Results in Control and Optimization 13 (2023) 100309

Contents lists available at ScienceDirect

Results in Control and Optimization

journal homepage: www.elsevier.com/locate/rico

On solving elliptic obstacle problems by constant abs-linearization

Olga Weiß [a,*], Monika Weymuth [b]

[a] Institut für Bauingenieurwesen, Mathematische Methoden in Statik und Dynamik, Helmut-Schmidt-Universität - Universität der Bundeswehr Hamburg, Germany

[b] Institut für Mathematik und Computergestützte Simulation, Universität der Bundeswehr München, Germany

ARTICLE INFO

Keywords:
Variational inequality
Obstacle problem
Constrained optimal control
Finite element method
Nonsmooth optimization
Abs-linearization
A priori error analysis

ABSTRACT

We consider optimal control problems governed by an elliptic variational inequality of the first kind, namely the obstacle problem. The variational inequality is treated by penalization, which leads to optimization problems governed by a nonsmooth semi-linear elliptic PDE. The CALi algorithm is then applied for the efficient solution of these nonsmooth optimization problems. The special feature of the optimization algorithm CALi is the treatment of the nonsmooth Lipschitz-continuous operators abs, max and min, which allows to explicitly exploit the nonsmooth structure. Stationary points are located by appropriate decomposition of the optimization problem into so-called smooth constant abs-linearized problems. Each of these constant abs-linearized problems can be solved by classical means. The comprehensive algorithmic concept is presented, and its performance is discussed through examples.

1. Introduction

In the present paper we consider an optimal control problem governed by an elliptic variational inequality of the first kind, more precisely the obstacle problem. There are various important applications which are modeled by means of variational inequalities as, e.g., elastoplasticity or piezo electricity. Due to the variational inequality constraint, these kinds of problems are nonsmooth and non-convex, which complicates their theoretical and numerical treatment.

The focus of this paper is put on a new algorithmic realization. Therefore, we will not discuss the topic of necessary and sufficient optimality conditions, which is already intensively investigated in the literature. We mention, e.g., [1–9] and the references therein.

Various solution algorithms for optimal control of the obstacle problem already exist in the literature. A commonly used approach is to regularize or penalize the variational inequality to get a semi-linear partial differential equation (PDE), where the nonlinearity depends on the regularization parameter, see e.g., [2,3,6,10–14] and the references therein. We will also follow this approach.

In [15] a new structure exploiting optimization method to solve optimal control problems subject to elliptic semi-linear and nonsmooth equations, the so called CALi algorithm, is proposed. In contrast to the usually applied smoothing and regularization techniques for nonsmooth optimization problems, this algorithm allows for an explicit exploitation of the structure caused by the nonsmoothness. For this purpose a special treatment of the absolute value operator, the so-called constant abs-linearization, is applied, depending on the level, where the nonsmoothness occurs.

The purpose of this paper is to show that this algorithm is versatile and multifunctional, so that it can also be applied to optimal control problems governed by elliptic variational inequalities (VIs) of the first kind. The main goal of this paper is to elaborate and illustrate the adjustment of this algorithm to exactly this class of nonsmooth optimization problems. Therefore, the considered optimization problems are reformulated into optimal control problems governed by nonsmooth elliptic PDEs. The CALi algorithm

* Corresponding author.
E-mail address: weiss@hsu-hh.de (O. Weiß).

https://doi.org/10.1016/j.rico.2023.100309
Received 29 December 2022; Received in revised form 19 September 2023; Accepted 28 September 2023
Available online 1 October 2023
2666-7207/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

is then applied for the efficient solution of these nonsmooth optimization problems with the absolute value operator as the only source of nonsmoothness. Due to reformulations based on the constant abs-linearization and well-known abs-linear reformulations, this covers also (but not exclusively) nonsmoothness given by the max and min operators.

The exploitation of the given data allows a targeted and optimal decomposition of the optimization problem in order to compute stationary points. This approach is able to solve the considered class of nonsmooth optimization problems in comparably less Newton steps and additionally maintains reasonable convergence properties. Numerical results for nonsmooth optimization problems illustrate the proposed approach and its performance.

An appropriate decomposition of the optimization problem into so-called smooth constant abs-linearized problems allows to compute the solution of the corresponding optimization problem constrained by VIs. Each of these constant abs-linearized problems can be solved by classical means. The comprehensive algorithmic concept is presented, and its performance is discussed through examples.

In order to solve our problem numerically, we discretize the semi-linear PDE arising by regularization of the VI-constraint with the help of continuous, piecewise linear finite elements for the state and piecewise constant functions for the control. Additional results of the present paper are the convergence rates with respect to the regularization parameter for the error in the control and the state. Our final error estimate contains information about the coupling of the regularization parameter and the mesh size. Similar error estimates for another smoothing-scheme are established in [13].

This paper has the following structure.

In Section 2, we introduce the considered problem class of optimal control of obstacle problems, discuss its properties and propose a suitable regularization which leads to an optimization problem with nonsmooth PDE constraint. Furthermore, the solution operators corresponding to the original and the penalized problem are also introduced and examined, as well as their relation. Section 3 presents a reformulation of the nonsmooth optimization problem into a smooth one using the constant abs-linearization together with a solution approach involving a penalty term, and introduces the resulting optimization algorithm. Moreover, the chosen discretization approach as well as the solution of the resulting finite dimensional optimization problems are discussed. Section 4 deals with an investigation of error estimates with respect to regularization and discretization. Numerical results for a collection of test problems are presented and analyzed in Section 5. Finally, a conclusion and an outlook are given in Section 6.

2. Preliminaries

2.1. Notation and problem statement

Throughout this work we use the standard notation $H_0^1(\Omega)$ and $W^{k,p}(\Omega)$, $k \in \mathbb{N}$, $1 \le p \le \infty$ for the Sobolev spaces on a domain $\Omega \subset \mathbb{R}^d$, $d \ge 1$. We refer to [16] for details of these spaces. As usual, the dual of $H_0^1(\Omega)$ w.r.t. the L^2-inner product is denoted by $H^{-1}(\Omega)$ and the symbol $\langle \cdot, \cdot \rangle$ denotes the dual pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. The L^2-scalar product is denoted by $(\cdot, \cdot)$.

Moreover, we introduce the bilinear form $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ by

$$a(y, v) := \int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x.$$

The coercivity constant of $a(\cdot, \cdot)$ will be denoted by β, i.e.,

$$a(v, v) \ge \beta \|v\|^2_{H^1(\Omega)} \quad \forall v \in H_0^1(\Omega). \tag{2.1}$$

We consider optimal control problems governed by an elliptic variational inequality of the first kind. These problems are also known as optimal control of the (elliptic) obstacle problem (2.2b).

$$\min_{(y,u) \in K \times L^2(\Omega)} J(y, u) := \frac{1}{2} \|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2} \|u\|^2_{L^2(\Omega)}, \tag{2.2a}$$

$$\text{s.t.} \quad (\nabla y, \nabla(v - y)) \ge \langle u + f, v - y \rangle \quad \forall v \in K \tag{2.2b}$$

where the closed convex set K is defined by $K := \{v \in H_0^1(\Omega) : v \ge \psi \text{ a.e. in } \Omega\}$ and ψ denotes a given obstacle. Note that in the case $\psi \equiv 0$ the set K forms a cone.

We impose the following assumptions on the data in Eq. (2.2):

(i) $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is a bounded domain that is either convex and polygonal or has a $C^{1,1}$-boundary.
(ii) The desired state satisfies $y_d \in L^2(\Omega)$ and $\alpha > 0$ is a fixed real number.
(iii) The obstacle ψ satisfies $\psi \in W^{2,\infty}(\Omega)$, $\gamma\psi \le 0$ a.e. on $\partial\Omega$, where $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ denotes the trace operator.
(iv) The given disturbance f is a function in $L^2(\Omega)$.

Note that the condition $\gamma\psi \le 0$ a.e. on $\partial\Omega$ is needed to ensure the existence of solutions for the obstacle problem.

2.2. Known results and penalization of the obstacle problem

In the following, we summarize some known results about the variational inequality (2.2b) and the optimal control problem (2.2).

We start with an existence and uniqueness result.

Lemma 2.1. *For every $u \in H^{-1}(\Omega)$, the variational inequality (2.2b) has a unique solution $y \in K$. Moreover, the associated solution operator $S : H^{-1}(\Omega) \to K \subset H_0^1(\Omega)$ mapping u to y, is globally Lipschitz continuous with Lipschitz constant $L = 1/\beta$ with β as in Eq. (2.1).*

The proof is standard and can, for instance, be found in [17, Chapter II]. It is important to note that the control-to-state operator S is Gâteaux differentiable, if and only if the active set $\{x \in \Omega : y(x) = \psi(x)\}$ coincides with the fine support of the regular Borel measure associated with $-\Delta y - u$ up to a set of zero H_0^1-capacity, see [1].

The continuity of S and the weak lower semicontinuity of J imply the following result which can be found, e.g., in [10, Chapter 3]:

Proposition 2.2. *There exists a globally optimal solution of (2.2) which is in general not unique due to the nonlinearity of S.*

In order to solve the variational inequality (2.2b) we use a common technique called penalization. The idea of this method is to approximate the variational inequality by a sequence of nonlinear equations. For details of penalization, we refer to [18, Chapters 1 and 2] and [17, Chapter IV].

Using the max-function as penalty operator the variational inequality (2.2b) can be approximated by the penalized equation

$$(\nabla y, \nabla v) - \tfrac{1}{\varepsilon}(\max(0, \psi - y), v) = \langle f + u, v\rangle \quad \forall v \in H_0^1(\Omega) \tag{2.3}$$

with a parameter $\varepsilon > 0$. For every $u \in H^{-1}(\Omega)$, Eq. (2.3) has a unique solution $y_\varepsilon(u)$ due to the monotonicity of the max-function (see e.g., [18, Chapter 1]). Therefore, the associated solution operator $S_\varepsilon : H^{-1}(\Omega) \to H_0^1(\Omega)$, mapping u to y_ε, is well-defined.

Lemma 2.3. *The operator S_ε is globally Lipschitz continuous with Lipschitz constant $1/\beta$ with β as in Eq. (2.1).*

Proof. The proof is straightforward. We set $y_\varepsilon^{(1)} = S_\varepsilon(u_1)$ and $y_\varepsilon^{(2)} = S_\varepsilon(u_2)$ and insert $v = y_\varepsilon^{(1)} - y_\varepsilon^{(2)} \in H_0^1(\Omega)$ in Eq. (2.3) with $u = u_1$ and $u = u_2$. Subtracting the arising equalities and using the coercivity of $a(\cdot,\cdot)$ leads to

$$\begin{aligned}\beta\|y_\varepsilon^{(1)} - y_\varepsilon^{(2)}\|_{H^1(\Omega)}^2 &\le a(y_\varepsilon^{(1)} - y_\varepsilon^{(2)}, y_\varepsilon^{(1)} - y_\varepsilon^{(2)})\\ &= \langle u_1 - u_2, y_\varepsilon^{(1)} - y_\varepsilon^{(2)}\rangle\\ &\quad + \frac{1}{\varepsilon}\left(\max(0, \psi - y_\varepsilon^{(1)}) - \max(0, \psi - y_\varepsilon^{(2)}), y_\varepsilon^{(1)} - y_\varepsilon^{(2)}\right).\end{aligned}$$

The monotonicity of the max-function implies the claim. □

The following result is well-known and can be found, e.g., in [18, Chapter 1].

Lemma 2.4. *It holds that $S_\varepsilon(u) \to S(u)$ in $H_0^1(\Omega)$ as $\varepsilon \to 0$, where $S(u)$ denotes the solution of the variational inequality (2.2b) associated with u.*

Theorem 2.5. *Let $\{u_\varepsilon\}_{\varepsilon>0} \subset L^2(\Omega)$ be a sequence that converges weakly in $L^2(\Omega)$ to $u \in L^2(\Omega)$ as $\varepsilon \to 0$. Then we have the strong convergence*

$$S_\varepsilon(u_\varepsilon) \xrightarrow{\varepsilon\to 0} S(u) \quad \text{in } H_0^1(\Omega).$$

Proof. By the triangle inequality we have

$$\|S_\varepsilon(u_\varepsilon) - S(u)\|_{H^1(\Omega)} \le \|S_\varepsilon(u_\varepsilon) - S_\varepsilon(u)\|_{H^1(\Omega)} + \|S_\varepsilon(u) - S(u)\|_{H^1(\Omega)}.$$

We observe that the second term tends to zero by Lemma 2.4. Moreover, due to Lemma 2.3 the first term can be estimated by

$$\|S_\varepsilon(u_\varepsilon) - S_\varepsilon(u)\|_{H^1(\Omega)} \le \frac{1}{\beta}\|u_\varepsilon - u\|_{H^{-1}(\Omega)}. \tag{2.4}$$

By compact embeddings the right-hand side of (2.4) tends to zero for $\varepsilon \to 0$. □

Applying Eq. (2.3) the optimal control problem (2.2) can be approximated by

$$\min_{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)} J(y,u) := \frac{1}{2}\|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2}\|u\|_{L^2(\Omega)}^2, \tag{2.5a}$$

$$\text{s.t. } (\nabla y, \nabla v) - \tfrac{1}{\varepsilon}(\max(0, \psi - y), v) = \langle f + u, v\rangle \quad \forall v \in H_0^1(\Omega). \tag{2.5b}$$

Due to the continuity of S_ε and the weak lower semicontinuity of J problem (2.5) has a globally optimal solution. Moreover, using Theorem 2.5 we can argue as in [6, proof of Theorem 3.14] to show that for each strictly locally optimal pair (y^*, u^*) of Eq. (2.2) there is a family of local solutions $(y_\varepsilon, u_\varepsilon)$ of (2.5) that converges strongly to (y^*, u^*) in $H_0^1(\Omega) \times L^2(\Omega)$.

3. The constant abs-linearization

In this section we show how the nonsmooth optimization problem (2.5) can be reformulated into a closely related but smooth one. The reformulation is done by using the constant abs-linearization (CALi) which is introduced in [15] and based on the idea described in [19,20]. Moreover, we explain how the algorithm CALi can be adapted to our class of problems, and finally we examine a finite element discretization of the reformulated optimization problem.

3.1. Reformulation of the penalized optimal control problem

Throughout this section, the operator $\ell : L^2(\Omega) \to L^2(\Omega)$ is defined by

$$\ell(y) := \max(0, y).$$

The nonsmooth operator ℓ has the following properties:

Lemma 3.1.

1. *The operator $\ell(y)(x) = \ell(y(x))$, $\ell(\cdot) : L^2(\Omega) \to L^2(\Omega)$ denotes an (autonomous) Nemytzkii operator induced by some nonlinear and nonsmooth function $\ell(.) : \mathbb{R} \to \mathbb{R}$ which satisfies the Carathéodory conditions, i.e., the mapping $t \mapsto \ell(t)$ is continuous on $\mathbb{R}$.*
2. *The function $\ell : \mathbb{R} \to \mathbb{R}$ is monotonically increasing, and satisfies the growth condition*

$$|\ell(t)| \le K + L_g|t| \quad \text{for all } t \in \mathbb{R} \tag{3.1}$$

for some constants $0 \le K, L_g < \infty$. Furthermore, ℓ is globally Lipschitz continuous, i.e.,

$$\exists\ L > 0 : |\ell(t_1) - \ell(t_2)| \le L|t_1 - t_2| \quad \forall t_i \in \mathbb{R}\,. \tag{3.2}$$

3. *The Nemytzkii operator ℓ is directionally differentiable, i.e.,*

$$\left\| \frac{\ell(y + \tau h) - \ell(y)}{\tau} - \ell'(y; h) \right\|_{L^2(\Omega)} \to 0 \quad \text{for } \tau \to 0_+, \quad \forall y, h \in L^2(\Omega), \tag{3.3}$$

with $\ell'(y)$ being locally Lipschitz continuous and monotone.
4. *The operator ℓ can be expressed as a finite composition of the absolute value function and Fréchet differentiable operators.*

Proof. The operator $\ell(y) = \max(0, y)$ is a Nemytzkii operator induced by the pointwise non-linear and nonsmooth function $\max(0, .) : \mathbb{R} \to \mathbb{R}_{\ge 0}, t \mapsto \max(0, t)$, which satisfies the Carathéodory conditions. The function $\max(0, .)$ is obviously monotonically increasing and satisfies the growth condition (3.1) with $|\ell(t)| = |\max(0, t)| \le |t|$, i.e., $K = 0, L_g = 1$. Furthermore, $\max(0, .)$ is globally Lipschitz-continuous with constant $L = 1$.

Since the inducing function $\ell(.) = \max(0, .) : \mathbb{R} \to \mathbb{R}_{\ge 0}$ is directionally differentiable with

$$\max'((0, t); h) = \begin{cases} h, & \text{if } t > 0 \\ \max(0, h), & \text{if } t = 0 \\ 0, & \text{if } t < 0\,, \end{cases}$$

it is also globally Lipschitz-continuous and monotone itself. Hence,

$$|\max'((0, t); h)| \le |h|$$

for all directions $h \in \mathbb{R}$. Then Lebesgue's dominated convergence theorem implies also the directional differentiability of the induced Nemytzkii operator. This proves the assertions 1.-3. Assertion 4. follows from Proposition 3.2, which is introduced below. □

Note that by Lemma 3.1. assertions 1. and . it follows that also the associated Nemytzkii operator $\ell : L^2(\Omega) \to L^2(\Omega)$ is Lipschitz-continuous and monotonically increasing.

For convenience of the reader, we briefly explain the basic ideas of the structured evaluation and the constant abs-linearization described in [15, Definitions 2.4 and 3.1]. For this purpose we introduce a new auxiliary function z, called the *switching function*, for the argument of the absolute value function and σ for the sign of z. The reformulation results in a representation, where all evaluations of the absolute value function can be clearly identified and exploited.

Before we introduce the structured evaluation, we want to recall the well-known reformulation

$$\max(v, u) = (v + u + \mathrm{abs}(v - u))/2\,. \tag{3.4}$$

The structured evaluation allows for a useful reformulation of the considered nonsmooth formulation by introducing additional functions z for the argument of the absolute value evaluations. Consider the nonsmooth Lipschitz-continuous operator $\ell : H_0^1(\Omega) \to L^2(\Omega)$, $\ell(y) = \max(v, y)$ with $v \in H^1(\Omega)$:

Proposition 3.2 (*Structured Evaluation for* $\max(v,y)$)**.** *An equivalent representation of* $\ell(y)=\max(v,y)=\frac{1}{2}(v+y+\operatorname{abs}(v-y))$ *denoted by* $\hat{\ell}$ *can be obtained using the following procedure known as the structured evaluation which is given by*

$$\hat{\ell}(y,\sigma z)=\ \tfrac{1}{2}(v+y+\sigma(z)z)\ ,$$

with $z=v-y$*, i.e. the argument of the absolute value evaluation, and* $\sigma(z)=\operatorname{sign}(z)$ *so that* $\sigma(z)z=\operatorname{abs}(z)$*.*

The structured evaluation yields an equivalent representation of ℓ by $\hat{\ell}(y,\sigma z)$ where the absolute value evaluation can be directly exploited. At this point it is directly evident that the nonsmooth dependence exists only between z and σ. For more complicated and also nested nonsmooth operators such a reformulation by means of structured evaluation proves to be extremely useful, see e.g. [15,19,20]. Note that for the setting considered here, one has that $z\in H^1(\Omega)$ denotes the argument of the absolute value evaluation in the corresponding nonsmooth formulation, such that in the case of $\ell(y)=\max(0,\psi-y)$ it holds that $z=\psi-y$. Furthermore, the function σ is a Nemytzkii operator defined by

$$\sigma\ :\ H^1(\Omega)\to L^\infty(\Omega),\qquad [\sigma(z)](x)=\operatorname{sign}(z(x))\quad \text{a.e. in } \Omega$$

as functions of z. This choice ensures that $\sigma(z)z=\operatorname{abs}(z)\in H^1(\Omega)$ holds. Note that the Nemytzkii operator $\sigma(z)$ takes the values $-1,0$ or 1, i.e., $\sigma(z)(.)\ :\ \Omega\to\{-1,0,1\}$ and depends nonsmoothly on the switching function z. However, by applying the constant abs-linearization, defined below, one reformulates the operator equation of the optimization problem at hand into a smooth one.

In the further course, we will for brevity denote the Nemytzkii operator $\sigma(z)$ at z by σ.

Example 3.3 (*Structured Evaluation for* $\max(0,y)$)**.** Following Proposition 3.2 the structured evaluation for $\ell(y)=\max(0,y)=\frac{1}{2}(y+\operatorname{abs}(y))$ is then given by

$$\hat{\ell}(y,\sigma z)=\ \tfrac{1}{2}(y+\sigma z)\ ,$$

with $z=y,\sigma=\operatorname{sign}(z),\sigma z=\operatorname{abs}(z)$.

Definition 3.4 (*Constant Abs-Linearization*)**.** For a given nonsmooth operator equation involving the abs-operator or reformulated by the structured evaluation in the fashion of [15, Definition 2.4] and Proposition 3.2, the *constant abs-linearization* of the nonsmooth operator equation is obtained by fixing the involved function $\sigma(z)$ to a given $\bar{\sigma}\in L^\infty(\Omega)$ with $\bar{\sigma}(x)\in\{-1,1\}$ for all $x\in\Omega$, releasing the nonsmooth dependency on z.

Hence, using the constant abs-linearization, the resulting operator equation is smooth in both arguments z and $\bar{\sigma}$, since the nonsmooth dependence of σ on z has been eliminated.

Note, that in the context of constant abs-linearization, the function $\bar{\sigma}$ takes only the values 1 and -1, but no longer 0. However, this does not influence the previous considerations, simply because if $z>0$ and $\bar{\sigma}=+1$, or $z<0$ and $\bar{\sigma}=-1$ respectively, then $\bar{\sigma}z=\operatorname{abs}(z)$ is still valid. If $z=0$, then even for $\bar{\sigma}\neq 0$ the relationship $\operatorname{abs}(z)=\bar{\sigma}z=0$ is guaranteed. Hence, no longer considering zero as a value for $\bar{\sigma}$ does not pose any limitations. However, fixing σ to a certain function $\bar{\sigma}$ according to Definition 3.4 provides a linearization in the following sense. Since the dependency of z and σ has been removed, the term $\bar{\sigma}z$ is now smooth and even linear in z.

Since in the further course we will only deal with the concrete case $\ell(y)=-\frac{1}{\varepsilon}\max(0,\psi-y)$, let us consider a retransformation $z=\psi-y$, so that $\hat{\ell}(y,\sigma z)=-\frac{1}{2\varepsilon}(\psi-y+\sigma z)=-\frac{1}{2\varepsilon}(1+\sigma)(\psi-y)$.

Using the reformulation Eq. (3.4) for the max-function in Eq. (2.5b) as well as applying the constant abs-linearization problem (2.5) can be transformed into the smooth optimization problem

$$\min_{y,u\in H_0^1\times L^2}\ \frac{1}{2}\|y-y_d\|^2_{L^2(\Omega)}+\frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \tag{3.5a}$$

$$\text{s.t.}\quad (\nabla y,\nabla v)-\tfrac{1}{2\varepsilon}((1+\bar{\sigma})(\psi-y),v)=\langle u+f,v\rangle\quad \forall v\in H_0^1(\Omega) \tag{3.5b}$$

$$\bar{\sigma}(\psi-y)\geq 0\ \ \text{a.e. in } \Omega. \tag{3.5c}$$

We will see in the further course that this way by careful choice of the fixed $\bar{\sigma}$ a very closely related but smooth optimization problem can be created, which yields the same optimal solution as Eq. (2.5).

Following the same procedure as in [15], we treat the inequality constraint (3.5c) with a penalty approach such that the objective functional (3.5a) is modified to

$$\min_{y,u}\ \ J(y,u)+\nu\int_\Omega\Big(\max(-\bar{\sigma}(\psi-y),0)\Big)^4\,dx \tag{3.6}$$

with a penalty factor $\nu>0$. In this framework, as well as in the remainder of this paper, ν describes a non-negative constant penalty parameter for the inequality condition on $\bar{\sigma}(\psi-y)$. Here, the exponent 4 ensures that the target function is twice continuously differentiable despite the max function that is used for the formulation of the penalty function. The modified optimization problem, i.e. the penalized CAL problem formulation, then reads as

$$\min_{y,u}\ \ J(y,u)+\nu\int_\Omega\Big(\max(-\bar{\sigma}(\psi-y),0)\Big)^4\,dx \tag{3.7a}$$

$$\text{s.t.} \quad (\nabla y, \nabla v) - \tfrac{1}{2\varepsilon}\left((1+\bar{\sigma})(\psi - y), v\right) = \langle u + f, v\rangle \quad \forall v \in H_0^1(\Omega). \tag{3.7b}$$

The modified objective functional is then coupled with the equality constraint (3.5b) using Lagrange multipliers, resulting in the following Lagrange function

$$\begin{aligned}\mathcal{L}(y, u, \lambda_{PDE}) = J(y,u) + \nu \int_\Omega \max(-\bar{\sigma}(\psi - y), 0)^4 \,\mathrm{d}\Omega + \left(\nabla \lambda_{PDE}, \nabla y\right) \\ + \left(\lambda_{PDE}, -\tfrac{1}{2\varepsilon}((1+\bar{\sigma})(\psi - y)) - (u + f)\right).\end{aligned} \tag{3.8}$$

It shall be pointed out that even if in this work the definition of the switching function was transformed back, so that the formulation of the operator $\hat{e}$ without z was considered, the reformulations by means of the switching function z are very helpful and especially useful. They allow to directly recognize and exploit the individual absolute value evaluations, even for more complicated and also nested nonsmooth operators. Moreover, in [15] the optimality discussion requires and exploits the Lagrange multipliers corresponding to the equality constraint which defines the switching functions.

3.2. The algorithm CALi

Before we introduce our algorithm for the solution of problem (3.5), we want to investigate a specific choice for the fixed $\bar{\sigma}$.

Definition 3.5. For some $\psi \in W^{2,\infty}(\Omega)$ and $y_d \in L^2(\Omega)$ we denote by $\bar{\sigma}_\psi$ which is defined by

$$\bar{\sigma}_\psi = \operatorname{sign}(\psi - y_d) = \begin{cases} +1, & \text{on } \Omega^+ := \{x \in \Omega : y_d(x) \le \psi(x)\} \\ -1, & \text{on } \Omega^- := \{x \in \Omega : y_d(x) > \psi(x)\}, \end{cases} \tag{3.9}$$

the fixed $\bar{\sigma}$ with respect to the desired state y_d and the obstacle ψ according to constant abs-linearization in Definition 3.4 by virtue of the respective structured evaluation of $\max(0, \psi - y)$.

Note that due to this choice of the sign function, it holds that $\bar{\sigma}_\psi = +1$ where $y_d = \psi$ and $\bar{\sigma}_\psi = -1$ where $y_d = -\psi$. This is done to ensure, that the regularization term does not vanish completely from the problem formulation.

As already discussed in [15], the choice of the specific $\bar{\sigma}$ is crucial, since it determines the decomposition of the domain of the underlying optimization problem. In the following, we want to emphasize the domain decomposition due to the constant abs-linearization and motivate the choice of $\bar{\sigma}_\psi$. For this purpose we will first consider the domain decomposition given by $-\operatorname{sign}(y_d)$ with the desired state y_d, i.e.,

$$-\operatorname{sign}(y_d(x)) = \begin{cases} -1, \text{ for } x \in \Omega_d^{\ge 0} := \{x \in \Omega | y_d(x) \ge 0\}, \\ +1, \text{ for } x \in \Omega_d^{<0} := \{x \in \Omega | y_d(x) < 0\}. \end{cases} \tag{3.10}$$

Note that $\bar{\sigma} = -\operatorname{sign}(y_d)$ defined by Eq. (3.10) corresponds exactly to $\bar{\sigma}_\psi$ for $\psi \equiv 0$, i.e.,

$$\bar{\sigma}_0 = \begin{cases} -1, \text{ for } x \in \Omega^-, \\ +1, \text{ for } x \in \Omega^+. \end{cases}$$

Hence, every fixed $\bar{\sigma}$ and especially $\bar{\sigma}_\psi$ analogous to Eq. (3.10) correspondingly decomposes the domain Ω into subdomains such that $\Omega = \Omega^+ \cup \Omega^-$ with $\bar{\sigma}(x) = +1$ on Ω^+ and $\bar{\sigma}(x) = -1$ on Ω^-. We consider the following example.

Example 3.6 (*Domain Decomposition by $\bar{\sigma}$*)**.** In order to motivate the choice of a specific $\bar{\sigma}$ and to illustrate the corresponding domain decomposition, we examine the following obstacle problem from [21], where we replaced the domain Ω_1 by Ω:

$$\begin{aligned}\Omega = (0,1)^2 \subseteq \mathbb{R}^2, \;\; y_d(x_1, x_2) = -\sin(\pi x_1)\sin(\pi x_2), \;\; f(x_1, x_2) = -2\pi^2 \sin(\pi x_1)\sin(\pi x_2) \\ \text{and} \;\; \psi(x_1, x_2) = -0.25\end{aligned}$$

Note that due to the chosen data the optimal state is given by

$$y^* = \begin{cases} \psi, & \text{on } \Omega^+ := \{x \in \Omega : y_d(x) < \psi(x)\} \\ y_d, & \text{on } \Omega^- := \{x \in \Omega : y_d(x) \ge \psi(x)\} \end{cases}$$

and due to the choice of f, y^* can be attained for all controls $u \le 0$ and hence especially for $u^* \equiv 0$. Consequently, $\psi - y^*$ corresponds exactly to the function

$$\begin{cases} 0, & \text{on } \Omega^+ \\ \psi - y_d, & \text{on } \Omega^-. \end{cases}$$

Therefore, we choose the signature function

$$\bar{\sigma} = \bar{\sigma}_\psi = \begin{cases} +1, & \text{on } \Omega^+ \\ -1, & \text{on } \Omega^- \end{cases}$$

correspondingly. Based on these considerations we usually set $\bar{\sigma} = \bar{\sigma}_\psi$ as defined in Definition 3.5. This choice is reasonable according to the specifications in [15] for the cases considered here, whenever the function $y^* = \max(\psi, y_d)$ can be reached as a feasible state.

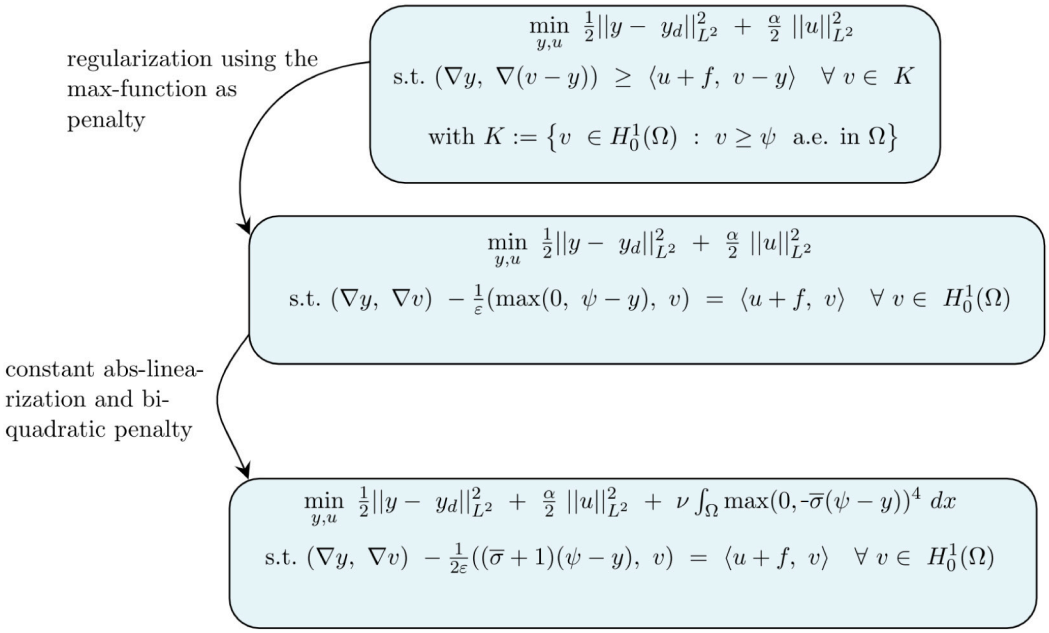

Fig. 1. The different optimal control problems and their relation.

Due to the decomposition of the domain Ω into $\Omega = \Omega^+ \cup \Omega^-$ provided by $\bar{\sigma}_\psi$ Eq. (3.5b) can be reformulated as

$$\begin{aligned} &\int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x - \frac{1}{2\varepsilon}\int_{\Omega^+}(1+\bar{\sigma}_\psi)(\psi - y)v \, \mathrm{d}x - \frac{1}{2\varepsilon}\int_{\Omega^-}(1+\bar{\sigma}_\psi)(\psi - y)v \, \mathrm{d}x = \int_\Omega (u+f)v \, \mathrm{d}x \\ \Leftrightarrow &\int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x - \frac{1}{\varepsilon}\int_{\Omega^+}(\psi - y)v \, \mathrm{d}x = \int_\Omega (u+f)v \, \mathrm{d}x \, . \end{aligned}$$

Motivated by [15] and the previous examinations, we propose the method stated in Algorithm 1 to solve the optimization problem (3.6) with constraint (3.5b).

Algorithm 1 CALi

Input: Initial values: $\bar{\sigma}_\psi, \mathbf{y}^0, \mathbf{u}^0$
Parameter: $\alpha, \nu, \varepsilon \geq 0$
Solve problem (3.6) subject to (3.5b) for $\bar{\sigma} = \bar{\sigma}_\psi$ to obtain $\mathbf{y}^*, \mathbf{u}^*, \boldsymbol{\lambda_{PDE}}^*$
Output: $\mathbf{y}^*, \mathbf{u}^*$

Since the proposed algorithm is essentially motivated by the special handling of the absolute value function, i.e., the constant abs-linearization, we call the resulting optimization algorithm *CALi* for *Constant Abs-Linearization*.

Fig. 1 shows the three main optimal control problems considered in this paper and how the original obstacle optimal control problem transforms into the regularized constant abs-linearized (CAL) problem formulation.

It should be noted that the solution of the smooth reformulated penalized problems can be accomplished with traditional methods of smooth optimization. For the numerical results shown in Section 5, we used a finite-element-approach based on FEniCS [22] to discretize the PDEs and to describe the other constraints in combination with a Newton method for the solution of the smooth modified constant abs-linearized problems. For the initial state, control and signature function $\bar{\sigma}_\psi$, the non-linear variational Lagrange problem is solved by Newton's method using the derivatives calculated within FEniCS.

Note that at this point the careful choice of $\bar{\sigma}$ allows the algorithm to get by without an update for $\bar{\sigma}$. However, in Section 5 we will also present a case where such an update could have been required. However strategies to efficiently updating this $\bar{\sigma}$, so that a sequence of related smooth optimization problems with associated solutions that further reduce the objective function value of the original problem and eventually lead to a minimal solution is obtained, are nontrivial. For the time being, these investigations remain the subject of current and future research. Nevertheless, such update strategies would always require the successive solution of CAL problems, which are studied and addressed in this paper.

3.3. The discrete system

We will discretize the considered systems with linear finite elements. For this purpose, we introduce a family of meshes $\{\mathbb{T}^h\}$. The mesh $\mathbb{T}^h$ consists of open triangles T and the mesh width is defined by

$$h := \max_{T\in\mathbb{T}^h} h_T \quad \text{with} \quad h_T := \operatorname{diam}(T).$$

Moreover, we assume that $\mathbb{T}^h$ is quasi-uniform in the sense of [23].

For the discretization of system (2.2) we introduce the space of piecewise linear functions

$$V^h := \{v^h \in H_0^1(\Omega) : \ v^h|_T \in \mathbb{P}_1(T) \ \forall T \in \mathbb{T}^h\},$$

where $\mathbb{P}_1$ denotes the space of polynomials of degree ≤ 1. The nodal basis of V^h is given by $\{\xi_1, \dots, \xi_n\}$. Moreover, we define the space of piecewise constant functions by

$$U^h := \operatorname{span}\{e_T : T \in \mathbb{T}^h\},$$

where $e_T : \Omega \to \mathbb{R}$ is the characteristic function for the simplex $T \in \mathbb{T}^h$.

For a given function $y^h \in V^h$ we denote by $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ its vector of coefficients with respect to the basis $\{\xi_1, \dots, \xi_n\}$, i.e.,

$$y^h(x) = \sum_{i=1}^{n} y_i \xi_i(x) \, .$$

Similarly, every discretized control function in the space U^h with $\mathbf{u} = (u_1, \dots, u_m)^T \in \mathbb{R}^m$ can be written as

$$u^h(x) = \sum_{i=1}^{m} u_i e_{T_i}(x) \, ,$$

where m is the total number of elements T in the triangulation $\mathbb{T}^h$.

One has to take into account that the operators max and abs are non-linear. The above representations yield the following discretization for some non-linear operator ℓ (once again we can consider $\ell(y) = \max(0, y)$ or $\ell(y) = \operatorname{abs}(y)$):

$$\ell(y^h) = \ell\left(\sum_{i=1}^{n} y_i \xi_i\right)$$

and

$$(\ell(y^h), v^h) = \int_\Omega \ell(y^h) v^h \, \mathrm{d}x \approx \sum_{T\in\mathbb{T}^h} \int_T \ell(y^h) v^h \, \mathrm{d}x \, . \tag{3.11}$$

The integrals over the elements $T \in \mathbb{T}^h$ are approximated by some quadrature formula

$$\sum_{T\in\mathbb{T}^h} \int_T \ell(y^h) v^h \, \mathrm{d}x \approx \sum_{T\in\mathbb{T}^h} \sum_{k=1}^{n_k} \omega_k \, \ell\left(\sum_{i=1}^{n} y_i \xi_i(x_k)\right) \sum_{j=1}^{n} \xi_j(x_k) \, , \tag{3.12}$$

with n_k quadrature points per element T and corresponding weights ω_k.

Hence, the naturally arising discretization for some nonsmooth operator equation like (2.5b) in the finite element context is per quadrature point, which increases the total number of absolute value evaluations.

The signature function $\bar{\sigma}$ is discretized similarly to the state y, such that $\bar{\sigma}^h \in V^h$ with the coefficient vector $\bar{\boldsymbol{\sigma}} = (\bar{\sigma}_1, \dots, \bar{\sigma}_n)^T \in \mathbb{R}^n$.

As seen in Eq. (3.12), we would like to point out that the number of nonsmooth functions ℓ in the discrete problem is per quadrature point. As already discussed in [15], this is not in perfect alignment with a representation of some finite element function like the state y. The consequent choice of this discretization leads to an increase of the polynomial degree due to the multiplication $\bar{\boldsymbol{\sigma}}(\boldsymbol{\psi} - \mathbf{y})$ in the discrete representation of the operator $\hat{\ell}$ opposed to the operator ℓ. However, this type of discretization allows for a straightforward implementation with FEniCS. Accordingly, the given expressions for the signature function is projected onto the space V^h. Thus, $\bar{\sigma}^h$ has as many entries as y^h. Consequently, $\bar{\sigma}^h$ attains only the values $+1$ and -1 on the corresponding mesh points. However, within an element $T \in \mathbb{T}^h$ the function $\bar{\sigma}^h$ can be linear and might go through zero to ensure the chosen sign constellation. See Example 3.7 for an illustrative example.

Example 3.7. For $\Omega = (0, 1) \times (0, 1)$ let $\bar{\sigma}$ be given by

$$\bar{\sigma} = \begin{cases} +1, & \text{if } x_1, x_2 < \frac{1}{2} \\ -1, & \text{else} \, . \end{cases} \tag{3.13}$$

Then Fig. 2 illustrates a very coarse discretization of Ω into triangles and the discretization of $\bar{\sigma}$ over the discretized domain. It reveals that $\bar{\sigma}^h$ attains only the values $+1$ and -1, respectively, at the corresponding vertices of the triangulation. On the elements of the triangulation, where one of the points takes a different sign than the other two, $\bar{\sigma}^h$ shows a linear progression between these values.

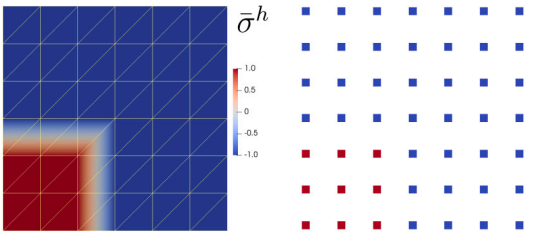

Fig. 2. (Left) surface plot of $\bar{\sigma}^h$ over Ω^h with edges of the mesh and (right) mesh point values of $\bar{\sigma}^h$ corresponding to Eq. (3.13).

Using the above notation the discretization of the optimization problem (2.5) where the variational inequality is approximated by the penalized equation Eq. (2.5b) is given by:

$$\min_{(y^h,u^h)\in V^h\times U^h} J(y^h,u^h), \tag{3.14a}$$

$$\text{s.t.} \quad (\nabla y^h,\nabla v^h) - \tfrac{1}{\varepsilon}(\max(\psi^h - y^h, 0^h), v^h) = (u^h, v^h) + (f^h, v^h) \quad \forall v^h \in V^h \tag{3.14b}$$

Here, we set $\psi^h = I^h\psi$, where $I^h : C(\bar{\Omega}) \to V^h$ denotes the standard Lagrange interpolation operator. Due to the continuous embedding $W^{2,\infty}(\Omega) \hookrightarrow C(\bar{\Omega})$, we have that $I^h\psi$ is well-defined. Further $f^h := P^h f$, where $P^h : L^2(\Omega) \to U^h$ is the L^2-projection onto the space of piecewise constant functions.

Inserting Eq. (3.11) into Eq. (3.14b) and replacing v by ξ leads to:

$$\begin{aligned}&\int_\Omega \sum_{k=1}^n \nabla\xi_j(x)\cdot\nabla\xi_k(x) y_k - \frac{1}{\varepsilon}\ell\left(\sum_{i=1}^n y_i\xi_i(x)\right)\xi_j(x)\,\mathrm{d}x\\ &= \int_\Omega\left(\sum_{s=1}^m u_s e_{T_s}(x)\right)\xi_j(x)\,\mathrm{d}x + \int_\Omega\left(\sum_{t=1}^m f_t e_{T_t}(x)\right)\xi_j(x)\,\mathrm{d}x\,,\end{aligned} \tag{3.15}$$

for $1 \le j \le n$. Here $\ell(y)$ is given by $\ell(y) = \max(\psi^h - y, 0)$ and hence the second term on the left-hand side is particularly given by

$$\int_\Omega \ell\left(\sum_{i=1}^n y_i\xi_i(x)\right)\xi_j(x)\,\mathrm{d}x = \int_\Omega \max\left(\sum_{i=1}^n(\psi_i - y_i)\xi_i(x), 0\right)\xi_j(x)\,\mathrm{d}x.$$

By defining

$$\begin{aligned}A_{jk} &:= \int_\Omega \nabla\xi_j(x)\cdot\nabla\xi_k(x)\,\mathrm{d}x = (\nabla\xi_j,\nabla\xi_k),\\ b_k(y^h) &:= \int_\Omega \ell\left(\sum_{i=1}^n y_i\xi_i(x)\right)\xi_k(x)\,\mathrm{d}x\end{aligned}$$

and

$$g_j := \int_\Omega\left(\sum_{s=1}^m u_s e_{T_s}(x)\right)\xi_j(x)\,\mathrm{d}x + \int_\Omega\left(\sum_{t=1}^m f_t e_{T_t}(x)\right)\xi_j(x)\,\mathrm{d}x\,,$$

Eq. (3.15) can be rewritten as

$$\sum_{k=1}^n A_{jk}y_k - \frac{1}{\varepsilon}b_k(y^h) = g_j \text{ for } 1 \le j \le n\,.$$

Here A_{jk} represent the entries of the stiffness matrix A. The discretization of the PDE results in a non-linear system of algebraic equations, which we abbreviate as

$$A\mathbf{y} - \frac{1}{\varepsilon}\mathbf{b}(\mathbf{y}) = (\mathbf{u}^T + \mathbf{f}^T)E\,, \tag{3.16}$$

with the control matrix $E_{ij} := (e_{T_i}, \xi_j)$ and $\mathbf{y} = (y_1, \dots, y_n)^T$ denoting the finite-element approximation belonging to the right-hand side given by the discrete control u and the discrete perturbation f. To this end, the function $y^h|_{T_k}$ on the linear element T_k is realized in terms of its values at the vertices of T_k. Note that in the above algebraic system the vectors $\boldsymbol{u}$ and $\boldsymbol{f}$ as well as the matrices A, E are constant since they are independent of the unknowns $y_1, \dots, y_n$. However, as previously mentioned, this non-linear algebraic equation is assumed to be based on a reasonable approximation of the integral via quadrature.

The resulting discrete objective functional reads as

$$\min_{(\mathbf{y},\mathbf{u})\in\mathbb{R}^n\times\mathbb{R}^m} J(\mathbf{y},\mathbf{u}) = \frac{1}{2}(\mathbf{y}-\mathbf{y}_d)^T M(\mathbf{y}-\mathbf{y}_d) + \frac{\alpha}{2}\mathbf{u}^T D\mathbf{u}\,.$$

Herein $M \in \mathbb{R}^{n\times n}$ denotes the mass matrix $M_{ij} = (\xi_i, \xi_j)$ and D the control mass matrix with the entries $D_{ij} = (e_{T_i}, e_{T_j})$, where D is a diagonal matrix because the interior of the triangles are disjunct to each other.

Analogous to the previously deduced discretization, the discrete version of the constant abs-linearized problem Eq. (3.7) with the added penalty approximation is given by:

$$\begin{aligned} &\min_{(y^h,u^h)\in V^h\times U^h} \; J(y^h,u^h) + \nu \int_\Omega \max\left(-\bar\sigma^h(\psi^h - y^h),0\right)^4 \,\mathrm{d}x \\ &\text{s.t.} \quad (\nabla y^h,\nabla v^h) - \tfrac{1}{2\varepsilon}\left((1+\bar\sigma^h)(\psi^h - y^h),v^h\right) = (u^h,v^h) + (f^h,v^h), \qquad \forall v^h \in V^h\,. \end{aligned} \tag{3.17}$$

Under consideration of Eq. (3.17), it becomes clear that the inequality constraint from Eq. (3.5c) is enforced per quadrature point via our penalty approach. We assume that the optimization problem Eq. (3.17) fulfills some kind of constraint qualification to ensure the existence of the Lagrange multipliers. The corresponding discrete Lagrange functional related to the penalized constant abs-linearized problem of system Eq. (3.17) is now given by

$$\begin{aligned} \mathcal{L}(y^h,u^h,\lambda^h_{PDE}) = J(y^h,u^h) &+ \left(\nabla\lambda^h_{PDE},\nabla y^h\right) + \nu\int_\Omega \left(\max(-\bar\sigma^h(\psi^h - y^h),0)\right)^4 \mathrm{d}x \\ &+ \left(\lambda^h_{PDE}, -\tfrac{1}{2\varepsilon}(1+\bar\sigma^h)(\psi^h - y^h) - u^h - f^h\right). \end{aligned} \tag{3.18}$$

The KKT system corresponding to Eq. (3.18) with e.g., $\bar\sigma^h = (\bar\sigma_\psi)^h$ is then solved with a non-linear variational Newton solver.

4. Error estimates

In this section we will prove an error estimate for the L^2-error of the control, i.e., for $\|u^*_{\varepsilon,h} - u^*\|_{L^2(\Omega)}$ under the assumption that a quadratic growth condition holds. Here u^* and $u^*_{\varepsilon,h} \in L^2(\Omega)$, respectively, denote locally optimal solutions of (2.2) and (3.14), respectively. In order to derive an error bound, we adapt the technique introduced in [14]. The proof is based on a quadratic growth condition and the L^2-error estimates for the state presented in [24,25], respectively.

In order to simplify the notation we introduce the reduced functionals

$$\begin{aligned} g &: L^2(\Omega)\to\mathbb{R}, \quad g(u) := J(S(u),u) \\ g_\varepsilon &: L^2(\Omega)\to\mathbb{R}, \quad g_\varepsilon(u) := J(S_\varepsilon(u),u) \\ g_{\varepsilon,h} &: L^2(\Omega)\to\mathbb{R}, \quad g_{\varepsilon,h}(u) := J(S_{\varepsilon,h}(u),u). \end{aligned}$$

Moreover, let u^*, u^*_ε and $u^*_{\varepsilon,h} \in L^2(\Omega)$ be locally optimal solutions of (2.2), (3.5) and (3.14), respectively.

At this point we want to emphasize that if $\bar\sigma$ is chosen suitably the discrete optimization problem

$$\begin{aligned} &\min_{(y^h,u^h)\in V^h\times U^h} \; J(y^h,u^h) \\ &\text{s.t.} \quad (\nabla y^h,\nabla v^h) - \tfrac{1}{2\varepsilon}\left((1+\bar\sigma^h)(\psi^h - y^h),v^h\right) = (u^h,v^h) + (f^h,v^h), \qquad \forall v^h \in V^h \\ &\qquad\qquad \bar\sigma^h(\psi^h - y^h) \ge 0 \quad \text{a.e. in } \Omega \end{aligned} \tag{4.1}$$

is just a reformulation of problem (3.14) and consequently also the optimal state as well as the optimal control of the two problems are the same. Since problem (3.14) is more convenient for the error analysis, we always consider problem (3.14) instead of (4.1) in this section. However, the reader should be aware that the error estimates are also valid for (4.1).

Fig. 3 presents the relation between the considered problems and reformulations together with their discrete counterpart.

We make the following assumptions for our error analysis:

Assumption 4.1.

(i) $u, f \in L^\infty(\Omega)$

(ii) There holds a quadratic growth condition, i.e., there are $\rho,\ \delta > 0$ such that

$$g(u^*) \le g(u) - \delta\|u - u^*\|^2_{L^2(\Omega)} \quad \forall u \in B_\rho(u^*), \tag{4.2}$$

where $B_\rho(u^*) := \{u \in L^2(\Omega) : \|u - u^*\|_{L^2(\Omega)} \le \rho\}$.

(iii) $\bar\sigma$ is chosen suitably such that the discrete problem (3.14) is just a reformulation of (4.1) and consequently the optimal state and the optimal control of the two problems coincide.

Remark 4.2. For the obstacle problem, the quadratic growth condition (4.2) holds if u^* satisfies some second-order sufficient optimality conditions (cf. [6]).

We start with an L^2-error estimate of the regularization error for the state, which is proven in [24] for a convex and polygonal domain and in [25] for a domain with $C^{1,1}$-boundary.

Theorem 4.3. *Let y and y_ε be the solutions of (2.2b) and (2.5b), respectively. Then there holds the estimate*

$$\|y - y_\varepsilon\|_{L^\infty(\Omega)} \le C\varepsilon$$

with a constant $C > 0$ independent of ε.

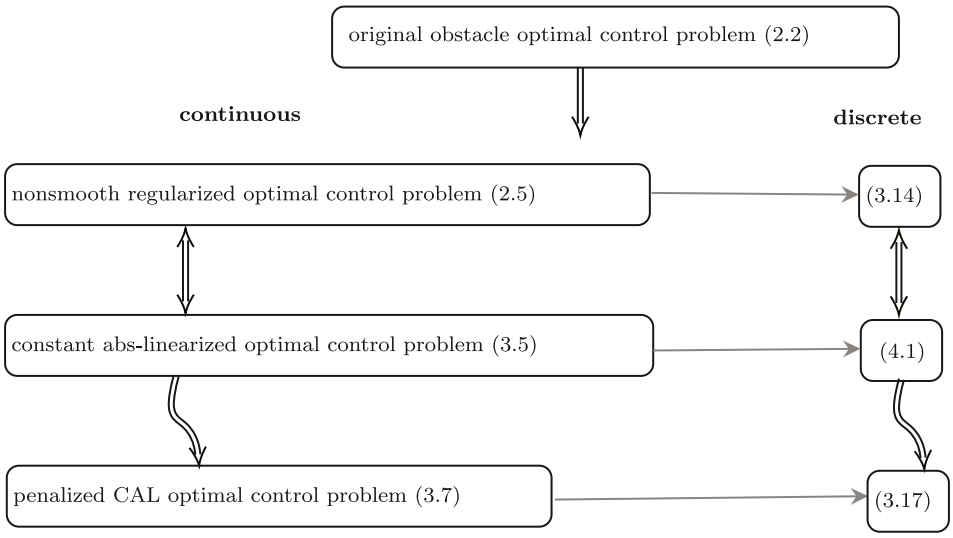

Fig. 3. Relation between the considered problems and their discrete counterpart.

Next, we state an L^2-error estimate of the discretization error for the state, which is proven in [24].

Theorem 4.4. *Let Ω be a bounded domain which is convex and polygonal. Moreover, let y_ε and $y_{\varepsilon,h}$ be the solutions of (2.5b) respectively (3.14b). In addition, assume that the obstacle satisfies $\psi < 0$ on the boundary. Then there exist constants $\varepsilon_d > 0$ and $h_d > 0$ such that for all $\varepsilon \le \varepsilon_d$ and $h \le h_d$ there holds*

$$\|y_\varepsilon - y_{\varepsilon,h}\|_{L^2(\Omega)} \le Ch^2 |\log h|^2 \left(\|f\|_{L^\infty(\Omega)} + \|u\|_{L^\infty(\Omega)} + \|\Delta\psi\|_{L^\infty(\Omega)} \right)$$

with a constant $C > 0$ independent of h.

Remark 4.5. The same convergence rate is also shown in [25], where the boundary is assumed to be smooth, i.e., of class $C^{1,1}$. It is worth noting that the condition $\psi < 0$ on the boundary is not necessary if the boundary is smooth.

We continue with some preparatory lemmas which are needed for our error analysis. Throughout the remainder of this paper, let $(\varepsilon_n, h_n)_{n\in\mathbb{N}} \subset \mathbb{R}^2_{>0}$ denote a sequence converging to zero.

Lemma 4.6. *Assume that $\psi < 0$ on the boundary, $\varepsilon_n \le \varepsilon_d$ and $h_n \le h_d$ with ε_d and h_d as in Theorem 4.4. Let $\{u_{\varepsilon_n,h_n}\}$ be a sequence that converges strongly in $H^{-1-1}(\Omega)$ to $u \in H^{-1}(\Omega)$ as $n \to \infty$. Denote the solution of the discretized Eq. (3.14b) corresponding to u_{ε_n,h_n} by y_{ε_n,h_n} and the solution of (2.2b) corresponding to u by y. Then $y_{\varepsilon_n,h_n} \to y$ in $H^1_0(\Omega)$.*

Proof. We have

$$\begin{aligned}\|y - y_{\varepsilon_n,h_n}\|_{H^1(\Omega)} &= \|S(u) - S_{\varepsilon_n,h_n}(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)} \\ &\le \|S(u) - S(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)} + \|S(u_{\varepsilon_n,h_n}) - S_{\varepsilon_n,h_n}(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)}.\end{aligned}$$

By Theorem 2.5 we have

$$\|S(u) - S(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)} \xrightarrow{n\to\infty} 0.$$

Moreover, applying the triangle inequality leads to

$$\begin{aligned}\|S(u_{\varepsilon_n,h_n}) - S_{\varepsilon_n,h_n}(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)} \le{}& \|S(u_{\varepsilon_n,h_n}) - S_{\varepsilon_n}(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)} \\ &+ \|S_{\varepsilon_n}(u_{\varepsilon_n,h_n}) - S_{\varepsilon_n,h_n}(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)}.\end{aligned}$$

Theorem 4.3 and the standard finite element error estimate for the H^1-norm yield

$$\|S(u_{\varepsilon_n,h_n}) - S_{\varepsilon_n,h_n}(u_{\varepsilon_n,h_n})\|_{H^1(\Omega)} \le C(\varepsilon_n + h_n) \xrightarrow{n\to\infty} 0,$$

which concludes the proof. □

Lemma 4.7. *Suppose that u^* satisfies the quadratic growth condition (4.2). Then there is a sequence $\{u^*_{\varepsilon_n,h_n}\}$ of locally optimal solutions to (3.17) with $u^*_{\varepsilon_n,h_n} \to u^*$ in $L^2(\Omega)$ as $n \to \infty$.*

Proof. Based on Lemma 4.6 the following proof is standard (see also [14, Lemma 5.5], where an analogous result is proven). Nevertheless, for later purpose and for convenience of the reader, we sketch the arguments. Following the classical localization argument from [26], we define the following discrete problems:

$$\min_{u\in B_\rho(u^*)} g_{\varepsilon_n,h_n}(u), \tag{4.3}$$

where $B_\rho(u^*)$ denotes the closed L^2-ball from (4.2). By standard arguments, the above problem admits a globally optimal solution for every $n<\infty$, denoted by $u^*_{\varepsilon_n,h_n}$. Due to the constraint this sequence is bounded in $L^2(\Omega)$ and thus admits a weakly convergent subsequence with limit $\tilde u\in L^2(\Omega)$, which, by compact embedding, converges strongly in $H^{-1}(\Omega)$. By Lemma 4.6 the associated states $y^*_{\varepsilon_n,h_n}:=S_{\varepsilon_n,h_n}(u^*_{\varepsilon_n,h_n})$ converge strongly to $\tilde y:=S(\tilde u)$. The weak lower semicontinuity of the objective functional along with the isolated local optimality of u^* implies $\tilde u=u^*$. Moreover, the Tikhonov term in the objective functional yields the norm convergence of $u^*_{\varepsilon_n,h_n}$ so that $u^*_{\varepsilon_n,h_n}\to u^*$ in $L^2(\Omega)$. This implies that $u^*_{\varepsilon_n,h_n}$ is in the interior of $B_\rho(u^*)$ for n sufficiently large and, therefore, $u^*_{\varepsilon_n,h_n}$ is a local solution of (3.14). □

Lemma 4.8. *Let $\{u^*_{\varepsilon_n,h_n}\}$ be the sequence of Lemma 4.7. Then $\{u^*_{\varepsilon_n,h_n}\}$ is uniformly bounded in $H^1(\Omega)$.*

Proof. Analogously to [27, Corollary 4.5] one can derive the following optimality system for (3.17): For every locally optimal solution $u^*_{\varepsilon_n,h_n}$ of (3.17) with associated state $y^*_{\varepsilon_n,h_n}$, there exist an adjoint state $p^*_{\varepsilon_n,h_n}\in V^h$ and a multiplier $\mu^*_{\varepsilon_n,h_n}\in L^\infty(\Omega)$ such that

$$\begin{aligned} a(p^*_{\varepsilon_n,h_n},v^h)+(\mu^*_{\varepsilon_n,h_n}p^*_{\varepsilon_n,h_n},v^h)&=(y^*_{\varepsilon_n,h_n}-y_d,v^h)\quad\forall v^h\in V^h \\ \mu^*_{\varepsilon_n,h_n}(x)&\in\partial_c\max(y^*_{\varepsilon_n,h_n}(x))\quad\text{a.e. in }\Omega \end{aligned} \tag{4.4}$$

$$p^*_{\varepsilon_n,h_n}(x)+\alpha u^*_{\varepsilon_n,h_n}(x)=0\quad\text{a.e. in }\Omega, \tag{4.5}$$

where $\partial_c\max:\mathbb{R}\rightrightarrows[0,\frac{1}{\varepsilon}]$ denotes the convex subdifferential of the function $\xi(y)=-\frac{1}{\varepsilon}\max(\psi-y,0)$. Testing Eq. (4.4) with $p^*_{\varepsilon_n,h_n}\in V^h$, the coercivity of $a(\cdot,\cdot)$ and Hölder's inequality leads to

$$\begin{aligned}\beta\|p^*_{\varepsilon_n,h_n}\|^2_{H^1(\Omega)}\le(\nabla p^*_{\varepsilon_n,h_n},\nabla p^*_{\varepsilon_n,h_n})&=(y^*_{\varepsilon_n,h_n}-y_d,p^*_{\varepsilon_n,h_n})-(\mu^*_{\varepsilon_n,h_n}p^*_{\varepsilon_n,h_n},p^*_{\varepsilon_n,h_n})\\&\le\|y^*_{\varepsilon_n,h_n}-y_d\|_{L^2(\Omega)}\|p^*_{\varepsilon_n,h_n}\|_{H^1(\Omega)}.\end{aligned}$$

Here we used that $(\mu^*_{\varepsilon_n,h_n}p^*_{\varepsilon_n,h_n},p^*_{\varepsilon_n,h_n})\ge0$ due to the monotonicity of the max-function. Due to Eq. (4.5) we arrive at

$$\|u^*_{\varepsilon_n,h_n}\|_{H^1(\Omega)}=\frac{1}{\alpha}\|p^*_{\varepsilon_n,h_n}\|_{H^1(\Omega)}\le\frac{1}{\alpha\beta}\|y^*_{\varepsilon_n,h_n}-y_d\|_{L^2(\Omega)}\le\frac{1}{\alpha\beta}\left(\|y^*_{\varepsilon_n,h_n}\|_{L^2(\Omega)}+\|y_d\|_{L^2(\Omega)}\right).$$

The boundedness of $y^*_{\varepsilon_n,h_n}$ in $L^2(\Omega)$ implies the claim. □

Theorem 4.9. *Suppose that u^* satisfies the quadratic growth condition (4.2) and $\psi<0$ on the boundary. Then there exist constants $\varepsilon_d>0$ and $h_d>0$ such that for all $\varepsilon\le\varepsilon_d$ and $h\le h_d$ one has*

$$\|u^*_{\varepsilon,h}-u^*\|_{L^2(\Omega)}\le C\left(\varepsilon^{1/2}+h|\log h|\right)$$

with constant $C>0$ independent of ε and h.

Proof. The proof follows the lines of [14, Theorem 5.8]. As seen in the proof of Lemma 4.7, $u^*_{\varepsilon,h}$ is a global solution of (4.3) and therefore

$$g_{\varepsilon,h}(u^*_{\varepsilon,h})\le g_{\varepsilon,h}(u^*). \tag{4.6}$$

Moreover, for ε and h sufficiently small, we have $u^*_{\varepsilon,h}\in B_\rho(u^*)$. Therefore, the quadratic growth condition (4.2) and (4.6) imply

$$\begin{aligned}\delta\|u^*_{\varepsilon,h}-u^*\|^2_{L^2(\Omega)}&\le g(u^*_{\varepsilon,h})-g_{\varepsilon,h}(u^*_{\varepsilon,h})+g_{\varepsilon,h}(u^*)-g(u^*)+g_{\varepsilon,h}(u^*_{\varepsilon,h})-g_{\varepsilon,h}(u^*)\\&\le|g(u^*_{\varepsilon,h})-g_{\varepsilon,h}(u^*_{\varepsilon,h})|+|g_{\varepsilon,h}(u^*)-g(u^*)|.\end{aligned} \tag{4.7}$$

We split the first term of (4.7) into two terms

$$|g(u^*_{\varepsilon,h})-g_{\varepsilon,h}(u^*_{\varepsilon,h})|\le|g(u^*_{\varepsilon,h})-g_\varepsilon(u^*_{\varepsilon,h})|+|g_\varepsilon(u^*_{\varepsilon,h})-g_{\varepsilon,h}(u^*_{\varepsilon,h})|. \tag{4.8}$$

For the first term in (4.8) we get

$$\begin{aligned}|g(u^*_{\varepsilon,h})-g_\varepsilon(u^*_{\varepsilon,h})|&=\frac12\left|\|S(u^*_{\varepsilon,h})-y_d\|^2_{L^2(\Omega)}-\|S_\varepsilon(u^*_{\varepsilon,h})-S(u^*_{\varepsilon,h})+S(u^*_{\varepsilon,h})-y_d\|^2_{L^2(\Omega)}\right|\\&\le\frac12\|S_\varepsilon(u^*_{\varepsilon,h})-S(u^*_{\varepsilon,h})\|^2_{L^2(\Omega)}\\&\quad+\|S_\varepsilon(u^*_{\varepsilon,h})-S(u^*_{\varepsilon,h})\|_{L^2(\Omega)}\|S(u^*_{\varepsilon,h})-y_d\|_{L^2(\Omega)}.\end{aligned}$$

The boundedness of $\{u^*_{\varepsilon,h}\}$ in $H^1(\Omega)$ by Lemma 4.8 and the Lipschitz continuity of the operator S (cf. Lemma 2.1) imply that $\|S(u^*_{\varepsilon,h}) - y_d\|_{L^2(\Omega)}$ is bounded. Hence, by Theorem 4.3 we obtain

$$|g(u^*_{\varepsilon,h}) - g_\varepsilon(u^*_{\varepsilon,h})| \leq C\varepsilon.$$

Analogously we obtain for the second term in Eq. (4.8) the estimate

$$\begin{aligned}|g_\varepsilon(u^*_{\varepsilon,h}) - g_{\varepsilon,h}(u^*_{\varepsilon,h})| \leq\ & \frac{1}{2}\|S_{\varepsilon,h}(u^*_{\varepsilon,h}) - S_\varepsilon(u^*_{\varepsilon,h})\|^2_{L^2(\Omega)} \\ &+ \|S_{\varepsilon,h}(u^*_{\varepsilon,h}) - S_\varepsilon(u^*_{\varepsilon,h})\|_{L^2(\Omega)}\|S_\varepsilon(u^*_{\varepsilon,h}) - y_d\|_{L^2(\Omega)}.\end{aligned}$$

Note that $\|S_\varepsilon(u^*_{\varepsilon,h}) - y_d\|_{L^2(\Omega)}$ is bounded due to the boundedness of $\{u^*_{\varepsilon,h}\}$ in $H^1(\Omega)$ and the Lipschitz continuity of the operator S_ε (cf. Lemma 2.3). Thus, Theorem 4.4 implies

$$|g_\varepsilon(u^*_{\varepsilon,h}) - g_{\varepsilon,h}(u^*_{\varepsilon,h})| \leq Ch^2|\log h|^2.$$

Applying the same arguments to the second term of (4.7) completes the proof. □

The previous theorem implies the following result:

Corollary 4.10. *Let $\varepsilon = Ch^2 \leq \varepsilon_d$ for $C > 0$ arbitrary. Under the assumptions that u^* satisfies the quadratic growth condition (4.2) and $\psi < 0$ on the boundary, there exists $h_d > 0$ such that for all $h \leq h_d$ it holds*

$$\|u^*_{\varepsilon,h} - u^*\|_{L^2(\Omega)} \leq Ch|\log h|.$$

Note that in the case of a suitable choice of $\bar\sigma$ the CAL problem formulation (3.5) and the nonsmooth penalty problem (2.5) (as well as their discrete pendants (3.14) and (4.1)) coincide, so that it remains to investigate error estimates for the discrete penalized CAL problem formulation (3.17) with increasing parameter ν.

Here one has to consider that the unsatisfied inequality constraint affects the argument by a bi-quadratic penalty of the violation multiplied by the penalty parameter ν. However, since this influence is compensated by the original objective functional $J(y,u)$, minimizing

$$J(y,u) + \nu \int_\Omega (\max(-\bar\sigma(\psi - y), 0))^4 \,\mathrm{d}x$$

can result in a solution that would not be feasible for the original optimal control problem, if the value of the penalty parameter ν is small relative to the original objective functional value $J(y,u)$. However, if the value of the penalty parameter ν is suitably large, the penalty for each violated constraint will increase the objective value so that minimizing the penalized objective functional will consequently yield a feasible solution for the non-smooth penalized optimal control problem (3.14). By increasing ν, the corresponding solution will therefore approach the feasible set of (3.14) and minimize the original objective functional $J(y,u)$. As a matter of principle, the solution corresponding to the penalized CAL problem then converges to a solution of (3.14) as $\nu \to \infty$.

5. Numerical results

In this section we test the performance of the algorithm CALi for the numerical solution of optimization problems of the form Eq. (2.2). For this purpose we present three different test examples taken from [9,12,21]. In all three examples the computational domain is chosen as the unit square $\Omega = (0,1)^2$ and for the fixed $\bar\sigma$ we use $\bar\sigma = \bar\sigma_\psi$ as introduced in Definition 3.4. For all tests we take $\nu = 1000$ and $y^0 = u^0 \equiv 0$ as initial guess for Newton's method. We initialize our ε-homotopy with $\varepsilon = 1.0$ and decrease the value of the penalization parameter constantly until the linear system in Newton's method is too ill-conditioned and Newton's method does not converge in under 20 steps, where an absolute error of 10^{-12} is pursued within the respective Newton procedure. For each $\varepsilon < 1.0$ we take the solution of the constant abs-linearized and penalized problem, i.e., Eq. (3.6), at the preceding value of ε as starting value for the current Newton iteration.

Besides the number of Newton steps, we also present the value $\left\|\bar\sigma z - |z|\right\|_{L^2(\Omega)}$, which measures the violation of the condition $\bar\sigma z = |z|$ with $z = \psi - y$. Furthermore, following [9] the value $\mu^- := \min_{k\in\mathcal{N}^-}(y_k - \psi_k)$, with $\mathcal{N}^- := \{k \in 1,\dots,n : y_k - \psi_k < 0\}$ is also documented, which denotes the violation of the obstacle constraint $y \geq \psi$ a.e. in Ω. As the penalty parameter ε decreases, μ^- typically should tend to zero.

We use the finite-element discretization introduced in Section 3. All the computations are done within the open source finite element environment FEniCS, version 2019.1.0, using the Python interface.

As a first example we consider once again Example 3.6.

Example 5.1. The obstacle problem is constructed with

$$y_d(x_1,x_2) = -\sin(\pi x_1)\sin(\pi x_2),\ \ f(x_1,x_2) = -2\pi^2\sin(\pi x_1)\sin(\pi x_2),\ \ \psi(x_1,x_2) = -0.25$$

We have already discussed that the optimal state is given by

$$y^* = \max(\psi, y_d) = \begin{cases}\psi, & \text{on } \Omega^+ := \{x \in \Omega : y_d(x) < \psi(x)\} \\ y_d, & \text{on } \Omega^- := \{x \in \Omega : y_d(x) \geq \psi(x)\}\end{cases}$$

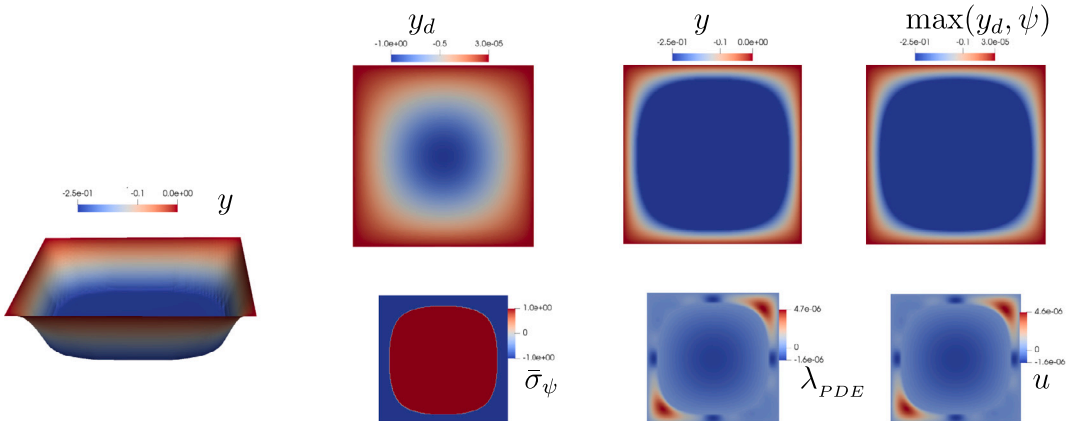

Fig. 4. Solution y with adjoint λ_{PDE} and control u for Example 5.1 with y_d and $\max(\psi, y_d)$ for $\bar{\sigma} \equiv \bar{\sigma}_\psi$.

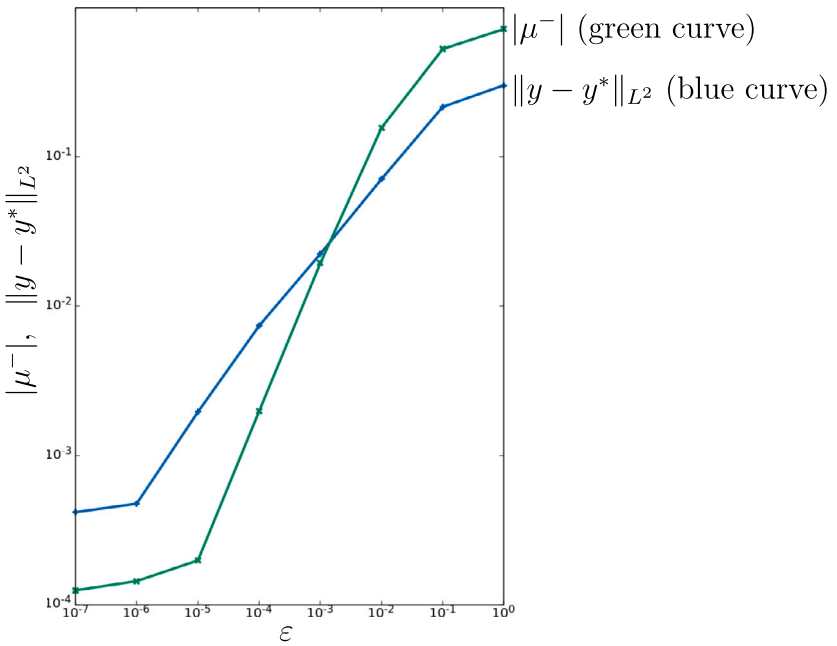

Fig. 5. Convergence plot of the obstacle violation $|\mu^-|$ (green curve) and the L^2-error for the state y (blue curve) for Example 5.1 with $h = 5.05\text{e-}03$ and $\alpha = 1.0$ for decreasing ε values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with $y^*|_{\partial\Omega} = 0$ and the optimal control is given by $u^* \equiv 0$. The numerical results for this example considering different values for the parameters α, ε and different mesh sizes are provided in Table 1. As expected, we see that μ^- tends to zero as ε decreases (see also Fig. 5) and the value $\|\bar{\sigma} z - |z|\|_{L^2(\Omega)}$ is quite small. Moreover, we observe that except for $\varepsilon = 1$ only 2 Newton steps are needed to solve the problem. Fig. 4 illustrates the desired state y_d, the fixed $\bar{\sigma}_\psi$, the solutions y, λ_{PDE} and u obtained with CALi using the parameters $\alpha = 1.0$, $\varepsilon = 1e-06$ and $h = 7.728e-03$ as well as the exact solution $y^* = \max(y_d, \psi)$.

Example 5.2. The data for the second example are chosen as:

$$y_d = y^* + \xi^* - \alpha \Delta y^*, \ f = -\Delta y^* - y^* - \xi^*, \ \psi = 0.0$$

with

$$y^* = \begin{cases} 160\left(x_1^3 - x_1^2 + 0.25x_1\right)\left(x_2^3 - x_2^2 + 0.25x_2\right), & \text{in } (0, 0.5)^2 \\ 0, & \text{else} \end{cases}$$

Table 1
Numerical results for Example 5.1.

h	α	ε	$\|y-y^*\|_{L^2(\Omega)}$	$\|\bar{\sigma}z-\|z\|\|_{L^2(\Omega)}$	μ^-	# Newton
3.009e−02	1.0	1.0	2.17e−01	1.3e−05	−5.295e−01	4
3.009e−02	1.0	1e−02	7.09e−02	1.8e−05	−1.570e−02	2
3.009e−02	1.0	1e−04	3.87e−03	2.2e−05	−1.935e−03	2
1.571e−02	1.0	1.0	2.16e−01	1.6e−06	−7.188e−01	4
1.571e−02	1.0	1e−02	7.09e−02	1.2e−04	−5.281e−02	2
1.571e−02	1.0	1e−04	1.55e−03	2.1e−05	−1.970e−03	2
7.728e−03	1.0	1.0	2.16e−01	6.6e−07	−7.188e−01	4
7.728e−03	1.0	1e−02	7.09e−02	1.4e−04	−5.281e−02	2
7.728e−03	1.0	1e−04	1.60e−03	2.4e−05	−1.970e−03	2
7.728e−03	1.0	1e−06	1.04e−03	3.0e−08	−7.130e−04	2
3.009e−02	1e−01	1.0	3.01e−01	1.3e−05	−7.181e−01	6
3.009e−02	1e−01	1e−02	7.09e−02	1.7e−04	−1.570e−02	2
3.009e−02	1e−01	1e−04	3.88e−03	2.3e−05	−1.969e−03	2
2.210e−02	1e−01	1.0	3.02e−01	2.7e−06	−7.193e−01	6
2.210e−02	1e−01	1e−02	7.10e−02	1.6e−04	−1.571e−02	2
2.210e−02	1e−01	1e−04	6.17e−03	2.0e−05	−1.971e−03	2
1.571e−02	1e−01	1.0	3.02e−01	1.5e−06	−7.194e−01	6
1.571e−02	1e−01	1e−02	7.10e−02	1.7e−03	−1.571e−02	2
1.571e−02	1e−01	1e−04	6.23e−03	2.1e−05	−1.971e−03	2
1.571e−02	1e−01	1e−06	4.44e−03	3.1e−08	−9.987e−04	2
7.728e−03	1e−01	1.0	3.02e−01	6.3e−07	−7.194e−01	6
7.728e−03	1e−01	1e−02	7.10e−02	1.7e−03	−1.570e−02	2
7.728e−03	1e−01	1e−04	7.24e−03	2.4e−05	−1.970e−03	2
7.728e−03	1e−01	1e−06	1.44e−03	3.0e−08	−7.130e−04	2
1.571e−02	1e−03	1.0	3.11e−01	9.2e−07	−7.397e−01	9
1.571e−02	1e−03	1e−02	8.37e−02	9.5e−04	−1.938e−01	2
1.571e−02	1e−03	1e−04	6.81e−03	2.1e−05	−1.978e−03	2
1.571e−02	1e−03	1e−06	4.44e−03	3.1e−08	−9.988e−04	2
1.571e−02	1e−05	1.0	3.16e−01	9.3e−07	−7.500e−01	10
1.571e−02	1e−05	1e−02	2.82e−01	1.3e−06	−6.679e−01	2
1.571e−02	1e−05	1e−04	6.13e−03	2.0e−05	−2.714e−03	2
1.571e−02	1e−05	1e−06	4.07e−03	3.2e−08	−1.006e−03	2

and

$$\xi^* = \max\Big(0, -2|x_1 - 0.8| - 2|x_1 x_2 - 0.3| + 0.5\Big)\,,$$

according to [4, Example 5.1]. Note that by construction the optimal control is $u^* = y^*$. The numerical results for this example considering different values for the parameters α, ε and different mesh sizes are provided in Table 2. Following [4] the results in Table 2 were computed with a Newton solver tolerance of $\frac{h^2}{2}$. We observe that for all combinations of parameter values except for $\varepsilon = 1.0$ only one Newton iteration is required to obtain a residual norm below 10^{-12}.

In [4] this example has the title "lack of strict complementarity" due to the fact that the active set at the solution contains a subset where strict complementarity fails to hold, i.e., the biactive set has a positive measure. It is precisely this lack of strict complementarity that poses a challenge, since the active constraint gradients at the solution are linearly dependent.

The numerical solutions for the parameters $\alpha = 1.0$ and $\varepsilon = 10^{-8}$ are displayed in Fig. 6. We would like to point out that the solution of the state in Fig. 6 differs from Figure 2 in [4] by a factor of 10^{-1} since the scaling factor is missing at the corresponding axis in [4, Figure 2].

Fig. 7 shows the decay of the obstacle violation $|\mu^-|$ for Example 5.2 with $h = 7.728\text{e-}03$ and $\alpha = 1.0$ for decreasing ε values in a log–log-scale.

The theoretical results of Section 4 show that the overall error consists of two contributions, namely the regularization error, i.e. $\|y_\varepsilon^* - y^*\|_{L^2(\Omega)}$ resp. $\|u_\varepsilon^* - u^*\|_{L^2(\Omega)}$, and the discretization error, i.e. $\|y_{\varepsilon,h}^* - y_\varepsilon^*\|_{L^2(\Omega)}$ resp. $\|u_{\varepsilon,h}^* - u_\varepsilon^*\|_{L^2(\Omega)}$. In order to numerically ascertain the approximation properties of the presented solution method, these errors were computed for different penalization parameters and different mesh sizes. Fig. 8 shows the convergence plot for the L^2-regularization-errors of the state and the control with fixed mesh size h and decreasing ε values for Example 5.2 in a log–log-scale. Representational for the numerical test, Fig. 8 suggests an approximation order of $\mathcal{O}(\varepsilon)$ for the L^2-error of the state and $\mathcal{O}(\varepsilon^{\frac{1}{2}})$ for the L^2-error of the control which confirms our theoretical results derived in Section 4.

Moreover, we observed that for a penalty parameter $\varepsilon \leq h^2$, the approximation error remained almost constant, i.e., the approximation error is dominated by the discretization error provided that ε is sufficiently small, see e.g., Fig. 9. This observation is in agreement with the theory of Section 4.

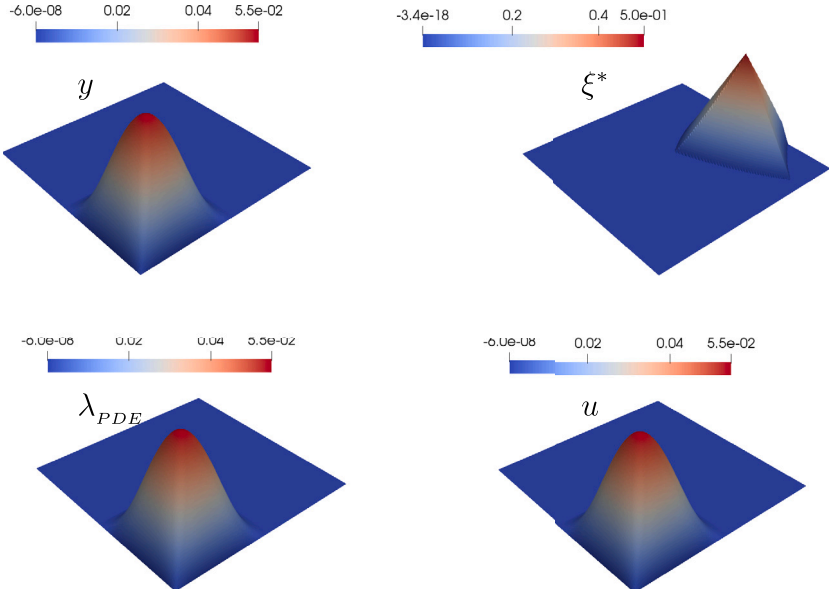

Fig. 6. Solution y (upper left), control u (lower right) and adjoint λ_{PDE} (lower left) for Example 5.2 with ξ^* (upper right) for $\alpha = 1.0$ and $\epsilon = 10^{-8}$.

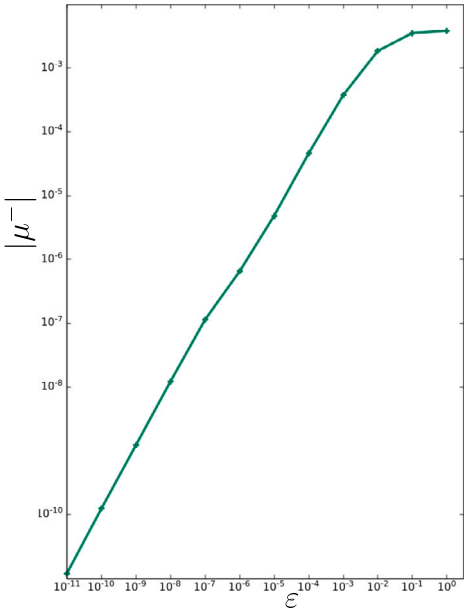

Fig. 7. Convergence plot of the obstacle violation for Example 5.2 with $h = 7.728\text{e-}03$ and $\alpha = 1.0$ for decreasing ϵ values.

Fig. 10 shows convergence plot of the L^2-error of the control u with $\alpha = 1.0$ for decreasing ϵ for different mesh sizes h. Fig. 11 shows convergence plots for the L^2-errors $\|y^*_{\epsilon,h} - y^*\|_{L^2(\Omega)}$ and $\|u^*_{\epsilon,h} - u^*\|_{L^2(\Omega)}$ for Example 5.2 in a log–log-scale for decreasing mesh size h and fixed parameter $\epsilon = 10^{-9}$. The observations in Fig. 11 suggest an approximation order of $\mathcal{O}(h)$ for the L^2-error of the control and $\mathcal{O}(h^2)$ for the L^2-error of the state. Similar observations were made for other test cases, such that once again the theoretically determined results of Section 4 have been numerically verified.

A successive approach with a heuristic switching strategy The following example has a special feature in contrast to the previous ones, since the function $\max(\psi, y_d)$ does not provide the optimal signature function $\bar{\sigma}$ and thus $\bar{\sigma}_\psi$ does not yield the constant abs-linearized (CAL) problem formulation which provides the optimal solution y^*. We present this example as an outlook for further research on optimal strategies for generating $\bar{\sigma}$ or a sequence of $\bar{\sigma}$-signature functions together with an efficient switching strategy such that the corresponding final CAL problem provides the optimal solution.

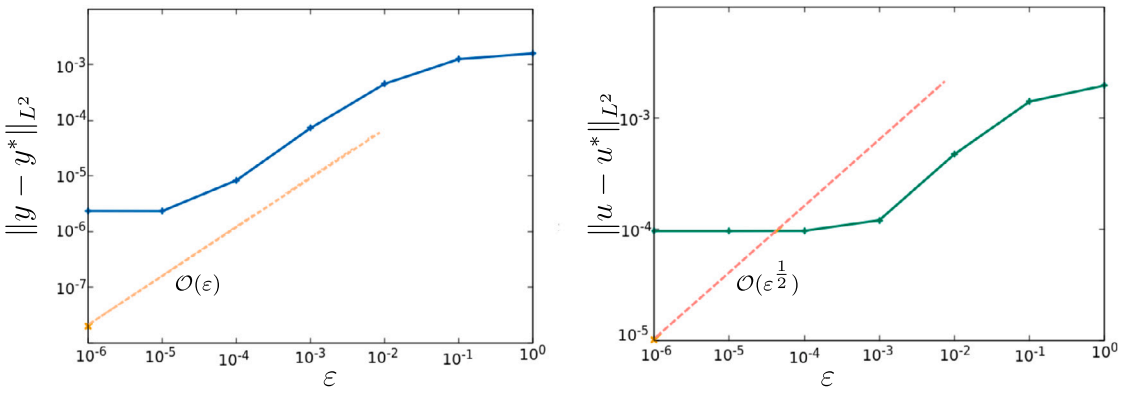

Fig. 8. Convergence plot of the L^2 error of the state (left, blue curve) and the control (right, green curve) for Example 5.2 with $\alpha = 1.0$ and $h = 6.4\text{e-}03$ for decreasing ε. The dashed lines show the rates of increase in magnitude, i.e., the approximation order with respect to ε for the L^2 error of the state (left, orange dashed line), which is of order $\mathcal{O}(\varepsilon)$ and the control (right, red dashed line), which is of order $\mathcal{O}(\varepsilon^{\frac{1}{2}})$, in the respective cases considered. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

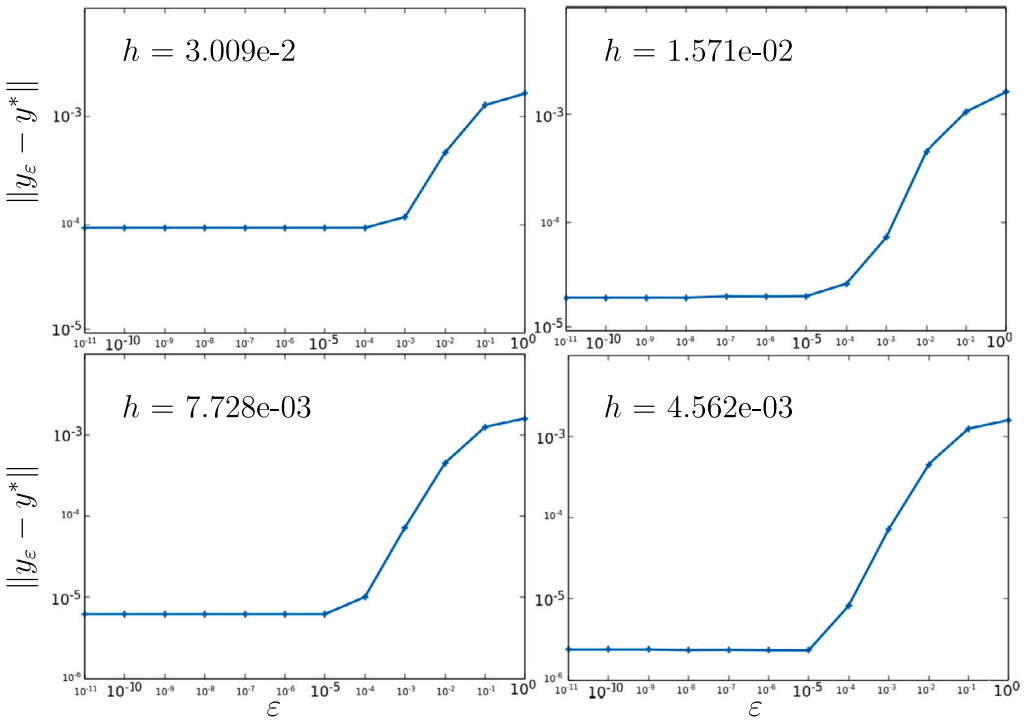

Fig. 9. Convergence plot of the L^2-error of the state y for Example 5.2 with $\alpha = 1.0$ for decreasing ε for different mesh sizes h.

However, this choice of $\bar{\sigma}$ is not always a priori straightforward or even goal-directed. In fact, the class of optimization problems where $\bar{\sigma}$ can be chosen based on the desired state is a special case. Thus, in other cases, an efficient successive switching of the sign choice in $\bar{\sigma}$ becomes necessary until an optimal $\bar{\sigma}$ is found which provides optimal solutions to the original problem. Such a switching strategy was developed and studied in [28]. In the following example, we adapt it and apply it in a purposeful manner for the here considered optimal control problems constrained by obstacle problems.

Example 5.3. For this example of an obstacle problem we choose the following data:

$$y_d(x_1, x_2) = -5x_1 - x_2 + 1,\ f(x_1, x_2) = -x_1 + 0.5,\ \psi(x_1, x_2) = 0.0\ .$$

This example corresponds to Example 2 from [9]. At this point it turns out that a choice of $\bar{\sigma}$ similar as before is not purposeful. Therefore, an adaptive strategy becomes necessary.

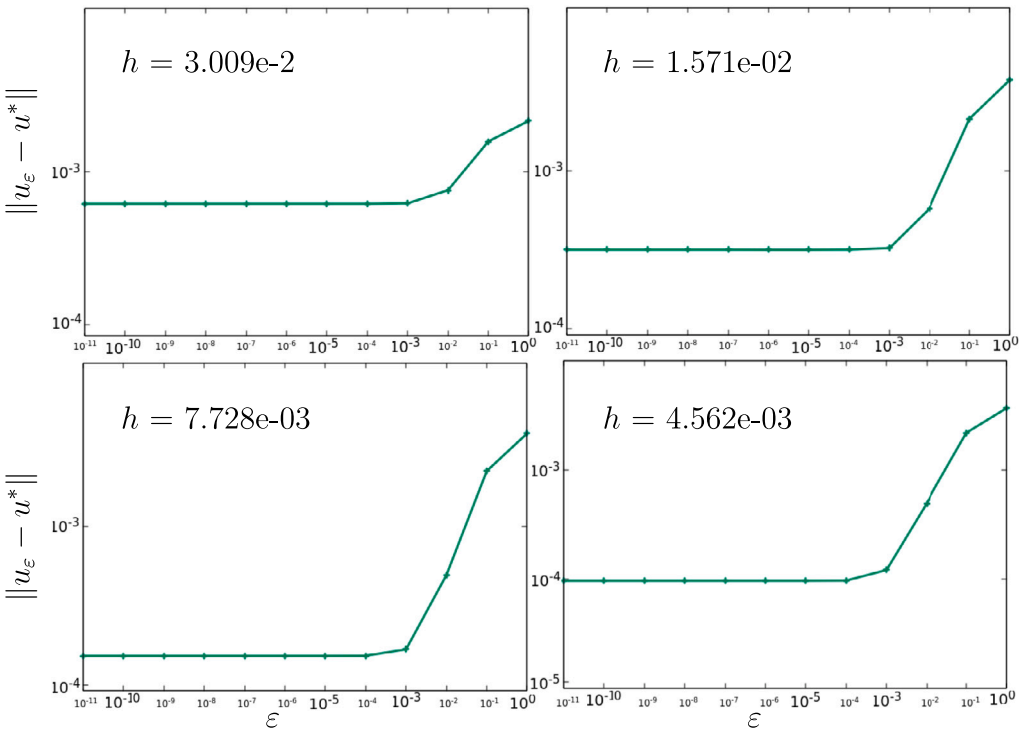

Fig. 10. Convergence plot of the L^2-error of the control u for Example 5.2 with $\alpha = 1.0$ for decreasing ε for different mesh sizes h.

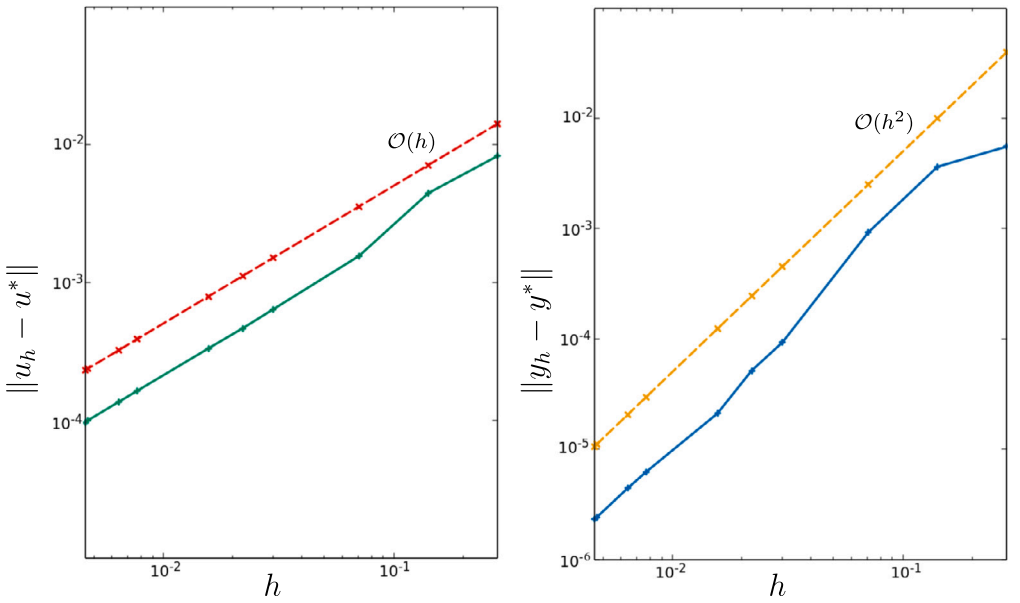

Fig. 11. Convergence plot of the discretization error for Example 5.2, i.e., L^2-error of the control u (left, green curve) and the state y (right, blue curve) with $\alpha = 1.0$ and $\varepsilon = 10^{-9}$ for decreasing mesh size h. The dashed lines show the rates of increase in magnitude, i.e., the approximation order with respect to the mesh size h for the control (left, red dashed line), which is of order $\mathcal{O}(h)$ and for the L^2 error of the state (right, orange dashed line), which is of order $\mathcal{O}(h^2)$, in the respective cases considered. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We follow and adapt here the heuristic strategy, presented in [28].

We initially choose $\bar{\sigma} = \bar{\sigma}_\psi$ and explore the optimality conditions in order to adopt $\bar{\sigma}$ and to find an optimal choice for $\bar{\sigma}$.

In [15,29] optimality conditions for the CAL problem are derived and a function $r(\bar{\sigma})$ is introduced, which especially provides information on whether the chosen $\bar{\sigma}$ leads to an optimal CAL problem formulation. The theory concerning the optimality conditions

Table 2
Numerical results for Example 5.2.

h	α	ε	$\|\bar{\sigma}z-\|z\|\|_{L^2(\Omega)}$	μ^-	# Newton
3.009e−02	1.0	1.0	1.4e−07	−5.162e−03	2
3.009e−02	1.0	1e−03	3.9e−11	−3.791e−04	1
3.009e−02	1.0	1e−06	2.2e−14	−1.437e−06	1
3.009e−02	1.0	1e−09	2.4e−20	−1.531e−09	1
3.009e−02	1.0	1e−12	2.4e−26	−1.373e−12	1
1.571e−02	1.0	1.0	7.5e−08	−5.220e−03	2
1.571e−02	1.0	1e−03	1.9e−11	−3.819e−04	1
1.571e−02	1.0	1e−06	1.3e−13	−4.688e−06	1
1.571e−02	1.0	1e−09	1.8e−19	−5.689e−09	1
1.571e−02	1.0	1e−12	2.5e−25	−6.061e−12	1
7.728e−03	1.0	1.0	1.2e−07	−5.019e−03	2
7.728e−03	1.0	1e−03	5.2e−11	−3.811e−04	1
7.728e−03	1.0	1e−06	1.4e−15	−1.390e−06	1
7.728e−03	1.0	1e−09	4.1e−21	−1.235e−09	1
7.728e−03	1.0	1e−12	4.1e−27	−1.236e−12	1
3.009e−02	1e−01	1.0	1.4e−08	−3.865e−03	2
3.009e−02	1e−01	1e−03	3.9e−11	−3.716e−04	1
3.009e−02	1e−01	1e−06	2.2e−14	−1.293e−06	1
3.009e−02	1e−01	1e−09	2.4e−20	−1.373e−09	1
3.009e−02	1e−01	1e−12	2.4e−26	−1.373e−12	1
2.210e−02	1e−01	1.0	2.2e−08	−3.751e−03	2
2.210e−02	1e−01	1e−03	7.6e−10	−3.768e−04	1
2.210e−02	1e−01	1e−06	2.4e−13	−5.291e−06	1
2.210e−02	1e−01	1e−09	2.8e−19	−5.876e−09	1
2.210e−02	1e−01	1e−12	3.7e−25	−6.124e−12	1
1.571e−02	1e−01	1.0	1.2e−08	−3.799e−03	2
1.571e−02	1e−01	1e−03	1.9e−11	−3.818e−04	1
1.571e−02	1e−01	1e−06	1.3e−13	−4.688e−06	1
1.571e−02	1e−01	1e−09	1.8e−19	−5.690e−09	1
1.571e−02	1e−01	1e−12	2.5e−25	−6.061e−12	1
7.728e−03	1e−01	1.0	1.9e−08	−3.900e−03	2
7.728e−03	1e−01	1e−03	1.0e−11	−3.816e−04	1
7.728e−03	1e−01	1e−06	1.7e−15	−9.199e−07	1
7.728e−03	1e−01	1e−09	4.1e−21	−1.235e−09	1
7.728e−03	1e−01	1e−12	4.1e−27	−1.236e−12	1
1.571e−02	1e−03	1.0	3.2e−07	−4.167e−03	6
1.571e−02	1e−03	1e−03	3.1e−11	−3.816e−04	1
1.571e−02	1e−03	1e−06	1.3e−13	−4.756e−06	1
1.571e−02	1e−03	1e−09	1.9e−19	−5.781e−09	1
1.571e−02	1e−05	1.0	4.5e−09	−4.568e−04	8
1.571e−02	1e−05	1e−03	9.6e−11	−3.557e−04	1
1.571e−02	1e−05	1e−06	1.9e−13	−5.187e−06	1
1.571e−02	1e−05	1e−09	2.7e−19	−6.385e−09	1

and the function $r(\bar{\sigma})$ can be adapted completely analogously to our setting. This leads to the function r defined by

$$r(\bar{\sigma}) := \lambda_P \frac{\partial \hat{\ell}(y,\bar{\sigma}z)}{\partial z_k} - \bar{\sigma}\lambda,$$

with Lagrange multiplier λ_P corresponding to the state equation with $\hat{\ell}$ and λ corresponding to the equality condition $(\psi - y = z)$ defining the switching function z. According to [15] the choice of the corresponding $\bar{\sigma}$ function is unfavorable, if $r(\bar{\sigma}) < 0$, since it does not lead to a CAL problem which also provides an optimal solution for the associated non-smooth optimization problem with PDE constraint (2.3). For details on the function $r(.)$, we refer to [15].

Hence for instances where $r(\bar{\sigma}) < 0$ one has to switch the signs of $\bar{\sigma}$ at the corresponding mesh points in order to approach an optimal $\bar{\sigma}$ and thus an optimal solution. In [28] a heuristic switching strategy was introduced. Accordingly, one switches the signs of $\bar{\sigma}$ exactly on those mesh points where the product $\bar{\sigma}_h\lambda_h$ is large. This leads to an iterative switching procedure of which an abstract illustration is given in Fig. 12.

Fig. 13 shows the evolution of the sign function $\bar{\sigma}$ by applying this heuristic strategy.

Indeed, calculations for the choice of $\bar{\sigma} = \bar{\sigma}^{\psi}$ show that $r(\bar{\sigma}^{\psi}) < 0$, e.g. $r(\bar{\sigma}^{\psi}) = -2.352202e+01$ for $\varepsilon = 1e-6$. Whereas, for $\bar{\sigma} = -\text{sign}(f)$, it holds that $r(\bar{\sigma}) \geq 0$, so e.g. $r(\bar{\sigma}) = 2.692266e+03$ for $\varepsilon = 1e-6$. Thus, it becomes clear that the choice $\bar{\sigma}^{\psi}$ is not goal-directed in every case, and thus further research is needed for the a priori optimal choice of $\bar{\sigma}$ or strategies for generating a sequence of $\bar{\sigma}$ functions that converges to a $\bar{\sigma}^*$ with corresponding optimal CAL problem in the sense that a computed solution is already optimal for the originally considered optimization problem constrained by an obstacle problem.

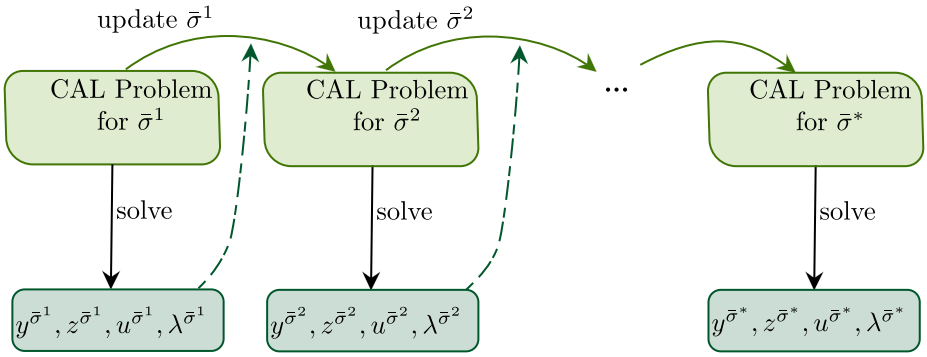

Fig. 12. Abstract illustration of the successive SCALi procedure via successive solution of CAL-problems [28].

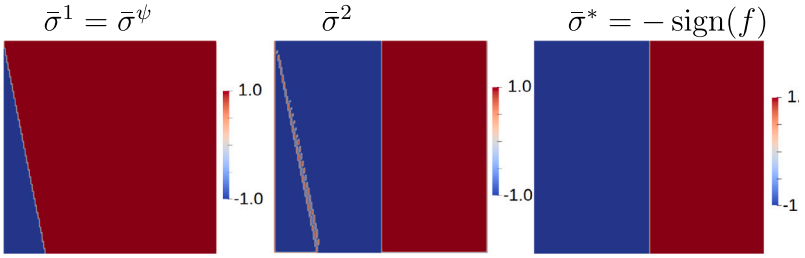

Fig. 13. Evolution of the sign function $\bar{\sigma}$ in Example 5.3 due to local sign-switchings according to the heuristic strategy [28].

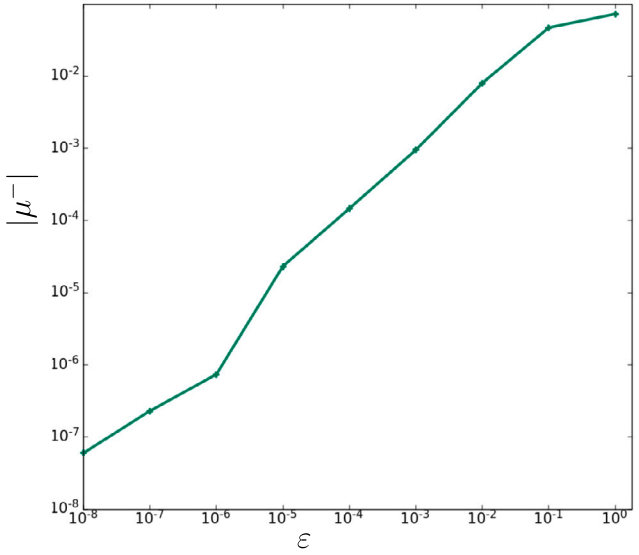

Fig. 14. Convergence plot of the absolute obstacle violation $|\mu^-|$ for decreasing ε for Example 5.3 with $\bar{\sigma}$ from Eq. (5.1) and $\alpha = 0.1$.

This heuristic swapping strategy yields as an optimal choice of $\bar{\sigma}$ as

$$\bar{\sigma}((x_1, x_2)) = \begin{cases} -1, & x_1 \leq 0.5 \\ +1, & \text{else,} \end{cases} \quad \text{for } (x_1, x_2) \in \Omega\,. \tag{5.1}$$

The numerical results for this example considering different values for the parameters α, ε and different mesh sizes are provided in Table 3. Once again, strict complementarity, however, is not satisfied, which makes this problem a further challenge.

Similar to [9] we observe a reduction in the absolute obstacle violation proportional to the reduction in the penalty parameter ε, see e.g., Fig. 14.

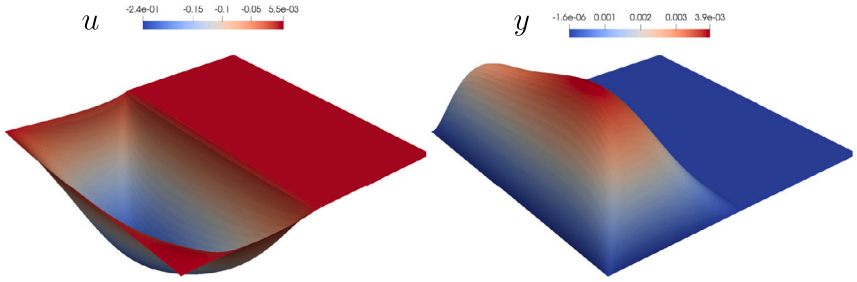

Fig. 15. Optimal control u and optimal state y for Example 5.3 with $\bar{\sigma}$ in Eq. (5.1) for $\alpha = 0.1$.

Table 3
Numerical results for Example 5.3 with $\bar{\sigma}$ from Eq. (5.1) and a Newton tolerance of 10^{-15}.

h	α	ϵ	$\|\bar{\sigma}z-\|z\|\|_{L^2(\Omega)}$	μ^-	# Newton
3.009e−02	1e−01	1.0	1.2e−03	−6.288e−02	10
3.009e−02	1e−01	1e−01	7.1e−04	−4.434e−02	4
3.009e−02	1e−01	1e−02	2.6e−05	−8.103e−03	4
3.009e−02	1e−01	1e−03	1.4e−07	−9.451e−04	3
3.009e−02	1e−01	1e−04	3.3e−10	−9.566e−05	2
3.009e−02	1e−01	1e−05	1.5e−12	−7.008e−06	2
3.009e−02	1e−01	1e−06	3.7e−14	−1.064e−06	2
1.571e−02	1e−01	1.0	1.1e−03	−6.290e−02	10
1.571e−02	1e−01	1e−01	6.9e−04	−4.409e−02	4
1.571e−02	1e−01	1e−02	2.3e−05	−7.672e−03	4
1.571e−02	1e−01	1e−03	8.4e−08	−7.816e−04	3
1.571e−02	1e−01	1e−04	9.1e−11	−5.458e−05	2
1.571e−02	1e−01	1e−05	1.4e−12	−5.906e−06	2
1.571e−02	1e−01	1e−06	5.8e−14	−1.626e−06	2
7.728e−03	1e−01	1.0	1.2e−03	−6.283e−02	10
7.728e−03	1e−01	1e−01	7.2e−04	−4.434e−02	4
7.728e−03	1e−01	1e−02	2.7e−05	−8.168e−03	4
7.728e−03	1e−01	1e−03	1.7e−07	−1.018e−03	3
7.728e−03	1e−01	1e−04	1.7e−09	−1.780e−04	2
7.728e−03	1e−01	1e−05	2.7e−11	−3.665e−05	2
7.728e−03	1e−01	1e−06	6.4e−14	−3.038e−06	2
4.714e−03	1e−01	1.0	1.2e−03	−6.283e−02	10
4.714e−03	1e−01	1e−01	7.1e−04	−4.426e−02	4
4.714e−03	1e−01	1e−02	2.6e−05	−8.023e−03	4
4.714e−03	1e−01	1e−03	1.4e−07	−9.521e−04	3
4.714e−03	1e−01	1e−04	1.0e−09	−1.464e−04	2
4.714e−03	1e−01	1e−05	8.6e−12	−2.326e−05	2
4.714e−03	1e−01	1e−06	3.0e−14	−7.392e−07	2

Fig. 15 shows the optimal control u as well as the optimal state y for $\bar{\sigma}$ from Eq. (5.1) and $\alpha = 0.1$.

Although the procedure presented here and the corresponding algorithm works impeccably for a large class of optimization problems with obstacle conditions and allows an explicit structure exploitation, which records a reduction in the number of required Newton steps, there are also optimization problems like Example 5.3 where the choice and fixation of $\bar{\sigma}$ does not lead directly to the optimal CAL problem without further analysis and effort.

In these cases the heuristic switching strategy proves to be useful, efficient and purposeful. However, a non heuristic approach and further development of the efficient strategy to switch between different $\bar{\sigma}$ and associated problem formulations constitutes subject of ongoing research.

6. Conclusion

We investigated a regularization approach for obstacle optimization problems, which results in optimal control problems constrained by a genuinely nonsmooth PDE. The presented and discussed solution method for this class of optimization problems is based on constant abs-linearization, which enables the optimization without any substitute assumptions for the nonsmoothness. The key idea is to generate a suitable reformulation of the nonsmooth PDE constrained regularized problem, the so-called constant abs-linearized problem, which can be solved using conventional methods for smooth optimization problems.

The type of discretization employed here was also presented and critically examined. Moreover, error estimates for the state and the control are derived, which contain information about the coupling of the regularization and the mesh size.

Finally, three different obstacle problems were considered and numerical results illustrating the performance of the presented algorithm were demonstrated and evaluated. The corresponding numerical results are very promising and also clearly confirm the theoretically derived error estimates.

The analysis and the numerical results have shown that it is useful as well as purposeful to rewrite the considered obstacle optimization problem by means of penalization and constant abs-linearization into a smooth but strongly related problem. By profoundly choosing the signature function $\bar{\sigma}$, one obtains a subproblem of the penalized nonsmooth problem, which is itself smooth and can be solved efficiently and effectively by means of standard optimization methods. In this paper we have shown that in certain cases the choice of $\bar{\sigma}_\psi = \operatorname{sign}(\psi - y_d)$ provides the optimal signature function. However, the choice of the signature function $\bar{\sigma}$ seems to be a delicate issue in general.

We therefore suggest and consider further research on a strategy for a generally advantageous choice of $\bar{\sigma}$ as well as a strategy for switching from one $\bar{\sigma}^i$, i.e., a concrete subproblem to the next $\bar{\sigma}^{i+1}$ by cleverly switching certain signs on certain areas in the underlying domain. By means of such a procedure, the constant abs-linearization can be successively applied. The resulting problems are smooth and can be solved as before with the usual methods of smooth optimization.

Another aspect of further research comprises the consideration of suitable regularization methods for optimization problems constrained by variational inequalities of the second kind into similar nonsmooth PDE constrained optimization problems and applying the algorithm CALi for their efficient solution.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

References

[1] Mignot F. Contrôle dans les inéquations variationelles elliptiques. J Funct Anal 1976;22(2):130–85.
[2] Mignot F, Puel J-P. Optimal control in some variational inequalities. SIAM J Control Optim 1984;22:466–76.
[3] Ito K, Kunisch K. Optimal control of elliptic variational inequalities. App Math Optim 2000;41(3):343–64.
[4] Hintermüller M, Kopacka I. Mathematical programs with complementarity constraints in function space: C- and strong stationarity and a path-following algorithm. SIAM J Optim 2009;20(2):868–902.
[5] Outrata J, Jarušek J, Stará J. On optimality conditions in control of elliptic variational inequalities. Set-Valued Var Anal 2011;19(1):23–42.
[6] Kunisch K, Wachsmuth D. Sufficient optimality conditions and semi-smooth Newton methods for optimal control of stationary variational inequalities. ESAIM Control Optim Calc Var 2012;18:520–47.
[7] Wachsmuth G. Strong stationarity for optimal control of the obstacle problem with control constraints. SIAM J Optim 2014;24(4):1914–32.
[8] Harder F, Wachsmuth G. Comparison of optimality systems for the optimal control of the obstacle problem. GAMM-Mitt 2018;40(4):312–38.
[9] Ali A, Deckelnick K, Hinze M. Global minima for optimal control of the obstacle problem. ESAIM Control Optim Calc Var 2020;26.
[10] Barbu V. Optimal control of variational inequalities. Boston: Pitman; 1984.
[11] Hintermüller M. Inverse coefficient problems for variational inequalities. ESAIM Math Model Numer Anal 2001;35:129–52.
[12] Hintermüller M, Kopacka I. A smooth penalty approach and a nonlinear multigrid algorithm for elliptic mpecs. Comput Optim Appl 2011;50:111–45.
[13] Schiela A, Wachsmuth D. Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints. ESAIM Math Model Numer Anal 2013;47(3):771–87.
[14] Meyer C, Thoma O. A priori finite element error analysis for optimal control of the obstacle problem. SIAM J Numer Anal 2013;51(1):605–28.
[15] Weiß O, Walther A. A structure exploiting algorithm for non-smooth semi-linear elliptic optimal control problems. Surv Math Appl 2022;17:139–79.
[16] Adams RA, Fournier JJF. Sobolev spaces. Pure and applied mathematics, 2nd ed. vol. 140, Amsterdam: Elsevier; 2003.
[17] Kinderlehrer D, Stampacchia G. An introduction to variational inequalities and their applications. Classics in applied mathematics, vol. 31, Philadelphia: Society for Industrial and Applied Mathematics (SIAM); 2000, Reprint of the 1980 original.
[18] Glowinski R, Lions J-L, Trémolières R. Numerical analysis of variational problems. Studies in mathematics and its applications, vol. 8, Amsterdam, New York: North-Holland Publishing Co.; 1981.
[19] Griewank A. On stable piecewise linearization and generalized algorithmic differentiation. Optim Methods Softw 2013;(6):1139–78.
[20] Griewank A, Walther A. First and second order optimality conditions for piecewise smooth objective functions. Optim Methods Softw 2016;(5):904–30.
[21] Meyer C, Rademacher A, Wollner W. Adaptive optimal control of the obstacle problem. SIAM J Sci Comput 2015;37(2):A918–45.
[22] Logg A, Wells G, Mardel K. In: Logg A, Mardal K, Wells G, editors. Automated solution of differential equations by the finite element method. Lecture notes in computational science and engineering, vol. 84, Berlin, Heidelberg: Springer Berlin Heidelberg; 2012.
[23] Brenner SC, Scott LR. The mathematical theory of finite element methods. Texts in applied mathematics, 3rd ed. vol. 15, New York: Springer; 2008.
[24] Hafemeyer D, Kahle C, Pfefferer J. Finite element error estimates in L^2 for regularized discrete approximations to the obstacle problem. Numer Math 2020;144(1):133–56.
[25] Nochetto RH. Sharp L^∞-error estimates for semilinear elliptic problems with free boundaries. Numer Math 1988;54:243–55.
[26] Casas E, Tröltzsch F. Error estimates for the finite-element approximation of a semilinear elliptic control problem. Control Cybern 2002;31:695–712.
[27] Christof C, Clason C, Meyer C, Walther S. Optimal control of a non-smooth semilinear elliptic equation. Math Control Relat Fields 2018;8(1):247–76.
[28] Weiß O. A heuristic extension to successive constant-abs-linearization for non-smooth elliptic PDE constrained optimization. Technical report, 2022.
[29] Weiß O. Non-smooth optimization by abs-linearization in reflexive function spaces [Ph.D. thesis], Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät; 2021, http://dx.doi.org/10.18452/24012.