A Structure Exploiting Algorithm for Non-Smooth Semi-Linear Elliptic Optimal Control Problems

Olga Weiß and Andrea Walther

Institut für Mathematik, Humboldt-Universität zu Berlin

September 29, 2021

Abstract

We investigate optimization problems with a non-smooth partial differential equation as constraint, where the non-smoothness is assumed to be caused by Nemytzkii operators generated by the functions abs, min and max. For the efficient as well as robust solution of such problems, we propose a new optimization method based on abs-linearization, i.e., a special handling of the non-smoothness with proficient exploitation of the non-smooth structure. 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 non-smooth optimization problems in very few Newton steps and additionally maintains reasonable convergence properties. Numerical results for non-smooth optimization problems illustrate the proposed approach and its performance.

Keywords: Non-Smooth Optimization, Constant Abs-Linearization, PDE Constrained Optimization, Non-Smooth PDE, Elliptic Optimal Control Problem

1 Motivation and Introduction

Non-smooth PDE-constrained optimization problems are known to be difficult to handle, theoretically as well as algorithmically. The challenge usually lies in the fact that no adjoint equation in the classical sense can be derived, which has a direct consequence on the development of algorithms, since no reduced gradient is available for first-order methods. In this paper we assume that the non-smoothness in the semi-linear elliptic state equation is caused by a non-smooth superposition operator which can be decomposed into a finite number of smooth functions and non-smooth Lipschitz-continuous operators abs, min and max. The presented algorithm takes advantage of this structural assumption and specifically exploits the non-smooth structure in the interest of solving the underlying optimization problem.
Non-smooth optimization problems with a partial differential equation (PDE) as constraint that involves the mentioned non-smooth non-linear functions arise in many modern applications. For example, a corresponding semi-linear elliptic partial differential equation describes the deflection of a stretched thin membrane partially covered by water, see [19]. Furthermore, a similar non-smooth partial differential equation arises in free boundary problems for a confined plasma, see, e.g. [19, 23]. Even nowadays, the optimization of such problems is challenging. Therefore, often either the non-smoothness is regularized, i.e., the non-differentiable term is replaced by a suitable smooth approximation to avoid dealing with the non-smoothness (see e.g. [4] and [10]) or the semi-smooth Newton method is used. For example, in [9] a variant of a semi-smooth Newton method is proposed to solve a specific non-smooth optimization problem including the max operator.

The intention of this paper is to propose an alternative idea for solving nonsmooth optimization problems. To motivate the algorithm presented in this paper, let us consider the following semi-linear non-smooth PDE constrained optimization problem with a typical tracking type objective functional presented in [9]:

$$\begin{aligned} \min_{(y,u)} \ & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + \ell(y) - u = f \ \text{ in } \Omega = (0,1)^2 \\ & y = 0 \ \text{ on } \partial\Omega \,, \end{aligned} \tag{1}$$

with desired state $y_d(x_1, x_2) = \sin(\pi x_1)\sin(2\pi x_2)$ and $\ell(y) = \max(0, y)$. Here $f \in L^2(\Omega)$ is chosen such that $-\Delta y_d + \max(0, y_d) = f$ holds. Obviously the non-differentiable operator $\max(0, y)$ represents a challenge for classical optimization methods. As presented in [9], due to the choice of y_d and f, the optimal state y^* coincides with y_d. Observing the desired state and hence also the optimal state, one can see that the sign of y_d, while only considering positive and negative signs and setting $\text{sign}(y_d)(x) = +1$ (or equally possible -1) for all $x \in \Omega$ with $y_d(x) = 0$ parts the domain $\Omega = (0,1)^2$ into two sections as can be seen in Fig. 1.

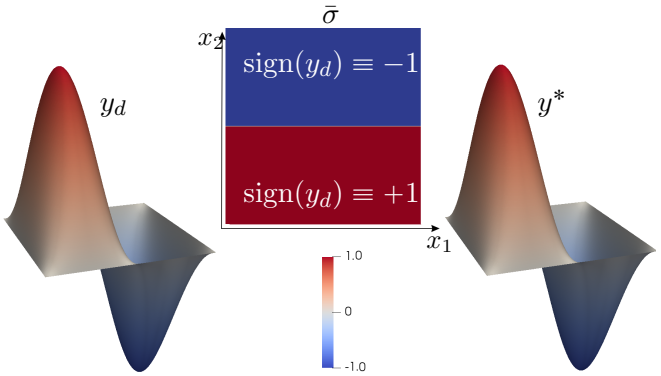


Figure 1: Desired state y_d with corresponding $\bar{\sigma} = \text{sign}(y_d)$ and the optimal state y^* for the optimal control problem (1).

This observation shows that the optimal state for Eq. (1) also optimally solves the following optimal control problem

$$\begin{aligned} \min_{(y,u)} \ & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + y - u = f \ \text{ in } \Omega^+ := (0,1) \times (0, \tfrac{1}{2}) \\ & -\Delta y - u = f \ \text{ in } \Omega^- := (0,1) \times (\tfrac{1}{2}, 1) \\ & y = 0 \ \text{ on } \partial\Omega \,, \end{aligned} \tag{2}$$

which can be abbreviated to

$$\begin{aligned} \min_{(y,u)} \ & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + \tfrac{1}{2}(y + \bar{\sigma} y) - u = f \ \text{ in } \Omega \\ & y = 0 \ \text{ on } \partial\Omega \,, \end{aligned} \tag{3}$$

where

$$\bar{\sigma}(x) = \text{sign}(y_d(x)) \quad \forall x \in \Omega$$

is defined by the desired state y_d (see Fig. 1) and thus a known parameter. However, considering the reformulation of the non-smooth max operator

$$\max(0, y) = \tfrac{1}{2}(y + |y|) = \tfrac{1}{2}(y + \text{sign}(y) y) \tag{4}$$

clearly reveals that the original non-smooth optimization problem Eq. (1) can be reformulated as Eq. (3) by Eq. (4) and fixing $\operatorname{sign}(y)$ to $\operatorname{sign}(y_d)$, which creates a smooth optimization problem, since the non-smooth dependency of $\operatorname{sign}(y)$ and y or $|y|$ is eliminated. Hence, the problem formulations given in Eqs. (2) and (3) are smooth in the conventional sense and thus do not raise the difficulties of the non-differentiable state equation given in Eq. (1). In fact, this kind of reformulation can be exploited in any tracking type optimization setting with certain properties of the non-smoothness in the state constraint, which we will utilize in this paper. The resulting fixing of the respective signs allows to effectively exploit the structure of the corresponding original non-smoothness and yields an optimization problem, which does not feature any non-smoothness. Motivated by this observation we present an efficient algorithm which allows to solve non-smooth optimization problems given by Eq. (1) in a proper way by exploiting the structure of non-smoothness of the operator ℓ.

In the finite dimensional setting the unconstrained minimization of piecewise smooth functions by successive abs-linearization without any regularization for the non-smoothness was studied by Griewank, Walther and co-authors in [11, 14, 15] and related work. There, it is always assumed that the non-smoothness of the considered optimization problem stems from evaluations of the absolute value function only. Using well-known reformulations, this covers the maximum and the minimum functions as well as complementarity problems. In [27] we already extended and adapted the algorithmic idea of the approach in finite dimensions to the infinite dimensional case, i.e., to PDE-constrained optimization problems with non-smooth objective functionals. Although the resulting algorithm SALMIN presented in [27] can also handle non-smooth optimization problems in function spaces by explicitly exploiting the non-smooth structure, it is not applicable to the optimization problems considered in this paper. The main difficulty involves the already mentioned challenge that the non-smoothness appears in the state equation and thus no adjoint equation in the classical sense as well as no classical reduced problem formulation can be derived. It is also important to note that the local model generated in [27] for the non-smooth case does not support the classical chain rule. Hence, one cannot directly handle the reduced unconstrained formulation. Therefore, we propose here a penalty-based approach to treat the PDE constraint explicitly. Nevertheless, we follow the idea for the finite dimensional case in that the key point of the optimization method under consideration is the location of stationary points by solving a closely related smooth problem exploiting the structure of the absolute value operator. Using classical methods for smooth PDE-constrained optimization.

The paper is organized as follows. In Sec. 2, we introduce the considered problem class, discuss its properties and propose a reformulation of the first order necessary optimality conditions. The closely related smooth problems will be presented in Sec. 3 together with a solution approach involving a penalty term and an analysis of the corresponding optimality conditions in the continuous setting. Sec. 4 summarizes the resulting optimization algorithm. Furthermore, the chosen discretization approach as well as the solution of the resulting finite dimensional optimization problems are discussed. Numerical results for a collection of test problems are presented and analyzed in Sec. 5. Finally, a conclusion and an outlook are given in Sec. 6.

2 The Problem Class, its Properties and a Reformulation

In order to illustrate our ideas, we consider the following optimization problem in the further course of this paper, where we focus on real valued functions defined on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n, n \in \mathbb{N}$. As a model problem we consider elliptic PDE-constrained optimization problems of the form

$$\begin{aligned} \min_{(y,u)\in H_0^1(\Omega)\times L^2(\Omega)} \quad & \mathcal{J}(y,u) \\ \text{s.t.} \quad & -\Delta y + \ell(y) - u = 0 \ \text{ in } \Omega \\ & y = 0 \ \text{ on } \partial\Omega\,, \end{aligned} \tag{5}$$

with a semi-linear elliptic PDE constraint and a tracking type objective functional

$$\mathcal{J} : H_0^1(\Omega) \to \mathbb{R},\ \mathcal{J}(y,u) := \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \ ,$$

where $y_d \in L^2(\Omega)$ denotes a given desired state. Hence, the functional $\mathcal{J}$ is twice continuously Fréchet differentiable, convex and bounded from below.
The special and at the same time challenging feature of Eq. (5) is caused by the non-smooth operator $\ell : H_0^1(\Omega) \to L^2(\Omega)$ in the state equation. Throughout the paper, we assume that the model problem Eq. (5) has the following properties:

Assumptions 2.1.

(i) The domain $\Omega \subseteq \mathbb{R}^n$ with $n \in \mathbb{N}$, is an open, bounded and measurable domain that is either convex and polygonal or has a $C^{1,1}$-boundary.

(ii) The constant $\alpha > 0$ is a given Tikhonov parameter.

For the special feature of the state equation, i.e., the non-smooth operator ℓ, we formulate an individual set of assumptions:

Assumptions 2.2.

(i) The operator $\ell : L^2(\Omega) \to L^2(\Omega), \ell(y)(x) = \ell(y(x))$ denotes an autonomous Nemytzkii operator induced by a non-linear and non-smooth function $\ell : \mathbb{R} \to \mathbb{R}$ which satisfies the Carathéodory conditions, i.e., the mapping $t \mapsto \ell(t)$ is continuous on $\mathbb{R}$.

(ii) The function $\ell : \mathbb{R} \to \mathbb{R}$ is monotonically increasing, and satisfies the growth condition

$$|\ell(t)| \leqslant K + L|t| \quad \text{for all } t \in \mathbb{R} \tag{6}$$

for some constants $0 \leqslant L, K < \infty$. Furthermore, ℓ is locally Lipschitz-continuous, i.e., for all constants $M > 0$ there exists a corresponding constant $L(M) > 0$ such that

$$|\ell(t_1) - \ell(t_2)| \leqslant L(M)|t_1 - t_2| \quad \text{for all } t_i \in [-M, M] \ . \tag{7}$$

(iii) The Nemytzkii operator ℓ is directionally differentiable in the sense of Hadamard, i.e.,

$$\left\| \frac{\ell(y + t\tilde{h}) - \ell(y)}{t} - \ell'(y;h) \right\|_{L^2} \to 0 \ , \tag{8}$$

for $t \to 0_+$ and $\tilde{h} \to h \in L^2(\Omega)$, at every $y \in H_0^1(\Omega)$ and in all directions $h \in L^2(\Omega)$, with $\ell'(y;.)$ being locally Lipschitz-continuous and monotone, as well.

(iv) The operator ℓ can be expressed as finite composition of the absolute value function and Fréchet differentiable operators.

Ass. 2.2 allows us to deal with nonlinear functions and their corresponding Nemytzkii operators. Note that the Nemytzkii operator is not assumed to be differentiable. Indeed all considered cases in this paper cover non-differentiable but Lipschitz-continuous Nemytzkii operators ℓ. Nevertheless, in the further course we will consider certain Nemytzkii operators derived from ℓ, and discuss their Fréchet differentiability. As already done in Ass. 2.2 (i) we will denote the Nemytzkii operator as well as the inducing function having other domain and image spaces with the same symbol, ℓ. The assumption Ass. 2.2 (iv) refers to the fact that the Lipschitz-continuous operator ℓ can be described by a so called *structured evaluation* presented in Ass. 2.8.

On the Operator ℓ

The well-definedness of Nemytzkii operators as well as their boundedness, continuity and differentiability depend substantially on the properties of the inducing function as well as on the considered function spaces. For a detailed study on Nemytzkii operators we refer to [3]. Nevertheless we want to point out, that the operator ℓ considered here fulfills the conditions of Thm. 3.1 in [3] due to Ass. 2.2, thus ensuring boundedness and continuity. Furthermore, a function $\ell : \mathbb{R} \to \mathbb{R}$ which satisfies Ass. 2.2 (i) and (ii) is bounded and satisfies the boundedness condition $|\ell(0)| \leqslant K$. Due to the assumptions on the generating function the Nemytzkii operator ℓ admits some relevant properties like Lipschitz-continuity and monotonicity, itself.

Proposition 2.3. *The Nemytzkii operator $\ell : L^2(\Omega) \to L^2(\Omega)$ induced by the function $\ell : \mathbb{R} \to \mathbb{R}$ which satisfies Ass. 2.2 (i) and (ii) is bounded, continuous and maximal monotone.*

Proof. To show the boundedness, we use both the growth condition and the equivalence of norms in finite dimensions. For $y \in L^2(\Omega)$ we have

$$\begin{aligned}\|\ell(y)\|_{L^2}^2 = \int_\Omega |\ell(y(x))|^2 \, dx \;&\leqslant \int_\Omega (L|y(x)| + K)^2 \, dx \\ &\leqslant \tilde{c} \int_\Omega L^2|y(x)|^2 + K^2 dx \leqslant \tilde{c} \left(L^2 \|y\|_{L^2}^2 + K^2|\Omega|\right) \, ,\end{aligned}$$

for a positive constant $\tilde{c}$. Consequently the Nemytzkii operator ℓ is bounded in $L^2(\Omega)$. To prove the continuity of $\ell : L^2(\Omega) \to L^2(\Omega)$, we consider some sequence $\{y_k\}_{k\in\mathbb{N}} \subseteq L^2(\Omega)$ with $y_k \to y$ in $L^2(\Omega)$. Hence, there exists some subsequence, denoted for simplicity also by y_k, with $y_k \to y$ almost everywhere on Ω. Then by exploiting the local Lipschitz-continuity (7), Lebesgue's dominated convergence theorem provides that the Nemytzkii operator ℓ is continuous on the reflexive Banach space $L^2(\Omega)$.
The monotonicity of the Nemytzkii operator follows directly from the monotonocity of the inducing function. From $(\ell(s) - \ell(t))(s - t) \geqslant 0$ for all $s, t \in \mathbb{R}$, one can conclude that $\langle \ell(y) - \ell(v), y - v \rangle_{L^2} = \int_\Omega (\ell(y(x)) - \ell(v(x)))(y(x) - v(x)) \, dx \geqslant 0$ holds for all $y, v \in L^2(\Omega)$, where we denote by $\langle \cdot, \cdot \rangle_{L^2}$ the duality pairing on the space $L^2(\Omega)$ and its dual. The Nemytzkii operator is hemicontinuous due to its continuity. Then it follows from [5] that the monotone and hemicontinuous Nemytzkii operator is maximal monotone. This proves the assertions. □

Due to the compact embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ the operator ℓ is also continuous from $H_0^1(\Omega)$ into $L^2(\Omega)$. With $\overline{B_M(0)} := \{y \in V : \|y\|_V \leqslant M\} \subset V$ we denote the closed ball with radius M in the considered Banach space V.

Proposition 2.4. *The Nemytzkii operator induced by the function ℓ which satisfies Ass. 2.2 (i) and (ii) is continuous in $L^\infty(\Omega)$ as well as in $L^2(\Omega)$ and locally Lipschitz for all $p \in [1, \infty]$, i.e.,*

$$\|\ell(y_1) - \ell(y_2)\|_{L^p} \leqslant L(M)\|y_1 - y_2\|_{L^p} \quad \forall y_i \in \overline{B_M(0)} \subset L^\infty(\Omega) \, , \tag{9}$$

with some $L(M) > 0$. Furthermore, the Nemytzkii operator $\ell = \ell(y)(x)$ is monotone in $y \in H_0^1(\Omega)$ for almost every $x \in \Omega$ and locally Lipschitz-continuous, such that for every constant $M > 0$ there exists some $\tilde{L}(M) > 0$ with

$$\|\ell(y_1) - \ell(y_2)\|_{L^2} \leqslant \tilde{L}(M)\|y_1 - y_2\|_{H^1} \quad \forall y_i \in \overline{B_M(0)} \subset H_0^1(\Omega) \, . \tag{10}$$

Proof. To prove the first assertion, suppose $y \in L^\infty(\Omega)$. From Eq. (6) we get for the Nemytzkii operator induced by ℓ that $\ell(y(\cdot)) \in L^\infty(\Omega)$. Hence, the considered Nemytzkii operator maps the Banach space $L^\infty(\Omega)$ into itself. For $y_1, y_2 \in L^\infty(\Omega)$ with $|y_1|, |y_2| \leqslant M$ almost everywhere on Ω, the local Lipschitz-continuity of ℓ implies for any $p \in [1, \infty)$

$$\begin{aligned}\|\ell(y_1) - \ell(y_2)\|_{L^p}^p &= \int_\Omega |\ell(y_1(x)) - \ell(y_2(x))|^p \, dx \\ &\leqslant L(M)^p \int_\Omega |y_1(x) - y_2(x)|^p \, dx = L(M)^p \|y_1 - y_2\|_{L^p}^p \, .\end{aligned}$$

Since $y_1, y_2 \in \overline{B_M(0)} \subset L^\infty(\Omega)$ were chosen arbitrary, this proves Eq. (9) for $p \in [1, \infty)$. For $p = \infty$ the assertion follows similarly.
The monotonicity of the Nemytzkii operator ℓ follows directly from

$$\langle \ell(y_1) - \ell(y_2), y_1 - y_2 \rangle = \int_\Omega (\ell(y_1(x)) - \ell(y_2(x)))(y_1(x) - y_2(x))\, dx \geqslant 0 \tag{11}$$

since $\ell(t_1) - \ell(t_2)(t_1 - t_2) \geqslant 0$ for all $t_1, t_2 \in \mathbb{R}$.
Due to the continuous embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ there exists some positive constant c such that $\|y\|_{L^2} \leqslant c\|y\|_{H^1}$ for all $y \in H^1(\Omega)$ and therefore $p = 2$ in Eq. (9) yields

$$\|\ell(y_1) - \ell(y_2)\|_{L^2} \leqslant \tilde{L}(M)\|y_1 - y_2\|_{L^2} \leqslant c\tilde{L}(M)\|y_1 - y_2\|_{H^1} \quad \forall y_1, y_2 \in \overline{B_M(0)} \subset H_0^1(\Omega)\ . \tag{12}$$

□

Existence of Solutions and Well-Posedness

Lemma 2.5. *For every $u \in L^2(\Omega)$ the PDE of the optimization problem (5) is non-linear, well posed and has a unique solution $y \in H_0^1(\Omega)$. If $n < 4$ then the unique solution to Eq. (5) is even continuous, i.e., $y \in H_0^1(\Omega) \cap C(\bar{\Omega})$.*

Proof. The proof applies standard arguments for monotone operators. For a detailed proof we refer the reader to [28]. Note that the existence of an unique solution $y \in H_0^1(\Omega)$ follows from the assumptions on the generating function ℓ with the Browder–Minty theorem. The boundedness and continuity of the solution follows from [8, Thm. 2.1] and [12]. □

Proposition 2.6. *The optimal control problem (5) admits at least one solution.*

Proof. See [28] and [24]. □

Due to our assumptions, the weak solution of the semi-linear PDE in Eq. (5) lies in the space

$$\mathcal{H}_\Delta := \{v \in H_0^1(\Omega) | \Delta v \in L^2(\Omega)\}\ . \tag{13}$$

The Control-to-State Operator

The solution or control-to-state operator $S(u) = y$ corresponding to the non-smooth state equation given in Eq. (5) plays an important role in the analysis of the overall optimization problem. Therefore we will state here some main properties and results:

Lemma 2.7 (The solution operator)**.** *Let $\ell : H_0^1(\Omega) \to L^2(\Omega)$ be a non-smooth operator satisfying Ass. 2.2 and $S(u) = y$ the solution operator associated with the PDE in Eq. (5). Then S has the following properties.*

i. The control-to-state operator $S(u) = y$ associated with Eq. (5) is a non-smooth operator.

ii. S is well-defined, bijective and globally Lipschitz-continuous as a function from $L^2(\Omega)$ to its image space $\mathcal{H}_\Delta$.

iii. S is directionally differentiable in the sense that the directional derivative defined by

$$S'(u; h) \equiv \lim_{t \to 0_+} \tfrac{1}{t}(S(u + th) - S(u)) \tag{14}$$

exists for all $u \in L^2(\Omega)$ and all directions $h \in L^2(\Omega)$. However S is not Gâteaux differentiable.

iv. $S : L^2(\Omega) \to \mathcal{H}_\Delta$ is Hadamard directional differentiable for all points and in all directions in $L^2(\Omega)$.

Proof. For the proof of these assertions, one can use similar arguments as in [9, Prop. 2.1]. For a detailed proof we refer the reader to [28]. □

As described in [9], the image space $\mathcal{H}_\Delta$ of the solution operator S equipped with the norm $\|\cdot\|_{\mathcal{H}_\Delta}$ induced by the scalar product

$$(v, w)_{\mathcal{H}_\Delta} := \int_\Omega \Delta v \Delta w + \nabla v \cdot \nabla w + vw \, dx$$

constitutes a complete normed vector space and thus a Hilbert space which in addition is also compactly embedded in $H_0^1(\Omega)$, due to the compact embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. Moreover, by regularity theory for elliptic differential operators, the space $\mathcal{H}_\Delta$ is isomorphic to $H_0^1(\Omega) \cap H^2(\Omega)$ (cf. [12, Lem. 9.17]). Furthermore if $n < 4$, $\mathcal{H}_\Delta$ is continuously embedded into $H_0^1(\Omega) \cap C(\bar{\Omega})$. Due to Lem. 2.7 (ii) the Hilbert space $\mathcal{H}_\Delta$ represents the appropriate image space for the considered solution operator S.
It should be noted that the algorithm proposed in Sec. 4 of this paper is not limited to this specific class of semi-linear PDEs. Instead, the arguments can easily be adapted to more general cases with, for example, a general linear elliptic differential operator of second order.
In addition to the assumptions on the non-smooth state equation given in Ass. 2.2, it can easily be observed that the tracking type objective functional $\mathcal{J} : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ in Eq. (5) is weakly lower semi-continuous and twice continuously Fréchet differentiable.
One particular example of this class of model problems of non-smooth semi-linear elliptic optimal control problems, where $\ell(y) = \max(0, y)$, can be found in [9]. There the authors also show that the resulting non-smooth control-to-state operator is directionally differentiable. They also precisely characterize its Bouligand subdifferentials, derive first-order optimality conditions using the Bouligand subdifferentials and use the directional derivative of the control-to-state mapping to establish strong stationarity conditions.
Although the objective functional itself is convex, the optimization problem (5) is not convex, which is why the existence of several locally optimal controls has to be taken into account. Furthermore, due to the non-convexity of the above optimal control problem, necessary first-order optimality conditions are no longer sufficient and the consideration of sufficient second-order optimality conditions becomes necessary if one wants to compute an actual minimizer. However, in this paper we limit our considerations to stationary points and an alternative way of ensuring a minimum, hence second-order conditions will not be the subject of this paper.
For the optimization we have to take into account that it is usually not possible to realize arbitrary controls $u \in L^2(\Omega)$. Therefore control constraints yielding a bounded and convex set of admissible controls can be introduced into the model problem. In addition to the methods presented here, the handling of such control constraints may include standard optimal control methods for control constraints [24] or the application of an additional penalty term similar to Eq. (46). However, this is not directly dealt with in this paper.
From now on we additionally assume $n < 4$ for $\Omega \subseteq \mathbb{R}^n$. Note that so far all of the above statements are valid regardless of the dimension n, except for the continuity of the solution of the semi-linear PDE, see Lem. 2.5.

Reformulating the PDE Constraint

Next, we introduce an essential reformulation of the PDE constraint based on the idea described in [13, 14]. For this purpose, we consider the Nemytzkii operator ℓ which is defined by the non-linear part of the PDE. Inspired by the finite dimensional approach of Griewank and Walther, we assume that the non-smooth operator ℓ can be described as a composition of elemental functions that are either continuously Fréchet differentiable or the absolute value operator. Subsequently, consecutive continuously Fréchet differentiable elemental functions can be conceptually combined to obtain a representation, where all evaluations of the absolute value functions can be clearly identified and exploited, see Ass. 2.8. For an illustration we refer to Exam. 2.9.

Assumption 2.8 (Structured evaluation)**.** *Let $\ell : H_0^1(\Omega) \to L^2(\Omega)$ be some non-smooth Lipschitz-continuous operator satisfying Ass. 2.2 and $s \in \mathbb{N}$. Furthermore, let $\psi_1 : H_0^1(\Omega) \to H^1(\Omega)$, $\psi_i : H_0^1(\Omega) \times [H^1(\Omega)]^{i-1} \to H^1(\Omega)$ with $2 \leqslant i \leqslant s+1$ be some Lipschitz-continuously Fréchet differentiable operators. We assume that the considered non-smooth operator ℓ can be*

reformulated such that an equivalent representation of ℓ *denoted by* $\hat{\ell}$ *can be obtained using the* structured evaluation *given by*

$$\left.\begin{array}{rcl} z_i & = & \psi_i(y, (\sigma_j z_j)_{j<i}) \\ \sigma_i & = & \operatorname{sign}(z_i) \end{array}\right\} i = 1, \ldots, s$$
$$\hat{\ell}(y, \sigma z) = \psi_{s+1}(y, (\sigma_i z_i)_{1 \leqslant i \leqslant s}) \quad \textit{with} \quad \sigma z = (\sigma_1 z_1, \ldots, \sigma_s z_s) \,,$$

where the σ_i *are called* signature functions.

Note that the functions z_i depend on the state y. Hence, whenever the dependency of a different but explicit function v is meant, we will simply write $z_i(v)$.
Let us point out, that the switching functions z_i in the structured evaluation procedure are introduced as auxiliary functions for each argument of the absolute value operator, which appear within the operator ℓ. The absolute value operator is then substituted by $\sigma_i z_i$. By applying this successively for all absolute value operators and their arguments one obtains an equivalent reformulation of the original operator ℓ in terms of y and $\sigma_i z_i$.
It is worth emphasizing that whenever the structured evaluation is applicable it provides an equivalent reformulation for the operator ℓ given by the operator $\hat{\ell}$. Therefore, $\hat{\ell}$ maps also into the same function space as ℓ, hence in the case under consideration into the Hilbert space $L^2(\Omega)$. Likewise, $\hat{\ell}$ also satisfies all assumptions in Ass. 2.2 itself.
Note that Ass. 2.8 ensures that the underlying non-smoothness does not contain any implicit dependencies for the switching functions z_i, such that the structured evaluation presented above can always be applied to obtain an equivalent reformulation. However the notation $(\sigma_j z_j)_{j<i}$ indicates that ψ_i might depend explicitly on the previously defined switching functions z_j with $j < i$. Hence, the switching function z_1 is defined as the argument of the first absolute value evaluation, i.e., as $\psi_1(y)$.
In the finite dimensional case, one has $z_i \in \mathbb{R}$ and therefore $\sigma_i \in \{-1, 0, 1\}$. For the infinite dimensional setting considered here, one obtains $z_i \in H^1(\Omega)$ and the functions σ_i are also Nemytzkii operators defined by

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

as functions of z_i. From now on we will omit the argument z_i whenever referring to σ_i, for brevity. The definition of σ_i ensures that $\sigma_i = \sigma_i(z_i) \in L^\infty(\Omega)$ holds. However, we recall that $\sigma_j z_j = \operatorname{abs}(z_j) \in H^1(\Omega)$ for the function $z_j \in H^1(\Omega)$ defined by the corresponding structured evaluation. Additionally, we will use the notation $\hat{\ell}(y, \sigma z) = \ell(y)$ for $\sigma z = (\sigma_1 z_1, \ldots, \sigma_s z_s)$ to refer explicitly to this particular representation of the non-smooth part $\ell(y)$ based on the corresponding auxiliary functions z_i and σ_i, $1 \leqslant i \leqslant s$ defined by the structured evaluation. It follows from the representation in Ass. 2.8 that ℓ is locally Lipschitz-continuous. Hence, ℓ and therefore also the equivalent $\hat{\ell}(y, \sigma z)$ are also continuous due to the assumed smoothness of ψ_i, $i = 1, \ldots, s$, [18, Theo. 3.15] and [29, Cha. 1]. It is important to note, that the new function $\hat{\ell}(y, w)$ is smooth i.e., Fréchet differentiable, in its two arguments y and $w = \sigma z = |z|$, due to the chosen formulation. This fact will be exploited later to define the closely related smooth problems.
Using the well-known reformulations

$$\begin{array}{l} \min(v, u) = (v + u - \operatorname{abs}(v - u))/2 \quad \text{and} \\ \max(v, u) = (v + u + \operatorname{abs}(v - u))/2 \,, \end{array} \tag{15}$$

a large class of non-smooth functions is covered by this function model.
Exam. 2.9 provides an illustration of the structured evaluation and also deals with the image and value spaces of the operators ψ_i.

Example 2.9. *We examine the commonly used non-smooth operator* max *as well as an operator which features nested non-smoothness in the form of nested absolute value functions.*

(i) *Consider the operator* $\ell(y) = \max(0, y)$*, which satisfies the Carathéodory condition and just like* abs(.) *is globally Lipschitz-continuous with Lipschitz constant* $L = 1$ *so that the associated Nemytskii-operator maps* $L^2(\Omega)$ *into itself. Exploiting the identities* (15)*, we can reformulate* ℓ *as a function in terms of the absolute value function and smooth elemental functions in the following way:*

$$\ell(y) = \max(0, y) = \tfrac{1}{2}(y + |y|) \, .$$

The corresponding structured evaluation for $\ell(y) = \max(0, y)$ *is given by*

$$\begin{array}{lll} z_1 & = \psi_1(y) & = y \\ \sigma_1 & = \operatorname{sign}(z_1) & \\ \hat{\ell}(y, \sigma z) & = \psi_2(y, \sigma z) & = \tfrac{1}{2}\big(y + \sigma_1 z_1\big) \end{array}$$

with $\psi_1 : H_0^1(\Omega) \to H_0^1(\Omega), y \mapsto y$ *and* $\psi_2 : H_0^1(\Omega) \times H_0^1(\Omega) \to H_0^1(\Omega), (y, v) \mapsto \tfrac{1}{2}\Big(y + v\Big)$.

(ii) *The non-smooth operator* $\ell(y) = -\max(|y - \beta_1| - \beta_2 y, 0)$ *admits nested absolute value operators. Nevertheless, for* $\beta_1, \beta_1 \geqslant 1$ *the inducing function* ℓ *is bounded, measurable, monotone and globally Lipschitz and hence also satisfies all assumptions required in Ass. 2.2. Then the corresponding structured evaluation is given by*

$$\begin{array}{lll} z_1 & = \psi_1(y) & = y - \beta_1 \\ \sigma_1 & = \operatorname{sign}(z_1) & \\ z_2 & = \psi_2(y, \sigma_1 z_1) & = \sigma_1 z_1 - \beta y \\ \sigma_2 & = \operatorname{sign}(z_2) & \\ \hat{\ell}(y, \sigma z) & = \psi_3(y, \sigma z) & = \tfrac{1}{2}(\bar{\sigma}_1 z_1 - \beta_2 y + \bar{\sigma}_2 z_2) \end{array}$$

with $\psi_1 : H_0^1(\Omega) \to H^1(\Omega), y \mapsto y - \beta_1$*,* $\psi_2 : H_0^1(\Omega) \times H^1(\Omega) \to H^1(\Omega), (y, v) \mapsto v - \beta_2 y$ *and* $\psi_3 : H_0^1(\Omega) \times H^1(\Omega) \times H^1(\Omega) \to H^1(\Omega), (y, v, w) \mapsto \tfrac{1}{2}(v - \beta_2 y + w)$.

Next we state the equivalent reformulation of the optimal control problem (5) by means of the structured evaluation. Inserting the formulation $\hat{\ell}(y, \sigma z)$ with the functions σ_i and z_i defined by the structured evaluation of ℓ into the original optimal control problem (5), one obtains for the functions $y \in H_0^1(\Omega), z \in [H^1(\Omega)]^s$, $u \in L^2(\Omega)$ and $\sigma_i = \operatorname{sign}(z_i)$ the optimization problem with state constraints

$$\begin{aligned} \min_{y,z,u,\sigma} \quad & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad (\nabla v, \nabla y)_{L^2} + (\hat{\ell}(y, \sigma z) - u, v)_{L^2} &= 0 \quad \forall v \in H_0^1(\Omega) \\ (\psi_i(y, (\sigma_j z_j)_{j<i}) - z_i, v)_{L^2} &= 0 \quad \forall v \in H_0^1(\Omega) \\ \sigma_i z_i &\geqslant 0 \quad \text{a.e. in } \Omega \\ \sigma_i : \Omega &\to \{-1, 0, 1\} \end{aligned} \Bigg\} \; \forall \; i = 1, \ldots, s \, . \tag{16}$$

Here, $[H^1(\Omega)]^s$ denotes the product $H^1(\Omega) \times \cdots \times H^1(\Omega)$ of the Hilbert spaces the switching function $z = (z_1, \ldots, z_s)$ lives in. Note that the last two conditions in Eq. (16) ensure that $\sigma_i z_i = \operatorname{abs}(z_i)$ holds.
At this point, the idea may arise that by reformulating and including functions with binary values, the underlying optimization problem was made considerably more difficult and complicated, since all derivative-based algorithms are not applicable due to the lack of continuity and differentiability of the integer-valued functions. However, as we will see in the further course of this paper and was already motivated in Sec. 1, we pursue a rather different strategy to exploit an efficient and proficient handling of the absolute value operator based on the reformulation given by the structured evaluation.

Definition 2.10 (Constant Abs-Linearization)**.** *For a given structured evaluation and the resulting operator* $\hat{\ell}$ *described in Ass. 2.8 the* constant abs-linearization *is obtained by fixing all* σ_i *for* $1 \leqslant i \leqslant s$ *to given* $\bar{\sigma}_i \in L^2(\Omega)$*,* $\bar{\sigma}_i : \Omega \to \{-1, 1\}$ *for all* $1 \leqslant i \leqslant s$*.*

In the context of constant abs-linearization, all $\bar{\sigma}_i$ only take the values 1 and -1, but not 0. This leads to the desired linearization since the dependency of z_i and σ_i has been removed and the term $\bar{\sigma}_i z_i$ is linear in z_i. The decision to assign only the values $+1$ or -1 to $\bar{\sigma}(x)$ does not influence the previous considerations, simply because if $z_i > 0$ and $\sigma_i = +1$, or $z_i < 0$ and $\sigma_i = -1$ respectively, then $\sigma_i z_i = \mathrm{abs}(z_i)$ is still valid. If $z_i = 0$, then even for $\sigma_i \neq 0$ the relationship $\mathrm{abs}(z_i) = \sigma_i z_i = 0$ is guaranteed. Hence, no longer considering zero as a value for $\bar{\sigma}$ does not pose any limitations.

Definition 2.11. *Let $\ell : H_0^1(\Omega) \to L^2(\Omega)$ be some Lipschitz-continuous operator satisfying Ass. 2.2, $s \in \mathbb{N}$ and $\hat{\ell}$ the resulting operator of structured evaluation described in Ass. 2.8. Then a given function y_d defines the switching function*

$$\begin{aligned} z(y_d) &:= (z_1(y_d), \dots, z_s(y_d)) \text{ with} \\ z_i(y_d) &:= \psi_i(y_d, (\sigma_j z_j(y_d))_{j<i}) \, . \end{aligned}$$

For some $y_d \in L^2(\Omega)$ we denote by

$$\begin{aligned} \bar{\sigma}^d &:= \sigma(z(y_d)) = (\sigma_1(z_1(y_d)), \dots, \sigma_s(z_s(y_d))) \\ &=: (\bar{\sigma}_1^d, \dots, \bar{\sigma}_s^d) \end{aligned}$$

the signature functions defined by the desired state according to the corresponding structured evaluation.

Next, we introduce the domain decomposition given by the signature functions $\bar{\sigma}^d$. To illustrate the approach we will first consider the domain decomposition given by the sign of the desired state y_d for the case $s = 1$ and $z_1 = y$

$$\bar{\sigma}^d = \mathrm{sign}(z_1(y_d)(x)) = \begin{cases} +1, \text{ for } x \in \Omega_d^+ := \{x \in \Omega | y_d(x) \geqslant 0\}, \\ -1, \text{ for } x \in \Omega_d^- := \{x \in \Omega | y_d(x) < 0\} \, . \end{cases} \tag{17}$$

Hence, for the case $s = 1$ the $\bar{\sigma}$ defined by Eq. (17) decomposes the domain Ω into subdomains such that $\overline{\Omega} = \overline{\Omega^+} \cup \overline{\Omega^-}$ with $\bar{\sigma}^d(x) = +1$ on Ω^+ and $\bar{\sigma}^d(x) = -1$ on Ω^-. The same applies for the case $s > 1$, i.e. the case with more than one absolute value evaluation in the formulation of ℓ. Let us consider the following example.

Example 2.12 (Domain Decomposition by $\bar{\sigma}^d$)**.**
We consider again $\Omega = (0,1)^2$ and

$$\ell(y) = -\max(|y - \beta_1| - \beta_2 y, 0) = -\tfrac{1}{2}(|y - \beta_1| - \beta_2 y + \big||y - \beta_1| - \beta_2 y\big|) \, ,$$

with $\beta_1, \beta_2 \geqslant 1$. Then by Ex. 2.9(ii) $\hat{\ell}(y, \sigma z) = -\frac{1}{2}(\sigma_1 z_1 - \beta_2 y + \sigma_2 z_2)$ with $z_1 = y - \beta_1$ and $z_2 = \sigma_1 z_1 - \beta_2 y$. For this particular example $\bar{\sigma}^d$ is given by

$$\begin{aligned} \sigma_1(y_d) &= \mathrm{sign}(y_d - \beta_1), \\ \sigma_2(y_d) &= \mathrm{sign}(\sigma_1(y_d) y_d - \beta_2 y_d) = \mathrm{sign}(\mathrm{sign}(y_d - \beta_1) y_d - \beta_2 y_d) \, . \end{aligned}$$

For $y_d(x_1, x_2) = \sin(\pi x_1)\sin(2\pi x_2)$, this yields

$$\bar{\sigma}_1^d(y_d) \equiv -1 \quad \text{and} \quad \bar{\sigma}_2^d(y_d) = \mathrm{sign}(-(1 + \beta_2) y_d) \equiv -\mathrm{sign}(y_d) \, .$$

Hence, the domain Ω is decomposed into $\overline{\Omega} = \overline{\Omega_1^+} \cup \overline{\Omega_1^-} = \overline{\Omega_2^+} \cup \overline{\Omega_2^-}$ with

$$\begin{aligned} \Omega_1^+ = \varnothing, \ \Omega_1^- &= \big((0,1) \times (0,1)\big) \\ \text{and } \ \Omega_2^+ = \big((0,1) \times (\tfrac{1}{2}, 1)\big), \ \Omega_2^- &= \big((0,1) \times (0, \tfrac{1}{2})\big) \, . \end{aligned}$$

In this regard we would like to recall the variational form of the underlying PDE and also point out the weak form of the equivalent reformulation using structured evaluation. One derives the weak formulation of a non-linear elliptic PDE the same way as in the linear case by multiplying

the equation with a test function, integrating the equation over the domain and then applying partial integration to transfer a derivative to the test function. In the case of the non-linear state equation in (5) we formally obtain the integral equation

$$\int_\Omega \nabla y \cdot \nabla v + \ell(y)v \; dx = \int_\Omega uv \; dx \; . \tag{18}$$

Due to the equivalent reformulation given by the structured evaluation Ass. 2.8, the decomposition of the domain Ω into $\Omega = \Omega^+ \cup \Omega^-$ and considering the case $s = 1$, Eq. (18) can be reformulated into

$$\begin{aligned} &\int_\Omega \nabla y \cdot \nabla v \; dx + \int_{\Omega^+} \hat{\ell}(y, \sigma z)v \; dx + \int_{\Omega^-} \hat{\ell}(y, \sigma z)v \; dx = \int_\Omega uv \; dx \\ \Leftrightarrow &\int_\Omega \nabla y \cdot \nabla v \; dx + \int_{\Omega^+} \hat{\ell}(y, +z)v \; dx + \int_{\Omega^-} \hat{\ell}(y, -z)v \; dx = \int_\Omega uv \; dx \; . \end{aligned}$$

Similarly, for a larger number $s \in \mathbb{N}$ of absolute value evaluations, the integrals can be partitioned according to the decomposition of the domain Ω and the terms depending on the respective $\bar{\sigma}_i^d$.

From now on we will use the notation $\bar{\ell}_{\bar{\sigma}}(y, z) := \hat{\ell}(y, \bar{\sigma} z)$ interchangeably. Applying the constant abs-linearization, the resulting operator $\hat{\ell}(\cdot, \bar{\sigma}\cdot) = \bar{\ell}_{\bar{\sigma}}(\cdot, \cdot)$ is smooth in both arguments. In addition to Ass. 2.2 we also assume that the given non-smooth operator ℓ fulfills the following property.

Assumptions 2.13 (The operator $\bar{\ell}_{\bar{\sigma}}$)**.** *Let $\ell : H_0^1(\Omega) \to L^2(\Omega)$ fulfill Ass. 2.2 and $\bar{\ell}_{\bar{\sigma}}(y, z)$ be the corresponding structured evaluation, see Ass. 2.8, with subsequently fixed $\bar{\sigma}$. We assume that $\bar{\ell}_{\bar{\sigma}}$ for fixed $\bar{\sigma}(x) \in \{-1, 1\}$ is affine-linear in y and z on the respective subdomains given by $\bar{\sigma}$.*

Note that an autonomous Nemytzkii operator ℓ is said to be affine linear, if its inducing function has the form

$$\ell(t) = f_0 + f_1 t \quad \text{for all } t \in \mathbb{R} \; ,$$

with $f_0 \in L^p(\Omega)$ and $f_1 \in L^\infty(\Omega)$; cf. [3].
At this point it should be pointed out that these requirements are indeed reasonable and can be satisfied by a large class of optimal control problems. Once again we consider the operator $\max(0, y)$ as an example.

Example 2.14 (Affine-Linearity of $\bar{\ell}_{\bar{\sigma}}$)**.** *Consider $\ell(y) = \max(0, y)$, then $s = 1$ and by Exam. 2.9 (i) $\bar{\ell}_{\bar{\sigma}}(y, z) = \hat{\ell}(y, \bar{\sigma} z) = \frac{1}{2}(y + \bar{\sigma}_1 z_1)$ with $z_1 = y$. Now fixing σ_1 for example to $\bar{\sigma}_1(x) \equiv +1$ on Ω yields $\bar{\ell}_{\bar{\sigma}}(y, z) = \frac{1}{2}(y + z_1) = y$ and therefore a linear operator. Alternatively, $\bar{\sigma}$ with $\bar{\sigma}_1(x) = +1$ on Ω^+, $\bar{\sigma}_1(x) = -1$ on Ω^- with $\Omega = \Omega^+ \cup \Omega^-$, e.g. Ω^+, Ω^- as in (2) yields*

$$\hat{\ell}(y, \bar{\sigma} z) = \begin{cases} \frac{1}{2}(y + z_1) = y, & \text{on } \Omega^+ \\ \frac{1}{2}(y - z_1) = 0, & \text{on } \Omega^- \; . \end{cases}$$

Proposition 2.15 (Differentiability of the Nemytzkii Operator $\bar{\ell}_{\bar{\sigma}}$)**.** *Let $\ell : H_0^1(\Omega) \to L^2(\Omega)$ fulfill Ass. 2.2 and $\hat{\ell}(y, \sigma z)$ be given by Ass. 2.8 with $\bar{\ell}_{\bar{\sigma}}(y, z) = \hat{\ell}(y, \bar{\sigma} z)$ satisfying Ass. 2.13 for some fixed $\bar{\sigma} \in L^2(\Omega)$ with $\bar{\sigma}(x) \in \{-1, +1\}$ for all $x \in \Omega$. Then the induced Nemytzkii operator $\bar{\ell}_{\bar{\sigma}}$ is Fréchet differentiable in $L^\infty(\Omega)$ as well as in $L^2(\Omega)$.*

Proof. See [24, Lemma 4.12] and [3]. □

Furthermore, according to [3] the Nemytzkii operator $\bar{\ell}_{\bar{\sigma}}$ is weakly continuous since the generating function $\bar{\ell}_{\bar{\sigma}}(y) = \hat{\ell}(y, \bar{\sigma} z)$ is affine.
Before characterizing the relation between local solutions of the model problem and its equivalent reformulation we define the notion of local solutions for the two considered problem formulations Eq. (5) and Eq. (16).

Definition 2.16 (Local Solution)**.** *Let $y^* \in H_0^1(\Omega)$, $z^* \in H^1(\Omega)$ and $u^*, \sigma^* \in L^2(\Omega)$ with $N_y \subseteq H_0^1(\Omega)$, $N_u \subseteq L^2(\Omega)$, $N_z \subseteq H^1(\Omega)$ and $N_\sigma \subseteq L^2(\Omega)$ local neighborhoods of y^*, u^*, z^* and σ^*, respectively. A pair $(y^*, u^*) \in N_y \times N_u$ is said to be a* local solution *to the original optimization problem (5) if (y^*, u^*) is feasible, i.e., satisfies the state constraint, and*

$$J(y,u) \geqslant J(y^*,u^*) \quad \text{for all } (y,u) \in N_y \times N_u \subseteq H_0^1(\Omega) \times L^2(\Omega) \text{ that are admissible.}$$

Similarly, $(y^*, z^*, u^*, \sigma^*) \in N_y \times N_z \times N_u \times N_\sigma$ is said to be a local solution *to the optimization problem (16) if $(y^*, z^*, u^*, \sigma^*)$ satisfies the constraints in (16) and*

$$J(y,u) \geqslant \hat{J}(y^*,z^*,u^*,\sigma^*) := J(y^*,u^*) \quad \text{for all } (y,u) \in N_y \times N_u \subseteq H_0^1(\Omega) \times L^2(\Omega)\,.$$

Note that the neighborhoods N_y and N_z are defined in the H^1-norm whereas N_u and N_σ are defined in the L^2-norm.
The following lemma characterizes the essential relation between the solutions of the original optimization problem (5) and the, according to Ass. 2.8 reformulated, optimization problem (16) with additional equality and inequality constraints for the auxiliary functions z and σ.

Lemma 2.17. *A pair $(y^*, u^*) \in H_0^1(\Omega) \times L^2(\Omega)$ with $y^* := y^*(u^*)$ is a local solution to the original optimization problem (5) if and only if $(y^*, z^*, u^*, \sigma^*) \in H_0^1(\Omega) \times H^1(\Omega) \times L^2(\Omega) \times L^2(\Omega)$ with $\sigma_i^* = \operatorname{sign}(z_i^*)$ and $z_i^* = \psi_i(y^*, (\sigma_j^* z_j^*)_{j<i})$ for $1 \leqslant i \leqslant s$ is a local solution of the optimization problem (16).*

Proof. Assume that u^* and the corresponding $y^* := y^*(u^*)$ are local solutions of the original optimization problem (5). Considering the equivalent reformulation of the operator ℓ into $\hat{\ell}$ by Ass. 2.8 and defining the auxiliary functions z_i^* and σ_i^* by

$$z_i^* = \psi_i(y^*, (\sigma_j^* z_j^*)_{j<i}), \qquad \sigma_i^* = \operatorname{sign}(z_i^*) \qquad \forall\ i = 1, \dots, s\,, \tag{19}$$

it follows that $(y^*, z^*, u^*, \sigma^*)$ is a local solution of the optimization problem (16). Here, the additional equality and inequality constraints for the definitions of the additional functions z_i^* and σ_i^*, $1 \leqslant i \leqslant s$, ensure that $\sigma_i^* z_i^* = \operatorname{abs}(z_i^*) \in L^2(\Omega)$ is valid for $1 \leqslant i \leqslant s$.
On the other hand, assume that $(y^*, z^*, u^*, \sigma^*)$, with σ_i^* defined by Eq. (19), is a local solution for optimization problem (16). Then $\sigma_i^* z_i^* = \operatorname{abs}(z_i^*) \in L^2(\Omega)$ is valid for $1 \leqslant i \leqslant s$ and one can replace in Eq. (16) σ_i accordingly, as well as z_i by $\psi_i(y, (\sigma_j z_j)_{j<i})$ for $1 \leqslant i \leqslant s$, taking the second equality condition in Eq. (16) into account. This then yields the optimal control problem (5) with solution (y^*, u^*). □

This observation motivates the optimization algorithm proposed in this paper. Note that the derivation of meaningful optimality conditions for Eq. (16) does not succeed with classical methods because of the non-smooth dependence of σ on z. However, if σ^* is known and $\sigma \equiv \sigma^*$ is fixed accordingly, the optimality conditions can be derived using the formal Lagrange technique and adapted approaches to [7, 6] and [24, Sec. 6.2.2] for optimal control problems with elliptic PDEs and pointwise state constraints since the non-smooth dependence of σ and z is removed.

Definition 2.18 (Optimally Reachable)**.**
Let $y_d \in L^2(\Omega)$ and $\bar{\sigma}^d$ be as defined in Def. 2.11. We call y_d an optimally reachable *desired state if for the optimal solution $(y^*, u^*, z^*, \sigma^*)$ one has $\bar{\sigma}^d = \sigma^*$ for almost all $x \in \Omega$ with $\sigma^*(x) \neq 0$.*

In the further course we will derive a necessary condition that allows us to easily check if $\bar{\sigma}^d \neq \sigma^*$. Nevertheless, we assume for this paper that the given desired state y_d is always optimally reachable:

Assumptions 2.19 (The optimally reachable desired state y_d)**.** *From now on we assume that the considered optimal control problem is such that the featured desired state y_d is optimally reachable in the sense of Def. 2.18 and hence $\bar{\sigma}^d = \sigma^*$.*

In fact, this assumption is satisfied by virtually all of our numerical test problems.
Next, we examine the first-order necessary optimality conditions for Eq. (16) with fixed functions $\sigma_i \equiv \sigma_i^*$ according to Eq. (19). In order to motivate the optimality system, a special case is considered. If the considered optimization problem and the semi-linearity given by the Nemytzkii operator ℓ is such that $s = 1$ and σ^* decomposes the domain into two separate non-overlapping connected measurable subdomains Ω^+, Ω^- such that $\overline{\Omega} = \overline{\Omega^+} \cup \overline{\Omega^-}$, then the optimization problem Eq. (16) reads as

$$\begin{aligned} \min_{y,z,u} \quad & \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + \hat{\ell}(y, \sigma z) - u = 0 && \text{in } \Omega \\ & \psi(y) - z = 0 && \text{in } \Omega \\ & y = 0 && \text{on } \partial\Omega \\ & z \geqslant 0 && \text{a.e. in } \Omega^+ \\ & z \leqslant 0 && \text{a.e. in } \Omega^- \end{aligned} \tag{20}$$

with $\sigma \equiv 0$ and hence $z \equiv 0$ on the common interface $\Gamma^0 := \overline{\Omega^+} \cap \overline{\Omega^-}$. Furthermore, $\Gamma^+ := \partial\Omega \cap \partial\Omega^+$ and $\Gamma^- := \partial\Omega \cap \partial\Omega^-$ denote the exterior boundary segment of Ω^+ and Ω^- respectively. Then by construction it holds that $y = 0$ on $\Gamma^+ \cup \Gamma^- = \partial\Omega$. See Fig. 2 for an illustration of the decomposition of Ω into the subdomains Ω^+, Ω^- and the resulting interface Γ^0.

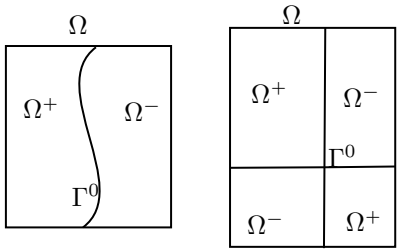


Figure 2: Non-overlapping decompositions of Ω into Ω^+, Ω^-.

Based on the assumptions on the original problem and the resulting conclusions, the underlying state problem yields a unique continuous state solution y for any control $u \in L^2(\Omega)$. Thus, due to the continuity of ψ, the switching function z is also uniquely determined and continuous as well. This yields in particular continuity of y and z across adjacent neighboring subdomains of Ω.
The reformulated optimal control problem (20) with coupled constrained systems over the subdomains is then given by

$$\min_{y,z,u} \quad \tfrac{1}{2}\|y - y_d\|_{L^2}^2 + \tfrac{\alpha}{2}\|u\|_{L^2}^2 \tag{21}$$

$$\text{s.t.} \quad -\Delta y + \hat{\ell}(y, +z) - u = 0 \quad \text{in } \Omega^+ \tag{22}$$

$$\psi(y) - z = 0 \quad \text{in } \Omega^+ \tag{23}$$

$$z > 0 \quad \text{a.e. in } \Omega^+ \tag{24}$$

$$z = 0 \quad \text{on } \Gamma^0 \tag{25}$$

$$y = 0 \quad \text{on } \Gamma^+ \tag{26}$$

$$-\Delta y + \hat{\ell}(y, -z) - u = 0 \quad \text{in } \Omega^- \tag{27}$$

$$\psi(y) - z = 0 \quad \text{in } \Omega^- \tag{28}$$

$$-z > 0 \quad \text{a.e. in } \Omega^- \tag{29}$$

$$z = 0 \quad \text{on } \Gamma^0 \tag{30}$$

$$y = 0 \quad \text{on } \Gamma^- \,. \tag{31}$$

The constraints Eqs. (22)–(26) and Eqs. (27)–(31) then provide local solutions y^+, z^+ and y^-, z^- with $y^+ = y^-$ on Γ^0, $y^+ = y^- = 0$ on $\partial\Omega$ and $z^+ = z^- = 0$ on Γ^0. The overall solution then satisfies $y(x) = \chi_{\Omega^+}(x)y^+(x) + \chi_{\Omega^-}(x)y^-(x)$ and $z(x) = \chi_{\Omega^+}(x)z^+(x) + \chi_{\Omega^-}(x)z^-(x)$. The well-posedness of the original PDE provides the well-posedness of the constraints restricted to the considered subdomains. Furthermore, the subproblem with the objective functional (21) restricted to Ω^+ and constrained by Eqs. (22)–(26) represents an optimization problem satisfying, e.g., the Slater constraint qualifications (CQs), such that the existence of associated Lagrange multipliers is ensured. The same applies to the optimization problem restricted to Ω^-.

Remark 2.20 (On the Slater CQ)**.** *Due to the affine-linearity assumption, Ass. 2.14 all the constraints of the problem formulations where σ is fixed are (affine-)linear, hence also in the case where σ is set to σ^*. Considering the cones which are spanned by the inequality constraints for z on Ω^- and Ω^+, the Slater condition requires the interior of these cones to be non-empty in $C(\overline{\Omega^-})$ and $C(\overline{\Omega^+})$, respectively. Note, that for admissible controls $u \in L^2(\Omega)$ which yield a state y and therefore switching functions z_i such that the inequality condition is a genuine inequality $\bar{\sigma}_i(x)z_i(x) > 0$ on the measurable subset $\Omega^+ \subseteq \Omega$ (and respectively $\Omega^- \subseteq \Omega$), the switching functions restricted to Ω^+ (and analogous to Ω^-) are already elements of the interior of the respective cone, due to the continuity of z_i. Hence the interior of these cones is not empty, satisfying the Slater CQ. If such a control u does not exist and only $\bar{\sigma}_i(x)z_i(x) = 0$ for all $x \in \Omega$ is valid for a feasible $\bar{\sigma}$, i.e., $z_i = 0$, then this condition can be ignored and z_i can be removed from the set of constraints. If this applies to all $1 \leqslant i \leqslant s$ then an optimal control problem with an affine-linear PDE (in the sense of Ass. 2.13) remains.*

The space $\mathcal{M}(\Omega)$ denotes the space of real regular Borel measures on Ω. Note that by the Riesz-representation theorem $\mathcal{M}(\Omega)$ can be identified with the dual of $C_0(\overline{\Omega})$, the space of continuous functions on Ω which vanish on the boundary $\partial\Omega$, and hence endowed with the norm

$$\|\mu\|_{\mathcal{M}} = \sup_{v \in C_0(\overline{\Omega}), \|v\|_{L^\infty} \leqslant 1} \int_{\overline{\Omega}} v \, d\mu \, . \tag{32}$$

Consequently, by considering the optimization problem partitioned into the subdomains, on which σ^* is constant, as well as the fact that $\sigma^* \equiv 0, z \equiv 0$ on the interface Γ^0, one can adapt the theory of [6, 7] and [24, Sec. 6.2.1] to provide the existence of well-defined Lagrange multipliers $\lambda_P^+, \lambda_P^-, \lambda^+, \lambda^-$ and regular Borel measures μ^+, μ^-, i.e., $\mu^\pm \in \mathcal{M}(\Omega^\pm)$. Composite construction then yields for the Lagrangian that there exists a real regular Borel measure μ on Ω, with $\mu = \mu_{|_{\Omega^+}} + \mu_{|_{\Omega^-}} = \mu^+ + \mu^-$, an adjoint state $\lambda_P = \lambda_{P|_{\Omega^+}} + \lambda_{P|_{\Omega^-}} = \lambda_P^+ + \lambda_P^-$ and a multiplier $\lambda = \lambda_{|_{\Omega^+}} + \lambda_{|_{\Omega^-}} = \lambda^+ + \lambda^- \in W^{1,p}(\Omega), p \in [1, \frac{n}{n-1})$ (see [24, Thm. 6.5]) such that the associated Lagrangian is given by

$$\begin{aligned}
\mathcal{L}(y, z, u, \lambda_P, \lambda, \mu) = {} & \mathcal{J}(y, u)_{|_{\Omega^+}} + \big(\nabla\lambda_P, \nabla y\big)_{L^2(\Omega^+)} + \big(\lambda_P, \hat{\ell}(y, z) - u\big)_{L^2(\Omega^+)} \\
& + (\lambda, \psi(y) - z)_{L^2(\Omega^+)} - \int_{\overline{\Omega^+}} z \, d\mu \\
& + \mathcal{J}(y, u)_{|_{\Omega^-}} + \big(\nabla\lambda_P, \nabla y\big)_{L^2(\Omega^-)} + \big(\lambda_P, \hat{\ell}(y, -z) - u\big)_{L^2(\Omega^-)} \\
& + (\lambda, \psi(y) - z)_{L^2(\Omega^-)} + \int_{\overline{\Omega^-}} z \, d\mu \\
= {} & \mathcal{J}(y, u) + \big(\nabla\lambda_P, \nabla y\big)_{L^2} + \big(\lambda_P, \hat{\ell}(y, \sigma z) - u\big)_{L^2} \\
& + (\lambda, \psi(y) - z)_{L^2} - \int_{\overline{\Omega}} \sigma z \, d\mu \, .
\end{aligned}$$

The associated Lagrangian $\mathcal{L}(y, z, u, \lambda_P, \lambda, \mu)$ then satisfies at the optimal solution, with σ

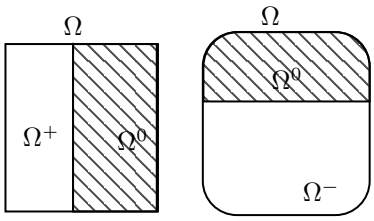


Figure 3: Non-overlapping decompositions of Ω into Ω^+, Ω^- and Ω^0.

fixed to σ^*, the following first-order necessary conditions

$$\begin{aligned}
0 = \mathrm{D}_y\mathcal{L}(\delta_y) &= \tfrac{\partial \mathcal{J}}{\partial y}\delta_y + \big(\nabla\lambda_P, \nabla\delta_y\big)_{L^2} + \big(\lambda_P \tfrac{\partial \hat{\ell}}{\partial y}, \delta_y\big)_{L^2} + \big(\lambda \tfrac{\partial \psi}{\partial y}, \delta_y\big)_{L^2} && \forall \delta_y \in H_0^1(\Omega),\\
0 = \mathrm{D}_u\mathcal{L}(\delta_u) &= \tfrac{\partial \mathcal{J}}{\partial u}\delta_u - \big(\lambda_P, \delta_u\big)_{L^2} && \forall \delta_u \in L^2(\Omega),\\
0 = \mathrm{D}_{\lambda_P}\mathcal{L}(\delta_{\lambda_P}) &= \big(\nabla\delta_{\lambda_P}, \nabla y\big)_{L^2} + \big(\hat{\ell} - u, \delta_{\lambda_P}\big)_{L^2} && \forall \delta_{\lambda_P} \in H_0^1(\Omega),\\
0 = \mathrm{D}_\lambda\mathcal{L}(\delta_\lambda) &= \big(\psi - z, \delta_\lambda\big)_{L^2} && \forall \delta_\lambda \in \mathcal{H}_\Delta,\\
0 = \mathrm{D}_z\mathcal{L}(\delta_z) &= \big(\lambda_P \sigma^* \tfrac{\partial \hat{\ell}}{\partial z}, \delta_z\big)_{L^2} - \big(\lambda, \delta_z\big)_{L^2} - \int_{\overline{\Omega}} \sigma^* \delta_z \, d\mu && \forall \delta_z \in \mathcal{H}_\Delta,\\
0 &= \int_{\overline{\Omega}} (\sigma^* z) \, d\mu, \qquad \mu \geqslant 0,
\end{aligned}$$

with the space $\mathcal{H}_\Delta$ as defined in Eq. (13). Note that the arguments of the functions and operators $\mathcal{L}(y, z, u, \lambda_P, \lambda, \mu), \hat{\ell}(y, \sigma^* z)$ and $\psi(y, \sigma^* z)$ are omitted for brevity. The multiplier μ is non-negative in the sense that

$$\int_{\overline{\Omega}} \xi(x) \, d\mu(x) \geqslant 0 \quad \text{for all } \xi \in C(\overline{\Omega}) \text{ with } \xi(.) \geqslant 0 \, .$$

In constellations where σ^* vanishes on a region of measure greater than zero (as illustrated in Fig. 3), i.e., $\sigma^* \equiv 0$ on $\overline{\Omega^0}$, the above considerations still apply in an adapted manner with homogeneous boundary conditions and adjusted regions.
Consequently, adapting the theory of [6, 7] and [24, Sec. 6.2.1] to the general case $s \in \mathbb{N}$, one derives for the Lagrangian

$$\begin{aligned}
\mathcal{L}(y, z, u, \lambda_P, \lambda, \mu) = \mathcal{J}(y, u) &+ \big(\nabla\lambda_P, \nabla y\big)_{L^2} + \big(\lambda_P, \hat{\ell}(y, \sigma z) - u\big)_{L^2} \\
&+ \sum_{i=1}^{s} \big(\lambda_i, \psi_i(y, (\sigma_j z_j)_{j<i}) - z_i\big)_{L^2} - \sum_{i=1}^{s} \int_{\overline{\Omega}} \sigma_i z_i \, d\mu
\end{aligned} \tag{33}$$

at the optimal point with σ_i fixed to σ_i^* the following first-order necessary conditions. By $\mathrm{D}_y\mathcal{L} = 0$ one obtains the adjoint equation and by $\mathrm{D}_u\mathcal{L} = 0$ combined with complementary conditions the optimality system. If the control u is assumed to be a solution of problem (5) and y the associated state, i.e., solves the semi-linear PDE with right hand side u, then there exist real regular Borel measures μ_i on Ω for $1 \leqslant i \leqslant s$, i.e., $\mu_i \in \mathcal{M}(\Omega)$. Furthermore, there exist an adjoint state λ_P and Lagrange multipliers λ_i for $1 \leqslant i \leqslant s$ which are elements of the Sobolev space $W^{1,p}(\Omega)$ for all $p \in [1, \frac{n}{n-1})$, such that the following optimality system is satisfied:

$$\begin{aligned}
0 = \mathrm{D}_y\mathcal{L}(\delta_y) &= \tfrac{\partial \mathcal{J}}{\partial y}\delta_y + \big(\nabla\lambda_P, \nabla\delta_y\big)_{L^2} \\
&\quad + \big(\lambda_P \tfrac{\partial \hat{\ell}}{\partial y}, \delta_y\big)_{L^2} + \sum_{i=1}^{s} \big(\lambda_i \tfrac{\partial \psi_i}{\partial y}, \delta_y\big)_{L^2} && \forall \delta_y \in H_0^1(\Omega) && (34)\\
0 = \mathrm{D}_u\mathcal{L}(\delta_u) &= \tfrac{\partial \mathcal{J}}{\partial u}\delta_u - \big(\lambda_P, \delta_u\big)_{L^2} && \forall \delta_u \in L^2(\Omega) && (35)\\
0 = \mathrm{D}_{\lambda_P}\mathcal{L}(\delta_{\lambda_P}) &= \big(\nabla\delta_{\lambda_P}, \nabla y\big)_{L^2} + \big(\hat{\ell} - u, \delta_{\lambda_P}\big)_{L^2} && \forall \delta_{\lambda_P} \in H_0^1(\Omega) && (36)
\end{aligned}$$

$$0 = \mathrm{D}_{\lambda_i}\mathcal{L}(\delta_{\lambda_i}) = \left(\psi_i - z_i, \delta_{\lambda_i}\right)_{L^2} \qquad \forall \delta_{\lambda_i} \in \mathcal{H}_\Delta, i \in \mathbb{N}_s \tag{37}$$

$$0 = \mathrm{D}_{z_k}\mathcal{L}(\delta_{z_k}) = \left(\lambda_P \sigma_k^* \tfrac{\partial \hat{\ell}}{\partial z_k}, \delta_{z_k}\right)_{L^2} - \left(\lambda_k, \delta_{z_k}\right)_{L^2} + \sum_{i=k+1}^{s} \left(\lambda_i \sigma_k^* \tfrac{\partial \psi_i}{\partial z_k}, \delta_{z_k}\right)_{L^2} - \int_{\overline{\Omega}} \sigma_k^* \delta_{z_k} d\mu_k \quad \forall \delta_{z_k} \in \mathcal{H}_\Delta, k \in \mathbb{N}_s, \tag{38}$$

$$0 = \int_{\overline{\Omega}} (\sigma_i^*(x) z_i(x))\, d\mu_i(x), \qquad \mu_i \geqslant 0 \qquad i \in \mathbb{N}_s\ , \tag{39}$$

with $\mathbb{N}_s := \{1, \dots, s\}$.
Note that the arguments of $\mathcal{L}(y, z, u, \lambda_P, \lambda, \mu), \hat{\ell}(y, \sigma^* z), \psi(y, \sigma^* z)$ are omitted for brevity. In these equations one obtains additional factors σ_k^* due to the chain rule. In Eqs. (34)–(38) σ^* is fixed eliminating the non-smooth dependency of σ_i on z_i. Therefore, no discussion of generalized stationarity concepts is required.
Furthermore, in the case of optimally reachable desired states, the signature function $\bar{\sigma}^d$ coincides almost everywhere with σ^* (except for regions where σ^* vanishes) and thus the above considerations also apply to $\bar{\sigma}^d$.

Definition 2.21 (Stationary Point)**.** *Let* $(y^*, u^*) \in (H_0^1(\Omega) \cup C(\bar{\Omega})) \times L^2(\Omega)$ *be such that Eqs.* (34)–(38) *hold. Then* (y^*, u^*) *is called* stationary point *for the optimal control problem* (5)*.*

Rearranging the terms in the integrals, condition (38) yields for $k \in \mathbb{N}_s$ that

$$\int_{\overline{\Omega}} \sigma_k^* \delta_{z_k} d\mu_k = \left(\lambda_P \sigma_k^* \tfrac{\partial \hat{\ell}(y, \sigma^* z)}{\partial z_k} - \lambda_k + \sum_{i=k+1}^{s} \lambda_i \sigma_k^* \tfrac{\partial \psi_i(y, (\sigma_j^* z_j)_{j<i})}{\partial z_k}, \delta_{z_k}\right)_{L^2} \quad \forall \delta_{z_k} \in \mathcal{H}_\Delta\ .$$

By applying σ_k^* as well as $\operatorname{sign}(\delta_{z_k})$ we define the function

$$r(\sigma_k) := |\sigma_k| \lambda_P \tfrac{\partial \hat{\ell}(y, \sigma z)}{\partial z_k} - \sigma_k \lambda_k + \sum_{i=k+1}^{s} |\sigma_k| \lambda_i \tfrac{\partial \psi_i(y, (\sigma_j z_j)_{j<i})}{\partial z_k}\ , \tag{40}$$

to represent the modified right hand side. For this equation the signature function σ does not depend on z and y and is treated as a separate variable. Exploiting the non-negativity of μ_k according to Eq. (39), one obtains with the previously defined function evaluated at σ_k^* that

$$0 \leqslant \int_{\overline{\Omega}} |\sigma_k^*| |\delta_{z_k}| d\mu_k = (r(\sigma_k^*), |\delta_{z_k}|)_{L^2}$$

for all $\delta_{z_k} \in \mathcal{H}_\Delta$ and $k \in \mathbb{N}_s$. Hence,

$$0 \leqslant r(\sigma_k^*) \quad \text{a.e. in } \Omega, k \in \mathbb{N}_s\ . \tag{41}$$

The inequality (41) will later be used to verify that the determined solution is a stationary point. Note that if y_d is not optimally reachable in the sense of Def. 2.18, the constant abs-linearized (CAL) problem given by $\bar{\sigma}^d$ is not optimal, i.e., $\bar{\sigma}^d \neq \sigma^*$. This can be verified by means of the function r defined in Eq. (40). If r is negative at $\bar{\sigma}_k^d$ for all $k \in \mathbb{N}_s$ the optimality condition is not satisfied. Therefore, it can be easily detected whether a computed solution for the CAL problem is stationary for the original nonsmooth problem (5).

3 Defining and Solving Constant Abs-Linearized Problems

Now, everything is prepared to introduce the new optimization algorithm. For fixed functions $\bar{\sigma}_i^d \in L^2(\Omega)$, $\bar{\sigma}_i^d : \Omega \to \{-1, 1\}$ for $1 \leqslant i \leqslant s$, we define for $(y, z, u) \in H_0^1(\Omega) \times [H^1(\Omega)]^s \times L^2(\Omega)$

the *CAL problem*

$$\min_{y,z,u} \quad \mathcal{J}(y,u) \tag{42}$$

$$\text{s.t.} \quad (\nabla v, \nabla y)_{L^2} + (\bar{\ell}_{\bar{\sigma}^d}(y,z) - u, v)_{L^2} = 0 \ \forall v \in H_0^1(\Omega) \tag{43}$$

$$(\psi_i(y, (\bar{\sigma}_j^d z_j)_{j<i}) - z_i, v_i)_{L^2} = 0 \ \forall v_i \in H_0^1(\Omega) \quad \forall\ i \in \mathbb{N}_s \tag{44}$$

$$\bar{\sigma}_i^d z_i \geqslant 0 \text{ a.e. in } \Omega \quad \forall\ i \in \mathbb{N}_s\ . \tag{45}$$

All functions occurring in this constant abs-linearized problem are smooth in the variables y, u and z because the function $\hat{\ell}(.,.)$ is smooth in its arguments as mentioned already in the last section. Therefore, standard smooth optimization methods can be used to solve the problem (42)–(45).

The Lagrangian with Bi-quadratic Penalty

As mentioned already above, so far the solution of the non-smooth optimization problem using a reduced formulation is not possible. The reason for this is the lack of the classical chain rule as well as the non-smoothness of the control-to-state operator associated with the non-smooth state equation. Due to the applied reformulation by means of the structured evaluation the considered optimal control problem exhibits pointwise state constraints which might lead to e.g. numerical difficulties due to the low regularity of the respective Lagrange multipliers, see e.g. [7]. Hence several regularization methods (in the smooth PDE case) like the Moreau–Yosida approximation [17], the Lavrentiev type regularization [21] or even barrier methods like the interior point method [22] were proposed to overcome this difficulty and to allow the application of e.g. semi-smooth Newton, SQP or active set methods. We propose for the special problem class considered here a penalty-based approach to solve the optimization problem (42)–(45), where the constraints (43) and (44) are handled explicitly. Methods based on a reduced formulation will be subject of future research.
From a formal point of view, we treat the inequality constraints (45) with a penalty approach such that the target function (42) is modified to obtain the augmented objective functional

$$\min_{y,z,u} \quad \mathcal{J}(y,u) + \nu \int_\Omega \sum_{i=1}^{s} \Big(\max(-\bar{\sigma}_i^d z_i, 0) \Big)^4 \, d\Omega \tag{46}$$

with a penalty factor $\nu > 0$. In this context, as well as in the further course, ν describes a non-negative constant penalty parameter for the inequality conditions on $\sigma_i z_i$. Here, we chose the exponent 4 to ensure that the target function is twice continuously differentiable despite the max function that is used for the formulation of the penalty function.
The modified target function (46) is then coupled with the equality constraints by means of Lagrange multipliers yielding the Lagrangian

$$\begin{aligned} \mathcal{L}^p(y,z,u,\lambda_P,\lambda_1,\ldots,\lambda_s) = \mathcal{J}(y,u) + \big(\nabla\lambda_P, \nabla y\big)_{L^2} + \big(\lambda_P, \hat{\ell}(y,\bar{\sigma}z) - u\big)_{L^2} \\ + \sum_{i=1}^{s} \big(\lambda_i, \psi_i(y,(\bar{\sigma}_j z_j)_{j<i}) - z_i\big)_{L^2} + \nu \int_\Omega \sum_{i=1}^{s} \Big(\max(-\bar{\sigma}_i z_i, 0) \Big)^4 \, d\Omega\ . \end{aligned} \tag{47}$$

A similar penalty approach was studied in [26], where the logarithm was used as barrier function. Here, we employ the max-function since we have to evaluate the penalty function also at $z = 0$.
Alternatively to CALi method together with the bi-quadratic penalty one could also apply a more common approach, the direct regularization and modification of the non-smooth Nemytzkii operator in the state equation in order to obtain a Gâteaux differentiable operator, which enables the standard way of deriving first-order optimality conditions by using adjoint calculus. However, this procedure does not only require several additional assumptions on the considered non-smooth optimization problem but also an asymptotic limit analysis for vanishing regularization parameters to eventually provide an optimality system of weak stationarity type. Hence, directly modifying the Nemytzkii operator does not result in a more efficient

approach compared to the method presented here. This fact is illustrated by the numerical results presented in Sec. 5 providing a comparison between a method based on such a regularization by direct modification and the CALi method.
Note, that in the case of such a direct regularization, the solution of a sequence of related regularized auxiliary problems with assigned regularization parameters is required in order to asymptotically approximate the solution of the original non-smooth problem. However, in the considered setting the CALi approach does not require asymptotic approximations or adaption of the parameter ν but instead only the solution of a single smooth problem as illustrated also by our numerical results given in Sec. 5. Hence, it constitutes a different approach compared to common regularization-based methods for non-smooth optimization.

Example 3.1. *We consider* $\ell(y) = \max(|y| - y, 0)$. *For the CAL optimization problem given by*

$$\begin{aligned}
\min_{(y,z,u)\in H_0^1\times H^1\times L^2} \quad & \tfrac{1}{2}\|y-y_d\|_{L^2}^2 + \tfrac{\alpha}{2}\|u\|_{L^2}^2 \\
\text{s.t.} \quad & (\nabla v, \nabla y)_{L^2} + \left(\tfrac{1}{2}\left(\bar\sigma_1 z_1 - y + \bar\sigma_2 z_2\right) - u, v\right)_{L^2} = 0 \ \forall v \in H_0^1(\Omega) \\
& (y - z_1, v)_{L^2} = 0 \ \forall v \in H_0^1(\Omega) \\
& (\bar\sigma_1 z_1 - y - z_2, v)_{L^2} = 0 \quad \forall v \in H_0^1(\Omega) \\
& \bar\sigma_1 z_1 \geqslant 0 \ \textit{a.e in } \Omega \\
& \bar\sigma_2 z_2 \geqslant 0 \ \textit{a.e in } \Omega \,,
\end{aligned}$$

one obtains the Lagrangian

$$\begin{aligned}
&\mathcal{L}^p(y,z,u,\lambda_P,\lambda_1,\lambda_2) \\
&\quad = \mathcal{J}(y,u) + \left(\nabla\lambda_P, \nabla y\right)_{L^2} + \left(\lambda_P, \tfrac{1}{2}\left(\bar\sigma_1 z_1 - y + \bar\sigma_2 z_2\right) - u\right)_{L^2} \\
&\qquad + \left(\lambda_1, y - z_1\right)_{L^2} + \left(\lambda_2, \bar\sigma_1 z_1 - y - z_2\right)_{L^2} \\
&\qquad + \nu \int_\Omega \sum_{i=1}^{2} \Big(\max(-\bar\sigma_i z_i, 0)\Big)^4 \, d\Omega \,.
\end{aligned}$$

Deriving Necessary Optimality Conditions for the Penalty Problem

A simple comparison of the equivalently reformulated problem and the penalty problem suggests that a solution to the penalty problem is feasible for the original problem, if it satisfies the condition

$$\sigma_i z_i = \operatorname{abs}(z_i) \text{ a.e. in } \Omega \,. \tag{48}$$

However, this provides no statement about the optimality of the solution. In addition, if the condition Eq. (48) is not met, only a statement about the selected penalty parameter can be made
For a CAL problem, the first-order necessary optimality conditions can now be derived from the Lagrangian (47) using standard KKT theory for smooth PDE-constrained optimization problems given that some regularity conditions (e.g. Slater CQ) are satisfied at the local minimum. This yields the following necessary first order conditions

$$
\begin{aligned}
0 = D_y\mathcal{L}^p(\delta_y) \quad &= \tfrac{\partial \mathcal{J}}{\partial y}\delta_y + \left(\nabla\lambda_P, \nabla\delta_y\right)_{L^2} + \left(\lambda_P \tfrac{\partial \hat{\ell}}{\partial y}, \delta_y\right)_{L^2} \\
&\quad + \sum_{i=1}^{s} \left(\lambda_i \tfrac{\partial \psi_i(y, (\bar{\sigma}_j^d z_j)_{j<i})}{\partial y}, \delta_y\right)_{L^2} \qquad \forall \delta_y \in H_0^1(\Omega) \qquad (49) \\
0 = D_u\mathcal{L}^p(\delta_u) \quad &= \tfrac{\partial \mathcal{J}}{\partial u}\delta_u - \left(\lambda_P, \delta_u\right)_{L^2} \qquad \forall \delta_u \in L^2(\Omega) \qquad (50) \\
0 = D_{\lambda_P}\mathcal{L}^p(\delta_{\lambda_P}) &= \left(\nabla y, \nabla\delta_{\lambda_P}\right)_{L^2} + \left(\hat{\ell} - u, \delta_{\lambda_P}\right)_{L^2} \qquad \forall \delta_{\lambda_P} \in H_0^1(\Omega) \qquad (51) \\
0 = D_{\lambda_i}\mathcal{L}^p(\delta_{\lambda_i}) \quad &= \left(\psi_i(y, (\bar{\sigma}_j^d z_j)_{j<i}) - z_i, \delta_{\lambda_i}\right)_{L^2} \qquad \forall \delta_{\lambda_i} \in H_0^1(\Omega), 1 \leqslant i \leqslant s \qquad (52) \\
0 = D_{z_k}\mathcal{L}^p(\delta_{z_k}) \quad &= \left(\lambda_P \bar{\sigma}_k^d \tfrac{\partial \hat{\ell}(y, \bar{\sigma}^d z)}{\partial z_k}, \delta_{z_k}\right)_{L^2} - \left(\lambda_k, \delta_{z_k}\right)_{L^2} \\
&\quad + \sum_{i=k+1}^{s} \left(\lambda_i \bar{\sigma}_k^d \tfrac{\partial \psi_i(y, (\bar{\sigma}_j^d z_j)_{j<i})}{\partial z_k}, \delta_{z_k}\right)_{L^2} \qquad (53) \\
&\quad + \nu \int_\Omega -4\bar{\sigma}_k^d \max(-\bar{\sigma}_k^d z_k, 0)^3 \, dx \qquad \forall \delta_{z_k} \in H_0^1(\Omega), 1 \leqslant k \leqslant s
\end{aligned}
$$

Definition 3.2 (KKT Point)**.** *Let $\bar{y} \in H_0^1(\Omega)$, $\bar{z} = (\bar{z}_1, \ldots, \bar{z}_s) \in [H^1(\Omega)]^s$, $\bar{u} \in L^2(\Omega)$. The pair $(\bar{y}, \bar{z}, \bar{u}, \bar{\lambda}_{PDE}, \bar{\lambda})$ is called a KKT point of the optimization problem (42)–(45) if there exist multipliers $\bar{\lambda}_P \in H^{-1}(\Omega)$ and $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_s) \in [(L^2)^*]^s$ such that Eq. (49)–(53) hold for $(\bar{y}, \bar{z}, \bar{u}, \bar{\lambda}_P, \bar{\lambda})$. In this case $(\bar{y}, \bar{z}, \bar{u})$ is a stationary point of (42)–(45).*

As one can easily see, the optimality conditions (34)–(37) coincide with the optimality conditions (49)–(52). The following relation between the KKT point of the penalty problem Eqs. (42)–(45) and the original optimization problem Eq. (5) can be derived.

Lemma 3.3. *Let $\bar{y} \in H_0^1(\Omega)$, $\bar{z} = (\bar{z}_1, \ldots, \bar{z}_s) \in [H^1(\Omega)]^s$, $\bar{u} \in L^2(\Omega)$, $\bar{\lambda}_{PDE} \in H^{-1}$ and $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_s) \in [(L^2)^*]^s$. Assume that the conditions Eq. (49)–(53) hold for $(\bar{y}, \bar{z}, \bar{u}, \bar{\lambda}_{PDE}, \bar{\lambda})$ together with Eq. (41) for all $1 \leqslant k \leqslant s$. Then the pair $(\bar{y}, \bar{z}, \bar{u}, \bar{\lambda}_{PDE}, \bar{\lambda})$ is a KKT point of the optimization problem Eq. (33). Furthermore, $(\bar{y}, \bar{u})$ is a stationary point of the original problem Eq. (5).*

Proof. We again point out that the optimality conditions (34)–(37) for the CAL problem with σ^* coincide with the optimality conditions (49)–(52) for the reformulated problem Eq. (16). Hence, if one computes a solution of the penalty problem with the target function (46) and the constraints (43)–(44), the necessary first order conditions (34)–(37) of the original optimization problem are already satisfied. Consequently, the only condition to verify is Eq. (41). Since the expressions on the right-hand side are completely independent of the Lagrange multiplier μ of the original optimization problem, one can compute this quantity also for the solution of the CAL problem. If it is non-negative, the computed solution $(\bar{y}, \bar{z}, \bar{u})$ of the CAL problem fulfills the necessary first order conditions of the original optimization problem for the chosen functions $\bar{\sigma}_i \in L^2(\Omega)$. Hence, $(\bar{y}, \bar{z}, \bar{u})$ is a stationary point of Eq. (16) and by Lem. 2.17 it is also a stationary point of Eq. (5). □

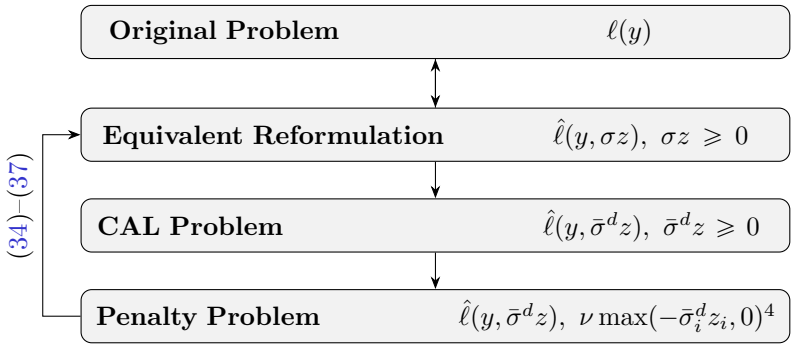


Figure 4: The different optimization problems and how they relate to each other

.

Fig. 4 illustrates the nature of the relationships between the different problem formulations that are derived and discussed in this paper. The acronym CAL abbreviates the term 'constant abs-linearized'.
Before we discuss the algorithm for solving these problems we would like to point out that also more general nonlinear and non-autonomous Nemytzkii operators can be considered as far as the generating function fulfills the growth condition

$$|\ell(x, y)| \leqslant c|y|^{p-1} + h(x) \text{ for a.e. } x \in \Omega$$

for all $y \in H_0^1(\Omega)$, with $c \geqslant 0$ and some $h \in L^q(\Omega)$, where the exponents p, q are chosen such that $H_0^1(\Omega)$ is compactly embedded into $L^p(\Omega)$ and $\ell(\cdot, y) \in L^q(\Omega)$, i.e., $1 < p < \frac{2n}{n-2}$ (and $p < \infty$), $\frac{1}{p} + \frac{1}{q} = 1$. For the derivation of optimality conditions a norm gap has to be considered for Nemytzkii operators that are not affine-linear, hence more caution should be devoted to the considered domain and image spaces.

4 The Resulting Optimization Algorithm

Motivated by the observations of the last section, we propose the method stated in Algo. 1 to solve optimal control problems with non-smooth PDEs of the class considered here as constraints.

Algorithm 1 CALi

Input: Initial values: $\bar{\sigma} \equiv \bar{\sigma}^d = (\bar{\sigma}_1, \ldots, \bar{\sigma}_s), y, z = (z_1, \ldots, z_s), u$
Parameter: $\alpha > 0, \nu > 0$
Solve smooth penalty problem (46) with constraints (43)–(44) to obtain $y, z, u, \lambda_P, \lambda$

if Eq. (41) holds for $k = 1, \ldots, s$ **then**
y, z, u stationary for original optimal control problem
end if

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*. Following standard practice for PDE-constrained optimization, we develop the algorithm in a function space setting. This has the advantage that the associated algorithms are often able to provide mesh-independent convergence for a variety of conform discretizations, see for example [1, 2, 16, 25]. As can be seen from the numerical results in Sec. 5, mesh independence is also an important feature of the algorithm presented here. Obviously, one can use the preferred method to solve the smooth modified CAL optimization problems.

For the numerical results shown in the next section, we used a Finite-Element-Approach based on FEniCS [20] to discretize the PDEs and to describe the other constraints in combination with a Newton method for the solution of the smooth CAL problems.
For the initial state, control and parameters $\bar{\sigma}_i$, the non-linear variational Lagrange problem is solved by Newton's method using the derivatives calculated within FEniCS.

Finite Dimensional Formulation

In the last paragraph, the algorithm was presented and explained in the continuous function space setting. Now, the natural question is how to put this into practice and, especially, how to solve the given penalty problems. For the numerical treatment of the optimal control problem (46) with the constraints given by Eqs. (43) and (44), the Lagrange equation (47) will be discretized. For this purpose we apply a standard finite element method with piecewise linear and continuous ansatz functions for the discretization of the functions y and $z_i\sigma_i,\ i = 1, \dots, s$, and piecewise constant ansatz functions for the control u. The resulting problem is solved by the Galerkin method within the open source finite element environment FEniCS.
Below we will discuss the spatial discretization of the constrained optimization problem (5), which will result in a large-scale non-linear optimization problem. We focus on finite element approaches with a quasi uniform triangulation $\mathbb{T}^h = \{T_1, \dots, T_m\}$, the vector space of test functions $V^h := \{v^h \in C^0(\bar{\Omega}) : v^h|_{T_j} \in \mathcal{P}_1(T)\ \forall T \in \mathbb{T}^h, v^h|_{\partial\Omega} = 0\} = \operatorname{span}\{\xi_1, \dots, \xi_n\}$ and the discrete control space $U^h := \operatorname{span}\{e_T : T \in \mathbb{T}^h\}$ where $e_T : \Omega \to \mathbb{R}$ denotes the characteristic function for the simplex $T \in \mathbb{T}^h$. The superscript h denotes the mesh size of the triangulation and is given by

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

In the following we denote by $(\cdot,\cdot)_\Omega$ the discrete finite dimensional scalar product over the domain Ω. Then the discretization of Eq. (5) can be stated as

$$\min_{(y^h,u^h)\in V^h\times U^h} \quad J(y^h, u^h) \tag{54}$$

$$\text{s.t.} \quad (\nabla y^h, \nabla v^h)_\Omega + \ (\ell(y^h), v^h)_\Omega - (u^h, v^h)_\Omega = 0 \quad \forall v^h \in V^h . \tag{55}$$

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 $\mathbb{T}^h$. Taking into account that the operator ℓ is non-linear, the above representations yield the following discretizations:

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

and

$$(\ell(y^h), v^h)_{L^2} = \int_\Omega \ell(y^h)v^h dx \approx \sum_{T\in\mathbb{T}^h} \int_T \ell(y^h)v^h dx . \tag{56}$$

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

$$\sum_{T\in\mathbb{T}^h} \int_T \ell(y^h)v^h dx \approx \sum_{T\in\mathbb{T}^h} \sum_{k=1}^{n_k} \omega_k\ \ell\left(\sum_{i=1}^{n} y_i\xi_i(x_k)\right) \sum_{j=1}^{n} \xi_j(x_k) , \tag{57}$$

with n_k quadrature points per element T and corresponding weights ω_k.
Hence, the naturally arising discretization for the non-smooth operator from Ass. 2.8 in the finite element context is per quadrature point. This increases the number of absolute value evaluations, but not the way they are nested.
As can be seen in Eq. (57), we would like to point out that the number of non-smooth functions ℓ in the discretized problem is per quadrature point. Compared with our substitution strategy Ass. 2.8, this is not in perfect alignment with a representation by a finite element function like the state y. Consequentially, the choice of this discretization and the execution of the equivalent reformulation according to Ass. 2.8 leads to an increase of the polynomial degree due to the multiplication $\bar{\sigma}_i \mathbf{z_i}$ in the discretized representation of the operator $\hat{\ell}$ in contrast to the operator ℓ. However, this specific discretization allows for a straight forward implementation with FEniCS.
Inserting Eq. (56) into Eq. (55) and replacing v by ξ leads to:

$$\int_\Omega \sum_{k=1}^{n} \nabla\xi_j(x) \cdot \nabla\xi_k(x) y_k + \ell\big(\sum_{i=1}^{n} y_i \xi_i(x)\big)\xi_j(x)dx = \int_\Omega \left(\sum_{s=1}^{m} u_s e_{T_s}(x)\right) \xi_j(x)dx \,, \tag{58}$$

for $1 \leqslant j \leqslant n$. By defining

$$A_{jk} := \int_\Omega \nabla\xi_j(x) \cdot \nabla\xi_k(x)dx = (\nabla\xi_j, \nabla\xi_k)_\Omega,$$
$$b_k(y^h) := \int_\Omega \ell\big(\sum_{i=1}^{n} y_i \xi_i(x)\big)\xi_k(x)dx$$

and

$$g_j := \int_\Omega u^h(x)\xi_j(x)dx = \int_\Omega \left(\sum_{s=1}^{m} u_s e_{T_s}(x)\right) \xi_j(x)dx$$

Eq. (58) can be rewritten as

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

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

$$A\mathbf{y} + \mathbf{b}(\mathbf{y}) = \mathbf{u}^T E \,, \tag{59}$$

with the control matrix $E_{ij} := (e_{T_i}, \xi_j)$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$ denoting the finite-element approximation belonging to the right-hand side given by the discrete control u. To this end, the function $y^h|_{T_k}$ on the linear element T_k, is realized in terms of its point values at preselected sets of nodes scattered along the boundary of T_k. Note that in the above algebraic system the vector $\boldsymbol{u}$ and the matrices A, E are constant since they are independent of the unknown $y_1, \ldots, y_n$. However, as previously mentioned, this non-linear algebraic equation is assumed to be based on a reasonable approximation of the integral via quadrature.
Hence, the resulting discretized 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 $M_{ij} = (\xi_i, \xi_j)_\Omega$ and D the control mass matrix with the entries $D_{ij} = (e_{T_i}, e_{T_j})_\Omega$, where D is a diagonal matrix because the interior of the triangles are disjunct to each other.
Similar to the previously derived discretization, the discrete counterpart to the CAL problem

Eq. (42)–(45) is given by:

$$\begin{aligned} \min_{(y^h,z^h,u^h)\in V^h\times[V^h]^s\times U^h} \quad & J(y^h,u^h)+\nu\int_\Omega \sum_{i=1}^{s} \max\left(-\bar\sigma_i^h z_i^h,0\right)^4 dx \\ \text{s.t.} \quad & (\nabla y^h,\nabla v^h)_\Omega + \left(\hat\ell(y^h,\bar\sigma^h z^h),v^h\right)_\Omega = (u^h,v^h)_\Omega, \qquad \forall v^h\in V^h \\ & (z_i^h-\psi_i(y^h,(\bar\sigma_j^h z_j^h)_{j<i}),v^h)_\Omega = 0 \quad \forall\ 1\leqslant i\leqslant s,\ \forall v^h\in V^h\ . \end{aligned} \tag{60}$$

Note that $z^h=(z_1^h,\dots,z_s^h)\in[V^h]^s$. Hence, the inequality constraint from Eq. (45) is enforced per quadrature point via our penalty approach.

The assumptions for the non-smooth operator ℓ are carried over from the continuous setting into the discrete and hence once again we assume that the optimization problem Eq. (60) fulfills some kind of constraint qualification to ensure that the Lagrange function and the Lagrange multipliers are well-defined, i.e., the existence of the Lagrange multipliers is ensured. The corresponding discrete Lagrange functional related to the penalty branch problem of system Eq. (60) is now given by

$$\begin{aligned} \mathcal{L}^p(y^h,z^h,u^h,\lambda_P^h,\lambda^h) = \mathcal{J}(y^h,u^h) + \left(\nabla\lambda_P^h,\nabla y^h\right)_\Omega + \nu\int_\Omega\sum_{i=1}^{s}\left(\max(-\bar\sigma_i^h z_i^h,0)\right)^4 dx \\ + \left(\lambda_P^h,\hat\ell(y^h,\bar\sigma z^h)-u^h\right)_\Omega + \sum_{i=1}^{s}\left(\lambda_i^h,\psi_i(y^h,(\bar\sigma_j^h z_j^h)_{j<i})-z_i^h\right)_\Omega . \end{aligned} \tag{61}$$

The KKT system corresponding to Eq. (61) is then solved with a non-linear variational Newton solver.

5 Numerical Results

For the numerical tests we considered two-dimensional examples defined below in Case 1 to Case 4. In each example Ω was chosen to be the unit square, and we take as an initial guess $y\equiv\ 0, u\equiv\ 0, z_1\equiv\ 0, z_2\equiv 0$. Furthermore, $\bar\sigma_1^d$ and $\bar\sigma_2^d$ are chosen such that they fit the ones defined by the desired state y_d.
We terminate the Newton iterations if the norm of the corresponding residuals becomes less than 10^{-12}.
All calculations were performed with FEniCS, version 2019.1.0, using the Python interface.

Case 1

$$\begin{aligned} \min_{(y,u)} \quad & \frac{1}{2}\|y-y_d\|_{L^2}^2+\frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y+\max(0,y)-u=f \ \text{ in } \Omega=(0,1)^2 \\ & y=0 \ \text{ on } \partial\Omega \end{aligned}$$

with desired state

(a)

$$y_d(x_1,x_2)=\sin(\pi x_1)\sin(2\pi x_2)\ ,$$

(b)

$$y_d(x_1,x_2)=\begin{cases}((x_1-\frac{1}{2})^4+\frac{1}{2}(x_1-\frac{1}{2})^3)\sin(\pi x_2), & \text{if } x_1\leqslant\frac{1}{2}\\ 0, & \text{otherwise}\ ,\end{cases}$$

where $f\in L^2(\Omega)$ is chosen on the right hand side such that $-\Delta y_d+\max(0,y_d)=f$ is fulfilled.

Case 2

$$\begin{aligned}\min_{(y,u)} \ & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + \max(0, y) + \beta \min(0, y) = u + f \ \text{ in } \Omega = (0,1)^2 \\ & y = 0 \ \text{ on } \partial\Omega \,,\end{aligned}$$

with the desired state chosen as

$$y_d(x_1, x_2) = 10^2(x_1 - \tfrac{1}{2})^3 \sin(2\pi x_1)\sin(\pi x_2)\cos(\pi x_2) \text{ and } \beta \in (0,1) \,.$$

Once again $f \in L^2(\Omega)$ is set to the right hand side for the homogeneous state constrained with $y = y_d$ and $u = 0$.

Case 3

$$\begin{aligned}\min_{(y,u)} \ & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y + y + |y| - u = f \ \text{ in } \Omega \\ & y = 0 \ \text{ on } \partial\Omega \,, \\ \text{with } y_d(x_1, x_2) = & \exp(-10(x_1 - \tfrac{1}{4})^3 + (x_2 - \tfrac{1}{2})^2)\sin(2\pi x_1)\sin(2\pi x_2) \,,\end{aligned}$$

and $f \in L^2(\Omega)$ set to the right hand side for the homogeneous state constrained.

Case 4

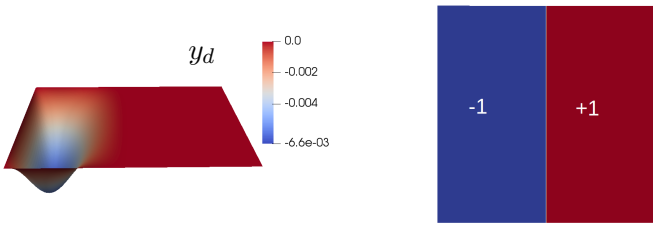
$$\begin{aligned}\min_{(y,u)} \ & \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 \\ \text{s.t.} \quad & -\Delta y - \max(|y - \beta_1| - \beta_2 y, 0) - u = f \ \text{ in } \Omega \\ & y = 0 \ \text{ on } \partial\Omega \,,\end{aligned}$$

with $\beta_1, \beta_2 \geqslant 1$, $y_d(x_1, x_2) = \sin(\pi x_1)\sin(2\pi x_2)$ and $f \in L^2(\Omega)$ set to the right hand side for the homogeneous PDE.

The numerical results for these cases, considering different values of the mesh size denoted by h, the penalty parameter α for the control in the objective functional, and the penalty parameter ν in the bi-quadratic penalty term, are presented in Tab. 2–5. A commonly used method for solving such non-smooth problems are semi-smooth Newton-like methods. Therefore, we also provide a comparison with results obtained with a semi-smooth Newton approach. It can be observed that in almost all cases only a few Newton iterations are needed to solve the problem and to compute the stationary point.
Case 1 (a) and (b) represent examples already studied in [9]. The parameters were adopted accordingly and the mesh size was reconstructed to match the one used in [9] in the best possible way. Tab. 2 shows a comparison between the non-regularized approach presented here and the proposed semi-smooth Newton's method in [9]. It can be observed that in the more involved example, according to [9], the approach presented here requires only one single Newton step to compute the optimal solution. The semi-smooth Newton method on the other hand requires an average of three to five Newton steps for the considered problem.
Furthermore, the quality of the resulting approximation is examined by the relative error $\|y_h - y\|_{L^2} / \|y\|_{L^2}$. The fact that CALi converges in very few Newton steps is mainly due to the fact that the reformulation described in Ass. 2.8 allows to exploit as much information as possible given by the optimization problem and in particular by the given desired state y_d. The initial choice of $\sigma = (\sigma_1, \ldots, \sigma_s)$ motivated by the optimally reachable desired state already provides the perfect guess for the σ_i.
It is important to highlight that having knowledge about the optimal CAL problem formulation given by σ^* for the examples considered here, since y_d is optimally reachable and using this to define the initial values accordingly, the operator $r(\bar{\sigma}^d)$ given by Eq. (40) is always non-negative.

			CALi		[9]
h	α	ν	$\frac{\|y_d - y_h\|_{L^2}}{\|y_d\|_{L^2}}$	# Newton	# Newton
3.009e-02	1e-4	100	1.787e-04	1	3
1.537e-02	1e-4	100	4.655e-05	1	3
7.728e-03	1e-4	100	1.176e-05	1	3
3.885e-03	1e-4	100	2.973e-06	1	3
2.828e-03	1e-4	100	1.576e-06	1	-
7.728e-03	1e-4	1e-3	1.176e-05	1	3
7.728e-03	1e-4	1e-2	1.176e-05	1	3
7.728e-03	1e-4	1e-1	1.176e-05	1	3
7.728e-03	1e-4	1.0	1.176e-05	1	3
7.728e-03	1e-4	50	1.176e-05	1	3
7.728e-03	1e-4	500	1.176e-05	1	3
7.728e-03	1e-2	100	1.202e-04	1	3
7.728e-03	1e-3	100	3.197e-05	1	3
7.728e-03	1e-6	100	5.697e-07	1	3
7.728e-03	1e-8	100	1.255e-07	1	3

Table 1: Numerical results in Case 1.



$\bar{\sigma}^d = \text{sign}(y_d)$

Figure 5: Desired state y_d which coincides with the optimal solution y and the corresponding sign on $\Omega = (0,1)^2$ for Eq. (1).

Furthermore, Tab. 1 and 2 also reveal that the penalty parameter ν has no apparent effect on the solution or even the number of Newton steps needed. This is due to the fact, that the desired state is optimally reachable and hence the corresponding choice of $\bar{\sigma}^d$ yields $\bar{\sigma}^d z = \text{abs}(z)$. Consequently, the penalty term is zero. This observation was also made in all other test cases. Hence, neither an asymptotic approximation nor a sophisticated selection of the penalty parameter ν is necessary. Thus, in all remaining examples considered here, we always choose $\nu \equiv 100$ without further adjustment.

			CALi		[9]
h	α	ν	$\frac{\|y_d - y_h\|_{L^2}}{\|y_d\|_{L^2}}$	# Newton	# Newton
3.009e-02	1e-4	50	5.764e-04	1	4
1.537e-02	1e-4	50	1.514e-04	1	5
7.728e-03	1e-4	50	3.790e-05	1	3
3.885e-03	1e-4	50	9.663e-06	1	3
3.009e-02	1e-4	100	5.764e-04	1	4
1.537e-02	1e-4	100	1.514e-04	1	5
7.728e-03	1e-4	100	3.790e-05	1	3
3.885e-03	1e-4	100	9.663e-06	1	3
7.728e-03	1e-4	1e-3	3.790e-05	1	3
7.728e-03	1e-4	1e-2	3.790e-05	1	3
7.728e-03	1e-4	1e-1	3.790e-05	1	3
7.728e-03	1e-4	1.0	3.790e-05	1	3
7.728e-03	1e-4	500	3.790e-05	1	3
7.728e-03	1e-2	100	8.106e-05	1	2
7.728e-03	1e-3	100	6.609e-05	1	2
7.728e-03	1e-5	100	1.237e-05	1	5
7.728e-03	1e-6	100	3.056e-06	1	no conv.

Table 2: Numerical results in Case 1.

It should be emphasized that in each test case, it is verified that the condition $\sigma z = \text{abs}(z)$ in the integral sense holds for the resulting z. This has been computed for the cases 2–4, where the maximal value of $\|\sigma_i z_i - |z_i|\|_{L^2}$ for $i = 1, 2$ is always fairly close to zero, i.e., usually less than $1e-7$.

h	α	Objective	$\frac{\|y - y_h\|_{L^2}}{\|y\|_{L^2}}$	$\|\sigma z - \|z\|\|_{L^2}$	#Newt.
3.009e-02	1e-04	6.466e-06	7.843e-03	7.8e-07	2
1.537e-02	1e-04	4.441e-07	2.051e-03	2.4e-06	2
7.727e-03	1e-04	2.868e-08	5.191e-04	1.2e-08	2
3.885e-03	1e-04	1.971e-09	1.325e-04	3.9e-08	2
2.828e-03	1e-04	6.431e-10	7.231e-05	1.5e-08	2
7.727e-03	1e-2	3.985e-08	6.658e-04	1.2e-08	2
7.727e-03	1e-3	3.758e-08	6.251e-04	1.2e-08	2
7.727e-03	1e-5	1.510e-08	3.314e-04	1.2e-08	2
7.727e-03	1e-6	6.464e-09	2.193e-04	1.2e-08	2
7.727e-03	1e-7	3.118e-09	1.591e-04	1.2e-08	2

Table 3: Numerical results for case 2 with $\beta = 0.01$.

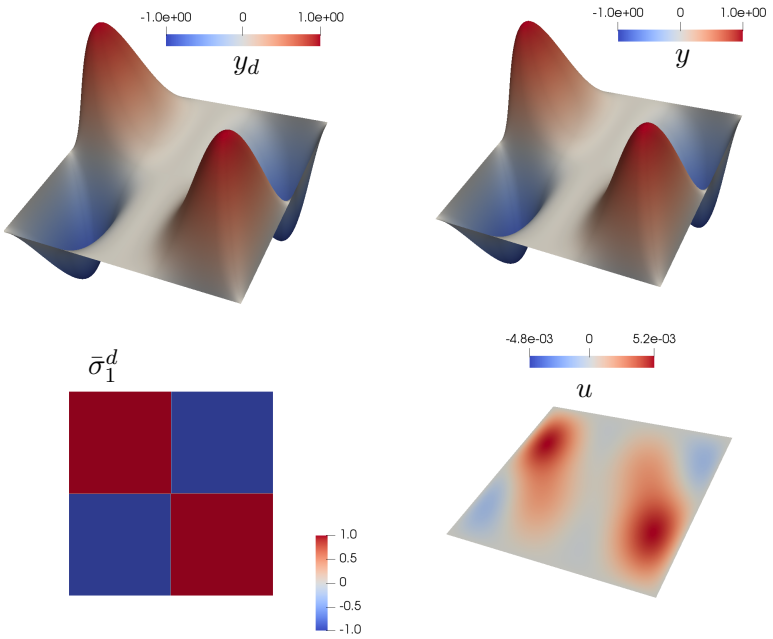


Figure 6: Desired state y_d (top left) which coincides with the optimal solution y_* (top right), the corresponding sign $\bar{\sigma}_1^d = \text{sign}(y_d)$ on $\Omega = (0,1)^2$ and the optimal control (bottom right) for Case 2.

h	α	$\frac{\|y-y_h\|_{L^2}}{\|y\|_{L^2}}$	$\|\sigma z - \|z\|\|_{L^2}$	#Newt.
3.009e-02	1e-04	4.426e-03	9e-07	1
1.537e-02	1e-04	1.056e-03	3e-06	1
7.728e-03	1e-04	2.920e-04	1e-08	1
3.885e-03	1e-04	7.382e-05	4e-08	1
2.828e-03	1e-04	3.912e-05	2e-08	1
7.728e-03	1e-02	4.340e-04	1e-08	1
7.728e-03	1e-03	4.064e-04	1e-08	1
7.728e-03	1e-05	1.138e-04	1e-08	1
7.728e-03	1e-06	4.336e-05	1e-08	1
7.728e-03	1e-07	2.989e-05	1e-08	1

Table 4: Numerical results for case 3.

h	α	$\frac{\|y-y_h\|_{L^2}}{\|y\|_{L^2}}$	$\max_{i=1,2}\{\|\sigma_i z_i - \|z_i\|\|_{L^2}\}$	#Newt.
3.009e-02	1e-04	1.985e-04	6e-05	2
1.537e-02	1e-04	4.528e-05	3e-05	2
7.728e-03	1e-04	1.132e-05	3e-06	2
3.885e-03	1e-04	2.858e-06	3e-07	2
2.828e-03	1e-04	1.017e-06	1e-07	2
7.728e-03	1e-02	4.125e-05	3e-06	2
7.728e-03	1e-03	3.030e-05	3e-06	2
7.728e-03	1e-05	2.641e-06	3e-06	2
7.728e-03	1e-06	6.348e-07	3e-06	2
7.728e-03	1e-07	4.287e-07	3e-06	2

Table 5: Numerical results for case 4 with nested nonsmoothness and $\beta_1 = 1.0, \beta_2 = 2.0$.

As additional observation, the numerical results suggest a further special property of the CAli algorithm, namely mesh independence. Regardless of the mesh size of the discretization, the behavior for the relative error $\|y_d - y\|_{L^2} / \|y_d\|_{L^2}$ with respect to different parameters α remains the same. The mesh independence can also be observed in Tab. 2 for Case 1 since independent of the mesh size only one Newton step is required. According to [1], for the Newton method in the case of smooth optimization, the fact that the algorithm always requires the same number of iterations, regardless of the mesh size, is known as *strong mesh independence*. Moreover, it is clearly evident that in each parameter setting the desired condition $\bar{\sigma}_i z_i = \text{abs}(z_i)$ for $i \in \mathbb{N}_s$ is met in the integral sense.

6 Conclusion and Outlook

We presented a new approach based on constant abs-linearization for the solution of optimization problems constrained by non-smooth PDEs. For the considered class of genuinely non-smooth problems, this approach enables the optimization without any substitute assumptions for the non-smoothness. The key idea is to consider closely related but smooth optimization problems and exploit the domain decomposition given by constant abs-linearization. Selective choice of the CAL problem, given by the sign of the optimally reachable desired state, generates a suitable reformulation of the original problem which yields the same optimal solution, but can be solved using conventional methods for smooth optimization problems. Optimality conditions for the considered formulations were derived and discussed. By treating the inequality condition with a bi-quadratic penalty approach the sign condition could easily be incorporated into the algorithmic framework. The type of discretization employed here was also presented and critically examined. Finally, several non-smooth PDE-constrained problems that fit into the considered setting were discussed. The corresponding numerical results clearly show also the resulting mesh independence of the presented method.
We have shown that with the proposed decomposition strategy, the underlying PDE with known optimal decomposition given by the signs of the desired state can be optimally solved in a reduced and minimal number of Newton steps. A part of continued research, remains the development of an efficient extension of the proposed approach for optimal control problems with given y_d which is not optimally reachable and hence unknown optimal $\bar{\sigma}$.
Furthermore, the existence and uniqueness of solutions to the PDE with ℓ such that $\hat{\ell}(y, \bar{\sigma} z)$ not necessary affine-linear remain the subject of current research. An additional aspect for future research is the investigation of a completely regularization-free solution method since the presented approach is not completely without regularization due to the reformulated inequality constraint on the switching and sign functions by means of a penalty regularization.
Moreover, it has been discussed, that in the optimally reachable case the penalty term vanishes

and the condition Eq. (48) satisfied, such that no limit analysis has to be performed. However, note that in the non-optimally reachable cases where the condition Eq. (48) might not be fulfilled, only a statement about the selected penalty parameter ν can be made, but no efficient strategy for switching the fixed σ_i can be derived. Hence, effective strategies for switching between different CAL problems remain subject of further research.

Acknowledgments

This work was partially funded by the DFG priority program SPP 1962 within the project "Shape Optimization for Maxwell's Equations Including Hysteresis Effects in the Material Laws (HyLa)".

References

[1] E.L. Allgower, K. Bohmer, F.A. Potra, and W.C. Rheinboldt. A mesh-independence principle for operator equations and their discretizations. *SIAM Journal on Numerical Analysis*, 23(1):160–169, 1986.

[2] I. Antal and J. Karátson. A mesh independent superlinear algorithm for some nonlinear nonsymmetric elliptic systems. *Computers & Mathematics with Applications*, 55(10):2185–2196, 2008.

[3] J. Appell and P.P Zabrejko. *Nonlinear Superposition Operators.* Cambridge Tracts in Mathematics. Cambridge University Press, 1990.

[4] V. Barbu. *Optimal Control of Variational Inequalities.* Pitman, Boston, 1984.

[5] F.E. Browder. Problèmes non linéaires, Séminaire de Mathématiques Supérieures. *Presses Univ. Montréal*, (15), 1966.

[6] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM Journal on Control and Optimization*, 24(6):1309–1318, 1986.

[7] E. Casas. Boundary control of semilinear elliptic equations with pointwise state constraints. *SIAM Journal on Control and Optimization*, 31(4):993–1006, 1993.

[8] E. Casas, De Los Reyes, J.C., and F. Tröltzsch. Sufficient second-order optimality conditions for semilinear control problems with pointwise state constraints. *SIAM Journal on Optimization*, 19(2):616–643, 2008.

[9] C. Christof, C. Clason, C. Meyer, and S. Walther. Optimal control of a non-smooth semilinear elliptic equation. *Mathematical Control and Related Fields*, 8(1):247–276, 2018.

[10] J.C. De Los Reyes. On the optimal control of some nonsmooth distributed parameter systems arising in mechanics. *GAMM-Mitteilungen*, 40(4):268–286, 2018.

[11] S. Fiege, A. Walther, and A. Griewank. An algorithm for nonsmooth optimization by successive piecewise linearization. *Mathematical Programming, Series A*, 2018. DOI: 10.1007/s10107-018-1273-5.

[12] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Classics in Mathematics. Springer-Verlag, Berlin, 2001.

[13] A. Griewank. On stable piecewise linearization and generalized algorithmic differentiation. *Optimization Methods and Software*, 28(6):1139–1178, 2013.

[14] A. Griewank and A. Walther. First and second order optimality conditions for piecewise smooth objective functions. *Optimization Methods and Software, 31(5):904-930*, 2016.

[15] A. Griewank and A. Walther. Relaxing kink qualifications and proving convergence rates in piecewise smooth optimization. *SIAM Journal on Optimization*, 29(1):262–289, 2019.

[16] M. Hintermüller and M. Ulbrich. A mesh-independence result for semismooth Newton methods. *Mathematical Programming*, 101(1):151–184, 2004.

[17] K. Ito and K. Kunisch. Semi-smooth newton methods for state-constrained optimal control problems. *Syst. Control Lett.*, 50:221–228, 2003.

[18] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer, 2007.

[19] F. Kikuchi, K. Nakazato, and T. Ushijima. Finite element approximation of a nonlinear eigenvalue problem related to MHD equilibria. *Japan Journal of Applied Mathematics*, 1(2):369–403, 1984.

[20] A. Logg, K.-A. Mardal, and G. Wells. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.

[21] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of pdes with regularized pointwise state constraints. *Computational Optimization and Applications*, 33:209–228, 2006.

[22] A. Schiela. Barrier methods for optimal control problems with state constraints. *SIAM Journal on Optimization*, 20(2):1002–1031, 2009.

[23] R. Temam. A non-linear eigenvalue problem: the shape at equilibrium of a confined plasma. *Archive for Rational Mechanics and Analysis*, 60:51–73, 1975.

[24] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods, and Applications*. Graduate Studies in Mathematics. American Mathematical Society, 2010.

[25] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Society for Industrial and Applied Mathematics, 2011.

[26] M. Ulbrich and S. Ulbrich. Primal-dual interior-point methods for PDE-constrained optimization. *Mathematical Programming*, 117(1):435–485, 2009.

[27] A. Walther, O. Weiß, A. Griewank, and S. Schmidt. Nonsmooth optimization by successive abs-linearisation in function spaces. *Applicable Analysis*, 2020. DOI: 10.1080/00036811.2020.1738397.

[28] O. Weiß. *Nonsmooth Optimization by Abs-Linearization in Reflexive Function Spaces*. PhD thesis, Humboldt-Universität zu Berlin, 2020. In preparation.

[29] E. Zeidler. *Applied Functional Analysis. Applications to Mathematical Physics*. Springer, 1995.