DFG Deutsche Forschungsgemeinschaft
Priority Programme 1962

Optimal Control of a Non-smooth Semilinear Elliptic Equation

Constantin Christof, Christian Clason, Christian Meyer, Stephan Walther

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

Preprint Number SPP1962-020

received on May 2, 2017

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

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

OPTIMAL CONTROL OF A NON-SMOOTH SEMILINEAR ELLIPTIC EQUATION

Constantin Christof* Christian Clason† Christian Meyer*,‡
Stephan Walther*

May 2, 2017

Abstract This paper is concerned with an optimal control problem governed by a non-smooth semilinear elliptic equation. We show that the control-to-state mapping is directionally differentiable and precisely characterize its Bouligand subdifferential. By means of a suitable regularization, first-order optimality conditions including an adjoint equation are derived and afterwards interpreted in light of the previously obtained characterization. In addition, the directional derivative of the control-to-state mapping is used to establish strong stationarity conditions. While the latter conditions are shown to be stronger, we demonstrate by numerical examples that the former conditions are amenable to numerical solution using a semi-smooth Newton method.

1 INTRODUCTION

In this paper, we consider the following non-smooth semilinear elliptic optimal control problem

$$\left.\begin{aligned} &\min_{u\in L^2(\Omega),\, y\in H_0^1(\Omega)} \quad J(y,u) \\ &\text{s.t.} \quad -\Delta y + \max(0,y) = u \ \text{ in } \Omega, \end{aligned}\right\} \tag{P}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is a bounded domain and J is a smooth objective; for precise assumptions on the data, we refer to Assumption 1.1 below.

The salient feature of (P) is of course the occurrence of the non-smooth max-function in the equality constraint in (P). This causes the associated control-to-state mapping $u \mapsto y$ to be non-smooth as well, and hence standard techniques for obtaining first-order necessary optimality conditions that are based on the adjoint of the Gâteaux-derivative of the control-to-state mapping cannot be applied. One remedy to cope with this challenge is to apply generalized differentiability concepts for the derivation of optimality conditions. Such concepts and the resulting conditions can be roughly grouped in two classes:

*TU Dortmund, Faculty of Mathematics, Vogelpothsweg 87, 44227 Dortmund, Germany, (constantin.christof@tu-dortmund.de, christian.meyer@math.tu-dortmund.de, stephan.walther@tu-dortmund.de)
†University of Duisburg-Essen, Faculty of Mathematics, Thea-Leymann-Str. 9, 45127 Essen, Germany (christian.clason@uni-due.de)
‡Corresponding author

(i) (generalized) directional derivatives, leading to "purely primal" optimality conditions stating that the directional derivatives of the reduced objective in feasible directions are non-negative;

(ii) various notions of subdifferentials, leading to abstract optimality conditions stating that zero is contained in a suitable subdifferential of the reduced objective.

For a general treatment of generalized derivatives and their relation, we only refer to [21, Chap. 10A], [4, Prop. 2.3.2], [23, Prop. 7.3.6 and 9.1.5], and the references therein. However, with the exception of the convex setting, concrete (and in particular, numerically tractable) characterizations of these abstract conditions are only available in a very restricted set of situations; see, e.g., [5, 7, 13, 18]. Therefore, an alternative approach to obtaining optimality conditions has become popular in optimal control of non-smooth partial differential equations (PDEs) and variational inequalities (VIs), which is based on regularization and relaxation schemes that allow standard adjoint-based optimality conditions combined with a limit analysis; see, e.g., [1, 2, 9, 10, 22, 24]. However, in particular in infinite dimensions, it is often unclear whether conditions obtained through such a limit process can be interpreted as optimality conditions in the sense of non-smooth analysis (whether of type (i) or (ii)), and hence their relative strength compared to those types of conditions is an open issue. To the best of our knowledge, the only contributions in this direction dealing with infinite-dimensional optimal control problems are [8, 14–17, 19, 26], where strong stationarity conditions are derived that are shown to be equivalent to purely primal optimality conditions of type (i). In particular, [15] is the only work treating optimal control of non-smooth PDEs, including a semilinear parabolic PDE similar to the one in (P). However, concerning optimality conditions obtained via regularization and their comparison to conditions of type (ii), we are not aware of any contribution dealing with the optimal control of either non-smooth PDEs or VIs.

The aim of our paper is to answer this question for the particular optimal control problem (P). For this purpose, we turn our attention to the Bouligand subdifferential of the control-to-state mapping, defined as the set of limits of Jacobians of smooth points in the spirit of, e.g., [20, Def. 2.12] or [11, Sec. 1.3]. Note that in infinite-dimensional spaces, one has to pay attention to the topology underlying these limit processes so that multiple notions of Bouligand subdifferentials arise, see Definition 3.1 below. We will precisely characterize these subdifferentials and use this result to interpret the optimality conditions arising in the regularization limit. We emphasize that the regularization and the associated limit analysis is fairly straightforward. The main contribution of our work is the characterization of the Bouligand subdifferential as a set of linear PDE solution operators; see Theorem 3.16. This characterization allows a comparison of the optimality conditions obtained by regularization with standard optimality conditions of type (ii) (specifically, involving Bouligand and Clarke subdifferentials), which shows that the former are surprisingly strong; cf. Theorem 4.7. On the other hand, it is well-known that one loses information in the regularization limit, and the same is observed in case of (P). In order to see this, we establish another optimality system, which is equivalent to a purely primal optimality condition of type (i). It will turn out that the optimality system derived in this way is indeed stronger than the one obtained via regularization, since it contains an additional sign condition for the adjoint state. It is, however, not clear how to solve these strong stationarity conditions numerically. In contrast to this, the optimality system arising in the regularization limit allows

a reformulation as a non-smooth equation that is amenable for semi-smooth Newton methods. We emphasize that we do not employ the regularization procedure for numerical computations, but directly solve the limit system instead. Our work includes first steps into this direction, but the numerical results are preliminary and give rise to future research.

Let us finally emphasize that our results and the underlying analysis are in no way limited to the PDE in (P). Instead the arguments can easily be adapted to more general cases, involving a piecewise C^1-function rather than the max-function and a (smooth) divergence-gradient-operator instead of the Laplacian. However, in order to keep the discussion as concise as possible and to be able to focus on the main arguments, we decided to restrict the analysis to the specific PDE under consideration.

The outline of the paper is as follows: This introduction ends with a short subsection on our notation and the standing assumptions. We then turn to the control-to-state mapping, showing that it is globally Lipschitz and directionally differentiable and characterizing points where it is Gâteaux-differentiable. Section 3 is devoted to the characterization of the Bouligand subdifferentials. We first state necessary conditions that elements of the subdifferentials have to fulfill. Afterwards, we prove that these are also sufficient, which is by far more involved compared to showing their necessity. In Section 4, we first shortly address the regularization and the corresponding limit analysis. Then, we compare the optimality conditions arising in the regularization limit with our findings from Section 3. The section ends with the derivation of the strong stationarity conditions. Section 5 deals with the numerical solution of the optimality system derived via regularization. The paper ends with an appendix containing some technical lemmas whose proofs are difficult to find in the literature.

1.1 NOTATION AND STANDING ASSUMPTIONS

Let us shortly address the notation used throughout the paper. By $\mathbb{1}_M$ we denote the characteristic function of a set $M \subset \mathbb{R}^d$. By λ^d we denote the d-dimensional Lebesgue measure. Given a (Borel-) measurable function $v : \Omega \to \mathbb{R}$, we abbreviate the set $\{x \in \Omega : v(x) = 0\}$ by $\{v = 0\}$. The sets $\{v > 0\}$ and $\{v < 0\}$ are defined analogously. As usual, the Sobolev space $H_0^1(\Omega)$ is defined as the closure of $C_c^\infty(\Omega)$ with respect to the H^1-norm. Moreover, we define the space

$$Y := \{y \in H_0^1(\Omega) : \Delta y \in L^2(\Omega)\}.$$

Equipped with the scalar product

$$(y, v)_Y := \int_\Omega \left(\Delta y\, \Delta v + \nabla y \cdot \nabla v + y\, v\right) \mathrm{d}x,$$

Y becomes a Hilbert space. Herein and throughout the paper, $\Delta = \operatorname{div} \circ \nabla : H_0^1(\Omega) \to H^{-1}(\Omega)$ denotes the distributional Laplacian. The space Y is compactly embedded in $H_0^1(\Omega)$, since, for any sequence $(y_n) \subset Y$ with $y_n \rightharpoonup y$ in Y, we have

$$\|y_n - y\|_{H^1(\Omega)}^2 = -\int_\Omega \Delta(y_n - y)\,(y_n - y)\,dx + \|y_n - y\|_{L^2(\Omega)}^2 \to 0$$

by the compact embedding of $H_0^1(\Omega)$ into $L^2(\Omega)$. Since Y is isometrically isomorphic to the subset

$$\{(y, \omega, \delta) \in L^2(\Omega; \mathbb{R}^{d+2}) : \exists\, v \in H_0^1(\Omega) \text{ with } y = v, \omega = \nabla v, \delta = \Delta v \text{ a.e. in } \Omega\}$$

of the separable space $L^2(\Omega;\mathbb{R}^{d+2})$, it is separable as well. With a little abuse of notation, we will denote the Nemytskii operator induced by the max-function (with different domains and ranges) by the same symbol. In the same way, we will denote by $\max'(y;h)$ the directional derivative of $y \mapsto \max(0,y)$ in the point y in direction h both considered as a scalar function and as the corresponding Nemytskii operator.

Throughout the paper, we will make the following standing assumptions.

Assumption 1.1. The set $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, is a bounded domain. The objective functional $J : Y \times L^2(\Omega) \to \mathbb{R}$ in (P) is weakly lower semicontinuous and continuously Fréchet-differentiable.

Note that we do not impose any regularity assumptions on the boundary of Ω.

2 DIRECTIONAL DIFFERENTIABILITY OF THE CONTROL-TO-STATE MAPPING

We start the discussion of the optimal control problem (P) by investigating its PDE constraint, showing that it is uniquely solvable and that the associated solution operator is directionally differentiable.

Proposition 2.1. *For all $u \in H^{-1}(\Omega)$, there exists a unique solution $y \in H^1_0(\Omega)$ to*

$$-\Delta y + \max(0,y) = u. \tag{PDE}$$

Moreover, the solution operator $S : u \mapsto y$ associated with (PDE) is well-defined and globally Lipschitz continuous as a function from $L^2(\Omega)$ to Y.

Proof. The arguments are standard. First of all, Browder and Minty's theorem on monotone operators yields the existence of a unique solution in $H^1_0(\Omega)$. If $u \in L^2(\Omega)$, then a simple bootstrapping argument implies $y \in Y$. The Lipschitz continuity finally follows from the global Lipschitz continuity of the max-operator. □

Theorem 2.2 (directional derivative of S). *Let $u, h \in L^2(\Omega)$ be arbitrary but fixed, set $y := S(u) \in Y$, and let $\delta_h \in Y$ be the unique solution to*

$$-\Delta \delta_h + \mathbb{1}_{\{y=0\}} \max(0,\delta_h) + \mathbb{1}_{\{y>0\}} \delta_h = h. \tag{2.1}$$

Then it holds

$$h_n \rightharpoonup h \ \text{ in } L^2(\Omega),\ t_n \to 0^+ \quad\Longrightarrow\quad \frac{S(u+t_n h_n) - S(u)}{t_n} \rightharpoonup \delta_h \ \text{ in } Y$$

and

$$h_n \to h \ \text{ in } L^2(\Omega),\ t_n \to 0^+ \quad\Longrightarrow\quad \frac{S(u+t_n h_n) - S(u)}{t_n} \to \delta_h \ \text{ in } Y.$$

In particular, the solution operator $S : L^2(\Omega) \to Y$ associated with (PDE) is Hadamard directionally differentiable in all points $u \in L^2(\Omega)$ in all directions $h \in L^2(\Omega)$, with $S'(u;h) = \delta_h \in Y$.

Proof. First observe that for every $h \in L^2(\Omega)$, (2.1) admits a unique solution $\delta_h \in Y$ by exactly the same arguments as in the proof of Proposition 2.1. Note moreover that (2.1) is equivalent to

$$-\Delta\delta_h + \max'(y;\delta_h) = h.$$

Now let $u, h \in L^2(\Omega)$ be arbitrary but fixed and let $(t_n) \in (0,\infty)$ and $(h_n) \in L^2(\Omega)$ be sequences with $t_n \to 0$ and $h_n \rightharpoonup h$ in $L^2(\Omega)$. We abbreviate $y_n := S(u + t_n h_n) \in Y$. Subtracting the equations for y and δ_h from the one for y_n yields

$$\begin{aligned} -\Delta\Big(\frac{y_n - y}{t_n} - \delta_h\Big) = h_n - h + \frac{\max(0, y + t_n\delta_h) - \max(0, y_n)}{t_n} \\ - \Big(\frac{\max(0, y + t_n\delta_h) - \max(0, y)}{t_n} - \max'(y;\delta_h)\Big). \end{aligned} \tag{2.2}$$

Testing this equation with $(y_n - y)/t_n - \delta_h$ and using the monotonicity of the max-operator leads to

$$\Big\|\frac{y_n - y}{t_n} - \delta_h\Big\|_{H^1(\Omega)} \le \|h_n - h\|_{H^{-1}(\Omega)} + \Big\|\frac{\max(0, y + t_n\delta_h) - \max(0, y)}{t_n} - \max'(y;\delta_h)\Big\|_{L^2(\Omega)}.$$

Now the compactness of $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ and the directional differentiability of $\max : L^2(\Omega) \to L^2(\Omega)$ (which directly follows from the directional differentiability of $\max : \mathbb{R} \to \mathbb{R}$ and Lebesgue's dominated convergence theorem) give

$$\frac{y_n - y}{t_n} - \delta_h \to 0 \quad \text{in } H^1_0(\Omega). \tag{2.3}$$

As $\max : L^2(\Omega) \to L^2(\Omega)$ is also Lipschitz continuous and thus Hadamard-differentiable, (2.3) implies

$$\frac{\max(0, y_n) - \max(0, y)}{t_n} - \max'(y;\delta_h) \to 0 \quad \text{in } L^2(\Omega). \tag{2.4}$$

Hence, (2.2) yields that the sequence $(y_n - y)/t_n - \delta_h$ is bounded in Y and thus (possibly after transition to a subsequence) converges weakly. Because of (2.3) the weak limit is zero and therefore unique, so that the whole sequence converges weakly. This implies the first assertion. If now h_n converges strongly to h in $L^2(\Omega)$, then (2.2) and (2.4) yield $(y_n - y)/t_n - \delta_h \to 0$ in Y, which establishes the second claim. □

Theorem 2.2 allows a precise characterization of points where S is Gâteaux-differentiable. This will be of major importance for the study of the Bouligand subdifferentials in the next section.

Corollary 2.3 (characterization of Gâteaux-differentiable points). *The solution operator $S : L^2(\Omega) \to Y$ is Gâteaux-differentiable in a point $u \in L^2(\Omega)$, i.e., $S'(u;\cdot) \in \mathcal{L}(L^2(\Omega), Y)$ if and only if the solution $y = S(u)$ satisfies $\lambda^d(\{y = 0\}) = 0$. If the latter is the case, then the directional derivative $\delta_h = S'(u;h) \in Y$ in a direction $h \in L^2(\Omega)$ is uniquely characterized as the solution to*

$$-\Delta\delta_h + \mathbb{1}_{\{y>0\}}\delta_h = h. \tag{2.5}$$

Proof. In view of (2.1), it is clear that if $\lambda^d(\{y = 0\}) = 0$, then S is Gâteaux-differentiable and the Gâteaux derivative is the solution operator for (2.5). It remains to prove that Gâteaux-differentiability implies $\lambda^d(\{y = 0\}) = 0$. To this end, let $u \in L^2(\Omega)$ be a point where S is Gâteaux-differentiable. Then it follows for all $h \in L^2(\Omega)$ that $S'(u;h) = -S'(u;-h)$, which by (2.1) in turn implies

$$\mathbb{1}_{\{y=0\}} \max(0, -S'(u;h)) + \mathbb{1}_{\{y=0\}} \max(0, S'(u;h)) = \mathbb{1}_{\{y=0\}} |S'(u;h)| = 0. \tag{2.6}$$

Consider now a function $\psi \in C^\infty(\mathbb{R}^d)$ with $\psi > 0$ in Ω and $\psi \equiv 0$ in $\mathbb{R}^d \setminus \Omega$, whose existence is ensured by Lemma A.1. Since $h \in L^2(\Omega)$ was arbitrary, we are allowed to choose

$$h := -\Delta\psi + \mathbb{1}_{\{y>0\}}\psi + \mathbb{1}_{\{y=0\}} \max(0, \psi) \in L^2(\Omega),$$

so that $S'(u;h) = \psi$ by virtue of (2.1). Consequently, we obtain from (2.6) that $\mathbb{1}_{\{y=0\}}\psi = 0$. Since $\psi > 0$ in Ω, this yields $\lambda^d(\{y = 0\}) = 0$ as claimed. □

3 BOULIGAND SUBDIFFERENTIALS OF THE CONTROL-TO-STATE MAPPING

This section is devoted to the main result of our work, namely the precise characterization of the Bouligand subdifferentials of the PDE solution operator S from Proposition 2.1.

3.1 DEFINITIONS AND BASIC PROPERTIES

We start with the rigorous definition of the Bouligand subdifferential. In the spirit of [20, Def. 2.12], it is defined as the set of limits of Jacobians of differentiable points. However, in infinite dimensions, we have of course to distinguish between different topologies underlying this limit process, as already mentioned in the introduction. This gives rise to the following

Definition 3.1 (Bouligand subdifferentials of S). Let $u \in L^2(\Omega)$ be given. Denote the set of smooth points of S by

$$D := \{v \in L^2(\Omega) : S : L^2(\Omega) \to Y \text{ is Gâteaux-differentiable in } v\}.$$

In what follows, we will frequently call points in D Gâteaux points.

(i) The *weak-weak Bouligand subdifferential* of S in u is defined by

$$\begin{aligned}\partial_B^{ww} S(u) := \{G \in \mathcal{L}(L^2(\Omega), Y) : &\text{ there exists } (u_n) \subset D \text{ such that} \\ &u_n \rightharpoonup u \text{ in } L^2(\Omega) \text{ and } S'(u_n)h \rightharpoonup G\,h \text{ in } Y \text{ for all } h \in L^2(\Omega)\}.\end{aligned}$$

(ii) The *weak-strong Bouligand subdifferential* of S in u is defined by

$$\begin{aligned}\partial_B^{ws} S(u) := \{G \in \mathcal{L}(L^2(\Omega), Y) : &\text{ there exists } (u_n) \subset D \text{ such that} \\ &u_n \rightharpoonup u \text{ in } L^2(\Omega) \text{ and } S'(u_n)h \to G\,h \text{ in } Y \text{ for all } h \in L^2(\Omega)\}.\end{aligned}$$

(iii) The *strong-weak Bouligand subdifferential* of S in u is defined by

$$\partial_B^{sw} S(u) := \{G \in \mathcal{L}(L^2(\Omega), Y) : \text{there exists } (u_n) \subset D \text{ such that } u_n \to u \text{ in } L^2(\Omega) \text{ and } S'(u_n)h \rightharpoonup G\,h \text{ in } Y \text{ for all } h \in L^2(\Omega)\}.$$

(iv) The *strong-strong Bouligand subdifferential* of S in u is defined by

$$\partial_B^{ss} S(u) := \{G \in \mathcal{L}(L^2(\Omega), Y) : \text{there exists } (u_n) \subset D \text{ such that } u_n \to u \text{ in } L^2(\Omega) \text{ and } S'(u_n)h \to G\,h \text{ in } Y \text{ for all } h \in L^2(\Omega)\}.$$

Remark 3.1. Based on the generalization of Rademacher's theorem to Hilbert spaces (see [16]) and the generalization of Alaoglu's theorem to the weak operator topology, one can show that $\partial_B^{ww} S(u)$ and $\partial_B^{sw} S(u)$ are non-empty for every $u \in L^2(\Omega)$; see also [6]. In contrast to this, it is not clear a priori if $\partial_B^{ws} S(u)$ and $\partial_B^{ss} S(u)$ are non-empty, too. However, Theorem 3.16 at the end of this section will imply this as a byproduct.

From the definitions, we obtain the following useful properties.

Lemma 3.2.

(i) For all $u \in L^2(\Omega)$ it holds

$$\partial_B^{ss} S(u) \subseteq \partial_B^{sw} S(u) \subseteq \partial_B^{ww} S(u) \quad \text{and} \quad \partial_B^{ss} S(u) \subseteq \partial_B^{ws} S(u) \subseteq \partial_B^{ww} S(u).$$

(ii) If S is Gâteaux-differentiable in $u \in L^2(\Omega)$, then it holds $S'(u) \in \partial_B^{ss} S(u)$.

(iii) For all $u \in L^2(\Omega)$ and all $G \in \partial_B^{ww} S(u)$, it holds

$$\|G\|_{\mathcal{L}(L^2(\Omega), Y)} \leq L,$$

where $L > 0$ is the Lipschitz constant of $S : L^2(\Omega) \to Y$.

Proof. Parts (i) and (ii) immediately follow from the definition of the Bouligand subdifferentials (to see (ii), just choose $u_n := u$ for all n). In order to prove part (iii), observe that the definition of $\partial_B^{ww} S(u)$ implies the existence of a sequence of Gâteaux points $u_n \in L^2(\Omega)$ such that $u_n \rightharpoonup u$ in $L^2(\Omega)$ and $S'(u_n)h \rightharpoonup Gh$ in Y for all $h \in L^2(\Omega)$. For each $n \in \mathbb{N}$, the global Lipschitz continuity of S according to Proposition 2.1 immediately gives $\|S'(u_n)\|_{\mathcal{L}(L^2(\Omega), Y)} \leq L$. Consequently, the weak lower semicontinuity of the norm implies

$$\|Gh\|_Y \leq \liminf_{n \to \infty} \|S'(u_n)h\|_Y \leq L\|h\|_{L^2} \quad \forall\, h \in L^2(\Omega).$$

This yields the claim. □

Remark 3.3. The Bouligand subdifferentials $\partial_B^{ww} S$ and $\partial_B^{sw} S$ do not change if the condition $S'(u_n)h \rightharpoonup G\,h$ in Y in Definition 3.1(i) and (iii) is replaced with either $S'(u_n)h \to G\,h$ in Z or $S'(u_n)h \rightharpoonup G\,h$ in Z, where Z is a normed linear space satisfying $Y \hookrightarrow Z$ such as, e.g., $Z = H^1(\Omega)$ or $Z = L^2(\Omega)$. This is seen as follows: By Lemma 3.2(iii), every sequence $(u_n) \subset D$ contains a subsequence $(S'(u_{n_k})h)$ converging weakly in Y. Thus, if such a sequence converges weakly or strongly in Z, the uniqueness of the (weak) limit in Z implies weak convergence of the whole sequence $(S'(u_n)h)$ in Y.

Next, we show closedness properties of the two strong subdifferentials.

Proposition 3.4 (strong-strong-closedness of $\partial_B^{ss} S$). *Let $u \in L^2(\Omega)$ be arbitrary but fixed. Suppose that*

(i) $u_n \in L^2(\Omega)$ *and* $G_n \in \partial_B^{ss} S(u_n)$ *for all* $n \in \mathbb{N}$,

(ii) $u_n \to u$ *in* $L^2(\Omega)$,

(iii) $G \in \mathcal{L}(L^2(\Omega), Y)$,

(iv) $G_n h \to Gh$ *in* Y *for all* $h \in L^2(\Omega)$.

Then G is an element of $\partial_B^{ss} S(u)$.

Proof. The definition of $\partial_B^{ss} S(u_n)$ implies that for all $n \in \mathbb{N}$, one can find a sequence $(u_{m,n}) \subset L^2(\Omega)$ of Gâteaux points with associated derivatives $G_{m,n} := S'(u_{m,n})$ such that $u_{m,n} \to u_n$ in $L^2(\Omega)$ as $m \to \infty$ and

$$G_{m,n} h \to G_n h \text{ in } Y \text{ for all } h \in L^2(\Omega) \text{ as } m \to \infty.$$

Since $L^2(\Omega)$ is separable, there exists a countable set $\{w_k\}_{k=1}^{\infty} \subseteq L^2(\Omega)$ that is dense in $L^2(\Omega)$. Because of the convergences derived above, it moreover follows that for all $n \in \mathbb{N}$, there exists an $m_n \in \mathbb{N}$ with

$$\|G_{m_n,n} w_k - G_n w_k\|_Y \le \frac{1}{n} \quad \forall\, k = 1, \dots, n \qquad \text{and} \qquad \|u_n - u_{m_n,n}\|_{L^2(\Omega)} \le \frac{1}{n}. \tag{3.1}$$

Consider now a fixed but arbitrary $h \in L^2(\Omega)$, and define

$$h_n^* := \operatorname{argmin}\{\|w_k - h\|_{L^2(\Omega)} : 1 \le k \le n\}.$$

Then the density property of $\{w_k\}_{k=1}^{\infty}$ implies $h_n^* \to h$ in $L^2(\Omega)$ as $n \to \infty$, and we may estimate

$$\begin{aligned} \|G_{m_n,n} h - Gh\|_Y &\le \|G_{m_n,n} h_n^* - G_n h_n^*\|_Y + \|(G_{m_n,n} - G_n)(h_n^* - h)\|_Y + \|G_n h - Gh\|_Y \\ &\le \frac{1}{n} + \|G_{m_n,n} - G_n\|_{\mathcal{L}(L^2,Y)} \|h_n^* - h\|_Y + \|G_n h - Gh\|_Y \to 0 \quad \text{as } n \to \infty, \end{aligned}$$

where the boundedness of $\|G_{m_n,n} - G_n\|_{\mathcal{L}(L^2,Y)}$ follows from Lemma 3.2(iii). The above proves that for all $h \in L^2(\Omega)$, we have $G_{m_n,n} h \to Gh$ in Y. Since $h \in L^2(\Omega)$ was arbitrary and the Gâteaux points $u_{m_n,n}$ satisfy $u_{m_n,n} \to u$ in $L^2(\Omega)$ as $n \to \infty$ by (3.1) and our assumptions, the claim finally follows from the definition of $\partial_B^{ss} S(u)$. □

Proposition 3.5 (strong-weak-closedness of $\partial_B^{sw} S$). *Let $u \in L^2(\Omega)$ be arbitrary but fixed. Assume that:*

(i) $u_n \in L^2(\Omega)$ *and* $G_n \in \partial_B^{sw} S(u_n)$ *for all* $n \in \mathbb{N}$,

(ii) $u_n \to u$ *in* $L^2(\Omega)$,

(iii) $G \in \mathcal{L}(L^2(\Omega), Y)$,

(iv) $G_n h \rightharpoonup Gh$ *in* Y *for all* $h \in L^2(\Omega)$.

Then G *is an element of* $\partial_B^{sw} S(u)$.

Proof. As in the proof before, for all $n \in \mathbb{N}$ the definition of $\partial_B^{sw} S(u_n)$ implies the existence of a sequence of Gâteaux points $u_{m,n} \in L^2(\Omega)$ with associated derivatives $G_{m,n} := S'(u_{m,n})$ such that $u_{m,n} \to u_n$ in $L^2(\Omega)$ as $m \to \infty$ and

$$G_{m,n} h \rightharpoonup G_n h \text{ in } Y \text{ for all } h \in L^2(\Omega) \text{ as } m \to \infty.$$

Now the compact embedding of Y in $H_0^1(\Omega)$ gives that $G_{m,n} h \to G_n h$ in $H_0^1(\Omega)$ as $m \to \infty$, and we can argue exactly as in the proof of Proposition 3.4 to show that there is a diagonal sequence of Gâteaux points $u_{m_n,n}$ such that $u_{m_n,n} \to u$ in $L^2(\Omega)$ and

$$G_{m_n,n} h \to Gh \text{ in } H_0^1(\Omega) \quad \text{for every } h \in L^2(\Omega). \tag{3.2}$$

On the other hand, by Lemma 3.2(iii), the operators $G_{m_n,n}$ are uniformly bounded in $\mathcal{L}(L^2(\Omega); Y)$. Therefore, for an arbitrary but fixed $h \in L^2(\Omega)$, the sequence $\|G_{m_n,n} h\|_Y$ is bounded in Y, so that a subsequence converges weakly to some $\eta \in Y$. Because of (3.2), $\eta = Gh$ and the uniqueness of the weak limit implies the weak convergence of the whole sequence in Y. As h was arbitrary, this implies the assertion. □

3.2 PRECISE CHARACTERIZATION OF THE BOULIGAND SUBDIFFERENTIALS

This section is devoted to an explicit characterization of the different subdifferentials in Definition 3.1 without the representation as (weak) limits of Jacobians of sequences of Gâteaux points. We start with the following lemma, which will be useful in the sequel:

Lemma 3.6. *Assume that*

(i) $j : \mathbb{R} \to \mathbb{R}$ *is a monotonically increasing and globally Lipschitz continuous,*

(ii) $(u_n) \subset L^2(\Omega)$ *is a sequence with* $u_n \rightharpoonup u \in L^2(\Omega)$,

(iii) $(\chi_n) \subset L^\infty(\Omega)$ *is a sequence satisfying* $\chi_n \geq 0$ *a.e. in* Ω *for all* $n \in \mathbb{N}$ *and* $\chi_n \rightharpoonup^* \chi$ *in* $L^\infty(\Omega)$ *for some* $\chi \in L^\infty(\Omega)$,

(iv) $w_n \in Y$ *is the unique solution to*

$$-\Delta w_n + \chi_n w_n + j(w_n) = u_n, \tag{3.3}$$

(v) $w \in Y$ *is the unique solution to*

$$-\Delta w + \chi w + j(w) = u. \tag{3.4}$$

Then it holds that $w_n \rightharpoonup w$ *in* Y, *and if we additionally assume that* $\chi_n \to \chi$ *pointwise a.e. and* $u_n \to u$ *strongly in* $L^2(\Omega)$, *then we even have* $w_n \to w$ *strongly in* Y.

Proof. First note that, due to the monotonicity and the global Lipschitz continuity of j, the equations (3.3) and (3.4), respectively, admit unique solutions in Y by the same arguments as in the proof of Proposition 2.1. Moreover, due to the weak and weak-$*$ convergence, the sequences (u_n) and (χ_n) are bounded in $L^2(\Omega)$ and $L^\infty(\Omega)$, respectively, so that (w_n) is bounded in Y. Hence there exists a weakly converging subsequence, w.l.o.g. denoted by the same symbol, such that $w_n \rightharpoonup \eta$ in Y; by passing to a further subsequence, we can assume due to the compact embedding $Y \hookrightarrow L^2(\Omega)$ that the convergence is strong in $L^2(\Omega)$. Together with the weak convergence of w_n and u_n, this allows passing to the limit in (3.3) to deduce that η satisfies

$$-\Delta\eta + \chi\eta + j(\eta) = u.$$

As the solution to this equation is unique, we obtain $\eta = w$. The uniqueness of the weak limit now gives convergence of the whole sequence, i.e., $w_n \rightharpoonup w$ in Y.

To prove the strong convergence under the additional assumptions, note that the difference $w_n - w$ satisfies

$$-\Delta(w_n - w) = (u_n - u) + (\chi\, w - \chi_n\, w_n) + (j(w) - j(w_n)). \tag{3.5}$$

For the first term on the right-hand side of (3.5), we have $u_n \to u$ in $L^2(\Omega)$ by assumption. The second term in (3.5) is estimated by

$$\|\chi\, w - \chi_n\, w_n\|_{L^2(\Omega)} \le \|\chi_n\|_{L^\infty(\Omega)}\|w - w_n\|_{L^2(\Omega)} + \|(\chi - \chi_n)w\|_{L^2(\Omega)}. \tag{3.6}$$

The first term in (3.6) converges to zero due to $w_n \rightharpoonup w$ in Y and the compact embedding, while the convergence of the second term follows from pointwise convergence of χ_n in combination with Lebesgue's dominated convergence theorem. The global Lipschitz continuity of j and the strong convergence of $w_n \to w$ in $L^2(\Omega)$ finally also give $j(w_n) \to j(w)$ in $L^2(\Omega)$. Therefore, the right-hand side in (3.5) converges to zero in $L^2(\Omega)$, and as $-\Delta$ induces the norm on Y, thanks to Poincaré's inequality, we finally obtain the desired strong convergence. □

By setting $j(x) = \max(0, x)$ and $\chi_n \equiv \chi \equiv 0$, we obtain as a direct consequence of the preceding lemma the following weak continuity of S.

Corollary 3.7. *The solution operator $S : L^2(\Omega) \to Y$ is weakly continuous, i.e.,*

$$u_n \rightharpoonup u \text{ in } L^2(\Omega) \quad \Longrightarrow \quad S(u_n) \rightharpoonup S(u) \text{ in } Y.$$

We will see in the following that all elements of the subdifferentials in Definition 3.1 have a similar structure. To be precise, they are solution operators of linear elliptic PDEs of a particular form.

Definition 3.2 (linear solution operator G_χ). Given a function $\chi \in L^\infty(\Omega)$ with $\chi \ge 0$, we define the operator $G_\chi \in \mathcal{L}(L^2(\Omega), Y)$ to be the solution operator of the linear equation

$$-\Delta\eta + \chi\,\eta = h. \tag{3.7}$$

We first address necessary conditions for an operator in $\mathcal{L}(L^2(\Omega), Y)$ to be an element of the Bouligand subdifferentials. Afterwards we will show that these conditions are also sufficient, which is more involved compared to their necessity.

Proposition 3.8 (necessary condition for $\partial_B^{ww} S(u)$). *Let $u \in L^2(\Omega)$ be arbitrary but fixed and set $y := S(u)$. Then for every $G \in \partial_B^{ww} S(u)$ there exists a unique $\chi \in L^\infty(\Omega)$ satisfying*

$$0 \le \chi \le 1 \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\} \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\} \tag{3.8}$$

such that $G = G_\chi$.

Proof. If $G \in \partial_B^{ww} S(u)$ is arbitrary but fixed, then there exists a sequence of Gâteaux points $u_n \in L^2(\Omega)$ such that $u_n \rightharpoonup u$ in $L^2(\Omega)$ and $S'(u_n)h \rightharpoonup Gh$ in Y for all $h \in L^2(\Omega)$. Now, let $h \in L^2(\Omega)$ be fixed but arbitrary and abbreviate $y_n := S(u_n)$, $\delta_{h,n} := S'(u_n)h$, and $\chi_n := \mathbb{1}_{\{y_n > 0\}}$. Then we know from Corollary 3.7 that $y_n \rightharpoonup y$ in Y and from Corollary 2.3 that $\delta_{h,n} = G_{\chi_n} h$. Moreover, from the Banach–Alaoglu Theorem it follows that, after transition to a subsequence (which may be done independently of h), it holds that $\chi_n \rightharpoonup^* \chi$ in $L^\infty(\Omega)$. Due to the weak-$*$ closedness of the set $\{\xi \in L^\infty(\Omega) : 0 \le \xi \le 1 \text{ a.e. in } \Omega\}$ and the pointwise almost everywhere convergence of (a subsequence of) y_n to y, we see that χ satisfies the conditions in (3.8). From Lemma 3.6, we may now deduce that $\delta_{h,n} \rightharpoonup G_\chi h$ in Y. We already know, however, that $\delta_{h,n} = S'(u_n)h \rightharpoonup Gh$ in Y. Consequently, since h was arbitrary, $G = G_\chi$, and the existence claim is proven. It remains to show that χ is unique. To this end, assume that there are two different functions $\chi, \tilde{\chi} \in L^\infty(\Omega)$ with $G = G_\chi = G_{\tilde{\chi}}$. If we then consider a function $\psi \in C^\infty(\mathbb{R}^d)$ with $\psi > 0$ in Ω and $\psi \equiv 0$ in $\mathbb{R}^d \setminus \Omega$ (whose existence is ensured by Lemma A.1) and define $h_\psi := -\Delta\psi + \chi\psi \in L^2(\Omega)$, then we obtain $\psi = G_\chi h_\psi = G_{\tilde{\chi}} h_\psi$, which gives rise to

$$-\Delta\psi + \chi\psi = h_\psi = -\Delta\psi + \tilde{\chi}\psi.$$

Subtraction now yields $(\chi - \tilde{\chi})\psi = 0$ a.e. in Ω and, since $\psi > 0$ in Ω, this yields $\chi \equiv \tilde{\chi}$. □

Proposition 3.9 (necessary condition for $\partial_B^{ws} S(u)$). *Let $u \in L^2(\Omega)$ be arbitrary but fixed with $y = S(u)$. Then for every $G \in \partial_B^{ws} S(u)$ there exists a unique function $\chi \in L^\infty(\Omega)$ satisfying*

$$\chi \in \{0,1\} \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\} \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\} \tag{3.9}$$

such that $G = G_\chi$.

Proof. Let $G \in \partial_B^{ws} S(u)$ be fixed but arbitrary. Since $\partial_B^{ws} S(u) \subseteq \partial_B^{ww} S(u)$, the preceding proposition yields that there is a unique function χ satisfying (3.8) such that $G = G_\chi$. It remains to prove that χ only takes values in $\{0,1\}$. To this end, first observe that the definition of $\partial_B^{ws} S(u)$ implies the existence of a sequence of Gâteaux points $(u_n) \subset L^2(\Omega)$ such that $u_n \rightharpoonup u$ in $L^2(\Omega)$ and $S'(u_n)h \to Gh$ in Y for every $h \in L^2(\Omega)$, where, according to Corollary 2.3, $S'(u_n) = G_{\chi_n}$ with $\chi_n := \mathbb{1}_{\{y_n > 0\}}$. As in the proof of Proposition 3.8, we choose the special direction $h_\psi := -\Delta\psi + \chi\psi \in L^2(\Omega)$, where $\psi \in C^\infty(\mathbb{R}^d)$ is again a function with $\psi > 0$ in Ω and $\psi \equiv 0$ in $\mathbb{R}^d \setminus \Omega$. Then $Gh_\psi = \psi$ and the strong convergence of $G_{\chi_n} h_\psi$ to Gh_ψ in Y allows passing to a subsequence to obtain $\Delta G_{\chi_n} h_\psi \to \Delta\psi$ and $G_{\chi_n} h_\psi \to \psi$ pointwise a.e. in Ω. From the latter, it follows that for almost all $x \in \Omega$ there exists an $N \in \mathbb{N}$ (depending on x) with $G_{\chi_n} h_\psi(x) > 0$ for all $n \ge N$ and consequently

$$\lim_{n\to\infty} \chi_n(x) = \lim_{n\to\infty} \frac{h_\psi(x) + \Delta G_{\chi_n} h_\psi(x)}{G_{\chi_n} h_\psi(x)} = \frac{h_\psi(x) + \Delta\psi(x)}{\psi(x)} = \chi(x) \text{ for a.a. } x \in \Omega.$$

But, as χ_n takes only the values 0 and 1 for all $n \in \mathbb{N}$, pointwise convergence almost everywhere is only possible if $\chi \in \{0,1\}$ a.e. in Ω. □

As an immediate consequence of the last two results, we obtain:

Corollary 3.10. *If S is Gâteaux-differentiable in a point $u \in L^2(\Omega)$, then it holds*

$$\partial_B^{ss} S(u) = \partial_B^{sw} S(u) = \partial_B^{ws} S(u) = \partial_B^{ww} S(u) = \{S'(u)\}.$$

Proof. The inclusion $\supseteq$ was already proved in Lemma 3.2. The reverse follows immediately from Propositions 3.8 and 3.9, and the fact that, in a Gâteaux point, it necessarily holds $\lambda^d(\{y = 0\}) = 0$ (see Corollary 2.3). □

Remark 3.11. Note that even in finite dimensions, the Bouligand and the Clarke subdifferential can contain in a Gâteaux point functionals other than the Gâteaux derivative; see, e.g., [4, Ex. 2.2.3]. Accordingly, the Bouligand subdifferentials from Definition 3.1 are better behaved than the Clarke subdifferential in this respect.

Remark 3.12. Similarly to Theorem 2.2, where the directional derivative of the max-function appears, Propositions 3.8 and 3.9 show that elements of $\partial_B^{ww} S(u)$ and $\partial_B^{ws} S(u)$ are characterized by PDEs which involve a measurable selection of the convex resp. Bouligand subdifferential of $\mathbb{R} \ni x \mapsto \max(0, x) \in \mathbb{R}$.

Now that we have found necessary conditions that elements of the subdifferentials $\partial_B^{ws} S(u)$ and $\partial_B^{ww} S(u)$ have to fulfill, we turn to sufficient conditions which guarantee that a certain linear operator is an element of these subdifferentials. Here we focus on the subdifferentials $\partial_B^{ss} S(u)$ and $\partial_B^{sw} S(u)$. It will turn out that a linear operator is an element of these subdifferentials if it is of the form G_χ with χ as in (3.8) and (3.9), respectively. Thanks to Lemma 3.2(i) and the necessary conditions in Propositions 3.8 and 3.9, this will finally give a sharp characterization of all Bouligand subdifferentials in Definition 3.1, see Theorem 3.16 below. We start with the following preliminary result.

Lemma 3.13. *Let $u \in L^2(\Omega)$ be arbitrary but fixed and write $y := S(u)$. Assume that $\varphi \in Y$ is a function satisfying*

$$\lambda^d(\{y = 0\} \cap \{\varphi = 0\}) = 0 \tag{3.10}$$

and define $\chi \in L^\infty(\Omega)$ via

$$\chi := \mathbb{1}_{\{y>0\}} + \mathbb{1}_{\{y=0\}} \mathbb{1}_{\{\varphi>0\}}. \tag{3.11}$$

Then G_χ is an element of the strong-strong Bouligand subdifferential $\partial_B^{ss} S(u)$.

Proof. We have to construct sequences of Gâteaux points converging strongly to u so that also their Gâteaux derivatives in an arbitrary direction $h \in L^2(\Omega)$ converge strongly in Y to $G_\chi h$. For this purpose, set $y_\varepsilon := y + \varepsilon\varphi$, $\varepsilon \in (0, 1)$, and $u_\varepsilon := -\Delta y_\varepsilon + \max(0, y_\varepsilon) \in L^2(\Omega)$. Then we obtain $S(u_\varepsilon) = y_\varepsilon$, $y_\varepsilon \to y$ in Y and $u_\varepsilon \to u$ in $L^2(\Omega)$ as $\varepsilon \to 0$. Choose now arbitrary but fixed representatives of y and φ and define $Z := \{y = 0\} \cap \{\varphi = 0\}$. Then, for all $\varepsilon \neq \varepsilon'$ and all $x \in \Omega$, we have

$$y(x) + \varepsilon\varphi(x) = 0 \quad \text{and} \quad y(x) + \varepsilon'\varphi(x) = 0 \iff x \in Z,$$

i.e., the sets in the collection $(\{y + \varepsilon\varphi = 0\} \setminus Z)_{\varepsilon\in(0,1)}$ are disjoint (and obviously Borel measurable). Furthermore, the underlying measure space $(\Omega, \mathcal{B}(\Omega), \lambda^d)$ is finite. Thus, we may apply Lemma B.1 to obtain a λ^1-zero set $N \subset (0,1)$ such that

$$\lambda^d(\{y + \varepsilon\varphi = 0\}) \leq \lambda^d(Z) + \lambda^d(\{y + \varepsilon\varphi = 0\} \setminus Z) = 0$$

for all $\varepsilon \in E := (0,1) \setminus N$. According to Corollary 2.3, this implies that S is Gâteaux-differentiable in u_ε for all $\varepsilon \in E$ with $S'(u_\varepsilon) = G_{\chi_\varepsilon}$ where $\chi_\varepsilon := \mathbb{1}_{\{y+\varepsilon\varphi>0\}}$. Consider now an arbitrary but fixed sequence $\varepsilon_n \in E$ with $\varepsilon_n \to 0$ as $n \to \infty$ and fix for the time being a direction $h \in L^2(\Omega)$. Then $\delta_{\varepsilon_n} := S'(u_{\varepsilon_n})h = G_{\chi_{\varepsilon_n}} h$ satisfies

$$-\Delta\delta_{\varepsilon_n} + \chi_{\varepsilon_n}\delta_{\varepsilon_n} = h,$$

and it holds that

$$\lim_{n\to\infty} \chi_{\varepsilon_n}(x) = \chi(x) \text{ f.a.a. } x \in \Omega \quad \text{and} \quad \chi_{\varepsilon_n} \rightharpoonup {*}\chi \text{ in } L^\infty(\Omega),$$

where χ is as defined in (3.11). Note that, according to assumption (3.10), the case $y(x) = \varphi(x) = 0$ is negligible here. Using Lemma 3.6, we now obtain that $\delta_{\varepsilon_n} = S'(u_{\varepsilon_n})h$ converges strongly in Y to $G_\chi h$. Since $h \in L^2(\Omega)$ was arbitrary, this proves the claim. □

In the following, we successively sharpen the assertion of Lemma 3.13 by means of Lemma 3.6 and the approximation results for characteristic functions proven in Appendix B.

Proposition 3.14 (sufficient condition for $\partial_B^{ss} S(u)$). *Let $u \in L^2(\Omega)$ be arbitrary but fixed and set $y := S(u)$. If $\chi \in L^\infty(\Omega)$ satisfies*

$$\chi \in \{0,1\} \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\} \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\}, \tag{3.12}$$

then $G_\chi \in \partial_B^{ss} S(u)$.

Proof. Since $\chi \in L^\infty(\Omega)$ takes only the values 0 and 1, there exists a Borel set $B \subseteq \Omega$ with $\chi = \mathbb{1}_B$. We now proceed in two steps:

(i) If $B \subseteq \Omega$ is open, then we know from Lemma B.2(i) that there exists a sequence $(\varphi_n) \subset Y$ of functions satisfying (3.10) such that

$$\mathbb{1}_{\{\varphi_n>0\}} \rightharpoonup^* \mathbb{1}_B \text{ in } L^\infty(\Omega) \quad \text{and} \quad \mathbb{1}_{\{\varphi_n>0\}} \to \mathbb{1}_B \text{ pointwise a.e. in } \Omega. \tag{3.13}$$

Let us abbreviate $\chi_n := \mathbb{1}_{\{y>0\}} + \mathbb{1}_{\{y=0\}}\mathbb{1}_{\{\varphi_n>0\}} \in L^\infty(\Omega)$. Then, thanks to (3.12) and (3.13), we arrive at

$$\chi_n \rightharpoonup^* \chi \text{ in } L^\infty(\Omega) \quad \text{and} \quad \chi_n \to \chi \text{ pointwise a.e. in } \Omega. \tag{3.14}$$

Moreover, due to Lemma 3.13, we know that $G_{\chi_n} \in \partial_B^{ss} S(u)$ for all $n \in \mathbb{N}$ and, from Lemma 3.6 and (3.14), we obtain $G_{\chi_n} h \to G_\chi h$ strongly in Y for all $h \in L^2(\Omega)$. Proposition 3.4 now gives $G_\chi \in \partial_B^{ss} S(u)$ as claimed.

(ii) Assume now that $\chi = \mathbb{1}_B$ for some Borel set $B \subseteq \Omega$. Then Lemma B.2(ii) implies the existence of a sequence of open sets $B_n \subseteq \Omega$ such that

$$\mathbb{1}_{B_n} \overset{*}{\rightharpoonup} \mathbb{1}_B \text{ in } L^\infty(\Omega) \quad \text{and} \quad \mathbb{1}_{B_n} \to \mathbb{1}_B \text{ pointwise a.e. in } \Omega. \tag{3.15}$$

Similarly to above, we abbreviate $\chi_n := \mathbb{1}_{\{y>0\}} + \mathbb{1}_{\{y=0\}}\mathbb{1}_{B_n} \in L^\infty(\Omega)$ so that (3.15) again yields the convergence in (3.14). Since from (i) we know that $G_{\chi_n} \in \partial_B^{ss} S(u)$ for all $n \in \mathbb{N}$, we can argue completely analogously to (i) to prove $G_\chi \in \partial_B^{ss} S(u)$ in the general case. □

Proposition 3.15 (sufficient condition for $\partial_B^{sw} S(u)$). *Let $u \in L^2(\Omega)$ be arbitrary but fixed and $y := S(u)$. If $\chi \in L^\infty(\Omega)$ satisfies*

$$0 \le \chi \le 1 \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\} \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\}, \tag{3.16}$$

then $G_\chi \in \partial_B^{sw} S(u)$.

Proof. We again proceed in two steps:

(i) If χ is a simple function of the form $\chi := \sum_{k=1}^K c_k \mathbb{1}_{B_k}$ with $c_k \in (0,1]$ for all $k, K \in \mathbb{N}$, $B_k \subseteq \Omega$ Borel and mutually disjoint, then we know from Lemma B.2(iv) that there exists a sequence of Borel sets $A_n \subseteq \Omega$ such that $\mathbb{1}_{A_n} \rightharpoonup^* \chi$ in $L^\infty(\Omega)$. In view of (3.16), this yields

$$\chi_n := \mathbb{1}_{\{y>0\}} + \mathbb{1}_{\{y=0\}}\mathbb{1}_{A_n} \rightharpoonup^* \chi \text{ in } L^\infty(\Omega)$$

so that, by Lemma 3.6, we obtain $G_{\chi_n} h \rightharpoonup G_\chi h$ in Y for all $h \in L^2(\Omega)$. Moreover, from Proposition 3.14, we already know that $G_{\chi_n} \in \partial_B^{ss} S(u) \subseteq \partial_B^{sw} S(u)$ for all $n \in \mathbb{N}$ and therefore, Proposition 3.5 gives $A_\chi \in \partial_B^{sw} S(u)$ as claimed.

(ii) For an arbitrary $\chi \in L^\infty(\Omega)$ satisfying (3.16), measurability implies the existence of a sequence of simple functions converging pointwise almost everywhere to χ. Due to (3.16), the pointwise projection of these simple functions on $[0,1]$ also converges pointwise a.e. to χ and thus weakly-$*$ in $L^\infty(\Omega)$, too. Then we can apply (i) and again Lemma 3.6 and Proposition 3.5 to obtain the claim. □

Thanks to Lemma 3.2(i), the necessary conditions in Propositions 3.8 and 3.9, respectively, in combination with the sufficient conditions in Propositions 3.14 and 3.15, respectively, immediately imply the following sharp characterization of the Bouligand subdifferentials of S.

Theorem 3.16 (precise characterization of the subdifferentials of S). *Let $u \in L^2(\Omega)$ be arbitrary but fixed and set $y := S(u)$.*

(i) It holds $\partial_B^{ws} S(u) = \partial_B^{ss} S(u)$. Moreover, $G \in \partial_B^{ss} S(u)$ if and only if there exists a function $\chi \in L^\infty(\Omega)$ satisfying

$$\chi \in \{0,1\} \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\} \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\}$$

so that $G = G_\chi$. Furthermore, for each $G \in \partial_B^{ss} S(u)$ the associated χ is unique.

(ii) *It holds* $\partial_B^{ww} S(u) = \partial_B^{sw} S(u)$. *Moreover,* $G \in \partial_B^{sw} S(u)$ *if and only if there exists a function* $\chi \in L^\infty(\Omega)$ *satisfying*

$$0 \le \chi \le 1 \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\} \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\}$$

so that $G = G_\chi$. *Furthermore, for each* $G \in \partial_B^{sw} S(u)$ *the associated* χ *is unique.*

Remark 3.17. Theorem 3.16 shows that it does not matter whether we use the weak or the strong topology for the approximating sequence $u_n \in L^2(\Omega)$ in the definition of the subdifferential, only the choice of the operator topology makes a difference. We further see that the elements of the strong resp. weak Bouligand subdifferential are precisely those operators G_χ generated by a function $\chi \in L^\infty(\Omega)$ that is obtained from a pointwise measurable selection of the Bouligand resp. convex subdifferential of the max-function, cf. Remark 3.12.

4 FIRST-ORDER OPTIMALITY CONDITIONS

In this section we turn our attention back to the optimal control problem (P). We are mainly interested in the derivation of first-order necessary optimality conditions in qualified form, i.e., involving dual variables. Due to the non-smoothness of the control-to-state mapping S caused by the max-function in (PDE), the standard procedure based on the adjoint of the Gâteaux derivative of S cannot be applied. Instead, regularization and relaxation methods are frequently used to derive optimality conditions, as already mentioned in the introduction. We will follow the same approach and derive an optimality system in this way in the next subsection. Since the arguments are rather standard, we keep the discussion concise. We again emphasize that the optimality conditions themselves are not remarkable at all. However, in Section 4.2, we will give a new interpretation of the optimality system arising through regularization by means of the Bouligand subdifferentials from Section 3, see Theorem 4.7, which is the main result of our paper.

4.1 REGULARIZATION AND PASSAGE TO THE LIMIT

For the rest of this section, let $\bar{u} \in L^2(\Omega)$ be an arbitrary local minimizer for (P). We follow a widely used approach (see, e.g., [17]) and define our regularized optimal control problem as follows:

$$\left.\begin{aligned} &\min_{u \in L^2(\Omega),\, y \in H_0^1(\Omega)} \quad J(y,u) + \frac{1}{2}\,\|u - \bar{u}\|^2_{L^2(\Omega)} \\ &\text{s.t.} \quad -\Delta y + \max\nolimits_\varepsilon(y) = u \ \text{ in } \Omega \end{aligned}\right\} \tag{P_ε}$$

with a regularized version of the max-function satisfying the following assumptions.

Assumption 4.1. The family of functions $\max_\varepsilon : \mathbb{R} \to \mathbb{R}$ fulfills the following conditions:

(i) $\max_\varepsilon \in C^1(\mathbb{R})$ for all $\varepsilon > 0$.

(ii) There is a constant $C > 0$ such that $|\max_\varepsilon(x) - \max(0,x)| \le C\,\varepsilon$ for all $x \in \mathbb{R}$.

(iii) For all $x \in \mathbb{R}$ and all $\varepsilon > 0$, there holds $0 \le \max_\varepsilon'(x) \le 1$.

(iv) For every $\delta > 0$, the sequence $(\max'_\varepsilon)_{\varepsilon>0}$ converges uniformly to 1 on $[\delta, \infty)$ and uniformly to 0 on $(-\infty, -\delta]$ as $\varepsilon \to 0^+$.

There are numerous possibilities to construct families of functions satisfying Assumption 4.1; we only refer to the regularized max-functions used in [17, 22]. As for the max-function, we will denote the Nemytskii-Operator associated with $\max_\varepsilon$ by the same symbol.

Lemma 4.2. *For every $u \in L^2(\Omega)$, there exists a unique solution $y_\varepsilon \in Y$ of the PDE in (P_ε). The associated solution operator $S_\varepsilon : L^2(\Omega) \to Y$ is weakly continuous and Fréchet-differentiable. Its derivative at $u \in L^2(\Omega)$ in direction $h \in L^2(\Omega)$ is given by the unique solution $\delta \in Y$ to*

$$-\Delta\delta + \max'_\varepsilon(y_\varepsilon)\delta = h, \tag{4.1}$$

where $y_\varepsilon = S_\varepsilon(u)$.

Proof. The arguments are standard. The monotonicity of $\max_\varepsilon$ by Assumption 4.1(iii) yields the existence of a unique solution, and bootstrapping implies that this is an element of Y. The weak continuity of S_ε follows from Lemma 3.6 in exactly the same way as Corollary 3.7. Due to Assumption 4.1(i) and (iii), the Nemytskii operator associated with $\max_\varepsilon$ is continuously Fréchet-differentiable from $H^1_0(\Omega)$ to $L^2(\Omega)$ and, owing to the non-negativity of $\max'_\varepsilon$, the linearized equation in (4.1) admits a unique solution $\delta \in Y$ for every $h \in L^2(\Omega)$. The implicit function theorem then gives the differentiability result. □

Lemma 4.3. *There exists a constant $c > 0$ such that, for all $u \in L^2(\Omega)$, there holds*

$$\|S(u) - S_\varepsilon(u)\|_Y \le c\,\varepsilon \quad \forall\, \varepsilon > 0. \tag{4.2}$$

Moreover, for every sequence $u_n \in L^2(\Omega)$ with $u_n \to u$ in $L^2(\Omega)$ and every sequence $\varepsilon_n \to 0^+$, there exists a subsequence $(n_k)_{k\in\mathbb{N}}$ and an operator $G \in \partial_B^{sw} S(u)$ such that

$$S'_{\varepsilon_{n_k}}(u_{n_k})h \rightharpoonup Gh \text{ in } Y \quad \forall\, h \in L^2(\Omega).$$

Proof. Given $u \in L^2(\Omega)$, let us set $y := S(u)$ and $y_\varepsilon := S_\varepsilon(u)$. Then it holds that

$$-\Delta(y - y_\varepsilon) + \max(0, y) - \max(0, y_\varepsilon) = \max_\varepsilon(y_\varepsilon) - \max(0, y_\varepsilon). \tag{4.3}$$

Testing this equation with $y - y_\varepsilon$ and employing the monotonicity of the max-function and Assumption 4.1(ii) gives $\|y - y_\varepsilon\|_{H^1(\Omega)} \le c\,\varepsilon$. Then, thanks to the Lipschitz continuity of the max-function and again Assumption 4.1(ii), a bootstrapping argument applied to (4.3) yields (4.2).

To obtain the second part of the lemma, let $u_n \in L^2(\Omega)$ and $\varepsilon_n \in (0, \infty)$ be sequences with $u_n \to u$ in $L^2(\Omega)$ and $\varepsilon_n \to 0^+$. Then (4.2) and Proposition 2.1 imply

$$\|S_{\varepsilon_n}(u_n) - S(u)\|_Y \le C\varepsilon_n + \|S(u_n) - S(u)\|_Y \to 0$$

as $n \to \infty$, i.e., $y_n := S_{\varepsilon_n}(u_n) \to y := S(u)$ in Y. Now, given an arbitary but fixed direction $h \in L^2(\Omega)$, we know that the derivative $\delta_n := S'_{\varepsilon_n}(u_n)h$ is characterized by

$$-\Delta\delta_n + \max'_{\varepsilon_n}(y_n)\delta_n = h.$$

Then, due to $y_n \to y$ pointwise a.e. in Ω (at least for a subsequence) and Assumption 4.1(iii) and (iv), there is a subsequence (not relabeled for simplicity) such that

$$\max'_{\varepsilon_n}(y_n) \rightharpoonup^* \chi \quad \text{in } L^\infty(\Omega)$$

with

$$0 \le \chi \le 1 \text{ a.e. in } \Omega, \quad \chi = 1 \text{ a.e. in } \{y > 0\}, \quad \text{and} \quad \chi = 0 \text{ a.e. in } \{y < 0\}.$$

Note that the transition to a subsequence above is independent of h. Using Lemma 3.6 and Theorem 3.16 then yields the second claim. □

Theorem 4.4 (optimality system after passing to the limit). *Let $\bar{u} \in L^2(\Omega)$ be locally optimal for (P) with associated state $\bar{y} \in Y$. Then there exists a multiplier $\chi \in L^\infty(\Omega)$ and an adjoint state $p \in L^2(\Omega)$ such that*

$$p = (G_\chi)^* \partial_y J(\bar{y}, \bar{u}) \tag{4.4a}$$

$$\chi(x) \in \partial_c \max(\bar{y}(x)) \quad \text{a.e. in } \Omega \tag{4.4b}$$

$$p + \partial_u J(\bar{y}, \bar{u}) = 0, \tag{4.4c}$$

where $\partial_c \max : \mathbb{R} \to 2^{\mathbb{R}}$ denotes the convex subdifferential of the max*-function. The solution operator G_χ is thus an element of $\partial_B^{sw} S(\bar{u})$.*

Proof. Based on the previous results, the proof follows standard arguments, which we briefly sketch for the convenience of the reader. We introduce the reduced objective functional associated with (P_ε) by

$$F_\varepsilon : L^2(\Omega) \to \mathbb{R}, \quad F_\varepsilon(u) := J(S_\varepsilon(u), u) + \frac{1}{2} \|u - \bar{u}\|^2_{L^2(\Omega)},$$

and consider the following auxiliary optimal control problem

$$\left.\begin{aligned} &\min_{u \in L^2(\Omega)} \quad F_\varepsilon(u) \\ &\text{s.t.} \quad \|u - \bar{u}\|_{L^2} \le r, \end{aligned}\right\} \tag{$P_{\varepsilon,r}$}$$

where $r > 0$ is the radius of local optimality of $\bar{u}$. Thanks to the weak continuity of S_ε by Lemma 4.2 and the weak lower semicontinuity of J by Assumption 1.1, the direct method of the calculus of variations immediately implies the existence of a global minimizer $\bar{u}_\varepsilon \in L^2(\Omega)$ of ($P_{\varepsilon,r}$). Note that due to the continuous Fréchet-differentiability of J, the global Lipschitz continuity of S, and (4.2), for all ε sufficiently small, there exists a constant $C' > 0$ independent of h and ε such that

$$|J(S(u), u) - J(S_\varepsilon(u), u)| \le C'\varepsilon \quad \forall\, u \in L^2(\Omega) \text{ with } \|u - \bar{u}\|_{L^2} \le r.$$

As a consequence, we obtain (with the same constant)

$$F_\varepsilon(\bar{u}) = J(S_\varepsilon(\bar{u}), \bar{u}) \le C'\varepsilon + J(S(\bar{u}), \bar{u})$$

and

$$\begin{aligned} F_\varepsilon(u) &= J(S_\varepsilon(u), u) + \|u - \bar{u}\|_{L^2}^2 \\ &\geq J(S(u), u) + \|u - \bar{u}\|_{L^2}^2 - C'\varepsilon \quad \forall\, u \in L^2(\Omega) \text{ with } \|u - \bar{u}\|_{L^2} \leq r, \end{aligned}$$

and therefore

$$F_\varepsilon(\bar{u}) \leq F_\varepsilon(u) \quad \forall\, u \in L^2(\Omega) \text{ with } \sqrt{2C'\varepsilon} \leq \|u - \bar{u}\|_{L^2} \leq r.$$

Thus, for every $\varepsilon > 0$ sufficiently small, any global solution $\bar{u}_\varepsilon$ of ($\mathrm{P}_{\varepsilon,r}$) must necessarily satisfy

$$\|\bar{u}_\varepsilon - \bar{u}\|_{L^2} \leq \sqrt{2C'\varepsilon}. \tag{4.5}$$

In particular, for ε small enough, $\bar{u}_\varepsilon$ is in the interior of the r-ball around $\bar{u}$ and, as a global solution of ($\mathrm{P}_{\varepsilon,r}$), also a local one of ($\mathrm{P}_\varepsilon$). It therefore satisfies the first-order necessary optimality conditions of the latter, which, thanks to the chain rule and Lemma 4.2, read

$$\big(\partial_y J(S_\varepsilon(\bar{u}_\varepsilon), \bar{u}_\varepsilon), S_\varepsilon'(\bar{u}_\varepsilon)h\big)_Y + \big(\partial_u J(S_\varepsilon(\bar{u}_\varepsilon), \bar{u}_\varepsilon), h\big)_{L^2} + (\bar{u}_\varepsilon - \bar{u}, h)_{L^2} = 0 \quad \forall\, h \in L^2(\Omega). \tag{4.6}$$

From Lemma 4.3 we obtain that there exists a sequence $\varepsilon_n \to 0^+$ and an operator $G \in \partial_B^{sw} S(\bar{u})$ such that

$$S_{\varepsilon_n}'(\bar{u}_{\varepsilon_n})h \rightharpoonup Gh \text{ in } Y \quad \forall\, h \in L^2(\Omega).$$

Further, we deduce from (4.5), the global Lipschitz continuity of S, and (4.2), that $S_{\varepsilon_n}(\bar{u}_{\varepsilon_n}) \to S(\bar{u})$ in Y. Combining all of the above and using our assumptions on J, we can pass to the limit $\varepsilon_n \to 0$ in (4.6) to obtain

$$\big(\partial_y J(S(\bar{u}), \bar{u}), Gh\big)_Y + \big(\partial_u J(S(\bar{u}), \bar{u}), h\big)_{L^2} = 0 \quad \forall\, h \in L^2(\Omega).$$

By setting $p := G^* \partial_y J(S(\bar{u}), \bar{u})$, this together with Theorem 3.16 and Remark 3.17 finally proves the claim. □

Corollary 4.5. *Assume that J is continuously Fréchet-differentiable from $H_0^1(\Omega) \times L^2(\Omega)$ to $\mathbb{R}$. If $\bar{u} \in L^2(\Omega)$ is locally optimal for (P) with associated state $\bar{y}$, then there exists a multiplier $\chi \in L^\infty(\Omega)$ and an adjoint state $p \in H_0^1(\Omega)$ such that*

$$-\Delta p + \chi\, p = \partial_y J(\bar{y}, \bar{u}) \tag{4.7a}$$

$$\chi(x) \in \partial_c \max(\bar{y}(x)) \quad \text{a.e. in } \Omega \tag{4.7b}$$

$$p + \partial_u J(\bar{y}, \bar{u}) = 0. \tag{4.7c}$$

If J is even Fréchet-differentiable from $L^2(\Omega) \times L^2(\Omega)$ to $\mathbb{R}$, then $p \in Y$.

Proof. According to Definition 3.2, G_χ is the solution operator of (3.7), which is formally self-adjoint. Thus we can argue as in [25, Sec. 4.6] to deduce (4.7a) and the H^1-regularity of p. The Y-regularity is again obtained by bootstrapping. □

4.2 INTERPRETATION OF THE OPTIMALITY CONDITIONS IN THE LIMIT

In classical non-smooth optimization, optimality conditions of the form $0 \in \partial_* f(x)$, where $\partial_* f$ denotes one of the various subdifferentials of f, frequently appear when a function $f : X \to \mathbb{R}$ is minimized over a normed linear space X; we only refer to [23, Secs. 7 and 9] and the references therein. With the help of the results of Section 3, in particular Theorem 3.16, we are now in the position to interpret the optimality system in (4.4) in this spirit. To this end, we first consider the reduced objective and establish the following result for its Bouligand subdifferential:

Proposition 4.6 (chain rule). *Let $u \in L^2(\Omega)$ be arbitrary but fixed, and let $F : L^2(\Omega) \to \mathbb{R}$ be the reduced objective for (P) defined by $F(u) := J(S(u), u)$. Moreover, set $y := S(u)$. Then it holds*

$$\begin{aligned}\left\{G^* \partial_y J(y,u) + \partial_u J(y,u) : G \in \partial_B^{sw} S(u)\right\} \\ \subseteq \partial_B F(u) := \{w \in L^2(\Omega) :\ &\text{there exists } (u_n) \subset L^2(\Omega) \text{ with } u_n \to u \text{ in } L^2(\Omega) \\ &\text{such that } F \text{ is Gâteaux in } u_n \text{ for all } n \in \mathbb{N} \\ &\text{and } F'(u_n) \rightharpoonup w \text{ in } L^2(\Omega) \text{ as } n \to \infty\}.\end{aligned}$$

Proof. Let $u \in L^2(\Omega)$ and $G \in \partial_B^{sw} S(u)$ be arbitrary but fixed. Then the definition of $\partial_B^{sw} S(u)$ guarantees the existence of a sequence $u_n \in L^2(\Omega)$ of Gâteaux points with $u_n \to u$ in $L^2(\Omega)$ and $S'(u_n)h \rightharpoonup Gh$ in Y for all $h \in L^2(\Omega)$. Since J is Fréchet- and thus Hadamard-differentiable and so is S by Theorem 2.2, we may employ the chain rule to deduce that F is Gâteaux-differentiable in the points $u_n \in L^2(\Omega)$ with derivative

$$F'(u_n) = S'(u_n)^* \partial_y J(y_n, u_n) + \partial_u J(y_n, u_n) \in L^2(\Omega)$$

for $y_n := S(u_n)$. As $y_n \to y$ in Y by Proposition 2.1 and $J : Y \times L^2(\Omega) \to L^2(\Omega)$ is continuously Fréchet-differentiable by Assumption 1.1, we obtain for every $h \in L^2(\Omega)$ that

$$\begin{aligned}\big(F'(u_n), h\big)_{L^2} &= \big(\partial_y J(y_n, u_n), S'(u_n)h\big)_Y + \big(\partial_u J(y_n, u_n), h\big)_{L^2} \\ &\to \big(\partial_y J(y, u), Gh\big)_Y + \big(\partial_u J(y, u), h\big)_{L^2}.\end{aligned}$$

Since $h \in L^2(\Omega)$ was arbitrary, this proves $G^* \partial_y J(y,u) + \partial_u J(y,u) \in \partial_B F(u)$. □

With the above result, we can now relate the optimality conditions obtained via regularization to the Bouligand subdifferential of the reduced objective, and in this way rate the strength of the optimality system in (4.4).

Theorem 4.7 (limit optimality system implies Bouligand-stationarity). *It holds:*

$$\begin{array}{c}\bar{u} \text{ is locally optimal for (P)} \\ \Downarrow \\ \text{there exists } \chi \in L^\infty(\Omega) \text{ and } p \in L^2(\Omega) \text{ so that (4.4) holds} \\ \Downarrow \\ \bar{u} \text{ is Bouligand-stationary for (P) in the sense that } 0 \in \partial_B F(u) \\ \Downarrow \\ \bar{u} \text{ is Clarke-stationary for (P) in the sense that } 0 \in \partial_C F(u)\end{array}$$

Here, $\partial_C F(u)$ denotes the Clarke subdifferential as defined in [4, Sec. 2.1].

Proof. The first two implications immediately follow from Theorem 4.4 and Proposition 4.6. For the third implication, observe that the weak closedness of $\partial_C F(u)$ (see [4, Prop. 2.1.5b]) and $F'(u) \in \partial_C F(u)$ in all Gâteaux points (cf. [4, Prop. 2.2.2]) result in $\partial_B F(u) \subseteq \partial_C F(u)$. □

Remark 4.8. The above theorem is remarkable for several reasons:

(i) Theorem 4.7 shows that $0 \in \partial_B F(u)$ is a necessary optimality condition for the optimal control problem (P). This is in general not true even in finite dimensions, as the minimization of the absolute value function shows.

(ii) The above shows that the necessary optimality condition in Theorem 4.4, which is obtained by regularization, is comparatively strong. It is stronger than Clarke-stationarity and even stronger than Bouligand-stationarity (which is so strong that it does not even make sense in the majority of problems).

Remark 4.9. It is easily seen that the limit analysis in Section 4.1 readily carries over to control constrained problems involving an additional constraint of the form $u \in U_{\mathrm{ad}}$ for a closed and convex $U_{\mathrm{ad}} \subset L^2(\Omega)$. The optimality system arising in this way is identical to (4.4) except for (4.4c), which is replaced by the variational inequality

$$(p + \partial_u J(\bar{y}, \bar{u}), u - \bar{u}) \geq 0 \quad \forall\, u \in U_{\mathrm{ad}}.$$

The interpretation of the optimality system arising in this way in the spirit of Theorem 4.7 is, however, all but straightforward, as it is not even clear how to define the Bouligand subdifferential of the reduced objective in the presence of control constraints. Intuitively, one would choose the approximating sequences in the definition of $\partial_B F$ from the feasible set U_{ad}, but then the arising subdifferential could well be empty. This gives rise to future research.

4.3 STRONG STATIONARITY

Although comparatively strong, the optimality conditions in Theorem 4.4 are not the most rigorous ones, as we will see in the sequel. To this end, we apply a method of proof which was developed in [15] for optimal control problems governed by non-smooth semilinear parabolic PDEs and inspired by the analysis in [16, 17]. We begin with an optimality condition without dual variables.

Proposition 4.10 (purely primal optimality conditions). *Let $\bar{u} \in L^2(\Omega)$ be locally optimal for (P) with associated state $\bar{y} = S(\bar{u}) \in Y$. Then there holds*

$$F'(\bar{u}; h) \geq 0 \quad \forall\, h \in L^2(\Omega) \quad \Longleftrightarrow \quad \partial_y J(\bar{y}, \bar{u}) S'(\bar{u}; h) + \partial_u J(\bar{y}, \bar{u}) h \geq 0 \quad \forall\, h \in L^2(\Omega). \tag{4.8}$$

Proof. As already argued above, $J : Y \times L^2(\Omega) \to \mathbb{R}$ and $S : L^2(\Omega) \to Y$ are Hadamard-differentiable so that the reduced objective $F : L^2(\Omega) \ni u \mapsto J(S(u), u) \in \mathbb{R}$ is Hadamard-differentiable by the chain rule for Hadamard-differentiable mappings with directional derivative $F'(u; h) = \partial_y J(S(u), u) S'(u; h) + \partial_u J(S(u), u) h$. Thus, by classical arguments, the local optimality of $\bar{u}$ implies $F'(\bar{u}; h) \geq 0$ for all $h \in L^2(\Omega)$. □

Lemma 4.11. *Let $p \in L^2(\Omega)$ fulfill (4.4a), i.e., $p = (G_\chi)^* \partial_y J(\bar y, \bar u)$ with some $\chi \in L^\infty(\Omega)$, $\chi \geq 0$. Then, for every $v \in Y$ there holds*

$$(-\Delta v + \chi v, p)_{L^2(\Omega)} = \langle \partial_y J(\bar y, \bar u), v\rangle_{Y',Y}. \tag{4.9}$$

Proof. Let $v \in Y$ be arbitary and define $g \in L^2(\Omega)$ by $g := -\Delta v + \chi\, v$ so that $v = G_\chi g$. Then $p = (G_\chi)^* \partial_y J(\bar y, \bar u)$ implies

$$(-\Delta v + \chi v, p)_{L^2(\Omega)} = (g, p)_{L^2(\Omega)} = \langle \partial_y J(\bar y, \bar u), G_\chi g\rangle_{Y',Y} = \langle \partial_y J(\bar y, \bar u), v\rangle_{Y',Y},$$

as claimed. □

Theorem 4.12 (strong stationarity). *Let $\bar u \in L^2(\Omega)$ be locally optimal for (P) with associated state $\bar y \in Y$. Then there exists a multiplier $\chi \in L^\infty(\Omega)$ and an adjoint state $p \in L^2(\Omega)$ such that*

$$p = (G_\chi)^* \partial_y J(\bar y, \bar u) \tag{4.10a}$$
$$\chi(x) \in \partial_c \max(\bar y(x)) \quad \text{a.e. in } \Omega \tag{4.10b}$$
$$p(x) \leq 0 \quad \text{a.e. in } \{\bar y = 0\} \tag{4.10c}$$
$$p + \partial_u J(\bar y, \bar u) = 0. \tag{4.10d}$$

Proof. From Theorem 4.4, we know that there exist $p \in L^2(\Omega)$ and $\chi \in L^\infty(\Omega)$ so that (4.4) is valid, which already gives (4.10a), (4.10b), and (4.10d). It remains to show (4.10c). To this end, let $v \in Y$ be arbitrary and define

$$h := -\Delta v + \mathbb{1}_{\{\bar y = 0\}} \max(0, v) + \mathbb{1}_{\{\bar y > 0\}} v = -\Delta v + \max'(\bar y; v) \in L^2(\Omega) \tag{4.11}$$

so that $v = S'(\bar u; h)$ by Theorem 2.2. By testing (4.11) with p and using (4.4c) and (4.8) from Proposition 4.10, we arrive at

$$\begin{aligned}(-\Delta v + \max'(\bar y; v), p)_{L^2(\Omega)} &= (h, p)_{L^2(\Omega)} \\ &= (-\partial_u J(\bar y, \bar u), h)_{L^2(\Omega)} \\ &\leq \langle \partial_y J(\bar y, \bar u), S'(\bar u; h)\rangle_{Y',Y} = \langle \partial_y J(\bar y, \bar u), v\rangle_{Y',Y}. \end{aligned} \tag{4.12}$$

On the other hand, we know from Lemma 4.11 that p and χ satisfy (4.9). Subtracting this equation from (4.12) and using the density of $Y \hookrightarrow L^2(\Omega)$ and the global Lipschitz continuity of $L^2(\Omega) \ni v \mapsto \max'(\bar y; v) \in L^2(\Omega)$ yields

$$\int_\Omega \big(\max'(\bar y; v) - \chi\, v\big) p \, dx \leq 0 \quad \forall\, v \in L^2(\Omega). \tag{4.13}$$

Note that, due to (4.4b), the bracket in (4.13) vanishes a.e. in $\{\bar y \neq 0\}$. Next, take an arbitrary $\varphi \in C_c^\infty(\Omega)$, $\varphi \geq 0$, and choose $v = \mathbb{1}_{\{\bar y = 0\}} \mathbb{1}_{\{\chi \leq 0.5\}} \varphi$ as test function in (4.13), which results in

$$\int_\Omega (1 - \chi) \mathbb{1}_{\{\bar y = 0\}} \mathbb{1}_{\{\chi \leq 0.5\}} p\, \varphi \, dx \leq 0 \quad \forall\, \varphi \in C_c^\infty(\Omega),\ \varphi \geq 0.$$

The fundamental lemma of the calculus of variations then implies that

$$p \leq 0 \quad \text{a.e. in } \{y = 0\} \cap \{\chi \leq 0.5\}.$$

On the set $\{y = 0\} \cap \{\chi \geq 0.5\}$, we can argue analogously by choosing non-positive test functions from $C_c^\infty(\Omega)$. This finally gives (4.10c). □

Proposition 4.13. *The strong stationarity conditions are equivalent to the purely primal optimality conditions, i.e., if $\bar{u} \in L^2(\Omega)$ together with its state $\bar{y}$ and a multiplier χ and an adjoint state p satisfies (4.10), then it also fulfills the variational inequality (4.8).*

Proof. Let $h \in L^2(\Omega)$ be arbitrary and define $\delta = S'(\bar{u}; h)$. Then the gradient equation in (4.10d) and Theorem 2.2 give

$$\begin{aligned}(-\partial_u J(\bar{y}, \bar{u}), h)_{L^2} = (p, h)_{L^2} &= (-\Delta\delta + \max'(\bar{y}; \delta), p)_{L^2} \\ &= (-\Delta\delta + \chi\,\delta, p)_{L^2} + (\max'(\bar{y}; \delta) - \chi\,\delta, p)_{L^2}. \end{aligned} \tag{4.14}$$

For the last term, (4.10b) and the sign condition in (4.10c) yield

$$(\max'(\bar{y}; \delta) - \chi\,\delta, p)_{L^2} = \int_{\{y=0\}\cap\{\delta\geq 0\}} (1-\chi)\delta\, p\, dx + \int_{\{y=0\}\cap\{\delta<0\}} (-\chi)\delta\, p\, dx \leq 0.$$

Together with Lemma 4.11, this implies that (4.14) results in

$$(-\partial_u J(\bar{y}, \bar{u}), h)_{L^2} \leq \langle \partial_y J(\bar{y}, \bar{u}), \delta\rangle_{Y',Y},$$

which is (4.8). □

Remark 4.14. As in case of the optimality system (4.4), the regularity of the adjoint state in Theorem 4.12 is again only limited by the mapping resp. differentiability properties of the objective functional. Thus, arguing as in Corollary 4.5, one shows that if J is differentiable from $H_0^1(\Omega) \times L^2(\Omega)$ or $Y \times L^2(\Omega)$ to $\mathbb{R}$, the adjoint state p satisfying (4.7a) is an element of $H_0^1(\Omega)$ or Y, respectively. (4.7a).

Remark 4.15. Although the optimality system (4.4) is comparatively strong by Theorem 4.7, it provides less information compared to the strong stationarity conditions in (4.10) since it lacks the sign condition (4.10c) for the adjoint state. The conditions (4.10) can be seen as the most rigorous qualified optimality conditions, as by Proposition 4.13 they are equivalent to the purely primal condition. We point out, however, that the method of proof of Theorem 4.12 can in general not be transferred to the case with additional control constraints (e.g., $u \in U_{\text{ad}}$ for a closed and convex set U_{ad}), since it requires the set $\{S'(\bar{u}; h) : h \in \text{cone}(U_{\text{ad}} - \bar{u})\}$ to be dense in $L^2(\Omega)$. In contrast to this, the adaptation of the limit analysis in Section 4.1 to the case with additional control constraints is straightforward as mentioned in Remark 4.9.

5 ALGORITHMS AND NUMERICAL EXPERIMENTS

One particular advantage of the optimality system in (4.4) is that it seems amenable to numerical solution, as we will demonstrate in the following. We point out, however, that we do not present a comprehensive convergence analysis for our algorithm to compute stationary points satisfying (4.4) but only a feasibility study. For the sake of presentation, we consider here an L^2 tracking objective of the form

$$J(y, u) := \frac{1}{2}\,\|y - y_d\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\,\|u\|^2_{L^2(\Omega)} \tag{5.1}$$

with a given desired state $y_d \in L^2(\Omega)$ and a Tikhonov parameter $\alpha > 0$.

5.1 DISCRETIZATION AND SEMI-SMOOTH NEWTON-METHOD

Let us start with a short description of our discretization scheme. For the discretization of the state and the control variable, we use standard continuous piecewise linear finite elements (FE), cf., e.g., [3]. Let us denote by $V_h \subset H_0^1(\Omega)$ the associated FE space spanned by the standard nodal basis functions $\varphi_1, \dots, \varphi_n$. The nodes of the underlying triangulation $\mathcal{T}_h$ are denoted by $x_1, \dots, x_n$. We then discretize the state equation in (P) by employing a mass lumping scheme for the non-smooth nonlinearity. Specifically, we consider the discrete state equation

$$\int_\Omega \nabla y_h \cdot \nabla v_h \, dx + \sum_{T \in \mathcal{T}_h} \frac{1}{3} |T| \sum_{x_i \in \overline{T}} \max(0, y_h(x_i)) \, v_h(x_i) = \int_\Omega u_h \, v_h \, dx \quad \forall \, v_h \in V_h, \tag{5.2}$$

where $y_h, u_h \in V_h$ denote the FE-approximations of y and u. With a slight abuse of notation, we from now denote the coefficient vectors $(y_h(x_i))_{i=1}^n$ and $(u_h(x_i))_{i=1}^n$ by $y, u \in \mathbb{R}^n$. The discrete state equation can then be written as the nonlinear algebraic equation

$$Ay + D \max(0, y) = Mu, \tag{5.3}$$

where $A := (\int_\Omega \nabla \varphi_i \cdot \nabla \varphi_j \, dx)_{ij=1}^n \in \mathbb{R}^{n \times n}$ and $M := (\int_\Omega \varphi_i \, \varphi_j \, dx)_{ij=1}^n \in \mathbb{R}^{n \times n}$ denote stiffness and mass matrix, $\max(0, .) : \mathbb{R}^n \to \mathbb{R}^n$ is the componentwise max-function, and

$$D := \operatorname{diag}\left(\frac{1}{3} |\omega_i|\right) \in \mathbb{R}^{n \times n}$$

with $\omega(x_i) = \operatorname{supp}(\varphi_i)$ is the lumped mass matrix. Due to the monotonicity of the max-operator, one easily shows that (5.2) and (5.3) admit a unique solution for every control vector u. The objective functional is discretized by means of a suitable interpolation operator I_h (e.g., the Clément interpolator or, if $y_d \in C(\overline{\Omega})$, the Lagrange interpolator). If, again by the abuse of notation, we denote the coefficient vector of $I_h y_d$ with respect to the nodal basis by y_d, we end up with the discretized objective

$$J_h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad J_h(y, u) := \frac{1}{2} (y - y_d)^\top M (y - y_d) + \frac{\alpha}{2} u^\top M u.$$

If we again regularize the max-function in (5.3), then a limit analysis analogous to Section 4.1 yields the following discrete counterpart to (4.4) with vectors $p, \chi \in \mathbb{R}^n$ as necessary optimality conditions for the discretized optimal control problem:

$$Ay + D \max(0, y) = -\frac{1}{\alpha} Mp, \tag{5.4a}$$

$$Ap + D \, \chi \circ p = M(y - y_d), \tag{5.4b}$$

$$\chi_i \in \partial_c \max(y_i), \quad i = 1, \dots, n. \tag{5.4c}$$

Here, $a \circ b := (a_i b_i)_{i=1}^n$ denotes the Hadamard product, and we have eliminated the control by means of the gradient equation $p + \alpha u = 0$.

Next, we reformulate (5.4) as a non-smooth system of equations. To this end, let us introduce for a given $\gamma > 0$ the proximal point mapping $\mathrm{prox}_\gamma : \mathbb{R} \to \mathbb{R}$ by

$$\mathrm{prox}_\gamma(x) := \mathrm{argmin}_{s\in\mathbb{R}} \left(\max(0,s) + \frac{1}{2\gamma}\,|s-x|^2 \right) = \begin{cases} x, & x < 0, \\ 0, & x \in [0,\gamma], \\ x-\gamma, & x > \gamma. \end{cases} \tag{5.5}$$

Since the proximal mapping of max coincides with the resolvent $(I + \gamma\partial_c \max)^{-1}$ of its convex subdifferential, it is straightforward to show that $g \in \partial_c \max(z)$ if and only if $z = \mathrm{prox}_\gamma(z + \gamma\, g)$. Thus, for every $\gamma > 0$, (5.4c) is equivalent to the non-smooth equation

$$y_i = \mathrm{prox}_\gamma(y_i + \gamma\,\chi_i), \quad i = 1, \dots, n. \tag{5.4c$'$}$$

Since prox_γ is Lipschitz continuous and piecewise continuously differentiable by (5.5), it is semi-smooth as a function from $\mathbb{R} \to \mathbb{R}$. As the same holds for the max-function, it seems reasonable to apply the semi-smooth Newton method to numerically solve the system consisting of (5.4a), (5.4b), and (5.4c$'$). However, the application of a Newton-like scheme to solve (5.4) is a delicate issue. This can already be seen by observing that χ_i is not unique in points where y_i and p_i vanish at the same time. Moreover, the Newton-matrix may well be singular. For a clearer notation, let us introduce the index sets $\mathcal{I}_+ := \{i : y_i > 0\}$ and $\mathcal{I}_\gamma := \{i : y_i + \gamma\,\chi_i \notin [0,\gamma]\}$, and denote by $\mathbb{1}_{\mathcal{I}_+}$ and $\mathbb{1}_{\mathcal{I}_\gamma}$ the characteristic vectors of these index sets. Then the Newton matrix associated with (5.4) is given by

$$\begin{pmatrix} A + D\ \mathrm{diag}(\mathbb{1}_{\mathcal{I}_+}) & \frac{1}{\alpha} M & 0 \\ -M & A + D\ \mathrm{diag}(\chi) & D\ \mathrm{diag}(p) \\ I - \mathrm{diag}(\mathbb{1}_{\mathcal{I}_\gamma}) & 0 & -\gamma\ \mathrm{diag}(\mathbb{1}_{\mathcal{I}_\gamma}) \end{pmatrix}.$$

It becomes singular if there is an index $i \in \{1, \dots, n\}$ such that $p_i = 0$ and $y_i + \gamma\,\chi_i \in [0,\gamma]$. Our ad hoc solution to resolve this issue is to remove the components of χ corresponding to these indices from the Newton equation and leave them unchanged in the Newton update. In the numerical tests, this naive procedure worked surprisingly well (at least for small values of γ), as we will demonstrate in the next subsection. Alternatively, one could of course also add a small negative multiple of the identity to the 3-3-block of the Newton matrix. These considerations show that there are plenty of open questions concerning an efficient and stable numerical solution of (5.4), which is however beyond the scope of this work and subject to future research.

Remark 5.1. In the context of optimal control of the obstacle problem, a problem of the type (P) arises after applying a quadratic penalization to the obstacle problem, see, e.g., [22]. Various authors use a regularization scheme as in Section 4.1 for its numerical solution; we only refer to [12]. Even though these schemes work well, in particular in combination with a suitable path following strategy for the regularization parameter, our numerical tests show that it might be possible to solve the limit conditions in (4.4) without a further regularization.

Remark 5.2. It is not clear how to numerically solve the strong stationarity conditions in (4.10), since the system may well become over-determined by the sign condition in (4.10c). This

corresponds to optimal control problems governed by the obstacle problem, where it is also not known how to use the strong stationarity conditions for numerical computations, as for instance observed in [17].

5.2 NUMERICAL RESULTS

We present two different two-dimensional examples. In both examples, the domain Ω was chosen to be the unit square, which is discretized by means of Friedrich–Keller triangulations. For the construction of exact solutions to (4.4), we introduce an additional inhomogeneity in the state equation, i.e., we replace the PDE in (P) by

$$y \in H_0^1(\Omega), \quad -\Delta y + \max(0, y) = u + f \ \text{ in } \Omega$$

with a given function $f \in L^2(\Omega)$. It is easy to see that this modification does not influence the analysis in the preceding sections. In all test cases, the semi-smooth Newton iteration was terminated if the relative difference between two iterates became less than 10^{-8}. The Newton equations in each iteration are solved by MATLAB's direct solver. In both examples, the state vanishes in parts of the domain so that the non-smoothness of the max-function becomes apparent.

5.2.1 *FIRST TEST CASE*

In the first example, state and adjoint state are set to

$$y(x_1, x_2) = \sin(\pi\, x_1) \sin(2\pi\, x_2), \quad p \equiv 0$$

and the data f and y_d are constructed such that (4.4) is fulfilled, i.e., the optimal control is $u \equiv 0$. We note that there is a subset where y and p vanish at the same time, but it is only of zero measure. This will be different in the second example. Table 1 presents the numerical results for different values of the mesh size h, the Tikhonov parameter α in the objective in (5.1), and the parameter γ in the proximal point mapping.

We observe that in all cases except one, only few Newton iterations are needed to solve the problem. Only in case of $\gamma = 10^{-6}$ does the semi-smooth Newton method fail to converge within the maximum number of 50 iterations. This is also observed in case of the second example as well as in other instances. Therefore smaller numbers of γ seem to be favorable, but it is to be noted that the condition number of the Newton matrix increases as γ is reduced. Moreover, Table 1 shows that the approximation of χ gets worse if γ decreases, while the approximation of y and p is not affected. Concerning the dependency on mesh size, we approximately observe a quadratic order of convergence for y and p in the first three mesh refinements. However, the last mesh refinement does not improve the result any more. A possible explanation is that the algorithm does not properly resolve the set $\{y = 0\}$, which calls for a more sophisticated treatment of the critical indices where $p_i = 0$ and $y_i + \gamma\, \chi_i \in [0, \gamma]$. Surprisingly, a reduction of the Tikhonov parameter leads to an improvement of the approximation results. This is a particularity of this example, and we observe a different behavior in the second test case. Note that the Tikhonov parameter enters the data through the construction of the example, which might be an explanation for this behavior.

Table 1: Numerical results in the first example.

h	α	γ	$\frac{\|y_h - y\|_{L^2}}{\|y\|_{L^2}}$	$\frac{\|p_h - p\|_{L^2}}{\|p\|_{L^2}}$	$\frac{\|\chi_h - \chi\|_{L^2}}{\|\chi\|_{L^2}}$	# Newton
$3.125 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$6.104 \cdot 10^{-4}$	$1.101 \cdot 10^{-5}$	$1.044 \cdot 10^{-1}$	3
$1.563 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$1.526 \cdot 10^{-4}$	$2.764 \cdot 10^{-6}$	$7.734 \cdot 10^{-2}$	3
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$3.575 \cdot 10^{-5}$	$8.361 \cdot 10^{-7}$	$2.674 \cdot 10^{-9}$	2
$3.906 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$3.651 \cdot 10^{-5}$	$3.297 \cdot 10^{-7}$	$2.656 \cdot 10^{-9}$	2
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-8}$	$1.659 \cdot 10^{-4}$	$3.292 \cdot 10^{-6}$	$6.110 \cdot 10^{-2}$	3
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-8}$	$1.241 \cdot 10^{-4}$	$2.443 \cdot 10^{-6}$	$6.683 \cdot 10^{-2}$	3
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$3.575 \cdot 10^{-5}$	$8.361 \cdot 10^{-7}$	$2.674 \cdot 10^{-9}$	2
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-8}$	$6.699 \cdot 10^{-6}$	$9.034 \cdot 10^{-8}$	$2.619 \cdot 10^{-9}$	2
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-6}$	–	–	–	no conv.
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$3.575 \cdot 10^{-5}$	$8.361 \cdot 10^{-7}$	$2.674 \cdot 10^{-9}$	2
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-10}$	$3.575 \cdot 10^{-5}$	$8.361 \cdot 10^{-7}$	$1.590 \cdot 10^{-7}$	2
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-12}$	$3.575 \cdot 10^{-5}$	$8.361 \cdot 10^{-7}$	$1.962 \cdot 10^{-5}$	2

5.2.2 *SECOND TEST CASE*

Here we choose

$$y(x_1, x_2) = p(x_1, x_2) = \begin{cases} \left((x_1 - \frac{1}{2})^4 + \frac{1}{2}(x_1 - \frac{1}{2})^3\right) \sin(\pi\, x_2), & x_1 < \frac{1}{2}, \\ 0, & x_1 \geq \frac{1}{2}. \end{cases}$$

Note that y and p are twice continuously differentiable and vanish both on the right half of the unit square. Therefore, the non-smoothness of the max-function comes into play on a set of positive measure in this example. Moreover, as y and p vanish at the same time, χ is not unique in this set. This example can thus be seen as a worst case scenario. Nevertheless, our algorithm is able to produce reasonable results, as Table 2 demonstrates. (Note that it does not make sense to list the relative L^2-error for χ_h, since χ is not unique as mentioned above.)

The algorithm shows a similar behavior as in the first example. Again, it does not converge if γ is chosen too large. Moreover, we observe approximately quadratic convergence with respect to mesh refinement, this time even in the last refinement step. In contrast to the first example, the numerical approximation now becomes worse if the Tikhonov parameter is reduced. This is a classical observation, which is also made in case of smooth optimal control problems.

In summary, one can conclude that our ad hoc realization of the semi-smooth Newton method seems to be able to solve (4.4) resp. its discrete counterpart (5.4) even in critical cases. A comprehensive convergence analysis is however still lacking, and the choice of the parameter γ appears to be a delicate issue. Moreover, as already mentioned in Remark 5.2, it is completely unclear how to incorporate the sign condition in (4.10c) into the algorithmic framework. This is subject to future research.

Table 2: Numerical results in the second example.

h	α	γ	$\frac{\|y_h - y\|_{L^2}}{\|y\|_{L^2}}$	$\frac{\|p_h - p\|_{L^2}}{\|p\|_{L^2}}$	# Newton
$3.125 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$9.402 \cdot 10^{-1}$	$1.708 \cdot 10^{-2}$	2
$1.563 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$2.389 \cdot 10^{-1}$	$4.697 \cdot 10^{-3}$	2
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$6.015 \cdot 10^{-2}$	$1.233 \cdot 10^{-3}$	1
$3.906 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$1.506 \cdot 10^{-2}$	$3.157 \cdot 10^{-4}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-8}$	$1.487 \cdot 10^{-3}$	$1.675 \cdot 10^{-3}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-8}$	$1.279 \cdot 10^{-2}$	$1.550 \cdot 10^{-3}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$6.015 \cdot 10^{-2}$	$1.233 \cdot 10^{-3}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-5}$	$1 \cdot 10^{-8}$	$1.725 \cdot 10^{-1}$	$8.777 \cdot 10^{-4}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-6}$	–	–	no conv.
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-8}$	$6.015 \cdot 10^{-2}$	$1.233 \cdot 10^{-3}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-10}$	$6.015 \cdot 10^{-2}$	$1.233 \cdot 10^{-3}$	1
$7.813 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-12}$	$6.015 \cdot 10^{-2}$	$1.233 \cdot 10^{-3}$	1

APPENDIX A SMOOTH "CHARACTERISTIC" FUNCTIONS OF OPEN SETS

Lemma A.1. *For every open set $D \subseteq \mathbb{R}^d$ there exists a function $\psi \in C^\infty(\mathbb{R}^d)$ such that $\psi > 0$ everywhere in D and $\psi \equiv 0$ in $\mathbb{R}^d \setminus D$.*

Proof. Since the collection of all open balls with rational radii and rational centers forms a base of the Euclidean topology on $\mathbb{R}^d$, given an arbitary but fixed open set D, we find (non-empty) open balls $B_n \subseteq D$, $n \in \mathbb{N}$, such that

$$D = \bigcup_{n=1}^{\infty} B_n.$$

For every ball B_n, there is a smooth rotational symmetric bump function $\psi_n \in C^\infty(\mathbb{R}^d)$ with $\psi_n > 0$ in B_n and $\psi_n \equiv 0$ in $\mathbb{R}^d \setminus B_n$. Define

$$\psi := \sum_{n=1}^{\infty} \frac{\psi_n}{2^n \|\psi_n\|_{H^n(\mathbb{R}^d)}}.$$

Then it holds that $\psi > 0$ in D, $\psi \equiv 0$ in $\mathbb{R}^d \setminus D$, and $\psi \in H^n(\mathbb{R}^d)$ for all $n \in \mathbb{N}$. Sobolev's embedding theorem then yield the claim. □

APPENDIX B APPROXIMATION OF FUNCTIONS IN $L^\infty(\Omega;\{0,1\})$ AND $L^\infty(\Omega;[0,1])$

Lemma B.1. *If (X, Σ, μ) is a finite measure space and if $\mathcal{A} := \{A_i\}_{i \in I}$, $I \subset \mathbb{R}$, is a collection of measurable disjoint sets $A_i \in \Sigma$, then there exists a countable set $N \subset I$ such that $\mu(A_i) = 0$ for all $i \in I \setminus N$.*

Proof. Define $\mathcal{A}_k := \{A_i \in \mathcal{A} : \mu(A_i) \geq 1/k\}$, $k \in \mathbb{N}$. Then every $A_i \in \mathcal{A}$ with $\mu(A_i) > 0$ is contained in at least one $\mathcal{A}_k$. Suppose $A_{i_1}, \ldots, A_{i_m}$ are contained in $\mathcal{A}_k$. Then their disjointness implies

$$\frac{m}{k} \leq \sum_{l=1}^{m} \mu(A_{i_l}) \leq \mu(X),$$

and thus $m \leq k\mu(X) < \infty$. Consequently, there can only be finitely many A_i in each $\mathcal{A}_k$ so that the set $\{i \in I : \mu(A_i) > 0\}$ is countable. □

Lemma B.2.

(i) If $A \subseteq \Omega$ is open, then there exists a sequence $\varphi_n \in Y$ with $\lambda^d(\{\varphi_n = 0\}) = 0$ such that

$$\mathbb{1}_{\{\varphi_n > 0\}} \to \mathbb{1}_A \text{ pointwise a.e. in } \Omega.$$

(ii) If $A \subseteq \Omega$ is Borel, then there exists a sequence of open sets $A_n \subseteq \Omega$ such that

$$\mathbb{1}_{A_n} \to \mathbb{1}_A \text{ pointwise a.e. in } \Omega.$$

(iii) If $A \subseteq \Omega$ is Borel and if $c \in (0,1]$ is arbitrary but fixed, then there exists a sequence of Borel sets $A_n \subseteq A$ such that

$$\mathbb{1}_{A_n} \overset{*}{\rightharpoonup} c\mathbb{1}_A \text{ in } L^\infty(\Omega).$$

(iv) If $\chi : \Omega \to \mathbb{R}$ is a simple function of the form

$$\chi := \sum_{k=1}^{K} c_k \mathbb{1}_{B_k}$$

with $K \in \mathbb{N}$, $c_k \in (0,1]$, and $B_k \subseteq \Omega$ Borel and mutually disjoint for all k, then there exists a sequence of Borel sets $A_n \subseteq \Omega$ such that

$$\mathbb{1}_{A_n} \overset{*}{\rightharpoonup} \chi \text{ in } L^\infty(\Omega).$$

Proof. Ad (i): Let $A \subseteq \Omega$ be an arbitrary but fixed open set. By Lemma A.1, there exist functions $\psi, \phi \in C^\infty(\mathbb{R}^d)$ with

$$\psi > 0 \text{ in } \Omega, \quad \psi \equiv 0 \text{ in } \mathbb{R}^d \setminus \Omega, \quad \text{and} \quad \phi > 0 \text{ in } A, \quad \phi \equiv 0 \text{ in } \mathbb{R}^d \setminus A.$$

So, if we define $\varphi_\varepsilon := \phi - \varepsilon\psi$, then $\varphi_\varepsilon \in Y \cap C(\overline{\Omega})$ for all $\varepsilon \in (0,1)$ and $\varphi_\varepsilon(x) \to \phi(x)$ for all $x \in \Omega$ as $\varepsilon \to 0$ (here and in the following, we always use the continuous representatives). Moreover, the sign conditions on ψ and ϕ imply that

$$\mathbb{1}_{\{\varphi_\varepsilon > 0\}} \to \mathbb{1}_A \text{ pointwise a.e. in } \Omega. \tag{B.1}$$

Consider now some $\varepsilon_1, \varepsilon_2 \in (0,1)$ with $\varepsilon_1 \neq \varepsilon_2$. Then it holds that

$$\begin{aligned}&\{x \in \Omega : \varphi_{\varepsilon_1}(x) = 0\} \cap \{x \in \Omega : \varphi_{\varepsilon_2}(x) = 0\} \\ &\quad = \{x \in \Omega : \phi(x) - \varepsilon_1\psi(x) = 0 \text{ and } \phi(x) - \varepsilon_2\psi(x) = 0\} = \{x \in \Omega : \varepsilon_1 = \varepsilon_2\} = \emptyset,\end{aligned}$$

showing that the collection $(\{x \in \Omega : \varphi_\varepsilon(x) = 0\})_{\varepsilon\in(0,1)}$ is disjoint. Analogously to the proof of Lemma 3.13, we now obtain by means of Lemma B.1 that there exists a sequence ε_n with $\varepsilon_n \to 0$ as $n \to \infty$ such that

$$\lambda^d(\{x \in \Omega : \varphi_{\varepsilon_n}(x) = 0\}) = 0 \quad \forall n \in \mathbb{N}.$$

Together with (B.1), this establishes the assertion of part (i).

Ad (ii): The outer regularity of the Lebesgue measure implies the existence of open sets $\tilde{A}_n \subseteq \Omega$ such that

$$A \subseteq \tilde{A}_n \quad \text{and} \quad \lambda^d(\tilde{A}_n \setminus A) < \frac{1}{n} \quad \forall n \in \mathbb{N}.$$

Let us define

$$A_n := \bigcap_{m=1}^{n} \tilde{A}_m \quad \forall\, n \in \mathbb{N}.$$

Then A_n is open for all $n \in \mathbb{N}$, and it holds that

$$A_{n+1} \subseteq A_n, \quad A \subseteq A_n, \quad \text{and} \quad \lambda^d(A_n \setminus A) < \frac{1}{n} \quad \forall n \in \mathbb{N}.$$

The above implies that

$$\mathbb{1}_{A_n}(x) \to \mathbb{1}_A(x) \quad \forall\, x \in A \cup \bigcup_{n\in\mathbb{N}} \Omega \setminus A_n = \Omega \setminus \left(\bigcap_{n\in\mathbb{N}} A_n \setminus A\right).$$

Since the exceptional set appearing above has measure zero, this proves (ii).

Ad (iii): We apply a homogenization argument. Given $n \in \mathbb{N}$, let us define

$$B_n := \bigcup_{k\in\mathbb{Z}^d} \frac{1}{n}k + \left[0, \frac{1}{n}\right]^{d-1} \times \left[0, \frac{1}{n}c\right] \subseteq \mathbb{R}^d.$$

The sequence $(\mathbb{1}_{B_n})_{n\in\mathbb{N}}$ is bounded in $L^\infty(\mathbb{R}^d)$, and we may extract a subsequence (not relabeled) such that $\mathbb{1}_{B_n} \rightharpoonup^* \chi$ in $L^\infty(\mathbb{R}^d)$ for some $\chi \in L^\infty(\mathbb{R}^d)$. Consider now an arbitrary but fixed $\varphi \in C_c^\infty(\mathbb{R}^d)$. Then it holds that

$$\begin{aligned}\int_{\mathbb{R}^d} \mathbb{1}_{B_n}\varphi \mathrm{d}\lambda &= \int_{\mathbb{R}^d} c\varphi \mathrm{d}\lambda + c\sum_{k\in\mathbb{Z}^d} \int_{\frac{1}{n}k+\left[0,\frac{1}{n}\right]^d} \varphi\left(\frac{1}{n}k\right) - \varphi(x)\mathrm{d}\lambda(x) \\ &\qquad + \sum_{k\in\mathbb{Z}^d} \int_{\frac{1}{n}k+\left[0,\frac{1}{n}\right]^{d-1}\times\left[0,\frac{1}{n}c\right]} \varphi(x) - \varphi\left(\frac{1}{n}k\right) \mathrm{d}\lambda(x) \\ &\to \int_{\mathbb{R}^d} c\varphi \mathrm{d}\lambda.\end{aligned}$$

Using standard density arguments and the uniqueness of the weak limit, we deduce from the above that $\mathbb{1}_{B_n} \rightharpoonup^* \chi \equiv c$ in $L^\infty(\mathbb{R}^d)$ for the whole original sequence $(\mathbb{1}_{B_n})_{n\in\mathbb{N}}$. But now for all $v \in L^1(\Omega)$, it holds that

$$\int_\Omega \mathbb{1}_{B_n\cap A} v \mathrm{d}\lambda = \int_{\mathbb{R}^d} \mathbb{1}_{B_n}\mathbb{1}_A v \mathrm{d}\lambda \to \int_\Omega c\mathbb{1}_A v \mathrm{d}\lambda,$$

i.e., $\mathbb{1}_{B_n \cap A} \rightharpoonup^* c\mathbb{1}_A$ in $L^\infty(\Omega)$. This gives the claim with $A_n := B_n \cap A$.

Ad (iv): According to part (iii), we can find for every $k \in \{1, \dots, K\}$ a sequence of Borel sets $A_{k,n} \subseteq B_k$ such that

$$\mathbb{1}_{A_{k,n}} \rightharpoonup^* c_k \mathbb{1}_{B_k} \text{ in } L^\infty(\Omega) \text{ as } n \to \infty.$$

Defining $A_n := \bigcup_{k=1}^K A_{n,k}$, the claim follows immediately by superposition. □

ACKNOWLEDGMENTS

C. Clason was supported by the DFG under grant CL 487/2-1, and C. Christof and C. Meyer were supported by DFG under grant ME 3281/6-1 resp. ME 3281/7-1, both within the priority programme SPP 1962 "Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization".

REFERENCES

[1] Barbu, *Optimal Control of Variational Inequalities*, Research notes in mathematics 100, Pitman, 1984.

[2] Bergounioux, Optimal control problems governed by abstract elliptic variational inequalities with state constraints, *SIAM Journal on Control and Optimization* 36 (1998), 273–289, DOI: 10.1137/s0363012996302615.

[3] Brenner & Scott, *The Mathematical Theory of Finite Element Methods*, Springer, 1994, DOI: 10.1007/978-0-387-75934-0.

[4] Clarke, *Optimization and Nonsmooth Analysis*, SIAM, 1990, DOI: 10.1137/1.9781611971309.

[5] Clason & Valkonen, Stability of saddle points via explicit coderivatives of pointwise subdifferentials, *Set-Valued and Variational Analysis* 25 (2017), 69–112, DOI: 10.1007/s11228-016-0366-7.

[6] Craven & Glover, An approach to vector subdifferentials, *Optimization* 38 (1996), 237–251, DOI: 10.1080/02331939608844251.

[7] Henrion, Mordukhovich & Nam, Second-order analysis of polyhedral systems in finite and infinite dimensions with applications to robust stability of variational inequalities, *SIAM Journal on Optimization* 20 (2010), 2199–2227, DOI: 10.1137/090766413.

[8] Herzog, Meyer & Wachsmuth, B- and strong stationarity for optimal control of static plasticity with hardening, *SIAM Journal on Optimization* 23 (2013), 321–352, DOI: 10.1137/110821147.

[9] Hintermüller, An active-set equality constrained Newton solver with feasibility restoration for inverse coefficient problems in elliptic variational inequalities, *Inverse Problems* 24 (2008), 034017, 23, DOI: 10.1088/0266-5611/24/3/034017.

[10] Ito & Kunisch, Optimal control of elliptic variational inequalities, *Applied Mathematics and Optimization* 41 (2000), 343–364, DOI: 10.1007/s002459911017.

[11] Klatte & Kummer, *Nonsmooth Equations in Optimization*, Kluwer, 2002, DOI: 10.1007/b130810.

[12] Kunisch & Wachsmuth, Path-following for optimal control of stationary variational inequalities, *Computational Optimization and Applications. An International Journal* 51 (2012), 1345–1373, DOI: 10.1007/s10589-011-9400-8.

[13] Mehlitz & Wachsmuth, The limiting normal cone to pointwise defined sets in Lebesgue Spaces, *Set-Valued and Variational Analysis* (2016), 1–19, DOI: 10.1007/s11228-016-0393-4.

[14] Mehlitz & Wachsmuth, Weak and strong stationarity in generalized bilevel programming and bilevel optimal control, *Optimization* 65 (2016), 907–935, DOI: 10.1080/02331934.2015.1122007.

[15] Meyer & Susu, Optimal control of nonsmooth, semilinear parabolic equations, tech. rep. 524, Ergebnisberichte des Instituts für Angewandte Mathematik, TU Dortmund, 2016, URL: http://www.mathematik.tu-dortmund.de/papers/MeyerSusu2015.pdf.

[16] Mignot, Contrôle dans les inéquations variationelles elliptiques, *Journal of Functional Analysis* 22 (1976), 130–185, DOI: 10.1016/0022-1236(76)90017-3.

[17] Mignot & Puel, Optimal control in some variational inequalities, *SIAM Journal on Control and Optimization* 22 (1984), 466–476, DOI: 10.1137/0322028.

[18] Outrata & Römisch, On optimality conditions for some nonsmooth optimization problems over L^p spaces, *Journal of Optimization Theory and Applications* 126 (2005), 411–438, DOI: 10.1007/s10957-005-4724-0.

[19] Outrata, Jarušek & Stará, On optimality conditions in control of elliptic variational inequalities, *Set-Valued and Variational Analysis* 19 (2011), 23–42, DOI: 10.1007/s11228-010-0158-4.

[20] Outrata, Kočvara & Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Kluwer, 1998, DOI: 10.1007/978-1-4757-2825-5.

[21] Rockafellar & Wets, *Variational Analysis*, Springer, 2004, DOI: 10.1007/978-3-642-02431-3.

[22] Schiela & Wachsmuth, Convergence analysis of smoothing methods for optimal control of stationary variational inequalities with control constraints, *ESAIM: M2AN* 47 (2013), 771–787, DOI: 10.1051/m2an/2012049.

[23] Schirotzek, *Nonsmooth Analysis*, Springer, 2007, DOI: 10.1007/978-3-540-71333-3.

[24] Tiba, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Springer, 1990, DOI: 10.1007/bfb0085564.

[25] Tröltzsch, *Optimal Control of Partial Differential Equations*, American Mathematical Society, Providence, RI, 2010, DOI: 10.1090/gsm/112.

[26] Wachsmuth, Towards M-stationarity for optimal control of the obstacle problem with control constraints, *SIAM Journal on Control and Optimization* 54 (2016), 964–986, DOI: 10.1137/140980582.