

COUPLED *VERSUS* DECOUPLED PENALIZATION OF CONTROL COMPLEMENTARITY CONSTRAINTS

YU DENG^{1,*}, PATRICK MEHLITZ² AND UWE PRÜFERT¹

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 non-negativity 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, some quantitative properties of the resulting algorithms are compared.

Mathematics Subject Classification. 49K20, 49M05, 49M25.

Received October 28, 2019. Accepted February 21, 2021.

1. 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 [27, 35, 41, 46, 52, 53, 57] 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 [10, 20, 38], see [26, 51, 52] as well for the numerical treatment of pointwise complementarity constraints in Lebesgue spaces with the aid of semismooth Newton-type methods. Furthermore, the interested reader is exemplary referred to [21–25, 41, 47, 54, 55] where complementarity constraints are considered which arise in the context of the optimal control of (quasi-)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

Keywords and phrases: Complementarity constraints, optimal control, parabolic PDE, penalty method.

¹ Faculty of Mathematics and Computer Science, Technische Universität Bergakademie Freiberg, Germany.

² Institute of Mathematics, Chair of Optimal Control, Brandenburgische Technische Universität Cottbus-Senftenberg, Germany.

* Corresponding author: yu.deng@math.tu-freiberg.de

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 [10], 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 MPCs, 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 non-negativity constraints $a, b \geq 0$ in the feasible set of the surrogate problem, see [28, 33, 45] 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 $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$ is called NCP-function whenever it satisfies

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

i.e. the set of roots associated with an NCP-function precisely recovers the complementarity set in \mathbb{R}^2 . A satisfying overview of NCP-functions can be found in [18, 31, 49]. For the construction of a solution method associated with a complementarity-constrained program, one can focus on penalizing the violation of the equality constraint $\varphi(a, b) = 0$ where $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$ is an NCP-function. This has been done with the aid of the (squared) Fischer–Burmeister function, see [17], in [29]. In [10], the authors used a similar idea in order to solve optimal control problems with control complementarity constraints. The theory of (partially) exact penalization in the context of finite-dimensional complementarity-constrained programming is discussed in [30, 34, 36, 48].

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 [10] 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-once approach, *i.e.* we solve the system for state, controls, and adjoint simultaneously. Generally, we rely on the so-called direct method, which is also known as first-discretize-then-optimize approach.

The paper is organized as follows: In Section 2, the notation and some function spaces used in this manuscript will be provided. Afterwards, the optimal control problem of interest will be introduced in Section 3. 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 4 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 non-negativity constraints as well as the 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 as 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 5. Furthermore, we present some comments on the practical implementation of the penalty algorithms. The different penalty approaches are compared by means of some numerical experiments in Section 6. Some conclusions in Section 7 close the paper.

2. NOTATION

2.1. 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}^* . The associated dual pairing is represented by $\langle \cdot, \cdot \rangle_{\mathcal{X}} : \mathcal{X}^* \times \mathcal{X} \rightarrow \mathbb{R}$. 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 $x_k \rightarrow \bar{x}$ and $x_k \rightharpoonup \bar{x}$, respectively. The polar and annihilator of a set $A \subset \mathcal{X}$ are defined by

$$A^\circ := \{x^* \in \mathcal{X}^* \mid \forall x \in A: \langle x^*, x \rangle_{\mathcal{X}} \leq 0\}, \quad A^\perp := \{x^* \in \mathcal{X}^* \mid \forall x \in A: \langle x^*, x \rangle_{\mathcal{X}} = 0\}.$$

2.2. Function spaces

For a nonempty, bounded domain $\Xi \subset \mathbb{R}^N$, a Banach space \mathcal{B} , and $1 \leq p \leq \infty$, $L^p(\Xi; \mathcal{B})$ denotes the common Lebesgue space of all (equivalence classes of) abstract measurable functions $u: \Xi \rightarrow \mathcal{B}$ such that $\Xi \ni \xi \mapsto \|u(\xi)\|_{\mathcal{B}} \in \mathbb{R}$ is p -integrable ($1 \leq p < \infty$) or essentially bounded ($p = \infty$). Similarly, $C(\bar{\Xi}; \mathcal{B})$ represents the space of all abstract continuous functions $u: \bar{\Xi} \rightarrow \mathcal{B}$. For brevity, we use $L^p(\Xi) := L^p(\Xi; \mathbb{R})$ for all $p \in [1, \infty]$ and $C(\bar{\Xi}) := C(\bar{\Xi}; \mathbb{R})$. For an arbitrary function $u \in L^1(\Xi)$, $\text{supp } u := \{\xi \in \Xi \mid u(\xi) \neq 0\}$ denotes the support of u . We would like to emphasize that this set is well-defined up to sets of Lebesgue measure zero. Recall that $\mathcal{M}(\bar{\Xi}) := C(\bar{\Xi})^*$ comprises all finite Borel measures on $\bar{\Xi}$. We use $\mathcal{M}_-(\bar{\Xi})$ to denote the set of all nonpositive measures from $\mathcal{M}(\bar{\Xi})$ 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 non-negative functions in $H^1(\Xi)$. Further information about these classical function spaces is presented in [1, 50].

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 $L^2(Q; \mathbb{R}^d)$. Similarly, 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 $W(0, T) := \{y \in L^2(I; H^1(\Omega)) \mid y' \in L^2(I; H^1(\Omega)^*)\}$ 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 about function spaces which are suitable for the discussion of parabolic differential equations can be found in [50, 56].

3. 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:

$$\begin{aligned} & \underset{u, v}{\text{minimize}} && J(u, v) := \frac{1}{2} \|S(u, v) - y_d\|_{L^2(\Omega)}^2 + \frac{\lambda_1}{2} \|u\|_{H^1(I)}^2 + \frac{\lambda_2}{2} \|v\|_{H^1(I)}^2 \\ & \text{subject to} && (u, v) \in \mathbb{C} := \{(w, z) \in H^1(I)^2 \mid 0 \leq w(t) \perp z(t) \geq 0 \text{ a.e. on } I\}. \end{aligned} \tag{OCP}$$

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

$$\begin{aligned} \partial_t y - \nabla \cdot (C \nabla y) + ay &= 0 && \text{a.e. on } Q \\ \vec{n} \cdot (C \nabla y) + qy &= bu + cv && \text{a.e. on } \Sigma \\ y(\cdot, 0) &= 0 && \text{a.e. on } \Omega, \end{aligned} \tag{PDE}$$

see Section 3.1. 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 $C: Q \rightarrow \mathbb{R}^{d \times d}$ represents thermal diffusivity while the functions $a: Q \rightarrow \mathbb{R}$ and $q: \Sigma \rightarrow \mathbb{R}$ 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 $b, c: \Gamma \rightarrow \mathbb{R}$ are used to model the distribution of the control effort on Γ provided by the controls $u, v: I \rightarrow \mathbb{R}$ which only live in time. Typically, b and c are characteristic functions of certain subsets of Γ which can be controlled individually. The precise assumptions on the data are stated in Assumption 3.1. Standard optimal boundary control of parabolic partial differential equations is discussed *e.g.* in Section 3 of [50].

In (OCP), we try to find *complementary* controls $(u, v) \in \mathbf{H}^1(I)^2$ such that the resulting terminal state $S(u, v)$ is close to a desired distribution of temperature $y_d: \Omega \rightarrow \mathbb{R}$ 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.*, [10, 12], \mathbf{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 \mathbf{H}^1 -norm of u and v . This leads to \mathbf{H}^1 -coercivity of the objective functional $J: \mathbf{H}^1(I)^2 \rightarrow \mathbb{R}$ and, thus, ensures the existence of a global minimizer. It needs to be noted that $\mathbf{H}^1(I)$ is compactly embedded into $C(\bar{I})$, see ([1], Thm. 6.3), which means that we actually have

$$\mathbb{C} = \{(w, z) \in \mathbf{H}^1(I)^2 \mid 0 \leq (\iota w)(t) \perp (\iota z)(t) \geq 0 \text{ for all } t \in I\},$$

where $\iota: \mathbf{H}^1(I) \rightarrow C(\bar{I})$ represents this embedding. These observations will be the base for our upcoming analysis. Let us stress that ι assigns to each function from $\mathbf{H}^1(I)$ a continuous *representative* in $C(\bar{I})$ and, thus, acts on $\mathbf{H}^1(I)$ by changing the input function on a set of measure zero. Subsequently, we will abstain from mentioning ι explicitly for brevity of notation.

The precise assumptions on (OCP) are stated below.

Assumption 3.1. Let the desired state $y_d \in \mathbf{L}^2(\Omega)$ and regularization parameters $\lambda_1, \lambda_2 > 0$ be fixed. Furthermore, let $b, c \in \mathbf{L}^\infty(\Gamma)$ be functions that do not vanish almost everywhere. Finally, we fix $C \in \mathbf{L}^\infty(Q; \mathbb{R}^{d \times d})$, $a \in \mathbf{L}^\infty(Q)$, and $q \in \mathbf{L}^\infty(\Sigma)$ such that C possesses symmetric images and satisfies the so-called condition of uniform ellipticity, *i.e.*

$$\exists \gamma > 0 \forall \eta \in \mathbb{R}^d: \quad \eta^\top C(x, t) \eta \geq \gamma |\eta|_2^2 \quad \text{for a.e. } (x, t) \in Q, \tag{3.1}$$

while q is non-negative almost everywhere on Σ .

3.1. The state equation

For the discussion of the state equation (PDE), we mainly follow classical arguments provided in Section 3.4 of [50] where the special setting $C \equiv \mathbb{I}$ is investigated. Here, \mathbb{I} denotes the identity matrix. However, due to (3.1) and the symmetry properties of the matrix function C , the arguments from [50] easily apply to the more general setting discussed here. Using test functions from $\mathcal{W} := \{\xi \in W_2^{1,1}(Q) \mid \forall x \in \Omega: \xi(x, T) = 0\}$ while exploiting Green's formula as well as the present boundary and initial condition, the associated variational formulation of (PDE) is given by

$$\iint_Q (-y \partial_t \xi + (C \nabla y) \cdot \nabla \xi + ay \xi) \, dx dt + \iint_\Sigma qy \xi \, ds dt = \iint_\Sigma (bu + cv) \xi \, ds dt \quad \forall \xi \in \mathcal{W}. \quad (3.2)$$

A solution of the variational problem (3.2) 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 Theorem 3.12 of [50] that (3.2) possesses a unique solution in $W_2^{1,0}(Q)$ which belongs to $W(0, T)$ after a suitable modification on a set of measure zero. Exploiting the continuity of the embedding $W(0, T) \hookrightarrow C(\bar{I}, L^2(\Omega))$, the observation operator $W(0, T) \ni y \rightarrow y(\cdot, T) \in L^2(\Omega)$ is linear and continuous. Thus, we obtain the following theorem from Lemma 3.13 of [50].

Lemma 3.2. *The control-to-observation-operator $S: H^1(I)^2 \rightarrow L^2(\Omega)$ 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

$$\begin{aligned} \partial_t y - \nabla \cdot (C \nabla y) + ay &= f && \text{a.e. on } Q \\ \vec{n} \cdot (C \nabla y) + qy &= bu + cv && \text{a.e. on } \Sigma \\ y(\cdot, 0) &= y_0 && \text{a.e. on } \Omega \end{aligned}$$

where $f \in L^2(Q)$ is a fixed heat source and $y_0 \in L^2(\Omega)$ is an initial distribution of temperature, see Section 7 of [50] and Section 4 of [56]. 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 $S^*: L^2(\Omega) \rightarrow (H^1(I)^2)^*$. In order to do that, the subsequently stated adjoint equation associated with (PDE) has to be considered for some $\eta \in L^2(\Omega)$:

$$\begin{aligned} -\partial_t p - \nabla \cdot (C \nabla p) + ap &= 0 && \text{a.e. on } Q \\ \vec{n} \cdot (C \nabla p) + qp &= 0 && \text{a.e. on } \Sigma \\ p(\cdot, T) &= \eta && \text{a.e. on } \Omega. \end{aligned} \quad (\text{APDE})$$

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

$$\iint_Q (p \partial_t \zeta + (C \nabla p) \cdot \nabla \zeta + ap \zeta) \, dx dt + \iint_\Sigma qp \zeta \, ds dt = \int_\Omega \eta \zeta(\cdot, T) \, dx \quad \forall \zeta \in \mathcal{W}'. \quad (3.3)$$

Again, a solution of (3.3) will be referred to as a weak solution of (APDE). Due to Lemma 3.17 of [50], 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 3.3. For $\eta \in L^2(\Omega)$, let $p \in W(0, T)$ be the weak solution of (APDE). Then, we have

$$\forall (u, v) \in H^1(I)^2: \quad \langle S^* \eta, (u, v) \rangle_{H^1(\Omega)^2} = \iint_{\Sigma} bpu \, dsdt + \iint_{\Sigma} cpv \, dsdt.$$

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, Theorem 3.18 of [50] yields

$$\langle S^* \eta, (u, v) \rangle_{H^1(I)^2} = \langle \eta, S(u, v) \rangle_{L^2(\Omega)} = \int_{\Omega} \eta y(\cdot, T) \, dx = \iint_{\Sigma} (bu + cv)p \, dsdt,$$

and this shows the claim. \square

3.2. Existence and optimality conditions

First, we combine Theorem 3.2 and Corollary 2.4 of [10] in order to infer the existence of an optimal solution to (OCP).

Proposition 3.4. Problem (OCP) possesses an optimal solution.

It has been shown in [10] that (OCP) can be transferred into an MPCC in Banach spaces where the complementarity condition is induced by the cone of all almost everywhere non-negative functions in $L^2(I)$, see [53] as well. Naturally, reasonable constraint qualifications 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 [10] were in position to derive necessary optimality conditions, which can be applied to problem (OCP), directly. Let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a feasible point of (OCP). Then, the sets

$$\begin{aligned} I^{+0}(\bar{u}, \bar{v}) &:= \{t \in I \mid \bar{u}(t) > 0 \wedge \bar{v}(t) = 0\}, \\ I^{0+}(\bar{u}, \bar{v}) &:= \{t \in I \mid \bar{u}(t) = 0 \wedge \bar{v}(t) > 0\}, \\ I^{00}(\bar{u}, \bar{v}) &:= \{t \in I \mid \bar{u}(t) = 0 \wedge \bar{v}(t) = 0\} \end{aligned} \tag{3.4}$$

are well-defined up to sets of measure zero. Particularly, after a suitable modification on a subset of measure zero, these sets provide a disjoint partition of I . Moreover, using continuous representatives of \bar{u} and \bar{v} , the latter property is inherent. However, we stress that in the subsequent results, all inclusions regarding these index sets or the support of a H^1 -function need to hold only up to sets of measure zero.

In Corollary 3.3 of [10], the following multiplier-free necessary optimality condition has been derived.

Proposition 3.5. Let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a locally optimal solution of (OCP). Furthermore, set

$$\mathcal{T}(\bar{u}, \bar{v}) := \{(z_u, z_v) \in H_+^1(I)^2 \mid \text{supp } z_u \subset I^{+0}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v}) \wedge \text{supp } z_v \subset I^{0+}(\bar{u}, \bar{v}) \cup I^{00}(\bar{u}, \bar{v})\}.$$

Then, the following conditions hold:

$$\langle S(\bar{u}, \bar{v}) - y_d, S(\bar{u}, \bar{v}) \rangle_{L^2(\Omega)} + \lambda_1 \langle \bar{u}, \bar{u} \rangle_{H^1(I)} + \lambda_2 \langle \bar{v}, \bar{v} \rangle_{H^1(I)} = 0, \tag{3.5a}$$

$$\langle S(\bar{u}, \bar{v}) - y_d, S(z_u, z_v) \rangle_{L^2(\Omega)} + \lambda_1 \langle \bar{u}, z_u \rangle_{H^1(I)} + \lambda_2 \langle \bar{v}, z_v \rangle_{H^1(I)} \geq 0 \quad \forall (z_u, z_v) \in \mathcal{T}(\bar{u}, \bar{v}). \tag{3.5b}$$

A dual counterpart of the above primal necessary optimality condition can be found in [10]. In the latter paper, it has been justified to call this system a strong stationarity-type necessary optimality condition. For numerical purposes, the system (3.5) from Proposition 3.5 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 Section 5.2 of [10].

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

Proposition 3.6. *Let $(\bar{u}, \bar{v}) \in \mathbf{H}^1(I)^2$ be a feasible point of (OCP). Furthermore, for $\varepsilon \geq 0$, set*

$$\mathcal{T}_\varepsilon(\bar{u}, \bar{v}) := \{(z_u, z_v) \in \mathbf{H}_+^1(I)^2 \mid \text{supp } z_u \subset \{t \in I \mid \bar{v}(t) \leq \varepsilon\} \wedge \text{supp } z_v \subset \{t \in I \mid \bar{u}(t) \leq \varepsilon\}\}.$$

Now, assume that (3.5a) is valid while

$$\langle S(\bar{u}, \bar{v}) - y_d, S(z_u, z_v) \rangle_{\mathbf{L}^2(\Omega)} + \lambda_1 \langle \bar{u}, z_u \rangle_{\mathbf{H}^1(I)} + \lambda_2 \langle \bar{v}, z_v \rangle_{\mathbf{H}^1(I)} \geq 0 \quad \forall (z_u, z_v) \in \mathcal{T}_\varepsilon(\bar{u}, \bar{v}) \quad (3.6)$$

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

$$\forall (u, v) \in \mathbf{C} \cap U: \quad J(u, v) \geq J(\bar{u}, \bar{v}) + \lambda_1 \|u - \bar{u}\|_{\mathbf{H}^1(I)}^2 + \lambda_2 \|v - \bar{v}\|_{\mathbf{H}^1(I)}^2 \quad (3.7)$$

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

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

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

By linearity of S , it holds

$$\langle S(\bar{u}, \bar{v}) - y_d, S(z_u, z_v) \rangle_{\mathbf{L}^2(\Omega)} + \lambda_1 \langle \bar{u}, z_u \rangle_{\mathbf{H}^1(I)} + \lambda_2 \langle \bar{v}, z_v \rangle_{\mathbf{H}^1(I)} + \langle \bar{\mu}, z_u \rangle_{\mathbf{H}^1(I)} + \langle \bar{\nu}, z_v \rangle_{\mathbf{H}^1(I)} = 0 \quad (3.8)$$

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

$$\langle \bar{\mu}, \bar{u} \rangle_{\mathbf{H}^1(I)} + \langle \bar{\nu}, \bar{v} \rangle_{\mathbf{H}^1(I)} = 0, \quad (3.9a)$$

$$\langle \bar{\mu}, z_u \rangle_{\mathbf{H}^1(I)} + \langle \bar{\nu}, z_v \rangle_{\mathbf{H}^1(I)} \leq 0 \quad \forall (z_u, z_v) \in \mathcal{T}_\varepsilon(\bar{u}, \bar{v}) \quad (3.9b)$$

from (3.5a) and (3.6). Let us define the MPCC-Lagrangian $L: \mathbf{H}^1(I) \times \mathbf{H}^1(I) \times \mathbf{H}^1(I)^* \times \mathbf{H}^1(I)^* \rightarrow \mathbb{R}$ of (OCP) by

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

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

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

for each $(u, v) \in \mathbf{H}^1(I)^2$. Due to (3.8), 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 (3.9a) yield

$$L(u, v, \bar{\mu}, \bar{\nu}) = J(\bar{u}, \bar{v}) + \frac{1}{2} \|S(u - \bar{u}, v - \bar{v})\|_{\mathbf{L}^2(\Omega)}^2 + \lambda_1 \|u - \bar{u}\|_{\mathbf{H}^1(I)}^2 + \lambda_2 \|v - \bar{v}\|_{\mathbf{H}^1(I)}^2. \quad (3.10)$$

Due to the 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

$$\forall (u, v) \in U: \quad \bar{u}(t) > \varepsilon \implies u(t) \geq \frac{\varepsilon}{2} \quad \bar{v}(t) > \varepsilon \implies v(t) \geq \frac{\varepsilon}{2} \quad \text{for a.e. } t \in I.$$

Now, fix $(u, v) \in \mathbb{C} \cap U$. For almost every $t \in I$ 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 $\text{supp } u \subset \{t \in I \mid \bar{v}(t) \leq \varepsilon\}$. Similarly, we have $\text{supp } v \subset \{t \in I \mid \bar{u}(t) \leq \varepsilon\}$. 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 (3.9b) and (3.10) in order to obtain

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

for each $(u, v) \in \mathbb{C} \cap U$, and this shows (3.7). \square

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.* Section 3.3 of [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 [46]. However, it has been remarked in [53] 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 [8, 32]. As discussed in Remark 3.4 of [10], the necessary optimality conditions provided in Proposition 3.5 are slightly weaker than the strong stationarity conditions of (OCP) in the sense of [53]. It is, thus, not surprising that our sufficient optimality condition from Proposition 3.6 is *not* based on the strong stationarity-type conditions from Proposition 3.5 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 Proposition 3.5 and 3.6 is somewhat small. On the other hand, one can easily check that the sufficient optimality condition from Proposition 3.6 guarantees that (\bar{u}, \bar{v}) is already a minimizer of the convex optimization problem

$$\begin{array}{ll} \underset{u, v}{\text{minimize}} & J(u, v) \\ \text{subject to} & u(t) \geq 0 \quad \text{a.e. on } \{t \in I \mid \bar{v}(t) \leq \varepsilon\} \\ & u(t) = 0 \quad \text{a.e. on } \{t \in I \mid \bar{v}(t) > \varepsilon\} \\ & v(t) \geq 0 \quad \text{a.e. on } \{t \in I \mid \bar{u}(t) \leq \varepsilon\} \\ & v(t) = 0 \quad \text{a.e. on } \{t \in I \mid \bar{u}(t) > \varepsilon\}, \end{array}$$

whose feasible set is in some sense *far away from* \mathbb{C} even in the setting of strict complementarity, *i.e.* when $I^{00}(\bar{u}, \bar{v}) = \emptyset$ holds, as long as $\varepsilon > 0$ is valid. In this regard, the presented sufficient optimality condition is quite restrictive. Finally, we would like to mention that the idea of deriving second-order sufficient optimality conditions for infinite-dimensional optimization problems by *enlarging* the standard critical cone, although classical, see [16, 37], is still exploited in order to infer sufficient second-order optimality conditions for optimal control problems, see *e.g.* [6, 7] and the references therein. Our result from Proposition 3.6 is related to this approach.

4. 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., [28, 29, 33, 45], and differ w.r.t. the question whether or not the complementarity condition is decomposed (into an equilibrium condition *and* non-negativity constraints).

4.1. 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 $k \rightarrow \infty$. First of all, it is clear that for any NCP-function $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$, we have

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

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

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

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

$$\forall (u, v) \in \mathbf{H}^1(I)^2: \quad \Phi(u, v) := \frac{1}{2} \int_I \varphi^2(u(t), v(t)) \, dt.$$

Due to $\mathbf{H}^1(I) \hookrightarrow C(\bar{I})$ and the continuity of φ , the integrand in the definition of Φ is continuous on \bar{I} for each choice of $(u, v) \in \mathbf{H}^1(I)^2$ (after some modifications on a set of measure zero if necessary) 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 non-negative but also enhances the smoothness of the penalty term. However, let us mention that there exist NCP-functions whose range is already a subset of the non-negative reals, see [18], and in this case, all the results of Section 4.2 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 Section 4 of [10]. An inherent advantage of this approach is that Φ is continuously Fréchet differentiable in this case since the squared Fischer–Burmeister function is continuously differentiable. This observation already has been used in the context of finite-dimensional complementarity programming in [29]. Further information about the use of NCP-functions in the context of optimal control can be found in [52]. 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

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

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

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

where $\mathbf{P}: \mathbf{H}^1(I) \rightarrow \mathbb{R}$ is given by

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

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

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

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

$$u(t)v(t) = 0 \quad \text{for a.e. } t \in I \quad (4.1)$$

on $\mathbf{H}_+^1(I)^2$. The restriction of the domain to $\mathbf{H}_+^1(I)^2$ is essential here. However, it needs to be noted that due to inequality constraints in \mathbf{H}^1 , the numerical solution of $(\mathbf{P}_{\ell_1}(\alpha_k))$ 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 4.7 and [14] for details.

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

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

where NSP abbreviates *Nonlinear Switching Penalty*, see Section 4.5 for details. Now, consider the unconstrained penalized program

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

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

$$\forall (u, v) \in \mathbf{H}^1(I)^2: \quad \Pi(u) := \frac{1}{2} \int_I \max^2(u(t), 0) dt \quad \Psi(u, v) := \int_I \psi(u(t), v(t)) dt.$$

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 Section 4.1 of [14]. Noting that the latter approach penalizes the violation of the equilibrium condition (4.1) individually, it is closely related to the considerations in [11–13] where optimal control problems with switching constraints on the controls which only live in time are investigated. In $(\mathbf{P}_\psi(\alpha_k, \beta_k))$, the equilibrium constraint as well as the non-negativity 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.

4.2. Abstract analysis of the penalty methods

We fix sequences $\{\alpha_k\}_{k \in \mathbb{N}}$ and $\{\beta_k\}_{k \in \mathbb{N}}$ of non-negative penalty parameters tending to ∞ as $k \rightarrow \infty$. Furthermore, we fix an arbitrary NCP-function $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$ and a continuous function $\psi: \mathbb{R}^2 \rightarrow \mathbb{R}$ which satisfies (NSP). The upcoming lemma, which shows that the penalty functionals from Section 4.1 are weakly sequentially continuous, will be important for our analysis.

Lemma 4.1. *The integral functions $\Phi, \mathbf{P}, \Psi: \mathbf{H}^1(I)^2 \rightarrow \mathbb{R}$ and $\Pi: \mathbf{H}^1(I) \rightarrow \mathbb{R}$ 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, \mathbf{P} , and Π .

Let $\{(u_k, v_k)\}_{k \in \mathbb{N}} \subset \mathbf{H}^1(I)^2$ be a sequence converging weakly to $(\bar{u}, \bar{v}) \in \mathbf{H}^1(I)^2$. Due to the compactness of $\mathbf{H}^1(I) \hookrightarrow \mathbf{C}(\bar{I})$, the strong convergences $u_k \rightarrow \bar{u}$ and $v_k \rightarrow \bar{v}$ hold true w.l.o.g. in $\mathbf{C}(\bar{I})$ and, thus, pointwise on \bar{I} . By continuity of φ , we have $\varphi^2(u_k(t), v_k(t)) \rightarrow \varphi^2(\bar{u}(t), \bar{v}(t))$ 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 $\mathbf{C}(\bar{I})$ and the continuity of φ , there is a constant $c > 0$ satisfying

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

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

Based on Lemma 4.1, we obtain the following results which generalize the considerations from Section 4 of [10].

Proposition 4.2. *For fixed $k \in \mathbb{N}$, each of the programs $(\mathbf{P}_\varphi(\alpha_k))$, $(\mathbf{P}_{\ell_1}(\alpha_k))$, and $(\mathbf{P}_\psi(\alpha_k, \beta_k))$ possesses a global minimizer.*

Proof. Noting that $S: \mathbf{H}^1(I)^2 \rightarrow \mathbf{L}^2(\Omega)$ is linear and continuous while squared norms are weakly sequentially lower semicontinuous, the functional J is weakly sequentially lower semicontinuous. Now, we invoke Lemma 4.1 in order to infer that $J_\varphi^k, J_{\ell_1}^k, J_\psi^k: \mathbf{H}^1(I)^2 \rightarrow \mathbb{R}$ are weakly sequentially lower semicontinuous as well. Furthermore, J_φ^k and J_ψ^k are coercive on $\mathbf{H}^1(I)^2$ while $J_{\ell_1}^k$ is coercive on $\mathbf{H}_+^1(I)^2$ which means that $(\mathbf{P}_\varphi(\alpha_k))$ and $(\mathbf{P}_\psi(\alpha_k, \beta_k))$ as well as $(\mathbf{P}_{\ell_1}(\alpha_k))$ possess a respective global minimizer. \square

Theorem 4.3. *For each $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in \mathbf{H}^1(I)^2$ be a global minimizer of $(\mathbf{P}_\varphi(\alpha_k))$ ($(\mathbf{P}_{\ell_1}(\alpha_k))$ or $(\mathbf{P}_\psi(\alpha_k, \beta_k))$, 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). Moreover, the weak limit of each weakly convergent subsequence of $\{(\bar{u}_k, \bar{v}_k)\}_{k \in \mathbb{N}}$ is a strong accumulation point of $\{(\bar{u}_k, \bar{v}_k)\}_{k \in \mathbb{N}}$ and a global minimizer of (OCP).*

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

$$\forall k \in \mathbb{N}: \quad \max(\Pi(-\bar{u}_k), \Pi(-\bar{v}_k), \Psi(\bar{u}_k, \bar{v}_k)) \leq \frac{1}{2 \min(\alpha_k, \beta_k)} \|y_d\|_{\mathbf{L}^2(\Omega)}^2,$$

i.e. taking the limit $k \rightarrow \infty$ and exploiting the weak sequential continuity of Π and Ψ , see Lemma 4.1, 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 w.l.o.g. continuous due to $\mathbf{H}^1(I) \hookrightarrow \mathbf{C}(\bar{I})$, 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 $(\bar{u}, \bar{v}) \in \mathbf{H}_+^1(I)^2$ and $\int_I \bar{u}(t) \bar{v}(t) dt = 0$. As a result, we have $(\bar{u}, \bar{v}) \in \mathbb{C}$. For an arbitrary point $(u, v) \in \mathbb{C}$, it holds $J_\psi^k(\bar{u}_k, \bar{v}_k) \leq J_\psi^k(u, v) = J(u, v)$ for all $k \in \mathbb{N}$ by feasibility of (u, v) for each of the programs $(\mathbf{P}_\psi(\alpha_k, \beta_k))$. Taking the limit $k \rightarrow \infty$ while observing that J is weakly sequentially lower semicontinuous, we have

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

i.e. (\bar{u}, \bar{v}) is a global minimizer of (OCP). The particular choice $(u, v) := (\bar{u}, \bar{v})$ yields $J(\bar{u}_k, \bar{v}_k) \rightarrow J(\bar{u}, \bar{v})$. Since we have $S(\bar{u}_k, \bar{v}_k) \rightarrow S(\bar{u}, \bar{v})$ in $\mathbf{L}^2(\Omega)$ due to Theorem 3.2, we obtain

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

Exploiting Lemma A.1 of [10] as well as $\lambda_1, \lambda_2 > 0$, we have the convergences $\|\bar{u}_k\|_{\mathbf{H}^1(I)} \rightarrow \|\bar{u}\|_{\mathbf{H}^1(I)}$ and $\|\bar{v}_k\|_{\mathbf{H}^1(I)} \rightarrow \|\bar{v}\|_{\mathbf{H}^1(I)}$. Combining this with the weak convergences $\bar{u}_k \rightharpoonup \bar{u}$ and $\bar{v}_k \rightharpoonup \bar{v}$, the desired convergences $\bar{u}_k \rightarrow \bar{u}$ and $\bar{v}_k \rightarrow \bar{v}$ follow from the fact that $\mathbf{H}^1(I)$ is a Hilbert space.

Observe that whenever we are given a weakly convergent subsequence of $\{(\bar{u}_k, \bar{v}_k)\}_{k \in \mathbb{N}}$, then we can partially reprise the above arguments in order to show that the associated weak limit point is, actually, a strong limit point and a global minimizer of (OCP).

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

Noting that the appearing surrogate problems $(\mathbf{P}_\varphi(\alpha_k))$, $(\mathbf{P}_{\ell_1}(\alpha_k))$, and $(\mathbf{P}_\psi(\alpha_k, \beta_k))$ 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

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

where $(\bar{u}, \bar{v}) \in \mathbf{H}^1(I)^2$ 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 4.4. *Let $(\bar{u}, \bar{v}) \in \mathbb{C}$ be chosen such that*

$$\forall (u, v) \in (\mathbb{C} \cap \mathbb{B}_\varepsilon(\bar{u}, \bar{v})) \setminus \{(\bar{u}, \bar{v})\}: \quad J(u, v) > J(\bar{u}, \bar{v}) \quad (4.2)$$

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 $J_\varphi^k(u, v) \geq J(\bar{u}, \bar{v}) + r$ ($J_{\ell_1}^k(u, v) \geq J(\bar{u}, \bar{v}) + r$ or $J_\psi^k(u, v) \geq J(\bar{u}, \bar{v}) + r$, 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 $(\mathbf{P}_\varphi(\alpha_k))$ ($(\mathbf{P}_{\ell_1}(\alpha_k))$ or $(\mathbf{P}_\psi(\alpha_k, \beta_k))$, 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 $r_l \downarrow 0$, $\{k_l\}_{l \in \mathbb{N}} \subset \mathbb{N}$ with $k_l \rightarrow \infty$, and $\{(u_l, v_l)\}_{l \in \mathbb{N}} \subset \mathbb{S}_\varepsilon(\bar{u}, \bar{v})$ such that

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

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 $J_\psi^{k_l}$, the above inequality leads to

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

Similar to the proof of Theorem 4.3, 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 (4.2). Now, we exploit the weak sequential lower semicontinuity of J in order to infer

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

This already yields $J(\tilde{u}, \tilde{v}) = J(\bar{u}, \bar{v})$, and thus $(\tilde{u}, \tilde{v}) = (\bar{u}, \bar{v})$ by (4.2). On the other hand, the above inequalities yield $J(u_l, v_l) \rightarrow J(\tilde{u}, \tilde{v})$ as $l \rightarrow \infty$, and similar arguments as in the proof of Theorem 4.3 guarantee that the strong convergences $u_l \rightarrow \tilde{u}$ and $v_l \rightarrow \tilde{v}$ 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. \square

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

Theorem 4.5. *Let $(\bar{u}, \bar{v}) \in H^1(I)^2$ be a strict local minimizer of (OCP) satisfying (4.2) for some $\varepsilon > 0$. Then, there is some $k_0 \in \mathbb{N}$ such that $(P_\varphi(\alpha_k))$ ($(P_{\ell_1}(\alpha_k))$ or $(P_\psi(\alpha_k, \beta_k))$, 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 prove 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

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

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 4.4. 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 $k \in \mathbb{N}$. \square

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 Section 3.2 of [28]. Observe that Proposition 3.6 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 determined 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 ∞ .

4.3. The coupled approach via the squared Fischer–Burmeister function

Let us briefly comment on the penalty method suggested in Section 4 of [10]. 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 [17], given by

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

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 Lemma 4.1 of [10] that Φ_{FB} is continuously Fréchet differentiable. Thus, with the aid of Lemma 3.3, we are in position to characterize the local minimizers of the associated penalized surrogate problems $(P_{\varphi_{\text{FB}}}(\alpha_k))$ where $\{\alpha_k\}_{k \in \mathbb{N}}$ is a sequence of non-negative penalty parameters tending to ∞ as $k \rightarrow \infty$, see Proposition 4.5 of [10] as well.

Proposition 4.6. *For fixed $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in \mathbf{H}^1(I)^2$ be a locally optimal solution of $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$. Then, we find an adjoint state $\bar{p}_k \in \mathbf{W}(0, T)$ which solves the following system:*

$$\langle b\bar{p}_k, z \rangle_{\mathbf{L}^2(\Sigma)} + \lambda_1 \langle \bar{u}_k, z \rangle_{\mathbf{H}^1(I)} + \alpha_k \langle \varphi_{\text{FB}}(\bar{u}_k, \bar{v}_k) \bar{\eta}_k, z \rangle_{\mathbf{L}^2(I)} = 0 \quad \forall z \in \mathbf{H}^1(I), \quad (4.3a)$$

$$\langle c\bar{p}_k, z \rangle_{\mathbf{L}^2(\Sigma)} + \lambda_2 \langle \bar{v}_k, z \rangle_{\mathbf{H}^1(I)} + \alpha_k \langle \varphi_{\text{FB}}(\bar{u}_k, \bar{v}_k) \bar{\zeta}_k, z \rangle_{\mathbf{L}^2(I)} = 0 \quad \forall z \in \mathbf{H}^1(I), \quad (4.3b)$$

$$-\partial_t \bar{p}_k - \nabla \cdot (C \nabla \bar{p}_k) + a \bar{p}_k = 0 \quad \text{a.e. on } Q, \quad (4.3c)$$

$$\vec{n} \cdot (C \nabla \bar{p}_k) + q \bar{p}_k = 0 \quad \text{a.e. on } \Sigma, \quad (4.3d)$$

$$\bar{p}_k(\cdot, T) = S(\bar{u}_k, \bar{v}_k) - y_d \quad \text{a.e. on } \Omega. \quad (4.3e)$$

Here, the adjoint equation (4.3c), (4.3d), (4.3e) has to be understood in the weak sense. Furthermore, the functions $\bar{\eta}_k, \bar{\zeta}_k \in \mathbf{L}^\infty(I)$ are given as stated below:

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

The biactive set $I^{00}(\bar{u}_k, \bar{v}_k)$ has been defined in (3.4).

In Remark 4.7 of [10], it already has been noted that taking the limit in the system (4.3) does not recover the strong stationarity-type conditions from Proposition 3.5 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 (3.5b) needs to be chosen much smaller.

4.4. The ℓ_1 -penalty approach

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

Proposition 4.7. *For fixed $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in \mathbf{H}^1(I)^2$ be a locally optimal solution of $(\mathbf{P}_{\ell_1}(\alpha_k))$. Then, we find measures $\bar{\mu}_k, \bar{\nu}_k \in \mathcal{M}(\bar{I})$ which solve the system*

$$\langle b\bar{p}_k, z \rangle_{\mathbf{L}^2(\Sigma)} + \lambda_1 \langle \bar{u}_k, z \rangle_{\mathbf{H}^1(I)} + \alpha_k \langle \bar{v}_k, z \rangle_{\mathbf{L}^2(I)} + \int_I z \, d\bar{\mu}_k = 0 \quad \forall z \in \mathbf{H}^1(I), \quad (4.4a)$$

$$\langle c\bar{p}_k, z \rangle_{\mathbf{L}^2(\Sigma)} + \lambda_2 \langle \bar{v}_k, z \rangle_{\mathbf{H}^1(I)} + \alpha_k \langle \bar{u}_k, z \rangle_{\mathbf{L}^2(I)} + \int_I z \, d\bar{\nu}_k = 0 \quad \forall z \in \mathbf{H}^1(I), \quad (4.4b)$$

$$\bar{\mu}_k, \bar{\nu}_k \leq 0, \quad \int_I \bar{u}_k \, d\bar{\mu}_k = 0, \quad \int_I \bar{v}_k \, d\bar{\nu}_k = 0 \quad (4.4c)$$

where the adjoint state $\bar{p}_k \in \mathbf{W}(0, T)$ is the uniquely determined weak solution of the system (4.3c), (4.3d), (4.3e).

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

$$\begin{aligned} 0 = \langle S^*(S(\bar{u}_k, \bar{v}_k) - y_d), (z_u, z_v) \rangle_{\mathbf{H}^1(I)^2} &+ \lambda_1 \langle \bar{u}_k, z_u \rangle_{\mathbf{H}^1(I)} + \lambda_2 \langle \bar{v}_k, z_v \rangle_{\mathbf{H}^1(I)} \\ &+ \alpha_k \langle \bar{v}_k, z_u \rangle_{\mathbf{L}^2(I)} + \alpha_k \langle \bar{u}_k, z_v \rangle_{\mathbf{L}^2(I)} \\ &+ \langle \bar{\mu}_k, z_u \rangle_{\mathbf{H}^1(I)} + \langle \bar{\nu}_k, z_v \rangle_{\mathbf{H}^1(I)} \end{aligned} \quad (4.5)$$

for all $(z_u, z_v) \in \mathbf{H}^1(I)^2$. By means of Lemma 3.1 of [14], we have $\bar{\mu}_k, \bar{\nu}_k \in \mathbf{H}^1(I)^* \cap \mathcal{M}_-(\bar{I})$. Due to $\mathbf{H}^1(I) \hookrightarrow \mathbf{C}(\bar{I})$, it holds $\mathcal{M}(\bar{I}) \hookrightarrow \mathbf{H}^1(I)^*$, *i.e.* we obtain the characterization $\bar{\mu}_k, \bar{\nu}_k \in \mathcal{M}_-(\bar{I})$. Since we have $\langle \bar{\mu}_k, \bar{u}_k \rangle_{\mathbf{H}^1(I)} = \langle \bar{\nu}_k, \bar{v}_k \rangle_{\mathbf{H}^1(I)} = 0$ from above, condition (4.4c) can be deduced. Next, let $\bar{p}_k \in \mathbf{W}(0, T)$ be the weak solution of (4.3c), (4.3d), (4.3e). Then, Lemma 3.3 yields

$$\langle S^*(S(\bar{u}_k, \bar{v}_k) - y_d), (z_u, z_v) \rangle_{\mathbf{H}^1(I)^2} = \iint_{\Sigma} (b\bar{p}_k z_u + c\bar{p}_k z_v) \, ds dt.$$

Putting this into (4.5) and decoupling the resulting condition w.r.t. z_u and z_v yields the conditions (4.4a) and (4.4b). This completes the proof. \square

The global convergence result of Theorem 4.3 only applies to the situation where $(\mathbf{P}_{\ell_1}(\alpha_k))$ can be solved to global optimality for each $k \in \mathbb{N}$. This, however, cannot be guaranteed in numerical practice since $(\mathbf{P}_{\ell_1}(\alpha_k))$ is a nonconvex program for large enough values of the penalty parameter. As a consequence, we need to study the situation where $(\mathbf{P}_{\ell_1}(\alpha_k))$ 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 4.8. *For each $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in \mathbf{H}^1(I)^2$ be a stationary point of $(\mathbf{P}_{\ell_1}(\alpha_k))$ in the sense of Proposition 4.7. Furthermore, let $(\bar{u}, \bar{v}) \in \mathbf{H}^1(I)^2$ 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 and that \bar{u} and \bar{v} are continuous. Noting that we have $\{(\bar{u}_k, \bar{v}_k)\}_{k \in \mathbb{N}} \subset \mathbf{H}_+^1(I)^2$ while $\mathbf{H}_+^1(I)$ is weakly sequentially closed, $(\bar{u}, \bar{v}) \in \mathbf{H}_+^1(I)^2$ is obtained. It remains to show $\int_I \bar{u}(t)\bar{v}(t) \, dt = 0$. Assume on the contrary that the measurable set

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

possesses positive measure. Due to the compactness of $\mathbf{H}^1(I) \hookrightarrow \mathbf{C}(\bar{I})$, we may assume w.l.o.g. the pointwise convergences $\bar{u}_k(t) \rightarrow \bar{u}(t)$ and $\bar{v}_k(t) \rightarrow \bar{v}(t)$ 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 $\bar{p}_k \in \mathbf{W}(0, T)$ and $\bar{\mu}_k, \bar{\nu}_k \in \mathcal{M}(\bar{I})$ be the Lagrange multipliers which solve the system (4.4). Testing (4.4a) with \bar{u}_k and (4.4b) with \bar{v}_k while exploiting (4.4c) yields

$$\begin{aligned} \langle b\bar{p}_k, \bar{u}_k \rangle_{\mathbf{L}^2(\Sigma)} + \lambda_1 \|\bar{u}_k\|_{\mathbf{H}^1(I)}^2 + \alpha_k \langle \bar{u}_k, \bar{v}_k \rangle_{\mathbf{L}^2(I)} &= 0, \\ \langle c\bar{p}_k, \bar{v}_k \rangle_{\mathbf{L}^2(\Sigma)} + \lambda_2 \|\bar{v}_k\|_{\mathbf{H}^1(I)}^2 + \alpha_k \langle \bar{u}_k, \bar{v}_k \rangle_{\mathbf{L}^2(I)} &= 0. \end{aligned} \tag{4.6}$$

Noting that the compactness of S , see Lemma 3.2, yields $S(\bar{u}_k, \bar{v}_k) \rightarrow S(\bar{u}, \bar{v})$, the continuity of the solution operator associated with (APDE) guarantees $\bar{p}_k \rightarrow \bar{p}$ in $\mathbf{W}(0, T)$ where $\bar{p} \in \mathbf{W}(0, T)$ is the uniquely determined weak solution of

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

Thus, due to the boundedness of $\{(\bar{u}_k, \bar{v}_k)\}_{k \in \mathbb{N}}$ in $\mathbf{H}^1(I)^2$, and the convergences $\bar{u}_k \rightarrow \bar{u}$ and $\bar{v}_k \rightarrow \bar{v}$ in $\mathbf{L}^2(I)$, (4.6) can be used to infer the boundedness of $\{\alpha_k \langle \bar{u}_k, \bar{v}_k \rangle_{\mathbf{L}^2(I)}\}_{k \in \mathbb{N}}$. On the other hand, the above considerations

show

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

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. This completes the proof. \square

From the finite-dimensional setting, see *e.g.* Theorem 2.1 of [28] or Theorem 3.4 of [33], 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 expect that taking the limit $k \rightarrow \infty$ in the system (4.4) recovers a strong stationarity-type condition in the sense of Proposition 3.5 which characterizes the limit point. However, following [28, 33], it might be possible to infer weaker stationarity-type conditions which provide some information on the biactive set. As we pointed out in Section 4.3, 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 $(P_{\ell_1}(\alpha_k))$. The associated result is an immediate consequence of similar considerations for switching-constrained optimal control problems, see [12].

Remark 4.9. For fixed $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in H^1(I)^2$ be a locally optimal solution of $(P_{\ell_1}(\alpha_k))$. Assume that the estimate $\alpha_k \geq \|S\|^2 + \max(\lambda_1, \lambda_2) + \pi^2$ holds for all $k \in \mathbb{N}$ 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 Theorem 3.2 of [12] (which, actually, applies directly since all functions appearing in the proof of this result are non-negative), 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 $J_{\ell_1}^k$ for small values of the penalty parameter α_k . Noting that we have

$$\begin{aligned} J_{\ell_1}^k(u, v) &= \frac{1}{2} \|S(u, v) - y_d\|_{L^2(\Omega)}^2 + \frac{\lambda_1}{2} \|\partial_t u\|_{L^2(I)}^2 + \frac{\lambda_2}{2} \|\partial_t v\|_{L^2(I)}^2 \\ &\quad + \frac{1}{2} \left((\lambda_1 - \alpha_k) \|u\|_{L^2(I)}^2 + (\lambda_2 - \alpha_k) \|v\|_{L^2(I)}^2 + \alpha_k \|u + v\|_{L^2(I)}^2 \right) \end{aligned}$$

for all $(u, v) \in H^1(I)^2$, $(P_{\ell_1}(\alpha_k))$ 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 4.7 are also sufficient. Particularly, for large regularization parameters λ_1 and λ_2 , there is some hope that solving the convex program $(P_{\ell_1}(\min(\lambda_1, \lambda_2)))$ yields a reasonable approximate of an optimal solution for (OCP). A related idea has been exploited in [13] in order to tackle switching-constrained optimal control problems. In order to avoid dealing with Lagrange multipliers from $\mathcal{M}_-(\bar{I})$, $(P_{\ell_1}(\min(\lambda_1, \lambda_2)))$ can be solved by the simple penalty approach from [14], *i.e.* one solves the sequence of programs

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

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

4.5. The decoupled ℓ_2 -penalty approach

Let us now focus on the decoupled penalty approach promoted *via* $(\mathbf{P}_\psi(\alpha_k, \beta_k))$ where $\{\alpha_k\}_{k \in \mathbb{N}}$ and $\{\beta_k\}_{k \in \mathbb{N}}$ are sequences of non-negative real numbers tending to ∞ as $k \rightarrow \infty$. In order to specify the penalty term Ψ , we have to fix a continuous function $\psi: \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 $\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}$ is an NCP-function with non-negative 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.* [9, 39], in order to derive a first-order optimality system for $(\mathbf{P}_\psi(\alpha_k, \beta_k))$ which seems to be an essential drawback in the 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 Theorem 5.9 of [19].

Another reasonable choice for an NSP-function is given by

$$\forall (a, b) \in \mathbb{R}^2: \quad \bar{\psi}(a, b) := \frac{1}{2}a^2b^2, \quad (4.7)$$

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 expect any promising dual convergence results beyond (if at all) weak stationarity, see Remark 4.7 of [10]. 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 $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$.

Lemma 4.10. *Let $\Psi: \mathbf{H}^1(I)^2 \rightarrow \mathbb{R}$ be the mapping associated with the functional $\bar{\psi}: \mathbb{R}^2 \rightarrow \mathbb{R}$ defined in (4.7). Then, Ψ is continuously Fréchet differentiable. For each $(\bar{u}, \bar{v}) \in \mathbf{H}^1(I)^2$, we obtain the subsequent formula for the associated Fréchet derivative $\Psi'(\bar{u}, \bar{v})$:*

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

Proof. For $(\bar{u}, \bar{v}), (\delta_u, \delta_v) \in \mathbf{H}^1(I)^2$, we have

$$\Psi(\bar{u} + \delta_u, \bar{v} + \delta_v) = \Psi(\bar{u}, \bar{v}) + \langle \bar{u}\bar{v}^2, \delta_u \rangle_{L^2(I)} + \langle \bar{u}^2\bar{v}, \delta_v \rangle_{L^2(I)} + o\left(\|(\delta_u, \delta_v)\|_{\mathbf{H}^1(I)^2}\right)$$

by definition of Ψ and continuity of the embedding $\mathbf{H}^1(I) \hookrightarrow L^\infty(I)$, see Theorem 4.12 of [1]. Thus, the assertion follows by definition of Fréchet differentiability. \square

Combining the above lemma with the smoothness of the penalty functional Π , we infer that the problem $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$ is smooth for each $k \in \mathbb{N}$ and $\bar{\psi}$ from (4.7). Thus, we obtain the following first-order optimality condition exploiting Lemma 4.2 of [14], Lemma 3.3, and Lemma 4.10.

Proposition 4.11. *For fixed $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in \mathbf{H}^1(I)^2$ be a locally optimal solution of $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$ where $\bar{\psi}$ is the function defined in (4.7). Then, the weak solution $\bar{p}_k \in \mathbf{W}(0, T)$ of the system (4.3c), (4.3d), (4.3e) satisfies the conditions*

$$\langle b\bar{p}_k, z \rangle_{L^2(\Sigma)} + \lambda_1 \langle \bar{u}_k, z \rangle_{\mathbf{H}^1(I)} + \langle \min(0, \alpha_k \bar{u}_k) + \beta_k \bar{u}_k \bar{v}_k^2, z \rangle_{L^2(I)} = 0 \quad \forall z \in \mathbf{H}^1(I), \quad (4.8a)$$

$$\langle c\bar{p}_k, z \rangle_{L^2(\Sigma)} + \lambda_2 \langle \bar{v}_k, z \rangle_{H^1(I)} + \langle \min(0, \alpha_k \bar{v}_k) + \beta_k \bar{u}_k^2 \bar{v}_k, z \rangle_{L^2(I)} = 0 \quad \forall z \in H^1(I). \quad (4.8b)$$

Unfortunately, the above necessary optimality condition is in general not sufficient since $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$ is not convex. However, we obtain the following result which addresses the situation where $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$ is solved only to stationarity for each $k \in \mathbb{N}$, see Proposition 4.8 as well.

Proposition 4.12. *For each $k \in \mathbb{N}$, let $(\bar{u}_k, \bar{v}_k) \in H^1(I)^2$ be a stationary point of $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$, where $\bar{\psi}$ is the function defined in (4.7), in the sense of Proposition 4.11. 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, $(\bar{u}, \bar{v}) \in \mathcal{C}$ holds.*

Proof. Let us assume w.l.o.g. that the weak convergences $\bar{u}_k \rightharpoonup \bar{u}$ and $\bar{v}_k \rightharpoonup \bar{v}$ hold true and that \bar{u} and \bar{v} are continuous. The compactness of $H^1(I) \hookrightarrow C(\bar{I})$ w.l.o.g. guarantees the pointwise convergences $\bar{u}_k(t) \rightarrow \bar{u}(t)$ and $\bar{v}_k(t) \rightarrow \bar{v}(t)$ for each $t \in \bar{I}$. We test (4.8a) with \bar{u}_k and (4.8b) with \bar{v}_k in order to obtain

$$\begin{aligned} \langle b\bar{p}_k, \bar{u}_k \rangle_{L^2(\Sigma)} + \lambda_1 \|\bar{u}_k\|_{H^1(I)}^2 + \alpha_k \int_I \min^2(0, \bar{u}_k(t)) dt + \beta_k \|\bar{u}_k \bar{v}_k\|_{L^2(I)}^2 &= 0, \\ \langle c\bar{p}_k, \bar{v}_k \rangle_{L^2(\Sigma)} + \lambda_2 \|\bar{v}_k\|_{H^1(I)}^2 + \alpha_k \int_I \min^2(0, \bar{v}_k(t)) dt + \beta_k \|\bar{u}_k \bar{v}_k\|_{L^2(I)}^2 &= 0. \end{aligned} \quad (4.9)$$

Similar to the proof of Proposition 4.8, we obtain the boundedness of the terms $\langle b\bar{p}_k, \bar{u}_k \rangle_{L^2(\Sigma)} + \lambda_1 \|\bar{u}_k\|_{H^1(I)}^2$ as well as $\langle c\bar{p}_k, \bar{v}_k \rangle_{L^2(\Sigma)} + \lambda_2 \|\bar{v}_k\|_{H^1(I)}^2$ in \mathbb{R} . Due to $\alpha_k \rightarrow \infty$ and $\beta_k \rightarrow \infty$ as $k \rightarrow \infty$, (4.9) now guarantees the convergences $\int_I \min^2(0, \bar{u}_k(t)) dt \rightarrow 0$, $\int_I \min^2(0, \bar{v}_k(t)) dt \rightarrow 0$, and $\|\bar{u}_k \bar{v}_k\|_{L^2(I)}^2 \rightarrow 0$. The pointwise convergences $\bar{u}_k \rightarrow \bar{u}$ and $\bar{v}_k \rightarrow \bar{v}$ as well as the boundedness of $\{\bar{u}_k\}_{k \in \mathbb{N}}$ and $\{\bar{v}_k\}_{k \in \mathbb{N}}$ in $C(\bar{I})$ can be used to infer $\int_I \min^2(0, \bar{u}(t)) dt = 0$ and $\int_I \min^2(0, \bar{v}(t)) dt = 0$ with the aid of the dominated convergence theorem, and, thus, $(\bar{u}, \bar{v}) \in H_+^1(I)^2$ follows. Finally, we exploit $\|\bar{u}_k \bar{v}_k\|_{L^2(I)}^2 \rightarrow 0$ in order to obtain that $I^{++}(\bar{u}, \bar{v})$ is of measure zero in similar fashion. This shows $(\bar{u}, \bar{v}) \in \mathcal{C}$ and completes the proof. \square

By construction of the penalty term Ψ via $\bar{\psi}$ from (4.7), it is clear that taking the limit $k \rightarrow \infty$ in the stationarity conditions (4.8) (if possible, see Proposition 4.12) 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.

5. COMPUTATIONAL IMPLEMENTATION OF THE PENALTY SCHEMES

In this section, we discuss the computational implementation of the three penalty schemes which were introduced in Section 4.2. Particularly, we investigate

- coupled FB:** the coupled penalty approach which exploits the Fischer–Burmeister function φ_{FB} , see Section 4.3,
- decoupled ℓ_1 :** the ℓ_1 -penalty approach from Section 4.4, and
- decoupled ℓ_2 :** the decoupled ℓ_2 -penalty approach from Section 4.5 where the underlying penalty term is induced by the function $\bar{\psi}$ from (4.7).

Thus, we have to investigate how the associated surrogate problems $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$, $(\mathbf{P}_{\ell_1}(\alpha_k))$, and $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$ 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 $k \rightarrow \infty$. Noting that $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$ and $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$ 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 $(\mathbf{P}_{\ell_1}(\alpha_k))$ is more challenging since we already observed in Proposition 4.7 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 [14], *i.e.* we solve $(\mathbf{P}_{\ell_1}(\alpha_k))$ with the aid of a penalty method where the penalty term addresses the non-negativity 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 $l \rightarrow \infty$, we consider

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

which is unconstrained. Observing that the penalty term $\mathbf{P}(u, v)$ may take negative values, the existence of solutions to $(\mathbf{P}_{\ell_1}(\alpha_k, \gamma_l))$ has to be discussed. Observing that \mathbf{P} and Π are weakly sequentially continuous, see Lemma 4.1, this would follow by \mathbf{H}^1 -coercivity of the objective functional associated with $(\mathbf{P}_{\ell_1}(\alpha_k, \gamma_l))$. The latter property is discussed in the subsequent lemma.

Lemma 5.1. *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 $(\mathbf{P}_{\ell_1}(\alpha_k, \gamma_l))$ is \mathbf{H}^1 -coercive.*

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

$$\alpha_k \int_{I_{-+} \cup I_{+-}} u(t)v(t) \, dt$$

where we used

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

Thus, we observe

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

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

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

Similarly, we obtain

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

Thus, we have

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

from the lemma's assumption, and the assertion follows. \square

Using standard arguments like in the proof of Theorem 4.3, one can show that a sequence $\{(u_k^l, v_k^l)\}_{l \in \mathbb{N}}$ of global minimizers associated with $(P_{\ell_1}(\alpha_k, \gamma_l))$ for fixed $k \in \mathbb{N}$ possesses a strongly convergent subsequence whose limit point is a global minimizer of $(P_{\ell_1}(\alpha_k))$. In this regard, the consideration of $(P_{\ell_1}(\alpha_k, \gamma_l))$ is meaningful for large enough penalty parameters γ_l .

5.1. 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-once", *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-once" counterpart for the indirect (first-optimize-then-discretize) 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-once" method, see [40, 42, 58]. 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 [40, 58]. However, having an FEM toolkit like OOPDE at hand, see [43], and together with MATLAB's capabilities to handle matrices, it is straightforward 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 $(P_{\varphi_{\text{FB}}}(\alpha_k))$, $(P_{\ell_1}(\alpha_k, \gamma_l))$, and $(P_{\bar{\psi}}(\alpha_k, \beta_k))$. 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 tetrahedra, 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 mass matrix $M_{\Omega_\Delta}^1$ and stiffness matrix K_{Ω_Δ} . Analogously, the discrete inner product in W_{Ω_Δ} can be obtained with the associated mass matrix $M_{\Omega_\Delta}^0$. 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 of $H^1(I)$, 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 of the transformation matrix E_{I_Δ} in order to represent the controls in the space W_{I_Δ} of piecewise constant functions. Similar as above, we introduce the mass matrices $M_{I_\Delta}^1$ and $M_{I_\Delta}^0$ 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 $y_d \in L^2(\Omega)$ will be discretized by functions from W_{Ω_Δ} . For simplicity, the resulting coefficient vector will be denoted by $y_d \in \mathbb{R}^{n_e}$ again. Consequently, the discretized objective functions of the surrogate problems $(P_{\varphi_{\text{FB}}}(\alpha_k))$, $(P_{\ell_1}(\alpha_k, \gamma_l))$, and $(P_{\bar{\psi}}(\alpha_k, \beta_k))$ are given by

$$\begin{aligned}
\tilde{J}_{\varphi_{\text{FB}}}^k(y, u, v) &:= \hat{J}(y, u, v) \\
&\quad + \frac{\alpha_k}{2} \left(((E_{I_\Delta} u)^2 + (E_{I_\Delta} v)^2)^{\frac{1}{2}} - E_{I_\Delta} u - E_{I_\Delta} v \right)^\top M_{I_\Delta}^0 \left(((E_{I_\Delta} u)^2 + (E_{I_\Delta} v)^2)^{\frac{1}{2}} - E_{I_\Delta} u - E_{I_\Delta} v \right), \\
\tilde{J}_{\ell_1}^{k,l}(y, u, v) &:= \hat{J}(y, u, v) \\
&\quad + \alpha_k (E_{I_\Delta} u)^\top M_{I_\Delta}^0 (E_{I_\Delta} v) \\
&\quad + \frac{\gamma_l}{2} \max(0, -E_{I_\Delta} u)^\top M_{I_\Delta}^0 \max(0, -E_{I_\Delta} u) + \frac{\gamma_l}{2} \max(0, -E_{I_\Delta} v)^\top M_{I_\Delta}^0 \max(0, -E_{I_\Delta} v), \\
\tilde{J}_{\bar{\psi}}^k(y, u, v) &:= \hat{J}(y, u, v) \\
&\quad + \frac{\alpha_k}{2} \max(0, -E_{I_\Delta} u)^\top M_{I_\Delta}^0 \max(0, -E_{I_\Delta} u) + \frac{\alpha_k}{2} \max(0, -E_{I_\Delta} v)^\top M_{I_\Delta}^0 \max(0, -E_{I_\Delta} v) \\
&\quad + \frac{\beta_k}{2} (u \bullet v)^\top E_{I_\Delta}^\top M_{I_\Delta}^0 E_{I_\Delta} (u \bullet v)
\end{aligned} \tag{5.1}$$

where we used

$$\hat{J}(y, u, v) := \frac{1}{2} (E_{\Omega_\Delta} y^n - y_d)^\top M_{\Omega_\Delta}^0 (E_{\Omega_\Delta} y^n - y_d) + \frac{\lambda_1}{2} u^\top (M_{I_\Delta}^1 + K_{I_\Delta}) u + \frac{\lambda_2}{2} v^\top (M_{I_\Delta}^1 + K_{I_\Delta}) v. \tag{5.2}$$

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. We would like to point out that these functions are continuously differentiable but possess nonsmooth first-order derivatives w.r.t. u and v . In order to discretize the weak formulation (3.2) 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 $C(t)$, $a(t)$, and $q(t)$ 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 $C(t)$ is denoted by $K_{\Omega_\Delta}(C(t))$, the mass matrix associated with $a(t)$ is represented by $M_{\Omega_\Delta}^1(a(t))$, $Q_{\Omega_\Delta}(q(t))$ is the matrix associated with the boundary integral involving $q(t)$, 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

$$\begin{aligned}
\partial_t M_{\Omega_\Delta}^1 y(t) + (K_{\Omega_\Delta}(C(t)) + Q_{\Omega_\Delta}(q(t)) + M_{\Omega_\Delta}^1(a(t))) y(t) &= G_{\Omega_\Delta}(b) u(t) + G_{\Omega_\Delta}(c) v(t), \\
M_{\Omega_\Delta}^1 y^0 &= 0.
\end{aligned} \tag{5.3}$$

The last step is now to discretize the ODE (5.3) 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

$$\Theta_{\Omega_\Delta}^i := M_{\Omega_\Delta}^1 + \delta t_i (K_{\Omega_\Delta}(C^{i+1}) + Q_{\Omega_\Delta}(q^{i+1}) + M_{\Omega_\Delta}^1(a^{i+1}))$$

for all $i = 0, \dots, n-1$ where the matrices $K_{\Omega_\Delta}(C^{i+1})$, $Q_{\Omega_\Delta}(q^{i+1})$, and $M_{\Omega_\Delta}^1(a^{i+1})$ are assembled with $C(t_{i+1})$, $q(t_{i+1})$, and $a(t_{i+1})$, respectively. Discretizing the derivative w.r.t. time in (5.3) with forward differences, we obtain the fully discretized system

$$\begin{aligned}
\Theta_{\Omega_\Delta}^i y^{i+1} - M_{\Omega_\Delta}^1 y^i &= \delta t_i (G_{\Omega_\Delta}(b) u^{i+1} + G_{\Omega_\Delta}(c) v^{i+1}) \quad i = 0, \dots, n-1, \\
M_{\Omega_\Delta}^1 y^0 &= 0.
\end{aligned}$$

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

$$[\mathbf{A} \mid -\mathbf{F}(b) \mid -\mathbf{F}(c)]z = 0, \quad (5.4)$$

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

$$\mathbf{A} := \begin{pmatrix} M_{\Omega_\Delta}^1 & 0 & \cdots & \cdots & \cdots & 0 \\ -M_{\Omega_\Delta}^1 & \Theta_{\Omega_\Delta}^0 & \ddots & & & \vdots \\ 0 & -M_{\Omega_\Delta}^1 & \Theta_{\Omega_\Delta}^1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -M_{\Omega_\Delta}^1 & \Theta_{\Omega_\Delta}^{n-2} & 0 \\ 0 & \cdots & \cdots & 0 & -M_{\Omega_\Delta}^1 & \Theta_{\Omega_\Delta}^{n-1} \end{pmatrix}$$

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

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

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

Equipping each of the potential objective functionals from (5.1) with the linear constraints (5.4), we obtain three finite-dimensional minimization problems which are the discrete counterparts of $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$, $(\mathbf{P}_{\ell_1}(\alpha_k, \gamma_l))$, and $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$, respectively. Let us note that the three problems possess the same constraints (5.4) which will be addressed *via* the discrete Lagrange multiplier $p \in \mathbb{R}^{(n+1)n_p}$. The corresponding Lagrange function for each objective functional $\tilde{J} \in \{\tilde{J}_{\varphi_{\text{FB}}}^k, \tilde{J}_{\ell_1}^{k,l}, \tilde{J}_{\bar{\psi}}^k\}$ reads as

$$L_{\tilde{J}}(y, u, v, p) := \tilde{J}(y, u, v) - p^\top (\mathbf{A}y - \mathbf{F}(b)u - \mathbf{F}(c)v).$$

Based on their individual KKT-conditions (which are indeed necessary optimality conditions since the constraints are linear) which read as a system of nonsmooth equations, respectively, a semismooth Newton method can be used to solve these problems. We would like to point out that the resulting adjoint equation on the multiplier p (*i.e.* the equation $\nabla_y L_{\tilde{J}}(y, u, v, p) = 0$) is the same for each choice $\tilde{J} \in \{\tilde{J}_{\varphi_{\text{FB}}}^k, \tilde{J}_{\ell_1}^{k,l}, \tilde{J}_{\bar{\psi}}^k\}$ and given by

$$\mathbf{A}^\top p = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ E_{\Omega_\Delta}^\top M_{\Omega_\Delta}^0 (E_{\Omega_\Delta} y^n - y_d) \end{pmatrix}. \quad (5.5)$$

Observe that we use the “first-discretize-then-optimize” approach here which is why the discrete Lagrange multiplier p is not generally equal to the solution of the discretized adjoint equation (APDE) which would come

Algorithm 1 Conceptual algorithm for solving (OCP) *via* $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$ or $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \alpha_k))$

- S0** Let $\{\alpha_k\}_{k \in \mathbb{N}}$ be a sequence of positive penalty parameters tending to ∞ as $k \rightarrow \infty$. Let a tolerance $\text{tol} > 0$ be given. Let (u_0, v_0) be a starting point of controls. Compute y_0 as a solution of the discretized state equation (5.4) with fixed $u := u_0$ and $v := v_0$. Compute p_0 as a solution of the discrete adjoint equation (5.5) with fixed $y := y_0$. Set $k := 0$.
- S1** Solve the discretized optimality system associated with $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$ and $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \alpha_k))$, 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 $\|(u_{k+1}, v_{k+1}) - (u_k, v_k)\|_M < \text{tol}$ and $\|\nabla L_{\tilde{j}}(y_{k+1}, u_{k+1}, v_{k+1}, p_{k+1})\| < \text{tol}$ for $\tilde{J} \in \{\tilde{J}_{\varphi_{\text{FB}}}^k, \tilde{J}_{\bar{\psi}}^k\}$, respectively, then return $(y_{k+1}, u_{k+1}, v_{k+1}, p_{k+1})$. Otherwise, set $k := k + 1$ and go to **S1**.
-

into play whenever the “first-optimize-then-discretize” approach would be used, see *e.g.* [44]. Let us note that the discrete first-order optimality system for a time-stationary counterpart of problem $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$ can be found in Section 5.1 of [10].

5.2. 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 5.1. As already mentioned, these systems will be solved with the aid of a Newton-type method. For globalization, 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 5.1. For brevity, however, we fix $\beta_k := \alpha_k$ for each $k \in \mathbb{N}$ for the consideration of $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \beta_k))$. The pseudo-code stated within Algorithm 1 presents a solution concept for (OCP) *via* the surrogates $(\mathbf{P}_{\varphi_{\text{FB}}}(\alpha_k))$ as well as $(\mathbf{P}_{\bar{\psi}}(\alpha_k, \alpha_k))$ and, thus, provides a possible realization of **coupled FB** and **decoupled** ℓ_2 . Let us briefly note that $\|\cdot\|_M$ denotes a weighted Euclidean norm where suitable choices for the matrix M are given by

$$\begin{bmatrix} M_{I_\Delta}^1 & \mathbb{0} \\ \mathbb{0} & M_{I_\Delta}^1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} M_{I_\Delta}^1 + K_{I_\Delta} & \mathbb{0} \\ \mathbb{0} & M_{I_\Delta}^1 + K_{I_\Delta} \end{bmatrix}$$

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* $(\mathbf{P}_{\ell_1}(\alpha_k))$, we exploit the nested penalty method Algorithm 2 which makes use of the surrogate $(\mathbf{P}_{\ell_1}(\alpha_k, \gamma_l))$. As already mentioned, this problem may fail to possess a solution if α_k is large while γ_l is small. Thus, we have to navigate the sequences $\{\alpha_k\}_{k \in \mathbb{N}}$ and $\{\gamma_l\}_{l \in \mathbb{N}}$ appropriately. Here, we exploit Lemma 5.1 and the subsequent comments for that purpose. Particularly, Algorithm 2 realizes **decoupled** ℓ_1 . 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 $\mathbf{P}_{\ell_1}(\alpha_0, \gamma_l)$ is a convex optimization problem for each $l \in \mathbb{N}$, see Section 4.4. This way, the first run of the inner loop of Algorithm 2 computes an approximate global minimizer of $\mathbf{P}_{\ell_1}(\alpha_0)$ which turns out to be a solid base for all remaining iterations, see Section 6.

In the subsequent remark, we comment on a possible strategy for the choice of a reasonable starting point in Algorithms 1 and 2.

Remark 5.2. Due to the strong local convergence properties of Newton-type methods, it might be important to choose a reasonable starting point (u_0, v_0) in Algorithms 1 and 2. In this regard, a discrete solution of the

Algorithm 2 Conceptual 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 $k \rightarrow \infty$. 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 starting point of controls. Compute y_0 as a solution of the discretized state equation (5.4) with fixed $u := u_0$ and $v := v_0$. Compute p_0 as a solution of the discrete adjoint equation (5.5) with fixed $y := y_0$. 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 $\|(u_k^{l+1}, v_k^{l+1}) - (u_k^l, v_k^l)\|_M < \text{tol}_1$ and $\|\nabla L_{\bar{y}_k^{l+1}}(y_k^{l+1}, u_k^{l+1}, v_k^{l+1}, p_k^{l+1})\| < \text{tol}_1$, 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 $(y_k^{l+1}, u_k^{l+1}, v_k^{l+1}, p_k^{l+1})$. Otherwise, carry out the updates $(y_{k+1}^0, u_{k+1}^0, v_{k+1}^0, p_{k+1}^0) := (y_k^{l+1}, u_k^{l+1}, v_k^{l+1}, p_k^{l+1})$, $l := 0$, $\gamma_0 := \frac{1}{2}\sigma\alpha_{k+1}^2 \max(1/\lambda_1, 1/\lambda_2)$, as well as $k := k + 1$ and go to **S1**.
-

optimal control problem

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

where the equilibrium condition in (OCP) is neglected, can be 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 [14]. In this regard, a direct (*cf.* Sect. 5.1) or indirect (see [14]) discretization strategy is possible. Observe that whenever a minimizer of (P) is already complementary, then it is a global minimizer of (OCP) as well.

5.3. Measuring feasibility

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

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

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 4.1.

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

$$\begin{aligned} \forall u, v \in \mathbb{R}^{n+1}: \quad \rho_{I_\Delta}(u, v) &:= |E_{I_\Delta} u|^\top M_{I_\Delta}^0 |E_{I_\Delta} v| \\ &+ \frac{1}{2} \max(0, -E_{I_\Delta} u)^\top M_{I_\Delta}^0 \max(0, -E_{I_\Delta} u) \\ &+ \frac{1}{2} \max(0, -E_{I_\Delta} v)^\top M_{I_\Delta}^0 \max(0, -E_{I_\Delta} v), \end{aligned} \tag{5.6}$$

see Section 5.1. Above, $|\eta|$ now denotes the componentwise absolute value of a vector η .

5.4. Quantifying robustness of the penalty methods

In order to check robustness of the three penalty methods **coupled FB**, **decoupled ℓ_1** , as well as **decoupled ℓ_2** , we cannot rely on only one computation process - *e.g.* by fixing the starting point for Algorithms 1 and 2 according to Remark 5.2 and comparing the resulting outputs. Instead, the three penalty methods are run with several different randomly chosen starting points. Based on the results of these computations, we use

performance profiles, see [15], to provide a convincing quantitative comparison of these methods. Here, we focus on two underlying criteria for evaluation of the performance, namely we investigate the objective function's value as well as the value of the feasibility measure from Section 5.3 at the computed points. Note that we do not rely on computation time here since the comparatively costly evaluations of the generalized second-order derivative of the Fischer–Burmeister function slow the method **coupled FB** down for spacial domains of dimension 2 or higher, see Section 6 as well.

Let \mathcal{S} be an abstract set of indices associated with the randomly chosen starting points. Furthermore, we set

$$\mathcal{A} := \{\text{coupled FB}, \text{decoupled } \ell_1, \text{decoupled } \ell_2\}.$$

For a starting point $(u_s, v_s) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$, $s \in \mathcal{S}$, let $(y_s^a, u_s^a, v_s^a, p_s^a)$ denote the output of algorithm $a \in \mathcal{A}$ realized *via* Algorithms 1 and 2, respectively. For the comparison of function values, we make use of the quantity

$$\forall s \in \mathcal{S} \forall a \in \mathcal{A}: \quad \mathcal{Q}_\theta^{\text{fv}}(s, a) := \begin{cases} \hat{J}(y_s^a, u_s^a, v_s^a) - \hat{J}_{\min} + \theta & \text{if } a \text{ succeeds using input } (u_s, v_s) \\ \infty & \text{otherwise} \end{cases}$$

as performance metric where \hat{J} is defined in (5.2) and \hat{J}_{\min} is a reasonable approximate of the objective function's value associated with a global minimizer of the finite-dimensional MPCC

$$\begin{aligned} & \underset{y, u, v}{\text{minimize}} && \hat{J}(y, u, v) \\ & \text{subject to} && (y, u, v) \in \mathbb{R}^{(n+1)n_p} \times \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \\ & && (y, u, v) \text{ solves (5.4)} \\ & && 0 \leq u^i \perp v^i \geq 0 \quad i = 0, \dots, n, \end{aligned} \tag{5.7}$$

see Remark 5.3 as well. For the quantitative comparison of feasibility, we exploit

$$\forall s \in \mathcal{S} \forall a \in \mathcal{A}: \quad \mathcal{Q}_\theta^{\text{feas}}(s, a) := \begin{cases} \rho_{I_\Delta}(u_s^a, v_s^a) + \theta & \text{if } a \text{ succeeds using input } (u_s, v_s) \\ \infty & \text{otherwise} \end{cases}$$

where ρ_{I_Δ} represents the feasibility measure from (5.6). Above, $\theta \geq 0$ is an additional parameter which is used to reduce sensitivity w.r.t. numerical accuracy. Using the performance metric \mathcal{Q}_θ^m , $m \in \{\text{fv}, \text{feas}\}$, we investigate the so-called performance ratio

$$\forall s \in \mathcal{S} \forall a \in \mathcal{A}: \quad r_{s,a} := \frac{\mathcal{Q}_\theta^m(s, a)}{\min\{\mathcal{Q}_\theta^m(s, \alpha) \mid \alpha \in \mathcal{A}\}}.$$

In the associated performance profile for the criterion $m \in \{\text{fv}, \text{feas}\}$, we plot the illustrative parts of the functions $v_a^m: [1, \infty) \rightarrow [0, 1]$, defined *via*

$$\forall \kappa \in [1, \infty): \quad v_a^m(\kappa) := \frac{\text{card}(\{s \in \mathcal{S} \mid r_{s,a} \leq \kappa\})}{\text{card}(\mathcal{S})}$$

for each algorithm $a \in \mathcal{A}$. Above, $\text{card}(X)$ denotes the cardinality of a set X .

Remark 5.3. In order to approximate the globally minimal objective function value \hat{J}_{\min} of the nonconvex program (5.7), we use the following heuristic procedure: Firstly, we discretize the time interval I using a relatively

rough time grid of $n' < n$ subintervals. For each index set $\mathcal{I} \subset \{0, \dots, n'\}$, we solve the *convex* problem

$$\begin{aligned} & \underset{y, u, v}{\text{minimize}} && \hat{J}(y, u, v) \\ & \text{subject to} && (y, u, v) \in \mathbb{R}^{(n'+1)n_p} \times \mathbb{R}^{n'+1} \times \mathbb{R}^{n'+1} \\ & && (y, u, v) \text{ solves (5.4)} \\ & && u^i \geq 0 \quad v^i = 0 \quad i \in \mathcal{I} \\ & && u^i = 0 \quad v^i \geq 0 \quad i \in \{0, \dots, n'\} \setminus \mathcal{I} \end{aligned}$$

to global optimality. Using linear interpolation, the control components of the associated global minimizers are lifted to the finer time grid of n subintervals. The resulting $2^{n'+1}$ pairs are then used as starting points of Algorithms 1 and 2 where the underlying parameters are specified in the context of the respective example, see Section 6. Finally, we determine \hat{J}_{\min} as the minimum of the associated objective function values of all $3 \cdot 2^{n'+1}$ obtained solutions.

6. NUMERICAL EXPERIMENTS

In this section, the three suggested penalty methods **coupled FB**, **decoupled ℓ_1** , and **decoupled ℓ_2** , whose numerical implementation has been discussed in Section 5, 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 [43]. All computations are performed on a standard computer possessing a 3.40GHz Intel(R) Core(TM) i5-3570K processor. 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 $\text{tol} = 10^{-8}$. 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 $\text{tol}_1 = \text{tol}_2 := 10^{-8}$ are used. For technical purposes, we stop Algorithm 1 or Algorithm 2 whenever the penalty parameter α_k exceeds $2 \cdot 10^5$ or 5, respectively, and use the latest iterate as the final output. In both algorithms, the matrix M which models the norm in the respective stopping criteria is chosen as the identity.

Example 6.1. The first example is given on a one-dimensional interval $\Omega := (0, 1)$ and the time interval $I := (0, 4)$. The heat source u is active at the boundary point $s = 0$ of Ω while v is active at $s = 1$, *i.e.* we use

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

The coefficient functions $C \equiv 0.0125$, $a \equiv 0$, and $q \equiv 1$ are fixed. Furthermore, we choose $\lambda_1 = \lambda_2 := 10^{-5}$ for the regularization parameters and $y_d(x) := \sin(5\pi x^2) + x$, $x \in \Omega$, for the desired state.

We subdivide the spacial domain Ω into 40 equidistant subintervals while the time interval I is discretized using 160 equidistant subintervals. To obtain a reasonable candidate for a global minimizer, we exploit the strategy proposed in Remark 5.3. More precisely, we use $n' := 10$ and choose equidistant subintervals for the rough discretization of I . The obtained solution with the best function value $\hat{J}_{\min} = 0.1400$ is visualized in Figure 1A. Exploiting Algorithms 1 and 2, we ran all three penalty methods for 100 random starting points from $\mathbb{R}^{161} \times \mathbb{R}^{161}$ ranging componentwise in $[0, 9]$. For the calculation of the associated performance profiles, which can be found in Figure 1B and 1C, we use $\theta := 1$. It turns out that **decoupled ℓ_1** outruns the other two algorithms regarding computed function values. Indeed, this algorithm computes a solution close to the heuristically found global minimizer in *all* runs. A potential reason seems to be the adjustment of penalty parameters in Algorithm 2. Recall that we solve a *convex* problem with a coercive objective function at the beginning of the computation process in the inner loop of Algorithm 2, see Section 4.4 and Lemma 5.1 as well, and obtain an associated global minimizer. In this example, the computed complementary controls can be found in a comparatively small neighborhood of this point. This observation is *independent* of the initial guess

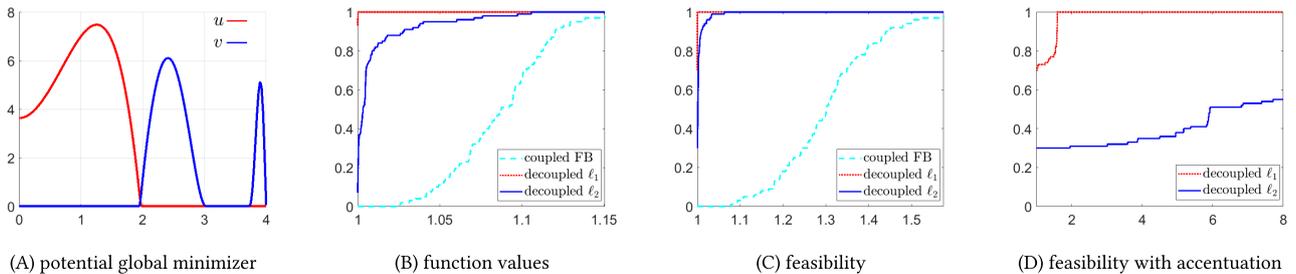


FIGURE 1. *Form left to right:* Heuristically determined global minimizer, performance profile for computed function values, performance profile for the feasibility measure, and performance profile for the feasibility measure with accentuation for Example 6.1.

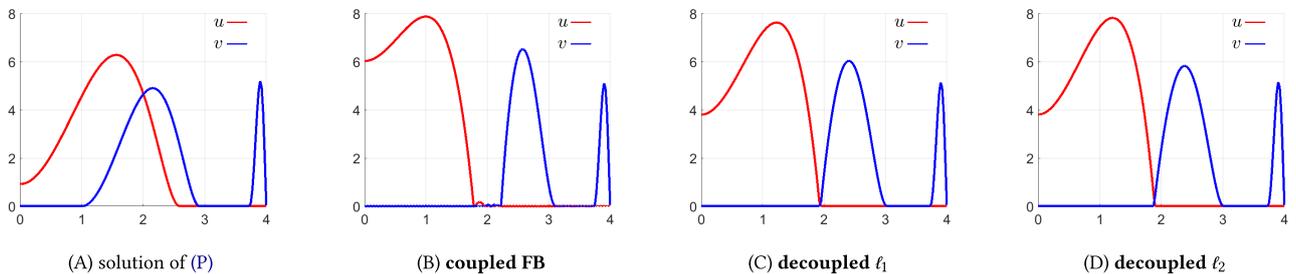


FIGURE 2. *From left to right:* Solution of the convex problem (P) as well as the corresponding outputs of **coupled FB**, **decoupled ℓ_1** , and **decoupled ℓ_2** for Example 6.1.

which explains the behavior of **decoupled ℓ_1** . Regarding feasibility, both decoupled methods **decoupled ℓ_1** and **decoupled ℓ_2** perform similarly good. Naturally, none of these methods exactly reaches perfect feasibility. We use $\theta := 10^{-4}$ in Figure 1D to visualize the fulfilment of the complementarity requirement regarding **decoupled ℓ_1** and **decoupled ℓ_2** in detail. Algorithm **coupled FB** is clearly outperformed by the other two methods regarding computed function values and feasibility. A possible explanation for this behavior seems to be hidden in the heavy nonlinearity of the squared Fischer–Burmeister function and its generalized second-order derivative. Let us mention that each individual run of the considered solution methods, applied to Example 6.1 for a fixed starting point, consumes at most 2 minutes of time.

In Figure 2, we visualize the output of the three algorithms resulting from the initial guess promoted in Remark 5.2, *i.e.* we use the solution of the discretized convex problem (P), presented in Figure 2A, as starting point. We observe that **decoupled ℓ_1** again recovers the heuristically determined global minimizer of the problem while the methods **coupled FB** and **decoupled ℓ_2** produce points which are somehow related but, nevertheless, different.

Example 6.2. Let us fix $I := (0, 1)$. We define two desired controls u_d and v_d by

$$\forall t \in I: \quad u_d(t) := \max(10 - 20t, 0), \quad v_d(t) := 20(1 - t) \sin(3.2\pi t).$$

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

$$B := \{s \in \Gamma \mid s_1 \geq 1/2\}, \quad C := \{s \in \Gamma \mid s_2 \leq 1/2\}.$$

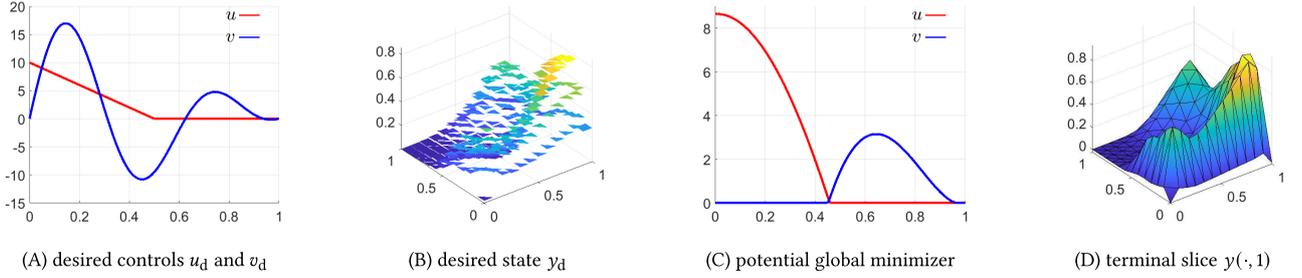


FIGURE 3. *From left to right:* Desired controls u_d and v_d , associated desired state y_d (discretized in W_{Ω_Δ}), heuristically determined global minimizer of controls (u, v) , and associated terminal state $y(\cdot, 1)$ for Example 6.2.

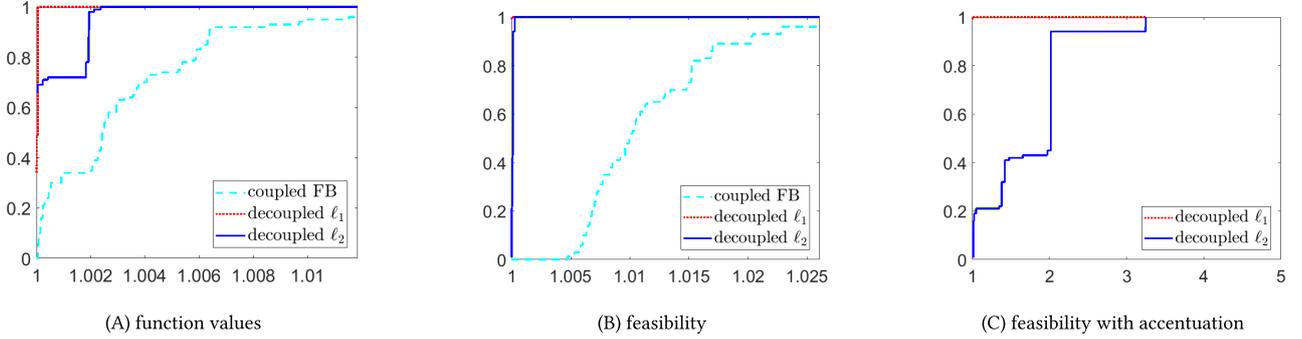


FIGURE 4. *From left to right:* Performance profile for computed function values, performance profile for the feasibility measure, and performance profile for the feasibility measure with accentuation for Example 6.2.

Here, $\chi_A: \Gamma \rightarrow \mathbb{R}$ 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 $y_d(x) := z_d(x, 1)$ for each $x \in \Omega$, see Figure 3B. Let us note that the controls u_d and v_d are not complementary, see Figure 3A, 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, we exploit $\alpha_0 := 1$ and $\alpha_k := 1.05 \cdot \alpha_{k-1}$ for all $k \in \mathbb{N}$. To compute the potential global minimizer, we subdivide the time interval roughly into $n' := 8$ equidistant subintervals. Note that the desired state y_d is still calculated exploiting the desired controls u_d and v_d on the time interval with finer discretization. Following the procedure sketched in Remark 5.3, one gets the best function value $\hat{J}_{\min} = 0.0032$. The potential global minimizer and its associated terminal state, visualized in Figure 3C and 3D, respectively, are outputs of Algorithm 1 when running the **decoupled** ℓ_2 method. 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} .

To check the robustness of all penalty methods, we fix 100 starting points from $\mathbb{R}^{101} \times \mathbb{R}^{101}$ whose components are randomly chosen from $[0, 9]$. The parameter $\theta := 1$ is used to generate the performance profiles of Figure 4A and 4B. At the first glance, **decoupled** ℓ_1 and **decoupled** ℓ_2 produce similarly good results, in particular w.r.t. the feasibility measure from Section 5.3. Choosing $\theta := 10^{-4}$, the minimal differences are presented in Figure 4C.

Finally, let us check the strategy from Remark 5.2 for the choice of the starting point. In Figure 5A, the solution of the associated problem (P) is presented which serves as a starting point for Algorithms 1 and 2. The associated outputs can be found in Figure 5B-5D. Obviously, the solution created by **coupled** FB with

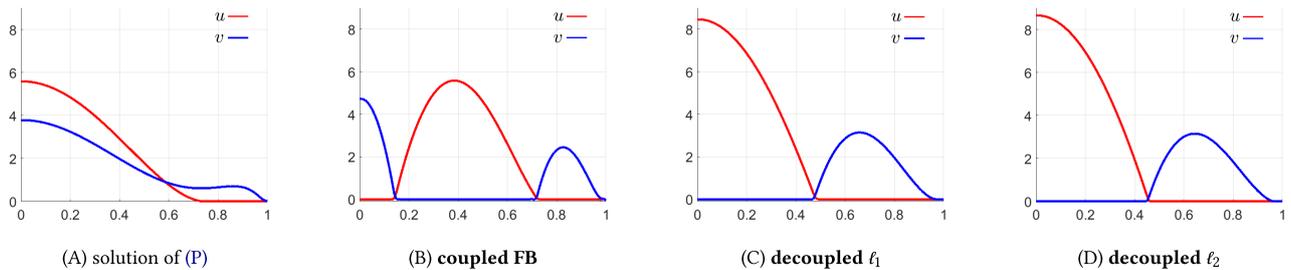


FIGURE 5. *From left to right:* Solution of the convex problem (P) and the corresponding outputs of **coupled FB**, **decoupled ℓ_1** , and **decoupled ℓ_2** for Example 6.2.

associated function value 0.0072 is not related to the potential global minimizers from Figure 3C with function value 0.0032. The other two methods approximately recover the heuristically determined global minimizer.

Summing up our experiments on Example 6.2, the method **coupled FB** is not competitive at all in comparison with **decoupled ℓ_1** and **decoupled ℓ_2** . This is not only reflected by means of computed function values and the feasibility of obtained solutions but also in the light of computation time. Exemplary, let us mention that in order to solve Example 6.2 for a fixed starting point, **coupled FB** takes more than 3 hours while each of the other two methods terminates after at most half an hour. Again, this seems to be caused by the quite expensive evaluation of the squared Fischer–Burmeister function’s generalized second-order derivative.

Summary. In both examples, the method **decoupled ℓ_1** as well as the method **decoupled ℓ_2** , which is based on the NSP-function $\bar{\psi}$ from (4.7), turn out to outrun the coupled Fischer–Burmeister penalty method **coupled FB** from the perspective of computed function values and feasibility of obtained solutions. As soon as the underlying spacial domain’s dimension is larger than 1, a similar behavior can be observed regarding computation time. These issues seem to be caused by the challenging structure of the squared Fischer–Burmeister function’s generalized second-order derivative. A potential approach to tackle this weakness might be the use of quasi-Newton-type ideas to accelerate **coupled FB**, but this is beyond the scope of this paper. We would like to mention (without presenting any detailed numbers) that the method **decoupled ℓ_2** consumes the least computation time due to its simple structure. However, with the aid of **decoupled ℓ_1** , we are in position to compute solutions which are close to the (heuristically determined) global minimizer in nearly all runs, and this observation is independent of the underlying starting point. This behavior might be caused by the neat choice of the penalty parameters in the inner loop of Algorithm 2 which always starts with the solution of a convex optimal control problem.

7. 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 non-negative, 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 non-negativity requirements are penalized individually is referred to as 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, all of them were implemented in MATLAB. In numerical practice, it turned out that the ℓ_1 -approach is the most robust one which yields good solutions w.r.t. computed objective values and feasibility of the obtained solutions in reasonable computation time.

Acknowledgements. We would like to thank two anonymous referees for their careful reading and valuable comments which helped us to enhance the presentation of the results in this paper. This work is partially 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. Mathematics in Industry: Challenges and Frontiers A Process View: Practice and Theory. Edited by D.R. Ferguson and T.J. Peters. 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. *J. Comput. Appl. Math.* **157** (2003) 365–382.
- [5] J.C. Butcher, Numerical Methods for Ordinary Differential Equations. Wiley & Sons, Chichester (2016).
- [6] E. Casas, Second order analysis for bang-bang control problems of PDEs. *SIAM J. Control Optim.* **50** (2012) 2355–2372.
- [7] E. Casas and M. Mateos, Critical cones for sufficient second order conditions in PDE constrained optimization. *SIAM J. Optim.* **30** (2020) 585–603.
- [8] C. Christof and G. Wachsmuth, On second-order optimality conditions for optimal control problems governed by the obstacle problem. *Optimization* **2020** (2020) 1–41.
- [9] F.H. Clarke, Optimization and Nonsmooth Analysis. Wiley (1983).
- [10] C. Clason, Y. Deng, P. Mehlitz and U. Prüfert, Optimal control problems with control complementarity constraints: existence results, optimality conditions, and a penalty method. *Optim. Methods Softw.* **35** (2020) 142–170.
- [11] C. Clason, K. Ito and K. Kunisch, A convex analysis approach to optimal controls with switching structure for partial differential equations. *ESAIM: COCV* **22** (2016) 581–609.
- [12] C. Clason, A. Rund and K. Kunisch, Nonconvex penalization of switching control of partial differential equations. *Syst. Control Lett.* **106** (2017) 1–8.
- [13] C. Clason, A. Rund, K. Kunisch and R.C. Barnard, A convex penalty for switching control of partial differential equations. *Syst. Control Lett.* **89** (2016) 66–73.
- [14] Y. Deng, P. Mehlitz and U. Prüfert, Optimal control in first-order Sobolev spaces with inequality constraints. *Comput. Optim. Appl.* **72** (2019) 797–826.
- [15] E.D. Dolan and J.J. Moré, Benchmarking optimization software with performance profiles. *Math. Program.* **91** (2002) 201–213.
- [16] J.C. Dunn, On Second order sufficient conditions for structured nonlinear programs in infinite-dimensional function spaces. In *Math. Program. with Data Perturbations*. Edited by A.V. Fiacco. Marcel Dekker Inc., New York (1998) 83–108.
- [17] A. Fischer, A special Newton-type optimization method. *Optimization* **24** (1992) 269–284.
- [18] A. Galántai, Properties and construction of NCP functions. *Comput. Optim. Appl.* **52** (2012) 805–824.
- [19] C. Geiger and C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin (2002).
- [20] L. Guo and J.J. Ye, Necessary optimality conditions for optimal control problems with equilibrium constraints. *SIAM J. Control Optim.* **54** (2016) 2710–2733.
- [21] F. Harder and G. Wachsmuth, Comparison of optimality systems for the optimal control of the obstacle problem. *GAMM-Mitteilungen* **40** (2018) 312–338.
- [22] F. Harder and G. Wachsmuth, The limiting normal cone of a complementarity set in Sobolev spaces. *Optimization* **67** (2018) 1579–1603.
- [23] M. Hintermüller, B.S. Mordukhovich and T.M. Surowiec, Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints. *Math. Program.* **146** (2014) 555–582.
- [24] M. Hintermüller and T.M. Surowiec, First-order optimality conditions for elliptic mathematical programs with equilibrium constraints via variational analysis. *SIAM J. Optim.* **21** (2011) 1561–1593.
- [25] M. Hintermüller and T.M. Surowiec, On the directional differentiability of the solution mapping for a class of variational inequalities of the second kind. *Set-Valued Variat. Anal.* **26** (2017) 631–642.
- [26] M. Hintermüller and M. Ulbrich, A mesh-independence result for semismooth Newton methods. *Math. Program.* **101** (2004) 151–184.
- [27] T. Hoheisel, C. Kanzow and A. Schwartz, Theoretical and numerical comparison of relaxation methods for mathematical programs with complementarity constraints. *Math. Program.* **137** (2013) 257–288.
- [28] X.M. Hu and D. Ralph, Convergence of a penalty method for mathematical programs with complementarity constraints. *J. Optim. Theory Appl.* **123** (2004) 365–390.
- [29] X.X. Huang, X.Q. Yang and D.L. Zhu, A sequential smooth penalization approach to mathematical programs with complementarity constraints. *Numer. Funct. Anal. Optim.* **27** (2006) 71–98.
- [30] C. Kanzow and A. Schwartz, Mathematical programs with equilibrium constraints: enhanced Fritz–John-conditions, new constraint qualifications, and improved exact penalty results. *SIAM J. Optim.* **20** (2010) 2730–2753.
- [31] C. Kanzow, N. Yamashita and M. Fukushima, New NCP-functions and their properties. *J. Optim. Theory Appl.* **94** (1997) 115–135.

- [32] K. Kunisch and D. Wachsmuth, Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities. *ESAIM: COCV* **18** (2012) 520–547.
- [33] S. Leyffer, G. López-Calva and J. Nocedal, Interior methods for mathematical programs with complementarity constraints. *SIAM J. Optim.* **17** (2006) 52–77.
- [34] G. Liu, J. Ye and J. Zhu, Partial exact penalty for mathematical programs with equilibrium constraints. *Set-Valued Anal.* **16** (2008) 785.
- [35] Z.-Q. Luo, J.-S. Pang and D. Ralph, *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press (1996).
- [36] Z.-Q. Luo, J.-S. Pang, D. Ralph and S.-Q. Wu, Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints. *Math. Program.* **75** (1996) 19–76.
- [37] H. Maurer and J. Zowe, First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems. *Math. Program.* **16** (1979) 98–110.
- [38] P. Mehlitz and G. Wachsmuth, The limiting normal cone to pointwise defined sets in Lebesgue spaces. *Set-Valued Variat. Anal.* **26** (2018) 449–467.
- [39] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation*. Springer-Verlag (2006).
- [40] I. Neitzel, U. Prüfert and T. Slawig, Strategies for time-dependent PDE control with inequality constraints using an integrated modeling and simulation environment. *Numerical Algor.* **50** (2009) 241–269.
- [41] J.-S. Pang and D.E. Stewart, Differential variational inequalities. *Math. Program. A* **113** (2008) 345–424.
- [42] L. Petzold, J.B. Rosen, P.E. Gill, L.O. Jay and K. Park, Numerical optimal control of parabolic PDEs using DASOPT. In *Large-Scale Optimization with Applications*. Edited by L.T. Biegler, T.F. Coleman, A.R. Conn, and F.N. Santosa. Vol. 93 of *The IMA Volumes in Mathematics and its Applications*. Springer, New York (1997).
- [43] U. Prüfert, OOPDE: An object oriented toolbox for finite elements in Matlab. TU Bergakademie Freiberg (2015).
- [44] U. Prüfert, *Solving optimal PDE control problems. Optimality conditions, algorithms and model reduction*. Habilitation thesis, TU Bergakademie Freiberg (2016).
- [45] D. Ralph and S.J. Wright, Some properties of regularization and penalization schemes for MPECs. *Optim. Methods Softw.* **19** (2004) 527–556.
- [46] S. Scheel and S. Scholtes, Mathematical programs with complementarity constraints: stationarity, optimality, and sensitivity. *Math. Oper. Res.* **25** (2000) 1–22.
- [47] A. Schiela and D. Wachsmuth, Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints. *ESAIM: M2AN* **47** (2013) 771–787.
- [48] S. Scholtes and M. Stöhr, Exact penalization of mathematical programs with equilibrium constraints. *SIAM J. Control Optim.* **37** (1999) 617–652.
- [49] D. Sun and L. Qi, On NCP-functions. *Comput. Optim. Appl.* **13** (1999) 201–220.
- [50] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. American Mathematical Society (2010).
- [51] M. Ulbrich, Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.* **13** (2002) 805–841.
- [52] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MOS-SIAM (2011).
- [53] G. Wachsmuth, Mathematical programs with complementarity constraints in Banach spaces. *J. Optim. Theory Appl.* **166** (2015) 480–507.
- [54] G. Wachsmuth, Towards M-stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Control Optim.* **54** (2016) 964–986.
- [55] G. Wachsmuth, Elliptic quasi-variational inequalities under a smallness assumption: uniqueness, differential stability and optimal control. *Calc. Variat. Partial Differ. Equ.* **59** (2020).
- [56] J. Wloka, *Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben*. Teubner (1982).
- [57] J.J. Ye, Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *J. Math. Anal. Appl.* **307** (2005) 350–369.
- [58] F. Yılmaz and B. Karasözen, An all-at-once approach for the optimal control of the unsteady Burgers equation. *J. Comput. Appl. Math.* **259** (2014) 771–779.