ESAIM: COCV 30 (2024) 16
https://doi.org/10.1051/cocv/2023091

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

ON THE IDENTIFICATION AND OPTIMIZATION OF NONSMOOTH SUPERPOSITION OPERATORS IN SEMILINEAR ELLIPTIC PDES

Constantin Christof* and Julia Kowalczyk

Abstract. We study an infinite-dimensional optimization problem that aims to identify the Nemytskii operator in the nonlinear part of a prototypical semilinear elliptic partial differential equation (PDE) which minimizes the distance between the PDE-solution and a given desired state. In contrast to previous works, we consider this identification problem in a low-regularity regime in which the function inducing the Nemytskii operator is a-priori only known to be an element of $H^1_{loc}(\mathbb{R})$. This makes the studied problem class a suitable point of departure for the rigorous analysis of training problems for learning-informed PDEs in which an unknown superposition operator is approximated by means of a neural network with nonsmooth activation functions (ReLU, leaky-ReLU, etc.). We establish that, despite the low regularity of the controls, it is possible to derive a classical stationarity system for local minimizers and to solve the considered problem by means of a gradient projection method. The convergence of the resulting algorithm is proven in the function space setting. It is also shown that the established first-order necessary optimality conditions imply that locally optimal superposition operators share various characteristic properties with commonly used activation functions: They are always sigmoidal, continuously differentiable away from the origin, and typically possess a distinct kink at zero. The paper concludes with numerical experiments which confirm the theoretical findings.

Mathematics Subject Classification. 35J61, 49J50, 49J52, 49K20, 49M05, 68T07.

Received June 8, 2023. Accepted December 13, 2023.

1. Introduction

This paper is concerned with the analysis and numerical solution of optimization problems of the following type:

$$\left.\begin{aligned}
\text{Minimize}\quad & \frac{1}{2}\|y-y_D\|^2_{L^2(\Omega)} + \nu_1\|g'\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|g'-u_D\|^2_{L^2(\mathbb{R})}\\
\text{w.r.t.}\quad & y \in H^1(\Omega),\quad g \in H^1_{loc}(\mathbb{R}),\\
\text{s.t.}\quad & -\Delta y + g(y) = f \text{ in } \Omega,\quad y = 0 \text{ on } \partial\Omega,\\
\text{and}\quad & g' \geq 0 \text{ a.e. in } \mathbb{R},\\
\text{and}\quad & g(0) = 0.
\end{aligned}\right\}\tag{1.1}$$

Keywords and phrases: Optimal control, superposition, Nemytskii operator, semilinear elliptic partial differential equation, data-driven models, learning-informed PDEs, inverse problems, Bouligand stationarity, Gâteaux differentiability, gradient projection method, artificial neural network, optimality condition, nonsmooth optimization, machine learning.

Technische Universität München, Chair of Optimal Control, School of Computation, Information and Technology, Department of Mathematics, Boltzmannstraße 3, 85748 Garching, Germany.

* Corresponding author: christof@cit.tum.de

Here, $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is a bounded domain which is convex or possesses a $C^{1,1}$-boundary; $y_D \in L^2(\Omega)$, $f \in L^2(\Omega) \setminus \{0\}$, and $u_D \in L^2(\mathbb{R})$ are given functions; g' denotes the weak derivative of g; $\nu_1 \geq 0$ and $\nu_2 > 0$ are regularization parameters; Δ denotes the Laplace operator; and the objective function is to be understood as an extended real-valued function on $H^1(\Omega) \times H^1_{loc}(\mathbb{R})$ with values in $[0, \infty]$. For the precise assumptions on the quantities in (1.1) and the definitions of the occurring spaces etc., we refer the reader to Sections 2.1 and 2.2. Note that, by introducing the variable $u := g'$, the problem (1.1) can also be formulated in the following, more convenient form:

$$\left.\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \nu_1 \|u\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(\mathbb{R})} \\ \text{w.r.t.} \quad & y \in H^1(\Omega), \quad u \in L^2(\mathbb{R}), \\ \text{s.t.} \quad & -\Delta y + g_u(y) = f \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega, \\ \text{and} \quad & g_u(t) = \int_0^t u(s)\,\mathrm{d}s \quad \forall t \in \mathbb{R}, \\ \text{and} \quad & u \geq 0 \text{ a.e. in } \mathbb{R}. \end{aligned}\right\} \tag{1.2}$$

The main results of this work – Theorems 4.5, 5.7, 6.17, 7.2, 7.5, 8.13 – establish improved regularity properties for local solutions of (1.1) and (1.2), first-order necessary optimality conditions of Bouligand- and primal-dual type, and convergence results for a gradient projection algorithm that makes it possible to solve (1.1) and (1.2) in a function space setting. For a more detailed overview of our findings, see Section 1.2 below.

1.1. Motivation, background, and relation to prior work

Before we present in detail what we prove for the local minimizers of (1.1) and (1.2) in this paper, let us briefly discuss why it is worthwhile to consider these problems: Our main motivation for the study of (1.1) is that this minimization problem can be interpreted as the continuous version of a training problem for a learning-informed partial differential equation (PDE) which aims to approximate an unknown nonlinearity in a semilinear model by means of an artificial neural network. Such problems have received considerable attention in recent years – see, *e.g.*, [1–9] – and typically take the following form if an all-at-once approach with a single right-hand side f is used for the network training and the same PDE as in (1.1) is studied:

$$\left.\begin{aligned} \text{Minimize} \quad & J_1(y) + J_2(\alpha) \\ \text{w.r.t.} \quad & y \in H^1(\Omega), \quad \alpha \in D, \\ \text{s.t.} \quad & -\Delta y + \psi(\alpha, y) = f \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega, \\ \text{and} \quad & \alpha \in D_{ad}. \end{aligned}\right\} \tag{1.3}$$

Here, J_1 is a fidelity term that measures the distance between the PDE-solution y and some given data either in the whole of Ω or in certain parts of it; $\psi \colon D \times \mathbb{R} \to \mathbb{R}$ is a neural network with parameter vector α and parameter space D (defined, *e.g.*, as in [10], Sect. 2); D_{ad} is an appropriately chosen subset of D; and $J_2 \colon D \to \mathbb{R}$ is a regularization term acting on α. See, for example, [3], Section 4, [7], Section 2, and [8], Section 3 for particular instances of problems of the type (1.3). What our model problem (1.1) now represents is – for a prototypical PDE and a special choice of the fidelity and regularization term that will be discussed in more detail in Section 2.3 – the continuous optimization problem that is obtained when the Nemytskii operator g is not assumed to be given in the form of a neural network but allowed to vary freely among all elements of $H^1_{loc}(\mathbb{R})$ that are nondecreasing and zero at the origin. (Note that the former of these conditions is required to ensure the unique solvability of the state-equation; see Section 3 below.) In other words, (1.1) is the continuous optimization problem that is present *before* a specific ansatz for the unknown nonlinearity g in the governing PDE – in the form of a neural network of a certain architecture or a comparable approximation instrument with a finite number of parameters – is decided on. This makes (1.1) an appropriate point of departure for the rigorous analysis and comparison of different approaches for the approximation of the unknown Nemytskii operator g in the PDE $-\Delta y + g(y) = f$: Its solutions are precisely the limit objects that approximate solutions obtained from a discretization of (1.1) by means of a neural network ansatz or a comparable approach should converge to when the number of degrees of freedom in the used discretization scheme goes to infinity. In future work, we plan to use the results established in this paper, for instance, for the derivation of a-priori error estimates for finite-element-based discretizations of (1.1), for the comparison of finite-element- and neural-network-based approaches for the numerical solution of (1.1), and for the development of hybrid methods which combine machine-learning components and more classical instruments to achieve an accurate, fast, and reliable identification of unknown nonlinear terms in partial differential equations.

We would like to emphasize at this point that the idea to study infinite-dimensional optimization problems in which an optimization variable enters a governing partial differential equation or variational inequality via superposition is not new. In the context of the heat equation, for example, this type of problem has already been considered for the calculation of unknown heat transfer laws in [11–17] and in the context of damage evolution in quasistatic elasticity in the recent [18]. For seminal works on the identification of Nemytskii operators in partial and ordinary differential equations, see also [19, 20], which analyze penalty approximation schemes for elliptic PDEs with convex analysis techniques, and [21], which is concerned with a classical boundary value problem in one dimension. Moreover, the task of identifying nonlinear model components has, of course, also been studied quite extensively in the field of inverse problems, *e.g.*, in relation to stability issues and iterative recovery algorithms; see [7, 22, 23] and the references therein.

What distinguishes our analysis from these earlier contributions is that, in our model problem (1.1) and its reformulation (1.2), the function $g = g_u$ inducing the Nemytskii operator in the governing PDE is a-priori only known to be in $H^1_{loc}(\mathbb{R})$ and, thus, merely Hölder continuous with exponent $1/2$. This nonsmoothness causes the control-to-state mapping $L^2_+(\mathbb{R}) \ni u \mapsto y \in H^1_0(\Omega)$ to lack the property of directional differentiability and significantly complicates the derivation of first-order optimality conditions that may serve as a starting point for the development of numerical solution algorithms – in particular as in (1.1) one also has to deal with the constraint $u = g' \geq 0$ a.e. in $\mathbb{R}$ which has to be imposed to ensure the well-posedness of the state-equation. At least to the best of our knowledge, the derivation of, for instance, primal-dual stationarity systems for local minimizers in such a low-regularity regime has not been accomplished so far in the literature. Compare, *e.g.*, with [11, 12, 15, 16, 18, 23] in this context which all assume (at least) C^1-regularity of the involved superposition operators. In view of the applications in the field of learning-informed PDEs outlined at the beginning of this section and the fact that various of the most popular activation functions in machine learning are nonsmooth (ReLU, leaky-ReLU, etc.), the requirement of continuous differentiability is very unsatisfactory. The only contributions we are aware of that contain results which go in a similar direction as our analysis are [21], [13], and [3]: In the first of these papers, [21], the author is able to establish a variant of a necessary optimality condition in terms of Clarke's generalized directional derivative for the identification of a nonsmooth Nemytskii operator in an ordinary differential equation that only requires Lipschitz continuity; see [21], Theorem 3. This necessary condition, however, is not amenable to any form of adjoint calculus and thus only of limited use; see the comments in [21], Section 3.3. It is moreover easy to check that the approach of [21] does not carry over to the PDE-setting. Second, in [13], Theorem 2, a Fréchet differentiability result is stated for the function that maps a merely Lipschitz continuous Nemytskii operator in a nonlinear heat transfer law to the solution of the corresponding heat equation. This theorem requires rather restrictive assumptions as, for instance, that the right-hand side of the heat equation is zero, that the considered Nemytskii operator only possesses a finite number of jumps in its derivative, and that the boundary temperature of the solution of the heat equation is strictly decreasing in time. Lastly, in [3], a necessary optimality condition for training problems of the type (1.3) involving neural networks with nonsmooth activation functions has recently been derived by means of a regularization approach and techniques of [24]; see [3], Theorem 6.5. It should be noted that, even for this special case, where the Nemytskii operator in the governing PDE is known to have a certain structure (namely that of a neural network), the derivation of stationarity systems is not straightforward as passages to the limit in compositions of the type $g_n'(y_n)$ are hard to handle unless convergences in very strong spaces are available for the sequences $\{g_n'\}$ and $\{y_n\}$; *cf.* [3], Remark 6.6. In the present work, we are able to circumvent these problems by exploiting certain regularity properties of the solutions of (1.1) and Stampacchia's lemma.

1.2. Summary of main results

The first main result of this work – Theorem 5.7 – is concerned with structural properties of local solutions of (1.2). It establishes that, if $u_D|_{[-r,r]} \in G[-r,r]$ (resp., $u_D|_{[-r,r]} \in BV[-r,r]$) holds for a representative of the desired control $u_D \in L^2(\mathbb{R})$ and a number $r > 0$ that is larger than a computable lower bound r_P, then every local minimizer $\bar{u} \in L^2(\mathbb{R})$ of (1.2) possesses a representative which satisfies

$$\bar{u} = \max\left(0, u_D - \frac{\nu_1}{\nu_2}\right) \text{ a.e. in } \mathbb{R} \setminus (-r,r) \quad \text{and} \quad \bar{u}|_{[-r,r]} \in G[-r,r] \text{ (resp., } \bar{u}|_{[-r,r]} \in BV[-r,r]). \tag{1.4}$$

Here, $G[-r,r]$ and $BV[-r,r]$ denote the spaces of regulated functions and functions of bounded variation on $[-r,r]$, respectively, and a vertical bar denotes a restriction. The properties in (1.4) are of fundamental importance for the analysis of the problem (1.2) for two reasons: First of all, the fact that all local minimizers of (1.2) are determined completely by u_D, ν_1, and ν_2 on $\mathbb{R} \setminus (-r,r)$ allows us to reformulate (1.2) as a minimization problem in $L^2(-r,r)$; see Theorem 4.5. This makes it possible to apply standard discretization techniques to solve (1.2) numerically; see Section 8. Second, the $G[-r,r]$-regularity in (1.4) implies that every local minimizer $\bar{u}$ of (1.2) admits a representative which has well-defined left and right limits on $[-r,r]$. As the function $g_{\bar{u}} \in H^1_{loc}(\mathbb{R})$ arises from $\bar{u}$ by integration, this entails that the map $g_{\bar{u}} \colon \mathbb{R} \to \mathbb{R}$ is directionally differentiable on $[-r,r]$ and that the control-to-state operator $L^2_+(\mathbb{R}) \ni u \mapsto y \in H^1_0(\Omega)$ of (1.2) possesses a directional derivative at all local solutions $\bar{u}$ of (1.2); see Corollary 6.11. Although the control-to-state mapping $u \mapsto y$ of (1.2) does not possess a directional derivative everywhere, it thus does so at all points that are relevant for the derivation of first-order necessary optimality conditions. By exploiting this effect, we are able to state a standard Bouligand stationarity condition for (1.2). This condition can be found in our second main result – Theorem 7.2.

Due to its implicit nature, the concept of Bouligand stationarity is typically not very useful in practice. To arrive at a more tangible stationarity system, in Sections 7 to 9, we focus on special instances of the problem (1.2), namely those for which the level sets $\{f = c\}$ of the given right-hand side $f \in L^2(\Omega)$ (defined up to sets of measure zero) have measure zero for all $c \in \mathbb{R}$. Our third main result – Theorem 6.17 – establishes that, under this assumption on f, the control-to-state mapping $u \mapsto y$ of (1.2) is not only directionally differentiable but even continuously Fréchet differentiable on the space of regulated functions. In combination with (1.4), this differentiability property allows us to derive the following explicit primal-dual first-order necessary optimality condition for local minimizers $\bar{u}$ of (1.2); see Theorem 7.5:

$$\begin{aligned}
&-\Delta \bar{y} + g_{\bar{u}}(\bar{y}) = f \text{ in } \Omega, \quad \bar{y} = 0 \text{ on } \partial\Omega, && -\Delta \bar{p}_1 + \bar{u}(\bar{y})\bar{p}_1 = \bar{y} - y_D \text{ in } \Omega, \quad \bar{p}_1 = 0 \text{ on } \partial\Omega,\\
&\bar{p}_2(s) = \begin{cases} \displaystyle\int_{\{\bar{y} \geq s\}} \bar{p}_1 \,\mathrm{d}x & \text{f.a.a. } s \geq 0, \\ \displaystyle -\int_{\{\bar{y} \leq s\}} \bar{p}_1 \,\mathrm{d}x & \text{f.a.a. } s < 0, \end{cases} && \bar{u} = \max\left(0, u_D + \frac{\bar{p}_2 - \nu_1}{\nu_2}\right) \quad \text{a.e. in } \mathbb{R}.
\end{aligned} \tag{1.5}$$

Here, $\bar{y}$ denotes the state associated with $\bar{u}$, the functions $\bar{p}_1 \in H^1_0(\Omega) \cap H^2(\Omega)$ and $\bar{p}_2 \in L^2(\mathbb{R})$ are adjoint variables, and the composition $\bar{u}(\bar{y})$ is defined in a suitable sense; see Definition 6.15. As we will see in Section 7, for sufficiently smooth u_D, the system (1.5) implies (among other things) that local solutions $\bar{u}$ of (1.2) give rise to maps $g_{\bar{u}} \colon \mathbb{R} \to \mathbb{R}$ that share various characteristic properties with commonly used activation functions from machine learning: The resulting $g_{\bar{u}}$ are always sigmoidal, continuously differentiable away from the origin, and typically possess a distinct kink at zero. We expect that this observation is of particular interest in the context of the learning-informed PDEs in (1.3). In Section 8, we moreover show that the continuous Fréchet differentiability of the control-to-state map established in Theorem 6.17 makes it possible to set up a gradient projection method for the numerical solution of (1.2) and to analyze the convergence of the resulting algorithm in function space. For the main result on this topic, see Theorem 8.13.

1.3. Structure of the remainder of the paper

We conclude this introduction with an overview of the content and the structure of the remainder of the paper: Section 2 is concerned with preliminaries. Here, we clarify the notation, state our standing assumptions, and briefly comment on the choice of the regularization term in the problems (1.1) and (1.2); see Section 2.3. In Section 3, we prove basic results on the control-to-state mapping $u \mapsto y$ of (1.2) that are needed for our analysis; see Theorem 3.4. Section 4 addresses the consequences that the properties in Section 3 have for the problem (1.2). This section in particular shows that (1.2) indeed admits an equivalent reformulation as an optimization problem in $L^2(-r, r)$ for all large enough r; *cf.* the comments on this topic in Section 1.2. In Section 5, we establish the regularity result in (1.4) for the local minimizers of (1.2); see Theorem 5.7. The proof of this key theorem relies on a relaxation of the nonnegativity constraint in (1.2) and a careful limit analysis. Section 6 demonstrates that the control-to-state operator of (1.2) is directionally differentiable and – under the additional assumption on f mentioned in Section 1.2 – even continuously Fréchet differentiable on the space of regulated functions; see Theorem 6.17. In Section 7, we use the differentiability properties from Section 6 and the regularity results from Section 5 to establish a Bouligand stationarity condition and the primal-dual system (1.5) for the problem (1.2). Section 8 is concerned with the solution of (1.2) by means of a gradient projection method. Here, we prove that such an algorithm terminates after finitely many iterations with an ϵ-stationary point. In Section 9, we conclude the paper with numerical experiments which confirm that the algorithm developed in Section 8 is also usable in practice.

2. Problem setting and notation

In this section, we introduce basic notation and discuss the assumptions on the quantities in the problems (1.1) and (1.2) that are supposed to hold throughout our analysis.

2.1. Basic notation

In what follows, we denote norms, scalar products, and dual pairings by the symbols $\|\cdot\|$, $(\cdot,\cdot)$, and $\langle\cdot,\cdot\rangle$, respectively, equipped with a subscript that clarifies the considered space. Given a normed space $(U, \|\cdot\|_U)$, we define $B_\rho^U(u)$ to be the closed ball of radius $\rho > 0$ in U centered at $u \in U$. The closure of a set in U is denoted by $\mathrm{cl}_U(\cdot)$. If a convex closed nonempty set $D \subset U$ and a point $u \in D$ are given, then we define $\mathcal{R}(u; D) := \{s(v-u) \mid s > 0, v \in D\}$ to be the radial cone to D at u and $\mathcal{T}(u; D) := \mathrm{cl}_U(\mathcal{R}(u; D))$ to be the tangent cone to D at u; *cf.* [25], Section 2.2.4. For the space of linear and continuous functions $H: U \to Y$ between two normed spaces $(U, \|\cdot\|_U)$ and $(Y, \|\cdot\|_Y)$, we use the symbol $\mathcal{L}(U, Y)$. In the special case $Y = \mathbb{R}$, we define $U^* := \mathcal{L}(U, \mathbb{R})$ to be the dual space of U. If U is continuously embedded into Y (in the sense of [26], Def. 4.19), then we write $U \hookrightarrow Y$. To denote the strong (respectively, weak) convergence of a sequence in a normed space, we use the symbol $\to$ (respectively, $\rightharpoonup$). Given an arbitrary nonempty set $D \subset U$ and a function $H: U \to Y$, we define $\mathrm{Range}(H)$ to be the range of H, $H|_D$ to be the restriction of H to D, and $\mathbb{1}_D: U \to \mathbb{R}$ to be the function that is equal to one on D and equal to zero on $U \setminus D$. When talking about the level sets of a map $H: U \to Y$, we use the shorthand notation $\{H = c\} := \{u \in U \mid H(u) = c\}$, $c \in Y$. Analogous abbreviations are also used for sublevel sets, superlevel sets, and general preimages. Fréchet, Gâteaux, and directional derivatives of a function $H: U \to Y$ are denoted by a prime; see Definition 6.1.

Given a Lebesgue measurable set $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we denote the vector space of (equivalence classes of) Lebesgue measurable real-valued functions on Ω by $L^0(\Omega)$. For the real Lebesgue spaces defined w.r.t. the d-dimensional Lebesgue measure λ^d on Ω, we use the standard notation $(L^q(\Omega), \|\cdot\|_{L^q(\Omega)})$, $1 \le q \le \infty$; see [27]. When working with elements of the spaces $L^q(\Omega)$, quantities like level sets etc. are understood as being defined up to sets of measure zero. For simplicity, we denote representatives of elements of $L^q(\Omega)$ by the same symbols as their associated equivalence classes. We further define $L^q_+(\Omega) := \{u \in L^q(\Omega) \mid u \ge 0 \text{ a.e. in } \Omega\}$. If $\Omega \subset \mathbb{R}^d$,

$d \in \mathbb{N}$, is an open set, then the symbols $W^{k,q}(\Omega)$, $H^k(\Omega)$, and $H_0^1(\Omega)$, $k \in \mathbb{N}$, $q \in [1, \infty]$, refer to the classical Sobolev spaces on Ω, endowed with their usual norms; see [28], Chapter 5. In this situation, we also define $L^1_{loc}(\Omega)$ and $H^1_{loc}(\Omega)$ to be the sets of functions that are locally in L^1 and H^1 on Ω, respectively, in the sense of [29]. For the spaces of continuous, k-times continuously differentiable, and smooth functions on Ω and $\overline{\Omega}$, we use the standard symbols $C(\Omega)$, $C^k(\Omega)$, $C^\infty(\Omega)$, $C(\overline{\Omega})$, $C^k(\overline{\Omega})$, and $C^\infty(\overline{\Omega})$, respectively; see [28, 29]. Here, $\overline{\Omega}$ is a shorthand notation for the closure of Ω w.r.t. the Euclidean norm $|\cdot|$, *i.e.*, $\overline{\Omega} := \Omega \cup \partial\Omega$, where $\partial(\cdot)$ denotes the boundary of a set. When talking about smooth functions with compact support on Ω, we write $C_c^\infty(\Omega)$. Recall that, for open bounded $\Omega \subset \mathbb{R}^d$, $C(\overline{\Omega})$ is a Banach space when equipped with the supremum-norm $\|\cdot\|_{C(\overline{\Omega})} := \|\cdot\|_\infty$. As usual, we denote the dual of $(H_0^1(\Omega), \|\cdot\|_{H^1(\Omega)})$ (interpreted as a space of distributions) by $H^{-1}(\Omega)$; see [28], Section 5.2. For the distributional Laplacian, we use the symbol $\Delta\colon H_0^1(\Omega) \to H^{-1}(\Omega)$, and for a weak gradient, the symbol ∇. Where appropriate, we consider Δ also as a function $\Delta\colon H^2(\Omega) \to L^2(\Omega)$. In the one-dimensional setting, weak and distributional derivatives are denoted by a prime.

Given $a, b \in \mathbb{R}$ with $a < b$, we define $BV[a,b]$ and $G[a,b]$ to be the spaces of real-valued functions of bounded variation and real-valued regulated functions on $[a,b]$, respectively. We emphasize that we understand $BV[a,b]$ and $G[a,b]$ as spaces of classical functions $v\colon [a,b] \to \mathbb{R}$ in this paper, analogously to [30], Defs. 2.1.1, 4.1.1. In cases in which the transition to equivalence classes in the sense of $L^\infty(a,b)$ becomes necessary, we highlight this by working with the symbols $L^\infty_G(a,b) := \{v \in L^\infty(a,b) \mid v \text{ has a representative in } G[a,b]\}$ and $L^\infty_{BV}(a,b) := \{v \in L^\infty(a,b) \mid v \text{ has a representative in } BV[a,b]\}$. For details on the relationship between the spaces $BV[a,b]$, $G[a,b]$, $L^\infty_G(a,b)$, and $L^\infty_{BV}(a,b)$, see [31], Section 3.2 and Section 6. Recall that $BV[a,b]$ is a Banach space when endowed with the norm $\|v\|_{BV[a,b]} := |v(a)| + \operatorname{var}(v;[a,b])$, where $\operatorname{var}(v;[a,b])$ denotes the variation of a function $v\colon [a,b] \to \mathbb{R}$ in the sense of [30], Definition 2.1.1. Analogously, $G[a,b]$ is a Banach space when equipped with the supremum-norm $\|\cdot\|_\infty$; see [30], Theorem 4.2.1. Due to [30], Remark 4.1.2, we have $C([a,b]) \subset G[a,b]$ and $BV[a,b] \subset G[a,b]$. From the definition of $G[a,b]$, it follows further that every $v \in G[a,b]$ possesses well-defined finite left limits at all points $t \in (a,b]$ and well-defined finite right limits at all points $t \in [a,b)$. We denote these limits by $v(t-)$ and $v(t+)$, respectively, *i.e.*, we define

$$v(t-) := \lim_{[a,t) \ni s \to t} v(s) \quad \forall t \in (a,b] \qquad \text{and} \qquad v(t+) := \lim_{(t,b] \ni s \to t} v(s) \quad \forall t \in [a,b). \tag{2.1}$$

At the endpoints of the interval $[a,b]$, we use the conventions $v(a-) := 0$ and $v(b+) := 0$ consistent with an extension by zero; see Section 6. We remark that additional symbols etc. are introduced in the following sections wherever necessary. This notation is explained on its first appearance in the text.

2.2. The problem under consideration

Since using the weak derivative $u := g'$ as an optimization variable offers various advantages (see Section 2.3 below for details), in the remainder of this paper, we focus on the reformulation (1.2) of the identification problem (1.1). The problem that we are henceforth concerned with thus reads as follows:

$$\left.\begin{aligned} \text{Minimize} \quad & J(y,u) := \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \nu_1 \|u\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(\mathbb{R})} \\ \text{w.r.t.} \quad & y \in H^1(\Omega), \quad u \in L^2(\mathbb{R}), \\ \text{s.t.} \quad & -\Delta y + g_u(y) = f \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega, \\ \text{and} \quad & g_u(t) = \int_0^t u(s)\,\mathrm{d}s \quad \forall t \in \mathbb{R}, \\ \text{and} \quad & u \geq 0 \text{ a.e. in } \mathbb{R}. \end{aligned}\right\} \tag{P}$$

Here, we have introduced the label (P) for the ease of reference and the letter J to denote the objective function of (P), *i.e.*,

$$J\colon L^2(\Omega) \times L^2(\mathbb{R}) \to [0,\infty], \qquad (y,u) \mapsto \frac{1}{2}\|y - y_D\|^2_{L^2(\Omega)} + \nu_1\|u\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(\mathbb{R})}. \tag{2.2}$$

For the sake of clarity, we collect our standing assumptions on the quantities in (P) in:

Assumption 2.1 (Standing Assumptions)**.**

- $d \in \{1, 2, 3\}$;
- $\Omega \subset \mathbb{R}^d$ is a bounded domain that is convex or of class $C^{1,1}$ (in the sense of [29], Sect. 6.2);
- $y_D \in L^2(\Omega)$ and $u_D \in L^2(\mathbb{R})$ are given (the desired state and control);
- $\nu_1 \geq 0$ and $\nu_2 > 0$ are given Tikhonov parameters;
- $f \in L^2(\Omega)$ is a given right-hand side that is not identical zero.

We emphasize that the above conditions are always assumed to hold in our analysis, even when not explicitly mentioned. Regarding the shorthand notation g_u in (P), we point out that we will use this symbol not only for controls $u \in L^2(\mathbb{R})$ but also for arbitrary $u \in L^1_{loc}(\mathbb{R})$, *i.e.*, we set

$$g_u\colon \mathbb{R} \to \mathbb{R}, \qquad g_u(t) := \int_0^t u(s)\,\mathrm{d}s \qquad \forall t \in \mathbb{R} \qquad \forall u \in L^1_{loc}(\mathbb{R}). \tag{2.3}$$

Here and in what follows, we define

$$\int_0^0 u(s)\,\mathrm{d}s := 0 \qquad \text{and} \qquad \int_b^a u(s)\,\mathrm{d}s := -\int_a^b u(s)\,\mathrm{d}s \qquad \forall a, b \in \mathbb{R}, a < b, \qquad \forall u \in L^1_{loc}(\mathbb{R}).$$

Note that these conventions ensure that $g_u' = u$ holds for all $u \in L^1_{loc}(\mathbb{R})$ in the sense of weak derivatives. With a slight abuse of notation, we also use the symbol g_u to denote Nemytskii operators (with varying domains of definition and image spaces) that are induced by the scalar functions in (2.3). The function spaces that these Nemytskii operators act on by superposition will always be clear from the context. For the sake of brevity, we further introduce the following notation for the differential operator in the governing PDE of (P):

$$A_u\colon H_0^1(\Omega) \to H^{-1}(\Omega), \qquad v \mapsto -\Delta v + g_u(v) \qquad \forall u \in L^q(\mathbb{R}), 2 \leq q \leq \infty. \tag{2.4}$$

It will be shown in Lemma 3.3 that the operator A_u in (2.4) is indeed well defined as a function from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$ for all $u \in L^q(\mathbb{R})$, $2 \leq q \leq \infty$.

2.3. Remarks on the problem setting and the choice of the regularization term

Before we begin with our analysis, we would like to emphasize that we consider (P) as a model problem in this paper. It is easy to check that the results that we derive in the following can be extended to, *e.g.*, more general elliptic partial differential operators, more complex objective functions (with some caveats regarding $H^2(\Omega)$-regularity results etc.; see [32], Theorem 3.2.1.2, [29], Theorem 9.15, and Sections 3 and 6), and problems which also involve an upper control constraint. We do not study these generalizations here since we are mainly interested in the nonstandard coupling between y and u in (P) and the consequences that this coupling has for the derivation of optimality conditions in the presence of nonsmooth controls and pointwise constraints.

Regarding the choice of the regularization term $\nu_1\|g'\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|g' - u_D\|^2_{L^2(\mathbb{R})}$ in (1.1), it should be noted that a control-cost term that acts solely on the weak derivative of the function $g = g_u$ is a very sensible thing to consider when the aim is to identify an "optimal" Nemytskii operator in a PDE like $-\Delta y + g(y) = f$. Indeed, including, for instance, an L^2-term $\nu_3\|g\|_{L^2(\mathbb{R})}$ with $\nu_3 > 0$ into the objective function of (1.1) would be completely nonsensical since, in combination with the constraint $g' \geq 0$ (which is required to ensure the well-posedness of the state-equation), such a norm causes the effective admissible set of (1.1) to consist only of the zero function. A possible work-around for this problem would be to estimate the range of the states y appearing in (1.1) a-priori, *i.e.*, to determine a constant $r > 0$ such that $\mathrm{Range}(y) \subset [-r, r]$ holds for all y, to pose the problem (1.1) in the space $H^1(-r, r)$, and to then include the regularization term $\nu_3\|g\|_{L^2(-r,r)}$ into the objective; *cf.* Lemma 3.5. However, if this approach is used, then the solutions of the resulting minimization problem depend on the initial guess for r (as one may easily check by letting r go to infinity) and the quality of the a-priori estimate for the range of the states has a significant impact on the obtained results. This effect is highly undesirable, both analytically and for the numerical realization.

By considering a regularization term of the form $\nu_1\|g'\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|g' - u_D\|^2_{L^2(\mathbb{R})}$ in the objective of (1.1), we can avoid the above problems. As we will see in Section 4, this regularizer automatically ensures that (1.1) can be reformulated as a problem in $L^2(-r, r)$ for all $r > 0$ that are bigger than a computable lower bound r_P without introducing an artificial dependence on the choice of r. Moreover, in the case $u_D \equiv 0$, the cost term in (1.1) naturally gives rise to locally optimal controls $\bar{u} = g'_{\bar{u}}$ that are "tight" in the sense that they are supported only on the range of the associated states $\bar{y}$ and, thus, on precisely those parts of the real line that are relevant for the PDE $-\Delta\bar{y} + g_{\bar{u}}(\bar{y}) = f$; *cf.* Corollary 7.3. As an additional benefit, working with the optimization variable $u = g'_u$ also makes it possible to transform the monotonicity constraint in (1.1) into the far easier to handle pointwise $L^2(\mathbb{R})$-control constraint in (P). We remark that a possible alternative to the inclusion of the regularization term $\nu_1\|g'\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|g' - u_D\|^2_{L^2(\mathbb{R})}$ is the introduction of an additional upper control constraint on g' in (1.1); *cf.* [16, 17, 21]. After changing to the variable $u = g'$, this approach gives rise to a problem with classical L^2-box constraints that is of bang-bang type. We leave the analysis of the problem class that is obtained along these lines and in particular the question of whether a regularity result analogous to (1.4) can be proven for such a variant of (P) for future research.

3. Basic properties of the control-to-state mapping

In this section, we prove basic results on the control-to-state mapping $u \mapsto y$ of the problem (P). We begin with the following elementary observation:

Lemma 3.1 (Hölder Continuity of g_u)**.** *For all $u \in L^q(\mathbb{R})$, $1 \le q \le \infty$, it holds*

$$|g_u(t_1) - g_u(t_2)| \le \|u\|_{L^q(\mathbb{R})}|t_1 - t_2|^{1-1/q} \qquad \forall t_1, t_2 \in \mathbb{R}. \tag{3.1}$$

Proof. This follows immediately from Hölder's inequality (with the usual conventions $1/\infty := 0$ and $0^0 := 1$):

$$|g_u(t_1) - g_u(t_2)| = \left|\int_{t_1}^{t_2} u(s)\,\mathrm{d}s\right| \le \|u\|_{L^q(\mathbb{R})}|t_1 - t_2|^{1-1/q} \qquad \forall t_1, t_2 \in \mathbb{R}.$$

□

Before we study which consequences the last result has for the differential operator A_u in (2.4), we introduce some additional notation that will be used throughout this work.

Definition 3.2 (Important Constants)**.** Henceforth, we denote:

i) by $c_P > 0$ the constant in the inequality of Poincaré-Friedrichs

$$c_P \int_\Omega v^2\,\mathrm{d}x \le \int_\Omega |\nabla v|^2\,\mathrm{d}x \qquad \forall v \in H^1_0(\Omega); \tag{3.2}$$

ii) by $\varepsilon_P > 0$ a constant $\varepsilon_P \in (0, c_P)$ whose value is fixed at this point once and for all;
iii) by $U_P \subset L^2(\mathbb{R})$ the set $U_P := \{u \in L^2(\mathbb{R}) \mid u \ge -\varepsilon_P \text{ a.e. in } \mathbb{R}\}$;
iv) by $y_P \in H^1_0(\Omega) \cap H^2(\Omega) \setminus \{0\}$ the solution of the Poisson problem

$$-\Delta y_P = f \text{ in } \Omega, \qquad y_P = 0 \text{ on } \partial\Omega; \tag{3.3}$$

v) by $r_P > 0$ the constant $r_P := 2\|y_P\|_\infty$.

Note that y_P is indeed an element of $H^1_0(\Omega) \cap H^2(\Omega) \setminus \{0\}$ and $C(\overline{\Omega})$ by our standing assumptions on f, Ω, and d, the regularity results in [29], Theorem 9.15, Lemma 9.17 and [32], Theorems 3.1.3.1, 3.2.1.2, and the Sobolev embeddings. More precisely, we obtain from [29, 32] that every function $v \in H_0^1(\Omega)$ satisfying $\Delta v \in L^2(\Omega)$ is an element of $H_0^1(\Omega) \cap H^2(\Omega)$ and that there exists a constant $C > 0$ such that the following estimate is true:

$$\|v\|_{H^2(\Omega)} \le C\|\Delta v\|_{L^2(\Omega)} \qquad \forall v \in H_0^1(\Omega) \cap H^2(\Omega). \tag{3.4}$$

We can now prove:

Lemma 3.3 (Properties of A_u)**.** *For every $u \in L^q(\mathbb{R})$, $2 \le q \le \infty$, the operator $A_u(v) := -\Delta v + g_u(v)$ in (2.4) is well defined as a function $A_u\colon H_0^1(\Omega) \to H^{-1}(\Omega)$. Moreover, the following is true:*

i) There exists a constant $C > 0$ such that, for all $u \in L^q(\mathbb{R})$, $2 \le q \le \infty$, we have

$$\|A_u(v_1) - A_u(v_2)\|_{H^{-1}(\Omega)} \le C\|v_1 - v_2\|_{H^1(\Omega)} + \|u\|_{L^q(\mathbb{R})}\|v_1 - v_2\|_{L^{2-2/q}(\Omega)}^{1-1/q} \quad \forall v_1, v_2 \in H_0^1(\Omega).$$

ii) There exists a constant $c > 0$ such that, for all $u \in U_P$, it holds

$$\langle A_u(v_1) - A_u(v_2), v_1 - v_2\rangle_{H_0^1(\Omega)} \ge c\|v_1 - v_2\|_{H^1(\Omega)}^2 \quad \forall v_1, v_2 \in H_0^1(\Omega).$$

Proof. For all $u \in L^q(\mathbb{R})$, $2 \le q \le \infty$, and $v \in H_0^1(\Omega)$, we have $-\Delta v \in H^{-1}(\Omega)$ and, due to (3.1) and $g_u(0) = 0$,

$$\|g_u(v)\|_{L^2(\Omega)}^2 = \int_\Omega g_u(v)^2 \,\mathrm{d}x \le \int_\Omega \left(\|u\|_{L^q(\mathbb{R})}|v|^{1-1/q}\right)^2 \mathrm{d}x = \|u\|_{L^q(\mathbb{R})}^2 \|v\|_{L^{2-2/q}(\Omega)}^{2-2/q}. \tag{3.5}$$

Because of the embeddings $H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow L^{2-2/q}(\Omega)$, $q \ge 2$, (3.5) yields that $g_u(v) \in L^2(\Omega) \subset H^{-1}(\Omega)$ holds and that A_u is well defined as a function from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$ for all $u \in L^q(\mathbb{R})$, $2 \le q \le \infty$. This proves the first assertion of the lemma. Next, we note that (3.1) implies

$$\begin{aligned}
\|A_u(v_1) - A_u(v_2)\|_{H^{-1}(\Omega)} &\le \|-\Delta(v_1 - v_2)\|_{H^{-1}(\Omega)} + \|g_u(v_1) - g_u(v_2)\|_{L^2(\Omega)} \\
&\le C\|v_1 - v_2\|_{H^1(\Omega)} + \left(\int_\Omega \left(\|u\|_{L^q(\mathbb{R})}|v_1 - v_2|^{1-1/q}\right)^2 \mathrm{d}x\right)^{1/2} \\
&= C\|v_1 - v_2\|_{H^1(\Omega)} + \|u\|_{L^q(\mathbb{R})}\|v_1 - v_2\|_{L^{2-2/q}(\Omega)}^{1-1/q}
\end{aligned}$$

for all $u \in L^q(\mathbb{R})$, $2 \le q \le \infty$, and $v_1, v_2 \in H_0^1(\Omega)$, where $C > 0$ denotes the operator norm of the Laplacian $\Delta\colon H_0^1(\Omega) \to H^{-1}(\Omega)$. This establishes i). It remains to prove ii). To this end, let us assume that $u \in U_P$ is given. From the definition of g_u, we obtain that the function $g_{\max(0,u)}\colon \mathbb{R} \to \mathbb{R}$ is nondecreasing, that the function $g_{\min(0,u)}\colon \mathbb{R} \to \mathbb{R}$ is nonincreasing, and that $g_u = g_{\max(0,u)} + g_{\min(0,u)}$ holds. By exploiting these properties, (3.1), and the definitions of U_P, c_P, and ε_P, we obtain

$$\begin{aligned}
\langle A_u(v_1) - A_u(v_2), v_1 - v_2\rangle_{H_0^1(\Omega)} &= \int_\Omega |\nabla(v_1 - v_2)|^2 \,\mathrm{d}x + \int_\Omega (g_u(v_1) - g_u(v_2))(v_1 - v_2)\,\mathrm{d}x \\
&= \|\nabla(v_1 - v_2)\|^2_{L^2(\Omega)^d} + \int_\Omega (g_{\min(0,u)}(v_1) - g_{\min(0,u)}(v_2))(v_1 - v_2)\,\mathrm{d}x \\
&\qquad + \int_\Omega (g_{\max(0,u)}(v_1) - g_{\max(0,u)}(v_2))(v_1 - v_2)\,\mathrm{d}x \\
&\geq \|\nabla(v_1 - v_2)\|^2_{L^2(\Omega)^d} - \int_\Omega \left|g_{\min(0,u)}(v_1) - g_{\min(0,u)}(v_2)\right| |v_1 - v_2|\,\mathrm{d}x \\
&\geq \frac{c_P - \varepsilon_P}{2c_P}\|\nabla(v_1 - v_2)\|^2_{L^2(\Omega)^d} + \frac{\varepsilon_P + c_P}{2}\|v_1 - v_2\|^2_{L^2(\Omega)} - \varepsilon_P \int_\Omega (v_1 - v_2)^2\,\mathrm{d}x \\
&\geq \frac{c_P - \varepsilon_P}{2\max(1, c_P)}\|v_1 - v_2\|^2_{H^1(\Omega)} \qquad \forall v_1, v_2 \in H_0^1(\Omega).
\end{aligned} \tag{3.6}$$

This establishes ii) and completes the proof. ☐

Using standard results and Lemma 3.3, we obtain:

Theorem 3.4 (Properties of the Control-to-State Mapping)**.** *The partial differential equation*

$$-\Delta y + g_u(y) = f \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega, \tag{3.7}$$

possesses a unique solution $y \in H_0^1(\Omega) \cap H^2(\Omega)$ for all $u \in U_P$. The solution operator $S\colon U_P \to H_0^1(\Omega) \cap H^2(\Omega)$, $u \mapsto y$, associated with the PDE (3.7) possesses the following properties:

i) The map S is completely continuous as a function from $U_P \subset L^2(\mathbb{R})$ to $H^2(\Omega)$, i.e.,

$$\{u_n\} \subset U_P, u_n \rightharpoonup u \text{ in } L^2(\mathbb{R}) \qquad \Rightarrow \qquad S(u_n) \to S(u) \text{ in } H^2(\Omega).$$

ii) There exists a constant $M > 0$ such that

$$S(U_P) \subset \left\{v \in H_0^1(\Omega) \cap H^2(\Omega) \;\middle|\; \|v\|_{H^2(\Omega)} \leq M\right\}. \tag{3.8}$$

iii) There exists a constant $c > 0$ such that

$$c\|S(u_1) - S(u_2)\|^2_{H^1(\Omega)} \leq \int_\Omega g_{u_2 - u_1}(S(u_1))(S(u_1) - S(u_2))\,\mathrm{d}x \quad \forall u_1, u_2 \in U_P.$$

iv) There exists a constant $C > 0$ such that, for all $2 \leq q \leq \infty$, it holds

$$\|S(u_1) - S(u_2)\|_{H^1(\Omega)} \leq C\|u_1 - u_2\|_{L^q(\mathbb{R})} \qquad \forall u_1, u_2 \in U_P \cap L^q(\mathbb{R}).$$

Proof. Let $u \in U_P$ be given. From Lemma 3.3, the continuous embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$, and the identity $A_u(0) = 0$, we obtain that $A_u\colon H_0^1(\Omega) \to H^{-1}(\Omega)$ is a continuous, coercive, and strongly monotone operator. In combination with [33], Theorem 4.1, this implies that there exists a unique $y \in H_0^1(\Omega)$ satisfying $A_u(y) = f$ in $H^{-1}(\Omega)$. The PDE (3.7) thus possesses a unique weak solution $y = S(u) \in H_0^1(\Omega)$. To see that we even have $y \in H^2(\Omega)$, we note that $g_u(y)$ is an element of $L^2(\Omega)$ by (3.5) and that $-\Delta y = f - g_u(y)$ holds in $H^{-1}(\Omega)$ by the definition of A_u. Since $f \in L^2(\Omega)$, this allows us to invoke (3.4) to obtain $y \in H^2(\Omega)$ as desired.

It remains to prove the properties of the solution map $S\colon u \mapsto y$ in i) to iv). To this end, we first note that (3.7), the identity $A_u(0) = 0$, and Lemma 3.3ii) imply that there exists a constant $c > 0$ such that

$$c\|y\|^2_{H^1(\Omega)} \leq \langle A_u(y), y\rangle_{H^1_0(\Omega)} = \int_\Omega f y \,\mathrm{d}x \leq \|f\|_{L^2(\Omega)}\|y\|_{H^1(\Omega)} \tag{3.9}$$

holds for all $u \in U_P$ and $y = S(u)$. This shows that the set $S(U_P)$ is bounded in $H^1(\Omega)$. Suppose now that a sequence $\{u_n\} \subset U_P$ satisfying $u_n \rightharpoonup u$ in $L^2(\mathbb{R})$ for some $u \in L^2(\mathbb{R})$ is given and define $y_n := S(u_n)$. Then it follows from the lemma of Mazur that u is an element of U_P. From the boundedness of the sequence $\{y_n\}$ in $H^1(\Omega)$, the theorem of Banach-Alaoglu, and the Rellich-Kondrachov embedding theorem, we further obtain that we can pass over to a subsequence (denoted the same) such that $\{y_n\}$ converges weakly in $H^1_0(\Omega)$ and strongly in $L^2(\Omega)$ to a function $y \in H^1_0(\Omega)$. In combination with (3.1), the embedding $L^2(\Omega) \hookrightarrow L^1(\Omega)$, the fact that $u_n \rightharpoonup u$ in $L^2(\mathbb{R})$ implies $g_{u_n-u}(t) \to 0$ for all $t \in \mathbb{R}$, the boundedness of $\{u_n\}$ in $L^2(\mathbb{R})$, the majorization estimate

$$|g_{u_n-u}(y)|^2 \leq \|u_n - u\|^2_{L^2(\mathbb{R})}|y| \qquad \text{a.e. in } \Omega$$

obtained from (3.1), and the dominated convergence theorem, this yields

$$\begin{aligned}\|g_u(y) - g_{u_n}(y_n)\|_{L^2(\Omega)} &= \|g_u(y) - g_{u_n}(y) + g_{u_n}(y) - g_{u_n}(y_n)\|_{L^2(\Omega)}\\ &\leq \|g_{u-u_n}(y)\|_{L^2(\Omega)} + \|g_{u_n}(y) - g_{u_n}(y_n)\|_{L^2(\Omega)}\\ &\leq \|g_{u-u_n}(y)\|_{L^2(\Omega)} + \|u_n\|_{L^2(\mathbb{R})}\|y - y_n\|^{1/2}_{L^1(\Omega)} \to 0\end{aligned} \tag{3.10}$$

for $n \to \infty$. Since the Laplacian $-\Delta\colon H^1_0(\Omega) \to H^{-1}(\Omega)$ is linear and continuous and since $A_{u_n}(y_n) = f$ holds by the definition of y_n, the above implies

$$0 = A_{u_n}(y_n) - f = -\Delta y_n + g_{u_n}(y_n) - f \rightharpoonup -\Delta y + g_u(y) - f = A_u(y) - f \text{ in } H^{-1}(\Omega).$$

The limit y of the sequence $\{y_n\}$ thus satisfies $A_u(y) - f = 0$ and we may conclude that $y = S(u)$ holds and that $S(u_n)$ converges weakly in $H^1_0(\Omega)$ to $S(u)$ for $n \to \infty$. As the above arguments can be repeated along arbitrary subsequences of the initial $\{u_n\} \subset L^2(\mathbb{R})$, it follows from a trivial contradiction argument that the convergence $S(u_n) \rightharpoonup S(u)$ in $H^1_0(\Omega)$ also holds for the whole original sequence of controls $\{u_n\}$; *cf.* [26], Lemma 4.16. This proves that S is weakly continuous as a function from $U_P \subset L^2(\mathbb{R})$ to $H^1_0(\Omega)$. To see that we even have strong convergence of the states in $H^2(\Omega)$, we note that the estimates (3.4) and (3.10) imply that there exists a constant $C > 0$ such that

$$\|y_n - y\|_{H^2(\Omega)} \leq C\| - \Delta y_n + \Delta y\|_{L^2(\Omega)} = C\|g_u(y) - g_{u_n}(y_n)\|_{L^2(\Omega)} \to 0$$

holds for $n \to \infty$. We thus indeed have $S(u_n) \to S(u)$ in $H^2(\Omega)$ for all $\{u_n\} \subset U_P$ that satisfy $u_n \rightharpoonup u$ in $L^2(\mathbb{R})$ for some $u \in L^2(\mathbb{R})$. This completes the proof of i).

To prove ii), let us assume that $u \in U_P \cap L^\infty(\mathbb{R})$ is given. From the identity $g_u = g_{\max(0,u)} + g_{\min(0,u)}$ and (3.7), we obtain that $y := S(u)$ is the (necessarily unique) weak solution of the semilinear PDE

$$-\Delta y + g_{\max(0,u)}(y) = \tilde{f} \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega, \tag{3.11}$$

whose right-hand side $\tilde{f}$ is defined by $\tilde{f} := f - g_{\min(0,u)}(y) \in L^2(\Omega)$. Note that the definition of U_P, (3.5), and the estimate (3.9) imply that $\|\min(0,u)\|_{L^\infty(\mathbb{R})} \leq \varepsilon_P$ holds and that $\tilde{f}$ satisfies

$$\|\tilde{f}\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} + \|g_{\min(0,u)}(y)\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} + \|\min(0,u)\|_{L^\infty(\mathbb{R})}\|y\|_{L^2(\Omega)} \leq \left(1 + \frac{\varepsilon_P}{c}\right)\|f\|_{L^2(\Omega)}. \tag{3.12}$$

From the L^∞-regularity of u, the definition of the function g_u, the regularity $y \in H_0^1(\Omega) \cap H^2(\Omega) \subset C(\overline{\Omega})$, and [34], Theorem 2.1.11, it follows further that the map $g_{\max(0,u)} \colon \mathbb{R} \to \mathbb{R}$ is globally Lipschitz continuous, that $g_{\max(0,u)}(0) = 0$ holds, and that $v := g_{\max(0,u)}(y)$ is an element of $H_0^1(\Omega)$ whose weak gradient is given by $\nabla v = \max(0, u(y))\nabla y$ a.e. in Ω. In combination with (3.11), the last identity yields

$$\|v\|^2_{L^2(\Omega)} \leq \int_\Omega \max(0,u(y))|\nabla y|^2 \,\mathrm{d}x + \|v\|^2_{L^2(\Omega)} = \int_\Omega \nabla y \cdot \nabla v + g_{\max(0,u)}(y) v \,\mathrm{d}x = \int_\Omega \tilde{f} v \,\mathrm{d}x \leq \|v\|_{L^2(\Omega)}\|\tilde{f}\|_{L^2(\Omega)}.$$

We thus have $\|v\|_{L^2(\Omega)} \leq \|\tilde{f}\|_{L^2(\Omega)}$. By revisiting (3.11), by exploiting (3.12), and by again invoking the estimate (3.4), we now obtain that there exists a constant $C > 0$ independent of u satisfying

$$\|y\|_{H^2(\Omega)} \leq C\|\tilde{f} - v\|_{L^2(\Omega)} \leq 2C\|\tilde{f}\|_{L^2(\Omega)} \leq 2C\left(1 + \frac{\varepsilon_P}{c}\right)\|f\|_{L^2(\Omega)}.$$

This proves the bound in ii) with $M := 2C\left(1 + \frac{\varepsilon_P}{c}\right)\|f\|_{L^2(\Omega)}$ for all $u \in U_P \cap L^\infty(\mathbb{R})$. Since the set $U_P \cap L^\infty(\mathbb{R})$ is dense in U_P w.r.t. convergence in $L^2(\mathbb{R})$ and since S is continuous as a function from $U_P \subset L^2(\mathbb{R})$ to $H^2(\Omega)$ by i), the assertion of ii) now follows immediately.

To establish iii), we note that, for all $u_1, u_2 \in U_P$ with associated states $y_1 := S(u_1)$ and $y_2 := S(u_2)$, we obtain from (3.7) and Lemma 3.3ii) that

$$\begin{aligned} c\|y_1 - y_2\|^2_{H^1(\Omega)} &\leq \langle A_{u_2}(y_1) - A_{u_2}(y_2), y_1 - y_2\rangle_{H_0^1(\Omega)} \\ &= \langle A_{u_1}(y_1) - A_{u_2}(y_2), y_1 - y_2\rangle_{H_0^1(\Omega)} + \langle A_{u_2}(y_1) - A_{u_1}(y_1), y_1 - y_2\rangle_{H_0^1(\Omega)} \\ &= 0 + \int_\Omega \left(g_{u_2}(y_1) - g_{u_1}(y_1)\right)(y_1 - y_2)\,\mathrm{d}x = \int_\Omega g_{u_2 - u_1}(y_1)(y_1 - y_2)\,\mathrm{d}x. \end{aligned} \tag{3.13}$$

This proves iii).

To finally obtain iv), we note that, for all $u_1, u_2 \in U_P \cap L^q(\mathbb{R})$, $2 \leq q \leq \infty$, we can use (3.5) to continue the estimate (3.13) as follows:

$$c\|y_1 - y_2\|^2_{H^1(\Omega)} \leq \|y_1 - y_2\|_{L^2(\Omega)}\|g_{u_2 - u_1}(y_1)\|_{L^2(\Omega)} \leq \|y_1 - y_2\|_{H^1(\Omega)}\|u_1 - u_2\|_{L^q(\mathbb{R})}\|y_1\|^{1-1/q}_{L^{2-2/q}(\Omega)}.$$

In combination with ii), this establishes iv) and completes the proof of the theorem. □

Note that the boundedness of the set $S(U_P) \subset H^2(\Omega)$ in Theorem 3.4ii), our assumption $d \in \{1,2,3\}$, and the Sobolev embeddings imply that there exists a constant $r > 0$ such that $\|S(u)\|_\infty \leq r$ holds for all $u \in U_P$. This property will later on allow us to restate (P) as a problem in $L^2(-r,r)$; see Theorem 4.5. Unfortunately, the precise value of the bound r on the maximum-norm of the states $S(u)$, $u \in U_P$, obtained from (3.8) is hard to calculate precisely as it depends on c_P and various embedding constants. This is a problem since, for an efficient numerical solution of (P), a bound on the range of the appearing states is required that is as sharp as possible. The following lemma shows that, for the nonnegative controls in (P), a much more tangible $\|\cdot\|_\infty$-estimate can be obtained in terms of the solution $y_P = S(0)$ of the Poisson equation in (3.3).

Lemma 3.5 ($\|\cdot\|_\infty$-Bound)**.** *The solution operator $S\colon U_P \to H_0^1(\Omega)\cap H^2(\Omega)$, $u \mapsto y$, of (3.7) satisfies*

$$\|S(u)\|_\infty \le r_P \qquad \forall\, 0 \le u \in L^2(\mathbb{R}). \tag{3.14}$$

Here, $r_P = 2\|y_P\|_\infty$ denotes the constant from Definition 3.2v) and y_P the solution of (3.3).

Proof. From Theorem 3.4 and the Sobolev embeddings, it follows that y_P satisfies $y_P = S(0) \in C(\overline{\Omega})$. Consider now a control $0 \le u \in L^2(\mathbb{R})$ with state $y := S(u) \in H_0^1(\Omega)\cap H^2(\Omega) \subset C(\overline{\Omega})$. We define $z := y - y_P$ and $w := z - \min(\|y_P\|_\infty, \max(-\|y_P\|_\infty, z))$. Due to their construction and the lemma of Stampacchia (see [35], Thm. II.A.1 and [28], Thm. 5.8.2), the functions z and w satisfy $z, w \in H_0^1(\Omega)$,

$$\int_\Omega \nabla z\cdot\nabla v + g_u(y)v \,\mathrm{d}x = 0 \qquad \forall v \in H_0^1(\Omega), \tag{3.15}$$

and

$$w = \begin{cases} z - \|y_P\|_\infty & \text{if } z > \|y_P\|_\infty, \\ 0 & \text{if } -\|y_P\|_\infty \le z \le \|y_P\|_\infty, \\ z + \|y_P\|_\infty & \text{if } z < -\|y_P\|_\infty. \end{cases}$$

By choosing w as the test function in (3.15) and by exploiting the formulas in [28], Theorem 5.8.2, we obtain

$$0 = \int_\Omega \nabla z\cdot\nabla w\,\mathrm{d}x + \int_\Omega g_u(y)w\,\mathrm{d}x = \int_\Omega |\nabla w|^2\,\mathrm{d}x + \int_\Omega g_u(y)w\,\mathrm{d}x \ge \int_\Omega |\nabla w|^2\,\mathrm{d}x. \tag{3.16}$$

Here, the last inequality follows from the implications

$$\begin{aligned} w > 0 \quad &\Rightarrow \quad y - y_P = z > \|y_P\|_\infty \quad &&\Rightarrow \quad y \ge y_P + \|y_P\|_\infty \quad &&\Rightarrow \quad y \ge 0, \\ w < 0 \quad &\Rightarrow \quad y - y_P = z < -\|y_P\|_\infty \quad &&\Rightarrow \quad y \le y_P - \|y_P\|_\infty \quad &&\Rightarrow \quad y \le 0, \end{aligned}$$

the fact that g_u is nondecreasing, and the identity $g_u(0) = 0$. Due to (3.2), (3.16) yields $w = 0$ and, by the definition of w, $y_P - \|y_P\|_\infty \le y \le y_P + \|y_P\|_\infty$. This establishes (3.14) and completes the proof. $\square$

Remark 3.6. Inequality (3.14) is not true anymore in general when the factor two in the definition of r_P is dropped. Indeed, numerical evidence suggests that the estimate $\|S(u)\|_\infty \le 2\|y_P\|_\infty$ for all $u \in L^2_+(\mathbb{R})$ is optimal.

4. First consequences for the optimization problem

Having studied the properties of the state-equation (3.7), we now turn our attention to the analysis of the optimization problem (P). Henceforth, with the letters S and J, we always refer to the control-to-state mapping $S\colon U_P \to H_0^1(\Omega)\cap H^2(\Omega)$ introduced in Theorem 3.4 and the function J defined in (2.2). For the definitions of the quantities ε_P, r_P, and U_P appearing in the following, see Definition 3.2. We first state:

Definition 4.1 (Local and Global Solutions)**.** A function $0 \le \bar{u} \in L^2(\mathbb{R})$ is called a local minimizer/local solution/locally optimal control of (P) if there exists a number $\rho > 0$ (the radius of optimality) such that

$$J(S(u), u) \ge J(S(\bar{u}), \bar{u}) \qquad \forall\, 0 \le u \in B_\rho^{L^2(\mathbb{R})}(\bar{u}). \tag{4.1}$$

If ρ can be chosen as ∞, then we call $\bar{u}$ a global minimizer/global solution/globally optimal control of (P).

It is easy to check that a control $\bar{u}$ can only be a local minimizer of (P) if $J(S(\bar{u}), \bar{u}) < \infty$ holds. In the case $\nu_1 > 0$, all local solutions of (P) thus satisfy $\bar{u} \in L^1(\mathbb{R})$. For the various variants of the problem (P) that appear in the following sections, the notions of local and global optimality are defined analogously to (4.1). By exploiting the complete continuity of the map S in Theorem 3.4i), we obtain:

Theorem 4.2 (Solvability of (P))**.** *The problem* (P) *possesses at least one globally optimal control* $\bar{u} \in L^2(\mathbb{R})$.

Proof. This follows immediately from the direct method of the calculus of variations. Indeed, the admissible set of the problem (P) is clearly nonempty (as it contains the zero function) and we can find nonnegative controls $\{u_n\} \subset L^2(\mathbb{R})$ with associated states $y_n := S(u_n)$ such that the function values $J(y_n, u_n)$ are finite and such that the sequence $\{J(y_n, u_n)\}$ converges to the infimal value of (P). Due to the structure of J, the sequence $\{u_n\}$ is bounded in $L^2(\mathbb{R})$ and we may assume w.l.o.g. that $u_n \rightharpoonup \bar{u}$ holds in $L^2(\mathbb{R})$ for some (necessarily nonnegative) $\bar{u} \in L^2(\mathbb{R})$. Since the function $L^2(\mathbb{R}) \ni u \mapsto \nu_1 \|u\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(\mathbb{R})} \in [0, \infty]$ is convex and lower semicontinuous by the lemma of Fatou and, as a consequence, weakly lower semicontinuous by the lemma of Mazur, and since the map $U_P \ni u \mapsto \frac{1}{2}\|S(u) - y_D\|^2_{L^2(\Omega)} \in \mathbb{R}$ is completely continuous by Theorem 3.4i), we obtain that $\liminf_{n\to\infty} J(S(u_n), u_n) \geq J(S(\bar{u}), \bar{u})$ holds. Due to the choice of $\{u_n\}$, this proves the claim. □

As already mentioned in Section 3, the $\|\cdot\|_\infty$-bound in (3.14) makes it possible to recast (P) as a minimization problem in $L^2(-r, r)$ for all large enough $r > 0$. To formulate this result rigorously, we introduce:

Definition 4.3 (Restriction and Extension by Zero)**.** Given a number $r > 0$, we define:

i) $\boldsymbol{u}_D \in L^2(-r, r)$ to be the function $\boldsymbol{u}_D := u_D|_{(-r,r)}$;
ii) $E_r \colon L^2(-r, r) \to L^2(\mathbb{R})$ to be the map that extends elements of $L^2(-r, r)$ by zero to elements of $L^2(\mathbb{R})$;
iii) $U_{P,r}$ to be the set $U_{P,r} := \left\{\boldsymbol{u} \in L^2(-r, r) \mid \boldsymbol{u} \geq -\varepsilon_P \text{ a.e. in } (-r, r)\right\} = \left\{\boldsymbol{u} \in L^2(-r, r) \mid E_r(\boldsymbol{u}) \in U_P\right\}$;
iv) S_r to be the operator $S_r \colon U_{P,r} \to H^1_0(\Omega) \cap H^2(\Omega)$, $S_r(\boldsymbol{u}) := S(E_r(\boldsymbol{u}))$;
v) F_r to be the function

$$F_r \colon U_{P,r} \to \mathbb{R}, \qquad F_r(\boldsymbol{u}) := \frac{1}{2}\|S_r(\boldsymbol{u}) - y_D\|^2_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 \boldsymbol{u} + \frac{\nu_2}{2}(\boldsymbol{u} - \boldsymbol{u}_D)^2 \, \mathrm{d}t.$$

Here and in what follows, the typesetting $\boldsymbol{u}$ is used to emphasize that we talk about functions on a subinterval $(-r, r)$ and not about controls u on the whole real line as appearing in the original problem formulation (P). Note that, due to the structure of the PDE (3.7), we have:

Lemma 4.4 (Truncation Invariance)**.** *The operators* S *and* S_r *satisfy*

$$S(u) = S\left(\mathbb{1}_{(-r,r)} u\right) = S_r\left(u|_{(-r,r)}\right) \qquad \forall\, 0 \leq u \in L^2(\mathbb{R}) \qquad \forall r \geq r_P. \tag{4.2}$$

Proof. Let $0 \leq u \in L^2(\mathbb{R})$ and $r \geq r_P$ be given. By the definition of S, the function $y = S(u) \in H^1_0(\Omega) \cap H^2(\Omega)$ is the weak solution of the PDE

$$-\Delta y + g_u(y) = f \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega,$$

and from (3.14), we obtain that the continuous representative of y satisfies $\mathrm{Range}(y) \subset [-r_P, r_P] \subset [-r, r]$. In combination with (2.3), this implies that $g_u(t) = g_{\mathbb{1}_{(-r,r)}u}(t)$ holds for all $t \in [-r, r]$ and that y is also the weak solution of

$$-\Delta y + g_{\mathbb{1}_{(-r,r)}u}(y) = f \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega.$$

Thus, $S(u) = S\left(\mathbb{1}_{(-r,r)} u\right)$ as claimed. The second identity $S\left(\mathbb{1}_{(-r,r)} u\right) = S_r\left(u|_{(-r,r)}\right)$ in (4.2) follows from the definition of S_r. This establishes the assertion. □

We can now prove:

Theorem 4.5 (Reformulation of (P))**.** *Let $r \geq r_P$ and $\bar{u} \in L^2(\mathbb{R})$ be given. Then $\bar{u}$ is a local solution of (P) with radius of optimality ρ if and only if the following two conditions are satisfied:*

i) It holds

$$\bar{u} = \max\left(0, u_D - \frac{\nu_1}{\nu_2}\right) \text{ a.e. in } \mathbb{R} \setminus (-r, r). \tag{4.3}$$

ii) The restriction $\bar{\boldsymbol{u}} := \bar{u}|_{(-r,r)} \in L^2(-r, r)$ is a local solution with radius of optimality ρ of the problem

$$\left.\begin{aligned} \text{Minimize} \quad & F_r(\boldsymbol{u}) = \frac{1}{2}\|S_r(\boldsymbol{u}) - y_D\|^2_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 \boldsymbol{u} + \frac{\nu_2}{2}(\boldsymbol{u} - u_D)^2 \,\mathrm{d}t \\ \text{w.r.t.} \quad & \boldsymbol{u} \in L^2(-r, r), \\ \text{s.t.} \quad & \boldsymbol{u} \geq 0 \text{ a.e. in } (-r, r). \end{aligned}\right\} \tag{P$_r$}$$

Proof. Suppose that $\bar{u} \in L^2(\mathbb{R})$ is a local solution of (P) and that $v \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ is an arbitrary function satisfying $v = 0$ a.e. in $(-r, r)$ and $\bar{u} + s_0 v \geq 0$ a.e. in $\mathbb{R}$ for some $s_0 > 0$. Then it holds $0 \leq \bar{u} + sv \in L^2(\mathbb{R})$ for all $0 < s < s_0$, it follows from the local optimality of $\bar{u}$ that, for all sufficiently small $s > 0$, we have $J(S(\bar{u} + sv), \bar{u} + sv) \geq J(S(\bar{u}), \bar{u})$, and we obtain from $\bar{u} + sv = \bar{u}$ a.e. in $(-r, r)$, $r \geq r_P$, and (4.2) that $S(\bar{u} + sv) = S(\bar{u})$ holds for all $0 < s < s_0$. In combination with the definition of J, this implies

$$\begin{aligned} J(S(\bar{u} + sv), \bar{u} + sv) &= \frac{1}{2}\|S(\bar{u}) - y_D\|^2_{L^2(\Omega)} + \nu_1\|\bar{u} + sv\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|\bar{u} + sv - u_D\|^2_{L^2(\mathbb{R})} \\ \geq J(S(\bar{u}), \bar{u}) &= \frac{1}{2}\|S(\bar{u}) - y_D\|^2_{L^2(\Omega)} + \nu_1\|\bar{u}\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|\bar{u} - u_D\|^2_{L^2(\mathbb{R})} \end{aligned} \tag{4.4}$$

for all sufficiently small $s > 0$ and, as a consequence,

$$\int_{\mathbb{R}} \nu_1 sv + \frac{\nu_2}{2}(\bar{u} + sv - u_D)^2 - \frac{\nu_2}{2}(\bar{u} - u_D)^2 \,\mathrm{d}x \geq 0. \tag{4.5}$$

Here, we have exploited the nonnegativity of the functions $\bar{u}$ and $\bar{u} + sv$ to rewrite the L^1-norms in (4.4). If we use the binomial identities in (4.5), divide by $s\nu_2$, and let s go to zero, then (4.5) yields

$$\int_{\mathbb{R}} \left(\frac{\nu_1}{\nu_2} - u_D + \bar{u}\right) v \,\mathrm{d}x \geq 0. \tag{4.6}$$

By considering functions v of the form $v = \mathbb{1}_{\{\bar{u} \geq \varepsilon\} \cap \mathbb{R} \setminus (-r,r)} w$ with arbitrary $\varepsilon > 0$ and $w \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ in (4.6), we obtain that $\bar{u}$ satisfies $\bar{u} = u_D - \nu_1/\nu_2$ a.e. in $\{\bar{u} > 0\} \cap \mathbb{R} \setminus (-r, r)$, and by considering functions of the form $v = \mathbb{1}_{\{\bar{u} = 0\} \cap \mathbb{R} \setminus (-r,r)} w$ with $0 \leq w \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, that $u_D - \nu_1/\nu_2 \leq 0$ holds a.e. in $\{\bar{u} = 0\} \cap \mathbb{R} \setminus (-r, r)$. This establishes the identity (4.3) in condition i).

Next, we prove that $\bar{u}$ also satisfies ii): Denote the radius of local optimality of $\bar{u}$ with $\rho > 0$ and assume that $0 \leq \boldsymbol{u} \in L^2(-r, r)$ is an arbitrary function such that $\|\bar{\boldsymbol{u}} - \boldsymbol{u}\|_{L^2(-r,r)} \leq \rho$ holds for $\bar{\boldsymbol{u}} := \bar{u}|_{(-r,r)}$. We define $u := E_r(\boldsymbol{u}) + \mathbb{1}_{\mathbb{R} \setminus (-r,r)} \bar{u}$. By construction, the function u satisfies $0 \leq u \in L^2(\mathbb{R})$ and $\|u - \bar{u}\|_{L^2(\mathbb{R})} = \|\bar{\boldsymbol{u}} - \boldsymbol{u}\|_{L^2(-r,r)} \leq \rho$. In combination with the definition of ρ, this yields

$$\frac{1}{2}\|S(u) - y_D\|^2_{L^2(\Omega)} + \nu_1\|u\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(\mathbb{R})} \geq \frac{1}{2}\|S(\bar{u}) - y_D\|^2_{L^2(\Omega)} + \nu_1\|\bar{u}\|_{L^1(\mathbb{R})} + \frac{\nu_2}{2}\|\bar{u} - u_D\|^2_{L^2(\mathbb{R})}. \tag{4.7}$$

Since u is identical to $\bar{u}$ a.e. in $\mathbb{R} \setminus (-r, r)$, since $S(\tilde{u}) = S_r(\tilde{u}|_{(-r,r)})$ holds for all $0 \leq \tilde{u} \in L^2(\mathbb{R})$ by (4.2), and since u and $\bar{u}$ are nonnegative, the inequality (4.7) can be recast as $F_r(\boldsymbol{u}) \geq F_r(\bar{\boldsymbol{u}})$. As $\boldsymbol{u}$ was arbitrary, this shows that $\bar{\boldsymbol{u}} = \bar{u}|_{(-r,r)} \in L^2(-r, r)$ is indeed a local solution of (P_r) (with the same radius of optimality ρ as $\bar{u}$). This completes the proof of the direction "$\Rightarrow$" in the asserted equivalence.

It remains to prove the reverse implication. To this end, let us assume that a function $\bar{u} \in L^2(\mathbb{R})$ satisfying the conditions i) and ii) is given. For such a function $\bar{u}$, we obtain from the properties in i) and ii) that $\bar{u} \geq 0$ holds a.e. in $\mathbb{R}$ and from (4.3) and a simple pointwise calculation that

$$\begin{aligned} &\nu_1 \|u\|_{L^1(\mathbb{R}\setminus(-r,r))} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(\mathbb{R}\setminus(-r,r))} \\ &\qquad \geq \nu_1 \|\bar{u}\|_{L^1(\mathbb{R}\setminus(-r,r))} + \frac{\nu_2}{2}\|\bar{u} - u_D\|^2_{L^2(\mathbb{R}\setminus(-r,r))} \quad \forall\, 0 \leq u \in L^2(\mathbb{R} \setminus (-r, r)). \end{aligned} \tag{4.8}$$

Consider now some $0 \leq u \in L^2(\mathbb{R})$ that satisfies $\|u - \bar{u}\|_{L^2(\mathbb{R})} \leq \rho$ for the radius of local optimality $\rho > 0$ of $\bar{\boldsymbol{u}} := \bar{u}|_{(-r,r)} \in L^2(-r, r)$ in (P_r). We define $\boldsymbol{u} := u|_{(-r,r)}$. For this $\boldsymbol{u}$, it clearly holds $0 \leq \boldsymbol{u} \in L^2(-r, r)$ and $\|\bar{\boldsymbol{u}} - \boldsymbol{u}\|_{L^2(-r,r)} \leq \rho$ and we obtain from condition ii) that

$$\begin{aligned} F_r(\boldsymbol{u}) &= \frac{1}{2}\|S_r(\boldsymbol{u}) - y_D\|^2_{L^2(\Omega)} + \nu_1 \|\boldsymbol{u}\|_{L^1(-r,r)} + \frac{\nu_2}{2}\|\boldsymbol{u} - \boldsymbol{u}_D\|^2_{L^2(-r,r)} \\ &\geq F_r(\bar{\boldsymbol{u}}) = \frac{1}{2}\|S_r(\bar{\boldsymbol{u}}) - y_D\|^2_{L^2(\Omega)} + \nu_1 \|\bar{\boldsymbol{u}}\|_{L^1(-r,r)} + \frac{\nu_2}{2}\|\bar{\boldsymbol{u}} - \boldsymbol{u}_D\|^2_{L^2(-r,r)}. \end{aligned}$$

Due to (4.2) and the definitions of $\boldsymbol{u}$, $\bar{\boldsymbol{u}}$, and $\boldsymbol{u}_D$, the above inequality can be rewritten as

$$\begin{aligned} &\frac{1}{2}\|S(u) - y_D\|^2_{L^2(\Omega)} + \nu_1 \|u\|_{L^1(-r,r)} + \frac{\nu_2}{2}\|u - u_D\|^2_{L^2(-r,r)} \\ &\quad \geq \frac{1}{2}\|S(\bar{u}) - y_D\|^2_{L^2(\Omega)} + \nu_1 \|\bar{u}\|_{L^1(-r,r)} + \frac{\nu_2}{2}\|\bar{u} - u_D\|^2_{L^2(-r,r)}. \end{aligned}$$

In combination with (4.8), this yields $J(S(u), u) \geq J(S(\bar{u}), \bar{u})$. Since $0 \leq u \in L^2(\mathbb{R})$ was an arbitrary function satisfying $\|u - \bar{u}\|_{L^2(\mathbb{R})} \leq \rho$, we may now conclude that $\bar{u}$ is indeed a local solution of (P) with radius of optimality ρ. This establishes the implication "$\Leftarrow$" of the asserted equivalence and completes the proof. $\square$

Note that the inequality $r \geq r_P$ in Theorem 4.5 is only needed for proving the equivalence of the problems (P) and (P_r). For (P_r) to be sensible, it suffices to assume that r is positive. In what follows, when studying the problem (P_r), we work with the relaxed assumption $r > 0$ whenever possible. For later use, we prove:

Proposition 4.6 (Enforcing Global Optimality)**.** *Let $r > 0$ be given and let $0 \leq \bar{\boldsymbol{u}} \in L^2(-r, r)$ be a local solution of the problem (P_r). Then there exists a constant $C_{\bar{\boldsymbol{u}}} > 0$ such that $\bar{\boldsymbol{u}}$ is the unique global solution of the problem*

$$\left.\begin{aligned} \textit{Minimize} \quad & F_r(\boldsymbol{u}) + \frac{C_{\bar{\boldsymbol{u}}}}{2}\|\boldsymbol{u} - \bar{\boldsymbol{u}}\|^2_{L^2(-r,r)} \\ \textit{w.r.t.} \quad & \boldsymbol{u} \in L^2(-r, r), \\ \textit{s.t.} \quad & \boldsymbol{u} \geq 0 \ \textit{a.e. in } (-r, r). \end{aligned}\right\} \tag{$\mathrm{P}_{r,\bar{\boldsymbol{u}}}$}$$

Proof. Due to the local optimality of $\bar{\boldsymbol{u}}$ in (P_r), we know that there exists $\rho > 0$ satisfying

$$F_r(\boldsymbol{u}) \geq F_r(\bar{\boldsymbol{u}}) \qquad \forall\, 0 \leq \boldsymbol{u} \in L^2(-r, r), \|\boldsymbol{u} - \bar{\boldsymbol{u}}\|_{L^2(-r,r)} \leq \rho. \tag{4.9}$$

Define $C_{\bar u} := 1 + 2F_r(\bar u)/\rho^2$. Then it follows from the nonnegativity of F_r on $L^2_+(-r,r)$ that

$$F_r(\bar u) < \frac{C_{\bar u}}{2}\rho^2 \leq \frac{C_{\bar u}}{2}\|u - \bar u\|^2_{L^2(-r,r)} \leq F_r(u) + \frac{C_{\bar u}}{2}\|u - \bar u\|^2_{L^2(-r,r)} \tag{4.10}$$

holds for all $0 \leq u \in L^2(-r,r)$ with $\|u - \bar u\|_{L^2(-r,r)} \geq \rho$. By combining (4.9) and (4.10) and by distinguishing cases, the assertion of the proposition follows. □

Since the function g_u is only Hölder continuous with exponent $1/2$ for arbitrary $u \in L^2(\mathbb{R})$, we cannot state first-order necessary optimality conditions for the problems (P), (P_r), and ($\mathrm{P}_{r,\bar u}$) at this point. To do so, we require additional information about the regularity properties of local minimizers.

5. Regularity of optimal controls

The purpose of this section is to prove that minimizers of the problems (P) and (P_r) are regulated if u_D is a smooth enough function. As we will see in Sections 6 to 8, this type of regularity is crucial for the derivation of stationarity conditions and the design of numerical solution algorithms. We consider the following situation:

Assumption 5.1 (Standing Assumptions for Section 5)**.** Throughout this section, we assume that:

- r_P, ε_P, and U_P are defined as before (see Definition 3.2);
- r is a given positive number;
- u_D, E_r, $U_{P,r}$, S_r, and F_r are defined as in Definition 4.3;
- u_D satisfies $u_D \in L^\infty_G(-r,r)$;
- $0 \leq \bar u \in L^2(-r,r)$ is a given local solution of the problem (P_r);
- $C_{\bar u} > 0$ is a constant such that $\bar u$ and $C_{\bar u}$ satisfy the assertion of Proposition 4.6;
- $\eta\colon (-\varepsilon_P,\infty) \to \mathbb{R}$ is defined by $\eta(t) := \min(0,t)^2/(t+\varepsilon_P)$ for all $t \in (-\varepsilon_P,\infty)$.

Note that, for the function η, we have:

Lemma 5.2 (Properties of η)**.** *The function η satisfies $\eta \in C^1((-\varepsilon_P,\infty))$, $\eta(t) = 0$ for all $t \in [0,\infty)$, $\eta(t) \geq \min(0,t)^2/\varepsilon_P$ for all $t \in (-\varepsilon_P,\infty)$, and $\eta(t) \to \infty$ for $(-\varepsilon_P,\infty) \ni t \to -\varepsilon_P$. Further, η is convex.*

Proof. This can be checked with trivial calculations and estimates. □

The main idea of the subsequent analysis is to relax the constraint $u \geq 0$ a.e. in $(-r,r)$ in ($\mathrm{P}_{r,\bar u}$) by means of the barrier function $\beta\eta\colon (-\varepsilon_P,\infty) \to \mathbb{R}$, $\beta > 0$, to establish a BV-regularity result for the adjoint variables of the resulting family of approximate problems, and to transfer this BV-regularity to the original problem ($\mathrm{P}_{r,\bar u}$) by passing to the limit $\beta \to \infty$. For a given $\beta > 0$, we thus consider the following problem:

$$\left.\begin{aligned} \text{Minimize}\quad & F_r(u) + \frac{C_{\bar u}}{2}\|u - \bar u\|^2_{L^2(-r,r)} + \int_{-r}^{r} \beta\eta(u)\,\mathrm{d}t \\ \text{w.r.t.}\quad & u \in L^2(-r,r), \\ \text{s.t.}\quad & u > -\varepsilon_P \text{ a.e. in } (-r,r). \end{aligned}\right\} \tag{$\mathrm{P}_{r,\bar u,\beta}$}$$

We remark that, for optimal control problems governed by variational inequalities, regularization approaches similar to that developed in this section have already been used, *e.g.*, in [36]. We first note:

Proposition 5.3 (Solvability of ($\mathrm{P}_{r,\bar u,\beta}$))**.** *For every $\beta > 0$, ($\mathrm{P}_{r,\bar u,\beta}$) possesses a global solution $\bar u_\beta \in L^2(-r,r)$.*

Proof. This follows completely analogously to the proof of Theorem 4.2 from the direct method of the calculus of variations, the results on the control-to-state mapping $S\colon U_P \to H^1_0(\Omega) \cap H^2(\Omega)$ in Theorem 3.4, the definitions of the functions S_r and F_r, and the properties of η. □

Due to the $C_{\bar u}$-term in the objective of ($\mathrm{P}_{r,\bar u,\beta}$) and Proposition 4.6, we have:

Proposition 5.4 (Convergence of Approximate Solutions)**.** *Consider for every $\beta > 0$ a global solution $\bar u_\beta \in L^2(-r,r)$ of* ($\mathrm{P}_{r,\bar u,\beta}$)*. Then it holds $\bar u_\beta \to \bar u$ in $L^2(-r,r)$ for $\beta \to \infty$.*

Proof. From the global optimality of $\bar u_\beta$ in ($\mathrm{P}_{r,\bar u,\beta}$), the inequality $\bar u_\beta > -\varepsilon_P$ a.e. in $(-r,r)$, the nonnegativity of $\bar u$, the definition of F_r, and the properties of η in Lemma 5.2, we obtain that

$$
\begin{aligned}
F_r(\bar u) &= F_r(\bar u) + \frac{C_{\bar u}}{2}\|\bar u - \bar u\|^2_{L^2(-r,r)} + \int_{-r}^{r} \beta\eta(\bar u)\,\mathrm{d}t\\
&\geq F_r(\bar u_\beta) + \frac{C_{\bar u}}{2}\|\bar u_\beta - \bar u\|^2_{L^2(-r,r)} + \int_{-r}^{r} \beta\eta(\bar u_\beta)\,\mathrm{d}t\\
&= \frac{1}{2}\|S_r(\bar u_\beta) - y_D\|^2_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 \bar u_\beta + \frac{\nu_2}{2}(\bar u_\beta - u_D)^2\,\mathrm{d}t + \frac{C_{\bar u}}{2}\|\bar u_\beta - \bar u\|^2_{L^2(-r,r)} + \int_{-r}^{r} \beta\eta(\bar u_\beta)\,\mathrm{d}t\\
&\geq -2r\nu_1\varepsilon_P + \frac{C_{\bar u}}{2}\|\bar u_\beta - \bar u\|^2_{L^2(-r,r)} + \frac{\beta}{\varepsilon_P}\int_{-r}^{r} \min(0,\bar u_\beta)^2\,\mathrm{d}t \qquad \forall \beta > 0.
\end{aligned}
$$

As $F_r(\bar u)$ is finite (*cf.* the comments after Definition 4.1), the above shows that the family $\{\bar u_\beta\}$ is bounded in $L^2(-r,r)$ and that the functions $\min(0,\bar u_\beta)$ converge strongly to zero in $L^2(-r,r)$ for $\beta\to\infty$. This implies in particular that, by passing over to a sequence $\{\beta_n\} \subset (0,\infty)$ with $\beta_n \to \infty$ for $n\to\infty$, we can achieve that $\bar u_{\beta_n} \rightharpoonup \hat u$ holds in $L^2(-r,r)$ for some $\hat u \in L^2(-r,r)$ as n tends to infinity. Due to the convergence $\min(0,\bar u_\beta) \to 0$ in $L^2(-r,r)$, we know that $\hat u$ is nonnegative a.e. in $(-r,r)$ and, thus, admissible for ($\mathrm{P}_{r,\bar u}$). Since $\bar u$ solves ($\mathrm{P}_{r,\bar u}$), since S is completely continuous by Theorem 3.4i), since convexity and lower semicontinuity imply weak lower semicontinuity, and again by the global optimality of $\bar u_\beta$ in ($\mathrm{P}_{r,\bar u,\beta}$), it now follows that

$$
\begin{aligned}
F_r(\hat u) + \frac{C_{\bar u}}{2}\|\hat u - \bar u\|^2_{L^2(-r,r)} &\geq F_r(\bar u) + \frac{C_{\bar u}}{2}\|\bar u - \bar u\|^2_{L^2(-r,r)}\\
&\geq \limsup_{n\to\infty}\left(F_r(\bar u_{\beta_n}) + \frac{C_{\bar u}}{2}\|\bar u_{\beta_n} - \bar u\|^2_{L^2(-r,r)} + \int_{-r}^{r}\beta_n\eta(\bar u_{\beta_n})\,\mathrm{d}t\right)\\
&\geq F_r(\hat u) + \frac{C_{\bar u}}{2}\limsup_{n\to\infty}\left(\|\bar u_{\beta_n} - \bar u\|^2_{L^2(-r,r)}\right) \geq F_r(\hat u) + \frac{C_{\bar u}}{2}\|\hat u - \bar u\|^2_{L^2(-r,r)}.
\end{aligned}
$$

The above can only be true if all appearing inequalities are identities. Since $\bar u$ is the unique global solution of ($\mathrm{P}_{r,\bar u}$), this implies, on the one hand, that $\bar u = \hat u$ holds and, on the other hand, that

$$
\limsup_{n\to\infty}\left(\|\bar u_{\beta_n} - \bar u\|^2_{L^2(-r,r)}\right) = \|\hat u - \bar u\|^2_{L^2(-r,r)} = 0.
$$

The sequence $\{\bar u_{\beta_n}\}$ thus converges strongly in $L^2(-r,r)$ to $\bar u$. Since the above arguments can be repeated for all sequences $\{\bar u_{\beta_n}\}$ satisfying $\bar u_{\beta_n} \rightharpoonup \hat u$ for some $\hat u \in L^2(-r,r)$ and $(0,\infty) \ni \beta_n \to \infty$ for $n\to\infty$, the convergence of the whole family $\{\bar u_\beta\}$ to $\bar u$ for $\beta\to\infty$ now follows from a trivial contradiction argument; *cf.* [26], Lemma 4.16. This completes the proof. ☐

By exploiting the estimate in Theorem 3.4iii), we can prove the following for the relaxed problems ($\mathrm{P}_{r,\bar u,\beta}$):

Proposition 5.5 (L^∞_{BV}-Regularity of Adjoint Variables)**.** *Let $\beta > 0$ be given and suppose that $\bar u_\beta$ is a local solution of* ($\mathrm{P}_{r,\bar u,\beta}$)*. Denote with $\bar p_\beta$ the element of $L^0(-r,r)$ defined by*

$$
\bar p_\beta := \nu_1 + \nu_2(\bar u_\beta - u_D) + C_{\bar u}(\bar u_\beta - \bar u) + \beta\eta'(\bar u_\beta).
$$

Then it holds $\bar p_\beta \in L^\infty_{BV}(-r,r)$ and $\bar p_\beta$ admits a $BV[-r,r]$-representative that satisfies

$$\|\bar p_\beta\|_{BV[-r,r]} \leq K \tag{5.1}$$

for a constant $K > 0$ which depends only on r, y_D, Ω, and the numbers $M, c > 0$ in Theorem 3.4.

Proof. Consider some $z \in L^\infty(-r,r)$ and define $z_n := \mathbb{1}_{\{\bar u_\beta \geq -\varepsilon_P + 1/n\}} z$, $n \in \mathbb{N}$. Then it holds

$$\bar u_\beta - s z_n \begin{cases} = \bar u_\beta > -\varepsilon_P & \text{a.e. in } \{\bar u_\beta < -\varepsilon_P + 1/n\}, \\ \geq -\varepsilon_P + 1/(2n) & \text{a.e. in } \{\bar u_\beta \geq -\varepsilon_P + 1/n\}, \end{cases} \qquad \forall\, 0 < s < (2n\|z\|_{L^\infty(-r,r)} + 1)^{-1}, \tag{5.2}$$

and it follows from the local optimality of $\bar u_\beta$ in ($\mathrm{P}_{r,\bar u,\beta}$) and the definition of F_r that $\bar u_\beta$ satisfies

$$\begin{aligned} &\frac{1}{2}\|S_r(\bar u_\beta) - y_D\|^2_{L^2(\Omega)} + \int_{-r}^r \nu_1 \bar u_\beta + \frac{\nu_2}{2}(\bar u_\beta - u_D)^2 + \frac{C_{\bar u}}{2}(\bar u_\beta - \bar u)^2 + \beta\eta(\bar u_\beta)\,\mathrm{d}t \\ &\qquad \leq \frac{1}{2}\|S_r(\bar u_\beta - s z_n) - y_D\|^2_{L^2(\Omega)} + \int_{-r}^r \nu_1(\bar u_\beta - s z_n) + \frac{\nu_2}{2}(\bar u_\beta - s z_n - u_D)^2 + \frac{C_{\bar u}}{2}(\bar u_\beta - s z_n - \bar u)^2 \\ &\qquad\qquad + \beta\eta(\bar u_\beta - s z_n)\,\mathrm{d}t \end{aligned}$$

for all small enough $s > 0$. If we use the binomial identities in the above, divide by s, and rearrange, then we obtain that

$$\begin{aligned} &\int_{\{\bar u_\beta \geq -\varepsilon_P + 1/n\}} \nu_1 z + \nu_2(\bar u_\beta - u_D) z - \frac{\nu_2}{2} s z^2 + C_{\bar u}(\bar u_\beta - \bar u) z - \frac{C_{\bar u}}{2} s z^2 + \beta \frac{\eta(\bar u_\beta) - \eta(\bar u_\beta - s z)}{s}\,\mathrm{d}t \\ &\qquad \leq \int_\Omega (S_r(\bar u_\beta) - y_D)\frac{S_r(\bar u_\beta - s z_n) - S_r(\bar u_\beta)}{s}\,\mathrm{d}x + \frac{s}{2}\int_\Omega \left(\frac{S_r(\bar u_\beta - s z_n) - S_r(\bar u_\beta)}{s}\right)^2 \mathrm{d}x. \end{aligned} \tag{5.3}$$

Due to Theorem 3.4iii), the linearity of g_u in u, the inequality of Cauchy-Schwarz, and the definition of S_r, we know that there exists a constant $c > 0$ independent of β and s such that

$$\|S_r(\bar u_\beta - s z_n) - S_r(\bar u_\beta)\|_{L^2(\Omega)} \leq \frac{s}{c}\|g_{E_r(z_n)}(S_r(\bar u_\beta))\|_{L^2(\Omega)} \qquad \forall\, 0 < s < (2n\|z\|_{L^\infty(-r,r)} + 1)^{-1}.$$

In combination with the properties of η, the second inequality in (5.2), the dominated convergence theorem, and (2.3), the above allows us to bound the right-hand side of (5.3) and to subsequently pass to the limit $s \to 0$. This yields

$$\int_{\{\bar u_\beta \geq -\varepsilon_P + 1/n\}} \nu_1 z + \nu_2(\bar u_\beta - u_D) z + C_{\bar u}(\bar u_\beta - \bar u) z + \beta\eta'(\bar u_\beta) z\,\mathrm{d}t \leq \frac{1}{c}\|S_r(\bar u_\beta) - y_D\|_{L^2(\Omega)}\|g_{E_r(z_n)}(S_r(\bar u_\beta))\|_{L^2(\Omega)}. \tag{5.4}$$

Since $\|S_r(\bar u_\beta) - y_D\|_{L^2(\Omega)} \leq M + \|y_D\|_{L^2(\Omega)}$ holds for the constant $M > 0$ in (3.8) and since the function $\nu_1 + \nu_2(\bar u_\beta - u_D) + C_{\bar u}(\bar u_\beta - \bar u) + \beta\eta'(\bar u_\beta) \in L^0(-r,r)$ appearing on the left-hand side of (5.4) is equal to $\bar p_\beta$ by definition, we may now conclude that, for all $z \in L^\infty(-r,r)$ and all $n \in \mathbb{N}$, we have

$$\int_{\{\bar u_\beta \geq -\varepsilon_P + 1/n\}} \bar p_\beta z\,\mathrm{d}t \leq \frac{M + \|y_D\|_{L^2(\Omega)}}{c}\left(\int_\Omega \left(\int_0^{S_r(\bar u_\beta)(x)} (E_r z_n)(t)\,\mathrm{d}t\right)^2 \mathrm{d}x\right)^{1/2}. \tag{5.5}$$

Recall that, due to Theorem 3.4ii) and the embedding $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$, there exists a constant $R = R(M, \Omega) > 0$ independent of β satisfying $\|S_r(\bar{u}_\beta)\|_\infty \le R$. If we use this bound in (5.5) and consider the special situation $z := \operatorname{sgn}(\bar{p}_\beta) \in L^\infty(-r, r)$, then it follows that

$$\int_{-r}^{r} |\bar{p}_\beta| \mathbb{1}_{\{\bar{u}_\beta \ge -\varepsilon_P + 1/n\}} \,\mathrm{d}t \le R \left(\frac{M + \|y_D\|_{L^2(\Omega)}}{c} \right) \lambda^d(\Omega)^{1/2} \qquad \forall n \in \mathbb{N}. \tag{5.6}$$

Due to the lemma of Fatou and since $\bar{u}_\beta > -\varepsilon_P$ holds a.e. in $(-r, r)$, we may let n go to ∞ in (5.6). This yields that $\|\bar{p}_\beta\|_{L^1(-r,r)} \le K_1$ holds for the constant

$$K_1 = K_1(c, M, \Omega, y_D) := R \left(\frac{M + \|y_D\|_{L^2(\Omega)}}{c} \right) \lambda^d(\Omega)^{1/2}.$$

By exploiting this L^1-regularity of $\bar{p}_\beta$, the dominated convergence theorem, and again the fact that $\bar{u}_\beta > -\varepsilon_P$ holds a.e. in $(-r, r)$, we can pass to the limit $n \to \infty$ in (5.5) for arbitrary $z \in L^\infty(-r, r)$ to obtain

$$\int_{-r}^{r} \bar{p}_\beta z \,\mathrm{d}t \le \frac{M + \|y_D\|_{L^2(\Omega)}}{c} \left(\int_\Omega \left(\int_0^{S_r(\bar{u}_\beta)(x)} (E_r z)(t) \,\mathrm{d}t \right)^2 \mathrm{d}x \right)^{1/2} \qquad \forall z \in L^\infty(-r, r). \tag{5.7}$$

Let us now consider functions z of the form $z = v'$, $v \in C^1([-r, r])$. For such z, it follows from the fundamental theorem of calculus, the definition of E_r, and a trivial distinction of cases that

$$\left| \int_0^{S_r(\bar{u}_\beta)(x)} (E_r v')(t) \,\mathrm{d}t \right| \le \begin{cases} |v\,(S_r(\bar{u}_\beta)(x))| + |v\,(0)| & \text{if } S_r(\bar{u}_\beta)(x) \in [-r, r], \\ |v\,(r)| + |v\,(0)| & \text{if } S_r(\bar{u}_\beta)(x) > r, \\ |v\,(-r)| + |v\,(0)| & \text{if } S_r(\bar{u}_\beta)(x) < -r, \end{cases} \qquad \text{f.a.a. } x \in \Omega.$$

In combination with (5.7), the above implies

$$\begin{aligned} \int_{-r}^{r} \bar{p}_\beta v' \,\mathrm{d}t &\le \frac{M + \|y_D\|_{L^2(\Omega)}}{c} \left(\int_\Omega \left(\int_0^{S_r(\bar{u}_\beta)(x)} (E_r v')(t) \,\mathrm{d}t \right)^2 \mathrm{d}x \right)^{1/2} \\ &\le \frac{M + \|y_D\|_{L^2(\Omega)}}{c} \left(\int_\Omega \left(2\|v\|_{C([-r,r])} \right)^2 \mathrm{d}x \right)^{1/2} \\ &= K_2 \|v\|_{C([-r,r])} \qquad \forall v \in C^1([-r, r]), \end{aligned} \tag{5.8}$$

where $K_2 = K_2(c, M, \Omega, y_D)$ is the constant defined by

$$K_2 := 2 \left(\frac{M + \|y_D\|_{L^2(\Omega)}}{c} \right) \lambda^d(\Omega)^{1/2}.$$

Using the theorem of Hahn-Banach (see [30], Thm. 8.1.1), we may deduce from (5.8) that the linear map $C^1([-r, r]) \ni v \mapsto \int_{-r}^{r} \bar{p}_\beta v' \,\mathrm{d}t \in \mathbb{R}$ can be extended to a linear and continuous functional on $C([-r, r])$. In combination with the Riesz representation theorem (see [30], Thm. 8.1.2), this yields that there exists a function $q \in BV[-r, r]$ which satisfies $q(-r) = 0$, $\operatorname{var}(q; [-r, r]) \le K_2$, and

$$\int_{-r}^{r} \bar{p}_\beta v' \,\mathrm{d}t = \int_{-r}^{r} v \,\mathrm{d}q \qquad \forall v \in C^1([-r, r]). \tag{5.9}$$

Here, the integral on the right has to be understood in the sense of Riemann-Stieltjes. If we choose test functions $v \in C_c^\infty(-r,r)$ in (5.9) and integrate by parts by means of [30], Theorem 5.5.1, then we obtain that

$$\int_{-r}^{r} (\bar{p}_\beta + q) v' \,\mathrm{d}t = 0 \qquad \forall v \in C_c^\infty(-r,r).$$

This shows that $\bar{p}_\beta + q = \text{const}$ holds a.e. in $(-r,r)$ and that $\bar{p}_\beta$ possesses a representative that satisfies $\bar{p}_\beta \in BV[-r,r]$ and $\operatorname{var}(\bar{p}_\beta; [-r,r]) = \operatorname{var}(q; [-r,r]) \leq K_2$. Henceforth, with the symbol $\bar{p}_\beta$, we always refer to this representative. Note that, due to the bound $\|\bar{p}_\beta\|_{L^1(-r,r)} \leq K_1$, we know that there exists at least one $s \in [-r,r]$ with $|\bar{p}_\beta(s)| \leq K_1/(2r)$. In combination with the definition of the variation $\operatorname{var}(\bar{p}_\beta; [-r,r])$ and our previous estimates, this allows us to conclude that

$$\|\bar{p}_\beta\|_{BV[-r,r]} = \left|\bar{p}_\beta(-r)\right| + \operatorname{var}(\bar{p}_\beta; [-r,r]) \leq |\bar{p}_\beta(s)| + 2\operatorname{var}(\bar{p}_\beta; [-r,r]) \leq \frac{K_1}{2r} + 2K_2.$$

If we define $K := \frac{K_1}{2r} + 2K_2$, then the last estimate takes precisely the form (5.1). This completes the proof of the proposition. □

We are now in the position to prove the main result of this section:

Theorem 5.6 (L^∞_G- and L^∞_{BV}-Regularity of $\bar{u}$)**.** *It holds $\bar{u} \in L^\infty_G(-r,r)$. If u_D is an element of $L^\infty_{BV}(-r,r)$, then we even have $\bar{u} \in L^\infty_{BV}(-r,r)$.*

Proof. Consider some globally optimal controls $\bar{u}_\beta$ of ($\mathrm{P}_{r,\bar{u},\beta}$). From Proposition 5.4, we know that $\bar{u}_\beta \to \bar{u}$ holds in $L^2(-r,r)$ for $\beta \to \infty$, and from Proposition 5.5, that there exist representatives of $\bar{u}$, $\bar{u}_\beta$, and u_D such that the functions $\bar{p}_\beta := \nu_1 + \nu_2(\bar{u}_\beta - u_D) + C_{\bar{u}}(\bar{u}_\beta - \bar{u}) + \beta\eta'(\bar{u}_\beta)$ satisfy $\|\bar{p}_\beta\|_{BV[-r,r]} \leq K$ for a constant K that is independent of β. In combination with the Helly selection theorem (see [30], Thm. 2.7.4) and classical results on L^2-convergence, this allows us to find a sequence $\{\beta_n\} \subset (0,\infty)$ satisfying $\beta_n \to \infty$ for $n \to \infty$ and a function $\bar{p} \in BV[-r,r]$ such that $\bar{u}_{\beta_n} \to \bar{u}$ holds pointwise a.e. in $(-r,r)$ and such that $\bar{p}_{\beta_n}(t) \to \bar{p}(t)$ holds for all $t \in [-r,r]$ as n tends to infinity. Due to the definition of $\bar{p}_{\beta_n}$, the fact that $\eta'(s) = \eta(s) = 0$ holds for all $s \geq 0$, and the fact that $\eta'(s) < 0$ holds for all $s \in (-\varepsilon_P, 0)$, the convergence properties of $\bar{u}_{\beta_n}$ and $\bar{p}_{\beta_n}$ imply that there exists a function $\zeta \in L^2(-r,r)$ such that $\bar{p}_{\beta_n} - \nu_1 - \nu_2(\bar{u}_{\beta_n} - u_D) - C_{\bar{u}}(\bar{u}_{\beta_n} - \bar{u}) = \beta_n \eta'(\bar{u}_{\beta_n}) \to \zeta$ holds pointwise a.e. in $(-r,r)$ for $n \to \infty$ with $\zeta \leq 0$ a.e. in $(-r,r)$ and $\zeta = 0$ a.e. in $\{\bar{u} > 0\}$. Passing to the limit now yields the identity $\bar{p} - \nu_1 + \nu_2 u_D = \nu_2 \bar{u} + \zeta$ a.e. in $(-r,r)$ and, after taking the positive part on the left and the right of this equality – keeping in mind the complementarity between $\bar{u}$ and ζ – that $\bar{u} = \max(0, \bar{p} - \nu_1 + \nu_2 u_D)/\nu_2$. Since $\bar{p} \in BV[-r,r] \subset G[-r,r]$ holds and since the function $\mathbb{R} \ni s \mapsto \max(0,s) \in \mathbb{R}$ is globally Lipschitz continuous, the asserted regularity properties of $\bar{u}$ now follow immediately; see [30], Definition 4.1.1 and [37], Theorem 4. □

Next, we state the consequences that Theorem 5.6 has for our original problem (P). For the ease of reference, we formulate this result such that it is independent of Assumption 5.1.

Theorem 5.7 (L^∞_G- and L^∞_{BV}-Regularity for (P))**.** *Suppose that $r \geq r_P$ is given and that $u_D|_{(-r,r)} \in L^\infty_G(-r,r)$ (resp., $u_D|_{(-r,r)} \in L^\infty_{BV}(-r,r)$) holds. Then every local solution $\bar{u} \in L^2(\mathbb{R})$ of (P) satisfies*

$$\bar{u} = \max\left(0, u_D - \frac{\nu_1}{\nu_2}\right) \text{ a.e. in } \mathbb{R} \setminus (-r,r) \qquad \text{and} \qquad \bar{u}|_{(-r,r)} \in L^\infty_G(-r,r) \text{ (resp., } \bar{u}|_{(-r,r)} \in L^\infty_{BV}(-r,r)\text{)}.$$

Proof. This follows immediately from Theorem 4.5 and Theorem 5.6. □

As we will make frequent use of Theorems 5.6, 5.7 in Sections 7 and 8, we conclude this section by introducing a shorthand notation for the regularity assumption on u_D needed in these results:

Definition 5.8 (Condition (A1))**.** Given $r > 0$, we say that condition (A1) holds if $u_D|_{(-r,r)} \in L^\infty_G(-r,r)$.

6. Differentiability properties of the control-to-state map

In this section, we show that the operator $S_r \colon U_{P,r} \to H^1_0(\Omega) \cap H^2(\Omega)$, $u \mapsto y$, introduced in Definition 4.3 is Hadamard directionally differentiable at all points $u \in L^\infty_G(-r,r) \cap U_{P,r}$ in all directions $z \in \mathcal{T}(u; U_{P,r})$. Note that, due to the identity (4.2), the differentiability results that we establish for S_r in the following carry over immediately to the solution map $S \colon u \mapsto y$ of (3.7); see Corollary 6.10 below. We focus on S_r in our analysis since this is the operator that appears in the gradient projection algorithm of Section 8. To clarify which notions of differentiability we are concerned with, we recall (see [25], Sect. 2.2.1, [38], Def. 1.1.2):

Definition 6.1 (Notions of Differentiability)**.** Let U and Y be normed spaces. Assume that a convex, closed, nonempty set $D \subset U$ and a function $H \colon D \to Y$ are given. Then H is called:

i) directionally differentiable at $u \in D$ in direction $z \in \mathcal{R}(u; D)$ if there exists $H'(u; z) \in Y$ (the directional derivative) such that, for all $\{s_n\} \subset (0, \infty)$ satisfying $s_n \to 0$ and $u + s_n z \in D$ for all n, we have

$$\lim_{n\to\infty} \frac{H(u + s_n z) - H(u)}{s_n} = H'(u; z);$$

ii) Hadamard directionally differentiable at $u \in D$ in direction $z \in \mathcal{T}(u; D)$ if there exists $H'(u; z) \in Y$ (the Hadamard directional derivative) such that, for all $\{s_n\} \subset (0, \infty)$ and $\{z_n\} \subset U$ satisfying $s_n \to 0$, $z_n \to z$, and $u + s_n z_n \in D$ for all n, we have

$$\lim_{n\to\infty} \frac{H(u + s_n z_n) - H(u)}{s_n} = H'(u; z);$$

iii) Hadamard-Gâteaux differentiable at $u \in D$ if H is Hadamard directionally differentiable in all directions $z \in \mathcal{T}(u; D)$ at u and if the derivative $H'(u; \cdot)$ can be extended to a function $H'(u; \cdot) =: H'(u) \in \mathcal{L}(U, Y)$;

iv) Fréchet differentiable at $u \in D$ if there exists a map $H'(u) \in \mathcal{L}(U, Y)$ satisfying

$$\lim_{0 < \|z\|_U \to 0,\, u+z\in D} \frac{\|H(u + z) - H(u) - H'(u)z\|_Y}{\|z\|_U} = 0.$$

If one of the above conditions holds for all $u \in D$ and all relevant $z \in U$, then we drop the reference to u and z and say that H possesses the respective differentiability property on D. If H is Fréchet differentiable on D and $D \ni u \mapsto H'(u) \in \mathcal{L}(U, Y)$ is a continuous function, then we call H continuously Fréchet differentiable on D.

For the discussion of the differentiability properties of the operator S_r, we require some preliminary results. We begin by introducing (*cf.* Def. 4.3ii)):

Definition 6.2 (Extension by Zero in $G[-r,r]$)**.** Given $r > 0$ and $u \in G[-r,r]$, we define $\mathcal{E}_r(u)$ to be the extension of u by zero, *i.e.*,

$$\mathcal{E}_r(u) \colon \mathbb{R} \to \mathbb{R}, \qquad \mathcal{E}_r(u)(t) := \begin{cases} u(t) & \text{if } t \in [-r, r], \\ 0 & \text{if } t \in \mathbb{R} \setminus [-r, r]. \end{cases} \tag{6.1}$$

The next lemma analyzes how the extensions of two different $G[-r,r]$-representatives of the same element of $L^\infty_G(-r,r)$ are related to each other.

Lemma 6.3 (Extension of Regulated Representatives)**.** *Let $r > 0$ be fixed and suppose that $u_1, u_2 \in G[-r,r]$ are given such that $u_1(t) = u_2(t)$ holds for a.a. $t \in (-r,r)$. Then the following is true:*

i) It holds $\mathcal{E}_r(u_1)(t-) = \mathcal{E}_r(u_2)(t-)$ *and* $\mathcal{E}_r(u_1)(t+) = \mathcal{E}_r(u_2)(t+)$ *for all* $t \in \mathbb{R}$.

ii) There exists an at most countable set $Z \subset \mathbb{R}$ *such that* $\mathcal{E}_r(u_1)(t) = \mathcal{E}_r(u_1)(t\pm) = \mathcal{E}_r(u_2)(t\pm) = \mathcal{E}_r(u_2)(t)$ *holds for all* $t \in \mathbb{R} \setminus Z$.

Here, with $\mathcal{E}_r(u_i)(t\pm)$, $i = 1, 2$, *we denote the left/right limit of the function* $\mathcal{E}_r(u_i) \colon \mathbb{R} \to \mathbb{R}$ *at* $t \in \mathbb{R}$*; cf.* (2.1).

Proof. Suppose that functions $u_1, u_2 \in G[-r, r]$ as in the lemma are given. Then it holds $\mathcal{E}_r(u_1)(t) = \mathcal{E}_r(u_2)(t)$ for a.a. $t \in \mathbb{R}$ by the definition of $\mathcal{E}_r$ and we obtain from (6.1) and [30], Definition 4.1.1 that the functions $\mathcal{E}_r(u_1)$ and $\mathcal{E}_r(u_2)$ have well-defined left and right limits at all points $t \in \mathbb{R}$. By combining these properties with the fact that the complement of a null set is dense, we arrive at the identities $\mathcal{E}_r(u_1)(t-) = \mathcal{E}_r(u_2)(t-)$ and $\mathcal{E}_r(u_1)(t+) = \mathcal{E}_r(u_2)(t+)$ for all $t \in \mathbb{R}$ as desired; *cf.* (2.1). This proves i). To establish ii), we note that (6.1) and [30], Theorem 4.1.8 entail that there exists an at most countable set $Z \subset \mathbb{R}$ such that $\mathcal{E}_r(u_1)$ and $\mathcal{E}_r(u_2)$ satisfy $\mathcal{E}_r(u_1)(t-) = \mathcal{E}_r(u_1)(t) = \mathcal{E}_r(u_1)(t+)$ and $\mathcal{E}_r(u_2)(t-) = \mathcal{E}_r(u_2)(t) = \mathcal{E}_r(u_2)(t+)$ for all $t \in \mathbb{R} \setminus Z$. In tandem with i), this yields ii) and completes the proof. ☐

Note that Lemma 6.3 implies that all $G[-r, r]$-representatives of a control $u \in L^\infty_G(-r, r)$ yield the same left and right limits after an extension by zero. This shows that it makes sense to talk about the functions $\mathcal{E}_r(u)(\cdot-) \colon \mathbb{R} \to \mathbb{R}$ and $\mathcal{E}_r(u)(\cdot+) \colon \mathbb{R} \to \mathbb{R}$ of a control $u \in L^\infty_G(-r, r)$. In what follows, we call $\mathcal{E}_r(u)(\cdot-)$ and $\mathcal{E}_r(u)(\cdot+)$ the left- and the right-continuous representative of the $L^2(\mathbb{R})$-extension $E_r(u)$, respectively.

Definition 6.4 (Left- and Right-Continuous Representative of $E_r(u)$)**.** Given $r > 0$ and $u \in L^\infty_G(-r, r)$, we define $\mathcal{E}_r(u)(\cdot-) \colon \mathbb{R} \to \mathbb{R}$ and $\mathcal{E}_r(u)(\cdot+) \colon \mathbb{R} \to \mathbb{R}$ to be the functions that are obtained by extending an arbitrary $G[-r, r]$-representative of u by zero in the sense of (6.1) and by subsequently taking left/right limits.

For the functions $\mathcal{E}_r(u)(\cdot-) \colon \mathbb{R} \to \mathbb{R}$ and $\mathcal{E}_r(u)(\cdot+) \colon \mathbb{R} \to \mathbb{R}$, we have:

Lemma 6.5 (Stability of Left- and Right-Continuous Representatives)**.** *Let* $r > 0$ *be given. Then it holds*

$$\sup_{t \in \mathbb{R}} |\mathcal{E}_r(u_1)(t\pm) - \mathcal{E}_r(u_2)(t\pm)| = \|u_1 - u_2\|_{L^\infty(-r,r)} \quad \forall u_1, u_2 \in L^\infty_G(-r, r). \tag{6.2}$$

Proof. The inequality "$\geq$" in (6.2) follows straightforwardly from (6.1), Lemma 6.3, and the definition of the norm $\|\cdot\|_{L^\infty(-r,r)}$. To establish "$\leq$", one can use the exact same arguments as in the proof of Lemma 6.3. ☐

We remark that (6.2) does not hold for arbitrary representatives of $E_r(u_1), E_r(u_2) \in L^2(\mathbb{R})$, even if they are regulated. Indeed, by modifying the function $\mathcal{E}_r(u_1)(\cdot\pm) - \mathcal{E}_r(u_2)(\cdot\pm)$ at a single point, one can make the supremum-norm on the left of (6.2) arbitrarily large while keeping the $L^\infty(-r, r)$-norm on the right of (6.2) constant. As a straightforward consequence of Lemma 6.5, we obtain:

Corollary 6.6 (Closedness of $L^\infty_G(-r, r)$)**.** *Let* $r > 0$*. The set* $L^\infty_G(-r, r)$ *is a closed subspace of* $L^\infty(-r, r)$*. In particular,* $(L^\infty_G(-r, r), \|\cdot\|_{L^\infty(-r,r)})$ *is a Banach space.*

Proof. That $L^\infty_G(-r, r)$ is a subspace is trivial. To see that $L^\infty_G(-r, r)$ is closed, suppose that $\{u_n\} \subset L^\infty_G(-r, r)$ is a sequence which satisfies $u_n \to u$ in $L^\infty(-r, r)$ for some $u \in L^\infty(-r, r)$. From (6.1) and (6.2), it follows that

$$\sup_{t \in [-r,r]} |\mathcal{E}_r(u_n)(t+) - \mathcal{E}_r(u_m)(t+)| = \|u_n - u_m\|_{L^\infty(-r,r)} \tag{6.3}$$

holds for all $m, n \in \mathbb{N}$. Since $\{u_n\}$ converges in $L^\infty(-r, r)$ and since the functions $\mathcal{E}_r(u_n)(\cdot+)|_{[-r,r]}$ are regulated by [30], Corollary 4.1.9, (6.3) shows that the sequence $\{\mathcal{E}_r(u_n)(\cdot+)|_{[-r,r]}\}$ is Cauchy in $(G[-r, r], \|\cdot\|_\infty)$. As $(G[-r, r], \|\cdot\|_\infty)$ is a Banach space, this implies that there exists $\hat{u} \in G[-r, r]$ such that $\mathcal{E}_r(u_n)(t+) \to \hat{u}(t)$ holds for all $t \in [-r, r]$. From Lemma 6.3, we obtain further that the functions $\mathcal{E}_r(u_n)(\cdot+)|_{[-r,r]}$ are representatives of u_n and, thus, that $\mathcal{E}_r(u_n)(t+) \to u(t)$ holds for a.a. $t \in (-r, r)$. In summary, this yields $u = \hat{u}$ a.e. in $(-r, r)$. The limit u of $\{u_n\}$ thus admits a $G[-r, r]$-representative, it holds $u \in L^\infty_G(-r, r)$, and the claim is proven. ☐

We are now in the position to study the differentiability properties of the operator S_r. The point of departure for our analysis is the following observation.

Lemma 6.7 (Hadamard Directional Differentiability of $g_{E_r(u)}$)**.** *Let $r > 0$ be fixed and let $u \in L^\infty_G(-r, r)$ be given. Then the function $g_{E_r(u)} \colon \mathbb{R} \to \mathbb{R}$ is Hadamard directionally differentiable on $\mathbb{R}$ with derivative*

$$g'_{E_r(u)}(t; z) = \begin{cases} \mathcal{E}_r(u)(t-)z & \text{if } z < 0, \\ \mathcal{E}_r(u)(t+)z & \text{if } z \geq 0, \end{cases} \qquad \forall z \in \mathbb{R} \qquad \forall t \in \mathbb{R}. \tag{6.4}$$

Proof. Choose an arbitrary $G[-r, r]$-representative of $u \in L^\infty_G(-r, r)$ (denoted by the same symbol) and suppose that $t \in \mathbb{R}$ and $z \geq 0$ are given. Then (2.3), the definition of the right limit $\mathcal{E}_r(u)(t+)$ of the function $\mathcal{E}_r(u) \colon \mathbb{R} \to \mathbb{R}$ at t, and the fact that $\mathcal{E}_r(u)$ is a representative of $E_r(u) \in L^2(\mathbb{R})$ imply that

$$\begin{aligned}
\limsup_{0<s\to 0} \left| \frac{g_{E_r(u)}(t + sz) - g_{E_r(u)}(t)}{s} - \mathcal{E}_r(u)(t+)z \right| &= \limsup_{0<s\to 0} \left| \frac{1}{s} \int_t^{t+sz} E_r(u)(\zeta) - \mathcal{E}_r(u)(t+) \,\mathrm{d}\zeta \right| \\
&= \limsup_{0<s\to 0} \left| \frac{1}{s} \int_t^{t+sz} \mathcal{E}_r(u)(\zeta) - \mathcal{E}_r(u)(t+) \,\mathrm{d}\zeta \right| \\
&\leq \limsup_{0<s\to 0} \sup_{\zeta\in(t,t+sz)} |\mathcal{E}_r(u)(\zeta) - \mathcal{E}_r(u)(t+)||z| = 0.
\end{aligned}$$

This shows that the function $g_{E_r(u)}$ is directionally differentiable in all nonnegative directions z at all $t \in \mathbb{R}$ with derivative $\mathcal{E}_r(u)(t+)z$. The directional differentiability in negative directions is obtained analogously. To see that $g_{E_r(u)}$ is even Hadamard directionally differentiable, it suffices to note that $g_{E_r(u)}$ is globally Lipschitz continuous due to the L^∞-regularity of u and to invoke [25], Proposition 2.49. This completes the proof. $\square$

In combination with Lemma 6.3, Lemma 6.7 yields:

Corollary 6.8 (Exceptional Set of $g_{E_r(u)}$)**.** *Let $r > 0$ be fixed and let $u \in L^\infty_G(-r, r)$ be given. Then there exists an at most countable set $Z \subset \mathbb{R}$ such that the function $g_{E_r(u)} \colon \mathbb{R} \to \mathbb{R}$ is Fréchet differentiable at all $t \in \mathbb{R} \setminus Z$.*

Proof. From Lemma 6.3ii) and Definition 6.4, we obtain that there exists an at most countable set $Z \subset \mathbb{R}$ such that $\mathcal{E}_r(u)(t-) = \mathcal{E}_r(u)(t+)$ holds for all $t \in \mathbb{R} \setminus Z$. In view of the properties in Lemma 6.7, this implies that $g_{E_r(u)}$ is Hadamard-Gâteaux differentiable at all $t \in \mathbb{R} \setminus Z$. As $g_{E_r(u)}$ is a map between finite-dimensional spaces, this implies the asserted Fréchet differentiability; see the comments after [25], Proposition 2.49. $\square$

Since the function S_r is, by definition, the solution operator $U_{P,r} \ni u \mapsto y \in H^1_0(\Omega) \cap H^2(\Omega)$ of the partial differential equation

$$-\Delta y + g_{E_r(u)}(y) = f \text{ in } \Omega, \qquad y = 0 \text{ on } \partial\Omega,$$

we can use the properties in Lemma 6.7 to prove:

Theorem 6.9 (Hadamard Directional Differentiability of S_r)**.** *Let $r > 0$ be given. The function $S_r \colon U_{P,r} \to H^1_0(\Omega) \cap H^2(\Omega)$ is Hadamard directionally differentiable at all points $u \in L^\infty_G(-r, r) \cap U_{P,r}$ in all directions $z \in \mathcal{T}(u; U_{P,r}) \subset L^2(-r, r)$ in the following sense: For all $u \in L^\infty_G(-r, r) \cap U_{P,r}$ and all $z \in \mathcal{T}(u; U_{P,r})$, there exists a unique $S'_r(u; z) \in H^1_0(\Omega) \cap H^2(\Omega)$ such that, for all $\{z_n\} \subset L^2(-r, r)$ and $\{s_n\} \subset (0, \infty)$ satisfying $z_n \to z$ in $L^2(-r, r)$, $s_n \to 0$ in $\mathbb{R}$, and $u + s_n z_n \in U_{P,r}$ for all $n \in \mathbb{N}$, it holds*

$$\frac{S_r(u + s_n z_n) - S_r(u)}{s_n} \to S'_r(u; z) \text{ in } H^2(\Omega) \text{ for } n \to \infty. \tag{6.5}$$

Further, the directional derivative $\delta := S_r'(u; z) \in H_0^1(\Omega) \cap H^2(\Omega)$ of S_r at a point $u \in L_G^\infty(-r,r) \cap U_{P,r}$ with state $y := S_r(u)$ in a direction $z \in \mathcal{T}(u; U_{P,r})$ is uniquely characterized by the condition that it is the weak solution of the partial differential equation

$$-\Delta\delta + \mathcal{E}_r(u)(y+)\max(0,\delta) + \mathcal{E}_r(u)(y-)\min(0,\delta) = -g_{E_r(z)}(y) \text{ in } \Omega, \qquad \delta = 0 \text{ on } \partial\Omega. \tag{6.6}$$

Here, $\mathcal{E}_r(u)(\,\cdot\,-)$ and $\mathcal{E}_r(u)(\,\cdot\,+)$ denote the left- and the right-continuous representative of $E_r(u) \in L^2(\mathbb{R})$.

Proof. Suppose that $u \in L_G^\infty(-r,r) \cap U_{P,r}$ and $z \in \mathcal{T}(u; U_{P,r})$ are given; define $y := S_r(u) \in H_0^1(\Omega) \cap H^2(\Omega)$; assume that $\{z_n\} \subset L^2(-r,r)$ and $\{s_n\} \subset (0,\infty)$ are sequences that satisfy $z_n \to z$ in $L^2(-r,r)$, $s_n \to 0$ in $\mathbb{R}$, and $u + s_n z_n \in U_{P,r}$ for all n; and denote the difference quotients on the left-hand side of (6.5) with δ_n. Then it follows from the definition of S_r and Theorem 3.4iv) that the sequence $\{\delta_n\} \subset H_0^1(\Omega) \cap H^2(\Omega)$ is bounded in $H_0^1(\Omega)$ and we obtain by subtracting the PDEs satisfied by $S_r(u + s_n z_n)$ and $S_r(u)$ and by rearranging the resulting equation that $\delta_n \in H_0^1(\Omega) \cap H^2(\Omega)$ is the weak solution of

$$-\Delta\delta_n + \frac{g_{E_r(u)}(y + s_n\delta_n) - g_{E_r(u)}(y)}{s_n} + g_{E_r(z_n)}(y + s_n\delta_n) = 0 \text{ in } \Omega, \qquad \delta_n = 0 \text{ on } \partial\Omega. \tag{6.7}$$

Because of the boundedness of $\{\delta_n\}$ in $H_0^1(\Omega)$ and the theorems of Banach-Alaoglu and Rellich-Kondrachov, we can pass over to a subsequence of $\{\delta_n\}$ (for simplicity denoted by the same symbol) which converges weakly in $H_0^1(\Omega)$ and strongly in $L^2(\Omega)$ to some $\delta \in H_0^1(\Omega)$. In combination with the fact that $E_r(u)$ is essentially bounded, the estimate (3.1), and the dominated convergence theorem, this yields

$$\begin{aligned}
&\left\| \frac{g_{E_r(u)}(y + s_n\delta_n) - g_{E_r(u)}(y)}{s_n} - g'_{E_r(u)}(y;\delta) \right\|_{L^2(\Omega)} \\
&\quad\le \left\| \frac{g_{E_r(u)}(y + s_n\delta_n) - g_{E_r(u)}(y + s_n\delta)}{s_n} \right\|_{L^2(\Omega)} + \left\| \frac{g_{E_r(u)}(y + s_n\delta) - g_{E_r(u)}(y)}{s_n} - g'_{E_r(u)}(y;\delta) \right\|_{L^2(\Omega)} \\
&\quad\le \|E_r(u)\|_{L^\infty(\mathbb{R})}\|\delta_n - \delta\|_{L^2(\Omega)} + \left\| \frac{g_{E_r(u)}(y + s_n\delta) - g_{E_r(u)}(y)}{s_n} - g'_{E_r(u)}(y;\delta) \right\|_{L^2(\Omega)} \to 0
\end{aligned}$$

and $\left\|g_{E_r(z_n)}(y + s_n\delta_n) - g_{E_r(z)}(y)\right\|_{L^2(\Omega)} \le \|E_r(z_n)\|_{L^2(\mathbb{R})} s_n^{1/2} \|\delta_n\|_{L^1(\Omega)}^{1/2} + \left\|g_{E_r(z_n - z)}(y)\right\|_{L^2(\Omega)} \to 0$ for $n \to \infty$. Due to (6.7), the above implies, on the one hand, that δ is a weak solution of

$$-\Delta\delta + g'_{E_r(u)}(y;\delta) + g_{E_r(z)}(y) = 0 \text{ in } \Omega, \qquad \delta = 0 \text{ on } \partial\Omega, \tag{6.8}$$

and, on the other hand, after subtracting (6.7) and (6.8) and invoking the regularity estimate (3.4), that δ is an element of $H_0^1(\Omega) \cap H^2(\Omega)$ which satisfies

$$\begin{aligned}
\|\delta_n - \delta\|_{H^2(\Omega)} &\le C\| - \Delta\delta_n + \Delta\delta\|_{L^2(\Omega)} \\
&\le C \left\| \frac{g_{E_r(u)}(y + s_n\delta_n) - g_{E_r(u)}(y)}{s_n} - g'_{E_r(u)}(y;\delta) \right\|_{L^2(\Omega)} + C \left\|g_{E_r(z_n)}(y + s_n\delta_n) - g_{E_r(z)}(y)\right\|_{L^2(\Omega)} \\
&\to 0
\end{aligned}$$

for $n \to \infty$ with a constant $C > 0$. The sequence $\{\delta_n\}$ thus converges even strongly in $H^2(\Omega)$ to δ. Next, we show that (6.8) can have at most one weak solution. To see this, let us suppose that $\delta, \hat{\delta} \in H_0^1(\Omega)$ are two weak solutions of (6.8). Then, by using $\hat{\delta} - \delta$ as a test function in the weak formulation for $\hat{\delta}$ and $\delta - \hat{\delta}$ as a test function in the weak formulation for δ, by adding the resulting identities, by employing (6.4), and by using that the bound $u \geq -\varepsilon_P$ a.e. in $(-r, r)$, Definition 6.4, and the definition of the left and right limit imply that $\mathcal{E}_r(u)(t\pm) \geq -\varepsilon_P$ holds for all $t \in \mathbb{R}$, we obtain that

$$\begin{aligned}
0 &= \int_\Omega |\nabla(\delta - \hat{\delta})|^2 + \mathcal{E}_r(u)(y+)\big(\max(0,\delta) - \max(0,\hat{\delta})\big)(\delta - \hat{\delta}) + \mathcal{E}_r(u)(y-)\big(\min(0,\delta) - \min(0,\hat{\delta})\big)(\delta - \hat{\delta})\,\mathrm{d}x \\
&\geq \int_\Omega |\nabla(\delta - \hat{\delta})|^2 - \varepsilon_P\big(\max(0,\delta) - \max(0,\hat{\delta})\big)(\delta - \hat{\delta}) - \varepsilon_P\big(\min(0,\delta) - \min(0,\hat{\delta})\big)(\delta - \hat{\delta})\,\mathrm{d}x \\
&= \int_\Omega |\nabla(\delta - \hat{\delta})|^2 - \varepsilon_P(\delta - \hat{\delta})^2\,\mathrm{d}x \geq (c_P - \varepsilon_P)\|\delta - \hat{\delta}\|^2_{L^2(\Omega)}.
\end{aligned}$$

Since $c_P - \varepsilon_P > 0$ holds by Definition 3.2, the above shows that δ and $\hat{\delta}$ are identical and that (6.8) possesses precisely one weak solution, namely the limit $\delta \in H_0^1(\Omega) \cap H^2(\Omega)$ of the difference quotients δ_n. Note that the unique solvability of the partial differential equation (6.8) also implies that the limit of the subsequence chosen after (6.7) is independent of the choice of the subsequence. In combination with a trivial contradiction argument (see [26], Lem. 4.16), this yields that the whole original sequence of difference quotients $\{\delta_n\}$ converges in $H_0^1(\Omega) \cap H^2(\Omega)$ to the weak solution δ of (6.8). By combining all of the above, by setting $S_r'(u; z) := \delta$, and by invoking Lemma 6.7, the claim of the theorem follows immediately. □

Note that the extension operator $\mathcal{E}_r$ in (6.6) can be dropped if the bound r and the control $u \in L_G^\infty(-r, r) \cap U_{P,r}$ satisfy $u \geq 0$ a.e. in $(-r, r)$ and $r \geq r_P$. Indeed, for such u and r, it follows from the definition of S_r and Lemma 3.5 that the range of the state $y = S_r(u) \in H_0^1(\Omega) \cap H^2(\Omega) \subset C(\overline{\Omega})$ is a subset of the interval $[-r_P, r_P]$ and we obtain from Definition 6.4 and, in the case $r = r_P$, our conventions for the left and the right limits of $G[-r, r]$-functions at $\pm r$ that $\mathcal{E}_r(u)(t\pm) = u(t\pm)$ holds for all $t \in \mathrm{Range}(y)$, where the left/right limit $u(t\pm)$ has to be understood as being defined w.r.t. an arbitrary $G[-r, r]$-representative of u. The last identity allows us to recast the PDE (6.6) as

$$-\Delta\delta + u(y+)\max(0,\delta) + u(y-)\min(0,\delta) = -\int_0^y z(t)\,\mathrm{d}t \text{ in } \Omega, \qquad \delta = 0 \text{ on } \partial\Omega, \tag{6.9}$$

without introducing any ill- or undefined terms. By exploiting this observation, we can also state the following tangible corollary for the original solution operator S of (3.7):

Corollary 6.10 (Hadamard Directional Differentiability of S)**.** *Suppose that a control $0 \leq u \in L^2(\mathbb{R})$ is given such that $u|_{(-r,r)} \in L_G^\infty(-r, r)$ holds for some $r \geq r_P$. Then the solution operator S of (3.7) is Hadamard directionally differentiable at u in all directions $z \in \mathcal{T}(u; L_+^2(\mathbb{R}))$. Moreover, the state $y := S(u) \in C(\overline{\Omega})$ satisfies $\mathrm{Range}(y) \subset [-r_P, r_P] \subset [-r, r]$ and the directional derivative $\delta := S'(u; z) \in H_0^1(\Omega) \cap H^2(\Omega)$ of S at u in a direction $z \in \mathcal{T}(u; L_+^2(\mathbb{R}))$ is uniquely characterized by the partial differential equation*

$$-\Delta\delta + u(y+)\max(0,\delta) + u(y-)\min(0,\delta) = -g_z(y) \text{ in } \Omega, \qquad \delta = 0 \text{ on } \partial\Omega. \tag{6.10}$$

Here, $u(\,\cdot\,-)\colon [-r, r] \to \mathbb{R}$ and $u(\,\cdot\,+)\colon [-r, r] \to \mathbb{R}$ denote the left and right limit function in the sense of (2.1) of an arbitrary $G[-r, r]$-representative of $u|_{(-r,r)} \in L_G^\infty(-r, r)$ (acting as a Nemytskii operator).

Proof. The assertions of the corollary follow straightforwardly from Lemmas 3.5, 4.4 and Theorem 6.9. □

As a direct consequence of Theorems 5.7, 6.9 and Corollary 6.10, we obtain:

Corollary 6.11 (Hadamard Directional Differentiability at Local Minimizers)**.** *Suppose that there exists $r \geq r_P$ such that $u_D \in L^2(\mathbb{R})$ satisfies condition (A1). Then the solution operator $S\colon U_P \to H_0^1(\Omega) \cap H^2(\Omega)$ of (3.7) is Hadamard directionally differentiable in the sense of Corollary 6.10 at all local minimizers $\bar{u} \in L_+^2(\mathbb{R})$ of (P) in all directions $z \in \mathcal{T}(\bar{u}; L_+^2(\mathbb{R}))$. Moreover, the map $S_r\colon U_{P,r} \to H_0^1(\Omega) \cap H^2(\Omega)$ is Hadamard directionally* *differentiable in the sense of Theorem 6.9 at all points* $\bar{u} \in L^2(-r,r)$*, that are obtained by restricting a local minimizer* $\bar{u}$ *of* (P) *to the interval* $(-r,r)$*, in all directions* $z \in \mathcal{T}(\bar{u}; U_{P,r})$*.*

Proof. This follows immediately by combining Theorems 5.7, 6.9 and Corollary 6.10. □

As $\mathcal{T}(\bar{u}; L^2_+(\mathbb{R}))$ is the tangent cone of the set of admissible controls of (P) at $\bar{u}$, Corollary 6.11 gives rise to a Bouligand stationarity condition for local solutions of (P) in a straightforward manner. Before we turn our attention to this topic, we study the differentiability properties of the maps S and S_r in a particularly nice situation, namely that in which the following additional assumption is satisfied:

Definition 6.12 (Condition (A2))**.** We say that condition (A2) holds if the given right-hand side $f \in L^2(\Omega)$ of the partial differential equation (3.7) satisfies $\lambda^d(\{f = c\}) = 0$ for all $c \in \mathbb{R}$.

Note that (A2) is an assumption on the *problem data* of (P) that can be checked a-priori and not on the *problem solution*. This distinguishes (A2) from similar conditions found in the literature on optimal control of nonsmooth PDEs, *cf.* [4–6]. Condition (A2) is, for example, automatically satisfied if f is a nonconstant, real-analytic function; see [39], Proposition 1. The main motivation for considering (A2) is the following observation:

Lemma 6.13 (Level Sets of States)**.** *Suppose that (A2) holds. Then, for every* $y \in S(U_P)$ *and every* $c \in \mathbb{R}$*, the level set* $\{y = c\}$ *has measure zero.*

Proof. Let $y \in S(U_P)$ be given and let $u \in U_P$ be a control that satisfies $y = S(u)$. From the definition of S and Theorem 3.4, we obtain that y is an element of $H^1_0(\Omega) \cap H^2(\Omega)$ that satisfies

$$-\Delta y + g_u(y) = f \text{ a.e. in } \Omega, \tag{6.11}$$

where the Laplacian is understood in the strong sense, *i.e.*, in the sense of second weak derivatives. Suppose now that (A2) holds and that $c \in \mathbb{R}$ is given. Then, by [40], Lemma 4.1 and due to the definition of g_u, we have $\Delta y = 0$ a.e. on $\{y = c\}$ and $g_u(y) = g_u(c) = \text{const}$ a.e. on $\{y = c\}$. In combination with (6.11), these equations imply that $-\Delta y + g_u(y) = g_u(c) = f = \text{const}$ holds a.e. on $\{y = c\}$. Because of (A2), the last identity can only be true if the measure of the set $\{y = c\}$ is zero. This proves the claim. □

As an immediate consequence of Lemmas 6.3, 6.13, we obtain:

Lemma 6.14 ($G[-r,r]$-Representatives and States)**.** *Suppose that (A2) holds and that* $r > 0$*,* $y \in S(U_P)$*, and* $\mathfrak{u}_1, \mathfrak{u}_2 \in G[-r,r]$ *satisfying* $\mathfrak{u}_1(t) = \mathfrak{u}_2(t)$ *for a.a.* $t \in (-r,r)$ *are given. Then it holds*

$$\mathcal{E}_r(\mathfrak{u}_1)(y) = \mathcal{E}_r(\mathfrak{u}_1)(y+) = \mathcal{E}_r(\mathfrak{u}_1)(y-) = \mathcal{E}_r(\mathfrak{u}_2)(y) = \mathcal{E}_r(\mathfrak{u}_2)(y+) = \mathcal{E}_r(\mathfrak{u}_2)(y-) \text{ a.e. in } \Omega. \tag{6.12}$$

Proof. From Lemma 6.3ii), we obtain that there exists an at most countable set $Z \subset \mathbb{R}$ satisfying $\mathcal{E}_r(\mathfrak{u}_1)(t) = \mathcal{E}_r(\mathfrak{u}_1)(t\pm) = \mathcal{E}_r(\mathfrak{u}_2)(t\pm) = \mathcal{E}_r(\mathfrak{u}_2)(t)$ for all $t \in \mathbb{R} \setminus Z$, and from Lemma 6.13, we know that $\lambda^d(\{y = c\}) = 0$ holds for all $c \in Z$. If we combine these two observations with the fact that a countable union of sets of measure zero is still of measure zero, then (6.12) follows immediately. Note that, as regulated functions are Borel measurable and (every representative of) y is Lebesgue measurable, no measurability problems arise here. □

Analogously to Definition 6.4, we obtain from Lemma 6.14 that, under condition (A2), it makes sense to talk about the composition $\mathcal{E}_r(\mathfrak{u})(y) \in L^\infty(\Omega)$ of a control $\mathfrak{u} \in L^\infty_G(-r,r)$ and a state $y \in S(U_P)$.

Definition 6.15 (Control-State-Composition)**.** Suppose that $r > 0$ is given and that (A2) holds. For all $\mathfrak{u} \in L^\infty_G(-r,r)$ and all $y \in S(U_P)$, we define $\mathcal{E}_r(\mathfrak{u})(y)$ to be the element of $L^\infty(\Omega)$ that is obtained by extending an arbitrary $G[-r,r]$-representative of $\mathfrak{u}$ by zero in the sense of (6.1) and by composing the resulting function with y. If the $C(\overline{\Omega})$-representative of y satisfies $\operatorname{Range}(y) \subset [-r,r]$, then we also write $\mathfrak{u}(y)$ instead of $\mathcal{E}_r(\mathfrak{u})(y)$.

By combining Theorem 6.9 and Lemma 6.14, keeping in mind Definitions 6.4, 6.15, we arrive at:

Proposition 6.16 (Hadamard-Gâteaux Differentiability of S_r)**.** *Let $r > 0$ be given and suppose that condition (A2) holds. Then the function $S_r\colon U_{P,r} \to H_0^1(\Omega) \cap H^2(\Omega)$ is Hadamard-Gâteaux differentiable at all points $u \in L^\infty_G(-r,r) \cap U_{P,r}$ in the sense that, for every $u \in L^\infty_G(-r,r) \cap U_{P,r}$, the Hadamard directional derivative $S_r'(u;\cdot)\colon \mathcal{T}(u;U_{P,r}) \to H_0^1(\Omega) \cap H^2(\Omega)$ in (6.5) can be extended to a linear and continuous function $S_r'(u)\colon L^2(-r,r) \to H_0^1(\Omega) \cap H^2(\Omega)$. Moreover, for every $u \in L^\infty_G(-r,r) \cap U_{P,r}$ with state $y := S_r(u)$, the Hadamard-Gâteaux derivative $S_r'(u) \in \mathcal{L}(L^2(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$ is precisely the solution operator $z \mapsto \delta$ of the linear elliptic partial differential equation*

$$-\Delta\delta + \mathcal{E}_r(u)(y)\delta = -g_{E_r(z)}(y) \text{ in } \Omega, \qquad \delta = 0 \text{ on } \partial\Omega. \tag{6.13}$$

Here, the composition $\mathcal{E}_r(u)(y) \in L^\infty(\Omega)$ is understood in the sense of Definition 6.15.

Proof. Consider some $u \in L^\infty_G(-r,r) \cap U_{P,r}$ with state $y := S_r(u)$. From Definitions 6.4, 6.15 and Lemma 6.14, we obtain that $\mathcal{E}_r(u)(y) = \mathcal{E}_r(u)(y+) = \mathcal{E}_r(u)(y-)$ holds a.e. in Ω. This identity implies that the weak formulations of the PDEs (6.6) and (6.13) are identical and that the directional derivatives $S_r'(u;z)$, $z \in \mathcal{T}(u;U_{P,r})$, are indeed characterized by (6.13). To complete the proof, it remains to check that the solution operator $z \mapsto \delta$ of the PDE (6.13) is linear, continuous, and well-defined as a function from $L^2(-r,r)$ to $H_0^1(\Omega) \cap H^2(\Omega)$. Since the inequality $u \geq -\varepsilon_P$ a.e. in $(-r,r)$ implies that $\mathcal{E}_r(u)(t+) \geq -\varepsilon_P$ holds for all $t \in \mathbb{R}$ (see the proof of Theorem 6.9), due to the definition of $\mathcal{E}_r(u)(y)$, and due to (6.12), we know that $\mathcal{E}_r(u)(y) \geq -\varepsilon_P$ holds a.e. in Ω. Analogously to (3.6), this allows us to conclude that the operator $H_0^1(\Omega) \ni v \mapsto -\Delta v + \mathcal{E}_r(u)(y)v \in H^{-1}(\Omega)$ satisfies the coercivity estimate

$$\langle -\Delta v + \mathcal{E}_r(u)(y)v, v\rangle_{H_0^1(\Omega)} \geq \int_\Omega |\nabla v|^2 \,\mathrm{d}x - \varepsilon_P \int_\Omega v^2 \,\mathrm{d}x \geq \frac{c_P - \varepsilon_P}{2\max(1,c_P)} \|v\|^2_{H^1(\Omega)} \qquad \forall v \in H_0^1(\Omega). \tag{6.14}$$

Since the mapping $H_0^1(\Omega) \ni v \mapsto -\Delta v + \mathcal{E}_r(u)(y)v \in H^{-1}(\Omega)$ is continuous and since

$$\|g_{E_r(z)}(y)\|_{L^2(\Omega)} \leq \|y\|^{1/2}_{L^1(\Omega)} \|z\|_{L^2(-r,r)} \tag{6.15}$$

holds for all $z \in L^2(-r,r)$ by (3.5), it follows from the lemma of Lax-Milgram (see [33], Lem. 2.2) that (6.13) possesses a linear and continuous solution operator $L^2(-r,r) \ni z \mapsto \delta \in H_0^1(\Omega)$. Due to (3.4) and again (6.15), we further obtain that there exists a constant $C > 0$ such that the solution δ of (6.13) satisfies

$$\|\delta\|_{H^2(\Omega)} \leq C\|\mathcal{E}_r(u)(y)\delta + g_{E_r(z)}(y)\|_{L^2(\Omega)} \leq C\left(\|\mathcal{E}_r(u)(y)\|_{L^\infty(\Omega)}\|\delta\|_{L^2(\Omega)} + \|y\|^{1/2}_{L^1(\Omega)}\|z\|_{L^2(-r,r)}\right). \tag{6.16}$$

In combination with the already established continuity and linearity of the map $z \mapsto \delta$ as a function from $L^2(-r,r)$ to $H_0^1(\Omega)$, the above shows that the solution mapping of the PDE (6.13) is indeed an element of $\mathcal{L}(L^2(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$. This completes the proof of the proposition. ☐

We remark that regularization effects similar to that in Proposition 6.16 have already been exploited in [40] for the analysis of multi-objective optimal control problems governed by nonsmooth semilinear elliptic equations and in [41] for the optimal control of nonsmooth quasilinear PDEs. (In both of these contributions, the appearing nonsmooth superposition operator is fixed, though, and the control enters *via* the right-hand side.)

We further would like to point out that it is crucial to work with the $G[-r,r]$-extension operator $\mathcal{E}_r$ in Proposition 6.16, and not with the $L^2(-r,r)$-extension operator E_r introduced in Definition 4.3ii). Indeed, only by restricting the attention to regulated representatives, we obtain from Lemmas 6.3, 6.13, 6.14 that the composition $\mathcal{E}_r(u)(y)$ of the control and the state in (6.13) is well defined. If we write, for example, $E_r(u)(y)$, then it is not clear anymore whether this expression is sensible as changing the $L^2(\mathbb{R})$-representative of $E_r(u)$ on an uncountable null set might affect the composition $E_r(u)(y)$ on a nonnegligible subset of Ω, even when all level sets of y have measure zero. Compare also with the comments in [13], Section 5 in this context.

By considering S_r as an operator on the Banach space $(L^\infty_G(-r,r), \|\cdot\|_{L^\infty(-r,r)})$, we can refine Proposition 6.16 as follows:

Theorem 6.17 (Continuous Fréchet Differentiability of S_r on $L^\infty_G(-r,r)\cap U_{P,r}$)**.** *Let $r>0$ be given and suppose that condition (A2) holds. Consider the map S_r as a function $S_r\colon L^\infty_G(-r,r)\cap U_{P,r} \to H_0^1(\Omega)\cap H^2(\Omega)$, i.e., as an operator defined on a subset of the space $(L^\infty_G(-r,r), \|\cdot\|_{L^\infty(-r,r)})$. Then S_r is continuously Fréchet differentiable on $L^\infty_G(-r,r)\cap U_{P,r}$ with the derivatives $S_r'(u)$ from Proposition 6.16 in the following sense:*

i) For all sequences $\{u_n\}\subset L^\infty_G(-r,r)\cap U_{P,r}$ satisfying $u_n\to u$ in $L^\infty(-r,r)$ for some $u\in L^\infty_G(-r,r)\cap U_{P,r}$, it holds $S_r'(u_n)\to S_r'(u)$ in $\mathcal{L}(L^2(-r,r), H_0^1(\Omega)\cap H^2(\Omega))$.
ii) For every $u\in L^\infty_G(-r,r)\cap U_{P,r}$, it holds

$$\lim_{0<\|z\|_{L^\infty(-r,r)}\to 0,\, u+z\in L^\infty_G(-r,r)\cap U_{P,r}} \frac{\|S_r(u+z)-S_r(u)-S_r'(u)z\|_{H^2(\Omega)}}{\|z\|_{L^2(-r,r)}} = 0. \tag{6.17}$$

Proof. We begin with the proof of i): Let $\{u_n\}\subset L^\infty_G(-r,r)\cap U_{P,r}$ be a sequence satisfying $u_n\to u$ in $L^\infty(-r,r)$ for some $u\in L^\infty_G(-r,r)\cap U_{P,r}$ and let $\{z_n\}\subset L^2(-r,r)$ be given such that $\|z_n\|_{L^2(-r,r)}\le 1$ holds for all n. Define $y_n := S_r(u_n)$ and $y := S_r(u)$. From Theorem 3.4i) and the definition of S_r, we obtain that $y_n\to y$ holds in $H^2(\Omega)$, and from (6.12), (6.13), and Definitions 6.4, 6.15, that the functions $\delta_n := S_r'(u_n)z_n$, $\eta_n := S_r'(u)z_n$, and $\mu_n := \delta_n-\eta_n$ satisfy (in the weak sense)

$$-\Delta\mu_n + \mathcal{E}_r(u)(y)\mu_n = \left[\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u_n)(y_n+)\right]\delta_n + g_{E_r(z_n)}(y) - g_{E_r(z_n)}(y_n) \text{ in } \Omega, \qquad \mu_n = 0 \text{ on } \partial\Omega. \tag{6.18}$$

Due to (3.1), the convergence $y_n\to y$ in $H^2(\Omega)$, and the bound $\|z_n\|_{L^2(-r,r)}\le 1$, we know that

$$\left\|g_{E_r(z_n)}(y) - g_{E_r(z_n)}(y_n)\right\|_{L^2(\Omega)} \le \|E_r(z_n)\|_{L^2(\mathbb{R})}\|y-y_n\|_{L^1(\Omega)}^{1/2} \le \|y-y_n\|_{L^1(\Omega)}^{1/2}\to 0$$

holds for $n\to\infty$. Using the estimates (6.14) and (6.15), the PDE (6.13), and again the bound $\|z_n\|_{L^2(-r,r)}\le 1$, it is further easy to check that $\delta_n = S_r'(u_n)z_n$ satisfies

$$\|\delta_n\|_{H^1(\Omega)} \le C\|g_{E_r(z_n)}(y_n)\|_{L^2(\Omega)} \le C\|y_n\|_{L^1(\Omega)}^{1/2}\|z_n\|_{L^2(-r,r)} \le C\|y_n\|_{L^1(\Omega)}^{1/2}$$

for all n with a constant $C>0$ that is independent of n. In combination with the same bootstrapping argument as in (6.16), Theorem 3.4ii), Lemma 6.5, and the convergence $u_n\to u$ in $L^\infty(-r,r)$, this shows that the sequence $\{\delta_n\}$ is bounded in $H^2(\Omega)$ and that

$$\begin{aligned}
\|\left[\mathcal{E}_r(u)(y_n+) - \mathcal{E}_r(u_n)(y_n+)\right]\delta_n\|_{L^2(\Omega)} &\le \|\mathcal{E}_r(u)(y_n+) - \mathcal{E}_r(u_n)(y_n+)\|_{L^\infty(\Omega)}\|\delta_n\|_{L^2(\Omega)}\\
&\le \sup_{t\in\mathbb{R}}|\mathcal{E}_r(u)(t+) - \mathcal{E}_r(u_n)(t+)|\|\delta_n\|_{L^2(\Omega)}\\
&= \|u-u_n\|_{L^\infty(-r,r)}\|\delta_n\|_{L^2(\Omega)}\to 0
\end{aligned}$$

holds for $n\to\infty$. Let us now consider the differences $[\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u)(y_n+)]\delta_n$. Since the function $\mathcal{E}_r(u)(\,\cdot+)\colon\mathbb{R}\to\mathbb{R}$ is continuous at all points $t\in\mathbb{R}\setminus Z$ with an at most countable exceptional set $Z\subset\mathbb{R}$ by Lemma 6.3 and Definition 6.4, since $y_n\to y$ holds in $C(\overline{\Omega})$ due to the convergence $y_n\to y$ in $H^2(\Omega)$ and the Sobolev embeddings, and since all level sets of y have measure zero by Lemma 6.13, we have

$$\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u)(y_n+) = \mathbb{1}_{\{y\notin Z\}}\left(\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u)(y_n+)\right)\to 0 \quad \text{ a.e. in } \Omega.$$

In combination with the boundedness of $\{\delta_n\}$ in $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$, the fact that $\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u)(y_n+)$ is essentially bounded, and the dominated convergence theorem, this yields that $[\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u)(y_n+)]\delta_n$ converges strongly to zero in $L^2(\Omega)$. In total, we have now proven that

$$\begin{aligned}
&\left\|[\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u_n)(y_n+)]\,\delta_n + g_{E_r(z_n)}(y) - g_{E_r(z_n)}(y_n)\right\|_{L^2(\Omega)} \\
&\quad\le \left\|g_{E_r(z_n)}(y) - g_{E_r(z_n)}(y_n)\right\|_{L^2(\Omega)} + \|[\mathcal{E}_r(u)(y_n+) - \mathcal{E}_r(u_n)(y_n+)]\,\delta_n\|_{L^2(\Omega)} \\
&\qquad\qquad + \|[\mathcal{E}_r(u)(y+) - \mathcal{E}_r(u)(y_n+)]\,\delta_n\|_{L^2(\Omega)} \\
&\quad\to 0
\end{aligned}$$

holds for $n \to \infty$. Due to the PDE (6.18), the estimate (6.14), and the lemma of Lax-Milgram (see [33], Lem. 2.2), this shows that μ_n converges to zero in $H^1(\Omega)$ and, again by the bootstrapping argument in (6.16), that $\mu_n \to 0$ holds in $H^2(\Omega)$. As μ_n satisfies $\mu_n = \delta_n - \eta_n = S_r'(u_n)z_n - S_r'(u)z_n$ by definition, this yields $\|S_r'(u_n)z_n - S_r'(u)z_n\|_{H^2(\Omega)} \to 0$. As $\{z_n\}$ was an arbitrary sequence satisfying $\|z_n\|_{L^2(-r,r)} \le 1$, it follows that $S_r'(u_n) \to S_r'(u)$ holds in $\mathcal{L}(L^2(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$. This completes the proof of i).

It remains to establish ii): Suppose that $u \in L_G^\infty(-r,r) \cap U_{P,r}$ and a sequence $\{z_n\} \subset L_G^\infty(-r,r)$ satisfying $0 < \|z_n\|_{L^\infty(-r,r)} \to 0$ and $u + z_n \in L_G^\infty(-r,r) \cap U_{P,r}$ for all n are given. From i), Proposition 6.16, and the convexity of the set $L_G^\infty(-r,r) \cap U_{P,r}$, we obtain that the map $[0,1] \ni s \mapsto S_r(u + sz_n) \in H_0^1(\Omega) \cap H^2(\Omega)$ is well-defined and Hadamard-Gâteaux differentiable on $[0,1]$ with a derivative that depends continuously on $s \in [0,1]$. By applying the mean value theorem to this function (see [42], Thm. 3.2.6), we obtain that

$$\begin{aligned}
&\frac{\|S_r(u+z_n) - S_r(u) - S_r'(u)z_n\|_{H^2(\Omega)}}{\|z_n\|_{L^2(-r,r)}} \\
&\quad= \frac{1}{\|z_n\|_{L^2(-r,r)}}\left\|\int_0^1 S_r'(u+sz_n)z_n\,\mathrm{d}s - S_r'(u)z_n\right\|_{H^2(\Omega)} \\
&\quad= \frac{1}{\|z_n\|_{L^2(-r,r)}}\left\|\int_0^1 [S_r'(u+sz_n) - S_r'(u)]z_n\,\mathrm{d}s\right\|_{H^2(\Omega)} \\
&\quad\le \sup_{z \in L_G^\infty(-r,r),\|z\|_{L^\infty(-r,r)} \le \|z_n\|_{L^\infty(-r,r)}, u+z \in L_G^\infty(-r,r)\cap U_{P,r}} \|S_r'(u+z) - S_r'(u)\|_{\mathcal{L}(L^2(-r,r),H^2(\Omega))}\,.
\end{aligned}$$

Since $\|z_n\|_{L^\infty(-r,r)} \to 0$ holds by assumption and due to i), the right-hand side of the last estimate goes to zero for $n \to \infty$. As the sequence $\{z_n\}$ was arbitrary, this establishes ii) and completes the proof. □

Note that, by combining the Fréchet differentiability property in point ii) of Theorem 6.17 with the elementary estimate $\|z\|_{L^2(-r,r)} \le \sqrt{2r}\|z\|_{L^\infty(-r,r)}$ for all $z \in L^\infty(-r,r)$, we obtain

$$\begin{aligned}
&\lim_{0<\|z\|_{L^\infty(-r,r)}\to 0,\, u+z\in L_G^\infty(-r,r)\cap U_{P,r}} \frac{\|S_r(u+z) - S_r(u) - S_r'(u)z\|_{H^2(\Omega)}}{\|z\|_{L^\infty(-r,r)}} \\
&\quad\le \sqrt{2r} \lim_{0<\|z\|_{L^\infty(-r,r)}\to 0,\, u+z\in L_G^\infty(-r,r)\cap U_{P,r}} \frac{\|S_r(u+z) - S_r(u) - S_r'(u)z\|_{H^2(\Omega)}}{\|z\|_{L^2(-r,r)}} \\
&\quad= 0.
\end{aligned}$$

Completely analogously, one can also check that the convergence $S_r'(u_n) \to S_r'(u)$ in $\mathcal{L}(L^2(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$ proven in point i) of Theorem 6.17 implies that $S_r'(u_n) \to S_r'(u)$ holds in $\mathcal{L}(L^\infty_G(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$. In summary, this shows that Theorem 6.17 establishes a notion of continuous Fréchet differentiability for the operator $S_r\colon L^\infty_G(-r,r) \cap U_{P,r} \to H_0^1(\Omega) \cap H^2(\Omega)$ that is stronger than the classical one for a function defined on a subset of $(L^\infty_G(-r,r), \|\cdot\|_{L^\infty(-r,r)})$ with values in $(H_0^1(\Omega) \cap H^2(\Omega), \|\cdot\|_{H^2(\Omega)})$; *cf.* Definition 6.1iv).

We remark that Proposition 6.16 and Theorem 6.17 again give rise to corresponding results for the solution operator S of (3.7), similarly to the relationship between Theorem 6.9 and Corollary 6.10. We omit stating these reformulations here and instead turn our attention to the consequences that Theorem 6.17 has for the reduced objective function $F_r\colon U_{P,r} \to \mathbb{R}$ of the problem (P_r):

Corollary 6.18 (Continuous Fréchet Differentiability of F_r on $L^\infty_G(-r,r) \cap U_{P,r}$)**.** *Let $r > 0$ be given and suppose that (A2) holds. Consider the map F_r introduced in Definition 4.3v) as a function $F_r\colon L^\infty_G(-r,r) \cap U_{P,r} \to \mathbb{R}$, i.e., as a map defined on a subset of the Banach space $(L^\infty_G(-r,r), \|\cdot\|_{L^\infty(-r,r)})$ with values in $\mathbb{R}$. Then F_r is continuously Fréchet differentiable on the set $L^\infty_G(-r,r) \cap U_{P,r}$ in the sense of Definition 6.1iv). Further, for all $u \in L^\infty_G(-r,r) \cap U_{P,r}$ with associated state $y := S_r(u)$, the evaluation $F_r'(u)z \in \mathbb{R}$ of the Fréchet derivative $F_r'(u) \in L^\infty_G(-r,r)^*$ of F_r at u in direction $z \in L^\infty_G(-r,r)$ is given by*

$$F_r'(u)z = \big(-p_2 + \nu_1 + \nu_2(u - u_D), z\big)_{L^2(-r,r)}, \tag{6.19}$$

where $p_2\colon [-r,r] \to \mathbb{R}$ is the function defined by the system

$$-\Delta p_1 + \mathcal{E}_r(u)(y)p_1 = y - y_D \text{ in } \Omega, \quad p_1 = 0 \text{ on } \partial\Omega, \qquad p_2(s) = \begin{cases} \displaystyle\int_{\{y \geq s\}} p_1 \,\mathrm{d}x & \text{if } s \in [0,r], \\ \displaystyle -\int_{\{y \leq s\}} p_1 \,\mathrm{d}x & \text{if } s \in [-r,0). \end{cases} \tag{6.20}$$

Proof. From Theorem 6.17 and the structure of F_r (or, alternatively, the chain rule), we obtain that the function $F_r\colon L^\infty_G(-r,r) \cap U_{P,r} \to \mathbb{R}$ is continuously Fréchet differentiable in the sense of Definition 6.1 (with $U := L^\infty_G(-r,r)$, $Y := \mathbb{R}$, and $D := L^\infty_G(-r,r) \cap U_{P,r}$) and that the derivatives of F_r satisfy

$$F_r'(u)z = (S_r(u) - y_D, S_r'(u)z)_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 z + \nu_2(u - u_D)z \,\mathrm{d}s \quad \forall u \in L^\infty_G(-r,r) \cap U_{P,r} \quad \forall z \in L^\infty_G(-r,r). \tag{6.21}$$

Consider now some $u \in L^\infty_G(-r,r) \cap U_{P,r}$ with associated state $y := S_r(u)$. Using the same arguments as in the proof of Proposition 6.16, one checks that the elliptic PDE in (6.20) possesses a unique solution $p_1 \in H_0^1(\Omega) \cap H^2(\Omega) \subset C(\overline{\Omega})$. If we choose functions $S_r'(u)z \in H_0^1(\Omega) \cap H^2(\Omega)$, $z \in L^\infty_G(-r,r)$, as test functions in the weak form of the PDE for p_1 and exploit (6.13), then we obtain

$$\int_\Omega (y - y_D) S_r'(u)z \,\mathrm{d}x = \int_\Omega \nabla p_1 \cdot \nabla S_r'(u)z + \mathcal{E}_r(u)(y) p_1 S_r'(u)z \,\mathrm{d}x = -\int_\Omega g_{E_r(z)}(y) p_1 \,\mathrm{d}x \qquad \forall z \in L^\infty_G(-r,r). \tag{6.22}$$

By plugging in the definition of $g_{E_r(z)}$, by invoking the theorem of Fubini (see [27], Thm. 3.4.4), by exploiting Lemma 6.13, and by using the definition of p_2, we can rewrite the right-hand side of (6.22) as follows:

$$\begin{aligned}
-\int_\Omega g_{E_r(z)}(y)p_1\,\mathrm{d}x &= -\int_{\{y>0\}}\int_0^{y(x)} E_r(z)(s)\,\mathrm{d}s\,p_1(x)\,\mathrm{d}x + \int_{\{y<0\}}\int_{y(x)}^0 E_r(z)(s)\,\mathrm{d}s\,p_1(x)\,\mathrm{d}x\\
&= -\int_{\{y>0\}}\int_{\mathbb{R}} \mathbb{1}_{[0,y(x)]}(s)E_r(z)(s)\,\mathrm{d}s\,p_1(x)\,\mathrm{d}x + \int_{\{y<0\}}\int_{\mathbb{R}} \mathbb{1}_{[y(x),0]}(s)E_r(z)(s)\,\mathrm{d}s\,p_1(x)\,\mathrm{d}x\\
&= \int_{\mathbb{R}} E_r(z)(s)\left(-\int_{\{y>0\}}\mathbb{1}_{[0,y(x)]}(s)p_1(x)\,\mathrm{d}x + \int_{\{y<0\}}\mathbb{1}_{[y(x),0]}(s)p_1(x)\,\mathrm{d}x\right)\mathrm{d}s\\
&= \int_{\mathbb{R}} E_r(z)(s)\left(-\mathbb{1}_{(0,\infty)}(s)\int_{\{y\geq s\}}p_1(x)\,\mathrm{d}x + \mathbb{1}_{(-\infty,0)}(s)\int_{\{y\leq s\}}p_1(x)\,\mathrm{d}x\right)\mathrm{d}s\\
&= -\int_{-r}^r z(s)p_2(s)\,\mathrm{d}s \qquad \forall z\in L^\infty_G(-r,r).
\end{aligned} \tag{6.23}$$

In combination with (6.21) and (6.22), this establishes (6.19) and completes the proof. □

Analogously to the map $S_r\colon L^\infty_G(-r,r)\cap U_{P,r}\to H^1_0(\Omega)\cap H^2(\Omega)$, the function $F_r\colon L^\infty_G(-r,r)\cap U_{P,r}\to\mathbb{R}$ is, in fact, continuously Fréchet differentiable in a sense that is slightly stronger than that in Definition 6.1. Indeed, from (6.21) and points i) and ii) of Theorem 6.17, it follows immediately that we not only have $F_r'(u_n)\to F_r'(u)$ in $L^\infty_G(-r,r)^*$ for all sequences $\{u_n\}\subset L^\infty_G(-r,r)\cap U_{P,r}$ which satisfy $u_n\to u$ for some $u\in L^\infty_G(-r,r)\cap U_{P,r}$ in $L^\infty(-r,r)$, but even $F_r'(u_n)\to F_r'(u)$ in $L^2(-r,r)^*\cong L^2(-r,r)$, and that we may even divide by the $L^2(-r,r)$-norm in the definition of Fréchet differentiability, and not only by the $L^\infty(-r,r)$-norm; *cf.* (6.17). We omit stating these properties in a separate result here because they are not needed for the subsequent analysis.

Note that, because of the $L^2(-r,r)$-scalar product in (6.19), the function $p_2\colon[-r,r]\to\mathbb{R}$ in (6.20) can be redefined on a set of measure zero without affecting the validity of the formula for the Fréchet derivative $F_r'(u)$. In particular, we could also interpret p_2 as an element of $L^\infty(-r,r)$ in Corollary 6.18 (where the essential boundedness of p_2 follows from the regularity properties of p_1). We have introduced p_2 as a classical function and not as an element of $L^\infty(-r,r)$ here because the representative in (6.20) enjoys very convenient additional regularity properties. We collect these in:

Lemma 6.19 (Regularity Properties of the Adjoint Variable p_2)**.** *Let $r>0$ be given and suppose that (A2) holds. Assume that a control $u\in L^\infty_G(-r,r)\cap U_{P,r}$ with associated state $y:=S_r(u)$ is given and that $p_1\in H^1_0(\Omega)\cap H^2(\Omega)$ and $p_2\colon[-r,r]\to\mathbb{R}$ are defined as in (6.20). Then the following is true:*

i) p_2 is continuous at every point $s\in[-r,r]\setminus\{0\}$ and right-continuous at $s=0$.
ii) p_2 is an element of the space $BV[-r,r]$.
iii) The distributional derivative of p_2 is the Borel measure on $[-r,r]$ given by

$$p_2' = \left(\int_\Omega p_1\,\mathrm{d}x\right)\delta_0 - ([y]^r_{-r})_*(p_1\lambda^d). \tag{6.24}$$

Here, δ_0 is the Dirac measure at zero and $([y]^r_{-r})_*(p_1\lambda^d)$ denotes the image (a.k.a. pushforward) of the weighted Lebesgue measure $p_1\lambda^d$ in the sense of [27], Section 3.6 under the continuous representative of

$$[y]^r_{-r} := \max(0,\min(r,y)) + \min(0,\max(-r,y)) = \begin{cases} r & \text{if } y\geq r,\\ y & \text{if } y\in(-r,r),\\ -r & \text{if } y\leq -r.\end{cases}$$

iv) If r and u satisfy $r \geq r_P$ and $u \geq 0$ a.e. in $(-r, r)$, then p_2 is identical zero in $[-r, r] \setminus (-r_P, r_P)$ and the formula (6.24) holds with $[y]_{-r}^{r}$ replaced by y.

Proof. Consider a sequence $\{s_n\} \subset [0, r]$ that satisfies $s_n \to s$ for some $s \in [0, r]$. We claim that the functions $\mathbb{1}_{\{y \geq s_n\}}$ converge pointwise a.e. in Ω to $\mathbb{1}_{\{y \geq s\}}$ for $n \to \infty$. To see this, let us assume that the maps $\mathbb{1}_{\{y \geq s_n\}}$ and $\mathbb{1}_{\{y \geq s\}}$ are defined w.r.t. some representative of y. Then, for every $x \in \Omega$ with $y(x) > s$, it holds $y(x) > s_n$ for all sufficiently large n and, thus, $\mathbb{1}_{\{y \geq s\}}(x) = \mathbb{1}_{\{y \geq s_n\}}(x) = 1$. Analogously, we obtain that, for every $x \in \Omega$ with $y(x) < s$, we have $y(x) < s_n$ for all large enough n and, thus, $\mathbb{1}_{\{y \geq s\}}(x) = \mathbb{1}_{\{y \geq s_n\}}(x) = 0$. Since the set $\{y = s\}$ has measure zero by Lemma 6.13, this shows that the sequence $\mathbb{1}_{\{y \geq s_n\}}$ indeed satisfies $\mathbb{1}_{\{y \geq s_n\}} \to \mathbb{1}_{\{y \geq s\}}$ a.e. in Ω for $n \to \infty$. By exploiting that $p_1 \in H_0^1(\Omega) \cap H^2(\Omega)$ is essentially bounded by the Sobolev embeddings and by invoking the dominated convergence theorem, we may now deduce that

$$\lim_{n \to \infty} p_2(s_n) = \lim_{n \to \infty} \int_\Omega \mathbb{1}_{\{y \geq s_n\}} p_1 \,\mathrm{d}x = \int_\Omega \mathbb{1}_{\{y \geq s\}} p_1 \,\mathrm{d}x = p_2(s).$$

This shows that p_2 is continuous on $(0, r]$ and right-continuous at $s = 0$. As the above arguments can also be used to establish the continuity of p_2 on $[-r, 0)$, i) now follows immediately.

To prove ii), we revisit (6.23) in the special case $z = v'$, $v \in C^1([-r, r])$. This yields

$$\begin{aligned}
\int_{-r}^{r} v'(s) p_2(s) \,\mathrm{d}s &= \int_{\{y>0\}} \int_0^{y(x)} E_r(v')(s) \,\mathrm{d}s \, p_1(x) \,\mathrm{d}x - \int_{\{y<0\}} \int_{y(x)}^{0} E_r(v')(s) \,\mathrm{d}s \, p_1(x) \,\mathrm{d}x \\
&= \int_{\{y>0\}} \int_0^{\min(r, y(x))} E_r(v')(s) \,\mathrm{d}s \, p_1(x) \,\mathrm{d}x - \int_{\{y<0\}} \int_{\max(-r, y(x))}^{0} E_r(v')(s) \,\mathrm{d}s \, p_1(x) \,\mathrm{d}x \\
&= \int_{\{y>0\}} \left[v(\min(r, y(x))) - v(0) \right] p_1(x) \,\mathrm{d}x - \int_{\{y<0\}} \left[v(0) - v(\max(-r, y(x))) \right] p_1(x) \,\mathrm{d}x \\
&= -v(0) \int_\Omega p_1 \,\mathrm{d}x + \int_\Omega v([y]_{-r}^{r}) p_1 \,\mathrm{d}x \qquad \forall v \in C^1([-r, r]).
\end{aligned} \tag{6.25}$$

Here, in the last step, we have exploited that the level set $\{y = 0\}$ has measure zero by Lemma 6.13 and that $[y]_{-r}^{r} = \max(0, \min(r, y)) + \min(0, \max(-r, y))$ holds by definition. By proceeding along the exact same lines as in the proof of Proposition 5.5, *i.e.*, by invoking the theorem of Hahn-Banach (see [30], Thm. 8.1.1), the Riesz representation theorem (see [30], Thm. 8.1.2), and the integration by parts formula for the Riemann-Stieltjes integral (see [30], Thm. 5.5.1), we obtain from (6.25) that there exists a function $q \in BV[-r, r]$ which satisfies $q = p_2$ a.e. in $(-r, r)$. Since we already know that p_2 is continuous in $[-r, r] \setminus \{0\}$, since elements of $BV[-r, r]$ are regulated, and since the complement of a set of Lebesgue measure zero is dense, the last equality implies that $q(t+) = p_2(t+) = p_2(t) = p_2(t-) = q(t-)$ holds for all $t \in (-r, r) \setminus \{0\}$. Recall that, as an element of $BV[-r, r]$, the function q can be written as the sum of a step function q_1 and a function $q_2 \in BV[-r, r] \cap C([-r, r])$; see [30], Theorem 2.6.1. Due to the identity $q(t+) = q(t-)$ for all $t \in (-r, r) \setminus \{0\}$ derived above, we know that $q(t+) - q(t-) = q_1(t+) - q_1(t-) = 0$ holds for all $t \in (-r, r) \setminus \{0\}$. This shows that there exist constants $c_1, c_2 \in \mathbb{R}$ which satisfy $q_1(t+) = q_1(t-) = c_1$ for all $t \in (-r, 0)$ and $q_1(t+) = q_1(t-) = c_2$ for all $t \in (0, r)$; see the formulas in [30], equations (2.5.3) and (2.5.7). We thus have $p_2(t) = q_2(t) + c_1$ for all $t \in (-r, 0)$ and $p_2(t) = q_2(t) + c_2$ for all $t \in (0, r)$. In summary, we may now conclude that

$$p_2 = q_2 + (p_2(-r) - q_2(-r))\mathbb{1}_{\{-r\}} + c_1 \mathbb{1}_{(-r,0)} + (p_2(0) - q_2(0))\mathbb{1}_{\{0\}} + c_2 \mathbb{1}_{(0,r)} + (p_2(r) - q_2(r))\mathbb{1}_{\{r\}}$$

holds on $[-r, r]$. This proves that p_2 is the sum of functions of bounded variation and that $p_2 \in BV[-r, r]$ holds as claimed in ii).

To establish iii), it suffices to note that (6.25) and [27], Theorem 3.6.1 (which is applicable here due to the comments on [27], p. 191) imply

$$\int_{-r}^{r} v'(s)p_2(s)\,\mathrm{d}s = -v(0)\int_\Omega p_1\,\mathrm{d}x + \int_\Omega v([y]_{-r}^r)p_1\,\mathrm{d}x = -v(0)\int_\Omega p_1\,\mathrm{d}x + \int_{-r}^{r} v\,\mathrm{d}\left[([y]_{-r}^r)_*(p_1\lambda^d)\right] \quad \forall v \in C^1([-r,r]).$$

This yields iii) as desired; see [28], Definition 2.2.4.

To finally prove iv), we recall that, in the case $r \geq r_P$ and $u \geq 0$ a.e. in $(-r,r)$, we obtain from Lemmas 3.5, 6.13 and the definition of S_r that the continuous representative of y satisfies $\mathrm{Range}(y) \subset [-r_P, r_P]$ and $\lambda^d(\{y = r_P\}) = \lambda^d(\{y = -r_P\}) = \lambda^d(\{y \geq r_P\}) = \lambda^d(\{y \leq -r_P\}) = 0$. In combination with (6.20) and the definition of $[y]_{-r}^r$, these properties immediately yield the assertions of iv). This completes the proof. □

As we will see in the next section, the properties of the adjoint variable p_2 established in Lemma 6.19 carry over to the local minimizers of (P) and (P_r) in many situations. Point ii) of Lemma 6.19 further makes it possible to interpret the adjoint variable p_2 in (6.20) as an element of the space $L^\infty_{BV}(-r,r) \subset L^\infty_G(-r,r)$. We will exploit this observation when setting up the gradient projection algorithm in Section 8.

7. First-order optimality conditions of Bouligand and primal-dual type

With the results of Sections 4 to 6 at hand, the derivation of first-order necessary optimality conditions for the problems (P) and (P_r) is straightforward. We begin by stating Bouligand stationarity conditions:

Theorem 7.1 (Bouligand Stationarity Condition for (P_r))**.** *Let $r > 0$ be given, suppose that u_D satisfies (A1), and assume that $\bar u$ is a local solution of (P_r) with state $\bar y := S_r(\bar u) \in H_0^1(\Omega) \cap H^2(\Omega)$. Then $\bar u$ is an element of $L^\infty_G(-r,r)$, the function $S_r\colon U_{P,r} \to H_0^1(\Omega) \cap H^2(\Omega)$ is Hadamard directionally differentiable at $\bar u$ in all directions $z \in \mathcal{T}(\bar u; U_{P,r})$ in the sense of Theorem 6.9, and it holds*

$$(\bar y - y_D, S_r'(\bar u; z))_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 z + \nu_2(\bar u - u_D) z\,\mathrm{d}t \geq 0 \qquad \forall z \in \mathcal{T}(\bar u; L^2_+(-r,r)). \tag{7.1}$$

Proof. The regularity $\bar u \in L^\infty_G(-r,r)$ has been proven in Theorem 5.6 and the Hadamard directional differentiability of S_r at $\bar u$ in all directions $z \in \mathcal{T}(\bar u; U_{P,r})$ follows from Theorem 6.9. To establish (7.1), let us suppose that $z \in \mathcal{T}(\bar u; L^2_+(-r,r))$ is given. From the definition of the tangent cone $\mathcal{T}(\bar u; L^2_+(-r,r))$, we obtain that there exist sequences $\{z_n\} \subset L^2(-r,r)$ and $\{s_n\} \subset (0,\infty)$ which satisfy $s_n \to 0$ and $z_n \to z$ in $L^2(-r,r)$ for $n \to \infty$ and $\bar u + s_n z_n \in L^2_+(-r,r)$ for all n; see [25], Section 2.2.4. In combination with the local optimality of $\bar u$ and the chain rule for Hadamard directionally differentiable functions (see [25], Prop. 2.47), this yields

$$0 \leq \lim_{n\to\infty} \frac{F_r(\bar u + s_n z_n) - F_r(\bar u)}{s_n} = (\bar y - y_D, S_r'(\bar u; z))_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 z + \nu_2(\bar u - u_D) z\,\mathrm{d}t. \tag{7.2}$$

As $z \in \mathcal{T}(\bar u; L^2_+(-r,r))$ was arbitrary, (7.2) implies (7.1) and the proof is complete. □

Theorem 7.2 (Bouligand Stationarity Condition for (P))**.** *Suppose that $u_D \in L^2(\mathbb{R})$ satisfies (A1) for some $r \geq r_P$ and assume that $\bar u \in L^2(\mathbb{R})$ is a local solution of (P) with state $\bar y := S(\bar u)$. Then it holds $\bar u|_{(-r,r)} \in L^\infty_G(-r,r)$, the solution operator $S\colon U_P \to H_0^1(\Omega) \cap H^2(\Omega)$ of (3.7) is Hadamard directionally differentiable at $\bar u$ in all directions $z \in \mathcal{T}(\bar u; L^2_+(\mathbb{R}))$, and $\bar u$ satisfies*

$$(\bar y - y_D, S'(\bar u; z))_{L^2(\Omega)} + \int_{\mathbb{R}} \nu_1 z + \nu_2(\bar u - u_D) z\,\mathrm{d}t \geq 0 \qquad \forall z \in L^1(\mathbb{R}) \cap \mathcal{T}(\bar u; L^2_+(\mathbb{R})). \tag{7.3}$$

Proof. The proof is along the lines of that of Theorem 7.1 and thus omitted. $\square$

Note that, in the case $\nu_1 = 0$, the assumption $z \in L^1(\mathbb{R})$ in (7.3) can be dropped. An immediate consequence of Theorem 7.2 is the following result:

Corollary 7.3 (Tight Controls)**.** *In the case $u_D \equiv 0$, every local solution $\bar{u} \in L^2(\mathbb{R})$ of (P) satisfies $\bar{u} = 0$ a.e. in $\mathbb{R} \setminus \mathrm{Range}(\bar{y})$. Here, $\mathrm{Range}(\bar{y})$ denotes the range of the continuous representative of the state $\bar{y} := S(\bar{u})$.*

Proof. This follows straightforwardly from the Bouligand stationarity condition (7.3) and the observation that $S'(\bar{u}; z) = 0$ holds for all $z \in \mathcal{T}(\bar{u}; L^2_+(\mathbb{R}))$ that vanish a.e. in $\mathrm{Range}(\bar{y})$; see (6.10) and (2.3). $\square$

As already mentioned in Section 1.2, the notion of Bouligand stationarity is rarely helpful in practice. Under assumption (A2), we can establish optimality systems that are far more suitable for applications.

Theorem 7.4 (Primal-Dual Stationarity System for (P_r))**.** *Let $r > 0$ be given, suppose that u_D and f satisfy (A1) and (A2), and assume that $\bar{u}$ is a local solution of (P_r). Then there exist adjoint variables $\bar{p}_1$ and $\bar{p}_2$ such that $\bar{u}$ and its state $\bar{y}$ satisfy the stationarity system*

$$
\begin{gathered}
\bar{u} \in L^\infty_G(-r, r), \qquad \bar{y}, \bar{p}_1 \in H^1_0(\Omega) \cap H^2(\Omega), \qquad \bar{p}_2 \in L^\infty_{BV}(-r, r),\\
-\Delta \bar{y} + g_{E_r(\bar{u})}(\bar{y}) = f \ \text{ in } \Omega, \quad \bar{y} = 0 \ \text{ on } \partial\Omega,\\
-\Delta \bar{p}_1 + \mathcal{E}_r(\bar{u})(\bar{y})\bar{p}_1 = \bar{y} - y_D \ \text{ in } \Omega, \quad \bar{p}_1 = 0 \ \text{ on } \partial\Omega,\\
\bar{p}_2(s) = \begin{cases} \displaystyle\int_{\{\bar{y} \geq s\}} \bar{p}_1 \,\mathrm{d}x & \text{if } s \geq 0,\\ \displaystyle -\int_{\{\bar{y} \leq s\}} \bar{p}_1 \,\mathrm{d}x & \text{if } s < 0, \end{cases} \qquad \text{f.a.a. } s \in (-r, r),\\
\bar{u}(s) = \max\left(0, u_D(s) + \frac{\bar{p}_2(s) - \nu_1}{\nu_2}\right) \qquad \text{f.a.a. } s \in (-r, r).
\end{gathered}
\tag{7.4}
$$

Here, E_r and $\mathcal{E}_r$ are defined as in Definitions 4.3, 6.2 and the composition $\mathcal{E}_r(\bar{u})(\bar{y})$ is understood in the sense of Definition 6.15. If $u_D = u_D|_{(-r,r)} \in L^\infty_{BV}(-r, r)$ holds, then we additionally have $\bar{u} \in L^\infty_{BV}(-r, r)$.

Proof. From the local optimality of $\bar{u}$ in (P_r), Theorem 5.6, Corollary 6.18, and the convexity of the set $L^\infty_G(-r, r) \cap L^2_+(-r, r)$, it follows that $\bar{u}$ is a solution of the variational inequality

$$
\tilde{u} \in L^\infty_G(-r, r) \cap L^2_+(-r, r), \qquad \big(-\bar{p}_2 + \nu_1 + \nu_2(\tilde{u} - u_D), u - \tilde{u}\big)_{L^2(-r,r)} \geq 0 \quad \forall u \in L^\infty_G(-r, r) \cap L^2_+(-r, r),
\tag{7.5}
$$

where $\bar{p}_2 \in L^\infty_{BV}(-r, r)$ is defined by the relations in (7.4). Using a simple contradiction argument and a pointwise distinction of cases, one easily checks that (7.5) can admit at most one solution $\tilde{u}$ and that the function $\max(0, u_D + (\bar{p}_2 - \nu_1)/\nu_2)$ solves (7.5). Thus, $\bar{u} = \max(0, u_D + (\bar{p}_2 - \nu_1)/\nu_2)$ as claimed. Note that this projection formula also implies that $\bar{u} \in L^\infty_{BV}(-r, r)$ holds if u_D satisfies $u_D|_{(-r,r)} \in L^\infty_{BV}(-r, r)$. By collecting all of the available information, (7.4) now follows and the proof is complete. $\square$

Theorem 7.5 (Primal-Dual Stationarity System for (P))**.** *Suppose that u_D satisfies (A1) for some $r \geq r_P$ and that (A2) holds. Then, for every local solution $\bar{u} \in L^2(\mathbb{R})$ of the problem (P) with state $\bar{y}$, the range of the continuous representative of $\bar{y}$ satisfies $\mathrm{Range}(\bar{y}) \subset [-r_P, r_P] \subset [-r, r]$ and there exist $\bar{p}_1$ and $\bar{p}_2$ such that the* *following stationarity system holds true:*

$$
\begin{gathered}
\bar{u}, \bar{p}_2 \in L^2(\mathbb{R}), \qquad \bar{y}, \bar{p}_1 \in H_0^1(\Omega) \cap H^2(\Omega), \qquad \bar{u}|_{(-r,r)} \in L_G^\infty(-r,r), \qquad \bar{p}_2|_{(-r,r)} \in L_{BV}^\infty(-r,r),\\
-\Delta \bar{y} + g_{\bar{u}}(\bar{y}) = f \text{ in } \Omega, \quad \bar{y} = 0 \text{ on } \partial\Omega,\\
-\Delta \bar{p}_1 + \bar{u}(\bar{y})\bar{p}_1 = \bar{y} - y_D \text{ in } \Omega, \quad \bar{p}_1 = 0 \text{ on } \partial\Omega,\\
\bar{p}_2(s) = \begin{cases} \displaystyle\int_{\{\bar{y} \geq s\}} \bar{p}_1 \,\mathrm{d}x & \text{if } s \geq 0,\\ \displaystyle -\int_{\{\bar{y} \leq s\}} \bar{p}_1 \,\mathrm{d}x & \text{if } s < 0, \end{cases} \qquad \text{f.a.a. } s \in \mathbb{R},\\
\bar{u}(s) = \max\left(0, u_D(s) + \frac{\bar{p}_2(s) - \nu_1}{\nu_2}\right) \qquad \text{f.a.a. } s \in \mathbb{R}.
\end{gathered} \tag{7.6}
$$

Here, the composition $\bar{u}(\bar{y}) = \bar{u}|_{(-r,r)}(\bar{y})$ is defined as in Definition 6.15. If $u_D|_{(-r,r)} \in L_{BV}^\infty(-r,r)$ holds, then the inclusion $\bar{u}|_{(-r,r)} \in L_G^\infty(-r,r)$ in (7.6) can be replaced by $\bar{u}|_{(-r,r)} \in L_{BV}^\infty(-r,r)$.

Proof. This follows straightforwardly from Lemma 3.5 and Theorems 4.5, 7.4. Note that $\bar{p}_2$ vanishes by definition a.e. in $\mathbb{R} \setminus [-r_P, r_P]$. □

The primal-dual stationarity systems (7.4) and (7.6) are not only appropriate points of departure for the design and analysis of numerical solution procedures but also provide interesting insights into the structural properties of locally optimal controls $\bar{u}$. From Lemma 6.19, we obtain, for example, the following:

Corollary 7.6 (Similarities Between Optimal Superposition Operators and Activation Functions)**.** *Suppose that (A2) is satisfied and that $u_D \in C(\mathbb{R}) \cap L^1(\mathbb{R})$ holds. Assume further that $\bar{u}$ is a local solution of (P). Then the function $g_{\bar{u}} : \mathbb{R} \to \mathbb{R}$ associated with $\bar{u}$ possesses the following properties:*

- *i) $g_{\bar{u}}$ is Hadamard directionally differentiable on $\mathbb{R}$;*
- *ii) $g_{\bar{u}}$ is continuously differentiable in $\mathbb{R} \setminus \{0\}$;*
- *iii) $g_{\bar{u}}$ is sigmoidal in the sense that it is nondecreasing, satisfies $g_{\bar{u}}(0) = 0$, and admits well-defined and finite limits $g_{\bar{u}}(-\infty) := \lim_{t\to-\infty} g_{\bar{u}}(t)$ and $g_{\bar{u}}(\infty) := \lim_{t\to\infty} g_{\bar{u}}(t)$.*

Proof. Due to the regularity $u_D \in C(\mathbb{R})$, we know that (A1) holds for all $r > 0$. This allows us to invoke Theorem 7.5 and points i), ii), and iv) of Lemma 6.19 to obtain that the function $\bar{p}_2 \in L^2(\mathbb{R})$ in (7.6) possesses a representative which is continuous in $\mathbb{R} \setminus \{0\}$, identical zero in $\mathbb{R} \setminus [-r_P, r_P]$, and regulated on $[-r_P, r_P]$. Since u_D is an element of $C(\mathbb{R})$ and since $\bar{u} = \max(0, u_D + (\bar{p}_2 - \nu_1)/\nu_2)$ holds by (7.6), the properties of $\bar{p}_2$ imply that $\bar{u}$ possesses a representative that is continuous in $\mathbb{R} \setminus \{0\}$ and regulated in every neighborhood of the origin. As $g_{\bar{u}}$ arises from $\bar{u}$ by integration, the assertions i) and ii) of the corollary now follow immediately; *cf.* Lemma 6.7. It remains to establish iii). To this end, we note that $g_{\bar{u}} : \mathbb{R} \to \mathbb{R}$ is trivially nondecreasing and zero at the origin by the definition of $g_{\bar{u}}$ and the inequality $\bar{u} \geq 0$ a.e. in $\mathbb{R}$. From (7.6) and the fact that $\bar{p}_2$ has a compact essential support, we further obtain that $\bar{u} \in L^1(\mathbb{R})$ holds. Again by the definition of $g_{\bar{u}}$ and due to the dominated convergence theorem, this yields

$$\lim_{t\to\infty} g_{\bar{u}}(t) = \lim_{t\to\infty} \int_{(0,\infty)} \mathbb{1}_{(0,t)} \bar{u} \,\mathrm{d}s = \int_0^\infty \bar{u} \,\mathrm{d}s$$

and

$$\lim_{t\to-\infty} g_{\bar{u}}(t) = \lim_{t\to\infty} -\int_{(-\infty,0)} \mathbb{1}_{(-t,0)} \bar{u} \,\mathrm{d}s = -\int_{-\infty}^0 \bar{u} \,\mathrm{d}s.$$

The limits $g_{\bar{u}}(-\infty)$ and $g_{\bar{u}}(\infty)$ thus exist and the proof is complete. □

Note that, from Lemma 6.13 and the stationarity system (7.6) (or, alternatively, from the expression for the distributional derivative in (6.24)), we obtain that $\bar{p}_2$ can only possess a continuous representative in the situation of Corollary 7.6 if $\bar{p}_1$ satisfies

$$\int_{\{\bar{y}\geq 0\}} \bar{p}_1 \,\mathrm{d}x + \int_{\{\bar{y}<0\}} \bar{p}_1 \,\mathrm{d}x = \int_{\Omega} \bar{p}_1 \,\mathrm{d}x = 0.$$

Since cases in which $\bar{p}_1$ has zero mean are exceptional, this shows that the control $\bar{u} = \max(0, u_D + (\bar{p}_2 - \nu_1)/\nu_2)$ tends to possess a proper jump at zero (unless, of course, the parameters u_D, ν_1, and ν_2 are such that the jump is removed by the projection formula). In particular, one has to expect that the function $g_{\bar{u}}$ arising from an optimal control $\bar{u}$ of (P) has a proper kink at the origin. In Section 9, we will see that such a nondifferentiability at zero is indeed what is typically observed in practice and that solving a problem of the type (P) usually gives rise to a "ReLU-like" sigmoidal function $g_{\bar{u}}\colon \mathbb{R} \to \mathbb{R}$.

Regarding the variable ν_1 in (7.6), we would like to point out that this parameter promotes sparsity properties of the optimal controls $\bar{u}$ of the problem (P). In fact, using the bound in Theorem 3.4ii), the form of the adjoint equation for $\bar{p}_1$ in (7.6), and the estimate (6.14), it is easy to check that there exists a constant $C > 0$, which depends only on Ω, y_D, ε_P, and the bound $M > 0$ in Theorem 3.4ii), such that every $\bar{p}_2 \in L^2(\mathbb{R})$ associated with a local solution $\bar{u} \in L^2(\mathbb{R})$ of (P) satisfies $\|\bar{p}_2\|_{L^\infty(\mathbb{R})} \leq C$. Due to the projection formula in (7.6), this implies that, in the case $u_D \in L^\infty(\mathbb{R})$, there exists a value $\bar{\nu}_1 > 0$ such that (P) is solved only by $\bar{u} = 0$ for all $\nu_1 \geq \bar{\nu}_1$. In the numerical experiments of Section 9, we will demonstrate that this effect is also present in practice and that the essential support of the optimal controls of (P) shrinks as ν_1 goes to infinity.

8. A GRADIENT PROJECTION ALGORITHM AND ITS CONVERGENCE ANALYSIS

In this section, we demonstrate that the differentiability results in Section 6 not only make it possible to derive first-order necessary optimality conditions for problems of the type (P) and (P_r) but also to set up numerical solution algorithms for the calculation of local minimizers. To be able to cover the control-discrete and the control-continuous case simultaneously, we henceforth consider the following variant of problem (P_r):

$$\left.\begin{aligned} \text{Minimize} \quad & F_r(u) = \frac{1}{2}\|S_r(u) - y_D\|^2_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 u + \frac{\nu_2}{2}(u - u_D)^2 \,\mathrm{d}t \\ \text{w.r.t.} \quad & u \in \mathcal{U}, \\ \text{s.t.} \quad & u \geq 0 \text{ a.e. in } (-r, r). \end{aligned}\right\} \tag{$\mathrm{P}_{r,\mathcal{U}}$}$$

We collect our assumptions on the quantities in ($\mathrm{P}_{r,\mathcal{U}}$) in:

Assumption 8.1 (Standing Assumptions for Section 8)**.** Throughout this section, we assume that:

- r is a given positive number;
- u_D, E_r, S_r, $U_{P,r}$, and F_r are defined as in Definition 4.3;
- u_D satisfies condition (A1);
- f satisfies condition (A2);
- $\mathcal{U}$ is either the Banach space $(L^\infty_G(-r,r), \|\cdot\|_{L^\infty(-r,r)})$ or a finite-dimensional control space of piecewise constant finite element functions subordinate to a grid $-r = r_0 < r_1 < r_2 < ... < r_N = r$, $N \in \mathbb{N}$, *i.e.*,

$$\mathcal{U} = \{u \in L^\infty(-r,r) \mid u = \text{const a.e. on } (r_{n-1}, r_n)\ \forall n = 1, ..., N\} \subset L^\infty_G(-r,r). \tag{8.1}$$

In addition, u_D, y_D, Ω, d, ν_1, ν_2, and f are, of course, still assumed to satisfy the conditions in Assumption 2.1. Note that, by setting up a numerical algorithm for the solution of ($\mathrm{P}_{r,\mathcal{U}}$), we also obtain a solution method for our original identification problem (P) due to Theorem 5.6 and the equivalence in Theorem 4.5. Using

Theorem 5.6 and the direct method of the calculus of variations, one further easily checks that ($\mathrm{P}_{r,\mathcal{U}}$) possesses at least one global solution $\bar{u} \in \mathcal{U}$. To discuss the first-order necessary optimality condition of ($\mathrm{P}_{r,\mathcal{U}}$) and the associated gradient projection method in an adequate manner, we introduce:

Definition 8.2 (L^2-Projection onto $\mathcal{U}$)**.** We define $\mathcal{P}_{\mathcal{U}}^{L^2}: L^\infty_G(-r,r) \to \mathcal{U}$ to be the function that maps a given $u \in L^\infty_G(-r,r)$ to the unique solution $\mathcal{P}_{\mathcal{U}}^{L^2}(u)$ of the variational problem

$$\mathcal{P}_{\mathcal{U}}^{L^2}(u) \in \mathcal{U}, \qquad \left(u - \mathcal{P}_{\mathcal{U}}^{L^2}(u), z\right)_{L^2(-r,r)} = 0 \quad \forall z \in \mathcal{U}.$$

Note that, in the case $\mathcal{U} = L^\infty_G(-r,r)$, the function $\mathcal{P}_{\mathcal{U}}^{L^2}$ is just the identity map on $L^\infty_G(-r,r)$. (In particular, it is not a problem in this case that $\mathcal{U} = L^\infty_G(-r,r)$ is not a closed subspace of $(L^2(-r,r), \|\cdot\|_{L^2(-r,r)})$ as would normally be required to define the orthogonal projection in $L^2(-r,r)$.) If $\mathcal{U}$ is a finite element space as in (8.1), then $\mathcal{U}$ is a closed subspace of $L^2(-r,r)$ and $\mathcal{P}_{\mathcal{U}}^{L^2}: L^\infty_G(-r,r) \to \mathcal{U}$ is the restriction of the orthogonal projection in $L^2(-r,r)$ onto $\mathcal{U}$ to the set $L^\infty_G(-r,r) \subset L^2(-r,r)$. As a further preparation for our analysis, we introduce an abbreviation for the L^2-gradient of the reduced objective function F_r of ($\mathrm{P}_{r,\mathcal{U}}$):

Definition 8.3 (L^2-Gradient of F_r)**.** Given a control $0 \le u \in \mathcal{U}$, we define $\nabla_{L^2} F_r(u) \in \mathcal{U}$ to be the function

$$\nabla_{L^2} F_r(u) := \mathcal{P}_{\mathcal{U}}^{L^2}\left(-p_2 + \nu_1 + \nu_2(u - u_D)\right). \tag{8.2}$$

Here, $p_2 \in L^\infty_{BV}(-r,r)$ is characterized by the system (6.20) with $y = S_r(u)$.

Recall that p_2 is indeed an element of $L^\infty_{BV}(-r,r)$ in the situation of Definition 8.3 by Lemma 6.19. In combination with the regularity of u and assumption (A1), this implies that $-p_2 + \nu_1 + \nu_2(u - u_D) \in L^\infty_G(-r,r)$ holds and that the right-hand side of (8.2) is well defined. Due to the properties of $\mathcal{P}_{\mathcal{U}}^{L^2}$ and the identity (6.19), we further have

$$\begin{aligned} F_r'(u)z &= \left(-p_2 + \nu_1 + \nu_2(u - u_D), z\right)_{L^2(-r,r)} \\ &= \left(\mathcal{P}_{\mathcal{U}}^{L^2}\left(-p_2 + \nu_1 + \nu_2(u - u_D)\right), z\right)_{L^2(-r,r)} = \left(\nabla_{L^2} F_r(u), z\right)_{L^2(-r,r)} \qquad \forall z \in \mathcal{U} \qquad \forall\, 0 \le u \in \mathcal{U}, \end{aligned}$$

where $F_r'(u) \in L^\infty_G(-r,r)^*$ denotes the Fréchet derivative of $F_r\colon L^\infty_G(-r,r) \cap U_{P,r} \to \mathbb{R}$ at u in the sense of Corollary 6.18. The function $\nabla_{L^2} F_r(u)$ is thus indeed the L^2-gradient of F_r at u. For later use, we note:

Lemma 8.4 (Boundedness of Adjoint Variables)**.** *There exists a constant $C > 0$ such that, for all $0 \le u \in \mathcal{U}$ with associated adjoint functions $p_1 \in H^1_0(\Omega) \cap H^2(\Omega)$ and $p_2 \in L^\infty_{BV}(-r,r)$ as in (6.20), it holds*

$$\|p_1\|_{H^1(\Omega)} + \|p_2\|_{L^\infty(-r,r)} \le C.$$

Proof. Let $0 \le u \in \mathcal{U}$ with associated $y, p_1 \in H^1_0(\Omega) \cap H^2(\Omega)$ and $p_2 \in L^\infty_{BV}(-r,r)$ as in (6.20) be given. From Lemma 3.5, we obtain that $\|y\|_{L^\infty(\Omega)} \le r_P$ holds. In combination with the PDE characterizing p_1 in (6.20), the fact that $0 \le \mathcal{E}_r(u)(y) \in L^\infty(\Omega)$ holds, and standard a-priori estimates for elliptic equations, this implies that there exists a constant $C_1 > 0$ satisfying $\|p_1\|_{H^1(\Omega)} \le C_1$. Due to the definition of p_2 and the boundedness of Ω, we may now deduce that there exists a constant $C_2 > 0$ satisfying $\|p_2\|_{L^\infty(-r,r)} \le C_2 \|p_1\|_{L^2(\Omega)} \le C_1 C_2$. In summary, this yields $\|p_1\|_{H^1(\Omega)} + \|p_2\|_{L^\infty(-r,r)} \le C_1(1 + C_2)$ which proves the claim. □

Lemma 8.5 (Continuity of Gradients)**.** *Suppose that a sequence $\{u_n\} \subset \mathcal{U}$ satisfying $0 \le u_n \to u$ in $L^\infty(-r,r)$ for some $u \in L^\infty(-r,r)$ is given. Then it holds $0 \le u \in \mathcal{U}$ and $\nabla_{L^2} F_r(u_n) \to \nabla_{L^2} F_r(u)$ in $L^2(-r,r)$.*

Proof. Assume that a sequence $\{u_n\} \subset \mathcal{U}$ as in the assertion of the lemma is given. Then it follows from Corollary 6.6 and the finite-dimensionality of spaces of the form (8.1) that $0 \leq u \in \mathcal{U}$. From Theorem 6.17, Corollary 6.18, and the structure of the objective F_r, we further obtain that

$$(S_r(u_n) - y_D, S_r'(u_n)z)_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 z + \nu_2(u_n - u_D)z \,\mathrm{d}t = F_r'(u_n)z = \big(-p_{2,n} + \nu_1 + \nu_2(u_n - u_D), z\big)_{L^2(-r,r)} \tag{8.3}$$

holds for all n and all $z \in L^\infty_G(-r,r)$, where $S_r'(u_n) \in \mathcal{L}(L^2(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$ denotes the Gâteaux derivative of S_r at u_n in the sense of Theorem 6.17 and $p_{2,n} \in L^\infty_{BV}(-r,r)$ the function that is associated with u_n *via* the system (6.20). Note that, as $S_r'(u_n)$ is a linear and continuous operator from $L^2(-r,r)$ to $H_0^1(\Omega) \cap H^2(\Omega)$ for all n, the equality between the left- and the right-hand side of (8.3) remains true for all test functions $z \in L^2(-r,r)$. Since $S_r'(u_n) \to S_r'(u)$ holds in $\mathcal{L}(L^2(-r,r), H_0^1(\Omega) \cap H^2(\Omega))$ by Theorem 6.17i) and since $S_r(u_n) \to S_r(u)$ holds in $H_0^1(\Omega) \cap H^2(\Omega)$ by Theorem 3.4i), this allows us to deduce that

$$p_{2,n} = -S_r'(u_n)^* \left(S_r(u_n) - y_D\right) \to -S_r'(u)^* \left(S_r(u) - y_D\right) \tag{8.4}$$

holds strongly in $L^2(-r,r)$, where a star denotes an adjoint operator. Using the exact same reasoning as in (8.3), one obtains that the right-hand side of (8.4) is identical to the function $p_2 \in L^\infty_{BV}(-r,r)$ associated with u *via* the system (6.20). We thus have $p_{2,n} \to p_2$ in $L^2(-r,r)$. The assertion of the lemma now follows immediately from the definition of the L^2-gradient of F_r and the properties of $\mathcal{P}_{\mathcal{U}}^{L^2}$. This completes the proof. $\square$

Using the L^2-gradient, we can formulate the first-order necessary optimality condition of ($\mathrm{P}_{r,\mathcal{U}}$) as follows:

Theorem 8.6 (First-Order Necessary Optimality Condition for ($\mathrm{P}_{r,\mathcal{U}}$))**.** *Suppose that $0 \leq \bar{u} \in \mathcal{U}$ is a local solution of ($\mathrm{P}_{r,\mathcal{U}}$). Then it holds*

$$\nabla_{L^2} F_r(\bar{u}) \begin{cases} \geq 0 & a.e. \text{ in } \{\bar{u} = 0\}, \\ = 0 & a.e. \text{ in } \{\bar{u} > 0\}. \end{cases} \tag{8.5}$$

Proof. From Corollary 6.18 and Definition 8.3, we obtain (analogously to the proof of Theorem 7.4) that $\bar{u} \in \mathcal{U}$ satisfies

$$\bar{u} \in \mathcal{U} \cap L^2_+(-r,r), \qquad (\nabla_{L^2} F_r(\bar{u}), u - \bar{u})_{L^2(-r,r)} \geq 0 \quad \forall u \in \mathcal{U} \cap L^2_+(-r,r).$$

Again analogously to the proof of Theorem 7.4, one checks that, for both the case $\mathcal{U} = L^\infty_G(-r,r)$ and the case where $\mathcal{U}$ is of the form (8.1), the above variational inequality implies $\bar{u} = \max\left(0, \bar{u} - \nabla_{L^2} F_r(\bar{u})\right)$ a.e. in $(-r,r)$. Using a distinction of cases, (8.5) now follows immediately. $\square$

Motivated by the necessary optimality condition (8.5), we introduce the following measures for the degree of stationarity of an admissible control:

Definition 8.7 (Stationarity Indicators)**.** Given a control $0 \leq u \in \mathcal{U}$ and a number $\epsilon > 0$, we define

$$\Theta_0(u) := \left\|\nabla_{L^2} F_r(u) \mathbb{1}_{\{u>0\}} + \min\left(0, \nabla_{L^2} F_r(u)\right) \mathbb{1}_{\{u=0\}}\right\|_{L^2(-r,r)}$$

and

$$\Theta_\epsilon(u) := \left(\int_{-r}^{r} \min\left(\frac{u}{\epsilon}, \nabla_{L^2} F_r(u)\right) \nabla_{L^2} F_r(u) \,\mathrm{d}t\right)^{1/2}.$$

Here, the min-function appearing in the integral/the norm acts by superposition.

Using the definitions of Θ_0 and Θ_ϵ, one easily establishes:

Lemma 8.8 (Properties of the Stationarity Measures)**.** *Let $0 \leq u \in \mathcal{U}$ be given. Then the following is true:*

i) u satisfies the first-order condition (8.5) if and only if $\Theta_0(u) = 0$;
ii) u satisfies the first-order condition (8.5) if and only if $\Theta_\epsilon(u) = 0$ for one (and thus all) $\epsilon > 0$;
iii) It holds $0 \leq \Theta_{\epsilon_1}(u) \leq \Theta_{\epsilon_2}(u) \leq \Theta_0(u)$ for all $\epsilon_1 \geq \epsilon_2 > 0$;
iv) It holds $\Theta_\epsilon(u) \to \Theta_0(u)$ for $\epsilon \to 0$.

Proof. Point i) is trivial. To prove the remaining assertions, we note that the nonnegativity of u implies that, for every $\epsilon > 0$, we have

$$\begin{aligned}
&\min\left(\frac{u}{\epsilon}, \nabla_{L^2} F_r(u)\right) \nabla_{L^2} F_r(u) \\
&\quad = \min\left(\frac{u}{\epsilon}, \nabla_{L^2} F_r(u)\right) \min(0, \nabla_{L^2} F_r(u)) + \min\left(\frac{u}{\epsilon}, \nabla_{L^2} F_r(u)\right) \max(0, \nabla_{L^2} F_r(u)) \\
&\quad = \min(0, \nabla_{L^2} F_r(u))^2 + \min\left(\frac{u}{\epsilon}, \max(0, \nabla_{L^2} F_r(u))\right) \max(0, \nabla_{L^2} F_r(u)) \quad \text{a.e. in } (-r, r)
\end{aligned}$$

and

$$\begin{aligned}
0 &\leq \min(0, \nabla_{L^2} F_r(u))^2 + \min\left(\frac{u}{\epsilon}, \max(0, \nabla_{L^2} F_r(u))\right) \max(0, \nabla_{L^2} F_r(u)) \\
&\leq \min(0, \nabla_{L^2} F_r(u))^2 + \mathbb{1}_{\{u>0\}} \max(0, \nabla_{L^2} F_r(u))^2 \quad \text{a.e. in } (-r, r).
\end{aligned}$$

Using the above identities and inequalities, integration, the dominated convergence theorem, and the definitions of Θ_0 and Θ_ϵ, the assertions in ii), iii), and iv) follow immediately. This completes the proof. $\square$

The stationarity measures Θ_ϵ, $\epsilon > 0$, are advantageous from both the analytical and the numerical point of view because they are more stable w.r.t. small perturbations of u than Θ_0. Compare also with the convergence analysis of the standard gradient projection method in finite dimensions in this context; see [43, 44]. We are now in the position to formulate the algorithm that we use for the numerical solution of the problem ($\mathrm{P}_{r,\mathcal{U}}$).

Algorithm 8.9 (Gradient Projection Method for the Solution of ($\mathrm{P}_{r,\mathcal{U}}$))**.**

```
1:  Choose an initial guess 0 ≤ u_0 ∈ U and parameters ε_1, ε_2 > 0, σ > 0, and ω, θ ∈ (0, 1).
2:  for i = 0, 1, 2, 3, ... do
3:      Calculate ∇_{L^2} F_r(u_i) ∈ U by means of the system (6.20) and formula (8.2).
4:      if Θ_{ε_1}(u_i) ≤ ε_2 then
5:          break
6:      end if
7:      Initialize σ_0 := σ and calculate a step size as follows:
8:      for j = 0, 1, 2, 3, ... do
9:          Calculate the quantity e_j := F_r(u_i) − F_r(max(0, u_i − σ_j ∇_{L^2} F_r(u_i))) − σ_j θ Θ_{ε_1}(u_i)^2.
10:         if e_j < 0 then
11:             Define σ_{j+1} := ω σ_j.
12:         else
13:             Define τ_i := σ_j and break
14:         end if
15:     end for
16:     Define u_{i+1} := max(0, u_i − τ_i ∇_{L^2} F_r(u_i)).
17: end for
```

We emphasize that the max-functions in steps 9 and 16 of Algorithm 8.9 again act by superposition. The next two lemmas show that Algorithm 8.9 is sensible.

Lemma 8.10 (Finite Termination of the Line-Search). *Suppose that numbers $\epsilon_1, \epsilon_2 > 0$, $\sigma > 0$, and $\omega, \theta \in (0,1)$ and a function $0 \leq u_i \in \mathcal{U}$ satisfying $\Theta_{\epsilon_1}(u_i) > \epsilon_2$ are given. Then the line-search procedure in lines 7 to 15 of Algorithm 8.9 terminates after finitely many steps with a step size $\tau_i > 0$.*

Proof. Suppose that a function $0 \leq u_i \in \mathcal{U}$ satisfying $\Theta_{\epsilon_1}(u_i) > \epsilon_2$ is given and that the line-search does not terminate after finitely many steps. Then it holds $e_j < 0$ for all $j \in \mathbb{N}_0$ and, as a consequence,

$$\frac{F_r(\max(0, u_i - \sigma_j \nabla_{L^2} F_r(u_i))) - F_r(u_i)}{\sigma_j} > -\theta\Theta_{\epsilon_1}(u_i)^2 \geq -\theta\Theta_0(u_i)^2 \qquad \forall j \in \mathbb{N}_0$$

with $0 < \sigma_j \to 0$ for $j \to \infty$. Here, the inequality involving $\Theta_0(u_i)$ follows from Lemma 8.8iii). Let us define

$$z_j := \frac{\max(0, u_i - \sigma_j \nabla_{L^2} F_r(u_i)) - u_i}{\sigma_j} = \frac{\max(0, u_i - \sigma_j \nabla_{L^2} F_r(u_i)) - \max(0, u_i)}{\sigma_j}, \qquad j \in \mathbb{N}_0. \tag{8.6}$$

Using Lebesgue's dominated convergence theorem and the directional differentiability and global Lipschitz continuity of the function $\mathbb{R} \ni s \mapsto \max(0, s) \in \mathbb{R}$, it is easy to check that $z_j \to z$ holds in $L^2(-r, r)$ with

$$z := -\nabla_{L^2} F_r(u_i) \mathbb{1}_{\{u_i > 0\}} + \max\left(0, -\nabla_{L^2} F_r(u_i)\right) \mathbb{1}_{\{u_i = 0\}}. \tag{8.7}$$

In view of Theorem 6.9 and Proposition 6.16, this implies that

$$\frac{S_r(u_i + \sigma_j z_j) - S_r(u_i)}{\sigma_j} \to S_r'(u_i) z$$

holds in $H^2(\Omega)$ for $j \to \infty$, where $S_r'(u_i) \in \mathcal{L}(L^2(-r, r), H_0^1(\Omega) \cap H^2(\Omega))$ denotes the Hadamard-Gâteaux derivative of S_r at u_i in the sense of Proposition 6.16. Due to the structure of F_r, it follows that

$$\begin{aligned}
-\theta\Theta_0(u_i)^2 &\leq \frac{F_r(\max(0, u_i - \sigma_j \nabla_{L^2} F_r(u_i))) - F_r(u_i)}{\sigma_j} \\
&= \frac{F_r(u_i + \sigma_j z_j) - F_r(u_i)}{\sigma_j} \\
&\to (S_r(u_i) - y_D, S_r'(u_i) z)_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 z + \nu_2 (u_i - u_D) z \,\mathrm{d}t \qquad \text{for } j \to \infty.
\end{aligned}$$

Note that, for all $\tilde{z} \in L^\infty_G(-r, r)$, the Fréchet differentiability properties in Theorem 6.17 and Corollary 6.18 imply that

$$(S_r(u_i) - y_D, S_r'(u_i)\tilde{z})_{L^2(\Omega)} + \int_{-r}^{r} \nu_1 \tilde{z} + \nu_2 (u_i - u_D) \tilde{z} \,\mathrm{d}t = F_r'(u_i)\tilde{z} = \left(-p_{2,i} + \nu_1 + \nu_2(u_i - u_D), \tilde{z}\right)_{L^2(-r,r)} \tag{8.8}$$

with $p_{2,i} \in L^\infty_{BV}(-r, r)$ defined *via* the system (6.20) for u_i. As $S_r'(u_i) \in \mathcal{L}(L^2(-r, r), H_0^1(\Omega) \cap H^2(\Omega))$ holds, by approximation, we obtain that the equality between the left- and the right-hand side of (8.8) remains true for all $\tilde{z} \in L^2(-r, r)$. In particular, we have

$$-\theta\Theta_0(u_i)^2 \leq \left(-p_{2,i} + \nu_1 + \nu_2(u_i - u_D), z\right)_{L^2(-r,r)}.$$

In the case $\mathcal{U} = L^\infty_G(-r,r)$, the above inequality, the definitions of the functions z and $\nabla_{L^2} F_r(u_i)$, and the identity $-p_{2,i} + \nu_1 + \nu_2(u_i - u_D) = \mathcal{P}_{\mathcal{U}}^{L^2}(-p_{2,i} + \nu_1 + \nu_2(u_i - u_D))$ imply that

$$-\theta\Theta_0(u_i)^2 \leq (\nabla_{L^2} F_r(u_i), z)_{L^2(-r,r)} = -\|z\|^2_{L^2(-r,r)} = -\Theta_0(u_i)^2. \tag{8.9}$$

In the case where $\mathcal{U}$ is a space of piecewise constant finite element functions as in (8.1), it follows from (8.7) that $z \in \mathcal{U}$ holds, and we may exploit the properties of $\mathcal{P}_{\mathcal{U}}^{L^2}$ and argue analogously to (8.9) to obtain that

$$-\theta\Theta_0(u_i)^2 \leq \left(-p_{2,i} + \nu_1 + \nu_2(u_i - u_D), z\right)_{L^2(-r,r)} = (\nabla_{L^2} F_r(u_i), z)_{L^2(-r,r)} = -\|z\|^2_{L^2(-r,r)} = -\Theta_0(u_i)^2.$$

In both of the above cases, one arrives at the inequality $-\theta\Theta_0(u_i)^2 \leq -\Theta_0(u_i)^2$ which is impossible due to the assumptions $\Theta_{\epsilon_1}(u_i) > \epsilon_2 > 0$ and $\theta \in (0,1)$ and the inequality $\Theta_{\epsilon_1}(u_i) \leq \Theta_0(u_i)$. The line-search algorithm thus has to terminate after finitely many steps as claimed. ☐

Note that it is necessary to take the detour *via* Theorem 6.9 and Proposition 6.16 in the proof of Lemma 8.10 as the sequence $\{z_j\}$ in (8.6) only converges in $L^q(-r,r)$ for all $q \in [1,\infty)$ but *not* in $L^\infty(-r,r)$ in general (as one may easily check by means of examples). This makes it impossible to invoke the Fréchet differentiability result in, *e.g.*, Corollary 6.18 directly. Next, we clarify that the update of u_i in Algorithm 8.9 is sensible.

Lemma 8.11 (Well-Definedness of Iterates)**.** *Suppose that a function $0 \leq u_i \in \mathcal{U}$ with associated gradient $\nabla_{L^2} F_r(u_i) \in \mathcal{U}$ and a number $\tau_i > 0$ are given. Then $u_{i+1} := \max(0, u_i - \tau_i \nabla_{L^2} F_r(u_i))$ satisfies $0 \leq u_{i+1} \in \mathcal{U}$.*

Proof. From (8.2) and the comments after Definition 8.3, we obtain that $\nabla_{L^2} F_r(u_i) \in \mathcal{U}$. This also implies that $u_i - \tau_i \nabla_{L^2} F_r(u_i) \in \mathcal{U}$. In the case $\mathcal{U} = L^\infty_G(-r,r)$, the inclusion $u_{i+1} = \max(0, u_i - \tau_i \nabla_{L^2} F_r(u_i)) \in \mathcal{U}$ now follows from the fact that the positive part of a regulated function is again regulated. In the case where $\mathcal{U}$ is a finite element space as in (8.1), it is trivially true that the Nemytskii operator induced by $\mathbb{R} \ni s \mapsto \max(0,s) \in \mathbb{R}$ maps $\mathcal{U}$ into itself, so $u_{i+1} \in \mathcal{U}$ again follows. Since we trivially have $u_{i+1} \geq 0$, this completes the proof. ☐

We emphasize that the above lemma is crucial as it shows that updates with the $L^2(-r,r)$-gradient do not cause the sequence of iterates $\{u_i\}$ to leave the set $L^\infty_G(-r,r) \cap U_{P,r}$ where the derivatives of the functions S_r and F_r are known to exist. Having checked that Algorithm 8.9 is sensible, we can now turn our attention to the convergence analysis. We start with:

Lemma 8.12 (Descent Behavior)**.** *Suppose that the termination criterion in line 4 of Algorithm 8.9 is not triggered in iteration $i \in \mathbb{N}_0$. Then it holds*

$$F_r(u_i) \geq F_r(u_{i+1}) + \tau_i \theta \Theta_{\epsilon_1}(u_i)^2 \geq F_r(u_{i+1}). \tag{8.10}$$

Proof. This follows immediately from the line-search procedure in Algorithm 8.9 and Lemma 8.8. ☐

We are now in the position to prove our main convergence result for Algorithm 8.9.

Theorem 8.13 (Finite Convergence of the Gradient Projection Method)**.** *Algorithm 8.9 – executed with arbitrary parameters $\epsilon_1, \epsilon_2, \sigma > 0$ and $\omega, \theta \in (0,1)$ and with an arbitrary initial guess $0 \leq u_0 \in \mathcal{U}$ – terminates after finitely many iterations with a final iterate u_i that satisfies $\Theta_{\epsilon_1}(u_i) \leq \epsilon_2$.*

Proof. We argue by contradiction: Suppose that Algorithm 8.9 does not terminate after finitely many steps. Then it produces a sequence of iterates $0 \leq u_i \in \mathcal{U}$ which satisfies $F_r(u_i) \geq F_r(u_{i+1}) + \tau_i \theta \Theta_{\epsilon_1}(u_i)^2$ and $\Theta_{\epsilon_1}(u_i) > \epsilon_2$ for all $i \in \mathbb{N}_0$ due to the termination criterion in line 4 of Algorithm 8.9 and Lemma 8.12. As $F_r(u_i) \geq 0$ holds for all $i \in \mathbb{N}_0$ by the definition of F_r and the nonnegativity of u_i, this yields

$$F_r(u_0) \geq F_r(u_n) + \sum_{i=0}^{n-1} \tau_i \theta \Theta_{\epsilon_1}(u_i)^2 \geq \sum_{i=0}^{n-1} \tau_i \theta \epsilon_2^2 \qquad \forall n \in \mathbb{N}_0$$

and, thus, $\sum_{i=0}^\infty \tau_i < \infty$.

We now first prove that the sequence of iterates $\{u_i\} \subset \mathcal{U}$ is bounded in $L^\infty(-r,r)$. To this end, we note that the update formula in line 16 of Algorithm 8.9, Definition 8.3, and the properties of $\mathcal{P}_{\mathcal{U}}^{L^2}$ imply that

$$\begin{aligned}\|u_{i+1}\|_{L^\infty(-r,r)} &= \|\max(0, u_i - \tau_i \nabla_{L^2} F_r(u_i))\|_{L^\infty(-r,r)} \\ &\leq \left\| u_i - \tau_i \mathcal{P}_{\mathcal{U}}^{L^2}\left(-p_{2,i} + \nu_1 + \nu_2(u_i - u_D)\right)\right\|_{L^\infty(-r,r)} \\ &= \left\|(1-\nu_2\tau_i)u_i - \tau_i \mathcal{P}_{\mathcal{U}}^{L^2}\left(-p_{2,i} + \nu_1 - \nu_2 u_D\right)\right\|_{L^\infty(-r,r)}\end{aligned}$$

holds for all $i \in \mathbb{N}_0$. Here, $p_{2,i} \in L^\infty_{BV}(-r,r)$ is again defined *via* the system (6.20) for u_i. Since $\sum_{i=0}^\infty \tau_i < \infty$ holds, we know that $\tau_i > 0$ converges to zero. This implies that there exists an index $m \in \mathbb{N}$ such that $|1-\nu_2\tau_i| \leq 1$ holds for all $i \geq m$. Using the fact that $\mathcal{P}_{\mathcal{U}}^{L^2}$ is just the identity map in the case $\mathcal{U} = L^\infty_G(-r,r)$ and given by the cell-wise averaging operator

$$\mathcal{P}_{\mathcal{U}}^{L^2}(v) = \frac{1}{r_n - r_{n-1}} \int_{r_{n-1}}^{r_n} v(s)\,\mathrm{d}s \qquad \text{a.e. in } (r_{n-1}, r_n) \qquad \forall n = 1, ..., N \qquad \forall v \in L^\infty_G(-r,r)$$

in the case where $\mathcal{U}$ is a finite element space of the form (8.1) (as one may easily check), we further obtain that

$$\left\|\mathcal{P}_{\mathcal{U}}^{L^2}(v)\right\|_{L^\infty(-r,r)} \leq \|v\|_{L^\infty(-r,r)} \tag{8.11}$$

holds for all $v \in L^\infty_G(-r,r)$. In combination with Lemma 8.4, all of this allows us to conclude that

$$\begin{aligned}\|u_{i+1}\|_{L^\infty(-r,r)} &\leq \left\|(1-\nu_2\tau_i)u_i - \tau_i \mathcal{P}_{\mathcal{U}}^{L^2}\left(-p_{2,i} + \nu_1 - \nu_2 u_D\right)\right\|_{L^\infty(-r,r)} \\ &\leq |1-\nu_2\tau_i|\,\|u_i\|_{L^\infty(-r,r)} + \tau_i \left\|-p_{2,i} + \nu_1 - \nu_2 u_D\right\|_{L^\infty(-r,r)} \\ &\leq \|u_i\|_{L^\infty(-r,r)} + C\tau_i \qquad \forall i \geq m\end{aligned}$$

holds with some constant $C > 0$. After a trivial induction, the above yields

$$\|u_i\|_{L^\infty(-r,r)} \leq \|u_m\|_{L^\infty(-r,r)} + C\sum_{j=0}^\infty \tau_j < \infty \qquad \forall i \geq m.$$

The sequence of iterates $\{u_i\} \subset \mathcal{U}$ is thus indeed bounded in $L^\infty(-r,r)$ as claimed.

Next, we show that the sequence $\{u_i\} \subset \mathcal{U}$ even converges in $L^\infty(-r,r)$ to a limit function $0 \leq \bar{u} \in \mathcal{U}$. To this end, we note that the update formula in line 16 of Algorithm 8.9, the global Lipschitz continuity of the function $\mathbb{R} \ni s \mapsto \max(0,s) \in \mathbb{R}$, the nonnegativity of the iterates u_i, Definition 8.3, the boundedness of the sequence $\{u_i\}$ in $L^\infty(-r,r)$, Lemma 8.4, and again the stability estimate (8.11) imply that there exists a constant $C > 0$ satisfying

$$\begin{aligned}\|u_{i+1} - u_i\|_{L^\infty(-r,r)} &= \|\max(0, u_i - \tau_i \nabla_{L^2} F_r(u_i)) - \max(0, u_i)\|_{L^\infty(-r,r)} \\ &\leq \|\tau_i \nabla_{L^2} F_r(u_i)\|_{L^\infty(-r,r)} \\ &\leq \tau_i \left\|\mathcal{P}_{\mathcal{U}}^{L^2}\left(-p_{2,i} + \nu_1 + \nu_2(u_i - u_D)\right)\right\|_{L^\infty(-r,r)} \\ &\leq C\tau_i \qquad \forall i \in \mathbb{N}_0,\end{aligned} \tag{8.12}$$

where $p_{2,i} \in L^\infty_{BV}(-r,r)$ is defined as before. Due to the estimate $\sum_{i=0}^\infty \tau_i < \infty$, the above shows that

$$\|u_m - u_n\|_{L^\infty(-r,r)} \le C \sum_{i=n}^\infty \tau_i \qquad \forall m,n \in \mathbb{N},\ m \ge n,$$

holds with $\sum_{i=n}^\infty \tau_i \to 0$ for $n \to \infty$. The sequence $\{u_i\} \subset \mathcal{U}$ is thus Cauchy in $L^\infty(-r,r)$. Since, for both the case $\mathcal{U} = L^\infty_G(-r,r)$ and the case where $\mathcal{U}$ is of the form (8.1), $(\mathcal{U}, \|\cdot\|_{L^\infty(-r,r)})$ is complete (see Cor. 6.6), it now follows that there indeed exists $\bar u \in \mathcal{U}$ such that $u_i \to \bar u$ holds in $L^\infty(-r,r)$ for $i \to \infty$ as claimed. That this limit $\bar u$ is nonnegative is obvious.

Having established that $\{u_i\} \subset \mathcal{U}$ satisfies $0 \le u_i \to \bar u$ in $L^\infty(-r,r)$ for a function $0 \le \bar u \in \mathcal{U}$, we can now produce a contradiction as desired: Since $\tau_i \to 0$ holds, we know that there exists $n \in \mathbb{N}$ such that $\tau_i < \sigma$ and $\tau_i/\omega < \epsilon_1$ holds for all $i \ge n$. In view of the line-search procedure in lines 7 to 15 of Algorithm 8.9, this implies that the step size τ_i/ω was rejected in line 10 and, as a consequence, that

$$F_r\left(\max\left(0, u_i - \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right) - F_r(u_i) > -\frac{\tau_i}{\omega}\theta\Theta_{\epsilon_1}(u_i)^2 \ge -\frac{\tau_i}{\omega}\theta\Theta_{\tau_i/\omega}(u_i)^2 \qquad \forall i \ge n, \tag{8.13}$$

where the last estimate follows from Lemma 8.8iii) and the inequality $\tau_i/\omega < \epsilon_1$. As $F_r \colon L^\infty_G(-r,r) \cap U_{P,r} \to \mathbb{R}$ is continuously Fréchet differentiable on $L^\infty_G(-r,r) \cap U_{P,r}$ by Corollary 6.18, we can apply the mean value theorem [42], Theorem 3.2.6 and Definition 8.3 to rewrite (8.13) as

$$\begin{aligned}
&-\theta\Theta_{\tau_i/\omega}(u_i)^2 \\
&\le \frac{\omega}{\tau_i}\left(F_r\left(\max\left(0, u_i - \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right)\right) - F_r(u_i)\right) \\
&= \int_0^1 F_r'\left((1-s)u_i + s\max\left(0, u_i - \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right)\right)\left(\frac{\max\left(0, u_i - \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right) - u_i}{\tau_i/\omega}\right)\mathrm{d}s \\
&= -\int_0^1 \left(\nabla_{L^2}F_r\left((1-s)u_i + s\max\left(0, u_i - \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right)\right), \min\left(\frac{\omega}{\tau_i}u_i, \nabla_{L^2}F_r(u_i)\right)\right)_{L^2(-r,r)}\mathrm{d}s \\
&= -\left(\nabla_{L^2}F_r(u_i), \min\left(\frac{\omega}{\tau_i}u_i, \nabla_{L^2}F_r(u_i)\right)\right)_{L^2(-r,r)} + \varrho_i \\
&= -\Theta_{\tau_i/\omega}(u_i)^2 + \varrho_i \qquad \forall i \ge n,
\end{aligned} \tag{8.14}$$

where the last line holds by Definition 8.7 and where ϱ_i denotes the remainder term

$$\varrho_i := \int_0^1 \left(\nabla_{L^2}F_r(u_i) - \nabla_{L^2}F_r\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right)\right), \min\left(\frac{\omega}{\tau_i}u_i, \nabla_{L^2}F_r(u_i)\right)\right)_{L^2(-r,r)}\mathrm{d}s. \tag{8.15}$$

Note that, due to the convergence $u_i \to \bar u$ in $L^\infty(-r,r)$, the estimates in (8.12), and the nonnegativity of u_i, we know that

$$\begin{aligned}
\left\|u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right) - \bar u\right\|_{L^\infty(-r,r)} &\le \|u_i - \bar u\|_{L^\infty(-r,r)} + \left\|\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2}F_r(u_i)\right)\right\|_{L^\infty(-r,r)} \\
&\le \|u_i - \bar u\|_{L^\infty(-r,r)} + \frac{1}{\omega}\|\tau_i\nabla_{L^2}F_r(u_i)\|_{L^\infty(-r,r)} \\
&\to 0 \qquad \text{for } i \to \infty \qquad \forall s \in [0,1].
\end{aligned} \tag{8.16}$$

In combination with Lemma 8.5, this implies that

$$\nabla_{L^2} F_r\left(u_i\right) - \nabla_{L^2} F_r\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right) \to \nabla_{L^2} F_r\left(\bar{u}\right) - \nabla_{L^2} F_r\left(\bar{u}\right) = 0$$

holds in $L^2(-r,r)$ for all $s \in [0,1]$ as i goes to infinity. By exploiting the formula (8.2), the stability estimate (8.11), Lemma 8.4, the boundedness of $\{u_i\}$ in $L^\infty(-r,r)$, and the same estimates as in (8.16), we further obtain that there exist constants $C_1, C_2, C_3 > 0$ independent of i and $s \in [0,1]$ satisfying

$$\begin{aligned}
&\left\|\nabla_{L^2} F_r\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right)\right\|_{L^\infty(-r,r)} \\
&\quad\le C_1 + \left\|\nu_1 + \nu_2\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right) - \nu_2 u_D\right\|_{L^\infty(-r,r)} \\
&\quad\le C_2 + \nu_2\left\|\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right\|_{L^\infty(-r,r)} \\
&\quad\le C_3 \qquad \forall s \in [0,1] \qquad \forall i \in \mathbb{N}.
\end{aligned}$$

Again due to the nonnegativity of the iterates u_i, we may now conclude that there exist constants $C_4, C_5, C_6 > 0$ independent of $s \in [0,1]$ and $i \in \mathbb{N}$ such that

$$\begin{aligned}
&\left|\left(\nabla_{L^2} F_r\left(u_i\right) - \nabla_{L^2} F_r\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right), \min\left(\frac{\omega}{\tau_i}u_i, \nabla_{L^2} F_r(u_i)\right)\right)_{L^2(-r,r)}\right| \\
&\le C_4\left\|\min\left(\frac{\omega}{\tau_i}u_i, \nabla_{L^2} F_r(u_i)\right)\right\|_{L^2(-r,r)} \\
&\le C_5
\end{aligned}$$

holds for all $s \in [0,1]$ and all $i \in \mathbb{N}$ and such that

$$\begin{aligned}
&\left|\left(\nabla_{L^2} F_r\left(u_i\right) - \nabla_{L^2} F_r\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right), \min\left(\frac{\omega}{\tau_i}u_i, \nabla_{L^2} F_r(u_i)\right)\right)_{L^2(-r,r)}\right| \\
&\le C_6\left\|\nabla_{L^2} F_r\left(u_i\right) - \nabla_{L^2} F_r\left(u_i - s\min\left(u_i, \frac{\tau_i}{\omega}\nabla_{L^2} F_r(u_i)\right)\right)\right\|_{L^2(-r,r)} \\
&\to 0
\end{aligned}$$

holds for all $s \in [0,1]$ as i goes to infinity. If we combine all of the above with (8.15) and invoke the dominated convergence theorem, then it follows that $\varrho_i \to 0$ holds for $i \to \infty$. Due to the estimate in (8.14) and the fact that $\epsilon_2^2 \le \Theta_{\epsilon_1}(u_i)^2 \le \Theta_{\tau_i/\omega}(u_i)^2$ holds for all large enough i (*cf.* (8.13)), this yields

$$0 < (1-\theta)\epsilon_2^2 \le (1-\theta)\Theta_{\tau_i/\omega}(u_i)^2 \le \varrho_i \to 0 \qquad \text{for } i \to \infty,$$

which is impossible. Algorithm 8.9 thus has to terminate after finitely many iterations as claimed in the theorem and the proof is complete. ☐

We would like to point out that – for the control space $\mathcal{U} = L^\infty_G(-r,r)$ – it is crucial to establish the existence of the limit $\bar{u}$ in the proof of Theorem 8.13 *via* the summability condition $\sum_{i=0}^\infty \tau_i < \infty$ that is obtained in the case of nonfinite convergence from the sufficient decrease condition (8.10). Indeed, only by exploiting this summability, we are able to prove that the sequence $\{u_i\}$ is convergent in $L^\infty(-r,r)$ and, thus, convergent in a topology for which continuity results for the gradient are available by Lemma 8.5. If one tries, for example, to exploit that the coercivity properties of the objective function F_r imply that sequences of iterates generated by a gradient descent method for the numerical solution of ($\mathrm{P}_{r,\mathcal{U}}$) possess weak accumulation points in $L^2(-r,r)$, then this turns out to be not very useful since it is completely unclear how to establish that these accumulation points are stationary in a meaningful sense if the space $\mathcal{U}$ is infinite-dimensional. (For a finite-dimensional space $\mathcal{U}$ of the type (8.1), these issues related to the topologies are, of course, completely irrelevant.)

9. Numerical experiments

We conclude this paper with numerical experiments that validate the analytical results of Sections 3 to 8. As examples, we consider two different instances of the problem

$$\left.\begin{aligned} \text{Minimize}\quad & \frac{1}{2}\|S_r(u)-y_D\|^2_{L^2(\Omega)}+\int_{-r}^{r}\nu_1 u+\frac{\nu_2}{2}(u-u_D)^2\,\mathrm{d}t\\ \text{w.r.t.}\quad & u\in L^\infty_G(-r,r),\\ \text{s.t.}\quad & u\geq 0 \text{ a.e. in } (-r,r), \end{aligned}\right\}\tag{9.1}$$

studied in Section 8.

9.1. Example 1: a test case with a known analytical solution

In our first test configuration, the parameters in (9.1) are chosen as

$$\begin{gathered} r=3,\qquad \Omega=(0,1)^2,\qquad \nu_1=0,\qquad \nu_2=10^{-3},\\ y_D(x_1,x_2)=\sin(2\pi x_1)\sin(2\pi x_2)\text{ f.a.a. }(x_1,x_2)^\top\in\Omega,\\ u_D(s)=1\text{ f.a.a. }s\in(-r,r),\\ f(x_1,x_2)=(8\pi^2+1)\sin(2\pi x_1)\sin(2\pi x_2)\text{ f.a.a. }(x_1,x_2)^\top\in\Omega. \end{gathered}\tag{9.2}$$

Note that, for the above choice of f, the solution y_P of the Poisson problem (3.3) is given by

$$y_P(x_1,x_2)=\frac{(8\pi^2+1)}{8\pi^2}\sin(2\pi x_1)\sin(2\pi x_2)\text{ f.a.a. }(x_1,x_2)^\top\in\Omega.$$

In particular, we have $r_P=2\|y_P\|_\infty=(8\pi^2+1)/(4\pi^2)\approx 2.0253$ and, as a consequence, $r\geq r_P$. In combination with Theorems 4.5 and 5.6, the $L^\infty_G(-r,r)$-regularity of u_D, and the fact that the right-hand side f in (9.2) satisfies (A2) by [39], Proposition 1, this shows that (9.1) can be reformulated as a problem of the type (P_r) or (P), respectively, (with $u_D\in L^2(\mathbb{R})$ as an arbitrary extension of u_D to the whole real line) and that the setting (9.2) is covered by the results of Sections 3 to 8. Using the formulas for y_D, u_D, and f, one further easily checks that (9.1) possesses the unique global solution $\bar u=u_D\equiv 1$ in the situation of (9.2) and that the optimal value of (9.1) is zero. A known analytical solution is thus indeed available under the assumptions of (9.2). We remark that the configuration in (9.2) is a rather exceptional one since, for the chosen u_D and y_D, the adjoint variables $\bar p_1$ and $\bar p_2$ in the stationarity system (7.4) satisfied by $\bar u$ and $\bar y$ are zero and the coefficient in front of the Dirac measure in the formula for the distributional derivative of $\bar p_2$ vanishes; see (6.24). This is a consequence of the construction that we have used in (9.2) to obtain an instance of (9.1) with a known analytical solution.

9.2. Example 2: a test case without a known analytical solution

In our second example, we suppose that the quantities in (9.1) are given by

$$
\begin{gathered}
r = 2, \qquad \Omega = (0,1)^2, \qquad \nu_1 \in \left\{0, \frac{1}{4096}, \frac{1}{2048}, \frac{1}{1024}, \frac{1}{512}, \frac{1}{256}, \frac{1}{128}\right\}, \qquad \nu_2 = 10^{-4},\\
y_D(x_1,x_2) = -0.125 + 0.275\,\sin(2\pi x_1)\sin(2\pi x_2) \text{ f.a.a. } (x_1,x_2)^\top \in \Omega,\\
u_D(s) = 0 \text{ f.a.a. } s \in (-r,r),\\
f(x_1,x_2) = 8\pi^2 \sin(2\pi x_1)\sin(2\pi x_2) \text{ f.a.a. } (x_1,x_2)^\top \in \Omega.
\end{gathered}
\tag{9.3}
$$

For this configuration, one easily checks that $r_P = r = 2$ holds, that u_D is an element of $L^\infty_G(-r,r)$, and that f satisfies condition (A2). The problem (9.1) can thus again be recast in the form (P_r) or (P), respectively, (with $u_D = 0 \in L^2(\mathbb{R})$) and all of the results in Sections 3 to 8 are applicable. To our best knowledge, an analytical solution is not available for the problem (9.1) in the situation of (9.3).

9.3. Details on the numerical realization

To be able to solve the problem (9.1) numerically, one has to discretize the control space $L^\infty_G(-r,r)$ and the governing PDE (3.7). For the discretization of the controls u, we consider piecewise constant finite element functions on an equidistant mesh of width $h_u = (2r)/N_u$, $N_u \in \mathbb{N}$, *i.e.*, we replace $L^\infty_G(-r,r)$ with the space

$$
\mathcal{U}_{h_u} := \{u \in L^\infty(-r,r) \mid u = \text{const a.e. on } ((n-1)h_u - r, nh_u - r)\ \forall n = 1,\dots,N_u\} \subset L^\infty_G(-r,r).
$$

Note that, if we discretize the controls u in (9.1) along the above lines and keep the exact PDE, then the resulting semidiscrete minimization problem still satisfies the assumptions of Section 8; *cf.* (8.1). This means that the convergence result established for Algorithm 8.9 in Theorem 8.13 applies to (9.1) in both the fully continuous and the control-discrete (but state-continuous) setting. For the discretization of the PDE, this is different. In fact, it is easy to check that the results in Sections 6 to 8 *do not* carry over to the state-discretized setting; see the arguments used in the proof of Lemma 6.13. At the end of the day, this means that the algorithm that is run in practice in the numerical experiments presented below should not be interpreted as a gradient descent algorithm for a control- and state-discrete version of (9.1) but rather as a gradient descent algorithm for a control-discrete and state-continuous version of (9.1) that is executed with inexact gradient and function value evaluations where the inexactness arises from the discretization of the governing PDE (*first optimize – then discretize*; *cf.* [45–47]). As we will see in Section 9.4, these peculiarities of the discretization are also reflected by the numerical results. To keep it simple, in what follows, we restrict our attention to the case where the state y and the governing PDE are discretized by means of globally continuous and piecewise affine finite element functions on Friedrichs-Keller triangulations of width $h_y = 1/N_y$, $N_y \in \mathbb{N}$. For the solution of the state equation and the PDE in (6.20), that has to be solved to evaluate the L^2-gradient $\nabla_{L^2} F_r(u_i) \in \mathcal{U}_{h_u}$ in step 3 of Algorithm 8.9, we use the exact same scheme as in [24, 40], *i.e.*, we employ mass lumping in the terms associated with the semilinearity and solve the resulting nonlinear system of equations with a semismooth Newton method. We remark that a very delicate point in the numerical realization of Algorithm 8.9 is the computation of the integrals appearing in (6.20). In our experiments, we evaluate the integrals in the formula for p_2 in (6.20) by rewriting them as $L^2(\Omega)$-scalar products and by subsequently employing a quadrature rule based on the mass-matrix of the finite element space used for the discretization of y. For the cell-wise averaging that is necessary to calculate the $L^2(-r,r)$-projection onto $\mathcal{U}_{h_u}$ in (8.2), we use the quadrature rule

$$
\frac{1}{h_u}\int_{(n-1)h_u - r}^{nh_u - r} v(s)\,\mathrm{d}s \approx \frac{1}{Q}\sum_{j=1}^{Q} v\left((n-1)h_u - r + \frac{j}{Q+1}h_u\right), \qquad n = 1,\dots,N_u, \qquad Q \in \mathbb{N}. \tag{9.4}
$$

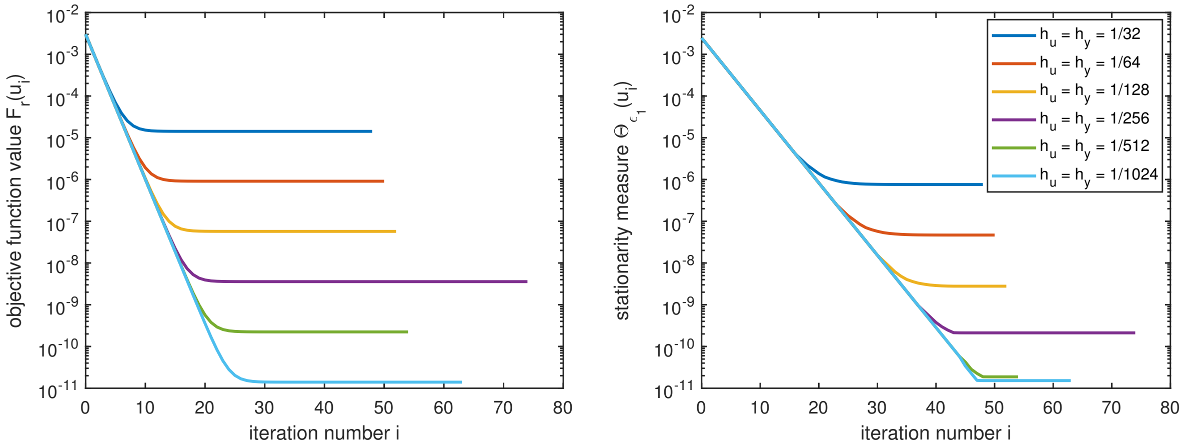

FIGURE 1. Value of the objective function (left) and the stationarity measure $\Theta_{\epsilon_1}(u_i)$ (right) as functions of the iteration counter i of Algorithm 8.9 in the case of Example 1 for different widths h_u and h_y. The legend refers to both figures. It can be seen that the reduction of the objective value and stationarity measure stagnates at a threshold that depends on the discretization level. This reflects that, by discretizing the involved PDEs, one introduces an error and, thus, causes Algorithm 8.9 to run with inexact gradient and function value evaluations; see Section 9.3.

Note that this rule avoids evaluating the integrand at the endpoints of the cells where the elements of $\mathcal{U}_{h_u}$ are discontinuous. Numerical experiments suggest that it is advantageous to determine the $L^2(-r, r)$-projection onto $\mathcal{U}_{h_u}$ in (8.2) by means of quadrature rules whose evaluation points are relatively far away from the discontinuities of the controls. (This is most likely caused by an interaction with the quadrature rule used in (6.20).)

9.4. Numerical results

The results that are obtained when Algorithm 8.9 is applied to problem (9.1) in the situation of Example 1 and the discretization described in Section 9.3 is used for the numerical realization can be seen in Figures 1 and 2 and Tables 1 and 2 below. In all of these experiments, the line-search in Algorithm 8.9 was performed with the parameters $\sigma = 512$ and $\omega = \theta = 0.8$, the semismooth Newton method used for the state equation was run with the tolerance 10^{-12}, the initial iterate u_0 was zero, the relaxation parameter ϵ_1 appearing in the stationarity measure Θ_{ϵ_1} was set to 10^{-16}, and the number of cell-wise quadrature points in (9.4) was chosen as $Q = 5$. Figures 1 and 2 and Table 1 depict the convergence behavior of Algorithm 8.9 in the case where the tolerance ϵ_2 in the termination criterion is set to zero. (We stopped the algorithm in these experiments when the step size τ_i in the line-search loop of Algorithm 8.9 became smaller than 10^{-10}). As Figure 1 shows, in this test configuration, the level that our gradient projection algorithm can drive the objective value and the stationarity measure to depends greatly on the widths used for the discretization. For the test case with the largest widths ($h_y = h_u = 1/32$), Algorithm 8.9 only achieves a final objective value of $\approx 1.424 \cdot 10^{-5}$ and a final stationarity measure of $\approx 7.580 \cdot 10^{-7}$. If the widths are reduced, then the ultimately obtained accuracy improves significantly until, for the smallest widths ($h_y = h_u = 1/1024$), we arrive at a final objective value of $\approx 1.397 \cdot 10^{-11}$ and a final stationarity measure of $\approx 1.522 \cdot 10^{-11}$. Note that the latter values are equal to the objective value/stationarity measure of the exact solution $\bar{u} \equiv 1$ (namely, zero) up to an error that is not much larger than the tolerance 10^{-12} of the semismooth Newton method used for the solution of the state equation. At this level of accuracy, the tolerance of the Newton method and the quadrature errors begin to limit the gains that are obtained by choosing smaller widths h_y and h_u; see the last two lines in the second plot of Figure 1.

We remark that the above observations are consistent with the interpretation of the implemented algorithm as a gradient descent method for a control-discrete and state-continuous version of (9.1) that is executed with

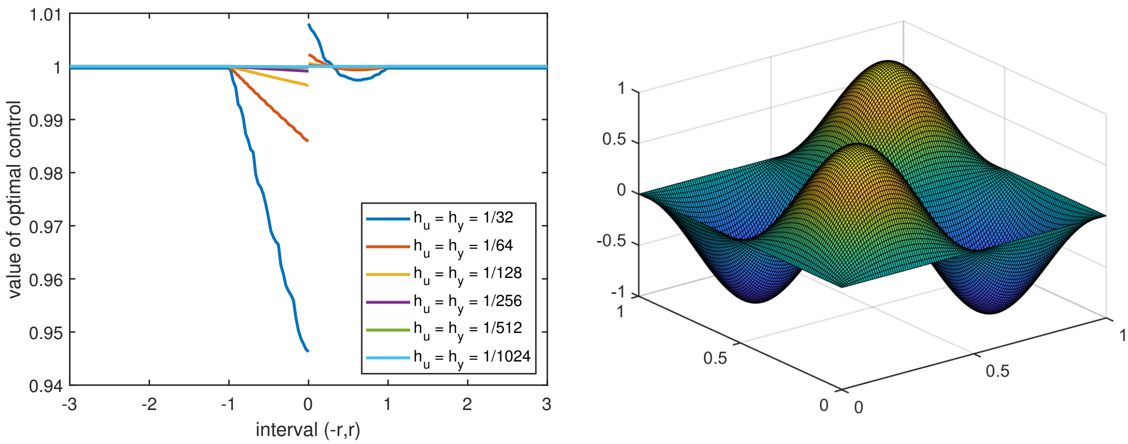

FIGURE 2. Approximations of the optimal control $\bar{u}$ obtained from Algorithm 8.9 at the end of the calculations depicted in Figure 1 for different mesh widths h_u and h_y (left) and resulting optimal state $\bar{y}$ for $h_u = h_y = 1/128$ (right).

TABLE 1. Final objective value (second row), stationarity measure (third row), and $L^\infty(-r,r)$-error between the approximate solution obtained from Algorithm 8.9 and $\bar{u} \equiv 1$ (fourth row) at the end of the computations depicted in Figure 1 for different widths h_u and h_y (first row). The measured $L^\infty(-r,r)$-errors suggest that, for this test case, the convergence is quadratic.

$h_u = h_y$	$1/32$	$1/64$	$1/128$	$1/256$	$1/512$	$1/1024$
F_r	$1.424 \cdot 10^{-5}$	$9.095 \cdot 10^{-7}$	$5.715 \cdot 10^{-8}$	$3.577 \cdot 10^{-9}$	$2.236 \cdot 10^{-10}$	$1.397 \cdot 10^{-11}$
Θ_{ϵ_1}	$7.580 \cdot 10^{-7}$	$4.672 \cdot 10^{-8}$	$2.781 \cdot 10^{-9}$	$2.123 \cdot 10^{-10}$	$1.875 \cdot 10^{-11}$	$1.522 \cdot 10^{-11}$
L^∞-error	$5.373 \cdot 10^{-2}$	$1.404 \cdot 10^{-2}$	$3.580 \cdot 10^{-3}$	$9.033 \cdot 10^{-4}$	$2.268 \cdot 10^{-4}$	$5.684 \cdot 10^{-5}$

inexact gradient and function value evaluations; see Section 9.3: If the discretization is too coarse, then the level of inexactness is too high and the perturbation of the gradients prevents the method from reducing Θ_{ϵ_1} to acceptable levels. Only once the widths are small enough, the approximations of the "real" gradients of the control-discrete and state-continuous objective function obtained from the FE-discretization of the involved PDEs are sufficiently accurate to achieve an adequate reduction. We remark that, by varying the width h_u in the situation of Figures 1 and 2, one can check that the width h_y is decisive for the final accuracy level while the choice of h_u only has little influence. This reflects that the discretization of the PDEs is indeed the source of the inexactness of the gradients as described in Section 9.3. Compare also with the analysis in Section 8 in this context which is completely unaffected by the discretization of the control space.

A further important feature of Algorithm 8.9 visible in Figure 1 is the property of asymptotic mesh independence: As can be seen in Figure 1, the graphs that depict how the objective value and the stationarity measure decrease with the iteration number initially all have the same slope and only branch off once the maximum accuracy of the respective discretization level is reached. In practice, this means that, if the goal is to calculate a control whose stationarity measure Θ_{ϵ_1} is smaller than a given tolerance $\epsilon_2 > 0$, then Algorithm 8.9 is able to accomplish this with a number of iterations that does not tend to infinity as h_u and h_y go to zero but is effectively constant for all widths that are sufficiently small; see Table 2.

As a final observation, we would like to point out that, even though the analytical solution $\bar{u}$ is continuous in the situation of Example 1, the approximate solutions calculated by Algorithm 8.9 possess discontinuities at the origin; see Figure 2. This illustrates that problems of the type (P) indeed promote a jump at zero and that

TABLE 2. Number of gradient projection steps needed to drive the stationarity measure Θ_{ϵ_1} below $\epsilon_2 = 10^{-8}$ in the situation of Example 1 for different h_y and h_u. It can be seen that the convergence is mesh-independent for all h_y that are smaller than $\approx 1/110$; *cf.* Figure 1.

h_y \ h_u	1/32	1/64	1/128	1/256	1/512	1/1024
1/80	∞	∞	∞	∞	∞	∞
1/90	37	38	∞	∞	∞	∞
1/100	32	32	33	34	34	34
1/110	32	32	32	32	32	32
1/120	32	32	32	32	32	32
1/130	32	32	32	32	32	32

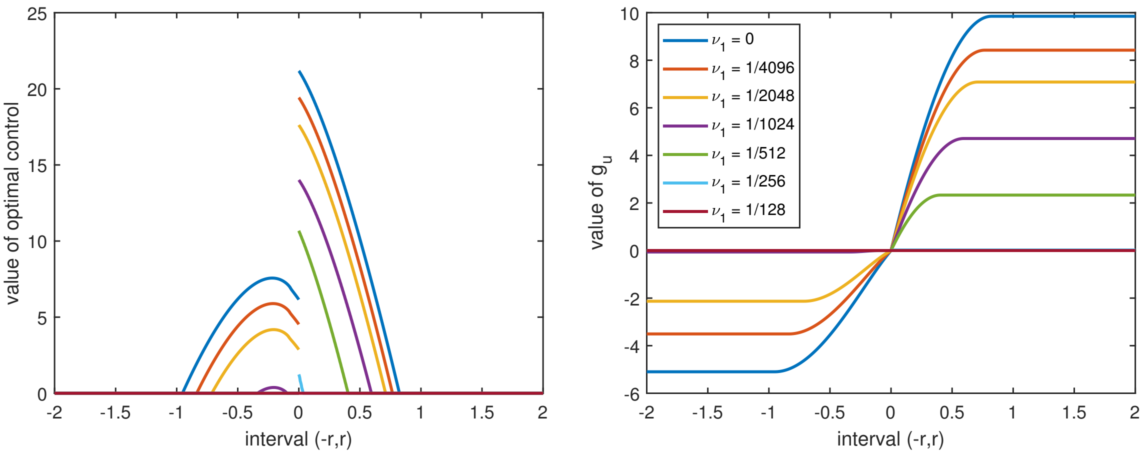

FIGURE 3. Approximations of the optimal control $\bar{u}$ (left) and the superposition function $g_{E_r(\bar{u})}$ (right) obtained from Algorithm 8.9 in the situation of Example 2 for different values of ν_1 and $h_y = h_u = 1/512$. The legend refers to both figures. It can be seen that the L^1-regularization promotes sparsity properties; *cf.* the stationarity system (7.4). For the highest value $\nu_1 = 1/128$, the optimal control is zero.

cases with continuous solutions are exceptional and typically instable w.r.t. small perturbations. We remark that this can also be seen in the formula (6.24) in which the coefficient in front of the Dirac measure only vanishes if p_1 has mean value zero. Note that, in the limit $h_y \to 0$ and $h_u \to 0$, the discontinuities in Figure 2 vanish and the approximate solutions converge to the exact optimal control $\bar{u} \equiv 1$. For the finest discretization, the $L^\infty(-r, r)$-error between the approximate and the exact solution is $\approx 5.684 \cdot 10^{-5}$; see Table 1.

Having demonstrated that Algorithm 8.9 works and is able to reproduce the analytical solution $\bar{u} = 1$ in the situation of Example 1, we now turn our attention to Example 2 for which no analytical solution is available. The results that we have obtained for Example 2 can be seen in Figures 3 and 4 and Table 3 below. In all of the depicted experiments, the parameters in Algorithm 8.9 were chosen as $\sigma = 2048$, $\omega = \theta = 0.8$, $\epsilon_1 = 10^{-16}$, and $\epsilon_2 = 10^{-8}$; the tolerance of the semismooth Newton method was set to 10^{-12}; the initial guess was $u_0 = 0$; and the controls and states were discretized as described in Section 9.3 with $Q = 10$.

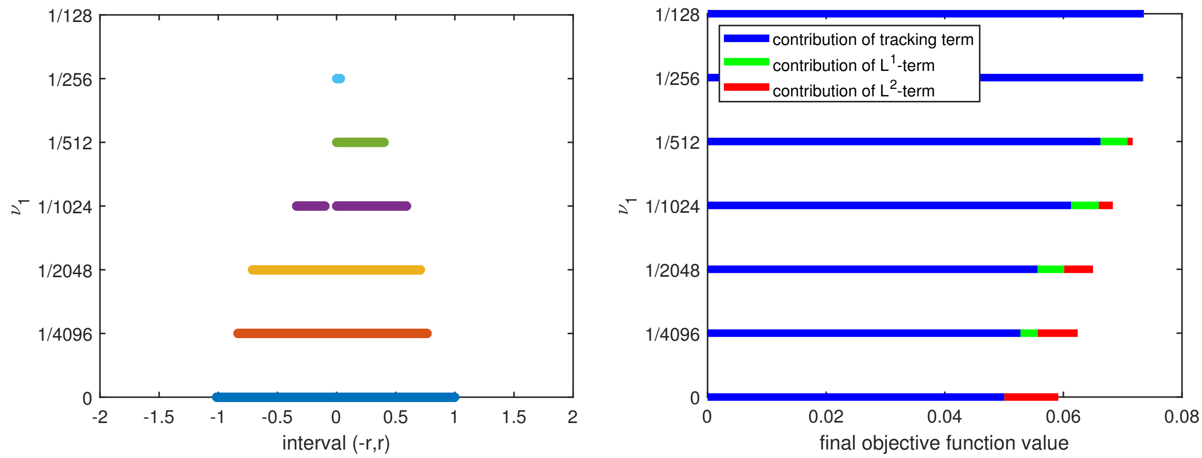

FIGURE 4. Depiction of the (essential) support of the optimal control $\bar{u}$ (left) and contributions of the tracking term and the $L^1(-r,r)$- and $L^2(-r,r)$-regularization terms to the final objective function value (right) in the situation of Figure 3. It can be seen that the essential support of the optimal control $\bar{u}$ shrinks as ν_1 increases. For the highest value $\nu_1 = 1/128$, it vanishes.

TABLE 3. Number of gradient projection steps needed to drive the stationarity measure Θ_{ϵ_1} below the tolerance $\epsilon_2 = 10^{-8}$ in the situation of Example 2 for different mesh widths h_y and h_u and $\nu_1 = 1/1024$. It can again be seen that the convergence is mesh-independent for all h_y that are sufficiently small. Compare with Figure 1 and Table 2 in this context.

h_y \ h_u	$1/32$	$1/64$	$1/128$	$1/256$	$1/512$	$1/1024$
$1/64$	45	∞	46	47	49	48
$1/128$	46	46	46	46	46	46
$1/256$	46	46	46	46	46	46
$1/512$	46	46	46	46	46	46
$1/1024$	46	46	46	46	46	46

As Table 3 shows, when applied to Example 2, Algorithm 8.9 exhibits an asymptotically mesh-independent convergence behavior that is analogous to that observed for Example 1 in Figure 1 and Table 2. In Figure 3, one can further see that the optimal superposition functions $g_{E_r(\bar{u})}$ calculated in Example 2 are indeed ReLU-like and "tight" in the sense that they are sigmoidal, continuously differentiable away from the origin, possess a distinct kink at zero, and have a vanishing derivative outside of the range of the optimal state. This confirms the predictions made by Corollaries 7.3, 7.6 and shows that (P) indeed promotes properties that are present in popular activation functions from machine learning. In Figures 3 and 4, one can also see that the L^1-regularization term in the objective function of (9.1) causes the optimal controls to be sparse in the sense that they are supported on a set whose measure decreases as ν_1 increases. This behavior is in good accordance with the influence that the parameter ν_1 has on the system (7.4); *cf.* [48]. We remark that the derivation of a-priori error estimates for problems of the type (9.1) is an interesting topic. We leave this field for future research.

References

[1] C. Aarset, M. Holler and T. T. N. Nguyen, Learning-informed parameter identification in nonlinear time-dependent PDEs. *Appl. Math. Optim.* **88** (2023) 53.

[2] S.L. Brunton, J.L. Proctor and J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Nat. Acad. Sci. U.S.A.* **113** (2016) 3932–3937.

[3] S. Court and K. Kunisch, Design of the monodomain model by artificial neural networks. *Discrete Contin. Dyn. Syst.* **42** (2022) 6031–6061.

[4] G. Dong, M. Hintermüller and K. Papafitsoros, A descent algorithm for the optimal control of ReLU neural network informed PDEs based on approximate directional derivatives. arXiv:2210.07900v1, 2022.

[5] G. Dong, M. Hintermüller and K. Papafitsoros, Optimization with learning-informed differential equation constraints and its applications. *ESAIM Control Optim. Calc. Var.* **28** (2022) 3.

[6] G. Dong, M. Hintermüller, K. Papafitsoros and K. Völkner, First-order conditions for the optimal control of learning-informed nonsmooth PDEs. arXiv:2206.00297v2, 2022.

[7] B. Kaltenbacher and T.T.N. Nguyen, Discretization of parameter identification in PDEs using neural networks. *Inverse Problems* **38** (2022) 124007.

[8] T. Qin, K. Wu and D. Xiu, Data driven governing equations approximation using deep neural networks. *J. Comput. Phys.* **395** (2019) 620–635.

[9] S.H. Rudy, S.L. Brunton, J.L. Proctor and J.N. Kutz, Data-driven discovery of partial differential equations. *Sci. Adv.* **3** (2017) e1602614.

[10] C. Christof and J. Kowalczyk, On the omnipresence of spurious local minima in certain neural network training problems. *Constr. Approx* (2023) to appear.

[11] D.N. Hào, B.V. Huong, P.X. Thanh and D. Lesnic, Identification of nonlinear heat transfer laws from boundary observations. *Appl. Anal.* **94** (2015) 1784–1799.

[12] A. Rösch, Identification of nonlinear heat transfer laws by optimal control. *Num. Funct. Anal. Optim.* **15** (1994) 417–434.

[13] A. Rösch, Fréchet differentiability of the solution of the heat equation with respect to a nonlinear boundary condition. *Z. Anal. Anwend.* **15** (1996) 603–618.

[14] A. Rösch, Second order optimality conditions and stability estimates for the identification of nonlinear heat transfer laws, in *Control and Estimation of Distributed Parameter Systems*, edited by W. Desch, F. Kappel, and K. Kunisch. Birkhäuser Basel, Basel (1998) 237–246.

[15] A. Rösch, A Gauss–Newton method for the identification of nonlinear heat transfer laws, in *Optimal Control of Complex Structures*, edited by K.-H. Hoffmann, I. Lasiecka, G. Leugering, J. Sprekels, and F. Tröltzsch. Birkhäuser Basel, Basel (2002) 217–230.

[16] A. Rösch and F. Tröltzsch, An optimal control problem arising from the identification of nonlinear heat transfer laws. *Arch. Control Sci.* **1** (1992) 4–183.

[17] A. Rösch, Stability estimates for the identification of nonlinear heat transfer laws. *Inverse Problems* **12** (1996) 743–756.

[18] S. Grützner and A. Muntean, Identifying processes governing damage evolution in quasi-static elasticity, part 1 – analysis. *Adv. Math. Sci. Appl.* **30** (2021) 305–334.

[19] V. Barbu and K. Kunisch, Identification of nonlinear elliptic equations. *Appl. Math. Optim.* **33** (1996) 139–167.

[20] V. Barbu, P. Neittaanmäki and A. Niemistö, A penalty method for the identification of nonlinear elliptic differential operator. *Numer. Funct. Anal. Optim.* **15** (1994) 503–530.

[21] M. Goebel, Smooth and nonsmooth optimal Lipschitz control – a model problem, in *Variational Calculus, Optimal Control and Applications*, edited by W. H. Schmidt, K. Heier, L. Bittner, and R. Bulirsch. Birkhäuser Basel (1998) 53–60.

[22] B. Kaltenbacher and W. Rundell, The inverse problem of reconstructing reaction-diffusion systems. *Inverse Problems* **36** (2020) 065011.

[23] Y. Kian, Lipschitz and Hölder stable determination of nonlinear terms for elliptic equations. *Nonlinearity* **36** (2023) 1302.

[24] C. Christof, C. Meyer, S. Walther and C. Clason, Optimal control of a non-smooth semilinear elliptic equation. *Math. Control Relat. Fields* **8** (2018) 247–276.

[25] J.F. Bonnans and A. Shapiro, Perturbation Analysis of Optimization Problems. Springer Series in Operations Research. Springer, New York (2000).

[26] B. Schweizer, Partielle Differentialgleichungen. Springer-Verlag, Berlin/Heidelberg (2013).

[27] V.I. Bogachev, Measure Theory. Springer (2007).

[28] H. Attouch, G. Buttazzo and G. Michaille, Variational Analysis in Sobolev and BV Spaces. SIAM, Philadelphia (2006).

[29] D. Gilbarg and N.S. Trudinger, Elliptic Partial Differential Equations of Second Order. Springer (2001).

[30] G.A. Monteiro, A. Slavík and M. Tvrdý, Kurzweil-Stieltjes Integral: Theory and Applications. Vol. 15 of *Series in Real Analysis*. World Scientific, Singapore (2019).

[31] L. Ambrosio, N. Fusco and D. Pallara, Functions of Bounded Variation and Free Discontinuity Problems. Oxford University Press, Oxford & New York (2000).

[32] P. Grisvard, Elliptic Problems in Nonsmooth Domains. Pitman (1985).

[33] F. Tröltzsch, Optimal Control of Partial Differential Equations. AMS (2010).

[34] W.P. Ziemer, Weakly Differentiable Functions. Springer Verlag, New York (1989).

[35] D. Kinderlehrer and G. Stampacchia, An Introduction to Variational Inequalities and Their Applications. Vol. 31 of *Classics in Applied Mathematics*. SIAM (2000).
[36] V. Barbu, Optimal Control of Variational Inequalities. Research Notes in Mathematics. Pitman (1984).
[37] M. Josephy, Composing functions of bounded variation. *Proc. Amer. Math. Soc* **83** (1981) 354–356.
[38] C. Christof, *Sensitivity Analysis of Elliptic Variational Inequalities of the First and the Second Kind*. Ph.D. thesis, Technische Universität Dortmund (2018).
[39] B.S. Mityagin, The zero set of a real analytic function. *Math. Notes* **107** (2020) 529–530.
[40] C. Christof and G. Müller, Multiobjective optimal control of a non-smooth semilinear elliptic partial differential equation. *ESAIM Control Optim. Calc. Var.* **27** (2021) Art. S13.
[41] C. Clason, V.H. Nhu and A. Rösch, No-gap second-order optimality conditions for optimal control of a non-smooth quasilinear elliptic equation. *ESAIM Control Optim. Calc. Var.* **27** (2021) Art. 62.
[42] P. Drábek and J. Milota, Methods of Nonlinear Analysis: Applications to Differential Equations. Birkhäuser Verlag (2007).
[43] A. Beck, Introduction to Nonlinear Optimization. MOS/SIAM Series on Optimization. SIAM (2014).
[44] C. Christof, Gradient-based solution algorithms for a class of bilevel optimization and optimal control problems with a nonsmooth lower level. *SIAM J. Optim.* **30** (2020) 290–318.
[45] M.B. Cohen, J. Diakonikolas and L. Orecchia, On acceleration with noise-corrupted gradients, in *Proceedings of the 35th International Conference on Machine Learning (PMLR 80)*. Stockholm, Sweden (2018) 1019–1028.
[46] L. Hertlein and M. Ulbrich, An inexact bundle algorithm for nonconvex nonsmooth minimization in Hilbert space. *SIAM J. Control Optim.* **57** (2019) 3137–3165.
[47] M. Hinze and A. Rösch, Discretization of optimal control problems, in *Constrained Optimization and Optimal Control for Partial Differential Equations*, edited by G. Leugering, S. Engell, A. Griewank, M. Hinze, R. Rannacher, V. Schulz, M. Ulbrich and S. Ulbrich. Springer Basel, Basel (2012) 391–430.
[48] G. Stadler, Elliptic optimal control problems with L1-control cost and applications for the placement of control devices. *Comput. Optim. Appl.* **44** (2009).