Numerical Optimization Methods for the Optimal Control of Elliptic Variational Inequalities

Thomas M. Surowiec

Abstract The optimal control of variational inequalities introduces a number of additional challenges to PDE-constrained optimization problems both in terms of theory and algorithms. The purpose of this article is to first introduce the theoretical underpinnings and then to illustrate various types of numerical methods for the optimal control of variational inequalities. For a generic problem class, sufficient conditions for the existence of a solution are discussed and subsequently, the various types of multiplier-based optimality conditions are introduced. Finally, a number of function-space-based algorithms for the numerical solution of these control problems are presented. This includes adaptive methods based on penalization or regularization as well as non-smooth approaches based on tools from non-smooth optimization and set-valued analysis. A new type of projected subgradient method based on an approximation of limiting coderivatives is proposed. Moreover, several existing methods are extended to include control constraints. The computational performance of the algorithms is compared and contrasted numerically.

1 Introduction

The optimal control of variational inequalities is a natural extension of PDE-constrained optimization in which the forward problem or underlying PDE is replaced by a variational inequality or convex variational problem. There are a vast number of applications in which variational inequalities or convex variational problems are used. Perhaps the most well-known disciplines are continuum mechanics

T. M. Surowiec (✉)
FB12 Mathematik und Informatik, Philipps-Universität Marburg, Marburg, Germany
e-mail: surowiec@mathematik.uni-marburg.de

[32, 52] and mathematical image processing [67, 70, 72]. However, one also finds variational inequalities in such diverse areas as petroleum engineering [85], digital microfluidics [84], and mathematical finance [51].

Just as in PDE-constrained optimization, we are interested in optimizing or controlling the solution to variational inequalities through one or several parameters. For example, in a packaging problem, we might be interested in determining the optimal force distribution to apply to an elastic membrane in order to obtain a desired shape without passing through a fixed obstacle. In the petroleum engineering application mentioned above, one obtains a variational inequality when using the variational formulation for the steady laminar flow of drill mud modelled as a Bingham fluid. Here, one might wish to determine the optimal pressure needed to affect a desired flow regime.

Due to a lack of Fréchet (or even Gâteaux) differentiability of the control-to-state mapping, a reduced space optimization approach will require either concepts from variational analysis or approximation theory not only to prove existence of a solution but also to derive meaningful first-order optimality conditions and develop efficient numerical methods.

The theoretical and algorithmic challenges remain intact when considering a full-space approach, since the feasible set (after reformulating the variational inequality as a complementarity problem) is both non-convex and fails to satisfy the usual constraint qualifications. For example, in PDE-constrained optimization, one typically uses the Slater or Robinson-Zowe-Kurcyusz constraint qualifications, cf. [48, 77], to ensure the existence and boundedness of the set of Lagrange multipliers. In order to handle these degenerate constraints, one may take a penalty or relaxation approach, though recent work suggests that there are viable theoretical approaches to this "full-space" setting that do not rely on smoothing, see [81].

We now give a brief historical development. We caution the reader that this is only a sampling of the sizeable amount of work published over the past five decades. The purpose is merely to provide an understanding of the current motivations in theory and numerical methods, especially the ideas presented in this paper. Variational inequalities were first introduced by Signorini and solved by Fichera [29]. For this article, the early works by Brezis, Lions, and Stampacchia [18, 60] are perhaps the most relevant as they provide us with a sufficient existence and regularity theory. As a starting point for further study, we also suggest the well-known monographs by Kinderlehrer and Stampacchia and Rodrigues [53, 71]. Perhaps the earliest work on the optimal control of variational inequalities is due to Yvon, in which the adaptive penalty technique for variational inequalities (see Section 4.1) was used in order to approximate the control problem by a more tractable, parameter-dependent problem [86, 87]. This was later used by Lions [58, 59] and fully developed by Barbu [7]. The monograph by Barbu [7] contains many important results and techniques that are still used today, see also related contributions in [9–12, 50, 68, 73]. In 1976, Mignot offered an alternative to the smoothing approach by developing concepts of generalized differentiability [63]. There, he was able to prove and obtain an explicit formula for the Hadamard directional differentiability of the control-to-state mapping of the variational inequality and, with this result,

derive first-order optimality conditions. This result later appeared in [64] and is rederived in [45] using techniques from variational analysis. More recent work in the control of variational inequalities has focused on the development of efficient, function-space-based numerical methods and ever more complex settings as in [2, 4, 21–23, 34–36, 39, 40, 45, 49, 56, 65, 66, 80, 83].

Parallel to the infinite-dimensional developments, a great deal of progress has been made on theory and numerical methods for finite-dimensional mathematical programs with equilibrium or complementarity constraints, as evidenced by the well-known manuscripts by Luo, Pang, and Ralph [61] and Outrata, Kočvara, and Zowe [69]. In addition to the references therein, we also mention the approaches in [3, 74], which have since been extended to infinite dimensions and are in part featured in this paper.

In reference to the existing finite-dimensional literature, recent works have begun referring to the problem class considered here as elliptic mathematical programs with equilibrium constraints (or elliptic MPECs). We will henceforth do the same for the sake of brevity.

The article is structured as follows. After introducing some necessary notation, data assumptions, and the canonical example in Section 2, we discuss sufficient conditions for the existence of a solution in Section 3. Afterwards, we give several kinds of multiplier-based first-order stationarity conditions similar to classical Karush-Kuhn-Tucker (KKT) conditions. These may or may not be first-order optimality conditions depending on the regularity and data of the given problem. For a newcomer to elliptic MPECs, this potentially puzzling array of possible stationarity conditions may seem strange. However, it is necessary to understand the gap in what is theoretically the "best" type of KKT point and what type of stationary point a given numerical method can theoretically guarantee (at least asymptotically).

Following the theoretical results in Section 3, we pass to the main focus of our discussion: numerical optimization methods for elliptic MPECs. First, in Section 4, we consider what we refer to as "regularization-based" methods. In Section 4.1, we present a typical adaptive smoothing method that makes use of approximation theory for variational inequalities and which is directly linked to the derivation of first-order stationarity conditions for the elliptic MPEC. Similarly, in Section 4.2, we use a simple penalization of the complementarity condition (in weak form) to obtain a simpler PDE-constrained optimization problem with control and state constraints. The numerical solution of the subproblems in the adaptive penalty method is thoroughly discussed for a canonical example problem and the numerical behavior is illustrated. We note that the smoothed problems become increasingly difficult to solve as the smoothing parameter tends to zero, which in part motivates the desire for "non-smooth" methods.

The structure of the non-smooth numerical methods section is devised to illustrate the parallels to some popular methods for smooth PDE-constrained optimization problems, e.g., projected-gradient methods (Section 5.1), direct solvers for the KKT system (Section 5.2), and globalization of the direct solvers via a line search (Section 5.3). There are obvious (and expected) limitations to these approaches, all of which arise from the non-smooth or degenerate nature of elliptic

MPECs. Nevertheless, they indicate the possibility of inventing new efficient numerical methods beyond the adaptive smoothing/penalty paradigm, which often outperform the smooth methods in practice.

In Section 5.1, we present a new subgradient method for the reduced space problem. The choice of "subgradient" is motivated by the limiting variational calculus found, e.g., in the book by Mordukhovich [66]. We discuss the asymptotic behavior of this method and illustrate the potential for future study. Following this in Section 5.2, an active-set-based solver with feasibility restoration suggested by Hintermüller in [36] as a direct solver for the C-stationarity system is considered. Finally, in Section 5.3, we consider the so-called bundle-free implicit programming approach [46], which can be understood as a globalization of the active-set solver using a line search. We extend the latter to include control constraints. The performance of the non-smooth methods is then compared and contrasted with that of the adaptive penalty method. All numerical examples have been solved in the Julia programming language [14] version 0.5.0 on a 2016 MacBook Pro Intel(R) Core(TM) i5 @ 3.1 GHz with 16 GB 2133 MHz LPDDR3 RAM.

2   Notation, Assumptions, and Preliminary Results

In this section, we fix our notation and analytical framework. Elliptic variational inequalities and elliptic MPECs are introduced.

2.1   Norms, Inner Products, and Convergence

All spaces are based on the real number field. The Euclidean norm and scalar product on $\mathbb{R}^m$ are denoted by $|x|$ and $x \cdot y$ for $x, y \in \mathbb{R}^m$, respectively. For $r \in \mathbb{R}$, we denote $\max(0, r)$ by $(r)_+$. All other norms are denoted by $\| \cdot \|_X$ for some space X. The topological dual of X is denoted by X^* and the dual pairing by $\langle \cdot, \cdot \rangle_{X, X^*}$. The inner product on a Hilbert space H is denoted by $(\cdot, \cdot)_H$. Given a sequence $\{x_k\} \subset X$, we denote strong convergence to $\overline{x} \in X$ by $x_k \xrightarrow{X} \overline{x}$; for weak convergence, we use $x_k \xrightharpoonup{X} \overline{x}$ and for weak-star convergence, $x_k \xrightharpoonup{X}{}^* \overline{x}$. In all these cases, we drop the sub- or superscripts if it is clear in context.

2.2   Extended Real-Valued Functionals and Convex Analysis

For a topological vector space U, the extended real-valued functional $G : U \to \overline{\mathbb{R}}$ is proper if $G(u) > -\infty$ for all $u \in U$ and $G(w) < +\infty$ for some $w \in U$. G is said to be closed or lower-semicontinuous if its epigraph

$$\text{epi } G := \{(u, \alpha) \in U \times \mathbb{R} \mid G(u) \le \alpha\}$$

is a closed subset in the product topology on $U \times \mathbb{R}$. Moreover, G is convex provided

$$G(\lambda u + (1 - \lambda)w)) \le \lambda G(u) + (1 - \lambda)G(w), \quad \forall u, w \in U, \forall \lambda \in [0, 1].$$

For a Banach space V and $F : V \to \overline{\mathbb{R}}$, the (convex) subdifferential of F at some point $x \in V$ such that $-\infty < F(x) < +\infty$ is defined by the potentially empty set:

$$\partial F(x) := \left\{ x^* \in V^* \mid F(y) \ge F(x) + \langle x^*, y - x \rangle, \forall y \in V \right\}.$$

If $C \subset V$, then the indicator functional i_C of C is defined by

$$i_C(x) = 0 \text{ if } x \in C, \text{ and } i_C(x) = +\infty \text{ otherwise.}$$

If C is non-empty, closed, and convex, then $\mathcal{N}_C(x) := \partial i_C(x)$ is called the convex normal cone to C at $x \in C$. If $K \subset V$ is a non-empty, closed, convex cone, then

$$\xi \in \mathcal{N}_K(x) \iff x \in K, \xi \in K^- := \left\{ \mu \in V^* \mid \langle \mu, y \rangle \le 0, \forall y \in K \right\} \text{ and } \langle \xi, x \rangle = 0.$$

Recall that K^- referred to as the polar cone to K. For more on convex analysis, see, e.g., [25].

2.3 *Function Spaces and 2-Capacity*

Unless noted, $\Omega \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$, is a non-empty, bounded, and open subset with Lipschitz boundary $\Gamma = \partial \Omega$. The Lebesgue measure of $E \subset \mathbb{R}^n$ is denoted by meas(E). For $p \in R$ with $1 \le p < +\infty$, we denote the usual Lebesgue space of p-integrable functions by $L^p(\Omega)$ and the Lebesgue space of essentially bounded functionals by $L^\infty(\Omega)$. The norms are given by

$$\|u\|_{L^p} = \left(\int_\Omega |u(\omega)|^p d\omega \right)^{1/p} \text{ and } \|u\|_{L^\infty} = \text{ess sup}_{\omega \in \Omega} |u(\omega)|,$$

respectively. For $k \in \mathbb{N}$, we denote the Sobolev space of L^p-functions u with $|D^\alpha u| = |(\partial^{\alpha_1} u, \dots, \partial^{\alpha_n} u)| \in L^p(\Omega)$ by $W^{k,p}(\Omega)$, where $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index with $|\alpha_1| + \cdots + |\alpha_n| \le k$ and $\partial^{\alpha_i} u$ is the α_i-th weak partial derivative of u with respect to x_i, $i \in \{1, 2, 3\}$. The $W^{k,p}$-norms are then defined by

$$\|u\|_{W^{k,p}} = \left(\sum_{|\alpha| \le k} \int_\Omega |D^\alpha u(\omega)|^p d\omega \right)^{1/p} \text{ and } \|u\|_{W^{k,\infty}} = \sum_{|\alpha| \le k} \text{ess sup}_{\omega \in \Omega} |D^\alpha u(\omega)|.$$

For our discussion, the most important case is when $p = 2$. Here, both $L^2(\Omega)$ and $H^k(\Omega) := W^{k,2}(\Omega)$ are Hilbert spaces with inner product defined using the norms given above. We denote the space of all H^1-functions with zero trace by $H_0^1(\Omega)$ and its dual by $H^{-1}(\Omega)$. It follows from the Poincaré inequality that

$$\|u\|_{H_0^1} = \|\nabla u\|_{L^2} = (\nabla u, \nabla u)_{L^2}^{1/2}.$$

is an equivalent norm on $H_0^1(\Omega)$. See, e.g., [1] for more. Finally, recall that the 2-capacity of an arbitrary subset $E \subset \Omega$ is given by

$$\text{Cap}_2(E, \Omega) := \inf \left\{ \|u\|_{H_0^1}^2 : v \in H_0^1(\Omega) \ v \geq 1 \text{ a.e. on open neighborhood } G \supset E \right\}.$$

cf. [6, Prop. 5.8.3]. Note that H_0^1-functions possess a representative that is continuous up to a set of positive capacity, cf. [26]. Moreover, there exist $E \subset \Omega$ with $\text{Cap}_2(E, \Omega) > 0$ and $\text{meas}(E) = 0$, e.g., the boundary of a smooth open set. Hence, an H_0^1-function in 2D is continuous across a smooth curve in the plane, but not at single points.

2.4　Elliptic Variational Inequalities

Let H be a Hilbert space and H^* its topological dual. For this subsection, $(\cdot, \cdot)$ denotes the inner product and $\|\cdot\|$ the norm on H. The pairing between H and H^* is denoted by $\langle \cdot, \cdot \rangle$. Let $a : H \times H \to \mathbb{R}$ be a bilinear form on H and $A : H \to H^*$ the associated bounded linear operator, i.e., $a(u, v) = \langle Au, v \rangle$, $u, v \in H$. We recall that a is said to be coercive/elliptic, if there exists some constant $c > 0$ such that $a(v, v) \geq c\|v\|^2$ for all $v \in H$. Clearly, the function $\|v\|_a := (a(v, v))^{1/2}$ defines an equivalent norm on H. Let $K \subset H$ be non-empty, closed, and convex and $w \in H^*$. Then, the variational problem

$$\text{Find } u \in K : a(u, v - u) \geq \langle w, v - u \rangle, \quad \text{for all } v \in K \tag{1}$$

is called an elliptic variational inequality. Note that problems of this type are often referred to as variational inequalities of the first kind. They can be equivalently written as a generalized equation:

$$\text{Find } u \in K : Au + \mathcal{N}_K(u) = Au + \partial i_K(u) \ni w. \tag{2}$$

If we replace i_K by a subdifferentiable proper closed convex functional φ, then we obtain a variational inequality of the second kind. The following is due to Lions and Stampacchia [60], see also [53]:

Theorem 1 *Let $a : H \times H \to \mathbb{R}$ be a coercive bilinear form, $K \subset H$ closed and convex, and $w \in H^*$. Then, (1) possesses a unique solution $S(w)$. Moreover, the solution mapping $w \mapsto S(w)$ is Lipschitz continuous: For all $w_1, w_2 \in H^*$, it holds that*

$$\|S(w_1) - S(w_2)\|_H \leq (1/c)\|w_1 - w_2\|_{H^*},$$

where c is the constant of coercivity of a.

Although the solution mapping S is Lipschitz continuous, it is not necessarily differentiable. In some special cases, S is directionally differentiable. For more on variational inequalities, see, e.g., [53]. The conditions on A and K will be considered standing assumptions on the variational inequality for the rest of this article. We conclude this subsection with an example.

Example 1 In the context of (1), let $H := H_0^1(\Omega)$ and $H^* = H^{-1}(\Omega)$. Moreover, define $\Psi \in H^1(\Omega)$ with $\Psi|_\Gamma \leq 0$ and $K \subset H_0^1(\Omega)$ such that

$$K := \left\{ u \in H_0^1(\Omega) \mid u(x) \geq \Psi(x), \ \text{ for almost every (a.e.) } x \in \Omega \right\}.$$

Clearly, since $\max(0, \Psi) \in H_0^1(\Omega)$, $K \neq \emptyset$. Moreover, one readily shows that K is closed and convex in $H_0^1(\Omega)$. The setting of (1) is general enough to allow for nonsymmetric bilinear forms $a(u, v)$, e.g.,

$$a(u, v) := \int_\Omega \sum_{i,j} a_{ij} \partial_i u \partial_j v - \sum_i b_i(\partial_i u)v + cuv\mathrm{d}x, \quad u, v \in H_0^1(\Omega),$$

with appropriate assumptions on a_{ij}, b_i, c. However, for our purposes it suffices to consider

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v\mathrm{d}x, \quad u, v \in H_0^1(\Omega)$$

Letting $f \in H^{-1}(\Omega)$ and combining the above, we obtain a classical obstacle problem

$$\text{Find } u \in K : \int_\Omega \nabla u \cdot \nabla[v - u]\mathrm{d}x \geq \langle f, v - u \rangle, \quad \text{for all } v \in K.$$

If $\Psi|_\Gamma \equiv 0$, then (without altering the boundary conditions), we obtain an equivalent problem:

$$\text{Find } u \in K_0 : \int_\Omega \nabla u \cdot \nabla[v - u]\mathrm{d}x \geq \langle f + \Delta\Psi, v - u \rangle, \quad \text{for all } v \in K_0,$$

where $K_0 = \left\{ u \in H_0^1(\Omega) \mid u(x) \geq 0, \ \text{a.e. } x \in \Omega \right\}$; see, e.g., [53, 71].

Of course, this set K represents perhaps the easiest kind of set that one can consider in a variational inequality. It is just one example of a so-called "polyhedric set," a notion introduced in the mid-1970s by Haraux [33] and Mignot [63] in the context of sensitivity analysis of variational inequalities. See the recent detailed study [82] for a state-of-the-art on polyhedricity. For more general closed, convex sets, e.g.,

$$\left\{ u \in H_0^1(\Omega) \mid |(\nabla u)(x)| \leq \Psi(x), \text{ a.e. } x \in \Omega \right\} \tag{3}$$

it is possible to prove existence, uniqueness, and continuity results under weak assumptions on Ψ. Since these sets are neither polyhedric nor cones, the differential sensitivity analysis of the solution map and subsequent derivation of optimality conditions or efficient numerical methods is extremely challenging. Elliptic MPECs with this type of constraint were considered in [45]. However, we caution the reader that the proofs for the derivation of the tangent cones are in fact erroneous. Thus, the differential sensitivity results for the solution mapping only hold under the assumption that the tangent cones do in fact have the form purported in the text.

2.5 Elliptic Mathematical Programs with Equilibrium Constraints

We present an abstract framework for a class of elliptic MPECs. Let V, H, and Z be separable Hilbert spaces such that the state space V is a dense subset of H and $V \subset H \subset V^*$ both algebraically and topologically. Moreover, assume that $f \in V^*$, $Z_{\text{ad}} \subset Z$ is a non-empty, closed, and convex set (the set of admissible controls/decision variables), $B : Z \to V^*$ is a bounded linear operator and $F : U \to \overline{\mathbb{R}}$ and $G : Z \to \overline{\mathbb{R}}$. We define an elliptic MPEC as follows:

$$\min J(z, u) := F(u) + G(z) \text{ over } (z, u) \in Z \times V, \tag{4a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \ u \text{ solves } (1) \text{ with } w := Bz + f. \tag{4b}$$

In light of Theorem 1, we can rewrite (4) in reduced form, analogously to standard PDE-constrained optimization problems:

$$\min \ \mathscr{J}(z) := F(S(Bz)) + G(z) \text{ over } z \in Z_{\text{ad}}. \tag{5}$$

Here, $f \in V^*$ is fixed. Therefore, we only write $S(Bz)$ for the solution mapping (instead of $S(Bz + f)$). Obtaining a meaningful full-space formulation of the optimization problem (4) is not always possible. However, if K in (1) is a cone, then we can also formulate a full-space version of the elliptic MPEC by introducing a slack variable $\xi \in V^*$:

$$\min J(z, u) \text{ over } (z, u, \xi) \in Z \times V \times V^*, \tag{6a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \tag{6b}$$

$$Au - \xi = Bz + f, \tag{6c}$$

$$u \in K, \ \xi \in K^+, \ \langle \xi, u \rangle = 0. \tag{6d}$$

Here, $K^+ := -K^-$. This follows from (2). Due to the presence of the complementarity condition (6d), problems of the type (6) are often referred to in the finite-dimensional literature (and more recently in the infinite-dimensional literature) as (elliptic) mathematical programs with complementarity constraints (MPCCs); see e.g., [61, 69, 74] (finite dimensions) and the recent work [81] (infinite dimensions).

Remark 1 We consider here the simplest case in which the control or decision variable z appears only on the "right-hand side" of the variational inequality. However, in many interesting applications, e.g., topology optimization [8], the control enters nonlinearly through the differential operator A, e.g., when $A(z)u = -\text{div}(z\nabla u)$. In the context of elliptic MPECs, this case is scarcely covered in the literature, see [36].

We finalize this section with a canonical example of an elliptic MPEC.

Example 2 In the notation of (4), we set $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, $V^* = H^{-1}(\Omega)$, and $Z = L^2(\Omega)$. For some $u_d \in L^2(\Omega)$, we let F be a standard tracking-type functional and G an L^2-Tikhonov regularization:

$$J(z, u) := F(u) + G(z) = \frac{1}{2}\|u - u_d\|_{L^2}^2 + \frac{\alpha}{2}\|z\|_{L^2}^2, \quad (z, u) \in L^2(\Omega) \times H_0^1(\Omega), \ \alpha > 0.$$

For the forward problem, we choose the simplest form of the obstacle problem with $\Psi \equiv 0$. Then, the following is an elliptic MPEC:

$$\min \frac{1}{2}\|u - u_d\|_{L^2}^2 + \frac{\alpha}{2}\|z\|_{L^2}^2, \text{ over } (z, u) \in L^2(\Omega) \times H_0^1(\Omega), \tag{7a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \ u \in K_0 \text{ solves } : \int_\Omega \nabla u \cdot \nabla[v - u]dx \geq \langle Bz + f, v - u \rangle, \quad \text{for all } v \in K_0, \tag{7b}$$

Some possibilities for Z_{ad} are local bilateral constraints: $-1 \leq w(x) \leq 1$, a.e. $x \in \Omega$ or global constraints such as $\|w\|_{L^2} \leq 1$.

3 Existence and Stationarity Conditions

We start this section in the abstract framework of Section 2.5. In order to provide insight into the deeper meaning of the various stationarity conditions, we ultimately restrict ourselves to the canonical example (Example 2).

3.1 Existence of a Solution

Under suitable conditions, on F, G, Z_{ad}, and B, we can use Weierstrass's existence theorem, see, e.g., [6, Thm. 3.2.5] to prove that the reduced form MPEC has a solution. For the following result, we appeal to the monograph by Barbu [7, Chap. 3].

Theorem 2 *In the context of problem* (4)*, we assume the following:*

1. $F : H \to \mathbb{R}$ *is locally Lipschitz and nonnegative.*
2. $G : Z \to \overline{\mathbb{R}}$ *is convex, lower-semicontinuous, and for some constants* $\kappa_1 > 0$, $\kappa_2 \in \mathbb{R}$. $G(z) \geq \kappa_1 \|z\|_Z + \kappa_2$, $\forall z \in Z$
3. B *is completely continuous.*

Then, (4)*, or equivalently* (5)*, admits a solution.*

Proof By replacing $G(z)$ with $G(z) + i_{Z_{ad}}(z)$, the assertion follows from [7, Prop. 3.1]. In particular, we note that by Barbu [7, Lem. 3.1] the mapping $(S \circ B) : Z \to V$ is completely continuous and, furthermore, the composite objective functional $(F \circ S \circ B) : Z \to \mathbb{R}_+$ is weakly lower-semicontinuous.

Since for any $z \in Z_{ad}$, (1) has a unique solution $u = S(Bz)$. There exists a unique $\xi \in \mathcal{N}_K(u)$ such that $\xi = Au - Bz - f$. Conversely, if (z, u, ξ) satisfies (6b)–(6d), then $z \in Z_{ad}$ and $u = S(Bz)$. In other words, there is a one-to-one correspondence between the feasible set of (5) and (6). Therefore, if (4), or equivalently (5), admits a solution, then so does (6).

Corollary 1 *Under the assumptions of Theorem 2,* (6) *admits a solution.*

Remark 2 Ignoring Theorem 2, we could also prove this assertion by appealing to [6, Thm. 3.2.5]. Indeed, under the hypotheses of Theorem 2, (6) can be viewed as an unconstrained problem in which the associated objective functional is radially unbounded, proper, and weakly lower-semicontinuous.

3.2 Primal Stationarity Conditions

In this subsection, we restrict ourselves to the setting of Example 2. As mentioned earlier, there are some situations in which the control-to-state mapping S is directionally differentiable. In particular, Mignot [63] demonstrated that the control-to-state mapping in Example 2 is directionally differentiable in the sense of Hadamard. The next result draws the parallel to PDE-constrained optimization problems.

Theorem 3 *Let* $(\overline{z}, \overline{u})$ *be a solution to* (7)*, then the following first-order necessary optimality condition holds:*

$$\alpha(\overline{z}, w - \overline{z}) + (\overline{u} - u_d, S'(\overline{u}; B(w - \overline{z}))) \geq 0, \quad \forall w \in Z_{ad}. \tag{8}$$

In the finite-dimensional literature, e.g., in [61, 74], primal first-order conditions of the type (8) are often referred to as B-stationarity conditions. There, the "B" comes from the fact that either the so-called Bouligand tangent cone or Bouligand differentiability is used. This motivates the next definition.

Definition 1 If $(\overline{z}, \overline{u})$ is a feasible point of (7) that satisfies (8), then $(\overline{z}, \overline{u})$ is B-stationary.

For the setting in (5), it is possible to derive B-stationarity conditions provided that F, G are smooth. Clearly, the main challenge is determining directional differentiability of S. Suppose for the sake of argument that $S'(\overline{u}; du)$ is linear in $du \in V^*$. Then, we can "dualize" the B-stationarity conditions (8):

$$\alpha(\overline{z}, w - \overline{z}) + (B^* p, w - \overline{z}) \geq 0, \quad \forall w \in Z_{\mathrm{ad}}, \quad p = S'(\overline{u})^*(u - u_d).$$

Note that the condition "$p = S'(\overline{u})^*(u - u_d)$" essentially means (in this ideal case) that p solves a type of adjoint equation. Using this "adjoint state" p, we could now develop solution algorithms. However, $S'(\overline{u}; du)$ is in general nonlinear (albeit Lipschitz) in du and thus, the standard adjoint state p does not exist. As a result, there are several different types of dual/multiplier-based first-order stationarity conditions reminiscent of classical KKT conditions in nonlinear programming.

3.3 Dual Stationarity Conditions

In this subsection, we again restrict ourselves to the setting in Example 2. Furthermore, we assume that Γ is regular enough to guarantee that $S(Bz) \in H^2(\Omega) \cap H_0^1(\Omega)$ for any $z \in Z_{\mathrm{ad}}$. This is possible if, e.g., $Bz + f \in L^2(\Omega)$ and Ω is a convex polyhedron, cf. [18, 71]. This regularity assumption ensures that $\xi \in L^2(\Omega)$ and, by the Sobolev embedding theorem, $S(Bz)$ is at least Hölder continuous on Ω. We now introduce two multiplier-based first-order stationarity conditions taken from [39, 40].

Definition 2 Let $z \in Z_{\mathrm{ad}}$ and $u = S(Bz)$ with associated slack variable $\xi \in L^2(\Omega)$. The set $\Omega^0 := \{x \in \Omega \mid u(x) = 0\}$ is called the active set or coincidence set and $I := \{x \in \Omega \mid u(x) > 0\}$ the inactive set. The set $\Omega^{00} := \Omega^0 \cap \{x \in \Omega \mid \xi(x) = 0\}$ is called the biactive or weakly active set and $\Omega^{0+} := \Omega^0 \cap \{x \in \Omega \mid \xi(x) > 0\}$ the strongly active set.

Note that the biactive and strongly active sets are only defined up to a set of Lebesgue measure zero. In contrast, the active and inactive sets may be more finely defined up to sets of positive capacity even in less regular settings.

Definition 3 The point $(z, u, \xi) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ is called a C-stationary point for (7) provided that there exist $p \in H_0^1(\Omega)$ and $\lambda \in H^{-1}(\Omega)$ such that the following system of equations is fulfilled:

$$\alpha(\overline{z}, w - \overline{z}) + (B^* p, w - \overline{z}) \geq 0, \quad \forall w \in Z_{\text{ad}} \tag{9a}$$

$$A^* p - \lambda = u_d - u \tag{9b}$$

$$Au - \xi = Bz + f \tag{9c}$$

$$\xi \geq 0, \text{ a.e. } x \in \Omega \tag{9d}$$

$$u \geq 0, \text{ a.e. } x \in \Omega \tag{9e}$$

$$(\xi, u) = 0 \tag{9f}$$

$$\langle \lambda, p \rangle \leq 0 \tag{9g}$$

$$p = 0, \text{ a.e. } x \in \Omega^{0+} \tag{9h}$$

$$\langle \lambda, \phi \rangle = 0, \forall \phi \in H_0^1(\Omega) : \phi = 0, \text{ a.e. } \Omega^0 \tag{9i}$$

If, in addition to (9), p and λ satisfy:

$$p \leq 0, \text{ a.e. } x \in \Omega^{00} \tag{10a}$$

$$\langle \lambda, \phi \rangle \geq 0, \forall \phi \in H_0^1(\Omega) : \phi \geq 0, \text{ a.e. } \Omega^{00}, \phi = 0, \text{ a.e. } \Omega \setminus (I \cup \Omega^{00}), \tag{10b}$$

then $(z, u, \xi) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ is called an S-stationary point.

Several comments on this system are necessary. To start, we note that if K_0 is replaced by the entire space $H_0^1(\Omega)$, then ξ, λ and the conditions (9d)–(10b) vanish and we obtain the usual first-order system for a linear-quadratic PDE-constrained optimization problem with convex control constraints. Moreover, if there is no biactive set, then both (in fact all) dual stationarity concepts coincide. Thus, biactivity is essentially the source of all difficulties in the study of MPECs. We will see later that it also relates to the Gâteaux differentiability of the control-to-state map.

Perhaps the most critical point here is the usage of almost everywhere versus quasi everywhere conditions for the multipliers. As defined above, these S-stationarity conditions are not equivalent to the necessary first-order optimality conditions in the pioneering works [63, 64]; unless certain regularity assumptions on the active/biactive/inactive sets are made. In [63, 64] (see also [43, 80]), the notion of capacity (in contrast to Lebesgue measure) is used. This is because capacity is needed for the correct representation of the tangent/contingent cone and its polar. When replacing q.e. by a.e. some conditions become stronger and others finer.

In order to see this, suppose that $\text{meas}(\Omega^0) = 0$, but $\text{Cap}_2(\Omega^0, \Omega) > 0$. Then, (9i) would imply $\lambda = 0 \in H^{-1}(\Omega)$. However, when using capacity, since the test functions ϕ admit quasi-continuous representatives, requiring $\phi = 0$ "quasi-everywhere" (q.e.) on Ω^0 would shrink the set of test functions so that λ is not necessarily zero. Similar problems arise by requiring a.e. instead of q.e. conditions on the adjoint variables p. Indeed, if Ω^{0+} has measure zero, then (9h) is trivially

satisfied by any p; this would not be the case if we use q.e. since $p \in H_0^1(\Omega)$ admits a quasi-continuous representative.

Therefore, the stationarity conditions of Mignot and Puel, see (29) below, should be considered the true strong/S-stationarity conditions for our canonical MPEC as opposed to (9)–(10). For more complex constraints in which the image space of the constraint mapping is an L^p-space, as in, e.g.,[34, 35, 45], the problems with defining active sets up to sets of capacity zero seem to be absent.

Nevertheless, (9)–(10) arose naturally through the limiting process of an adaptive penalty scheme. As such, they are highly relevant for the study of numerical methods for elliptic MPECs, especially for methods utilizing smoothing plus continuation. In contrast, it is unclear how to properly include notions of capacity into efficient numerical methods (without simplifying assumptions as in [46]).

Finally, in many complex real-world applications, it might be impossible to obtain even C-stationarity conditions of this type, due to a lack of compactness and regularity properties. For example, in the optimal control of electrowetting on dielectrics[4], Allen-Cahn [27, 28], Cahn-Hilliard [20], or Cahn-Hilliard-Navier-Stokes system with obstacle potentials [47], one can usually only derive an approximation of (9g). Here, this would be equivalent to replacing (9g) by

$$\limsup_{k \to \infty} \langle \lambda_k, p_k \rangle \leq 0$$

for sequences $\{\lambda_k\}$ and $\{p_k\}$ with $\lambda_k \rightharpoonup \lambda$ and $p_k \rightharpoonup p$.

In conclusion, the various notions of dual stationarity and the theory needed to derive them are still active areas of research. The main point here is that there is a stratification of concepts that one should also be aware of when considering the design and convergence of numerical methods. That is, it is not enough to prove that a scheme converges but also to what kind of stationary point. Of course, many of the issues involving capacity, weak topologies, products of weakly convergent sequences, and weak lim-inf or lim-sup-type sign conditions will not necessarily appear in numerical experiments.

4 Regularization-Based Methods

In this and the coming sections, we present a number of numerical optimization methods for elliptic MPECs. These are split into two classes: Methods that employ adaptive smoothing, relaxation, or penalization of the forward problem or complementarity constraints (regularization-based methods) and those that do not (non-smooth methods).

We also note that the proper discretization of elliptic MPECs using (adaptive) finite elements schemes that take into account the additional difficulties due to the inherent degeneracy/non-smoothness has only been considered in a handful of papers. For example, we mention the most recent works [16, 17, 24, 37, 62]. For

our numerical study, we make use of a simple finite difference scheme to discretize the operators along with a nested grid approach to simulate mesh refinements. This allows us to easily compare the methods.

The regularization-based methods all follow a similar scheme. First, the variational inequality is approximated by a parameter-dependent semilinear elliptic PDE or the complementarity constraint is relaxed or penalized. This yields in both cases a more tractable family of approximating optimization problems. Next, the smooth PDE-constrained problems are solved, yielding a parameter-dependent KKT point. Finally, continuation is performed on the regularization parameter (passing to 0 or $+\infty$).

The non-smooth methods are rather different and there is still plenty of room for new ideas. We only mention here that with these methods the emphasis is placed on directly solving the original, non-smooth problem without changing the forward problem. We postpone further details until later.

4.1 An Adaptive Penalty Method

We begin in the abstract framework of Section 2.5 and present a general approximation result as found in [7, Thm. 2.2]. For a comprehensive study on the numerical analysis and approximation of variational inequalities, see, e.g., [31]. In the following, ϕ^ϵ refers to some penalty functional for the constraint K.

Definition 4 For any constant $\epsilon > 0$, let $\phi^\epsilon : V \to \mathbb{R}$ such that ϕ^ε is convex and Fréchet differentiable on V and satisfies

1. There exists a C, independent of ϵ, u with $\phi^\epsilon(u) \geq -C(\|u\|_V + 1)$ for all $\varepsilon > 0$, $u \in V$.
2. $\phi^\epsilon(u) \to i_K(u)$ as $\epsilon \downarrow 0$ for all $u \in V$.
3. For all $u \in V$ and for all $\{u_\epsilon\}$ such that $u_\epsilon \rightharpoonup u$ as $\epsilon \downarrow 0$, $\liminf_{\epsilon \downarrow 0} \phi^\epsilon(u_\epsilon) \geq i_K(u)$.

Remark 3 See also [7, Thm 2.4] for functionals on H, which is more relevant for (13).

In the abstract setting, we approximate (1) by

$$Au + \nabla\phi^\epsilon(u) = Bz + f \tag{11}$$

We denote the approximate solution mapping by S_ϵ. For the next result, we restrict ourselves to the case when A is symmetric (as in Example 2), see [7, Thm. 2.2].

Theorem 4 *Let $w_\epsilon \to w$ in V^* as $\epsilon \downarrow 0$, then the sequence $\{u_\epsilon\} \subset V$ with $u_\epsilon := S_\epsilon(w_\epsilon)$ converges weakly to $u = S(w)$ as $\epsilon \downarrow 0$. If in addition,*

$$\langle \nabla\phi^\epsilon(y) - \nabla\phi^\delta(v), y - v \rangle \geq -C(1 + (\|\nabla\phi^\epsilon(y)\|_{V^*}^2 + \|\nabla\phi^\delta(v)\|_{V^*}^2))(\epsilon + \delta),$$

$$\forall \epsilon, \delta > 0, \forall y, v \in V, \qquad (12)$$

then $u_\epsilon \to u$ (strongly in V) as $\epsilon \downarrow 0$.

In the setting of Example 2, one possibility for ϕ^ϵ is:

$$\phi^\epsilon(u) := (2\epsilon)^{-1}\|(-u)_+\|_{L^2}^2. \qquad (13)$$

Redefining K_0 for L^2-functions, ϕ^ϵ in (13) is the Moreau-Yosida regularization of the indicator functional i_{K_0} with respect to the L^2-topology, cf. [5, Sec. 2.7]. However, in order to solve the smoothed MPECs numerically, we will require more smoothness of ϕ^ϵ below. We now approximate (5) by:

$$\min \mathscr{J}_\epsilon(z) := F(S_\epsilon(Bz)) + G(z) \text{ over } z \in Z_{\text{ad}}, \qquad (14)$$

Thus, the existence theory, optimality conditions, and numerical methods for (14) reduces to results from PDE-constrained optimization.

Turning to Example 2, we employ a smoothed plus function in (11) characterized by:

$$\nabla\phi^\epsilon(u) = \begin{cases} \epsilon^{-1}u - 0.5, & \text{if } u \geq \epsilon, \\ \frac{u^3}{\epsilon^3} - \frac{u^4}{2\epsilon^4}, & \text{if } u \in (0, \epsilon), \\ 0, & \text{else,} \end{cases} \quad \nabla^2\phi^\epsilon(u) = \begin{cases} \epsilon^{-1}, & \text{if } u \geq \epsilon, \\ \frac{3u^2}{\epsilon^3} - \frac{2u^3}{\epsilon^4}, & \text{if } u \in (0, \epsilon), \\ 0. & \text{else.} \end{cases}$$

$$(15)$$

We can then prove that the approximate solution mapping S_ϵ is Fréchet differentiable. Then, for a solution (z_ϵ, u_ϵ), there exists an adjoint state $p_\epsilon \in V$ such that

$$\alpha(z_\epsilon, w - z_\epsilon) + (B^*p_\epsilon, w - z_\epsilon) \geq 0, \quad \forall w \in Z_{\text{ad}} \qquad (16a)$$

$$A^*p_\epsilon - \lambda_\epsilon = u_d - u_\epsilon, \qquad (16b)$$

$$Au_\epsilon - \xi_\epsilon, = Bz_\epsilon + f, \qquad (16c)$$

$$\xi_\epsilon = \nabla\phi^\epsilon(-u_\epsilon), \text{ a.e. } x \in \Omega \qquad (16d)$$

$$\lambda_\epsilon = -\nabla^2\phi^\epsilon(-u_\epsilon)p_\epsilon, \text{ a.e. } x \in \Omega, \qquad (16e)$$

holds; cf. the techniques in [57, 77].

We note here that the case of gradient constraints mentioned in (3) can also be treated using such a penalty method. However, using a standard quadratic penalty/Moreau-Yosida-type approach yields a penalized state equation of the form:

$$-\text{div}((1 + \xi_\epsilon)\nabla u_\epsilon) = Bz_\epsilon + f, \qquad (17a)$$

$$\xi_\epsilon = \frac{1}{\epsilon}\left(1 - \frac{\Psi}{|\nabla u_\epsilon|}\right)_+, \qquad (17b)$$

which is considerably more challenging due to the bilinear dependence on u_ϵ in the PDE and accompanying non-smooth first-order PDE.

From (16), it is possible to derive C-stationarity conditions and, under further assumptions, even S-stationarity conditions, cf. [39, 40], by passing to the limit in ϵ. More specifically, we show that a subsequence of stationary points for (16) converges to a point that satisfies C-stationarity for (7). Therefore, if we have a numerical method that solves (16), then by performing continuation on $\epsilon \downarrow 0$ we have a means of numerically approximating C-stationary points (or better) for (7). This furnishes a convergence proof in function space for the "outer loop" of the method, depicted in Algorithm 4.1. For the interested reader, we provide a short discussion of the limiting arguments in Detail 5.

Algorithm 4.1 Adaptive penalty method: outer loop

Input: $u_d, f \in L^2(\Omega)$, $\beta \in (0, 1)$;
1: Choose $\epsilon^0 > 0$, $(z^0, u^0, \xi^0, p^0, \lambda^0) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega)$ and set $k := 0$;
2: **repeat**
3:   Compute a stationary point $(z^{k+1}, u^{k+1}, \xi^{k+1}, p^{k+1}, \lambda^{k+1})$ of (16) with $\epsilon = \epsilon^k$ using an iterative scheme with initial value $(z^k, u^k, \xi^k, p^k, \lambda^k)$;
4:   Set $\epsilon^{k+1} = \beta \epsilon^k$
5: **until** some stopping rule is satisfied.

Detail 5 (Sketch of the Limiting Technique) We first note that (16a) is equivalent to

$$z_\epsilon = \text{Proj}_{Z_{\text{ad}}}(-\frac{1}{\alpha} B^* p_\epsilon). \tag{18}$$

Suppose Z_{ad} is bounded, then $\{z_\epsilon\}_{\epsilon > 0}$ is bounded in Z. Let $\epsilon_k \downarrow 0$. Then, there exists a subsequence $\{z_l\}$ with $z_l = z_{\epsilon_{k_l}} \overset{Z}{\rightharpoonup} z^\star \in Z_{ad}$. Hence, $S_{\epsilon_{k_l}}(Bz_l) =: u_l \overset{V}{\to} u^\star = S(Bz^\star)$ by Theorem 4 due to the complete continuity of B. Therefore, $\xi_l \to \xi^\star := Bz^\star + f - Au^\star \in -\mathcal{N}_{K_0}(u^\star)$, i.e., $\xi^\star, u^\star$ satisfy. For the adjoint state p_ϵ and multiplier λ_ϵ, we test the adjoint equation (16b) with p_ϵ:

$$c_1\|p_\epsilon\|_V^2 \leq \langle A^* p_\epsilon - \lambda_\epsilon, p_\epsilon \rangle = (u_d - u_\epsilon, p_\epsilon) \leq$$
$$\|u_d - u_\epsilon\|_H \|p_\epsilon\|_H \leq c_2 \|u_d - u_\epsilon\|_H \|p_\epsilon\|_V$$

for scalars $c_1, c_2 > 0$, independent of ϵ. Since $u_l \overset{V}{\to} u^\star$, $\{p_l\}$ is bounded in V. Hence, there exists $\{p_{l_m}\}$ with $p_{l_m} \overset{V}{\rightharpoonup} p^\star$ and, by substitution, $\lambda_{l_m} \overset{V^*}{\rightharpoonup} \lambda^\star = u^\star - u_d + A^* p^\star$.

Without knowledge of the best possible system, one might stop at this point; however, the theory indicates that a much more refined system is possible. We quickly demonstrate (9g). The remaining conditions (9h), (9i) require lengthy

arguments that go beyond the scope of this work. Returning to (16b), it is clear that

$$\langle A^* p_\epsilon, p_\epsilon \rangle - (u_d - u_\epsilon, p_\epsilon) = \langle \lambda_\epsilon, p_\epsilon \rangle = - \int_\Omega \nabla^2 \phi^\epsilon(-u_\epsilon)|p_\epsilon|^2 dx \leq 0.$$

Since A is symmetric, the bilinear form $a(\cdot, \cdot)$ is weakly lower-semicontinuous. Moreover, due to the Rellich-Kondrachov theorem, both $u_\epsilon \overset{H}{\to} u^\star$, $p_\epsilon \overset{H}{\to} p^\star$. Thus,

$$\langle \lambda^\star, p^\star \rangle = \langle A^* p^\star, p^\star \rangle - (u_d - u^\star, p^\star) \leq \liminf_{m \to +\infty} \langle A^* p_{l_m}, p_{l_m} \rangle - (u_d - u_{l_m}, p_{l_m}) \leq 0. \tag{19}$$

In [7], a more general setting allows a similar argument. However in general, e.g., if A is not symmetric, we only have:

$$\limsup_{\epsilon \downarrow 0} \langle \lambda_\epsilon, p_\epsilon \rangle \leq 0$$

as mentioned at the end of Section 3.3.   □

A possible stopping rule in Algorithm 4.1 could be the residual of the C-stationarity system (9) or when a minimum value of ϵ^{k+1} is reached, e.g., machine precision. Obviously, the most computationally demanding part here is step 3. Assuming that $\text{Proj}_{Z_{ad}}$ is relatively simple to calculate, e.g., for local bilateral constraints, then (16) reduces to

$$A^* p_\epsilon + \nabla^2 \phi^\epsilon(-u_\epsilon)p_\epsilon = u_d - u_\epsilon, \tag{20a}$$

$$Au_\epsilon - \nabla \phi^\epsilon(-u_\epsilon) = B\text{Proj}_{Z_{ad}}(-\frac{1}{\alpha}B^* p_\epsilon) + f. \tag{20b}$$

One could then solve (20) using, e.g., a semismooth Newton method. Alternatively, an interior point approach for the projection might be employed. We briefly recall the semismooth Newton method in infinite-dimensional spaces as discussed in [19, 38, 79]. Let X, Y be Banach spaces, $D \subset X$ an open subset of X, and $\mathscr{F} : D \to Y$.

Definition 5 The mapping $\mathscr{F} : D \subset X \to Y$ is said to be Newton-differentiable on the open subset $U \subset D$, if there exists a family of mappings $\mathscr{G} : U \to \mathscr{L}(X, Y)$ such that

$$\|\mathscr{F}(x + h) - \mathscr{F}(x) - \mathscr{G}(x + h)h\|_Y = o(\|h\|_X), \forall x \in U.$$

$\mathscr{G}$ is called the Newton derivative for $\mathscr{F}$ on U. In [38], it is shown that

$$\mathscr{G}_\delta(y)(x) = \begin{cases} 1 \text{ if } y(x) > 0 \\ 0 \text{ if } y(x) < 0 \\ \delta \text{ if } y(x) = 0 \end{cases} \tag{21}$$

for every $y \in X$ and $\delta \in \mathbb{R}$ is a Newton derivative of the $\max(0, \cdot)$, under the condition that $\max(0, \cdot) : L^p(\Omega) \to L^q(\Omega)$ with $1 \leq q < p \leq \infty$.

Therefore, if a mapping $\mathscr{F}$ is Newton-differentiable, then using the concepts above leads to a generalized Newton step for the equation $\mathscr{F}(x) = 0$, see, e.g., [19, 38].

Theorem 6 *Suppose that $\mathscr{F}(x^*) = 0$ and that $\mathscr{F}$ is Newton-differentiable on an open neighborhood U of x^* with Newton derivative $\mathscr{G}$. If $\mathscr{G}(x)$ is nonsingular for all $x \in U$ and the set $\left\{ ||\mathscr{G}(x)^{-1}||_{\mathscr{L}(Y,X)} : x \in U \right\}$ is bounded, then the semismooth Newton iteration*

$$x_{l+1} = x_l - \mathscr{G}(x_l)^{-1}\mathscr{F}(x_l), \quad l = 0, 1, 2, \ldots \tag{22}$$

converges superlinearly to x^*, provided that $||x_0 - x^*||_X$ is sufficiently small.

Under the assumption that $Z_{\mathrm{ad}} := \left\{ v \in L^2(\Omega) \,|\, a \leq v \leq b, \text{ a.e. } x \in \Omega \right\}$ with $a, b \in L^2(\Omega)$ and $a < b$, we can derive a semismooth Newton step for the solution of (20) in function space: Fix some $(u, p) \in H_0^1(\Omega) \times H_0^1(\Omega)$ and define the following subsets of Ω:

$$\Omega^a := \left\{ x \in \Omega \,\Big|\, a(x) + \alpha^{-1}(B^*p)(x) > 0 \right\},$$

$$\Omega^b := \left\{ x \in \Omega \,\Big|\, -\alpha^{-1}(B^*p)(x) - b(x) > 0 \right\}.$$

Moreover, let $\Omega^{ina} := \Omega \setminus (\Omega^a \cup \Omega^b)$ (up to a set of Lebesgue measure zero) and define the residual $\mathscr{F}_\epsilon(u, p)$ of (20) by:

$$\mathscr{F}_\epsilon^1(u, p) := Au - \nabla\phi^\epsilon(-u) - B\mathrm{Proj}_{Z_{\mathrm{ad}}}\left(-\frac{1}{\alpha}B^*p\right) - f,$$

$$\mathscr{F}_\epsilon^2(u, p) := A^*p + \nabla^2\phi^\epsilon(-u)p - u_d + u.$$

Since

$$z = \mathrm{Proj}_{Z_{\mathrm{ad}}}(-\alpha^{-1}B^*p) = -\alpha^{-1}B^*p - (-\alpha^{-1}B^*p - b)_+ + (a + \alpha^{-1}B^*p)_+, \tag{23}$$

we can use (21) to obtain a Newton derivative $\mathscr{G}$ for $\mathscr{F}$:

$$\mathscr{G}_\epsilon(u, p) = \begin{bmatrix} A + \nabla^2\phi^\epsilon(-u) & \alpha^{-1}\chi_{\Omega^{ina}}BB^* \\ I - \nabla^3\phi^\epsilon(-u)p & A^* + \nabla^2\phi^\epsilon(-u) \end{bmatrix},$$

where $\chi_{\Omega^{ina}}$ is the characteristic function for the set Ω^{ina}. If $(\delta u, \delta p)$ denotes the difference between the new iterate and the current iterate in the semismooth Newton step, then at each iteration, we solve

$$\begin{bmatrix} A + \nabla^2 \phi^\epsilon(-u) & \alpha^{-1}\chi_{\Omega^{ina}}BB^* \\ I - \nabla^3 \phi^\epsilon(-u)p & A^* + \nabla^2 \phi^\epsilon(-u) \end{bmatrix} \begin{bmatrix} \delta u \\ \delta p \end{bmatrix} = -\mathscr{F}_\epsilon(u, p). \qquad (24)$$

If we can show that $\mathscr{G}_\epsilon(u, p)$ is invertible independently of (u, p) (for fixed $\epsilon > 0$) so that the set $\{ \|\mathscr{G}_\epsilon(u, p)^{-1}\| : (u, p) \in H_0^1(\Omega) \times H_0^1(\Omega)\}$ is bounded, then we are guaranteed to have local superlinear convergence. This leads to Algorithm 4.2.

Algorithm 4.2 Adaptive penalty method: inner loop

Input: tol > 0, $a, b, u_d, f \in L^2(\Omega)$, $\epsilon_k > 0$, $(u_0^k, p_0^k) \in H_0^1(\Omega) \times H_0^1(\Omega)$, $l := 0$;
1: **repeat**
2:     Compute a step $(\delta u_l^k, \delta p_l^k)$ by solving (24) with $(u, p) := (u_l^k, p_l^k)$, $\epsilon = \epsilon_k$.
3:     Set $u_{l+1}^k := u_l^k + \delta u_l^k$, $p_{l+1}^k := p_l^k + \delta p_l^k$, $l := l + 1$;
4: **until** $\|\mathscr{F}(u_l^k, p_l^k)\| <$ tol

We conclude this subsection with a numerical experiment. This is used in part to compare to the non-smooth numerical methods in later sections.

Example 3 Let $\Omega = (0, 1)^2$, $\alpha = 1$, $b \equiv 0.035$, $a \equiv 0$, and $A = -\Delta$ (associated with $H_0^1(\Omega)$). Defining

$$y^\dagger(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} 160(\mathbf{x}_1^3 - \mathbf{x}_1^2 + 0.25\mathbf{x}_1)(\mathbf{x}_2^3 - \mathbf{x}_2^2 + 0.25\mathbf{x}_2) & \text{in } (0, 0.5) \times (0, 0.5), \\ 0 & \text{else,} \end{cases}$$

$$\xi^\dagger(\mathbf{x}_1, \mathbf{x}_2) = \max(0, -2|\mathbf{x}_1 - 0.8| - 2|\mathbf{x}_1\mathbf{x}_2 - 0.3| + 0.5),$$

we set

$$f = -\Delta y^\dagger - y^\dagger - \xi^\dagger, \text{ and } u_d = y^\dagger + \xi^\dagger - \alpha \Delta y^\dagger.$$

The example is chosen due to the nontrivial biactive set and overlap of active sets for the control and state constraints. The only change to [40, Exp. 5.1] is the addition of control constraints, which were set to $\pm\infty$ there. Concerning the discretization and solution, we use a standard five-point stencil to discretize the negative Laplace operator with finite differences. The problem is solved on a uniform mesh with 512^2 grid points.

We start the algorithm at $(z^0, u^0, \xi^0) = (0, 0, 0)$. The stopping criterion is based on the L^2-norm of the residual with stopping tolerance of 10^{-9}. For this example, the ϵ-update in 4.1 proved to be extremely sensitive, meaning that a reasonably aggressive update strategy, e.g., $\epsilon_{k+1} = \epsilon_k/2$ failed once $\epsilon_k = \mathcal{O}(10^{-4})$. To be fair, one would normally not cold-start this algorithm on such a fine grid. Opting instead

for either an adaptive FEM strategy, multigrid scheme (as in [40]), or a nested grid strategy (as in [39, 46]) would certainly improve the performance and allow for a more aggressive update strategy for the smoothing parameter. Nevertheless, we see that as the penalized problem approaches the original non-smooth non-convex problem, the nonlinear system becomes increasingly more difficult to solve (eventually even failing). See Table 1 for the convergence history and Figure 1 for plots of the solution.

4.2 An ℓ^1 Penalty Method

In this subsection, we present a technique originating in the finite-dimensional MPEC literature [3]. The extension to infinite dimensions can be found here [43]. The idea is elegant in its simplicity and allows us to approximate the elliptic MPEC by a sequence of PDE-constrained optimization problems with control and state constraints. Moreover, instead of a semilinear elliptic PDE, as in the previous method, we have a linear elliptic PDE. We begin in the abstract framework of Section 2.5 under the assumption that K is a cone and then pass to the problem in Example 2.

Using an ℓ^1-penalty for the condition $\langle u, \xi \rangle = 0$, we approximate (6) by

$$\min J(z, u) + \frac{1}{\epsilon} |\langle u, \xi \rangle| \text{ over } (z, u, \xi) \in Z \times V \times V^*, \tag{25a}$$

$$\text{s.t. } z \in Z_{\text{ad}}, \tag{25b}$$

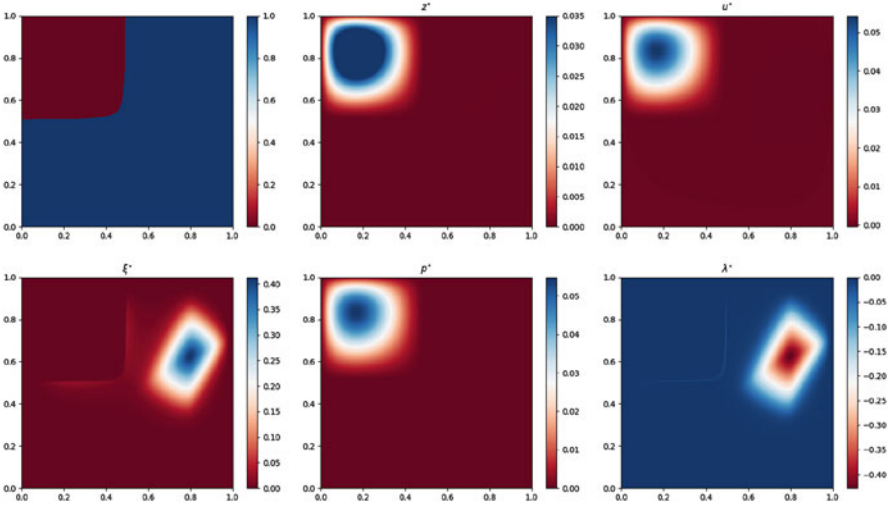$$Au - \xi = Bz + f, \tag{25c}$$

$$u \in K, \ \xi \in K^+. \tag{25d}$$

By definition of K^+, $\langle u, \xi \rangle \geq 0$. Hence, for $(u, \xi) \in K \times K^+$, $\frac{1}{\epsilon} |\langle u, \xi \rangle| = \frac{1}{\epsilon} \langle u, \xi \rangle$, which yields the smooth objective $J_\epsilon(z, u, \xi) := J(z, u) + \frac{1}{\epsilon} \langle u, \xi \rangle$.

The analysis for this problem requires several technical results. Nevertheless, under appropriate regularity and boundedness assumptions, one can still show existence of a solution, consistency of the approximation, derive first-order conditions and (after passing to the limit in ϵ) obtain a (weak) form of C-stationarity. We briefly sketch the ideas here and refer the reader to [43, Section 2] for the detailed technical analysis in the context of Example 2.

Detail 7 (Sketch of Existence and Consistency Arguments) In addition to the assumptions in Theorem 2, let F be weakly lower-semicontinuous, Z_{ad} bounded, and A symmetric.

To show that (25) has a solution, we prove the boundedness of infimizing sequence $\{(z_k, u_k, \xi_k)\}$ (despite the unboundedness of K^+ and lack of coercivity of $J_\epsilon(z, u, \xi)$). Let (z_0, u_0, ξ_0) be feasible for (6). Then, for all sufficiently large $k \in N$,

Table 1 Convergence history of the adaptive penalty method Algorithm 4.1 with stopping tolerance tol $= 10^{-9}$, starting $\epsilon_1 = 1$ using $\epsilon_k = 2^{-(k-1)}$ with $k = 1, \ldots, 12$, l represents the iterations per inner loop of Algorithm 4.2

l \ ϵ	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	2^{-11}
1	3.037e−04	2.284e−03	1.547e−02	7.270e−02	1.581e−01	2.013e−01	2.247e−01	2.381e−01
2	2.120e−04	1.169e−02	4.510e−01	5.556e+00	1.852e+01	3.929e+01	7.506e+01	1.407e+02
3	1.524e−09	4.459e−06	5.586e−03	6.919e−01	6.179e+00	1.755e+01	3.830e+01	7.712e+01
4	1.550e−12	1.856e−12	8.915e−07	8.776e−03	4.555e−01	2.449e+00	7.160e+00	1.683e+01
5			1.561e−12	1.441e−06	2.453e−03	4.973e−02	2.747e−01	9.366e−01
6				1.824e−12	8.096e−08	2.439e−05	4.920e−04	3.762e−03
7					1.817e−12	7.139e−12	2.038e−09	9.028e−08
8							4.209e−12	6.215e−06
9								4.915e−06
. . .								. . .
33								1.649e−09
34								1.193e−09
35								8.641e−10

Fig. 1 Solution plots for adaptive penalty method Algorithms 4.1 and 4.2, clockwise from upper left: characteristic function $\chi_{\Omega_\star^0}$ optimal control $z^\star$, state $u^\star$, multiplier $\lambda^\star$, adjoint $p^\star$, and multiplier $\xi^\star$. Notice the nontrivial biactive set for the variational inequality and the upper and lower active sets for the control constraints

$$J(z_0, u_0) = J_\epsilon(z_0, u_0, \xi_0) \geq$$

$$J_\epsilon(z_k, u_k, \xi_k) = F(u_k) + G(z_k) + \frac{1}{\epsilon}\langle u_k, \xi_k \rangle \geq G(z_k) + \frac{1}{\epsilon}\langle u_k, \xi_k \rangle. \tag{26}$$

Hence, $J(z_0, u_0) - G(z_k) \geq \frac{1}{\epsilon}\langle u_k, \xi_k \rangle \geq 0$. Since Z_{ad} is bounded, there exists a weakly convergent subsequence $\{z_{k_l}\}$ with $z_{k_l} \rightharpoonup z^\star \in Z_{\mathrm{ad}}$. Then, by the weak lower-semicontinuity of G, we have

$$0 \leq \liminf_l \langle u_{k_l}, \xi_{k_l} \rangle \leq \limsup_l \langle u_{k_l}, \xi_{k_l} \rangle \leq J(z_0, u_0) - G(z^\star).$$

Thus, testing (25c) with u_{k_l}, we first obtain the boundedness of $\{u_{k_l}\}$ in V and then $\{\xi_{k_l}\}$ in V^*. Given A is symmetric, we use an argument as in (19) to prove that (along a further subsequence $\{k_{l_m}\}$) $\liminf_m \langle u_{k_{l_m}}, \xi_{k_{l_m}} \rangle \geq \langle u^\star, \xi^\star \rangle$, where $u^\star$ and $\xi^\star$ are the weak limit points. Since K and K^+ are weakly closed, $u^\star \in K$ and $\xi^\star \in K^+$. This suffices to prove that (25) has a solution. A proof for consistency of the approximation, i.e., that global optimizers $\{(z_\epsilon, u_\epsilon, \xi_\epsilon)\}$ converge as $\epsilon \downarrow 0$ (at least along a subsequence) to a global optimizer $\{(z^\star, u^\star, \xi^\star)\}$ of (6) can be derived analogously using the boundedness of $\{z_\epsilon\} \subset Z_{\mathrm{ad}}$ and subsequently, the inequality $\epsilon(J(z_0, u_0) - G(z_\epsilon)) \geq \langle u_\epsilon, \xi_\epsilon \rangle \geq 0$. $\qquad\square$

As in Section 4.1, we restrict ourselves to the setting of Example 2 for the derivation of stationarity conditions. Here, the derivation of first-order optimality conditions requires the verification of a constraint qualification, in this case that

of Robinson-Zowe-Kurcyusz, see [88]. According to Propositions 2.5 and 2.6 in [43], we have the following ϵ-dependent first-order optimality conditions: For any solution $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ to (25) (under the data assumptions of Example 2), there exists a multiplier-tuple $(p_\epsilon, \vartheta_\epsilon, \tau_\epsilon)$ such that

$$\alpha(z_\epsilon, w - z_\epsilon) + (B^* p_\epsilon, w - z_\epsilon) \geq 0, \quad \forall w \in Z_{ad} \tag{27a}$$

$$A^* p_\epsilon + \frac{1}{\epsilon} \xi_\epsilon - \vartheta_\epsilon = u_d - u_\epsilon, \tag{27b}$$

$$A u_\epsilon - \xi_\epsilon = B z_\epsilon + f, \tag{27c}$$

$$u_\epsilon \in K_0, \ \xi_\epsilon \in K_0^+, \tag{27d}$$

$$\frac{1}{\epsilon} u_\epsilon - p_\epsilon - \tau_\epsilon = 0, \tag{27e}$$

$$\vartheta_\epsilon \in K_0^+, \ \langle \vartheta_\epsilon, u_\epsilon \rangle = 0, \ \tau_\epsilon \in K_0, \ \langle \xi_\epsilon, \tau_\epsilon \rangle = 0. \tag{27f}$$

In comparison, this system is much larger than (16) and contains additional information. Nevertheless, under certain boundedness assumptions, passing to the limit in ϵ yields, in fact, a weaker C-stationarity system, see [43, Thm. 2.9]:

$$\alpha(z^\star, w - z^\star) + (B^* p^\star, w - z^\star) \geq 0, \quad \forall w \in Z_{ad} \tag{28a}$$

$$A^* p^\star - \lambda^\star = u_d - u^\star, \tag{28b}$$

$$A u^\star - \xi^\star = B z^\star + f, \tag{28c}$$

$$u^\star \in K_0, \ \xi^\star \in K_0^+, \ \langle u^\star, \xi^\star \rangle = 0, \tag{28d}$$

$$\langle \lambda^\star, u^\star \rangle = 0, \ \langle p^\star, \xi^\star \rangle = 0, \ \langle \lambda^\star, p^\star \rangle \leq 0. \tag{28e}$$

To see that (28e) relates to (9h), (9i) suppose for the sake of argument that $\xi^\star, p^\star, \lambda^\star$ are merely vectors of length n and the conditions in (28e) are understood as the componentwise (Hadamard) products. Then, by complementarity, $\xi^\star \geq 0$ and $\xi_i^\star = 0$ if i is an inactive or biactive index and $\langle p^\star, \xi^\star \rangle = 0$ in turn implies that the strongly active components of $p^\star$ are zero (as in (9h)). The same applies to the inactive components of $\lambda^\star$ due to $\langle \lambda^\star, u^\star \rangle = 0$ (as in (9i)).

Though there is a significant gap, the derivation of (28) is related to the convergence of a function-space-based numerical method. Indeed, using known results for linear elliptic PDE-constrained optimization problems with control and state constraints, we have viable efficient algorithms that can guarantee convergence to a KKT point, which satisfies (27). By performing continuation on ϵ we can be assured to converge (along a subsequence) to a weak C-stationary point. Furthermore, Theorem 2.12 in [43] provides a more compelling argument.

Theorem 8 (Thm 2.12 [43]) *Suppose $(z_\epsilon, u_\epsilon, \xi_\epsilon, p_\epsilon, \vartheta_\epsilon, \tau_\epsilon)$ satisfies (27) and that $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ is feasible for (7). Then, $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ is strongly stationary in the sense of Mignot and Puel, i.e., conditions (9h)–(10b) are replaced by*

$$p = 0, \; q.e. \; x \in \Omega^{0+} \tag{29a}$$

$$\langle \lambda, \phi \rangle = 0, \forall \phi \in H_0^1(\Omega) : \phi = 0, \; q.e. \; \Omega^0 \tag{29b}$$

$$p \leq 0, \; q.e. \; x \in \Omega^{00} \tag{29c}$$

$$\langle \lambda, \phi \rangle \geq 0, \forall \phi \in H_0^1(\Omega) : \phi \geq 0, \; q.e. \; \Omega^{00}, \phi = 0, \; q.e. \; \Omega \setminus (I \cup \Omega^{00}) \tag{29d}$$

Since we are using an ℓ^1 penalty, which often amounts to an exact penalty function, there is a good chance that a stationary point $(z_\epsilon, u_\epsilon, \xi_\epsilon)$ is feasible for (7) for sufficiently small ϵ.

Algorithm 4.3 ℓ^1 penalty method: outer loop

Input: $u_d, f \in L^2(\Omega), \beta \in (0, 1)$;
1: Choose $\epsilon^0 > 0, (z^0, u^0, \xi^0, p^0, \vartheta^0, \tau^0) \in L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \times H_0^1(\Omega) \times H^{-1}(\Omega) \times H_0^1(\Omega)$ and set $k := 0$;
2: **repeat**
3:      Compute a stationary point $(z^{k+1}, u^{k+1}, \xi^{k+1}, p^{k+1}, \vartheta^{k+1}, \tau^{k+1})$ of (27) with $\epsilon = \epsilon^k$ using an iterative scheme with initial value $(z^k, u^k, \xi^k, p^k, \vartheta^k, \tau^k)$;
3:      Set $\epsilon^{k+1} = \beta \epsilon^k$
4: **until** some stopping rule is satisfied.

Again we might choose as a stopping criterion the residual of C-stationarity, taking $\lambda := \epsilon^{-1}\xi - \vartheta$ and substituting $\tau = \epsilon^{-1}u - p$. Since the theory only guarantees $\vartheta \in H^{-1}(\Omega)$, one will need to treat the discrete quantities carefully, cf. [43] for more details. In addition, the solution of the subproblems (25) does not reduce to the solution of a system similar to (20). Therefore, the solution of the subproblems can be more difficult than in the adaptive penalty framework. Nevertheless, the theoretical result in Theorem 8 indicates the potential of this algorithm to generate a better stationary point.

5 Non-Smooth Numerical Methods

In this section, we present several methods that do not require a smoothing or penalization of the original MPEC. We present a new approximate projected subgradient method alongside a direct solver for the C-stationarity system presented in [36] and a recent method from [46] that may serve as a globalization of the direct solver. In all of these methods, we need to solve the variational inequality (1). This can be done using the semismooth Newton methods as in [38, 41] or special monotone multigrid methods as in [55].

5.1 An Approximate Projected Subgradient Method

The subgradient method is perhaps the simplest non-smooth optimization algorithm. Suppose X is a real separable Hilbert space. Given a proper, convex, lower-semicontinuous, and subdifferentiable functional $f : X \to \mathbb{R}$, $x^0 \in X$, $g^0 \in \partial f(x^0)$, and a sequence $\{v_k\}$ with $v_k > 0$, compute a sequence of iterates $\{x^k\}$ according to the rule

$$x^{k+1} := x^k - v_k g^k, \quad g^{k+1} \in \partial f(x^k).$$

Here, and unless otherwise noted below, we would need to apply the Riesz map to g^k before using it in this iteration.

Similarly, given a non-empty, closed, and convex subset $C \subset X$ the projected-gradient method replaces the previous rule by

$$x^{k+1} := \text{Proj}_C (x^k - v_k g^k), \quad g^{k+1} \in \partial f(x^{k+1}). \tag{30}$$

The subgradient method was invented by N.Z. Shor in the 1970s, see [76], and although it does not guarantee descent of the objective functional and can be quite slow to converge, it still finds a wide array of applications due to its simplicity and ability to be combined with distributed algorithm techniques, as in, e.g., [78].

Consider the reduced elliptic MPEC in (5):

$$\min \; \mathscr{J}(z) := F(S(Bz)) + G(z) \text{ over } z \in Z_{\text{ad}}.$$

Since S is nonlinear, the reduced objective functional is typically non-convex. Therefore, we cannot directly apply the projected subgradient method. However, there exist a number of generalized subdifferentials for non-convex functions. In our case, we will initially make use of the limiting subdifferential (also known as the Mordukhovich subdifferential) for the reduced objective. We will restrict ourselves to the framework of Example 2. First, we recall several definitions from variational analysis, see [66].

Definition 6 (Normal Cones to Arbitrary Sets) Let X be a Hilbert space and $C \subset X$. Then, the multifunction $\widehat{\mathscr{N}_C} : X \rightrightarrows X^*$ defined by

$$\widehat{\mathscr{N}_C}(x) := \left\{ x^* \in X^* \; \middle| \; \langle x^*, x'-x \rangle_X \leq o(\|x'-x\|_X), \; \forall x' \xrightarrow{X} x, x' \in C \right\}, \quad x \in C, \tag{31}$$

and $\widehat{\mathscr{N}_C}(x) := \emptyset$ for $x \notin C$ is called the regular (Fréchet) normal cone to C. The multifunction $\mathscr{N}_C : X \rightrightarrows X^*$ defined by

$$\mathscr{N}_C(x) := \left\{ x^* \in X^* \; \middle| \; \exists x_k \xrightarrow{X} x, \; \exists x_k^* \xrightarrow{X^*} x^* : x_k^* \in \widehat{\mathscr{N}_C}(x_k), \; \forall k \in \mathbb{N} \right\} \tag{32}$$

is called the limiting (Mordukhovich) normal cone to C.

Although $\widehat{\mathcal{N}_C}$ is convex, it fails to admit a satisfactory calculus needed for most non-smooth, non-convex problems. In contrast, the limiting normal cone enjoys a robust calculus. Note that for closed convex sets C, both cones agree, and in general $\widehat{\mathcal{N}_C}(x) \subsetneqq \mathcal{N}_C(x)$. We will use the limiting normal cone to define a generalized subgradient, needed in part for our proposed numerical method.

Definition 7 (Limiting Subdifferential) Let X be a Hilbert space, $\phi : X \to \overline{\overline{\mathbb{R}}}$, and $x \in X$ such that $|\phi(x)| < +\infty$. The set

$$\partial\phi(x) := \left\{x^* \in X^* \mid (x^*, -1) \in \mathcal{N}_{\text{epi}\,\phi}(x, \phi(x))\right\} \tag{33}$$

is called the limiting (Mordukhovich) subdifferential. If $|\phi(x)| = \infty$, we set $\partial\phi(x) = \emptyset$.

Therefore, if we know the limiting subdifferential $\partial\mathcal{J}(z)$ in (5), then we could design a projected subgradient iteration along the lines of (30):

$$z^{k+1} := \text{Proj}_{Z_{\text{ad}}}(z^k - \nu_k g^k), \quad g^{k+1} \in \partial\mathcal{J}(z^{k+1}).$$

If $\mathcal{J}$ were smooth, then $\partial\mathcal{J}$ is just the gradient of the reduced objective functional, which we usually calculate in PDE-constrained optimization by solving an adjoint equation. However here, $\mathcal{J}$ is non-smooth and non-convex. In order to obtain a generalized adjoint state for the reduced objective functional we require the so-called "coderivatives."

Definition 8 (Coderivatives) Let X be a Hilbert space, $\Phi : X \rightrightarrows X^*$, and $y \in \Phi(x)$, i.e., $(x, y) \in \text{Graph}\,\Phi$. The regular (Fréchet) coderivative of Φ at (x, y) is the multifunction $\widehat{D}^*\Phi(x, y) \colon Y^* \rightrightarrows X^*$ defined by

$$h^* \in \widehat{D}^*\Phi(x, y)(d^*) \Longleftrightarrow (h^*, -d^*) \in \widehat{\mathcal{N}_{\text{Graph}\,\Phi}}(x, y). \tag{34}$$

The limiting (Mordukhovich) coderivative $D^*\Phi(x, y)$ of Φ at $(x, y) \in \text{Graph}\,\Phi$ is similarly defined by

$$h^* \in D^*\Phi(x, y)(d^*) \Longleftrightarrow (h^*, -d^*) \in \mathcal{N}_{\text{Graph}\,\Phi}(x, y). \tag{35}$$

For example, if $\Phi = S_\epsilon$ from Section 4.1, then the coderivatives coincide and we have:

$$\widehat{D}^*S_\epsilon(w, u)(w^*) = D^*S_\epsilon(w, u)(w^*) = S_\epsilon'(w)^*w^*, \quad w^* \in X^*,$$

i.e., $\widehat{D}^*S_\epsilon(w, u)(w^*)$ yields the usual adjoint state p obtained by solving the associated linear elliptic PDE with $-w^*$ on the right-hand side.

For the tracking-type objective in Example 2, it was argued in [44, Prop. 1] that

$$\partial\mathcal{J}(z) \subset \alpha z + B^*D^*S(Bz, u)(u - u_d). \tag{36}$$

Here, it follows from [66, Thm 4.44] that $p \in D^*S(Bz, u)(u - u_d)$ is a solution to the generalized adjoint equation:

$$A^*p + D^*\mathcal{N}_{K_0}(u, \xi)(p) \ni u_d - u, \tag{37}$$

where $\xi = Bz + f - Au \in \mathcal{N}_{K_0}(u)$. Therefore, assuming that (37) were solvable, we could fashion our projected subgradient method as in Algorithm 5.4.

Algorithm 5.4 Limiting projected subgradient algorithm

Input: $\{\nu_k\}$ with $\nu_k > 0$; $z^0 \in Z$; $u^0 = S(Bz^0)$; $\xi^0 = Bz^0 + f - Au^0$; Find a solution p^0 to

$$A^*p + D^*\mathcal{N}_{K_0}(u^0, \xi^0)(p) \ni u_d - u^0.$$

1: **for** $k = 0, 1, \ldots$ **do**
2:    Set $z^{k+1} := \mathrm{Proj}_{Z_{\mathrm{ad}}}(z^k - \nu_k g^k)$ with $g^k = \alpha z^k + B^*p^k$.
3:    Set $u^{k+1} := S(Bz^{k+1})$, $\xi^{k+1} := Bz^{k+1} + f - Au^{k+1}$.
4:    Find p^{k+1} a solution to

$$A^*p + D^*\mathcal{N}_{K_0}(u^{k+1}, \xi^{k+1})(p) \ni u_d - u^{k+1}.$$

5: **end for**

Since we have no efficient means of handling $D^*\mathcal{N}_{K_0}(u^{k+1}, \xi^{k+1})(p)$, Algorithm 5.4 is impractical, especially when we consider that subgradient methods potentially require many iterations even for favorable convex problems. We therefore propose an alternative in Algorithm 5.5. The formal derivation for the approximate generalized adjoint state is based on simple geometric observations for a related finite-dimensional setting.

Consider that the (convex) normal cone $\mathcal{N}_{K_0}$ is generated by the (convex) subdifferential of the indicator functional i_{K_0}. Since the functionals ϕ^ϵ in (13) converge in a variational sense to i_{K_0}, they provide a viable candidate for approximating elements of $\mathcal{N}_{K_0}$. Moreover, for any $u \in V$, $\nabla\phi^\epsilon(u) = -\epsilon^{-1}(-u)_+$. Comparing as $\epsilon \downarrow 0$, it appears (at least in finite dimensions) that Graph $\nabla\phi^\epsilon \to$ Graph $-\mathcal{N}_{K_0}$, see Figure 2. This behavior transfers to $\widehat{\mathcal{N}}_{\mathrm{Graph}-\mathcal{N}_{K_0}}(u, \xi)$ and $\widehat{\mathcal{N}}_{\mathrm{Graph}\,\nabla\phi^\epsilon}(u, \xi)$, cf. Figure 2 with $\Theta := \mathrm{Graph}-\mathcal{N}_{\mathbb{R}_+}$ and $\Lambda := \mathrm{Graph}\,\nabla\phi^{0.1}$.

Using the information in Figure 2, we can first calculate the limiting normal cones $\mathcal{N}_\Theta(u, \xi)$, from which we obtain: $D^*\mathcal{N}_\Theta(1, 0)(p) = \{0\}$, for all p, and

$$D^*\mathcal{N}_\Theta(0, 1)(p) = \mathbb{R}, \text{ if } p = 0,$$

otherwise $D^*\mathcal{N}_\Theta(0, 1)(p) = \emptyset$. The most interesting case is:

$$D^*\mathcal{N}_\Theta(0, 0)(p) = \begin{cases} \{0\}, & \text{for all } p < 0, \\ \mathbb{R}, & \text{if } p = 0 \\ \mathbb{R}_-, & \text{else.} \end{cases}$$
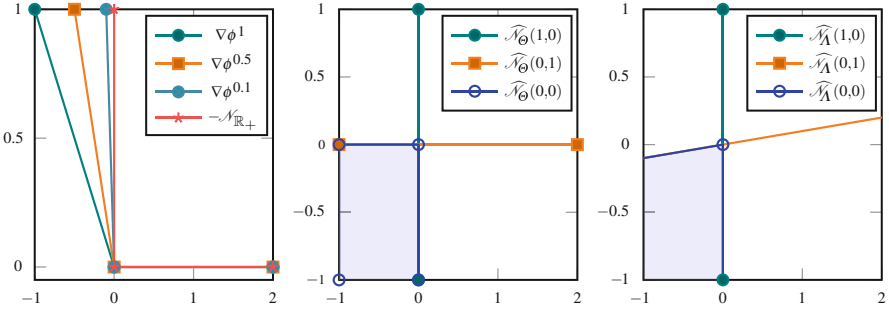
Fig. 2 From left to right: Graph $\nabla\phi^\epsilon$ ($\epsilon = 1, 0.5, 0.1$) versus Graph $\mathscr{N}_{\mathbb{R}_+}$; $\widehat{\mathscr{N}}_{\text{Graph}-\mathscr{N}_{\mathbb{R}_+}}(u,\xi)$ for $(u,\xi) = (1,0), (0,1), (0,0)$; $\widehat{\mathscr{N}}_{\text{Graph}\,\nabla\phi^{0.1}}(u,\xi)$ for $(u,\xi) = (1,0), (0,1), (0,0)$

Similarly, we have for all p

$$D^*\mathscr{N}_\Lambda(1,0)(p) = \{0\}, \quad D^*\mathscr{N}_\Lambda(0,1)(p) = -\epsilon^{-1}p;$$

and

$$D^*\mathscr{N}_\Lambda(0,0)(p) = \begin{cases} \{0\} \cup \{-\epsilon^{-1}p\}, & \text{for all } p < 0, \\ \{0\}, & \text{if } p = 0, \\ [-\epsilon^{-1}p, 0] & \text{else.} \end{cases}$$

Though certainly more tractable numerically, $D^*\mathscr{N}_\Lambda$ is still a set-valued mapping with non-convex images. We therefore suggest the following single-valued mapping:

$$\widetilde{D}^*\mathscr{N}_\Lambda(0,0)(p) := \left\{ q : q = -\epsilon^{-1}\chi_{\{u=0\}}p \right\}$$

where $\chi_{\{u=0\}}$ is the characteristic function for the active set. This mapping coincides with the limiting coderivative $D^*\mathscr{N}_\Theta$ on the inactive set and strongly active set, whenever $D^*\mathscr{N}_\Theta$ is non-empty. On the biactive set, it is either contained in $D^*\mathscr{N}_\Theta$ or approaches it for $\epsilon \downarrow 0$, cf. Figure 3.

By extrapolating these ideas from this simple one-dimensional geometric study to the infinite-dimensional setting, we arrive at our proposed algorithm.
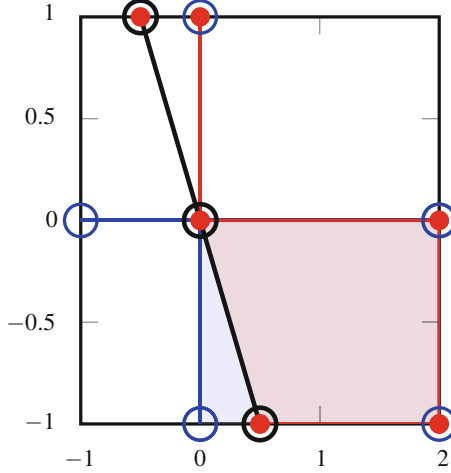
Fig. 3 Blue: $D^* \mathcal{N}_\Theta(0,0)(p)$, Red: $D^* \mathcal{N}_\Lambda(0,0)(p)$, Black: $\widetilde{D}^* \mathcal{N}_\Lambda(0,0)(p)$; $\epsilon = 0.5$

Algorithm 5.5 Approximate projected subgradient algorithm

Input: $\{\epsilon_k\}$ such that $\epsilon_k \downarrow 0$; $\{\nu_k\}$ such that $\nu_k > 0$.
Input: $z^0 \in Z$; $u^0 = S(Bz^0)$; $\xi^0 = Bz^0 + f - Au^0$; Solve for p^0:

$$A^* p + \frac{1}{\epsilon_0} \chi_{\{u^0 \leq 0\}} p = u_d - u^0.$$

1: **for** $k = 0, 1, \ldots$ **do**
2:     Set $z^{k+1} := \text{Proj}_{Z_{\text{ad}}} (z^k - \nu_k g^k)$ with $g^k = \alpha z^k - B^* p^k$.
3:     Set $u^{k+1} := S(Bz^{k+1})$, $\xi^{k+1} := Bz^{k+1} + f - Au^{k+1}$
4:     Solve for p^{k+1}:

$$A^* p + \frac{1}{\epsilon_{k+1}} \chi_{\{u^{k+1} \leq 0\}} p = u_d - u^{k+1}.$$

5: **end for**

Remark 4 In fact, $-\chi_{\{u \leq 0\}} p$ is nothing more than the adjoint of the Newton derivative for the non-smooth Nemytskii operator $(-\cdot)_+ : H_0^1(\Omega) \to L^2(\Omega)$.

Using the analytical techniques described throughout the text, we have the next result.

Theorem 9 *Suppose that Z_{ad} is bounded. Then, any sequence $\{(z^k, u^k, \xi^k, p^k, \lambda^k)\} \subset Z \times V \times V^* \times V \times V^*$ generated by Algorithm 5.5 is bounded. Here,*

$$\lambda_k := -\frac{1}{\epsilon_k} \chi_{\{u^k \leq 0\}} p^k.$$

Moreover, any weak accumulation point of $(z^\star, u^\star, \xi^\star, p^\star, \lambda^\star)$ will be feasible for (7) and we have

$$\lim_{k\to\infty} \int_{\{u^k=0\}} |p^k|^2 \mathrm{d}x = 0, \quad \langle \lambda^\star, p^\star \rangle \le 0, \quad \langle \lambda^\star, u^\star \rangle = 0. \tag{38}$$

Proof Since Z_{ad} is bounded, $\{z^k\}$ is bounded in Z. It immediately follows that $\{u^k\}$ is bounded, since

$$c\|u^k\|_V^2 \le \langle Au^k, u^k \rangle - \langle \xi^k, u^k \rangle = \langle Bz^k + f, u^k \rangle \Rightarrow c\|u^k\|_V \le \|Bz^k + f\|_{V^*}$$

and $\xi^k = Bz^k + f - Au^k$. Similarly, we obtain the boundedness of $\{p^k\}$ in V and $\{\lambda^k\}$ in V^*:

$$c\|p^k\|_V^2 \le \langle A^* p^k, p^k \rangle + \langle \lambda^k, p^k \rangle = \langle u_d - u^k, p^k \rangle \Rightarrow c\|p^k\|_V \le \|u_d - u^k\|_{V^*}$$

and $\lambda^k = u_d - u^k - A^* p^k$. Furthermore, since

$$0 \le \epsilon_k \langle A^* p^k, p^k \rangle + \int_{\{u^k=0\}} |p^k|^2 \mathrm{d}x = \epsilon_k \langle u_d - u^k, p^k \rangle$$

the limit condition in (38) holds. Moreover, will can again show that $\langle \lambda^\star, p^\star \rangle \le 0$ using the same argument as in (19) and by definition $\langle \lambda^k, u^k \rangle = 0$. Since B is compact, $u^k \to u^\star$ in V (along a subsequence). This yields $\langle \lambda^\star, u^\star \rangle = 0$.

The purpose of Theorem 9 is to show that Algorithm 5.5 produces a sequence with a weak accumulation point that satisfies a kind of limiting C-stationarity system. However, the lag in indices prevents us from closing the argument by proving that $\{z^k\}$ fulfils (9) (regardless of whether we choose fixed, bounded, or diminishing step sizes). Nevertheless, we will see later in the bundle-free approach that a variant of our approximate adjoint equation can under certain circumstances yield descent directions for the reduced objective.

We now demonstrate the performance of the algorithm on an example with a nontrivial biactive set: Example 3. In order to assure feasibility at every step, we solve the variational inequality with the primal-dual active set (PDAS)/semismooth Newton method from [38]. Note that although the solver for the variational inequality is mesh dependent, the majority of the linear solves are done within the first four iterations, see Table 2. For a graph of the behavior of the residuals as well as plots of the solution, see Figure 4.

We again use a uniform grid with 512^2 grid points and start the algorithm at $(z^0, u^0, \xi^0) = (0, 0, 0)$. We choose the a priori step sizes $\nu_k := (k)^{-1/2}$ and update ϵ_k according to $\epsilon_k := 10^{-4}/2^k$. The inner PDAS solver stops once the residual of the non-smooth system of equations reaches a tolerance of 10^{-10}. Though the theory does not provide a stopping criterion, we check the residual of strong stationarity, which reaches $\mathcal{O}(10^{-9})$ after 30 iterations. The residual is calculated using a discrete approximation of the following quantity:

Table 2 Outer loop k vs. inner loop iterations "iter" for PDAS with tol = 10^{-10} used in Algorithm 5.5

k	1	2	3	4	5	$\cdots$	31	32
iter	84	9	2	1	1	$\cdots$	1	1
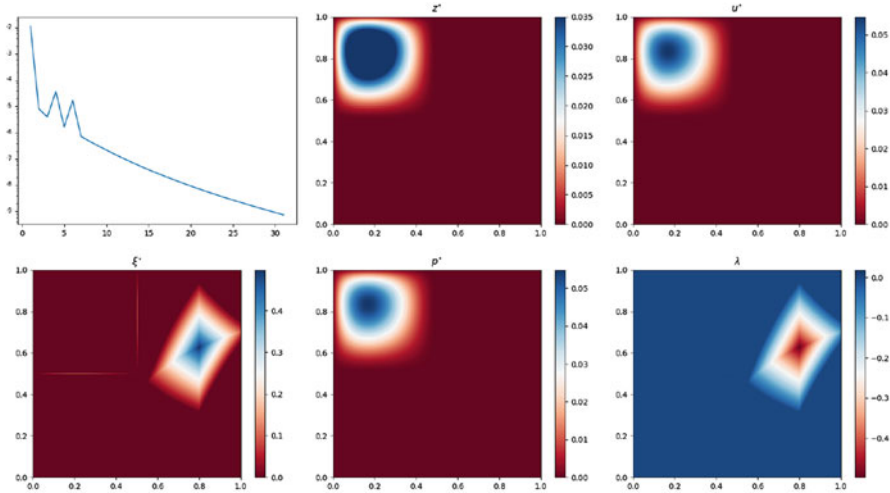
Fig. 4 Clockwise from upper left: residual of strong stationarity for Algorithm 5.5, optimal control $z^\star$, state $u^\star$, multiplier $\lambda^\star$, adjoint $p^\star$, and multiplier $\xi^\star$

$$\text{res}^k := \|z^k - \text{Proj}_{Z_{\text{ad}}}(z^k - g^k)\|_{L^2} + \|Au^k - \xi^k - z^k - f\|_{L^2} + \|\max(0, -u^k)\|_{L^2} +$$

$$\|\max(0, -\xi^k)\|_{L^2} + |(\xi^k, u^k)_{L^2}| + \|Ap^k - \lambda^k + u^k - u_d\|_{L^2} + |\max(0, \langle \lambda^k, p^k \rangle)| +$$

$$\|\chi_{\Omega_k^{0+}} p^k\|_{L^2} + \|\chi_{\mathscr{I}_k} \lambda^k\|_{L^2} + \|\chi_{\Omega_k^{00}} \max(0, p^k)\|_{L^2} + \|\chi_{\Omega_k^{00}} \max(0, -\lambda^k)\|_{L^2}$$

$$(39)$$

Despite lacking a convergence theory, the algorithm performs very well on this large-scale nontrivial problem. Indeed, counting all the free variables (z, u, ξ, etc.) there are over 10^6 degrees of freedom. Moreover, the presence of biactivity means that the example considered is genuinely non-smooth and non-convex. This is particularly encouraging, as the algorithm clearly outperforms the adaptive penalty method in terms of ease of implementation, size and structure of the systems of linear equations, and order of accuracy (almost reaching even strong stationarity). In particular, we note the stark contrast to the "sharpness" of the solution in comparison to the smooth method.

5.2 A Direct Solver for C-Stationarity Conditions

In this section, we adapt a method from [36] to the canonical MPEC (7). We work in the setting of Example 2 and assume that

$$Z_{\text{ad}} := \left\{ v \in L^2(\Omega) \mid a \le v \le b, \text{ a.e. } x \in \Omega \right\}$$

with $a, b \in L^2(\Omega)$, $a < b$. We first state the algorithm. Then, we motivate the steps of the algorithm and discuss its convergence properties.

The formal derivation of Algorithm 5.6, which we describe in detail below, follows several basic steps: Fix a control and estimate the active and inactive sets; ignoring biactivity, use (28e) to approximate (28b)–(28e) as a system of equations; solve the reduced system (including (28a)) to obtain an update; test the residual of C-stationarity, if necessary, return to 1.

Algorithm 5.6 An active-set equality-constrained newton solver w. Feasibility restoration

---

Input: $a, b, u_d, f \in L^2(\Omega)$, $\alpha > 0$, $z^0 \in L^2(\Omega)$, $p^0 \in H_0^1(\Omega)$, $k := 0$;

1: **repeat**
2:     Compute (u^k, ξ^k) by solving

$$Au - \xi = Bz^k + f, \quad \xi = (\xi - cu)_+, \quad c > 0,$$

and set

$$\Omega_k^0 := \left\{ x \in \Omega \mid u^k(x) = 0 \right\},$$

$$\Omega_k^+ := \left\{ x \in \Omega \mid u^k(x) > 0 \right\};$$

3:     Compute

$$\Omega_k^a := \left\{ x \in \Omega \mid a(x) + \alpha^{-1}(B^* p^k)(x) > 0 \right\},$$

$$\Omega_k^b := \left\{ x \in \Omega \mid -\alpha^{-1}(B^* p^k)(x) - b(x) > 0 \right\}.$$

and $\Omega_k^{ina} := \Omega \setminus (\Omega_k^a \cup \Omega_k^b)$.

4:     Compute $(\delta u^k, \delta p^k)$ by solving

$$A^* \delta p + \delta u = u_d - u^k - A^* p^k, \qquad\qquad \text{on } \Omega_k^+,$$

$$\delta p = 0, \qquad\qquad \text{on } \Omega_k^0,$$

$$A\delta u + \alpha^{-1} B \chi_{\Omega_k^{ina}} B^* \delta p = B\text{Proj}_{Z_{\text{ad}}}(-\alpha^{-1} B^* p^k) + f - Au^k, \qquad \text{on } \Omega_k^+,$$

$$\delta u = 0, \qquad\qquad \text{on } \Omega_k^0,$$

and set $u^{k+1} := u^k + \delta u^k$, $p^{k+1} := p^k + \delta p^k$,

$$z^{k+1} := z^k + \delta z = \text{Proj}_{Z_{\text{ad}}}(-\alpha^{-1} B^* p^k) - \alpha^{-1} \chi_{\Omega_k^{ina}} B^* \delta p^k$$

$k := k + 1$;
5: **until** some stopping criterion is satisfied

More specifically, suppose we are in the infinite-dimensional setting. Assuming that $\xi^\star$ is sufficiently regular, then Equations (28c), (28d) can be understood as the following system of smooth and non-smooth equations:

$$Au^\star - \xi^\star = Bz^\star + f, \tag{40a}$$

$$\xi^\star = (\xi^\star - cu^\star)_+, \tag{40b}$$

where $c > 0$ is some scaling constant. Since $\xi^\star \in L^2(\Omega)$, the Newton derivative for the plus function described at the end of Section 4.1 is not valid in this setting (since both the domain and range here must be taken as $L^2(\Omega)$). On a discrete level, this is not an issue. For some fixed $z^\star$, solving (40) gives a pair $(u^\star, \xi^\star)$ along with active and inactive sets:

$$\Omega^0(u^\star) := \left\{x \in \Omega \mid u^\star(x) = 0\right\}, \quad \Omega^+(u^\star) := \left\{x \in \Omega \mid u^\star(x) > 0\right\}.$$

As in our discussion of (28), if we treat the variables $u^\star$, $\lambda^\star$ as finite-dimensional vectors and the complementarity condition as the pointwise product of $u^\star$ and $\lambda^\star$, then the complementarity condition would indicate that $\lambda^\star = 0$ on the inactive set. Analogously, we take $p^\star = 0$ on the (entire) active set, thus ignoring biactivity. Consequently the remaining sign condition in (28) holds. Finally, using the projection formula (23) along with a semismooth Newton step, we can handle the variational inequality (28a). We recall the sets

$$\Omega^a := \left\{x \in \Omega \,\middle|\, a(x) + \alpha^{-1}(B^*p^\star)(x) > 0\right\},$$

$$\Omega^b. := \left\{x \in \Omega \,\middle|\, -\alpha^{-1}(B^*p^\star)(x) - b(x) > 0\right\}.$$

and $\Omega^{ina} := \Omega \setminus (\Omega^a \cup \Omega^b)$.

Now, supposing we want a new approximation $(u^\star, z^\star, p^\star)$ via $u^\star + \delta u$, $z^\star + \delta z$, and $p^\star + \delta p$, we consider the reduced system by eliminating the dual variables:

$$\delta z + \alpha^{-1}\chi_{\Omega^{ina}}B^*\delta p = \text{Proj}_{Z_{ad}}(-\alpha^{-1}B^*p^\star) - z^\star \qquad \text{on } \Omega, \tag{41a}$$

$$A^*\delta p + \delta u = u_d - u^\star - A^*p^\star, \qquad \text{on } \Omega^+(u^\star), \tag{41b}$$

$$\delta p = 0, \qquad \text{on } \Omega^0(u^\star), \tag{41c}$$

$$A\delta u - B\delta z = Bz^\star + f - Au^\star, \qquad \text{on } \Omega^+(u^\star), \tag{41d}$$

$$\delta u = 0, \qquad \text{on } \Omega^0(u^\star), \tag{41e}$$

If we replace δz in (41d), then we get the smaller system in $(\delta u, \delta p)$:

$$A^*\delta p + \delta u = u_d - u^\star - A^*p^\star, \qquad \text{on } \Omega^+(u^\star), \tag{42a}$$

$$\delta p = 0, \qquad \text{on } \Omega^0(u^\star), \tag{42b}$$

$$A\delta u + \alpha^{-1}B\chi_{\Omega^{ina}}B^*\delta p = B\text{Proj}_{Z_{ad}}(-\alpha^{-1}B^*p^\star) + f - Au^\star, \qquad \text{on } \Omega^+(u^\star), \tag{42c}$$

$$\delta u = 0, \qquad \text{on } \Omega^0(u^\star), \tag{42d}$$

Table 3 Residuals of strong stationarity for Algorithm 5.6 with stopping tolerance tol = 10^{-7}

k	1	2	3	4	5
res^k	8.7287e−2	1.6175e−3	1.1190e−5	3.8548e−7	2.2375e−8

which is remarkably similar to the semismooth Newton step in the adaptive penalty method. Unlike the smooth methods, however, a proof of convergence remains elusive. The main culprit here is clearly the sequence $\{\Omega_k^+\}$, which need not converge with respect to any notion of set convergence, e.g., Painlevé-Kuratowski or in the sense of characteristic functions.

Note also that this formally derived system is well-defined in function space provided that we have enough regularity. For example, provided that the sets $\Omega^+(u^\star)$, $\Omega^0(u^\star)$ are sufficiently regular, we could reduce the search for $(\delta u, \delta p) \in \mathscr{H} \times \mathscr{H}$ with $\mathscr{H} := H_0^1(\Omega^+(u^\star))$, solve the associated weak form of (42a), (42b), and extend the solutions by zero on $\Omega^0(u^\star)$.

An obvious stopping criterion for Algorithm 5.6 would be the residual of (28) up to some user-defined tolerance (as suggested in [36]). Moreover, we see that the computational effort is roughly that of the adaptive penalty method. In fact, there is much less nonlinearity here due to a lack of the penalty terms ϕ^ε. We also mention that the MPECs in [36] are much more challenging than (7), as the controls there arise inside the differential operator. Nevertheless, the algorithm seems to perform very well, even on examples with nontrivial biactive sets.

We demonstrate the performance of Algorithm 5.6 on Example 3. As expected, Algorithm 5.6 behaves like a second-order method, see Table 3. We once again used the PDAS/semismooth Newton method in [38] to restore feasibility at every step. Moreover, since the multiplier λ is eliminated from the algorithm, we artificially reintroduce it for the calculation of the residuals.

The solutions look identical to those plotted in Figure 5. We therefore only provide images of the biactive sets for $y^\star$ and upper and lower active sets for $z^\star$. We note that this algorithm also performs quite well, reaching a residual of strong stationarity on the order of $\mathcal{O}(10^{-8})$ within $k = 5$ iterations, though it never reaches $\mathcal{O}(10^{-9})$ in contrast to the approximating subgradient algorithm. In addition, the effort to solve each step is higher, as seen in Table 4.

Our next non-smooth method seeks to overcome the theoretical deficiencies of Algorithm 5.5 and 5.6. In some sense, it takes a step towards bridging the gap between this active-set-based solver and the approximate projected subgradient method in the previous subsection.

5.3   The Bundle-Free Implicit Programming Method

We now present the bundle-free implicit programming approach from [46], which we extend for control constraints. In contrast to the active-set method in the previous subsection, this method is based off of B-stationarity conditions. We must therefore assume that S is directionally differentiable.
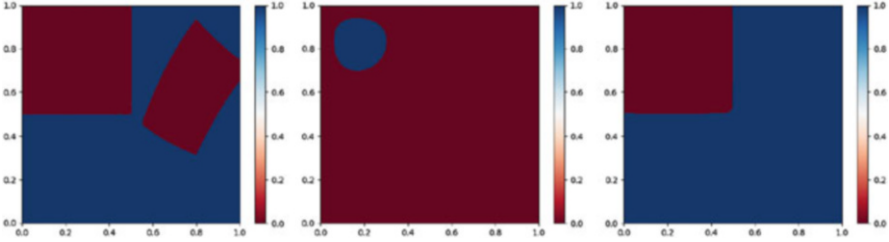
Fig. 5 Plot of solutions using Algorithm 5.6. Left to right: characteristic functions $\chi_{\Omega^{00}_\star}$ (set of all indices with $u_i^\star \leq 1e{-}8$ and $\xi_i^\star \leq 0$), $\chi_{\Omega^b_\star}$ (set of all indices with $z_i^\star \leq b_i - 1e{-}6$), and $\chi_{\Omega^a_\star}$ (set of all indices with $z_i^\star \geq a_i + 1e{-}6$)

Table 4 Outer loop k vs. inner loop "iter" iterations for PDAS with tol = 10^{-10} used in Algorithm 5.6

k	1	2	3	4	5
iter	84	21	19	1	1

We first state the basic assumptions and the algorithm. Afterwards, we discuss the motivations for the steps and examine the convergence properties. Throughout this subsection, let $Z_{\mathrm{ad}} := \left\{ z \in L^2(\Omega) \mid a \leq z \leq b \right\}$ with $a, b \in L^2(\Omega)$ and $a < b$ and assume that the variational inequality is defined as in (7). We otherwise work in the abstract framework of (6), where $\mathscr{J}(z) := F(S(Bz)) + G(z)$ and F and G are continuously Fréchet differentiable. By assumption, $\xi \in H$.

For some constant $c_q > 0$, let $q(\cdot) := c_q \| \cdot \|^2/2$ and for $z \in Z_{\mathrm{ad}}$ define the local quadratic models:

$$\mathscr{M}(\cdot) := q(\cdot) + F'(z; S'(Bz; B\cdot)) + G'(z; \cdot),$$

$$\mathscr{M}_\epsilon(\cdot) := q(\cdot) + F'(z; d_\epsilon(B\cdot)) + G'(z; \cdot), \quad \epsilon > 0,$$

where $d_\epsilon(\cdot)$ is a smooth approximation of $S(Bz; \cdot)$ such that $d_\epsilon(0) = 0$. Here, we suggest letting $d_\epsilon(w) = d$, the solution to

$$Ad + \epsilon^{-1}\chi_{\Omega^{0+}}d - \chi_{\Omega^0}\nabla^2\phi^\epsilon(-d) = w,$$

where $w \in V^*$, $\epsilon > 0$, and $u = S(Bz)$. This is related to the true formula for $d = S'(Bz; w)$ given by

$$\text{Find } d \in \mathscr{K}(u, \xi) : \int_\Omega \nabla d \cdot \nabla[v - d]\mathrm{d}x \geq \langle w, v - d \rangle, \quad \text{for all } v \in \mathscr{K}(u, \xi), \tag{43}$$

where $\mathscr{K}(u, \xi) := \left\{ u \in H_0^1(\Omega) \mid d(x) \geq 0, \text{ q.e. } x \in \Omega^0, d(x) = 0. \text{ a.e. } x \in \Omega^{0+} \right\}$; see [63] (or [15, Chap. 6] for an English summary). As this formula makes use of quasi-continuity and potentially non-negligible sets of Lebesgue measure zero, it is unclear how to make direct use of the nonnegativity condition

in a numerical method. To circumvent this issue, a certain regularity condition is assumed throughout (as in [46]):

1. $\Omega^0 = \overline{\mathrm{int}(\Omega^0)}$,
2. If $m(\Omega^{00}) = 0$, then $\mathrm{cap}(\Omega^{00}) = 0$ or $S'(Bu; Bh) = d = 0$ q.e. $x \in \Omega^0$.
3. If $m(\Omega^{00}) > 0$, then $\exists \nu > 0, \exists \gamma' > 0 : \forall \gamma \geq \gamma', d_{\gamma,\varepsilon} \geq 0, \ a.e. \ \mathscr{A}_\nu \setminus \Omega^0$ where $\mathscr{A}_\nu := \left\{ x \in \Omega \,\middle|\, \mathrm{dist}(x, \Omega^0) < \nu \right\}$.

In light of this assumption, we speak of "negligible" biactive sets whenever $m(\Omega^{00}) = 0$. Note that S is in fact Gâteaux differentiable, whenever this condition holds. This once again highlights a fundamental difference between the optimal control of PDEs and variational inequalities, even for the simplest variational inequality of interest.

We also suggest the smooth approximation of the directional derivative in order to guarantee that $p_\epsilon = d'_\epsilon(0)^* w$ solves

$$A^* p + \epsilon^{-1} \chi_{\Omega^{0+}} p = w$$

and that $\{p_\epsilon\}_{\epsilon > 0}$ is uniformly bounded in V. Notice the similarity to the approximating limiting subgradient suggested earlier.

Algorithm 5.7 A bundle-free implicit programming approach

Input: $\epsilon^0 > 0, \beta, \rho, \sigma \in (0,1), s > 0, (z^0, u^0, \xi^0) \in Z \times V \times H, k := 0$;
1: **repeat**
2:     Set $\Omega_k^0 := \left\{ x \in \Omega \,\middle|\, u^k(x) = 0 \right\}$ and $\Omega_k^{00} := \left\{ x \in \Omega_k^0 \,\middle|\, \xi^k(x) = 0 \right\}$;
3:     **if** Ω_k^{00} is negligible **then**
4:         Compute δz^k by

$$\delta z = \mathrm{Proj}_{\mathscr{T}_{Z_{\mathrm{ad}}}(z^k)} \left[-c_q^{-1} \nabla \mathscr{M}^k(0) \right]; \tag{44}$$

5:         Compute step size τ^k using a line search;
6:         Set $z^{k+1} = z^k(\tau^k), u^{k+1} := S(Bz^{k+1}), \xi^{k+1} := Au^{k+1} - Bz^{k+1} - f, k := k+1$;
7:     **else**
8:         Compute δz^k by

$$\delta z = \mathrm{Proj}_{\mathscr{T}_{Z_{\mathrm{ad}}}(z^k)} \left[-c_q^{-1} \nabla \mathscr{M}_{\epsilon^k}^k(0) \right]; \tag{45}$$

9:         **while** descent criterion fails **do**
10:             Choose $\epsilon^k \in (0, \rho \epsilon^k)$, update model $\mathscr{M}_{\epsilon^k}^k$ go to 8:;
11:         **end while**
12:         Choose $\epsilon^{k+1} \in (0, \rho \epsilon^k)$;
13:         Compute τ^k as in 5:, update (z^k, u^k, ξ^k) as in 6:;
14:     **end if**
15:     (Robustification Step), $k := k+1$;
16: **until** some stopping criterion is satisfied

In Algorithm 5.7, $\mathscr{T}_{Z_{\mathrm{ad}}}(z)$ is the tangent cone to Z_{ad} at $z \in Z_{\mathrm{ad}}$, which is defined by

$$\mathscr{T}_{Z_{\mathrm{ad}}}(z) := -[\mathscr{N}_{\mathrm{ad}}(z)]^+ = \overline{\mathbb{R}_+(Z_{\mathrm{ad}} - z)}.$$

We use the generalized Armijo line search (cf. [13].): Set $\tau^k := \beta^{m_k} s$ where m_k is the first nonnegative integer such that

$$\mathcal{J}(z^k) - \mathcal{J}(z^k(\beta^m s)) \geq \frac{\sigma}{\beta^m s} \|z^k - z^k(\beta^m s))\|_{L^2}^2, \tag{46}$$

where $\beta, \sigma \in (0, 1)$, $s > 0$ and, given a step δz, we set

$$z(\tau) := \mathrm{Proj}_{Z_{\mathrm{ad}}}(z + \tau \delta z), \ \ \tau > 0.$$

Since much of Algorithm 5.7 is derived from B-stationarity conditions, we restate them here:

$$\mathcal{J}'(z; w - z) = F'(S(Bz))S'(Bz; B(w - z)) + G'(z)(w - z) \geq 0, \quad \forall w \in Z_{\mathrm{ad}}. \tag{47}$$

From (47), it is clear that we can equivalently reformulate B-stationarity as:

$$\mathcal{J}'(z; \delta z) \geq 0, \ \ \forall \delta z \in \mathcal{T}_{Z_{\mathrm{ad}}}(z).$$

Moreover, since $\mathcal{J}'(z; \delta z)$ is positively homogeneous in δz, it follows that $\delta z = 0$ is a minimizer of $\mathcal{J}'(z; \delta z)$ over $\mathcal{T}_{Z_{\mathrm{ad}}}(z)$, whenever z is B-stationary. Finally, given that $\mathcal{T}_{Z_{\mathrm{ad}}}(z)$ is a non-empty, closed, and convex cone, we can add the coercive quadratic form q to the definition of B-stationarity without altering the characterization, cf. the general analysis in [46, Section 2]. Therefore, if z is B-stationary, then $0 \in Z$ solves the auxiliary problem

$$\min \left\{ \mathcal{M}(\delta z) := q(\delta z) + \mathcal{J}'(z; \delta z) \ \text{over} \ \delta z \in \mathcal{T}_{Z_{\mathrm{ad}}}(z) \right\}. \tag{48}$$

Now, if the biactive set is negligible, then S is in fact Gâteaux differentiable. In this case, (48) has a unique solution given by

$$(\delta z^* + c_q^{-1}(B^* S'(Bz)^* \nabla F'(S(Bz)) + \nabla G(z)), w - \delta z^*) \geq 0, \ \ \forall w \in \mathcal{T}_{Z_{\mathrm{ad}}}(z),$$

which, noting that $\nabla \mathcal{M}(0) = B^* S'(Bz)^* \nabla F'(S(Bz)) + \nabla G(z)$, is equivalent to (45). Furthermore, in this smooth setting, we can prove the following result.

Proposition 1 *Let $z \in Z_{\mathrm{ad}}$, $u = S(Bz)$, $\xi = Au - Bz - f$, and suppose that $\mathcal{J}$ is Gâteaux differentiable at z. If z is not B-stationary, then (46) stops in a finite number of steps.*

Proof Suppose δz is given by the projection formula (45). Since δz is the unique global optimum and $0 \in \mathcal{T}_{Z_{\mathrm{ad}}}(z)$, we have

$$q(\delta z) + \mathcal{J}'(z; \delta z) < 0 \Longrightarrow \mathcal{J}'(z; \delta z) \leq -\frac{c_q}{2} \|\delta z\|_{L^2}^2 \tag{49}$$

Moreover, for any $\tau > 0$, we have $z(\tau) := \mathrm{Proj}_{Z_{ad}}(z - \tau \delta z)$ and since $\mathrm{Proj}_{Z_{ad}}$ is non-expansive:

$$\|z - z(\tau)\|_{L^2} = \|\mathrm{Proj}_{Z_{ad}}(z + \tau \delta z) - \mathrm{Proj}_{Z_{ad}}(z)\|_{L^2} \leq \tau \|z\|_{L^2}. \tag{50}$$

Furthermore, since $z \in Z_{ad}$, $\delta z \in \mathscr{T}_{Z_{ad}}(z)$, and Z_{ad} is defined by simple pointwise bound constraints in $L^2(\Omega)$, we appeal to the proof of Lemma 6.34 [15], which shows that

$$\tau^{-1}(z(\tau) - z) \to \delta z \text{ as } \tau \downarrow 0. \tag{51}$$

Finally, suppose that (46) fails for all $\tau > 0$. Then,

$$\tau^{-1}(\mathscr{J}(z(\tau)) - \mathscr{J}(z)) > -\sigma \tau^{-2}\|z - z(\tau)\|_{L^2}^2, \quad \forall \tau > 0.$$

On the right side of the inequality, we can estimate from below using (50):

$$-\sigma \tau^{-2}\|z - z(\tau)\|_{L^2}^2 \geq -\sigma \|\delta z\|_{L^2}^2$$

Now, letting $z(\tau) = z + \tau \tau^{-1}(z(\tau) - z) = z + \tau d_\tau$, where $d_\tau \to \delta z$ by (51), we have by (49):

$$\tau^{-1}(\mathscr{J}(z(\tau)) - \mathscr{J}(z)) = \tau^{-1}(\mathscr{J}(z + \tau d_\tau) - \mathscr{J}(z)) \to \mathscr{J}'(z; \delta z) \leq -\frac{c_q}{2}\|\delta z\|_{L^2}^2.$$

But, then

$$-\frac{c_q}{2}\|\delta z\|_{L^2}^2 \geq \mathscr{J}'(z; \delta z) \geq -\sigma \|\delta z\|_{L^2}^2,$$

a contradiction, since $\sigma \in (0, c_q/2)$.

Proposition 1 provides a justification for steps 4:–6: in Algorithm 5.7. Note that the calculation of the gradient $\nabla \mathscr{M}^k(0)$ requires the solution of an adjoint equation. For example, in the setting of Example 2, $\nabla \mathscr{M}^k(0) = \alpha z^k - B^* p^k$, where p^k solves (here in strong form):

$$\begin{aligned} A^* p &= u_d - u^k \text{ on } \Omega_k^+, \\ p &= 0, \qquad \text{on } \Omega_k^0, \end{aligned} \tag{52}$$

Turning now to the case when the biactive set is non-negligible, we can easily adjust the proof of Proposition (1) for the non-smooth setting.

Corollary 2 *Let $z \in Z_{ad}$, $u = S(Bz)$, $\xi = Au - Bz - f$, and suppose that δz is minimizer for (48). If z is not B-stationary, then (46) stops in a finite number of steps.*

Nevertheless, whenever the biactive set is non-negligible, the directional derivative is nonlinear in δz. In particular, (48) is non-convex, which was a key assumption in the arguments. We therefore need an alternative procedure to calculate the step δz.

In Algorithm 5.7, we suggested the step $\delta z = \text{Proj}_{\mathscr{T}_{Z_{ad}}(z^k)}\left[-c_q^{-1}\nabla\mathscr{M}_{\epsilon^k}^k(0)\right]$, which is related to the smoothed auxiliary problem using the ϵ-dependent model $\mathscr{M}_\epsilon$:

$$\min\left\{\mathscr{M}_\epsilon(\delta z) := q(\delta z) + F'(z; d_\epsilon(B\delta z)) + G'(z)\delta z \text{ over } \delta z \in \mathscr{T}_{Z_{ad}}(z)\right\}. \quad (53)$$

Note if the approximation was chosen so that $d_\epsilon \to S$ and 0 solves (53) for all sufficiently small $\epsilon > 0$ (or at least along some null sequence), then z must be B-stationary as 0 solves (48), as well. In fact, in the current setting, A is symmetric so we can use the approximation results in [7] (discussed above), see also [46], to argue that if $\delta z_\epsilon \to \delta z$ weakly in $L^2(\Omega)$, then

$$d_\epsilon(B\delta z_\epsilon) \to S'(Bz; B\delta z), \text{ as } \epsilon \downarrow 0,$$

provided that B is completely continuous, e.g., when B is the embedding of $L^2(\Omega)$ into $H^{-1}(\Omega)$.

On the other hand, if there exists some $\delta z \in \mathscr{T}_{Z_{ad}}(z)$ such that $(\nabla\mathscr{M}_\epsilon(0), \delta z) < 0$, then by continuity, there is some $\eta_\epsilon > 0$ such that $\mathscr{M}_\epsilon(\eta_\epsilon\delta z) < 0$. Since $\mathscr{T}_{Z_{ad}}(z)$ is a cone, $\eta_\epsilon\delta z \in \mathscr{T}_{Z_{ad}}(z)$ and 0 does not solve (53). If this persists as $\epsilon \downarrow 0$, then z cannot be B-stationary.

In light of this, consider our δz update. Let $w := d_\epsilon'(0)^*F'(z) + G'(z) = \nabla\mathscr{M}_\epsilon(0)$ and note that

$$\mathscr{T}_{Z_{ad}}(z) = \left\{\delta z \in L^2(\Omega) \,\middle|\, \delta z \geq 0, \text{ a.e. on } \Omega^a, \ \delta z \leq 0, \text{ a.e. on } \Omega^b\right\},$$

where $\Omega^a := \{x \in \Omega \mid z(x) = a(x)\}$, $\Omega^b := \{x \in \Omega \mid z(x) = b(x)\}$ and $\Omega^{ina} := \Omega \setminus (\Omega^a \cup \Omega^b)$. Then, using the basic properties of $\text{Proj}_{\mathscr{T}_{Z_{ad}}(z)}$, we have

$$(\nabla\mathscr{M}_\epsilon(0), \delta z) =$$

$$-c_q\left[\int_{\Omega^{ina}}|\delta z|^2 + \int_{\Omega^a \cap \{-c_q^{-1}w \geq 0\}}|\delta z|^2 + \int_{\Omega^b \cap \{-c_q^{-1}w \leq 0\}}|\delta z|^2\right] \leq 0.$$

Assuming that the latter term is nonzero and expanding $\mathscr{M}_\epsilon$ at zero in direction δz, we obtain:

$$\mathscr{M}_\epsilon(0 + \eta\delta z) = \eta\left(\frac{\eta c_q}{2}\|\delta z\|_{L^2}^2 + (\nabla\mathscr{M}_\epsilon(0), \delta z) + o(1)\right). \quad (54)$$

We may then choose a sufficiently small $\eta_\epsilon > 0$ such that

$$\frac{\eta_\epsilon c_q}{2}\|\delta z\|_{L^2}^2 + (\nabla \mathcal{M}_\epsilon(0), \delta z) + o(1) < 0.$$

Hence, $\mathcal{M}_\epsilon(\eta_\epsilon \delta z) \leq 0$ and by definition

$$F'(z; d_\epsilon(B(\eta_\epsilon \delta z))) + G'(z)(\eta_\epsilon \delta z) \leq -\eta_\epsilon^2 q(\delta z_\epsilon).$$

Therefore, if a uniform lower bound $\eta > 0$ with $\eta_\epsilon \geq \eta > 0$ exists, then we can prove that δz_ϵ is a descent direction for some sufficiently small $\epsilon > 0$.

Proposition 2 *Let* $z \in Z_{\mathrm{ad}}$, $u = S(Bz)$, *and* $\xi = Au - Bz - f$ *and suppose that* $\mathcal{J}$ *is only directionally differentiable at* z. *Furthermore, let* $\delta z_\epsilon = \mathrm{Proj}_{\mathscr{T}_{Z_{\mathrm{ad}}}(z)} \left[-c_q^{-1} \nabla \mathcal{M}_\epsilon(0) \right]$ *and assume that* $\limsup_{\epsilon \downarrow 0}(\nabla \mathcal{M}_\epsilon(0), \delta z_\epsilon) < 0$. *If there exists an* $\eta > 0$ *such that* $\mathcal{M}_\epsilon(\eta \delta z_\epsilon) \leq 0$ *for all sufficiently small* $\epsilon > 0$, *then there exists some* $\widehat{\epsilon} > 0$ *such that* $\delta z_{\widehat{\epsilon}}$ *is a descent direction for* $\mathcal{J}$ *at* z.

Proof By assumption, $F'(z; d_\epsilon(B(\eta \delta z_\epsilon)) + G'(z)(\eta \delta z_\epsilon) \leq -\eta^2 q(\delta z_\epsilon)$. Then,

$$\mathcal{J}'(z; \eta \delta z_\epsilon) = (\mathcal{J}'(z; \eta \delta z_\epsilon) - F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) - G'(z)(\eta \delta z_\epsilon)) +$$

$$F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) + G'(z)(\eta \delta z_\epsilon) \leq$$

$$(\mathcal{J}'(z; \eta \delta z_\epsilon) - F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) - G'(z)(\eta \delta z_\epsilon)) - \eta^2 q(\delta z_\epsilon) =$$

$$F'(z; S(Bz; B(\eta \delta z_\epsilon))) - F'(z; d_\epsilon(B(\eta \delta z_\epsilon))) - \eta^2 q(\delta z_\epsilon) =$$

$$\langle \nabla F(z), S(Bz; B(\eta \delta z_\epsilon)) - d_\epsilon(B(\eta \delta z_\epsilon)) \rangle - \eta^2 q(\delta z_\epsilon).$$

Now, since $\{p_\epsilon\}$ is bounded, we can show that $\{\delta z_\epsilon\}$ is bounded. Hence, there is a subsequence (denoted still by ϵ) such that $\delta z_\epsilon \rightharpoonup \delta z^*$. Then, for sufficiently small $\widehat{\epsilon} > 0$

$$\langle \nabla F(z), S(Bz; B(\eta \delta z_{\widehat{\epsilon}})) - d_{\widehat{\epsilon}}(B(\eta \delta z_{\widehat{\epsilon}})) \rangle \leq \eta^2 q(\delta z_{\widehat{\epsilon}})/2, \tag{55}$$

Hence, $\mathcal{J}'(z; \delta z_{\widehat{\epsilon}}) \leq -c_q \eta \|\delta z_{\widehat{\epsilon}}\|^2/4$, as was to be shown.

Since a direct verification of the hypotheses is potentially too expensive from a computational standpoint, we suggest a heuristic. Fix a lower bound $\underline{\eta} > 0$. If $(\nabla \mathcal{M}_\epsilon(0), \delta z) < -q(\delta z)$ and (55) (or an approximation as in [46, Remark 3.13]) holds with $\eta = 1$ (or $\eta = \underline{\eta}$), we use the current δz in the line search. Thus, the "descent condition" in 9: holds. If $-q(\delta z) \leq (\nabla \mathcal{M}_\epsilon(0), \delta z) < 0$, then we choose $\eta > 0$ such that $\eta q(\delta z) + (\nabla \mathcal{M}_\epsilon(0), \delta z) < 0$, set $\delta z := \eta \delta z$, update $\underline{\eta} := \min(\eta, \underline{\eta})$, and check (55) or an approximation. Here, the descent condition also holds, but the model $\mathcal{M}_\epsilon$ might be failing. If (55) fails in the latter, then we go to step 10:.

Finally, if $(\nabla \mathcal{M}_\epsilon(0), \delta z) = 0$, then we go to step 10: . In practice, one might also attempt to circumvent the verification of (55) by using a sufficiently small $\epsilon > 0$ and decreasing at every step.

This heuristic has its limitations, e.g., it cannot prohibit $\eta \downarrow 0$. Therefore, if it appears that this is the case, we break the while loop and go to a "robustification step" in 15:. This just means that if the quadratic model $\mathcal{M}_\epsilon$ appears to be ineffective, then we should calculate a new control z by using an alternative method that is guaranteed to converge and "restart" the algorithm. For instance, we could use the adaptive penalty method in Section 4.1 (for a fixed γ, increasing each time the robustification step is used).

Finally, we recall that in the classical projected-gradient approaches, the Lipschitz continuity of the gradient of the objective is essential for convergence proofs, cf. [13]. There, one can show that the line search will always stop provided that the step size is below a certain threshold which only depends on σ and the Lipschitz modulus L. This is, of course, not possible for the control of variational inequalities as the reduced objective is non-smooth (even when the biactive set is negligible we only have Gâteaux differentiability). Therefore, we must also monitor the behavior of the accepted step sizes at each iteration. If, as with η, the step sizes τ^k appear to be rapidly decreasing with each iteration, then we also make use of a robustification step. For a full convergence proof in the case of the canonical MPEC (excluding control constraints), we refer the interested reader to [46].

We conclude by demonstrating the performance of the bundle-free method on two examples, Example 3 and the following example (adapted from [39, Ex. 6.1] by adding control constraints).

Example 4 Here, we let $a \equiv 0$, $b \equiv 0.8$ and set $\alpha = 1$. In addition, we define f and u_d as follows:

$$z_1(\mathbf{x}_1, \mathbf{x}_2) = -4096\mathbf{x}_1^6 + 6144\mathbf{x}_2^5 - 3072\mathbf{x}_1^4 + 512\mathbf{x}_2^3,$$

$$z_2(\mathbf{x}_1, \mathbf{x}_2) = -244.140625\mathbf{x}_1^6 + 585.9375\mathbf{x}_2^5 - 468.75\mathbf{x}_2^4 + 125\mathbf{x}_2^3,$$

$$y^*(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} z_1(\mathbf{x}_1, \mathbf{x}_2)z_2(\mathbf{x}_1, \mathbf{x}_2) & (\mathbf{x}_1, \mathbf{x}_2) \in]0, 0.5[\times]0, 0.8[, \\ 0 & \text{otherwise,} \end{cases}$$

$$u^* = y^*, \quad \xi^*(\mathbf{x}_1, \mathbf{x}_2) = 2\max(0, -|\mathbf{x}_1 - 0.8| - |\mathbf{x}_1\mathbf{x}_2 - 0.2| - 0.3 + 0.35),$$

$$f = -\Delta y^* - u^* - \xi^*, \quad u_d = y^* + \xi^* - \alpha\Delta u^*.$$

In order to compare to the other methods, we again use a uniform grid with 512^2 grid points and start the algorithm at $(z^0, u^0, \xi^0) = (0, 0, 0)$ for Example 3. In contrast, we use a random starting point when solving Example 4. We start the algorithm with $\epsilon_0 = 10^{-10}$ and subsequently set $\epsilon_{k+1} = \epsilon_k/2$.

For Example 3 we obtain the same solution as in all the previous algorithms. Likewise, the algorithm performs very well on Example 4, see Table 5 and Figure 6. We note, however, that Example 4 (when starting with a random initial guess) is

Table 5 Residuals, step sizes, and number of PDAS iterations ("iter") for Algorithm 5.7 when used to solve Examples 3 and 4

k	1	2	3	4
Example 3				
res^k	0.01018	8.7843e−7	9.2082e−11	
τ_k	1.0	1.0	1.0	
iter	84	84	84	
Example 4				
res^k	0.5388	0.00375	3.0727e−7	1.5366e−7
τ_k	1.0	1.0	0.5	0.00390625
iter	133	133	133	133
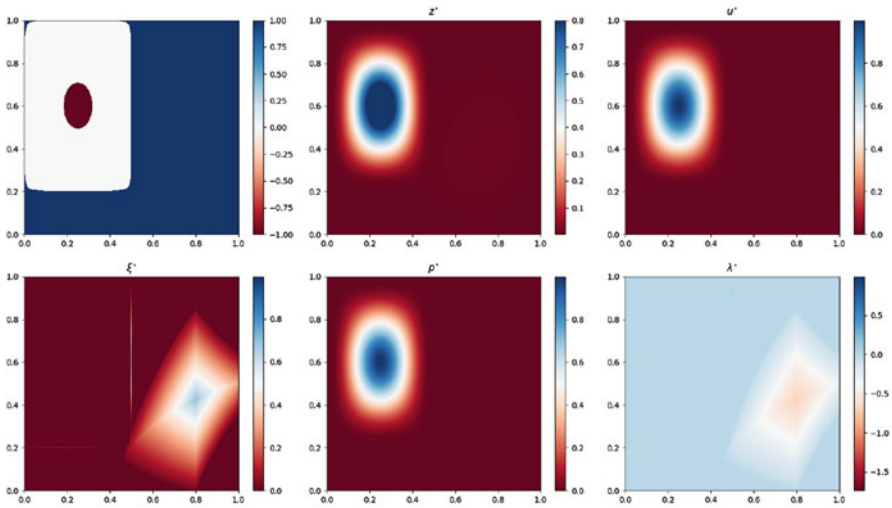
Fig. 6 Clockwise from upper left: characteristic functions $\chi_{\Omega^b_\star}$ (red) and $\chi_{\Omega^a_\star}$ (blue) (set of all indices with $z_i^\star \geq b_i - 1\text{e}{-}6$ and $z_i^\star \leq a_i + 1\text{e}{-}6$), optimal control $z^\star$, state $u^\star$, multiplier $\lambda^\star$, adjoint $p^\star$, and multiplier $\xi^\star$ for Example 4

more difficult to solve. In fact, once the difference in function values becomes negligible, i.e., on the order of $\mathcal{O}(10^{-13})$, the residual of S-stationarity stagnates at $\mathcal{O}(10^{-7})$. Finally, one important aspect of the theory for the bundle-free method was the relation $(\nabla \mathcal{M}_\epsilon(0), \delta z) < -q(\delta z)$. To see how the choice of ϵ_k influences this, see Table 6. There, we observe that far from the solution, a larger value of ϵ_k seems to yield a good approximation. However, the choice becomes critical once we close in on the solution.

Table 6 Behavior of Taylor expansion of $\mathscr{M}_\epsilon(0 + \delta z_k)$ at each iteration in Algorithm 5.7 when applied to Example 3

ϵ	$q(\delta z_0) + (\nabla \mathscr{M}_\epsilon(0), \delta z_0)$	$q(\delta z_1) + (\nabla \mathscr{M}_\epsilon(0)\delta z_1)$	$q(\delta z_2) + (\nabla \mathscr{M}_\epsilon(0), \delta z_2)$
1e−1	−19.1512	−0.264565	−0.264625
1e−2	−18.6723	−0.0392292	−0.0392487
1e−3	−18.5176	−0.000749805	−0.000750887
1e−4	−18.5087	−8.50068e−6	−8.37251e−6
1e−5	−18.5082	−2.36252e−7	−8.53014e−8
1e−6	−18.5082	−1.5286e−7	−8.54905e−10
1e−7	−18.5081	−1.52175e−7	−8.53968e−12
1e−8	−18.5081	−1.52185e−7	−8.5597e−14
1e−9	−18.5081	−1.52187e−7	−2.26899e−15
1e−10	−18.5081	−1.52187e−7	−1.55728e−15

6 Conclusion

Despite being a topic of interest for several decades, the optimal control of variational inequalities continues to be an active field of research. The rapidly growing interest in the past decade appears to be a result of the many theoretical, algorithmic, and computational advances in PDE-constrained optimization to date. We have seen here that both reduced space and full-space approaches are possible; however, the difficulties due to either a non-smooth control-to-state mapping or degenerate complementarity constraints persist. The various techniques for deriving optimality conditions in the presence of non-smoothness or degeneracy have led both in finite and infinite dimensions to a hierarchy of first-order optimality conditions. Some of these conditions are directly related to function-space-based numerical methods (C-stationarity) whereas other conditions (e.g., those of Mignot and Puel) are derived using concepts of generalized differentiation and a fine analysis of the regularity properties of the underlying functions and multipliers. These facts should therefore always be taken into account when developing numerical optimization algorithms.

In our numerical studies, we considered two main types of solution methods: smooth and non-smooth. Using approximation techniques for variational inequalities as in [30, 31], the smooth methods are almost always available and allow us to immediately take advantage of existing solvers for (smooth) PDE-constrained optimization problems. For the smoothed/regularized problems, we are only limited by our knowledge of the corresponding parameter-dependent PDE-constrained optimization problem. In our study, we make use of a smooth continuation approach and solve the first-order conditions directly using a semismooth Newton method. For small penalty parameters ϵ, the solution of the linear system (24) needed to calculate the updates is relatively well-behaved. However, the lower off-diagonal block becomes increasingly problematic as we attempt to approach the limiting problem for $\epsilon \downarrow 0$. Thus, we eventually pay a major price for smoothing the original problem.

As mentioned in the introduction, we present here several possible non-smooth methods that mirror related approaches in smooth PDE-constrained optimization. The first method is inspired by the classical projected subgradient methods in [54, 76]. Just as in these classical methods, the relative cost of each iteration is roughly as cheap as a standard projected-gradient approach sans line search: solve the VI, then a linear elliptic PDE, and compute a pointwise projection. As with all subgradient gradient methods, the strong convergence statements are essentially limited to convex problems. Nevertheless, the method presented here behaves quite well in practice and thus, warrants a deeper study in future research.

The active-set method has been taken from [36] and adapted to the setting of the canonical example. At every iteration, the cost of solving the nonlinear system is slightly more than in the smooth continuation approach due to the feasibility restoration step, which requires the solution of a mixed complementarity problem. However, the conditioning issues are now absent as the potentially problematic perturbation in the lower off-diagonal blocks has been removed. Just as in [36], this method performs exceptionally well, even on a problem with persistent biactivity throughout the iterations. However, as mentioned in the text, proving convergence of this method is rather difficult. One possibility would be to make regularity and monotonicity arguments on the data and active sets throughout the iterations (as was done in [42] for a related problem).

Finally, we considered the bundle-free implicit programming approach, which can be thought of as a globalization of the active-set strategy since we typically choose a step size of $\tau = 1$ at the beginning of every line search. Though the theory does contain several strong assumptions to guarantee unconditional global convergence to a stationary point, there appear to be no other genuinely non-smooth function-space-based algorithms for non-smooth non-convex problems currently in the literature. As noted in [46], if one can prove a kind of semismooth property of the reduced objective, then the convergence theory is greatly simplified. In comparison to the other non-smooth methods, it is clearly more costly due to the usage of the line search. The caveats for this method are the need to monitor the step sizes (essentially restarting if they get too small) and a meaningful heuristic for the convergence criteria. Nevertheless, the theory stands in contrast to non-convex bundle methods as outlined in, e.g., [75]. There, in the ideal non-convex setting, the algorithm terminates at a point at which 0 lies up to some tolerance $\varepsilon > 0$ in a convex hull of certain subgradients corresponding to points y_i that are not "far away" from the current iterate x_k. The connection to the various MPEC stationarity concepts is therefore unclear.

References

1. R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.
2. S. Albrecht and M. Ulbrich. Mathematical programs with complementarity constraints in the context of inverse optimal control for locomotion. *Optimization Methods and Software*, pages 1–29, 2017.
3. M. Anitescu, P. Tseng, and S.J. Wright. Elastic-mode algorithms for mathematical programs with equilibrium constraints: global convergence and stationarity properties. *Math. Program*, 110:337–371, 2005.
4. H. Antil, M. Hintermüller, R. H. Nochetto, T. M. Surowiec, and D. Wegner. Finite horizon model predictive control of electrowetting on dielectric with pinning. *Interfaces Free Bound.*, 19(1):1–30, 2017.
5. H. Attouch. *Variational Convergence for Functions and Operators*. Pitman Advanced Publishing Program, Boston, London, Melbourne, 1984.
6. H. Attouch, G. Buttazzo, and G. Michaille. *Variational analysis in Sobolev and BV spaces*, volume 6 of *MPS/SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. Applications to PDEs and optimization.
7. V. Barbu. *Optimal control of variational inequalities*, volume 100 of *Research Notes in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1984.
8. M. P. Bendsøe and O. Sigmund. *Topology Optimization. Theory, methods and Applications*. Springer Verlag, Berlin, Heidelberg, New York, 2003.
9. M. Bergounioux. Optimal control of variational inequalities: a mathematical programming approach. In *Modelling and optimization of distributed parameter systems (Warsaw, 1995)*, pages 123–130. Chapman & Hall, New York, 1996.
10. M. Bergounioux. Optimal control of an obstacle problem. *Appl. Math. Optim.*, 36(2):147–172, 1997.
11. M. Bergounioux. Use of augmented Lagrangian methods for the optimal control of obstacle problems. *J. Optim. Theory Appl.*, 95(1):101–126, 1997.
12. A. Bermúdez and C. Saguez. Optimal control of a Signorini problem. *SIAM J. Control Optim.*, 25(3):576–582, 1987.
13. D. P. Bertsekas. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control*, pages 174–184, 1976.
14. J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.*, 59(1):65–98, 2017.
15. J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Verlag, Berlin, Heidelberg, New York, 2000.
16. M. Boukrouche and D. A. Tarzia. Convergence of distributed optimal control problems governed by elliptic variational inequalities. *Comput. Optim. Appl.*, 53(2):375–393, 2012.
17. C. Brett, C. M. Elliott, M. Hintermüller, and C. Löbhard. Mesh adaptivity in optimal control of elliptic variational inequalities with point-tracking of the state. *Interfaces Free Bound.*, 17(1):21–53, 2015.
18. H. R. Brezis and G. Stampacchia. Sur la régularité de la solution d'inéquations elliptiques. *Bull. Soc. Math. France*, 96:153–180, 1968.
19. X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216 (electronic), 2000.
20. P. Colli, M. H. Farshbaf-Shaker, G. Gilardi, and J. Sprekels. Optimal boundary control of a viscous Cahn–Hilliard system with dynamic boundary condition and double obstacle potentials. *SIAM Journal on Control and Optimization*, 53(4):2696–2721, 2015.
21. J. C. De Los Reyes. Optimal control of a class of variational inequalities of the second kind. *SIAM J. Control Optim.*, 49(4):1629–1658, 2011.
22. J. C. De los Reyes, R. Herzog, and C. Meyer. Optimal control of static elastoplasticity in primal formulation. *SIAM J. Control Optim.*, 54(6):3016–3039, 2016.

23. J. C. De los Reyes and C. Meyer. Strong stationarity conditions for a class of optimization problems governed by variational inequalities of the second kind. *J. Optim. Theory Appl.*, 168(2):375–409, 2016.
24. A. K. Dond, T. Gudi, and N. Nataraj. A nonconforming finite element approximation for optimal control of an obstacle problem. *Comput. Methods Appl. Math.*, 16(4):653–666, 2016.
25. I. Ekeland and R. Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, English edition, 1999. Translated from the French.
26. L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in advanced mathematics. CRC Press, Boca Raton (Fla.), 1992.
27. M. H. Farshbaf-Shaker. A penalty approach to optimal control of Allen-Cahn variational inequalities: MPEC-view. *Numerical Functional Analysis and Optimization*, 33(11):1321–1349, 2012.
28. M. H. Farshbaf-Shaker and C. Hecht. Optimal control of elastic vector-valued Allen–Cahn variational inequalities. *SIAM Journal on Control and Optimization*, 54(1):129–152, 2016.
29. G. Fichera. Problemi elastostatici con vincoli unilaterali: Il problema di Signorini con ambigue condizioni al contorno. *Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. I (8)*, 7:91–140, 1963.
30. R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer Series in Computational Physics. Springer-Verlag, New York, 1984.
31. R. Glowinski, J. L. Lions, and R. Trémolières. *Numerical analysis of variational inequalities*, volume 8 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1981.
32. W. Han and B. D. Reddy. *Plasticity*, volume 9 of *Interdisciplinary Applied Mathematics*. Springer, New York, second edition, 2013. Mathematical theory and numerical analysis.
33. A. Haraux. How to differentiate the projection on a convex set in Hilbert space. some applications to variational inequalities. *J. Math. Soc. Japan*, 29(4):615–631, 1977.
34. R. Herzog, C. Meyer, and G. Wachsmuth. C-stationarity for optimal control of static plasticity with linear kinematic hardening. *SIAM J. Control Optim.*, 50(5):3052–3082, 2012.
35. R. Herzog, C. Meyer, and G. Wachsmuth. B- and strong stationarity for optimal control of static plasticity with hardening. *SIAM J. Optim.*, 23(1):321–352, 2013.
36. M. Hintermüller. An active-set equality constrained Newton solver with feasibility restoration for inverse coefficient problems in elliptic variational inequalities. *Inverse Problems*, 24(3):23pp., 2008.
37. M. Hintermüller, R. H. W. Hoppe, and C. Löbhard. Dual-weighted goal-oriented adaptive finite elements for optimal control of elliptic variational inequalities. *ESAIM Control Optim. Calc. Var.*, 20(2):524–546, 2014.
38. M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth newton method. *SIAM Journal on Optimization*, 13(3):865–888, 2002.
39. M. Hintermüller and I. Kopacka. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. *SIAM J. Optim.*, 20(2):868–902, 2009.
40. M. Hintermüller and I. Kopacka. A smooth penalty approach and a nonlinear multigrid algorithm for elliptic MPECs. *Comput. Optim. Appl.*, 50(1):111–145, 2011.
41. M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM Journal on Optimization*, 17(1):159–187, 2006.
42. M. Hintermüller and A. Laurain. Optimal shape design subject to elliptic variational inequalities. *SIAM Journal on Control and Optimization*, 49(3):1015–1047, 2011.
43. M. Hintermüller, C. Löbhard, and M. H. Tber. An ℓ^1-penalty scheme for the optimal control of elliptic variational inequalities. In Mehiddin Al-Baali, Lucio Grandinetti, and Anton Purnama, editors, *Numerical Analysis and Optimization*, volume 134 of *Springer Proceedings in Mathematics & Statistics*, pages 151–190. Springer International Publishing, 2015.

44. 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(1):555–582, 2014.
45. M. Hintermüller and T. Surowiec. First-order optimality conditions for elliptic mathematical programs with equilibrium constraints via variational analysis. *SIAM J. Optim.*, 21(4):1561–1593, 2011.
46. M. Hintermüller and T. Surowiec. A bundle-free implicit programming approach for a class of elliptic MPECs in function space. *Math. Program.*, 160(1):271–305, 2016.
47. M. Hintermüller and D. Wegner. Optimal control of a semidiscrete Cahn–Hilliard–Navier–Stokes system. *SIAM Journal on Control and Optimization*, 52(1):747–772, 2014.
48. M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
49. B. Horn and S. Ulbrich. Shape optimization for contact problems based on isogeometric analysis. *Journal of Physics: Conference Series*, 734(3):032008, 2016.
50. K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Applied Mathematics and Optimization*, 41(3):343–364, 2000.
51. P. Jaillet, D. Lamberton, and B. Lapeyre. Variational inequalities and the pricing of American options. *Acta Appl. Math.*, 21(3):263–289, 1990.
52. N. Kikuchi and J. T. Oden. *Contact problems in elasticity: a study of variational inequalities and finite element methods*, volume 8 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1988.
53. D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*, volume 88 of *Pure and Applied Mathematics*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980.
54. Krzysztof C. Kiwiel. Methods of descent for nondifferentiable optimization. Lecture Notes in Mathematics. 1133. Berlin etc.: Springer-Verlag, 1985.
55. R. Kornhuber. Monotone multigrid methods for elliptic variational inequalities. I. *Numer. Math.*, 69(2):167–184, 1994.
56. K. Kunisch and T. Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM J. Imaging Sci.*, 6(2):938–983, 2013.
57. J. L. Lions. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Dunod, Paris, 1968.
58. J.-L. Lions. Various topics in the theory of optimal control of distributed systems. In *Optimal control theory and its applications (Proc. Fourteenth Biennial Sem. Canad. Math. Congr., Univ. Western Ontario, London, Ont., 1973), Part I*, pages 116–309. Lecture Notes in Econom. and Math. Systems, Vol. 105. Springer, Berlin, 1974.
59. J.-L. Lions. *Contrôle des systèmes distribués singuliers*, volume 13 of *Méthodes Mathématiques de l'Informatique [Mathematical Methods of Information Science]*. Gauthier-Villars, Montrouge, 1983.
60. J.-L. Lions and G. Stampacchia. Variational inequalities. *Comm. Pure Appl. Math.*, 20:493–519, 1967.
61. Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, 1996.
62. C. Meyer and O. Thoma. A priori finite element error analysis for optimal control of the obstacle problem. *SIAM J. Numer. Anal.*, 51(1):605–628, 2013.
63. F. Mignot. Contrôle dans les inéquations variationelles elliptiques. *J. Functional Analysis*, 22(2):130–185, 1976.
64. F. Mignot and J.-P. Puel. Optimal control in some variational inequalities. *SIAM J. Control and Optimization*, 22(3):466–476, 1984.
65. K. Mombaur, A. Truong, and J.-P. Laumond. From human to humanoid locomotion–an inverse optimal control approach. *Autonomous Robots*, 28(3):369–383, 2010.
66. B. S. Mordukhovich. *Variational analysis and generalized differentiation. I: Basic Theory*, volume 330 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2006.

67. D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42(5):577–685, 1989.
68. P. Neittaanmäki, J. Sprekels, and D. Tiba. *Optimization of Elliptic Systems*. Springer Monographs in Mathematics. Springer, New York, 2006.
69. J. V. Outrata, M. Kočvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, volume 28 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1998.
70. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE transactions on pattern analysis and machine intelligence*, 12(7):174–184, 1990.
71. J.-F. Rodrigues. *Obstacle Problems in Mathematical Physics*. Number 134 in North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam, 1984.
72. L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1–4):259–268, 1992. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991).
73. C. Saguez. Optimal control of free boundary problems. In *System modelling and optimization (Budapest, 1985)*, volume 84 of *Lecture Notes in Control and Inform. Sci.*, pages 776–788. Springer, Berlin, 1986.
74. H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.
75. H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, 1992.
76. N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag, New York, 1985.
77. F. Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
78. J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Trans. Automat. Control*, 31(9):803–812, 1986.
79. M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13(3):805–842 (2003), 2002.
80. G. Wachsmuth. Strong stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Optim.*, 24(4):1914–1932, 2014.
81. G. Wachsmuth. Mathematical programs with complementarity constraints in Banach spaces. *J. Optim. Theory Appl.*, 166(2):480–507, 2015.
82. G. Wachsmuth. A guided tour of polyhedric sets: Basic properties, new results on intersections and applications. Technical report, TU Chemnitz, 2016.
83. G. Wachsmuth. Towards M-stationarity for optimal control of the obstacle problem with control constraints. *SIAM J. Control Optim.*, 54(2):964–986, 2016.
84. S. W. Walker, A. Bonito, and R. H. Nochetto. Mixed finite element method for electrowetting on dielectric with contact line pinning. *Interfaces Free Bound.*, 12(1):85–119, 2010.
85. Y-S. Wu, K. Pruess, and P. A. Witherspoon. Flow and displacement of Bingham non-Newtonian fluids in porous media. *SPE Reservoir Engineering*, 7(3), 1992.
86. J.-P. Yvon. *Etude de quelques problèmes de contrôle pour des systèmes distribués*. PhD thesis, Paris VI, 1973.
87. J.-P. Yvon. Optimal control of systems governed by variational inequalities. In *Fifth Conference on Optimization Techniques (Rome, 1973), Part I*, pages 265–275. Lecture Notes in Comput. Sci., Vol. 3. Springer, Berlin, 1973.
88. J. Zowe and S. Kurcyusz. Regularity and stability for the mathematical programming problem in Banach spaces. *Appl. Math. Optim.*, 5(1):49–62, 1979.