v})$ is nonempty and weakly sequentially compact. Clearly, $(\bar{u},\bar{v})$ is feasible to ($P_\varphi(\alpha_k,\varepsilon)$) for each $k \in \mathbb{N}$ which yields $J_\varphi^k(\tilde{u}_k,\tilde{v}_k) \leq J(\bar{u},\bar{v})$. Furthermore, we have $J_\varphi^k(u,v) > J(\bar{u},\bar{v})$ for all sufficiently large $k \in \mathbb{N}$ and all $(u,v) \in \mathbb{S}_\varepsilon(\bar{u},\bar{v})$ due to Lemma 3.4. Consequently, $(\tilde{u}_k,\tilde{v}_k)$ lies in the interior of $\mathbb{B}_\varepsilon(\bar{u},\bar{v})$. Thus, $(\tilde{u}_k,\tilde{v}_k)$ is a local minimizer of ($P_\varphi(\alpha_k)$) for sufficiently large $k \in \mathbb{N}$. □

Note that a strict local minimizer of (OCP) of radius $\varepsilon > 0$ is a strict local minimizer of (OCP) of radius ε' for each $\varepsilon' \in (0,\varepsilon)$ as well. Thus, in each neighborhood of a strict local minimizer of (OCP), the surrogate problems ($P_\varphi(\alpha_k)$), ($P_{\ell_1}(\alpha_k)$), and ($P_\psi(\alpha_k,\beta_k)$) possess a local minimizer for sufficiently large $k \in \mathbb{N}$. Particularly, all the suggested penalty methods are in position to identify strict local minimizers of (OCP). In this regard, the above considerations generalize similar results for standard MPCCs, see [20, Section 3.2]. Observe that Proposition 2.6 provides a condition which guarantees that a given feasible point of (OCP) is a strict local minimizer of that problem. Note that local minimizers of the penalized surrogate problems can be characterized with the aid of first-order optimality conditions as soon as the appearing penalty functions are sufficiently smooth, and this will be discussed exemplary for three different penalty schemes in the upcoming subsections. Particularly, due to the above observations, there is a reasonable hope that one can identify strict local minimizers of (OCP) by solving the penalized surrogate problems to stationarity while the penalty parameter tends to ∞.

3.3 The coupled approach via the squared Fischer–Burmeister function

Let us briefly comment on the penalty method suggested in [8, Section 4]. There, the coupled penalty approach has been discussed in the setting where the underlying NCP-function is chosen to be the popular Fischer–Burmeister function, see [13], given by

$$\forall (a,b) \in \mathbb{R}^2 : \quad \varphi_{\mathrm{FB}}(a,b) := \sqrt{a^2+b^2} - a - b.$$

Noting that the range of φ_{FB} is $\mathbb{R}$, the square in the definition of the associated penalty functional Φ_{FB} is important in order to obtain a meaningful penalty method. Furthermore, it has been shown in [8, Lemma 4.1] that Φ_{FB} is continuously Fréchet differentiable. Thus, with the aid of Lemma 2.3, we are in position to characterize the local minimizers of the associated penalized surrogate problems ($P_{\varphi_{\mathrm{FB}}}(\alpha_k)$) where $\{\alpha_k\}_{k\in\mathbb{N}}$ is a sequence of nonnegative penalty parameters tending to ∞ as $k \to \infty$, see [8, Proposition 4.5] as well.

Proposition 3.6. *For fixed $k \in \mathbb{N}$, let $(\bar{u}_k,\bar{v}_k) \in H^1(I)^2$ be a locally optimal solution of ($P_{\varphi_{\mathrm{FB}}}(\alpha_k)$). Then, we find an adjoint state $\bar{p}_k \in W(0,T)$ which solves the following system:*

$$\langle b\bar{p}_k, z\rangle_{L^2(\Sigma)} + \lambda_1 \langle \bar{u}_k, z\rangle_{H^1(I)} + \alpha_k \langle \varphi_{\mathrm{FB}}(\bar{u}_k,\bar{v}_k)\bar{\eta}_k, z\rangle_{L^2(I)} = 0 \quad \forall z \in H^1(I), \tag{3.3a}$$

$$\langle c\bar{p}_k, z\rangle_{L^2(\Sigma)} + \lambda_2 \langle \bar{v}_k, z\rangle_{H^1(I)} + \alpha_k \langle \varphi_{\mathrm{FB}}(\bar{u}_k,\bar{v}_k)\bar{\zeta}_k, z\rangle_{L^2(I)} = 0 \quad \forall z \in H^1(I), \tag{3.3b}$$

$$-\partial_t \bar{p}_k - \nabla\cdot(\mathrm{C}\nabla\bar{p}_k) + \mathrm{a}\bar{p}_k = 0 \quad \text{in } Q, \tag{3.3c}$$

$$\vec{n}\cdot(\nabla \mathrm{C}\bar{p}_k) + \mathrm{q}\bar{p}_k = 0 \quad \text{in } \Sigma, \tag{3.3d}$$

$$\bar{p}_k(\cdot,T) = S(\bar{u}_k,\bar{v}_k) - y_\mathrm{d} \quad \text{in } \Omega. \tag{3.3e}$$

Here, the adjoint equation (3.3c), (3.3d), (3.3e) has to be understood in weak sense. Furthermore, the functions $\bar{\eta}_u, \bar{\zeta}_k \in L^\infty(I)$ are given as stated below:

$$\forall t \in I: \quad \bar{\eta}_k(t) := \begin{cases} \dfrac{\bar{u}_k(t)}{\sqrt{\bar{u}_k^2(t)+\bar{v}_k^2(t)}} - 1 & t \notin I^{00}(\bar{u}_k,\bar{v}_k), \\ 0 & t \in I^{00}(\bar{u}_k,\bar{v}_k), \end{cases} \quad \bar{\zeta}_k(t) := \begin{cases} \dfrac{\bar{v}_k(t)}{\sqrt{\bar{u}_k^2(t)+\bar{v}_k^2(t)}} - 1 & t \notin I^{00}(\bar{u}_k,\bar{v}_k), \\ 0 & t \in I^{00}(\bar{u}_k,\bar{v}_k). \end{cases}$$

The biactive set $I^{00}(\bar{u}_k,\bar{v}_k)$ has been defined in (2.3).

In [8, Remark 4.7], it already has been noted that taking the limit in the system (3.3) does not recover the strong stationarity-type conditions from Proposition 2.5 but only a problem-tailored system of weak stationarity which does not provide information on the biactive set $I^{00}(\bar{u},\bar{v})$, i.e. the set of test functions which satisfy (2.4b) needs to be chosen much smaller.

3.4 The ℓ_1-penalty approach

Here, we want to present some more facts on the ℓ_1-penalty approach promoted via ($\mathrm{P}_{\ell_1}(\alpha_k)$). Again, let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence of nonnegative penalty parameters tending to ∞ as $k \to \infty$. First of all, we note that the bilinear penalty term P is continuously Fréchet differentiable which allows us to infer the following necessary optimality condition for ($\mathrm{P}_{\ell_1}(\alpha_k)$).

Proposition 3.7. *For fixed $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in H^1(I)^2$ be a locally optimal solution of ($\mathrm{P}_{\ell_1}(\alpha_k)$). Then, we find measures $\bar{\mu}_k, \bar{\nu}_k \in \mathcal{M}(\bar{I})$ which solve the system*

$$\langle b\bar{p}_k, z\rangle_{L^2(\Sigma)} + \lambda_1 \langle \bar{u}_k, z\rangle_{H^1(I)} + \alpha_k \langle \bar{v}_k, z\rangle_{L^2(I)} + \int_I z \mathrm{d}\bar{\mu}_k = 0 \quad \forall z \in H^1(I), \tag{3.4a}$$

$$\langle c\bar{p}_k, z\rangle_{L^2(\Sigma)} + \lambda_2 \langle \bar{v}_k, z\rangle_{H^1(I)} + \alpha_k \langle \bar{u}_k, z\rangle_{L^2(I)} + \int_I z \mathrm{d}\bar{\nu}_k = 0 \quad \forall z \in H^1(I), \tag{3.4b}$$

$$\bar{\mu}_k, \bar{\nu}_k \leq 0, \quad \int_I \bar{u}_k \mathrm{d}\bar{\mu}_k = 0, \quad \int_I \bar{v}_k \mathrm{d}\bar{\nu}_k = 0 \tag{3.4c}$$

where the adjoint state $\bar{p}_k \in W(0,T)$ is the uniquely determined weak solution of the system (3.3c), (3.3d), (3.3e).

Proof. By standard arguments, we find multipliers $\bar{\mu}_k \in H^1_+(I)^\circ \cap \{\bar{u}_k\}^\perp$ and $\bar{\nu}_k \in H^1_+(I)^\circ \cap \{\bar{v}_k\}^\perp$ which satisfy

$$\begin{aligned} 0 = \left\langle S^\star(S(\bar{u}_k, \bar{v}_k) - y_\mathrm{d}), (z_u, z_v)\right\rangle_{H^1(I)^2} &+ \lambda_1 \langle \bar{u}_k, z_u\rangle_{H^1(I)} + \lambda_2 \langle \bar{v}_k, z_v\rangle_{H^1(I)} \\ &+ \alpha_k \langle \bar{v}_k, z_u\rangle_{L^2(I)} + \alpha_k \langle \bar{u}_k, z_v\rangle_{L^2(I)} \\ &+ \langle \bar{\mu}_k, z_u\rangle_{H^1(I)} + \langle \bar{\nu}_k, z_v\rangle_{H^1(I)} \end{aligned} \tag{3.5}$$

for all $(z_u, z_v) \in H^1(I)^2$. By means of [12, Lemma 3.1], we have $\bar{\mu}_k, \bar{\nu}_k \in H^1(I)^\star \cap \mathcal{M}_-(\bar{I})$. Due to $H^1(I) \hookrightarrow C(\bar{I})$, it holds $\mathcal{M}(\bar{I}) \hookrightarrow H^1(I)^\star$, i.e. we obtain the characterization $\bar{\mu}_k, \bar{\nu}_k \in \mathcal{M}_-(\bar{I})$. Since we have $\langle \bar{\mu}_k, \bar{u}_k\rangle_{H^1(I)} = \langle \bar{\nu}_k, \bar{u}_k\rangle_{H^1(I)} = 0$ from above, the condition (3.4c) can be deduced. Next, let $\bar{p}_k \in W(0,T)$ be the weak solution of (3.3c), (3.3d), (3.3e). Then, Lemma 2.3 yields

$$\left\langle S^\star(S(\bar{u}_k, \bar{v}_k) - y_\mathrm{d}), (z_u, z_v)\right\rangle_{H^1(I)^2} = \iint_\Sigma (b\bar{p}_k z_u + c\bar{p}_k z_v) \mathrm{d}s\mathrm{d}t.$$

Putting this into (3.5) and decoupling the resulting condition w.r.t. z_u and z_v yields the conditions (3.8a) and (3.8b). This completes the proof. □

The global convergence result of Theorem 3.3 only applies to the situation where ($\mathrm{P}_{\ell_1}(\alpha_k)$) can be solved to global optimality for each $k \in \mathbb{N}$. This, however, cannot be guaranteed in numerical practice since ($\mathrm{P}_{\ell_1}(\alpha_k)$) is a nonconvex program for large enough values of the penalty parameter. As a consequence, we need to study the situation where ($\mathrm{P}_{\ell_1}(\alpha_k)$) is only solved to stationarity for each $k \in \mathbb{N}$. As the subsequent result shows, this guarantees at least feasibility of weak accumulation points without further assumptions.

Proposition 3.8. *For each $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in H^1(I)^2$ be a stationary point of ($\mathrm{P}_{\ell_1}(\alpha_k)$) in the sense of Proposition 3.7. Furthermore, let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a weak accumulation point of $\{(\bar{u}_k, \bar{v}_k)\}_{k\in\mathbb{N}}$. Then, $(\bar{u}, \bar{v}) \in \mathbb{C}$ holds.*

Proof. We assume w.l.o.g. that the weak convergences $\bar{u}_k \rightharpoonup \bar{u}$ and $\bar{v}_k \rightharpoonup \bar{v}$ hold true. Noting that we have $\{(\bar{u}_k, \bar{v}_k)\}_{k\in\mathbb{N}} \subset H^1_+(I)^2$ while $H^1_+(I)$ is weakly sequentially closed, $(\bar{u}, \bar{v}) \in H^1_+(I)^2$ is obtained. It remains to show $\int_I \bar{u}(t)\bar{v}(t)\mathrm{d}t = 0$. Assume on the contrary that the measurable set

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

possesses positive measure. Due to the compactness of $H^1(I) \hookrightarrow C(\bar{I})$, we obtain the pointwise convergences $\bar{u}_k(t) \to \bar{u}(t)$ and $\bar{v}_k(t) \to \bar{v}(t)$ for all $t \in \bar{I}$. Particularly, the continuity of $\bar{u}$ and $\bar{v}$ yields the existence of a measurable set $\mathcal{I} \subset I^{++}(\bar{u}, \bar{v})$ of positive measure $|\mathcal{I}|$ as well as of a constant $\varepsilon > 0$ such that $\bar{u}_k(t), \bar{v}_k(t) \geq \varepsilon$ is valid for all $k \in \mathbb{N}$ and all $t \in \mathcal{I}$.

For each $k \in \mathbb{N}$, let $\bar{p}_k \in W(0,T)$ and $\bar{\mu}_k, \bar{\nu}_k \in \mathcal{M}(\bar{I})$ be the Lagrange multipliers which solve the system (3.4). Testing (3.4a) with $\bar{u}_k$ and (3.4b) with $\bar{v}_k$ while exploiting (3.4c) yields

$$\begin{aligned} \langle b\bar{p}_k, \bar{u}_k\rangle_{L^2(\Sigma)} + \lambda_1 \|\bar{u}_k\|^2_{H^1(I)} + \alpha_k \langle \bar{u}_k, \bar{v}_k\rangle_{L^2(I)} &= 0, \\ \langle c\bar{p}_k, \bar{v}_k\rangle_{L^2(\Sigma)} + \lambda_2 \|\bar{v}_k\|^2_{H^1(I)} + \alpha_k \langle \bar{u}_k, \bar{v}_k\rangle_{L^2(I)} &= 0. \end{aligned} \tag{3.6}$$

Noting that the compactness of S, see Lemma 2.2, yields $S(\bar{u}_k, \bar{v}_k) \to S(\bar{u}, \bar{v})$, the continuity of the solution operator associated with (APDE) guarantees $\bar{p}_k \to \bar{p}$ in $W(0,T)$ where $\bar{p} \in W(0,T)$ is the uniquely determined weak solution of

$$\begin{aligned} -\partial_t \bar{p} - \nabla \cdot (\mathrm{C}\nabla\bar{p}) + \mathrm{a}\bar{p} &= 0 && \text{in } Q \\ \vec{n} \cdot (\mathrm{C}\nabla\bar{p}) + \mathrm{q}\bar{p} &= 0 && \text{in } \Sigma \\ \bar{p}(\cdot, T) &= S(\bar{u}, \bar{v}) - y_\mathrm{d} && \text{in } \Omega. \end{aligned}$$

Thus, due to the boundedness of $\{(\bar{u}_k, \bar{v}_k)\}_{k\in\mathbb{N}}$ in $H^1(I)^2$, and the convergences $\bar{u}_k \to \bar{u}$ and $\bar{v}_k \to \bar{v}$ in $L^2(I)$, (3.6) can be used to infer the boundedness of $\{\alpha_k \langle \bar{u}_k, \bar{v}_k\rangle_{L^2(I)}\}_{k\in\mathbb{N}}$. On the other hand, the above considerations show

$$\alpha_k \langle \bar{u}_k, \bar{v}_k\rangle_{L^2(I)} \geq \alpha_k \int_{\mathcal{I}} \bar{u}_k(t)\bar{v}_k(t)\mathrm{d}t \geq \alpha_k \varepsilon^2 |\mathcal{I}|$$

for each $k \in \mathbb{N}$. Since $\{\alpha_k\varepsilon^2|\mathcal{I}|\}_{k\in\mathbb{N}}$ is not bounded due to $\varepsilon^2|\mathcal{I}| > 0$, this is a contradiction. Hence, the set $I^{++}(\bar{u}, \bar{v})$ must be of measure zero. The continuity of $\bar{u}$ and $\bar{v}$ gives the stronger condition $I^{++}(\bar{u}, \bar{v}) = \emptyset$. This completes the proof. □

From the finite-dimensional setting, see e.g. [20, Theorem 2.1] or [24, Theorem 3.4], it is well known that the ℓ_1-penalty approach does not yield strongly stationary points in general if one takes the limit in the stationarity system associated with the penalized surrogate problems. Consequently, one cannot await that taking the limit $k \to \infty$ in the system (3.4) recovers a strong stationarity-type condition in the sense of Proposition 2.5 which characterizes the limit point. However, following [20, 24], it might be possible to infer weaker stationarity-type conditions which provide some information on the biactive set. As we pointed out in Section 3.3, this is not possible for the coupled Fischer–Burmeister penalty method. A detailed investigation of this issue is, however, beyond the scope of this paper.

Below, we comment on the length of non-complementary control arcs associated with locally optimal solutions of ($\mathrm{P}_{\ell_1}(\alpha_k)$). The associated result is an immediate consequence of similar considerations for switching-constrained optimal control problems, see [10].

Remark 3.9. *For fixed $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in H^1(I)^2$ be a locally optimal solution of ($\mathrm{P}_{\ell_1}(\alpha_k)$). Assume that $\alpha_k \geq \|S\|^2 + \max(\lambda_1, \lambda_2) + \pi^2$ holds where $\|S\|$ denotes the norm of the linear operator S w.r.t. the space of all bounded, linear operators mapping from $H^1(I)^2$ to $L^2(\Omega)$ and π represents the popular mathematical constant. Adapting the arguments provided in the proof of [10, Theorem 3.2] (which, actually, applies directly since all functions appearing in the proof of this result are nonnegative), one obtains that the controls $\bar{u}_k$ and $\bar{v}_k$ are pointwise complementary apart from intervals of length at most $\sqrt{\max(\lambda_1, \lambda_2)}$.*

Finally, we would like to comment on the convexity of the functional $J^k_{\ell_1}$ for small values of the penalty parameter α_k. Noting that we have

$$\begin{aligned} J^k_{\ell_1}(u, v) = \tfrac{1}{2}\|S(u, v) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|\partial_t u\|^2_{L^2(I)} + \tfrac{\lambda_2}{2}\|\partial_t v\|^2_{L^2(I)} \\ + \tfrac{1}{2}\left((\lambda_1 - \alpha_k)\|u\|^2_{L^2(I)} + (\lambda_2 - \alpha_k)\|v\|^2_{L^2(I)} + \alpha_k\|u + v\|^2_{L^2(I)}\right) \end{aligned}$$

for all $(u, v) \in H^1(I)^2$, ($\mathrm{P}_{\ell_1}(\alpha_k)$) is a convex, smooth program for all $k \in \mathbb{N}$ such that $\alpha_k \leq \min(\lambda_1, \lambda_2)$ holds. In this case, the necessary optimality conditions from Proposition 3.7 are also sufficient. Particularly, for large regularization parameters λ_1 and λ_2, there is some hope that solving the convex program ($\mathrm{P}_{\ell_1}(\min(\lambda_1, \lambda_2))$) yields a reasonable approximate of an optimal solution for (OCP). A related idea has been exploited in [11] in order to tackle switching-constrained optimal control problems. In order to avoid dealing with Lagrange multipliers from $\mathcal{M}_-(\bar{I})$, ($\mathrm{P}_{\ell_1}(\min(\lambda_1, \lambda_2))$) can be solved by the simple penalty approach from [12], i.e., one solves the sequence of programs

$$\begin{aligned} \underset{u,v}{\text{minimize}} \quad & \tfrac{1}{2}\|S(u, v) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|^2_{H^1(I)} + \tfrac{\lambda_2}{2}\|v\|^2_{H^1(I)} \\ & + \min(\lambda_1, \lambda_2)\mathrm{P}(u, v) + \tfrac{\alpha_k}{2}(\Pi(-u) + \Pi(-v)) \\ \text{subject to} \quad & (u, v) \in H^1(I)^2, \end{aligned}$$

e.g. by applying a semismooth Newton method to the associated (nonsmooth) system of first-order (necessary and sufficient) optimality condition.

3.5 The decoupled ℓ_2-penalty approach

Let us now focus on the decoupled penalty approach promoted via ($\mathrm{P}_\psi(\alpha_k, \beta_k)$) where $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\beta_k\}_{k\in\mathbb{N}}$ are sequences of nonnegative real numbers tending to ∞ as $k \to \infty$. In order to specify the penalty term Ψ, we have to fix a continuous function $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ which satisfies (NSP). In what follows, such functions will be referred to as NSP-functions. Possible choices for ψ are given e.g. by

- $\mathbb{R}^2 \ni (a, b) \mapsto |ab| \in \mathbb{R}$,
- $\mathbb{R}^2 \ni (a, b) \mapsto \min(|a|, |b|) \in \mathbb{R}$,
- $\mathbb{R}^2 \ni (a, b) \mapsto |a| + |b| - \sqrt{a^2 + b^2} \in \mathbb{R}$,

where the last two functions are constructed by exploiting the fact that whenever $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ is an NCP-function with nonnegative values on $\mathbb{R}^2_+$, then the map $\mathbb{R}^2 \ni (a,b) \mapsto \varphi(|a|,|b|) \in \mathbb{R}$ is an NSP-function. Clearly, the functions mentioned above are nonsmooth which leads to nonsmoothness of the respective associated penalty function Ψ. As a consequence, one needs to exploit subdifferential constructions from nonsmooth analysis, see e.g. [7, 28], in order to derive a first-order optimality system for $(\mathrm{P}_\psi(\alpha_k,\beta_k))$ which seems to be an essential drawback in light of our idea to solve the first-order systems of the penalized problems with a Newton-type method. Namely, this would then require the use of second-order subdifferential constructions which we want to avoid here. On the other hand, it has to be admitted that only nonsmooth penalty functions are likely to provide exact penalization in general, see [15, Theorem 5.9].

Another reasonable choice for an NSP-function is given by

$$\forall (a,b) \in \mathbb{R}^2\colon \quad \bar\psi(a,b) := \tfrac{1}{2}a^2b^2, \tag{3.7}$$

and we will focus on this particular function in the following. Clearly, $\bar\psi$ is smooth but highly nonlinear. Furthermore, its derivative at the origin simply vanishes which is why we cannot await any promising dual convergence results beyond (if at all) weak stationarity, see [8, Remark 4.7]. On the other hand, the subsequent lemma shows that the associated ℓ_2-penalty functional Ψ is smooth and, thus, we are in position to easily infer a first-order necessary optimality condition for the associated sequence of penalized problems $(\mathrm{P}_{\bar\psi}(\alpha_k,\beta_k))$.

Lemma 3.10. *Let $\Psi\colon H^1(I)^2 \to \mathbb{R}$ be the mapping associated with the functional $\bar\psi\colon \mathbb{R}^2 \to \mathbb{R}$ defined in (3.7). Then, Ψ is continuously Fréchet differentiable. For each $(\bar u,\bar v) \in H^1(I)^2$, we obtain the subsequent formula for the associated Fréchet derivative $\Psi'(\bar u,\bar v)$:*

$$\forall (\delta_u,\delta_v) \in H^1(I)^2\colon \quad \Psi'(\bar u,\bar v)(\delta_u,\delta_v) = \left\langle \bar u\bar v^2, \delta_u\right\rangle_{L^2(I)} + \left\langle \bar u^2\bar v, \delta_v\right\rangle_{L^2(I)}.$$

Proof. For the proof, we exploit the chain rule for Fréchet differentiable mappings, see [37, Theorem 2.20]. First of all, we note that the embedding $H^1(I) \hookrightarrow L^4(I)$ is continuous, see [1, Theorem 4.12]. Let $\mathrm{E}\colon H^1(I) \to L^4(I)$ represent this embedding. Second, we observe that the Nemytskii-operator Ψ^{N} associated with $\bar\psi$ from (3.7) maps from $L^4(I)^2$ to $L^1(I)$. Exploiting Hölder's inequality, one can easily check that the Nemytskii-operator associated with $\nabla\bar\psi\colon \mathbb{R}^2 \to \mathbb{R}^2$ maps $L^4(I)^2$ to $L^{4/3}(I)^2$. Due to [16, Theorems 4 and 7], $\Psi^{\mathrm{N}}\colon L^4(I)^2 \to L^1(I)$ is continuously Fréchet differentiable at $(\mathrm{E}\bar u, \mathrm{E}\bar v)$ and the following formula holds:

$$\forall (\theta_u,\theta_v) \in L^4(I)\colon \quad (\Psi^{\mathrm{N}})'(\mathrm{E}\bar u,\mathrm{E}\bar v)(\theta_u,\theta_v) = \bar u\bar v^2\theta_u + \bar u^2\bar v\theta_v.$$

Third, setting $\mathrm{L}(f) := \int_I f(t)\,\mathrm{d}t$ for each $f \in L^1(I)$, $\mathrm{L}\colon L^1(I) \to \mathbb{R}$ is a continuous linear operator. Thus, observing that $\Psi = \mathrm{L}\circ\Psi^{\mathrm{N}}\circ(\mathrm{E},\mathrm{E})$ holds true, the desired result follows from the chain rule. □

Combining the above lemma with the smoothness of the penalty functional Π, we infer that the problem $(\mathrm{P}_{\bar\psi}(\alpha_k,\beta_k))$ is smooth for each $k \in \mathbb{N}$ and $\bar\psi$ from (3.7). Thus, we obtain the following first-order optimality condition exploiting [12, Lemma 4.2], Lemma 2.3, and Lemma 3.10.

Proposition 3.11. *For fixed $k \in \mathbb{N}$, let $(\bar u_k,\bar v_k) \in H^1(I)^2$ be a locally optimal solution of $(\mathrm{P}_{\bar\psi}(\alpha_k,\beta_k))$ where $\bar\psi$ is the function defined in (3.7). Then, the weak solution $\bar p_k \in W(0,T)$ of the system (3.3c), (3.3d), (3.3e) satisfies the conditions*

$$\langle b\bar p_k, z\rangle_{L^2(\Sigma)} + \lambda_1\langle \bar u_k, z\rangle_{H^1(I)} + \left\langle \min(0,\alpha_k\bar u_k) + \beta_k\bar u_k\bar v_k^2, z\right\rangle_{L^2(I)} = 0 \quad \forall z \in H^1(I), \tag{3.8a}$$

$$\langle c\bar p_k, z\rangle_{L^2(\Sigma)} + \lambda_2\langle \bar v_k, z\rangle_{H^1(I)} + \left\langle \min(0,\alpha_k\bar v_k) + \beta_k\bar u_k^2\bar v_k, z\right\rangle_{L^2(I)} = 0 \quad \forall z \in H^1(I). \tag{3.8b}$$

Unluckily, the above necessary optimality condition is in general not sufficient since $(\mathrm{P}_{\bar\psi}(\alpha_k,\beta_k))$ is not convex. However, we obtain the following result which addresses the situation where $(\mathrm{P}_{\bar\psi}(\alpha_k,\beta_k))$ is solved only to stationarity for each $k \in \mathbb{N}$, see Proposition 3.8 as well.

Proposition 3.12. *For each $k \in \mathbb{N}$, let $(\bar u_k,\bar v_k) \in H^1(I)^2$ be a stationary point of $(\mathrm{P}_{\bar\psi}(\alpha_k,\beta_k))$, where $\bar\psi$ is the function defined in (3.7), in the sense of Proposition 3.11. Furthermore, let $(\bar u,\bar v) \in H^1(I)^2$ be a weak accumulation point of $\{(\bar u_k,\bar v_k)\}_{k\in\mathbb{N}}$. Then, $(\bar u,\bar v) \in \mathbb{C}$ holds.*

Proof. Let us assume w.l.o.g. that the weak convergences $\bar u_k \rightharpoonup \bar u$ and $\bar v_k \rightharpoonup \bar u$ hold true. The compactness of $H^1(I) \hookrightarrow C(\bar I)$ guarantees the convergences $\bar u_k(t) \to \bar u(t)$ and $\bar v_k(t) \to \bar v(t)$ for each $t \in \bar I$. For later use, let us define measurable sets $I^{u-}(\bar u,\bar v)$ and $I^{v-}(\bar u,\bar v)$ by

$$I^{u-}(\bar u,\bar v) := \{t \in I \,|\, \bar u(t) < 0\}, \qquad I^{v-}(\bar u,\bar v) := \{t \in I \,|\, \bar v(t) < 0\}.$$

Furthermore, we test (3.8a) with $\bar{u}_k$ and (3.8b) with $\bar{v}_k$ in order to obtain

$$\begin{aligned}\langle b\bar{p}_k, \bar{u}_k\rangle_{L^2(\Sigma)} + \lambda_1 \|\bar{u}_k\|^2_{H^1(I)} + \alpha_k \int_{I^{u-}(\bar{u},\bar{v})} \bar{u}_k^2(t)\,\mathrm{d}t + \beta_k \|\bar{u}_k\bar{v}_k\|^2_{L^2(I)} = 0,\\ \langle c\bar{p}_k, \bar{v}_k\rangle_{L^2(\Sigma)} + \lambda_2 \|\bar{v}_k\|^2_{H^1(I)} + \alpha_k \int_{I^{v-}(\bar{u},\bar{v})} \bar{v}_k^2(t)\,\mathrm{d}t + \beta_k \|\bar{u}_k\bar{v}_k\|^2_{L^2(I)} = 0.\end{aligned} \tag{3.9}$$

Similar to the proof of Proposition 3.8, we obtain the boundedness of $\langle b\bar{p}_k, \bar{u}_k\rangle_{L^2(\Sigma)}+\lambda_1 \|\bar{u}_k\|^2_{H^1(I)}$ and $\langle c\bar{p}_k, \bar{v}_k\rangle_{L^2(\Sigma)}+\lambda_2 \|\bar{v}_k\|^2_{H^1(I)}$ in $\mathbb{R}$. Due to $\alpha_k \to \infty$ and $\beta_k \to \infty$ as $k \to \infty$, (3.9) now guarantees $\int_{I^{u-}(\bar{u},\bar{v})} \bar{u}_k^2 \mathrm{d}t \to 0$, $\int_{I^{v-}(\bar{u},\bar{v})} \bar{v}_k^2 \mathrm{d}t \to 0$, and $\|\bar{u}_k\bar{v}_k\|^2_{L^2(I)} \to 0$. The pointwise convergences $\bar{u}_k \to \bar{u}$ and $\bar{v}_k \to \bar{v}$ as well as the continuity of $\bar{u}$ and $\bar{v}$ can thus be used to infer $I^{u-}(\bar{u},\bar{v}) = I^{v-}(\bar{u},\bar{v}) = \emptyset$, i.e. $(\bar{u},\bar{v}) \in H^1_+(I)^2$. Finally, we exploit $\|\bar{u}_k\bar{v}_k\|^2_{L^2(I)} \to 0$ in order to obtain $I^{++}(\bar{u},\bar{v}) = \emptyset$ as in the proof of Proposition 3.8. This shows $(\bar{u},\bar{v}) \in \mathbb{C}$ and completes the proof. □

By construction of the penalty term Ψ via $\bar{\psi}$ from (3.7), it is clear that taking the limit $k \to \infty$ in the stationarity conditions (3.8) (if possible, see Proposition 3.12) does not provide any information on the biactive set $I^{00}(\bar{u},\bar{v})$. Consequently, at most weak stationarity-type conditions can be inferred for this method at the limit point.

4 Computational implementation of the penalty schemes

In this section, we discuss the computational implementation of the three penalty schemes which were introduced in Section 3.2. Particularly, we investigate

- the coupled penalty approach which exploits the Fischer–Burmeister function φ_{FB}, see Section 3.3,
- the ℓ_1-penalty approach from Section 3.4, and
- the decoupled ℓ_2-penalty approach from Section 3.5 where the underlying penalty term is induced by the function $\bar{\psi}$ from (3.7).

Thus, we have to investigate how the associated surrogate problems ($\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k)$), ($\mathrm{P}_{\ell_1}(\alpha_k)$), and ($\mathrm{P}_{\bar{\psi}}(\alpha_k,\beta_k)$) can be solved numerically where $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\beta_k\}_{k\in\mathbb{N}}$ are sequences of positive real penalty parameters tending to ∞ as $k \to \infty$. Noting that ($\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k)$) and ($\mathrm{P}_{\bar{\psi}}(\alpha_k,\beta_k)$) are unconstrained, smooth problems, this is not an issue for the coupled penalty approach using φ_{FB} and the decoupled penalty approach which exploits $\bar{\psi}$. On the other hand, the computational solution of the constrained problem ($\mathrm{P}_{\ell_1}(\alpha_k)$) is more challenging since we already observed in Proposition 3.7 that first-order optimality conditions for this problem comprise Lagrange multipliers from a measure space. In order to handle this issue, we adapt the approach from [12], i.e. we solve ($\mathrm{P}_{\ell_1}(\alpha_k)$) with the aid of a penalty method where the penalty term addresses the nonnegativity constraints on the controls. More precisely, for fixed $k \in \mathbb{N}$ and a sequence $\{\gamma_l\}_{l\in\mathbb{N}}$ of positive penalty parameters tending to ∞ as $l \to \infty$, we consider

$$\begin{aligned}\underset{u,v}{\text{minimize}} \quad & \tfrac{1}{2}\|S(u,v)-y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|^2_{H^1(I)} + \tfrac{\lambda_2}{2}\|v\|^2_{H^1(I)} \\ & \qquad + \alpha_k \mathrm{P}(u,v) + \gamma_l(\Pi(-u)+\Pi(-v)) \\ \text{subject to} \quad & (u,v) \in H^1(I)^2,\end{aligned} \tag{$\mathrm{P}_{\ell_1}(\alpha_k,\gamma_l)$}$$

which is unconstrained. Observing that the penalty term $\mathrm{P}(u,v)$ may take negative values, the existence of solutions to ($\mathrm{P}_{\ell_1}(\alpha_k,\gamma_l)$) has to be discussed. Observing that P and Π are weakly sequentially continuous, see Lemma 3.1, this would follow by H^1-coercivity of the objective functional associated with ($\mathrm{P}_{\ell_1}(\alpha_k,\gamma_l)$). The latter property is discussed in the subsequent lemma.

Lemma 4.1. *For fixed $k \in \mathbb{N}$ and each $l \in \mathbb{N}$ such that $\gamma_l > \frac{1}{2}\alpha_k^2 \max(1/\lambda_1, 1/\lambda_2)$ holds, the objective functional of ($\mathrm{P}_{\ell_1}(\alpha_k,\gamma_l)$) is H^1-coercive.*

Proof. Fix a pair of functions $(u,v) \in H^1(I)^2$. The negative part of the objective map associated with ($\mathrm{P}_{\ell_1}(\alpha_k,\gamma_l)$) is given by

$$\alpha_k \int_{I_{-+}\cup I_{+-}} u(t)v(t)\mathrm{d}t$$

where we used

$$I_{-+} := \{t \in I \,|\, u(t) < 0 \ \wedge\ v(t) > 0\}, \qquad I_{+-} := \{t \in I \,|\, u(t) > 0 \ \wedge\ v(t) < 0\}.$$

Thus, we observe

$$\alpha_k \mathrm{P}(u,v) + \gamma_l(\Pi(-u) + \Pi(-v)) \geq \int_{I_{-+}} \big(\alpha_k u(t)v(t) + \gamma_l u^2(t)\big)\mathrm{d}t + \int_{I_{+-}} \big(\alpha_k u(t)v(t) + \gamma_l v^2(t)\big)\mathrm{d}t.$$

Let us investigate the first integral on the right hand side. Due to $v(t) > 0$ for almost every $t \in I_{-+}$, the integrand is negative only in situations where $-\frac{\alpha_k}{\gamma_l} v(t) < u(t) < 0$ holds and the minimum value of the integrand is achieved for $u(t) = -\frac{\alpha_k}{2\gamma_l} v(t)$. Thus, it holds

$$\int_{I_{-+}} \big(\alpha_k u(t)v(t) + \gamma_l u^2(t)\big)\mathrm{d}t \geq -\frac{\alpha_k^2}{4\gamma_l} \int_{I_{-+}} v^2(t)\mathrm{d}t \geq -\frac{\alpha_k^2}{4\gamma_l} \int_I v^2(t)\mathrm{d}t.$$

Similarly, we obtain

$$\int_{I_{+-}} \big(\alpha_k u(t)v(t) + \gamma_l v^2(t)\big)\mathrm{d}t \geq -\frac{\alpha_k^2}{4\gamma_l} \int_{I_{-+}} u^2(t)\mathrm{d}t \geq -\frac{\alpha_k^2}{4\gamma_l} \int_I u^2(t)\mathrm{d}t.$$

Thus, we have

$$\begin{aligned}
&\tfrac{\lambda_1}{2} \|u\|^2_{H^1(I)} + \tfrac{\lambda_2}{2} \|v\|^2_{H^1(I)} + \alpha_k \mathrm{P}(u,v) + \gamma_l(\Pi(-u) + \Pi(-v)) \\
&\quad \geq \tfrac{\lambda_1}{2} \|\partial_t u\|^2_{L^2(I)} + \tfrac{\lambda_2}{2} \|\partial_t v\|^2_{L^2(I)} + \underbrace{\left(\tfrac{\lambda_1}{2} - \tfrac{\alpha_k^2}{4\gamma_l}\right)}_{>0} \|u\|^2_{L^2(I)} + \underbrace{\left(\tfrac{\lambda_2}{2} - \tfrac{\alpha_k^2}{4\gamma_l}\right)}_{>0} \|v\|^2_{L^2(I)}
\end{aligned}$$

from the lemma's assumption, and the assertion follows. □

Using standard arguments like in the proof of Theorem 3.3, one now can show that a sequence $\{(u_k^l, v_k^l)\}_{l\in\mathbb{N}}$ of global minimizers associated with $(\mathrm{P}_{\ell_1}(\alpha_k, \gamma_l))$ for fixed $k \in \mathbb{N}$ possesses a strongly convergent subsequence whose limit point is a global minimizer of $(\mathrm{P}_{\ell_1}(\alpha_k))$. In this regard, the consideration of $(\mathrm{P}_{\ell_1}(\alpha_k, \gamma_l))$ is meaningful for large enough penalty parameters γ_l.

4.1 A Discretization method

For the numerical treatment, the so-called direct or first-discretize-then-optimize approach is used. We exploit this approach in the flavor "all-at-ones", i.e. the respective overall optimality system will be discretized as one (potentially very large) nonlinear system. It should be mentioned that by using this approach, we fix the time step sizes of both parabolic PDEs (for the state equation directly and for the adjoint equation indirectly) once and for all times. The advantage of this approach is that the difficulty of dealing with the backward-in-time adjoint equation coupled with the forward-in-time state equation is no longer present. Note that this approach results in a slightly different discrete system than its "all-at-ones" counterpart for the indirect (discretize-then-optimize) method. For a discussion of direct vs. indirect method, we refer to [2]. There is also a wide variety of papers devoted to the numerical solution of parabolic optimal control problems using a specific software implementing the direct "all-at-ones" method, see [29, 31, 42]. For (linear) parabolic optimal control problems, there are multi-grid methods available, see e.g. [4]. Note that it is not necessary to deal with the linear/nonlinear system directly (as we will do it here), see again [29, 42]. However, having an FEM toolkit like OOPDE at hand, see [32], and together with MATLAB's capabilities to handle matrices, it is straight-forward to transfer the linearized optimality conditions almost one-to-one into (prototype) software.

For simplicity, we restrict our considerations to the implicit Euler or BDF-1 scheme for time discretization since it is the easiest unconditional A-stable discretization scheme, see [5]. Consequently, we will transfer the discretized state equation into matrix-vector form, where the unknown controls, the associated state, and some adjoint state will appear on the left-hand-side while only the initial value for the state is on the right-hand-side.

In the following, we describe the tools we use to construct the discrete optimality system associated with $(\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k))$, $(\mathrm{P}_{\ell_1}(\alpha_k, \gamma_l))$, and $(\mathrm{P}_{\tilde{\psi}}(\alpha_k, \beta_k))$. For our computations, we choose a tessellation Ω_Δ of Ω. In case where Ω is one-dimensional, Ω_Δ is chosen to be a family of subintervals. If $d = 2$ or $d = 3$ holds, Ω_Δ is a family of triangles or triangular pyramids, respectively. For simplicity, we assume that Ω is a bounded polygon which guarantees $\Omega = \Omega_\Delta$. For later use, let n_p and n_e denote the number of vertices and elements in Ω_Δ, respectively. All H^1-functions will be discretized by piecewise linear elements (ansatz space V_{Ω_Δ}) while piecewise constant elements are used in order to represent L^2- and L^∞-functions (ansatz space W_{Ω_Δ}). Thus, a mixed finite element method is used for the discretization w.r.t. spacial variables. The discrete inner product in V_{Ω_Δ} can be computed with the aid of the associated the mass matrix $M^1_{\Omega_\Delta}$ and stiffness matrix K_{Ω_Δ}. Analogously, the discrete inner product in W_{Ω_Δ} can be obtained with the associated mass matrix $M^0_{\Omega_\Delta}$. The matrix $E_{\Omega_\Delta} \in \mathbb{R}^{n_e \times n_p}$, which maps from V_{Ω_Δ} to W_{Ω_Δ}, represents the finite-dimensional counterpart of the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$. The time interval

$I = (0,T)$ is partitioned into a family $I_\Delta := \{[t_{i-1}, t_i]\}_{i=1}^n$ of n subintervals with $t_0 := 0$ and $t_n := T$. Since the controls u and v are elements $H^1(I)$, they will be discretized in V_{I_Δ}. In the case where controls have to be measured w.r.t. the L^2-norm, we make use the transformation matrix E_{I_Δ} in order to represent the controls in space W_{I_Δ} of piecewise constant functions. Similar as above, we introduce the mass matrices $M^1_{I_\Delta}$ and $M^0_{I_\Delta}$ as well as the stiffness matrix K_{I_Δ} in order to represent the inner product in V_{I_Δ} and W_{I_Δ}. By definition, it holds $y(t) \in H^1(\Omega)$ for each $y \in W(0,T)$ and $t \in I$. Without a new notation, we will identify $y(t)$ (as an abstract function) with its coefficient vector w.r.t. the spatial discretization. If reasonable, we apply this convention to all other functions as well. Furthermore, we assume that all coefficient functions are column vectors. In this sense, the state at time t_i, $i = 0, \dots, n$, is denoted by y^i. Particularly, y^0 represents the discretized initial state $y(0)$ while y^n is used for the discretized terminal state $y(T)$. Similarly, u^i and v^i are exploited to represent the discretized controls at time t_i, $i = 0, \dots, n$. The discretized state $y := (y^0, \dots, y^n)$ will be interpreted as a vector of length $(n+1)n_p$ while the discretized controls $u := (u^0, \dots, u^n)$ and $v := (v^0, \dots, v^n)$ are vectors of length $n+1$.

Due to the above comments, the desired state $y_\mathrm{d} \in L^2(\Omega)$ will be discretized by functions from W_{Ω_Δ}. For simplicity, the resulting coefficient vector will be denoted by $y_\mathrm{d} \in \mathbb{R}^{n_e}$ again. Consequently, the discretized objective functionals of the surrogate problems ($\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k)$), ($\mathrm{P}_{\ell_1}(\alpha_k, \gamma_l)$), and ($\mathrm{P}_{\tilde{\psi}}(\alpha_k, \beta_k)$) are given as stated below:

$$
\begin{aligned}
\tilde{J}^k_{\varphi_{\mathrm{FB}}}(y,u,v) &:= \tfrac{1}{2}(E_{\Omega_\Delta} y^n - y_\mathrm{d})^\top M^0_{\Omega_\Delta}(E_{\Omega_\Delta} y^n - y_\mathrm{d}) + \tfrac{\lambda_1}{2} u^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right) u + \tfrac{\lambda_2}{2} v^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right) v \\
&\quad + \tfrac{\alpha_k}{2} \left(\left(E_{I_\Delta} u^2 + E_{I_\Delta} v^2\right)^{\frac{1}{2}} - E_{I_\Delta} u - E_{I_\Delta} v\right)^\top M^0_{I_\Delta} \left(\left(E_{I_\Delta} u^2 + E_{I_\Delta} v^2\right)^{\frac{1}{2}} - E_{I_\Delta} u - E_{I_\Delta} v\right), \\
\tilde{J}^{k,l}_{\ell_1}(y,u,v) &:= \tfrac{1}{2}(E_{\Omega_\Delta} y^n - y_\mathrm{d})^\top M^0_{\Omega_\Delta}(E_{\Omega_\Delta} y^n - y_\mathrm{d}) + \tfrac{\lambda_1}{2} u^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right) u + \tfrac{\lambda_2}{2} v^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right) v \\
&\quad + \alpha_k (E_{I_\Delta} u)^\top M^0_{I_\Delta}(E_{I_\Delta} v) \\
&\quad + \tfrac{\gamma_l}{2} \max(0, -E_{I_\Delta} u)^\top M^0_{I_\Delta} \max(0, -E_{I_\Delta} u) + \tfrac{\gamma_l}{2} \max(0, -E_{I_\Delta} v)^\top M^0_{I_\Delta} \max(0, -E_{I_\Delta} v), \\
\tilde{J}^k_{\tilde{\psi}}(y,u,v) &:= \tfrac{1}{2}(E_{\Omega_\Delta} y^n - y_\mathrm{d})^\top M^0_{\Omega_\Delta}(E_{\Omega_\Delta} y^n - y_\mathrm{d}) + \tfrac{\lambda_1}{2} u^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right) u + \tfrac{\lambda_2}{2} v^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right) v \\
&\quad + \tfrac{\alpha_k}{2} \max(0, -E_{I_\Delta} u)^\top M^0_{I_\Delta} \max(0, -E_{I_\Delta} u) + \tfrac{\alpha_k}{2} \max(0, -E_{I_\Delta} v)^\top M^0_{I_\Delta} \max(0, -E_{I_\Delta} v) \\
&\quad + \tfrac{\beta_k}{2} (u \bullet v)^\top E^\top_{I_\Delta} M^0_{I_\Delta} E_{I_\Delta} (u \bullet v).
\end{aligned}
\tag{4.1}
$$

Above, $u \bullet v$ denotes the componentwise product of $u, v \in \mathbb{R}^{n+1}$. Furthermore, $u^2 := u \bullet u$ holds and the square root as well as the maximum have to be interpreted componentwise. In order to discretize the weak formulation (2.1) of the state equation (PDE), we first have to deal with the data functions C, a, q, b, and c. Spacial discretization is carried out w.r.t. W_{Ω_Δ}. Following our convention, we use $\mathrm{C}(t)$, $\mathrm{a}(t)$, and $\mathrm{q}(t)$ for the semi-discrete approximations of C, a, and q, respectively. Additional matrices and vectors will be used (in the semi-discretized form): the stiffness matrix associated with the diffusion matrix $\mathrm{C}(t)$ is denoted by $K_{\Omega_\Delta}(\mathrm{C}(t))$, the mass matrix associated with $\mathrm{a}(t)$ is represented by $M^1_{\Omega_\Delta}(\mathrm{a}(t))$, $Q_{\Omega_\Delta}(\mathrm{q}(t))$ is the matrix associated with the boundary integral of involving $\mathrm{q}(t)$, and $G_{\Omega_\Delta}(b)$ as well as $G_{\Omega_\Delta}(c)$ are vectors representing the boundary condition depending on b and c, respectively. Using this notation, we obtain the semi-discretized state equation, now a system of ODEs, as

$$
\begin{aligned}
\partial_t M^1_{\Omega_\Delta} y(t) + (K_{\Omega_\Delta}(\mathrm{C}(t)) + Q_{\Omega_\Delta}(\mathrm{q}(t)) + M^1_{\Omega_\Delta}(\mathrm{a}(t)))y(t) &= G_{\Omega_\Delta}(b)u(t) + G_{\Omega_\Delta}(c)v(t), \\
M^1_{\Omega_\Delta} y^0 &= 0.
\end{aligned}
\tag{4.2}
$$

The last step is now to discretze the ODE (4.2) in time. Let $\delta t_i := t_{i+1} - t_i$, $i = 0, \dots, n-1$, be the step sizes associated with I_Δ. Furthermore, we set

$$
\Theta^i_{\Omega_\Delta} := M^1_{\Omega_\Delta} + \delta t_i \left(K_{\Omega_\Delta}(\mathrm{C}^{i+1}) + Q_{\Omega_\Delta}(\mathrm{q}^{i+1}) + M^1_{\Omega_\Delta}(\mathrm{a}^{i+1})\right)
$$

for all $i = 0, \dots, n-1$ where the matrices $K_{\Omega_\Delta}(\mathrm{C}^{i+1})$, $Q_{\Omega_\Delta}(\mathrm{q}^{i+1})$, and $M_{\Omega_\Delta}(\mathrm{a}^{i+1})$ are assembled with $\mathrm{C}(t_{i+1})$, $\mathrm{q}(t_{i+1})$, and $\mathrm{a}(t_{i+1})$, respectively. Discretizing the derivative w.r.t. time in (4.2) with forward differences, we obtain the fully discretized system

$$
\begin{aligned}
\Theta^i_{\Omega_\Delta} y^{i+1} - M^1_{\Omega_\Delta} y^i &= \delta t_i \left(G_{\Omega_\Delta}(b) u^{i+1} + G_{\Omega_\Delta}(c) v^{i+1}\right) \qquad i = 0, \dots, n-1, \\
M^1_{\Omega_\Delta} y^0 &= 0.
\end{aligned}
$$

Collecting the discretized state and controls in a column vector $z := [y, u, v]$ and rearranging all appearing matrices, the above iterative solution process can be represented as the linear system

$$
[\mathrm{A} \,|\, -\mathrm{F}(b) \,|\, -\mathrm{F}(c)]\, z = 0, \tag{4.3}
$$

where the block matrix $\mathrm{A} \in \mathbb{R}^{(n+1)n_p \times (n+1)n_p}$ is given by

$$\mathrm{A} := \begin{pmatrix} M^1_{\Omega_\Delta} & \mathbb{0} & \cdots & \cdots & \cdots & \mathbb{0} \\ -M^1_{\Omega_\Delta} & \Theta^0_{\Omega_\Delta} & \ddots & & & \vdots \\ \mathbb{0} & -M^1_{\Omega_\Delta} & \Theta^1_{\Omega_\Delta} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -M^1_{\Omega_\Delta} & \Theta^{n-2}_{\Omega_\Delta} & \mathbb{0} \\ \mathbb{0} & \cdots & \cdots & \mathbb{0} & -M^1_{\Omega_\Delta} & \Theta^{n-1}_{\Omega_\Delta} \end{pmatrix}$$

while the block matrix $\mathrm{F}(\xi) \in \mathbb{R}^{(n+1)n_p \times (n+1)}$ reads as

$$\mathrm{F}(\xi) := \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & \delta t_0 G_{\Omega_\Delta}(\xi) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \delta t_{n-2} G_{\Omega_\Delta}(\xi) & 0 \\ 0 & \cdots & \cdots & 0 & \delta t_{n-1} G_{\Omega_\Delta}(\xi) \end{pmatrix}$$

for $\xi \in \{b, c\}$. Above, $\mathbb{0}$ denotes an all-zero matrix of the dimension $n_p \times n_p$, and 0 represents the all-zero vector of length n_p.

Equipping each of the potential objective functionals from (4.1) with the linear constraints (4.3), we obtain three finite-dimensional minimization problems which are the discrete counterparts of ($\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k)$), ($\mathrm{P}_{\ell_1}(\alpha_k, \gamma_l)$), and ($\mathrm{P}_{\tilde{\psi}}(\alpha_k, \beta_k)$), respectively. Based on their individual first-order necessary optimality conditions which read as a system of nonsmooth equations, respectively, a semismooth Newton method can be used to solve these problems. Let us note that the three problems possess the same constraints (4.3) which will be addressed via the discrete Lagrange multiplier $p \in \mathbb{R}^{(n+1)n_p}$. Observe that we use the "first-discretize-then-optimize" approach here which is why this discrete Lagrange multiplier is not generally equal to the solution of the discretized adjoint equation (APDE) which would come into play whenever the "first-optimize-then-discretize" approach would be used, see e.g. [33]. Let us note that the discrete first-order optimality system for a time-stationary counterpart of problem ($\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k)$) can be found in [8, Section 5.1].

4.2 Conceptual algorithms

In this subsection, we describe some suitable algorithms which can be used to solve (OCP) via the discrete optimality systems obtained in Section 4.1. As already mentioned, these systems will be solved with the aid of a Newton-type method. Due to the strong local convergence properties of Newton-type methods, it is important to choose a reasonable starting point. In this regard, a discrete solution of the optimal control problem

$$\begin{aligned} &\underset{u,v}{\text{minimize}} && J(u,v) := \tfrac{1}{2}\,\|S(u,v) - y_{\mathrm{d}}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\,\|u\|^2_{H^1(I)} + \tfrac{\lambda_2}{2}\,\|v\|^2_{H^1(I)} \\ &\text{subject to} && (u,v) \in H^1_+(I)^2, \end{aligned} \tag{P}$$

where the equilibrium condition in (OCP) is neglected, is computed for the initial guess. Noting that (P) is convex, this can be done efficiently with the aid of e.g. a penalty method again, see [12]. In this regard, a direct (cf. Section 4.1) or indirect (see [12]) discretization strategy is possible. Observe that whenever a minimizer of (P) is already complementary, then it is a global minimizer of (OCP) as well. For globalization of the Newton-type methods under consideration, we exploit classical damping.

Due to the different properties of the applied penalty schemes, a proper choice for the penalty parameters is very important, see e.g. Lemma 4.1. For brevity, however, we fix $\beta_k := \alpha_k$ for each $k \in \mathbb{N}$ for the consideration of ($\mathrm{P}_{\tilde{\psi}}(\alpha_k, \beta_k)$). The pseudo-code stated within Algorithm 1 presents a solution concept for (OCP) via the surrogates ($\mathrm{P}_{\varphi_{\mathrm{FB}}}(\alpha_k)$) and ($\mathrm{P}_{\tilde{\psi}}(\alpha_k, \alpha_k)$). Let us briefly note that $\|\cdot\|_M$ denotes a weighted Euclidean norm where suitable choices for the matrix M are given by

$$\begin{bmatrix} M^1_{I_\Delta} & \mathbb{O} \\ \mathbb{O} & M^1_{I_\Delta} \end{bmatrix} \qquad \text{or} \qquad \begin{bmatrix} M^1_{I_\Delta} + K_{I_\Delta} & \mathbb{O} \\ \mathbb{O} & M^1_{I_\Delta} + K_{I_\Delta} \end{bmatrix}$$

in order to represent the discretized L^2- or H^1-norm. Another reasonable choice for M is, for sure, the identity.

For the numerical solution of (OCP) via ($\mathrm{P}_{\ell_1}(\alpha_k)$), we exploit a nested penalty algorithm which makes use of the surrogate ($\mathrm{P}_{\ell_1}(\alpha_k, \gamma_l)$). As already mentioned, this problem may fail to possess a solution if α_k is large while γ_l is small. Thus, we have

Algorithm 1 Conceptional algorithm for solving (OCP) via ($P_{\varphi_{FB}}(\alpha_k)$) or ($P_{\bar{\psi}}(\alpha_k, \alpha_k)$)

S0 Let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence of positive penalty parameters tending to ∞ as $k \to \infty$. Let a tolerance tol > 0 be given. Let (u_0, v_0) be a discrete globally optimal solution of (P). Compute y_0 as a solution of the discretized state equation (4.3) with source (u_0, v_0). Compute p_0 as a solution of the discretized adjoint equation with the terminal condition $y_0(\cdot, T) - \vec{y}_d$. Set $k := 0$.

S1 Solve the discretized optimality system associated with ($P_{\varphi_{FB}}(\alpha_k)$) and ($P_{\bar{\psi}}(\alpha_k, \alpha_k)$), respectively, for fixed α_k using a damped Newton method with starting point (y_k, u_k, v_k, p_k). Let $(y_{k+1}, u_{k+1}, v_{k+1}, p_{k+1})$ be the associated solution.

S2 If $\|(u_{k+1}, v_{k+1}) - (u_k, v_k)\|_M <$ tol, then return (u_{k+1}, v_{k+1}). Otherwise, set $k := k + 1$ and go to **S1**.

Algorithm 2 Conceptional algorithm for solving (OCP) via ($P_{\ell_1}(\alpha_k)$)

S0 Let $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence of positive penalty parameters tending to ∞ as $k \to \infty$. Let tolerances $\text{tol}_1, \text{tol}_2 > 0$ be given. Fix $\sigma > 1$ and $\gamma_0 := \frac{1}{2}\sigma\alpha_0^2 \max(1/\lambda_1, 1/\lambda_2)$. Let (u_0, v_0) be a discrete globally optimal solution of (P). Compute y_0 as a solution of the discretized state equation (4.3) with source (u_0, v_0). Compute p_0 as a solution of the discretized adjoint equation with the terminal condition $y_0(\cdot, T) - \vec{y}_d$. Set $k := 0$, $l := 0$, and $(y_k^l, u_k^l, v_k^l, p_k^l) := (y_0, u_0, v_0, p_0)$.

S1 Solve the discretized optimality system of ($P_{\ell_1}(\alpha_k, \gamma_l)$) for fixed α_k and γ_l using a damped Newton method with starting point $(y_k^l, u_k^l, v_k^l, p_k^l)$. Let $(y_k^{l+1}, u_k^{l+1}, v_k^{l+1}, p_k^{l+1})$ be the associated solution.

S2 If $\|(u_k^{l+1}, v_k^{l+1}) - (u_k^l, v_k^l)\|_M < \text{tol}_1$, then go to **S3**. Otherwise, set $\gamma_{l+1} := \sigma\gamma_l$ as well as $l := l + 1$ and go to **S1**.

S3 If $\|(u_k^{l+1}, v_k^{l+1}) - (u_k^0, v_k^0)\|_M < \text{tol}_2$, then return (u_k^{l+1}, v_k^{l+1}). Otherwise, set $(y_{k+1}^0, u_{k+1}^0, v_{k+1}^0, p_{k+1}^0) := (y_k^{l+1}, u_k^{l+1}, v_k^{l+1}, p_k^{l+1})$, $l := 0$, $\gamma_0 := \frac{1}{2}\sigma\alpha_{k+1}^2 \max(1/\lambda_1, 1/\lambda_2)$, as well as $k := k + 1$ and go to **S1**.

to navigate the sequences $\{\alpha_k\}_{k\in\mathbb{N}}$ and $\{\gamma_l\}_{l\in\mathbb{N}}$ appropriately. Here, we exploit Lemma 4.1 and the subsequent comments for that purpose.

In practice, it turned out that choosing a rough tolerance tol_1 is advisable in order to bound the number of inner iterations where the penalty parameter addressing the inequality constraints is increased. Let us note that the lower bound $\frac{1}{2}\sigma\alpha_k^2 \max(1/\lambda_1, 1/\lambda_2)$ is very large whenever α_k is large or one of the regularization parameters λ_1 or λ_2 is small. In order to avoid numerical difficulties, we therefore focus on sequences $\{\alpha_k\}_{k\in\mathbb{N}}$ which increase quite slowly. In our simulations, we use $\alpha_0 := \min(\lambda_1, \lambda_2)$ since this guarantees that $P_{\ell_1}(\alpha_0, \gamma_l)$ is a convex optimization problem for each $l \in \mathbb{N}$, see Section 3.4. This way, the first run of the inner loop of Algorithm 2 computes an approximate global minimizer of $P_{\ell_1}(\alpha_0)$ which turns out to be a solid base for all remaining iterations.

4.3 Measuring feasibility

If is clear that for a pair $(u, v) \in H^1(I)^2$, it holds $\mathrm{P}(|u|, |v|) + \Pi(-u) + \Pi(-v) \geq 0$ as well as

$$(u, v) \in \mathbb{C} \quad \Longleftrightarrow \quad \mathrm{P}(|u|, |v|) + \Pi(-u) + \Pi(-v) = 0$$

where $|u|$ and $|v|$ denote the pointwise absolute values of u and v, respectively. Recall that the operators Π and P have been defined in Section 3.1.

In order to measure the violation of complementarity associated with the outputs of Algorithms 1 and 2, it is, thus, reasonable to exploit

$$\begin{aligned} \forall u, v \in \mathbb{R}^{n+1}: \quad \rho_{I_\Delta}(u, v) := {} & |E_{I_\Delta} u|^\top M^0_{I_\Delta} |E_{I_\Delta} v| \\ & + \tfrac{1}{2} \max(0, -E_{I_\Delta} u)^\top M^0_{I_\Delta} \max(0, -E_{I_\Delta} u) \\ & + \tfrac{1}{2} \max(0, -E_{I_\Delta} v)^\top M^0_{I_\Delta} \max(0, -E_{I_\Delta} v), \end{aligned} \tag{4.4}$$

see Section 4.1. Above, $|\eta|$ now denotes the componentwise absolute value of a vector η.

4.4 Checking strong stationarity

Observing that the objective functionals of ($P_{\varphi_{FB}}(\alpha_k)$), ($P_{\ell_1}(\alpha_k, \gamma_l)$), and ($P_{\bar{\psi}}(\alpha_k, \alpha_k)$) are not convex, only approximate stationary points of these problems are computed in the inner iterations of Algorithms 1 and 2 in general. Therefore, the convergence

result from Theorem 3.3 does not apply. Particularly, it is not clear that the output of Algorithms 1 and 2 is related to a local minimizer of (OCP). However, we can use the optimality conditions from Proposition 2.5 in order to check whether the computed points are at least (approximately) strongly stationary. Observing that each local minimizer of (OCP) is stationary in that sense, a failed stationarity test indicates that Algorithms 1 and 2 did not found a local minimizer. This idea is taken from [8].

Let $\bar{u}, \bar{v} \in \mathbb{R}^{n+1}$ be two discrete controls computed by Algorithm 1 or Algorithm 2 and let $\bar{y} \in \mathbb{R}^{(n+1)n_p}$ be the associated state, i.e. $(\bar{y}, \bar{u}, \bar{v})$ is a solution of (4.3). Furthermore, let $z_u, z_v \in \mathbb{R}^{n+1}$ be arbitrarily chosen and let $z_y \in \mathbb{R}^{(n+1)n_p}$ be chosen such that (z_y, z_u, z_v) is a solution of (4.1). We define

$$\begin{aligned}\Theta &:= \left(E_{\Omega_\Delta}\bar{y}^n - y_\mathrm{d}\right)^\top M^0_{\Omega_\Delta}\left(E_{\Omega_\Delta}\bar{y}^n\right) + \lambda_1 \bar{u}^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right)\bar{u} + \lambda_2 \bar{v}^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right)\bar{v}\\ \Sigma(z_u, z_v) &:= \left(E_{\Omega_\Delta}\bar{y}^n - y_\mathrm{d}\right)^\top M^0_{\Omega_\Delta}\left(E_{\Omega_\Delta}z_y^n\right) + \lambda_1 \bar{u}^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right)z_u + \lambda_2 \bar{v}^\top \left(M^1_{I_\Delta} + K_{I_\Delta}\right)z_v,\end{aligned}$$

where $\bar{y}^n$ and z_y^n denotes the respective discrete states at terminal time T. Furthermore, we set

$$\begin{aligned}I^{+0}_{I_\Delta}(\bar{u}, \bar{v}) &:= \left\{i \in \{0, 1, \ldots, n\} \,\middle|\, \bar{u}^i \geq \epsilon \,\wedge\, |\bar{v}^i| < \epsilon\right\},\\ I^{0+}_{I_\Delta}(\bar{u}, \bar{v}) &:= \left\{i \in \{0, 1, \ldots, n\} \,\middle|\, |\bar{u}^i| < \epsilon \,\wedge\, \bar{v}^i \geq \epsilon\right\},\\ I^{00}_{I_\Delta}(\bar{u}, \bar{v}) &:= \left\{i \in \{0, 1, \ldots, n\} \,\middle|\, |\bar{u}^i| < \epsilon \,\wedge\, |\bar{v}^i| < \epsilon\right\}\end{aligned}$$

for some tolerance $\epsilon > 0$ in order to define discrete counterparts for the sets from (2.3). Observe that $I^{+0}_{I_\Delta}(\bar{u}, \bar{v})$, $I^{0+}_{I_\Delta}(\bar{u}, \bar{v})$, and $I^{00}_{I_\Delta}(\bar{u}, \bar{v})$ do not necessarily provide a partition of $\{0, 1, \ldots, n\}$ whenever ϵ is small. With the aid of these sets, we define

$$\mathcal{T}_{I_\Delta}(\bar{u}, \bar{v}) := \left\{(z_u, z_v) \in \mathbb{R}^{n+1}_+ \times \mathbb{R}^{n+1}_+ \,\middle|\, \begin{aligned}&\forall i \notin I^{+0}_{I_\Delta}(\bar{u}, \bar{v}) \cup I^{00}_{I_\Delta}(\bar{u}, \bar{v}):\ z_u^i = 0\\ &\forall i \notin I^{0+}_{I_\Delta}(\bar{u}, \bar{v}) \cup I^{00}_{I_\Delta}(\bar{u}, \bar{v}):\ z_v^i = 0\end{aligned}\right\}.$$

Then, a reasonable counterpart of the stationarity system (2.4) is given by

$$\Theta \approx 0 \qquad \text{and} \qquad \forall (z_u, z_v) \in \mathcal{T}_{I_\Delta}(\bar{u}, \bar{v}):\quad \Sigma(z_u, z_v) \gtrsim 0 \tag{4.5}$$

where $\approx$ and $\gtrsim$ mean that the relations $=$ and $\geq$ have to be fulfilled up to some pre-defined tolerance, see [8] for details. We say that $(\bar{u}, \bar{v})$ passes the stationarity test whenever these conditions hold. Due to the free choice of all appearing tolerances, we have to admit, however, that the result of this stationarity test might be quite subjective. In this paper, we will not focus on details regarding the stationarity test. For its implementation, we mainly follow the strategy provided in [8].

5 Numerical examples

In this section, the three suggested penalty methods whose numerical implementation has been discussed in Section 4 will be tested in terms of two (academic) examples. These numerical experiments are implemented using the object oriented finite element MATLAB class library OOPDE, see [32]. If not stated otherwise, in Algorithm 1, we exploit $\alpha_0 := 1$ and $\alpha_k := 1.2 \cdot \alpha_{k-1}$ for all $k \in \mathbb{N}$. Furthermore, we use $\mathrm{tol} = 0.05$. For Algorithm 2, we set $\alpha_0 := \min(\lambda_1, \lambda_2)$ and $\alpha_k := 1.2 \cdot \alpha_{k-1}$ for all $k \in \mathbb{N}$. Additionally, $\sigma := 2$ and $\mathrm{tol}_1 = \mathrm{tol}_2 := 0.05$ are used. In both algorithms, the matrix M which models the norm in the respective stopping criteria is chosen as the identity. In order to document our results, let us denote the outcome of Algorithm 1 for the Fischer–Burmeister penalty approach by (u_φ, v_φ) and for the decoupled ℓ_2-penalty approach with the NSP-function from (3.7) by (u_ψ, v_ψ). Finally, we denote the output of Algorithm 2 by (u_{ℓ_1}, v_{ℓ_1}). Recall that the initial guess (u_0, v_0) for Algorithms 1 and 2 is the discrete solution of (P).

Example 1: The first example is given on a one-dimensional interval $\Omega := (0, 1)$ and the time interval $I := (0, 4)$ which are both discretized in 160 equidistant subintervals, respectively. The heat sources u is active at the boundary point $s = 0$ of Ω while v is active at $s = 1$, i.e. we use

$$\forall s \in \{0, 1\}:\quad b(s) := \begin{cases}1 & \text{if } s = 0,\\ 0 & \text{if } s = 1,\end{cases} \qquad c(s) := \begin{cases}0 & \text{if } s = 0,\\ 1 & \text{if } s = 1.\end{cases}$$

The coefficient functions $\mathrm{C} \equiv 0.0125$, $\mathrm{a} \equiv 0$, and $\mathrm{q} \equiv 1$ are fixed. Furthermore, we choose $\lambda_1 = \lambda_2 := 10^{-5}$ for the regularization parameters and $y_\mathrm{d}(x) := \sin(5\pi x^2) + x$, $x \in \Omega$, for the desired state. In Figure 1, the initial guess (u_0, v_0) as well as the outputs of Algorithms 1 and 2 are illustrated.

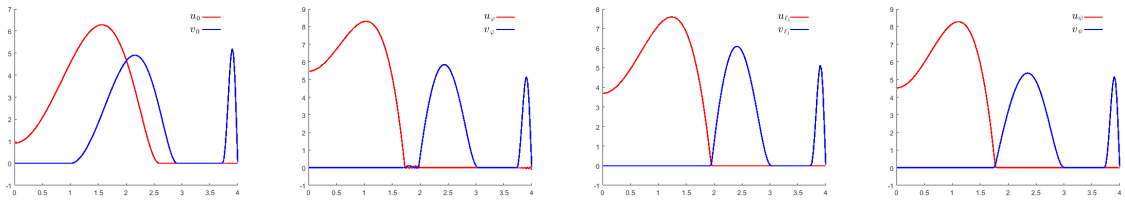

Figure 1: Example 1 - From left to right: initial guess (u_0, v_0); obtained solution (u_φ, v_φ) for $\alpha_k \approx 2.99$ (6 outer iterations); obtained solution (u_{ℓ_1}, v_{ℓ_1}) for $\alpha_k \approx 0.11$ (51 outer iterations) and $\gamma_k \approx 4.27 \cdot 10^3$; obtained solution (u_ψ, v_ψ) for $\alpha_k \approx 1.07 \cdot 10^1$ (13 outer iterations).

In Table 1, some numbers regarding computation time, objective value, and the feasibility measure from (4.4) are presented in order to compare the different outputs of Algorithms 1 and 2. Regarding computation time, these results are not surprising noting that an additional penalty loop is used in Algorithm 2 while the evaluation of the generalized second-order derivative of the squared Fischer–Burmeister function is quite expensive, see [8]. Let us note that the all three computed points pass the stationarity test in terms of a reasonably chosen tolerance.

	computation time	objective value	feasibility measure
coupled FB-penalty	62 seconds	0.1408	0.0291
ℓ_1-penalty	105 seconds	0.1401	0.0025
decoupled ℓ_2-penalty	38 seconds	0.1404	0.0011

Table 1: Example 1 - Comparison of the three penalty approaches.

Example 2: Let us fix $I := (0,1)$. We define two desired controls u_d and v_d by

$$\forall t \in I: \quad u_\mathrm{d}(t) := \max(10 - 20t, 0), \qquad v_\mathrm{d}(t) = 20(1-t)\sin(3.2\pi t).$$

The state on the unit-square $\Omega := (0,1) \times (0,1)$ is determined via the parabolic equation (PDE), where the coefficients $\mathrm{C} \equiv 0.0125$, $\mathrm{a} \equiv 0$, and $\mathrm{q} \equiv 1$ are constant while the source term is characterized via $b := \chi_B$ and $c := \chi_C$ with

$$B := \{s \in \Gamma \,|\, s_1 \geq \tfrac{1}{2}\}, \qquad C := \{s \in \Gamma \,|\, s_2 \leq \tfrac{1}{2}\}.$$

Here, $\chi_A \colon \Gamma \to \mathbb{R}$ denotes the characteristic function of a set $A \subset \Gamma$ which equals 1 on A and vanishes otherwise. Let us set $\lambda_1 = \lambda_2 := 10^{-5}$. Furthermore, let z_d be the solution of (PDE) for the controls u_d and v_d from above and set $y_\mathrm{d}(x) := z_\mathrm{d}(x,1)$ for each $x \in \Omega$. Let us note that the controls u_d and v_d are not complementary, see Figure 2, which leads to a meaningful associated problem (OCP). The spacial domain Ω is discretized with the mesh-width $h = 0.0825$ while the time interval is partitioned into 100 equidistant intervals. In order to run Algorithm 1 for the coupled penalty approach based on the squared Fischer–Burmeister function, we use $\alpha_0 := 1$ and $\alpha_k := 1.05 \cdot \alpha_{k-1}$ for all $k \in \mathbb{N}$.

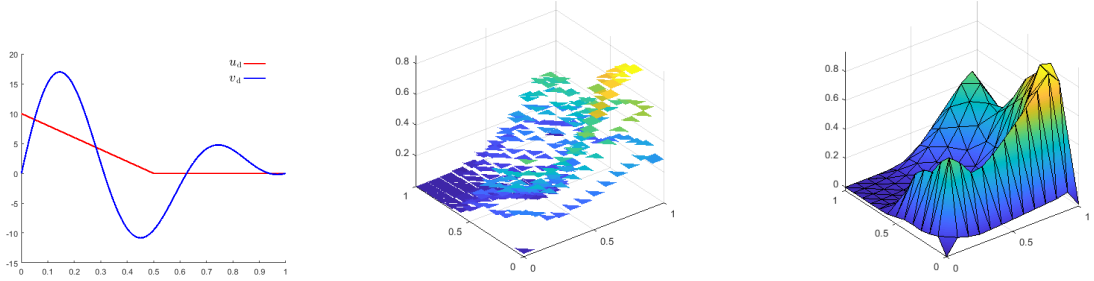

Figure 2: Example 2 - From left to right: desired controls u_d and v_d; associated desired state y_d (discretized in W_{Ω_Δ}) at the terminal time $t = 1$; terminal slice $y(\cdot, 1)$ associated with (u_{ℓ_1}, v_{ℓ_1}) (discretized in V_{Ω_Δ}).

In Figure 3, the initial guess as well as the outputs of Algorithms 1 and 2 are visualized. Let us note that the terminal slices of the resulting state variables are quite similar for all three penalty methods. Exemplary, the resulting terminal slice for the state associated with (u_{ℓ_1}, v_{ℓ_1}) is visualized in Figure 2. In contrast to the desired state, which is formally an L^2-function and, thus, discretized in W_{Ω_Δ}, the terminal state possesses Sobolev regularity and, thus, is represented in V_{Ω_Δ}.

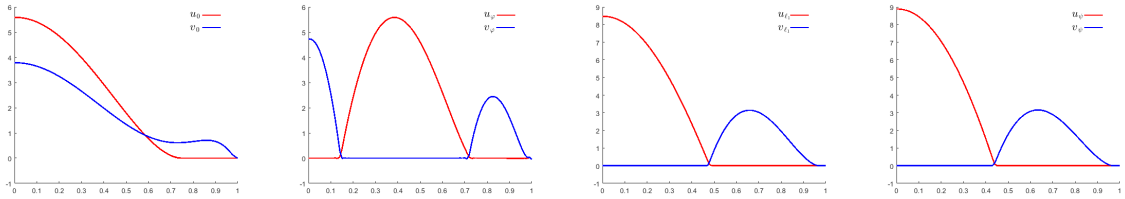

Figure 3: Example 2 - From left to right: initial guess (u_0, v_0); obtained solution (u_φ, v_φ) for $\alpha_k \approx 1.89$ (13 outer iterations); obtained solution (u_{ℓ_1}, v_{ℓ_1}) for $\alpha_k \approx 1.57 \cdot 10^{-1}$ (53 outer iterations) and $\gamma_k \approx 2.97 \cdot 10^3$; obtained solution (u_ψ, v_ψ) for $\alpha_k \approx 1.85 \cdot 10^1$ (16 outer iterations).

In Table 2, we present some numbers related to the numerical treatment of this second example. Again, all computed solutions pass the stationarity test.

	computation time	objective value	feasibility measure
coupled FB-penalty	3.3 hours	0.0072	0.0154
ℓ_1-penalty	245 seconds	0.0033	0.0016
decoupled ℓ_2-penalty	307 seconds	0.0032	0.0003

Table 2: Example 2 - Comparison of the three penalty approaches.

Summary In both examples, the ℓ_1-penalty method as well as the decoupled ℓ_2-penalty method based on the NSP-function $\bar{\psi}$ from (3.7) turned out to outrun the coupled Fischer–Burmeister penalty method not only w.r.t. obtained feasibility but also w.r.t. resulting function values. As soon as the underlying spacial domain's dimension is larger than 1, the computation time for the coupled Fischer–Burmeister penalty approach explodes. This seems to be caused by the quite expensive evaluation of the squared Fischer–Burmeister function's derivatives. On the other hand, we observed that the ℓ_1-penalty approach and the decoupled ℓ_2-penalty method computed almost the same solutions which are not far from being complementary. In the one-dimensional setting, all three methods compute solutions quite fast. However, we observe that as soon as the dimension grows, the numerical solution of the Newton systems in Algorithm 1 involving the generalized second-order derivative of the squared Fischer–Burmeister function gets far more time consuming. Based on our results, we, thus, have to recommend the use of the ℓ_1- or decoupled ℓ_2-penalty approach for the numerical solution of (OCP), at least if the underlying domain Ω is of dimension 2 or larger, since all the three introduced penalty methods share the same theoretical properties, see Section 3.

6 Conclusions

In this paper, we investigated three different penalty approaches for the solution of optimal control problems with pointwise control complementarity constraints. In the so-called coupled penalty method, the overall complementarity condition is penalized as a whole with the aid of an NCP-function. Second, the so-called ℓ_1-penalty approach only penalizes the violation of the pointwise equilibrium condition. Noting that feasible controls are needed to be nonnegative, the resulting penalty term can be chosen to be the L^2-inner product of the controls. Third, the situation where the equilibrium condition as well as the nonnegativity requirements are penalized individually is referred to a decoupled penalty approach. It has been shown that, theoretically, all these penalty methods share the same convergence properties. In order to check the quantitative properties of all these methods, we investigated their computational implementation. In numerical practice, it turned out that both, the decoupled penalty approach where the violation of the equilibrium condition is penalized with the aid of an ℓ_2-penalty term and the ℓ_1-approach, yield comprehensive results in reasonable time. These methods clearly outran the coupled penalty method based on the squared Fischer–Burmeister function w.r.t. computation time, objective value of the computed solutions, and validity of the complementarity condition as soon as the underlying domain's dimension is at least 2. In the one-dimensional case, all three methods are competitive.

Acknowledgments

This work partially is supported by the DFG grant *Bilevel Optimal Control: Theory, Algorithms, and Applications* under grant number DE 650/10-2 within the second phase of the Priority Program SPP 1962 (Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization).

References

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

[2] J. T. Betts and S. L. Campbell. Discretize then Optimize. In D. R. Ferguson and T. J. Peters, editors, *Mathematics in Industry: Challenges and Frontiers A Process View: Practice and Theory*. SIAM Publications, Philadelphia, 2005.

[3] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2000.

[4] A. Borzi. Multigrid methods for parabolic distributed optimal control problems. *Journal of Computational and Applied Mathematics*, 157(2):365–382, 2003.

[5] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. Wiley & Sons, Chichester, 2016.

[6] C. Christof and G. Wachsmuth. On Second-Order Optimality Conditions for Optimal Control Problems Governed by the Obstacle Problem. *arXiv*, 2019.

[7] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley, 1983.

[8] C. Clason, Y. Deng, P. Mehlitz, and U. Prüfert. Optimal control problems with control complementarity constraints: existence results, optimality conditions, and a penalty method. *Optimization Methods and Software*, pages 1–29, 2019.

[9] 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, 2016.

[10] C. Clason, A. Rund, and K. Kunisch. Nonconvex penalization of switching control of partial differential equations. *Systems & Control Letters*, 106:1–8, 2017.

[11] 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, 2016.

[12] Y. Deng, P. Mehlitz, and U. Prüfert. Optimal control in first-order Sobolev spaces with inequality constraints. *Computational Optimization & Applications*, 72(3):797–826, 2019.

[13] A. Fischer. A special Newton-type optimization method. *Optimization*, 24(3-4):269–284, 1992.

[14] A. Galántai. Properties and construction of NCP functions. *Computational Optimization and Applications*, 52(3):805–824, 2012.

[15] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin, 2002.

[16] H. Goldberg, W. Kampowsky, and F. Tröltzsch. On Nemytskij operators in L_p-spaces of abstract functions. *Mathematische Nachrichten*, 155:127–140, 1992.

[17] 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.

[18] F. Harder and G. Wachsmuth. Comparison of optimality systems for the optimal control of the obstacle problem. *GAMM-Mitteilungen*, 40(4):312–338, 2018.

[19] F. Harder and G. Wachsmuth. The limiting normal cone of a complementarity set in Sobolev spaces. *Optimization*, 67(10):1579–1603, 2018.

[20] X. M. Hu and D. Ralph. Convergence of a Penalty Method for Mathematical Programming with Complementarity Constraints. *Journal of Optimization Theory and Applications*, 123(2):365–390, 2004.

[21] X. X. Huang, X. Q. Yang, and D. L. Zhu. A Sequential Smooth Penalization Approach to Mathematical Programs with Complementarity Constraints. *Numerical Functional Analysis and Optimization*, 27(1):71–98, 2006.

[22] C. Kanzow, N. Yamashita, and M. Fukushima. New NCP-Functions and Their Properties. *Journal of Optimization Theory and Applications*, 94(1):115–135, 1997.

[23] K. Kunisch and D. Wachsmuth. Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2):520–547, 2012.

[24] S. Leyffer, G. López-Calva, and J. Nocedal. Interior Methods for Mathematical Programs with Complementarity Constraints. *SIAM Journal on Optimization*, 17(1):52–77, 2006.

[25] G. Liu, J. Ye, and J. Zhu. Partial Exact Penalty for Mathematical Programs with Equilibrium Constraints. *Set-Valued Analysis*, 16(5):785, 2008.

[26] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996.

[27] 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.

[28] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation*. Springer-Verlag, 2006.

[29] I. Neitzel, U. Prüfert, and T. Slawig. Strategies for time-dependent PDE control with inequality constraints using an integrated modeling and simulation environment. *Numerical Algorithms*, 50(3):241–269, 2009.

[30] J.-S. Pang and D. E. Stewart. Differential variational inequalities. *Mathematical Programming A*, 113(2):345–424, Jun 2008.

[31] L. Petzold, J. B. Rosen, P. E. Gill, L. O. Jay, and K. Park. Numerical Optimal Control of Parabolic PDEs Using DASOPT. In L. T. Biegler, T. F. Coleman, A. R. Conn, and F. N. Santosa, editors, *Large-Scale Optimization with Applications.*, volume 93 of *The IMA Volumes in Mathematics and its Applications*. Springer, New York, 1997.

[32] U. Prüfert. *OOPDE: An object oriented toolbox for finite elements in Matlab*. TU Bergakademie Freiberg, 2015.

[33] U. Prüfert. *Solving optimal PDE control problems. Optimality conditions, algorithms and model reduction.* TU Bergakademie Freiberg, 2016.

[34] D. Ralph and S. J. Wright. Some properties of regularization and penalization schemes for MPECs. *Optimization Methods and Software*, 19(5):527–556, 2004.

[35] S. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.

[36] D. Sun and L. Qi. On NCP-Functions. *Computational Optimization and Applications*, 13(1):201–220, 1999.

[37] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. American Mathematical Society, 2010.

[38] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM, 2011.

[39] G. Wachsmuth. Mathematical programs with complementarity constraints in Banach spaces. *Journal on Optimization Theory and Applications*, 166:480–507, 2015.

[40] J. Wloka. *Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben*. Teubner, 1982.

[41] J. J. Ye. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *J. Math. Anal. Appl.*, 307:350–369, 2005.

[42] F. Yılmaz and B. Karasözen. An all-at-once approach for the optimal control of the unsteady Burgers equation. *Journal of Computational and Applied Mathematics*, 259:771 – 779, 2014.