

## **First-order conditions for the optimal control of learning-informed nonsmooth PDEs**

Guozhi Dong<sup>1</sup>, Michael Hintermüller<sup>2,3</sup>, Kostas Papafitsoros<sup>2</sup>, Kathrin Völkner<sup>4</sup>

submitted: June 2, 2022

<sup>1</sup> School of Mathematics and Statistics  
Central South University  
Lushan South Road 932  
410083 Changsha, China  
E-Mail: guozhi.dong@csu.edu.cn

<sup>2</sup> Weierstrass Institute  
Mohrenstr. 39  
10117 Berlin, Germany  
E-Mail: michael.hintermueller@wias-berlin.de  
kostas.papafitsoros@wias-berlin.de

<sup>3</sup> Humboldt-Universität zu Berlin  
Unter den Linden 6  
10099 Berlin, Germany  
E-Mail: hint@math.hu-berlin.de

<sup>4</sup> Martin-Luther-Universität Halle-Wittenberg  
Institut für Mathematik  
06099 Halle, Germany  
E-Mail: kathrin.voelkner@mathematik.uni-halle.de

No. 2940  
Berlin 2022



---

2020 *Mathematics Subject Classification.* 49M15, 65J15, 65K10, 35J61, 68T07, 49J52.

*Key words and phrases.* Nonsmooth partial differential equations, data-driven models, neural networks, ReLU activation function, optimal control, PDE constrained optimization.

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). The work of GD is also supported by an NSFC grant, No. 12001194. The work of MH is partially supported by the DFG SPP 1962, project-145r. The work of KV was supported by the WIAS Female Master Students Program.

Edited by  
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)  
Leibniz-Institut im Forschungsverbund Berlin e. V.  
Mohrenstraße 39  
10117 Berlin  
Germany

Fax: +49 30 20372-303  
E-Mail: [preprint@wias-berlin.de](mailto:preprint@wias-berlin.de)  
World Wide Web: <http://www.wias-berlin.de/>

# First-order conditions for the optimal control of learning-informed nonsmooth PDEs

Guozhi Dong, Michael Hintermüller, Kostas Papafitsoros, Kathrin Völkner

## Abstract

In this paper we study the optimal control of a class of semilinear elliptic partial differential equations which have nonlinear constituents that are only accessible by data and are approximated by nonsmooth ReLU neural networks. The optimal control problem is studied in detail. In particular, the existence and uniqueness of the state equation are shown, and continuity as well as directional differentiability properties of the corresponding control-to-state map are established. Based on approximation capabilities of the pertinent networks, we address fundamental questions regarding approximating properties of the learning-informed control-to-state map and the solution of the corresponding optimal control problem. Finally, several stationarity conditions are derived based on different notions of generalized differentiability.

## 1 Introduction

### 1.1 Context and motivation

In this paper we study the following prototypical optimal control problem:

$$\begin{aligned} & \text{minimize} && J(y, u) := \frac{1}{2} \|y - g\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2, && \text{over } (y, u) \in H_0^1(\Omega) \times L^2(\Omega), \\ & \text{subject to} && \begin{cases} -\Delta y + f(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} && \text{and } u \in \mathcal{C}_{ad}. \end{aligned} \tag{P_f}$$

Here  $\Omega$  denotes an open, bounded Lipschitz domain in  $\mathbb{R}^d$ ,  $d \geq 2$ , with boundary  $\partial\Omega$ ,  $g \in L^2(\Omega)$  is a given desired state, and  $\alpha > 0$  is fixed. Further,  $\mathcal{C}_{ad}$  is the admissible set of controls  $u$ , which is a nonempty, bounded, closed and convex subset of the Lebesgue space  $L^p(\Omega)$  for some  $p \geq 2$ . The state is given by  $y$  and is supposed to lie in the Sobolev space  $H_0^1(\Omega)$ ; see, e.g., [1] for the latter. Moreover,  $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$  is assumed to be a rather general nonlinear and possibly nonsmooth function. A particular motivating example for this work is given by  $f$  represented (or approximated) by a so-called ReLU artificial neural network. In this case  $f$  is replaced by  $\mathcal{N}$ , a neural network function that has the Rectified Linear Unit (ReLU)  $\sigma(t) := \max(t, 0)$  as the underlying activation function; see Section 2 for more details and definitions. When such an  $\mathcal{N}$  approximates  $f$ , the semilinear partial differential equation PDE in  $(P_f)$  is replaced by the following (nonsmooth) *learning-informed* PDE

$$\begin{cases} -\Delta y + \mathcal{N}(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega. \end{cases} \tag{1.1}$$

The concept of learning-informed PDEs was recently introduced in [16]. It aims at representing an (explicitly) unknown physical law  $f$  by a neural network (map)  $\mathcal{N}$  which is learned from given data.

For instance, assume that a data set

$$D := \{(y_i, u_i) : y_i \text{ (approximately) solves the original unknown PDE for } u_i, i = 1 \dots, n_D\},$$

of pre-specified controls and associated state responses is at our disposal. Such a set may, for example, arise from measurements or computations. Then this data set can be used towards evaluation instances of  $f$  via  $f(x_j, y_i(x_j)) \simeq u_i(x_j) + \Delta y_i(x_j)$ , where  $\{x_j\}_{j=1}^\ell$ ,  $\ell \in \mathbb{N}$ , is a finite collection of evaluation locations in the domain  $\Omega$ . Using these instances as a training set, a neural network  $\mathcal{N}$  can be computed in the context of supervised learning and take the role of an approximation of the unknown  $f$ . Via solving the associated PDE (1.1), this results in a learning-informed control-to-state map  $S_{\mathcal{N}} : u \mapsto y$ . In view of this, such a set-up can also be regarded as an example in the currently rather limited field of operator learning, which is the task of learning a map between infinite dimensional spaces; here from the control to the state space. In particular, it provides a case that depicts how low dimensional learning could help to identify infinite dimensional control-to-state operators. We also note that in this set-up, the training of the network  $\mathcal{N}$  is assumed to be done in an *offline* phase, meaning that the minimization problem that constitutes the training of  $\mathcal{N}$  and the optimal control problem ( $P_f$ ) are decoupled. In [16], a smooth version of the above framework was considered in order to learn the physical law that governs the separation of a (initially homogeneous) material mixture into two pure states. This law can be related to (a derivative of) a double-well potential type function  $f$  in a stationary Allen-Cahn equation. In the same work, an ordinary differential equation (ODE) version of ( $P_f$ ) was used to approximate the physics behind magnetic resonance imaging (MRI) [14] with the goal of reconstructing quantitative maps of biophysical parameters in the context of quantitative MRI.

The regularity of the neural network function  $\mathcal{N}$  is determined by the regularity of the underlying activation function. In this vein, we note that the set of functions represented by ReLU neural networks coincides with the family of piecewise affine maps; see Section 2. Since the main objective in [16] was to introduce the concept of learning-informed PDEs, only smooth neural networks were considered, corresponding to some smooth, typically sigmoidal activation function, such as, e.g.,  $\tanh$  or  $\arctan$ . This eases the subsequent optimization-theoretic treatment of the optimal control problem like the derivation of first-order optimality conditions. However, when it comes to deep learning applications nowadays, the nonsmooth ReLU is the most popular choice among activation functions. Its main advantages lie in its sparsification properties, its easy computation and further advantages during training, such as for instance with regards to remedy the issue of vanishing gradients [6, 18]. The study of approximation capabilities of ReLU neural networks is another active field of research [20, 27]. All these aspects are also relevant to our problem, in particular with respect to the capability of the learning-informed optimal control problem to approximate the original one. We mention already here that we consider the nonsmooth  $\mathcal{N}$  to be monotonically increasing in the variable  $y$ . This guarantees uniqueness of solutions to the learning-informed state equation, resulting in a well-defined control-to-state map. Conceptually, this assumes that the ground truth map  $f$  is also monotonically increasing and the network has been trained sufficiently well, thus preserving this monotonicity. However, we point out that for homogeneous Dirichlet boundary conditions (like the one in (1.1)) one can still show uniqueness of the solution of the state equation if the negative part of  $\partial_y \mathcal{N}$  is sufficiently small (which is often the case for good enough approximations of monotone functions). In this case it can still be shown that the corresponding PDE operator is strongly monotone and the Browder-Minty theorem can be applied. Let us also point out that there are ways to enforce monotonicity during training [13, 25, 29] via a sufficiently good derivative approximation (Sobolev training) [12].

Since  $f$  and its approximation  $\mathcal{N}$  are nonsmooth maps, one does not expect the associated control-to-state map  $S$  to be Fréchet differentiable. This poses difficulties in the derivation of first-order optimality respectively stationarity conditions for the optimal control problem as well as its numerical treatment.

Therefore, one needs to resort to more general notions of differentiability (than Fréchet differentiability); we refer to the introduction of [10] for a short review. In the same paper, the optimal control of a semilinear PDE with the special choice  $f(\cdot, y) = \max(0, y)$  and in the absence of control constraints was studied, characterizing the Bouligand subdifferential of the control-to-state map and deriving so-called strong stationarity conditions. We remark that in our set-up this would correspond to a very simple and shallow ReLU network  $\mathcal{N}$ . Here, following [10] and also adapting certain results from a recent preprint [8], we study in depth a version of  $(P_f)$  with a more general  $f$  that still covers the ReLU neural network case. This function  $f$  is assumed to be measurable in the first variable and locally Lipschitz and directionally differentiable in the second one. For this general problem, we establish a series of stationarity respectively necessary first-order conditions. We begin by deriving purely primal optimality conditions ( $B$ -stationarity) which are equivalent to the directional derivative of the reduced objective in feasible directions being non-negative, as well as conditions that are derived as a limit of a regularization scheme [5, 10, 26] (weak stationarity). The latter conditions are refined in the case that  $f$  is a piecewise differentiable function yielding  $C$ -stationarity. Under additional constraint qualifications, we further improve  $C$ -stationarity by establishing certain sign conditions on the dual variables. This leads to strong stationarity which is shown to be equivalent to  $B$ -stationarity in specific cases.

## 1.2 Structure of the paper

In Section 2 we start by reviewing basic properties of ReLU neural networks. In Section 3, we study the state equation of the optimal control problem  $(P_f)$  with a rather general nonsmooth, nonlinear function  $f$ . Existence and uniqueness of solutions to this PDE are discussed, and continuity as well as directional differentiability properties of the corresponding control-to-state map are shown. We then study approximation results of this PDE on two levels: (i)  $f$  is approximated by a ReLU neural network  $\mathcal{N}$ ; (ii)  $f$ , or a network  $\mathcal{N}$ , is approximated by some smoothed version of it. The corresponding general optimal control problem is studied in Section 4. Existence and approximation results stemming from the analogous results of the state equation are shown. In Section 5, we focus on different stationarity systems that are necessary for local minimizers of the optimal control problem, namely  $B$ -, weak,  $C$ - and strong stationarity.

We finally mention that in the companion paper [15], by employing a descent algorithm inspired by the bundle-free method in [23], a numerical algorithm that treats the nonsmooth optimal control problem  $(P_f)$  directly is proposed, analyzed and tested.

## 2 ReLU neural networks

In this work, we rely on ReLU neural networks as efficient approximators of an unknown and potentially complex function  $f : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_L}$ , with  $n_0, n_L \in \mathbb{N}$ , for which only a few (approximate) data values are available, namely  $\{f_j \approx f(x_j)\}_{j=1}^{n_D}$  where  $n_D \in \mathbb{N}$  and  $\{x_j\}_{j=1}^{n_D} \subset \Omega$ . Given this setting, we assume that a network function  $\mathcal{N}_\theta$  with parameters  $\theta$  – see Section 2.1 below for definitions – has been trained to “learn” the function  $f$  in the context of supervised learning. This training is typically achieved by (approximately) solving

$$\min_{\theta \in \mathbb{R}^p} \sum_{j=1}^{n_D} \ell(\mathcal{N}_\theta(x_j), f_j) + \mathcal{R}(\theta). \quad (2.1)$$

Here,  $\ell$  is a suitable loss function and  $\mathcal{R}$  is an optional regularization term inducing some a priori properties on  $\theta$ . Clearly, the study of (2.1) is an important research area in the field of deep learning [19], but here we rather assume that this learning process has been successfully completed such that we are equipped with a ReLU network  $\mathcal{N}$  which approximates  $f$  sufficiently well. As a consequence, in what follows we merely focus on the properties of such networks.

## 2.1 Definition and basic properties

Let  $\mathbb{N} := \{1, 2, \dots\}$ .

**Definition 2.1** (Standard feedforward multilayer neural network). *Let  $L \in \mathbb{N}$ , network parameters  $\theta = ((W_1, b_1), \dots, (W_L, b_L))$  with  $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$ ,  $b_i \in \mathbb{R}^{n_i}$ , for  $i = 1, \dots, L$  and  $n_i \in \mathbb{N}$  for  $i = 0, \dots, L$ . Furthermore let  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  be an arbitrary function. We say that a function  $\mathcal{N} : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_L}$  is a neural network with weight matrices  $(W_i)_{i=1}^L$ , bias vectors  $(b_i)_{i=1}^L$  (the network parameters) and activation function  $\sigma$  if  $\mathcal{N}(x)$  can be defined through the following recursive relation for any  $x \in \mathbb{R}^{n_0}$ :*

$$z_0 = x, \tag{2.2}$$

$$z_\ell = \sigma(W_\ell z_{\ell-1} + b_\ell), \quad \ell = 1, \dots, L-1, \tag{2.3}$$

$$\mathcal{N}(x) = W_L z_{L-1} + b_L. \tag{2.4}$$

The action of the activation function  $\sigma$  in (2.3) is understood in a component-wise sense, i.e., for a vector  $y = (y^1, \dots, y^n) \in \mathbb{R}^n$  we have  $\sigma(y) := (\sigma(y^1), \dots, \sigma(y^n))$ . More compactly,  $\mathcal{N}$  reads

$$\mathcal{N}(x) = T_L \circ \sigma(T_{L-1}) \circ \dots \circ \sigma(T_2) \circ \sigma(T_1)(x), \quad x \in \mathbb{R}^{n_0}, \tag{2.5}$$

where, for every  $\ell = 1, \dots, L$ ,  $T_\ell$  denotes the affine transformation  $\mathbb{R}^{n_{\ell-1}} \ni z \mapsto W_\ell z + b_\ell \in \mathbb{R}^{n_\ell}$ .

We say that  $\mathcal{N}$  is a ReLU neural network if  $\sigma$  is the ReLU (Rectified Linear Unit) activation function

$$\sigma(t) = \max(t, 0), \quad t \in \mathbb{R}. \tag{2.6}$$

Following the standard neural network terminology, we say that a neural network defined as in (2.2)–(2.4), has  $L$  layers and  $L-1$  hidden layers, with the latter denoting the operations in (2.3). The final operation (2.4) is called the *output layer*. Furthermore,  $n_i$  is called the number of *neurons* in the  $i$ -th layer,  $i = 1, \dots, L$ , which is the number of rows of the weight matrix  $W_i$ . The number of neurons of a given layer is also called the *width* of that very layer, while the number of layers is called the *depth* of the network.

We note that a neural network as a function does not necessarily admit a unique representation with respect to the weight matrices, the bias vectors and the activation functions. Furthermore in Definition 2.1, the input of the  $\ell$ -th layer consists only of the output  $z_{\ell-1}$  of the previous layer. A more general neural network definition would allow the input for each layer to depend on the output of all previous layers. In that case, every  $W_\ell$  would be a weight matrix of size  $\mathbb{R}^{n_\ell \times (\sum_{k=0}^{\ell-1} n_k)}$ . However, since every network of that type can be realized by a network as in Definition 2.1 (see also [20]) we will always relate to the more classical definition given above.

We are interested in the regularity and structural properties of the class of functions that are realized by ReLU neural networks. It turns out that the latter coincides with the class of *continuous piecewise affine functions*, which we define next.

**Definition 2.2** (Continuous piecewise affine functions). *Let  $n_0 \in \mathbb{N}$ . We say that a function  $\mathcal{F} : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$  is continuous piecewise affine if the following condition holds:*

- $\mathcal{F}$  is continuous and there exist finitely many affine maps  $f_1, \dots, f_p : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$  for some  $p \in \mathbb{N}$  such that for every  $x \in \mathbb{R}^{n_0}$ , there exists an  $i \in \{1, \dots, p\}$  such that  $\mathcal{F}(x) = f_i(x)$ .

Note that such a map  $\mathcal{F}$  is not necessarily Fréchet differentiable, and see [2, 3, 31] for characterizations and properties of continuous piecewise affine functions.

**Theorem 2.3.** *A function  $\mathcal{N} : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$  is a ReLU neural network if and only if it is a continuous piecewise affine function.*

From the definition (2.2)–(2.4) it is clear that  $\mathcal{N} : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^{n_L}$ ,  $n_L \geq 1$ , is a  $\mathbb{R}^{n_L}$ -valued ReLU neural network if and only if  $\mathcal{N} = (\mathcal{N}_1, \dots, \mathcal{N}_L)$  with each  $\mathcal{N}_i : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, L$ , is a scalar-valued ReLU neural network. Thus,  $\mathcal{N}$  is an  $\mathbb{R}^{n_L}$ -valued ReLU neural network if and only if it is an  $\mathbb{R}^{n_L}$ -valued continuous piecewise affine function, with the latter defined exactly as in Definition 2.2 with the only difference that the affine maps  $f_i$  are  $\mathbb{R}^{n_L}$ -valued.

## 2.2 Approximation results

Next we recall several basic properties regarding the approximation capabilities of ReLU neural networks in the spirit of the several versions of the *universal approximation theorem*; see [28] for a thorough review. In fact, in the case where the activation function is  $k$ -times differentiable and not a polynomial, the set of the corresponding one-hidden layer networks is dense in the set of  $k$ -times continuously differentiable functions in the topology of uniform convergence (for function values and derivatives up to order  $k$ ) on compact sets. In the case of the ReLU activation function, the corresponding approximation results are less standard and are stated here for the ease of access.

Indeed, there exists a growing amount of literature that studies properties of ReLU networks with regards to their approximation of functions of a given Sobolev regularity. Typically, special emphasis is given to the upper bound on the number of layers, neurons and nonzero weights needed to achieve approximation of a certain accuracy with respect to a Sobolev norm; see for instance [7, 20, 32]. For our purposes, the following result taken from the aforementioned references will be of interest: Given an open, bounded domain  $U \subset \mathbb{R}^{n_0}$  with Lipschitz boundary we have that

$$\text{for every } \epsilon > 0 \text{ and } f \in W^{1,\infty}(U) \text{ there exists a ReLU network } \mathcal{N}_\epsilon : \mathbb{R}^{n_0} \rightarrow \mathbb{R} \text{ such that} \quad (2.7)$$

$$\|\mathcal{N}_\epsilon - f\|_{W^{1,\infty}(U)} < \epsilon.$$

Here, the Sobolev space  $W^{1,\infty}(U)$  relates to Lipschitz functions on  $U$ ; see [1]. We also note that the density result in (2.7) is typically stated for  $U = (-M, M)^{n_0}$  for a given  $M > 0$ , but it is easy to see that it holds for every open, bounded  $U \subset \mathbb{R}^{n_0}$  with Lipschitz boundary. Indeed, given such a domain  $U$ , consider a bounded, linear extension operator  $T : W^{1,\infty}(U) \rightarrow W^{1,\infty}(\mathbb{R}^{n_0})$  such that  $Tg$  has compact support, which is common for every  $g \in W^{1,\infty}(U)$ . Then we readily get

$$\|\mathcal{N}_\epsilon - f\|_{W^{1,\infty}(U)} \leq \|\mathcal{N}_\epsilon - Tf\|_{W^{1,\infty}((-M,M)^{n_0})}$$

for some large enough  $M > 0$ .

### 3 Nonsmooth and ReLU neural network informed PDE

We recall once again the structure of the learning-informed optimal control problem that motivates us in the following:

$$\begin{aligned} & \text{minimize} && J(y, u) := \frac{1}{2} \|y - g\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2, && \text{over } (y, u) \in H_0^1(\Omega) \times L^2(\Omega), \\ & \text{subject to} && \begin{cases} -\Delta y + \mathcal{N}(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} && \text{and } u \in \mathcal{C}_{ad}. \end{aligned} \quad (P_{\mathcal{N}})$$

Note that the governing state equation in  $(P_{\mathcal{N}})$  is a semilinear elliptic PDE with the neural network  $\mathcal{N}$  (approximating some unknown  $f$ ) as a constituent. Concerning  $\mathcal{N} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$  we assume that it is a ReLU neural network which is monotonically increasing in the second variable. The objective  $J$  is of tracking type and may be replaced by other suitable candidates allowing to prove existence of a solution to the associated optimal control problem. Such tracking objectives often arise, e.g., in engineering applications where one wishes to find a system state  $y$  which is sufficiently close to some given desired state  $g$  and which arises through some control action  $u$  in the governing equation. The associated control is constrained by  $\mathcal{C}_{ad}$ , and its pertinent average  $L^2$ -cost is  $\alpha > 0$ .

As outlined above, our underlying assumption is that  $\mathcal{N}$  is an approximation of an unknown, potentially nonsmooth function  $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ , with  $(P_f)$  the corresponding optimal control problem. Thus, for most of our analysis we will adopt a more general perspective by considering functions  $f$  that belong to a larger family than the one defined by ReLU neural networks only. In such a setting  $f = \mathcal{N}$  becomes one particular instance. Concerning such a general  $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ , throughout we invoke the following:

For every  $y \in \mathbb{R}$  the function  $x \mapsto f(x, y)$  is Lebesgue measurable. Moreover (A<sub>1</sub>)

$$f(\cdot, 0) \in L^\infty(\Omega).$$

$f$  is Lipschitz continuous in  $y$  on bounded sets, that is, for every  $M > 0$ , there exists a (A<sub>2</sub>)  
constant  $L = L(M) > 0$ , such that for all  $y_1, y_2 \in (-M, M)$  :

$$|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2|, \quad \text{for almost every } x \in \Omega.$$

For almost every  $x \in \Omega$  the function  $y \mapsto f(x, y)$  is monotone increasing. (A<sub>3</sub>)

For almost every  $x \in \Omega$  the function  $y \mapsto f(x, y)$  is directionally differentiable with (A<sub>4</sub>)  
its directional derivative at  $y \in \mathbb{R}$  in direction  $h \in \mathbb{R}$  given by

$$f'_x(y; h) = \lim_{t_n \rightarrow 0^+} \frac{f(x, y + t_n h) - f(x, y)}{t_n}.$$

Note that we do not necessarily require  $y \mapsto f(x, y)$  to be strictly monotonically increasing. It is further clear that a ReLU neural network  $\mathcal{N} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ , restricted to  $\Omega \times \mathbb{R}$ , which is monotonically increasing in the second variable satisfies assumptions (A<sub>1</sub>)–(A<sub>4</sub>). In fact, it is even globally Lipschitz continuous in the second variable.

Observe further that  $\mathcal{N}$  is even Hadamard directionally differentiable with respect to the second variable. Then, using the chain rule for Hadamard directionally differentiable functions [9, Proposition 2.47], we can state a recursion formula for  $\mathcal{N}'_x(y; h)$ . For this, we first confine ourselves to a two-layer



architecture and generalize afterwards. Indeed, let  $z := (x, y)$  and  $N^{(2)}(z) = W_2 \cdot \sigma(W_1 z + b_1) + b_2$ ,  $W_2 \in \mathbb{R}^{1 \times n_1}$ ,  $W_1 \in \mathbb{R}^{n_1 \times (d+1)}$ ,  $b_1 \in \mathbb{R}^{n_1}$ ,  $b_2 \in \mathbb{R}$ . Then we have for any  $y, h \in \mathbb{R}$  that

$$(N^{(2)})'_x(y; h) = W_2 \cdot \left( \mathbb{1}_{(0, \infty)}(W_1 z + b_1) W_1(:, n_0) h + \mathbb{1}_{\{0\}}(W_1 z + b_1) \max(0, W_1(:, n_0) h) \right). \quad (3.1)$$

Here,  $W_1(:, n_0)$  denotes the last column of  $W_1$ , and  $\mathbb{1}_{(0, \infty)}(W_1 z + b_1)$  is a diagonal matrix, whose diagonal consists of the vector resulting from the componentwise action of the characteristic function  $\mathbb{1}_{(0, \infty)} : \mathbb{R} \rightarrow \{0, 1\}$ , with  $\mathbb{1}_{(0, \infty)}(t) = 1$  if  $t \in \mathcal{M} := (0, \infty)$  and  $\mathbb{1}_{(0, \infty)}(t) = 0$  otherwise, on the vector  $W_1 z + b_1$  – similarly for the second summand in (3.1). In general, recursively for  $N^{(\ell)} = W_\ell \sigma(N^{\ell-1}(z)) + b_\ell$  we have

$$(N^{(\ell)})'_x(y; h) = W_\ell \cdot \left( \mathbb{1}_{(0, \infty)}(N^{(\ell-1)}(z)) (N^{(\ell-1)})'_x(y; h) + \mathbb{1}_{\{0\}}(N^{(\ell-1)}(z)) \max(0, (N^{(\ell-1)})'_x(y; h)) \right). \quad (3.2)$$

Comparing (3.1)–(3.2) with

$$\nabla \mathcal{N}(x) = W_L \cdot \sigma'(N^{(L-1)}(x)) \cdot W_{L-1} \cdot \dots \cdot \sigma'(N^{(1)}(x)) \cdot W_1, \quad (3.3)$$

the weak gradient of  $\mathcal{N}$  [7], we note that while (3.3) holds almost everywhere, the formulas for the directional derivatives hold at every point.

For our analysis we will make use of the space

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

which is a separable Hilbert space equipped with the inner product  $(y, v)_Y := \int_\Omega \Delta y \Delta v + \nabla y \nabla v + yv \, dx$  and it is compactly embedded in  $H_0^1(\Omega)$  [10].

We will also denote by  $F$  and  $N$  the Nemytskii operators  $y \mapsto F(y)$  and  $y \mapsto N(y)$ , with  $F(y)(x) = f(x, y)$ ,  $N(y)(x) = \mathcal{N}(x, y)$  for  $y$  in some  $L^p$  space. Note that under the assumptions  $(A_1)$ – $(A_2)$ ,  $F$  is a well-defined operator from  $L^\infty(\Omega) \rightarrow L^\infty(\Omega)$  which is Lipschitz continuous on bounded sets. Moreover  $N : L^p(\Omega) \rightarrow L^p(\Omega)$  is Lipschitz continuous for every fixed  $1 \leq p \leq \infty$ .

### 3.1 Existence and uniqueness

The next proposition deals with the existence of solutions for the state equation of  $(P_f)$  as well as continuity properties of the corresponding control-to-state map. As the proof is rather standard, we will only provide a sketch. We also note that here we closely follow [10, Proposition 2.1] where  $f(x, y) = \max(y, 0)$ .

**Proposition 3.1.** *Let  $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$  satisfy the assumptions  $(A_1)$ – $(A_3)$ , and let  $u \in L^p(\Omega)$  for some  $p \geq 2$  and  $p > \frac{d}{2}$ . Then the state equation*

$$\begin{cases} -\Delta y + f(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \quad (E)$$

*admits a unique solution  $y \in Y \cap C^{0,a}(\overline{\Omega})$ , for some Hölder exponent  $a > 0$  depending only on  $p, d$  and  $\Omega$ . Furthermore, for every  $M > 0$  there exists a constant  $c_a > 0$  (that depends on  $M$ ) such that*

$$\|y\|_{C^{0,a}(\overline{\Omega})} \leq c_a \|u - f(\cdot, 0)\|_{L^p(\Omega)}, \quad \text{for all } \|u\|_{L^p(\Omega)} \leq M. \quad (3.4)$$

*Finally, the control-to-state map  $S : u \mapsto y$  has the following properties:*

- (i)  $S : L^p(\Omega) \rightarrow H_0^1(\Omega)$  is globally Lipschitz and weakly-strongly continuous.
- (ii)  $S : L^p(\Omega) \rightarrow Y$  is weakly-weakly continuous.
- (iii)  $S : L^p(\Omega) \rightarrow Y \cap C^{0,\alpha}(\overline{\Omega})$  is Lipschitz continuous on bounded sets of  $L^p(\Omega)$ .

Whenever the nonlinearity is represented by a ReLU network  $\mathcal{N}$  (i.e.  $f = \mathcal{N}$ ), then the last Lipschitz continuity is also global and the constant  $c_\alpha$  does not depend on  $M$ .

*Proof.* We only provide a general sketch of proof for the sake of completeness, since the arguments are standard. The first step is to consider a family of truncations  $f_k$  of  $f$  where for  $k > 0$  we define

$$f_k(x, y) := \begin{cases} f(x, -k), & \text{if } y < -k, \\ f(x, y), & \text{if } |y| \leq k, \\ f(x, k), & \text{if } y > k. \end{cases} \quad (3.5)$$

The associated state equation reads

$$\begin{cases} -\Delta y + f_k(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega. \end{cases} \quad (E_k)$$

An application of the Browder-Minty theorem gives that  $(E_k)$  admits a unique solution  $y \in H_0^1(\Omega)$ . Since  $u \in L^2(\Omega)$  and  $f_k$  is bounded, in particular  $F_k : L^2(\Omega) \rightarrow L^2(\Omega)$  so we have that  $\Delta y_k \in L^2(\Omega)$  as well, and hence  $y_k \in Y$ . The same theorem gives that the corresponding control-to-state map  $S_k : L^p(\Omega) \rightarrow H_0^1(\Omega)$  is Lipschitz continuous (with a Lipschitz constant that does not depend on  $k$ ) and also weakly-strongly continuous. In fact  $S_k : L^p(\Omega) \rightarrow Y$  is also globally Lipschitz by using the fact  $F_k$  is so as well (but here the Lipschitz constant depends on  $k$ ). Moreover, one can easily check that the map from  $S_k : L^p(\Omega) \rightarrow Y$  is also weakly-weakly continuous.

By adapting standard results of Stampachia's method, e.g. [30, Theorem 4.5], and crucially also using that  $p > \frac{d}{2}$ , one can show that  $y_k \in L^\infty(\Omega)$  and there exists a constant  $c_\infty > 0$  independent of  $u$  such that

$$\|y_k\|_{L^\infty(\Omega)} \leq c_\infty \|u - f_k(\cdot, 0)\|_{L^p(\Omega)} = c_\infty \|u - f(\cdot, 0)\|_{L^p(\Omega)}. \quad (3.6)$$

By choosing a large enough  $k > 0$ , it is clear that  $y_k = y$  is the solution of the problem  $(E)$ , which is also unique from the monotonicity of  $f$ . The Lipschitz continuity of  $S : L^p(\Omega) \rightarrow H_0^1(\Omega)$  as well as its weak-strong continuity follow readily. However, for the Lipschitz continuity of  $S : L^p(\Omega) \rightarrow Y$  one needs to restrict to bounded sets of  $L^p(\Omega)$  since the corresponding Lipschitz constant for  $F_k$  depends on  $k$ . Note that this is not the case when  $f$  is globally Lipschitz, e.g., a ReLU neural network.

Finally, regarding the Hölder regularity and the corresponding estimate (3.4), these follow from an application of [17, Theorem 8.29]. Thus, it remains to show the Lipschitz continuity with respect to the  $C^{0,\alpha}$  semi-norm. For this purpose, let  $u_i \in L^p(\Omega)$  and  $y_i = S(u_i)$ ,  $i = 1, 2$ , and define

$$\xi := \begin{cases} \frac{F(y_1) - F(y_2)}{y_1 - y_2} & \text{on } \{x \in \Omega : y_1(x) \neq y_2(x)\}, \\ 0 & \text{on } \{x \in \Omega : y_1(x) = y_2(x)\}. \end{cases}$$

Since  $y_1, y_2 \in C^{0,\alpha}(\overline{\Omega})$ ,  $\xi \geq 0$  is essentially bounded by a local Lipschitz constant of  $F$  depending on  $\max\{\|y_1\|_\infty, \|y_2\|_\infty\}$ . Then, by definition of  $\xi$  we have that  $y_1 - y_2$  solves the equation

$$\begin{cases} -\Delta y + \tilde{f}(y) = u_1 - u_2, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \quad (3.7)$$

where  $\tilde{f}(y) = \xi y$ . Obviously, the function  $\tilde{f}$  satisfies the assumptions of this proposition, and in particular we can apply the estimate (3.4) to obtain

$$\|y_1 - y_2\|_{C^{0,\alpha}(\overline{\Omega})} \leq \tilde{c}_a \|u_1 - u_2 - \tilde{f}(0)\|_{L^p(\Omega)} = \tilde{c}_a \|u_1 - u_2\|_{L^p(\Omega)}.$$

Note that, also in view of the discussion right after, the constant  $\tilde{c}_a$  depends on the Lipschitz constant of  $\tilde{f}$ , that is  $\|\xi\|_{L^\infty(\Omega)}$ , which is controlled by  $\|y_1\|_{L^\infty(\Omega)} + \|y_2\|_{L^\infty(\Omega)}$ . The latter can be controlled by  $\|u_1\|_{L^p(\Omega)}, \|u_2\|_{L^p(\Omega)}$ . Thus, in the end we get the required Lipschitz continuity on bounded sets of  $L^p(\Omega)$ . Again for a globally Lipschitz  $f$ , as it is the case for a ReLU network  $\mathcal{N}$ , the Lipschitz continuity  $S : L^p(\Omega) \rightarrow C^{0,\alpha}(\overline{\Omega})$  is also global.  $\square$

We add a remark regarding the constant  $M > 0$  in the statement of Proposition 3.1. We note that in the case of locally Lipschitz  $f$ , the constant  $c_a$  in (3.4) will also depend on  $\|u\|_{L^p(\Omega)}$ . This is because  $c_a$  depends on the local Lipschitz constant  $L > 0$  such that  $\|f(\cdot, y) - f(\cdot, 0)\|_{L^p(\Omega)} \leq L\|y\|_{L^p}$ . This constant  $L$  depends on  $\|y\|_{L^\infty(\Omega)}$  which is controlled by  $\|u\|_{L^p(\Omega)}$ . More specifically, for  $u$  in a bounded subset of  $L^p(\Omega)$ , one can take a uniform constant  $c_a$  in (3.4). If  $f$  is globally Lipschitz, then  $c_a$  can be chosen independently of  $\|u\|_{L^p(\Omega)}$ .

### 3.2 Control-to-state map (differentiability)

We will proceed by stating differentiability properties of the solution operator  $S$  of the state equation (E), under the assumptions (A<sub>1</sub>)–(A<sub>4</sub>).

We start by mentioning that it is easy to check after an application of the dominated convergence theorem that under these assumptions,  $F : L^\infty(\Omega) \rightarrow L^p(\Omega)$  is Hadamard directionally differentiable for  $1 \leq p < \infty$ . The directional derivative  $F'(y; h) \in L^p(\Omega)$  is given by

$$F'(y; h)(x) = f'_x(y(x); h(x)). \quad (3.8)$$

In fact if  $y, h \in L^\infty(\Omega)$  and  $(h_n)_{n \in \mathbb{N}}$  is a bounded sequence in  $L^\infty(\Omega)$  with  $h_n \rightarrow h$  in  $L^p(\Omega)$ , and  $t_n \rightarrow 0^+$ , then we have

$$\frac{F(y + t_n h_n) - F(y)}{t_n} \rightarrow F'(y; h) \quad \text{in } L^p(\Omega). \quad (3.9)$$

In the case of a ReLU neural network, we also have that  $N : L^p(\Omega) \rightarrow L^p(\Omega)$  for  $1 \leq p < \infty$  is Hadamard directional differentiable with the directional derivative  $N'_x(y; h) \in L^p(\Omega)$ , given by the analogous formula to (3.8). The following proposition also is a simple adaptation of [10, Theorem 2.2] to our case.

**Proposition 3.2.** *The control-to-state map  $S : L^p(\Omega) \rightarrow Y$  of (E) is Hadamard directional differentiable, that is, given  $u \in L^p(\Omega)$ , a direction  $h \in L^p(\Omega)$ ,  $(h_n)_{n \in \mathbb{N}} \subset L^p(\Omega)$  with  $h_n \rightarrow h$  in  $L^p(\Omega)$  and  $t_n \rightarrow 0^+$  we have*

$$\frac{S(u + t_n h_n) - S(u)}{t_n} \rightarrow S'(u; h) \quad \text{in } Y.$$

Moreover,  $S'(u; h) := z_h \in Y \cap C^{0,\alpha}(\overline{\Omega})$  and it is the unique solution of

$$\begin{cases} -\Delta z_h + F'(y; z_h) = h, & \text{in } \Omega, \\ z_h = 0, & \text{on } \partial\Omega, \end{cases} \quad (K)$$

where  $y = S(u)$ .

*Proof.* We denote  $u_n := u + t_n h_n \in L^p(\Omega)$ ,  $y_n := S(u_n)$  and  $z_n := \frac{y_n - y}{t_n}$  for every  $n \in \mathbb{N}$ . From the Lipschitz continuity of  $S$ , we deduce the existence of some  $z_h \in Y$  such that  $z_n \rightharpoonup z_h$  (weakly) in  $Y$  and  $z_n \rightarrow z_h$  (strongly) in  $H_0^1(\Omega)$  (due to the compact embedding) along a subsequence. Furthermore we have

$$-\Delta z_n + \frac{F(y_n) - F(y)}{t_n} = h_n. \quad (3.10)$$

According to (3.9), the second term of the left hand side in (3.10) converges to  $F'(y, z_h)$  strongly in  $L^2(\Omega)$ , and since  $h_n \rightarrow h$  in  $L^2(\Omega)$ , we also get that  $-\Delta z_n$  converges strongly to  $-\Delta z_h$ . Hence  $z_n \rightarrow z_h$  strongly in  $Y$ . Taking the weak limit in  $L^2(\Omega)$  in (3.10), we get that  $z_h$  satisfies  $(K)$ . Note now that fixing  $y \in L^\infty(\Omega)$ , we have that  $F'(y; \cdot) : L^p(\Omega) \rightarrow L^p(\Omega)$  is the Nemytskii operator induced by  $(x, z) \mapsto f'_x(y(x); z(x))$ . It can be easily checked that the latter map satisfies the assumptions  $(A_1)$ – $(A_3)$ . Hence, Theorem 3.1 applies and thus  $z_h \in C^{0,a}(\bar{\Omega})$  as well. Since this solution is unique, the convergence  $z_n \rightarrow z_h$  holds along the whole sequence.  $\square$

### 3.3 Approximation results

In this section we are concerned with certain approximation results regarding the state equation  $(E)$ . In particular, we consider approximating sequences for the nonlinear nonsmooth function  $f$ , and we show approximation results for the sequence of solutions of the corresponding approximating state equations to the solution of the limit problem. We will consider approximations on two levels: The first level of approximation arises from the approximation of  $f$  by a sequence of ReLU neural networks  $\mathcal{N}_n$  in the sense of (2.7) and can be thought of as the capability of ReLU network-informed PDEs to approximate some ground truth nonsmooth model. The second level of approximation considers the approximation of a ReLU network-informed PDE by (a sequence of) PDEs that correspond to a smoothing of the network. Concerning the latter we will also briefly discuss some complications which arise when the smoothed network results from simply smoothing the ReLU activation function. We recall that smoothing a general function  $f$  is a typical first step in standard methods that study the optimal control of PDEs that contain nonsmooth components. In this context, one often regularizes these components and subsequently considers the limit behaviour as regularization vanishes; compare, e.g., [5, 10, 26].

We start by studying the approximation of a general function  $f$  by a sequence of ReLU neural networks.

**Proposition 3.3.** *Suppose that  $f : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$  satisfies  $(A_1)$ – $(A_4)$  with the additional assumption that  $f \in W_{loc}^{1,\infty}(\mathbb{R}^d \times \mathbb{R})$ . Moreover, let  $u \in L^p(\Omega)$  for some  $p \geq 2$  and  $p > \frac{d}{2}$ . Then, given  $K > 0$ , there exists a sequence  $(\mathcal{N}_n)_{n \in \mathbb{N}}$  of ReLU neural networks  $\mathcal{N}_n : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ ,  $n \in \mathbb{N}$ , such that*

$$\mathcal{N}_n \rightarrow f, \quad \text{in } W^{1,\infty}(\Omega \times (-K, K)). \quad (3.11)$$

Furthermore, there exists  $K > 0$  such that for sufficiently large  $n \in \mathbb{N}$ , the approximating learning-informed PDE

$$\begin{cases} -\Delta y + \mathcal{N}_n(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \quad (E_{\mathcal{N}_n})$$

has a unique solution  $y_n \in Y \cap C^{0,a}(\bar{\Omega})$  satisfying (3.4), with the corresponding control-to-state-map  $S_n$  satisfying (i)–(iii) of Proposition 3.1 (with global Lipschitz constants). Furthermore, it holds that

$$y_n \rightarrow y, \quad \text{strongly in } Y \text{ and in } C^{0,a}(\bar{\Omega}), \quad (3.12)$$

where  $y \in Y \cap C^{0,a}(\overline{\Omega})$  is the unique solution of (E).

*Proof.* Given  $K > 0$ , the  $W^{1,\infty}(\Omega \times (-K, K))$  convergence (3.11) for some sequence  $(\mathcal{N}_n)_{n \in \mathbb{N}}$  follows directly from (2.7). Note that every  $\mathcal{N}_n$  satisfies assumptions  $(A_1)$ – $(A_4)$  with the possible exception of  $(A_3)$ , that is, monotonicity in the second variable. We note however that this monotonicity was only used previously in the application of Browder-Minty's theorem in order to show that the operator  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$  with

$$\langle A(y), z \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} := \int_{\Omega} \nabla y \nabla z \, dx + \int_{\Omega} f(x, y) z \, dx,$$

is strongly monotone. We argue that this is the case also for the corresponding operator  $A_{n,k}$  (where  $f$  has been substituted by the truncation  $\mathcal{N}_{n,k}$  of  $\mathcal{N}_n$  as in (3.5)). Indeed, let  $c_{\Omega}$  to be the Poincaré (inequality) constant, set

$$K := c_a(\|u\|_{L^p(\Omega)} + (\|f(\cdot, 0)\|_{L^\infty(\Omega)} + 1)|\Omega|^{1/p}) > 0, \quad (3.13)$$

and fix  $k > K$ . Then we have for  $y_1, y_2 \in H_0^1(\Omega)$

$$\begin{aligned} \langle A_{n,k}(y_1) - A_{n,k}(y_2), y_1 - y_2 \rangle &\geq \frac{1}{(c_{\Omega} + 1)^2} \|y_1 - y_2\|_{H_0^1(\Omega)}^2 \\ &\quad + \int_{\Omega} (\mathcal{N}_{n,k}(x, y_1) - \mathcal{N}_{n,k}(x, y_2))(y_1 - y_2) \, dx. \end{aligned} \quad (3.14)$$

Due to the fact that  $\mathcal{N}_{n,k} \rightarrow f_k$  in  $W^{1,\infty}(\Omega \times \mathbb{R})$ , we have that for large enough  $n \in \mathbb{N}$

$$\begin{aligned} &\int_{\Omega} (\mathcal{N}_{n,k}(x, y_1) - \mathcal{N}_{n,k}(x, y_2))(y_1 - y_2) \, dx \\ &= \int_{\Omega} ((\mathcal{N}_{n,k} - f_k)(x, y_1) - (\mathcal{N}_{n,k} - f_k)(x, y_2))(y_1 - y_2) \, dx \\ &\quad + \underbrace{\int_{\Omega} (f_k(x, y_1) - f_k(x, y_2))(y_1 - y_2) \, dx}_{\geq 0} \\ &\geq -\|\nabla(\mathcal{N}_{n,k} - f_k)\|_{L^\infty(\Omega \times \mathbb{R})} \|y_1 - y_2\|_{L^2(\Omega)}^2. \end{aligned}$$

Since the term  $\|\nabla(\mathcal{N}_{n,k} - f_k)\|_{L^\infty(\Omega \times \mathbb{R})}$  can be arbitrarily small for large enough  $n \in \mathbb{N}$ , we have that the last negative term can be absorbed into  $\frac{1}{(c_{\Omega} + 1)^2} \|y_1 - y_2\|_{H_0^1(\Omega)}^2$  in (3.14) and hence the operator  $A_{n,k}$  is strongly monotone for large enough  $n \in \mathbb{N}$ . Thus the corresponding approximating truncated problem  $(E_{\mathcal{N}_{n,k}})$  has a solution. In order to show that  $(E_{\mathcal{N}_n})$  has a solution  $y_n \in Y \cap C^{0,a}(\overline{\Omega})$  as well, we argue exactly as in the proof of Proposition 3.1, keeping in mind that since  $\|\mathcal{N}_{n,k}(\cdot, 0) - f(\cdot, 0)\|_{L^\infty(\Omega)} = \|\mathcal{N}_n(\cdot, 0) - f(\cdot, 0)\|_{L^\infty(\Omega)} \rightarrow 0$  for large enough  $k \in \mathbb{N}$  the  $L^\infty$  norm of  $y_{n_k}$  will be bounded by  $K$ . The estimate (3.4) and the corresponding continuity properties of the control-to-state map  $S_n$  follow analogously. We note that since the Lipschitz constants of  $\mathcal{N}_n$  are uniformly bounded (in particular their  $k$ -truncation), the same holds for the Lipschitz constants of the maps  $S_n : L^p(\Omega) \rightarrow H_0^1(\Omega)$  and  $S_n : L^p(\Omega) \rightarrow Y \cap C^{0,a}(\overline{\Omega})$ .

Using the convergence of  $\mathcal{N}_n$  to  $f$ , it is easy to check that the sequence  $(y_n)_{n \in \mathbb{N}}$  is bounded in  $Y$  and hence there exists an (unrelabeled) subsequence and  $\zeta \in Y$  such that  $y_n \rightharpoonup \zeta$  weakly in  $Y$  and

$y_n \rightarrow \zeta$  strongly in  $H_0^1(\Omega)$ . Note that in view of (3.4), we also have  $\|\zeta\|_{L^\infty(\Omega)} \leq K$ . Observe now that for every  $z \in L^2(\Omega)$

$$\begin{aligned} \left| \int_{\Omega} \mathcal{N}_n(x, y_n) z \, dx - \int_{\Omega} f(x, \zeta) z \, dx \right| &\leq \int_{\Omega} |\mathcal{N}_n(x, y_n) z - \mathcal{N}_n(x, \zeta) z| \, dx \\ &\quad + \int_{\Omega} |\mathcal{N}_n(x, \zeta) z - f(x, \zeta) z| \, dx \\ &\leq L \|y_n - \zeta\|_{L^2(\Omega)} \|z\|_{L^2(\Omega)} \\ &\quad + \|\mathcal{N}_n - f\|_{L^\infty(\Omega \times (-K, K))} \|z\|_{L^2(\Omega)} \rightarrow 0 \end{aligned}$$

as  $n \rightarrow \infty$ , where we also used the fact that the Lipschitz constants  $L_n$  of  $\mathcal{N}_n$  on  $\Omega \times [-K, K]$  are uniformly bounded (say by  $L$ ). Hence  $\mathcal{N}_n(x, y_n) \rightharpoonup f(x, \zeta)$  in  $L^2(\Omega)$  and thus by taking weak limits in  $(E_{\mathcal{N}_n})$ , we have that  $\zeta \in Y$  is equal to the solution  $y$  of  $(E)$ . By uniqueness the weak convergence  $y_n \rightharpoonup y$  (in  $Y$ ) holds for the whole sequence. Note that the estimate

$$\|\Delta y_n - \Delta y\|_{L^2(\Omega)} = \|\mathcal{N}_n(\cdot, y_n) - f(\cdot, y)\|_{L^2(\Omega)} \leq L \|y_n - y\|_{L^2(\Omega)} + |\Omega| \|\mathcal{N}_n - f\|_{L^\infty(\Omega \times (-K, K))} \quad (3.15)$$

yields the strong convergence  $y_n \rightarrow y$  in  $Y$ , keeping in mind that  $\Delta$  induces the norm on  $Y$ .

Finally we would like to show that in addition we also have  $\|y_n - y\|_{C^{0,a}(\bar{\Omega})} \rightarrow 0$  as  $n \rightarrow \infty$ . For that purpose note that  $w_n := y_n - y$  solves the following PDE:

$$\begin{cases} -\Delta w_n = f(\cdot, y) - \mathcal{N}_n(\cdot, y_n), & \text{in } \Omega, \\ w_n = 0, & \text{on } \partial\Omega. \end{cases} \quad (3.16)$$

Using the same arguments as in Proposition 3.1, we have that

$$\|y_n - y\|_{C^{0,a}(\bar{\Omega})} = \|w_n\|_{C^{0,a}(\bar{\Omega})} \leq c_a \|f(\cdot, y) - \mathcal{N}_n(\cdot, y_n)\|_{L^p(\Omega)},$$

and thus it suffices to show that the latter is going to zero. We have

$$\|f(\cdot, y) - \mathcal{N}_n(\cdot, y_n)\|_{L^p(\Omega)} \leq \|f(\cdot, y) - f(\cdot, y_n)\|_{L^p(\Omega)} + \|\mathcal{N}_n(\cdot, y_n) - f(\cdot, y_n)\|_{L^p(\Omega)}.$$

The first term on the right hand side tends to zero since  $y_n \rightarrow y$  a.e. (up to a subsequence) and by applying the dominated convergence theorem, while the second term vanishes since  $\|\mathcal{N}_n - f\|_{L^\infty(\Omega \times (-K, K))} \rightarrow 0$ .  $\square$

We will now pass to the second level of approximation, which is more related to techniques in the optimal control of nonsmooth PDEs, where an optimality system is derived as a limit of optimality systems corresponding to smooth approximations for the PDE. Here, we will initially consider the learning-informed case and study smoothings of the involved ReLU network  $\mathcal{N}$ , also providing a word of caution about some complications that can result from such smoothings.

In particular, here we will consider the following ReLU informed semilinear PDE

$$\begin{cases} -\Delta y + \mathcal{N}(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \quad (E_{\mathcal{N}})$$

and its corresponding smoothed version

$$\begin{cases} -\Delta y + \mathcal{N}_\epsilon(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega. \end{cases} \quad (E_{\mathcal{N}_\epsilon})$$

A very natural way to produce smooth approximations  $\mathcal{N}_\epsilon$  of  $\mathcal{N}$  is simply via smoothing its activation function. This type of smoothing, here referred to as *canonical smoothing* of  $\mathcal{N}$ , is studied in detail in [15]. Note, however, that nonmonotonicity of  $\mathcal{N}_\epsilon$  may arise after such a smoothing. In fact, as we saw in Proposition 3.3, we were able to prove strong monotonicity for the operator  $A_{n,k}$  (for large enough  $n$ ) despite the fact that  $\mathcal{N}_n$  was not monotone itself. This was done by taking advantage of the  $L^\infty$  convergence of  $\nabla \mathcal{N}_{n,k}$  to  $\nabla f_k$ , the latter being positive almost everywhere. But in the case of a canonical smoothing  $\mathcal{N}_\epsilon$ , the convergence of  $\nabla \mathcal{N}_\epsilon$  to  $\nabla \mathcal{N}$  as  $\epsilon \rightarrow 0$  can only be guaranteed to hold with respect to the  $L^p$  norm for every  $1 \leq p < \infty$ , see [15, Remark 2.9] making the application of the Browder-Minty theorem problematic, see the corresponding section in [15].

We stress here that our purpose of using  $\mathcal{N}_\epsilon$  is to obtain a limiting optimality system as  $\epsilon \rightarrow 0$  for the optimal control with respect to the learning-informed PDE ( $P_{\mathcal{N}}$ ). Hence, for the following proposition, as well as for the passage to the limit later on, we will assume that  $\mathcal{N}_\epsilon$  remains monotone with respect to the second variable, a property that can be achieved theoretically (but perhaps not practically in implementations) via convolution; see also Remark 3.5. We thus set here

$$\mathcal{N}_\epsilon(x, y) := \int_{\mathbb{R}} \rho_\epsilon(y - z) \mathcal{N}(x, z) dz, \quad (3.17)$$

where  $(\rho_\epsilon)_{\epsilon>0}$  is the family of standard mollifiers [33]. Note that we have  $\mathcal{N}_\epsilon \rightarrow \mathcal{N}$  uniformly as  $\epsilon \rightarrow 0$ ,  $\|\nabla \mathcal{N}_\epsilon\|_{L^\infty(\Omega \times \mathbb{R})}$  is uniformly bounded, and each  $\mathcal{N}_\epsilon$  satisfies the assumptions  $(A_1)$ – $(A_4)$ .

Due to the linear growth of  $\mathcal{N}_\epsilon$  and the  $L^\infty$  bound for its gradient, we have that the Nemytskii operator  $N_\epsilon$  with  $N_\epsilon(y)(x) = \mathcal{N}_\epsilon(x, y)$  is continuously Fréchet differentiable from  $L^p(\Omega) \rightarrow L^q(\Omega)$  for every  $1 \leq q < p < \infty$ , with derivative  $N'_\epsilon : L^p(\Omega) \rightarrow \mathcal{L}(L^p(\Omega), L^q(\Omega))$  defined as

$$(N'_\epsilon(y)h)(x) = \nabla_y \mathcal{N}_\epsilon(x, y(x))h(x), \quad \text{for a.e. } x \in \Omega. \quad (3.18)$$

Consequently,  $N_\epsilon$  is also continuously Fréchet differentiable as a function from  $H_0^1(\Omega) \rightarrow L^q(\Omega)$ , for any  $1 \leq q < \frac{2d}{d-2}$  ( $d \geq 3$ ), for any  $1 \leq q < \infty$ , ( $d = 2$ ), and for any  $1 \leq q \leq \infty$  ( $d = 1$ ). In particular,  $N_\epsilon$  is continuously Fréchet differentiable from  $H_0^1(\Omega) \rightarrow L^2(\Omega)$  for any dimension  $d \geq 1$ . Recall that  $\nabla_y \mathcal{N}_\epsilon(x, y(x)) \in L^\infty(\Omega \times \mathbb{R})$  with its explicit form derived

$$\nabla \mathcal{N}_\epsilon(x) = W_L \cdot \sigma'_\epsilon(N_\epsilon^{(L-1)}(x)) \cdot W_{L-1} \cdot \dots \cdot \sigma'_\epsilon(N_\epsilon^{(1)}(x)) \cdot W_1. \quad (3.19)$$

**Proposition 3.4.** *Let  $\mathcal{N} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$  be a ReLU network that satisfies assumptions  $(A_1)$ – $(A_4)$  and let  $(\mathcal{N}_\epsilon)_{\epsilon>0}$  be a smoothing of  $\mathcal{N}$  defined as in (3.17). Moreover, let  $u \in L^p(\Omega)$  for some  $p \geq 2$  and  $p > \frac{d}{2}$ . Then the smoothed state equation  $(E_{\mathcal{N}_\epsilon})$  has a unique solution  $y_\epsilon \in Y \cap C^{0,\alpha}(\bar{\Omega})$ . Furthermore the following hold:*

- (i) *The corresponding control-to-state map  $S_\epsilon$  satisfies the continuity properties (i)–(iii) of Proposition 3.1 (with global Lipschitz constants independent of  $\epsilon > 0$ ).*
- (ii)  *$S_\epsilon : L^p(\Omega) \rightarrow Y$  is Fréchet differentiable. Its derivative at  $u \in L^p(\Omega)$  in direction  $h \in L^p(\Omega)$  is given by the unique solution of*

$$\begin{cases} -\Delta z_h + N'_\epsilon(y_\epsilon)z_h = h, & \text{in } \Omega, \\ z_h = 0, & \text{on } \partial\Omega, \end{cases} \quad (K_\epsilon)$$

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

*Proof.* Since each  $\mathcal{N}_\epsilon$  satisfies the assumptions  $(A_1)$ – $(A_4)$  and has at most linear growth in the second variable, we can directly apply the Browder-Minty theorem, i.e. without the truncation argument, and as before show that the smoothed state equation  $(E_{\mathcal{N}_\epsilon})$  has a unique solution  $y_\epsilon \in Y \cap C^{0,\alpha}(\overline{\Omega})$ .

For (i) note that here the continuity properties (i)–(iii) of Proposition 3.1 follow as before. The fact that the Lipschitz constants are global, i.e. independent of  $\epsilon > 0$ , follows from the fact that  $\|\nabla \mathcal{N}_\epsilon\|_{L^\infty}$  and hence the Lipschitz constants of  $\mathcal{N}_\epsilon$  are uniformly bounded in  $\epsilon$ .

For (ii), note first that since  $N'_\epsilon(y_\epsilon)$  is nonnegative, the equation  $(K_\epsilon)$  has a unique solution  $z_h \in Y \cap C^{0,\alpha}(\overline{\Omega})$ . Then, given the continuous Fréchet differentiability of  $N_\epsilon$ , an application of the implicit function theorem gives the Fréchet differentiability of  $S_\epsilon : L^p(\Omega) \rightarrow Y$ .

□

**Remark 3.5.** One can get results analogous to Proposition 3.4 for the problem involving the function  $f$  under assumptions  $(A_1)$ – $(A_4)$ . Indeed, using the smoothing  $f_\epsilon(x, y) := \int_{\mathbb{R}} \rho_\epsilon(y - z) f(x, z) dz$  for almost every  $x \in \Omega$ , we have that for almost every  $x \in \Omega$ ,  $f_\epsilon(x, \cdot) \in C^\infty(\mathbb{R})$ , and that it is monotonically increasing. Moreover,  $f_\epsilon$  is uniformly Lipschitz continuous in  $y$  on bounded sets, i.e., for every  $M > 0$ , there exists a constant  $L = L(M)$  independent of  $\epsilon > 0$  such that for every  $y_1, y_2 \in (-M, M)$

$$|f_\epsilon(x, y_1) - f_\epsilon(x, y_2)| \leq L|y_1 - y_2|, \quad \text{for almost every } x \in \Omega. \quad (3.20)$$

Furthermore, for every  $M > 0$ , there exists a constant  $c > 0$  such that for all  $y \in (-M, M)$

$$|f_\epsilon(x, y) - f(x, y)| \leq c\epsilon \quad \text{for almost every } x \in \Omega. \quad (3.21)$$

The analogous inequalities in  $L^p$  hold for the superposition operators  $F_\epsilon$  when restricting  $y$  to a  $L^\infty(\Omega)$  ball. Then by considering

$$\begin{cases} -\Delta y + f_\epsilon(\cdot, y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \quad (E_{f_\epsilon})$$

we have that analogous results to the ones of Proposition 3.4 hold with the only difference, that the Lipschitz continuity of  $S_\epsilon : L^p(\Omega) \rightarrow Y \cap C^{0,\alpha}(\overline{\Omega})$  is local (on bounded sets of  $L^p(\Omega)$ ), but still with Lipschitz constants independent of  $\epsilon$ . Furthermore, in that case we can only show Gâteaux differentiability for  $S_\epsilon : L^p(\Omega) \rightarrow Y$ , with the derivative at  $u \in L^p(\Omega)$  in direction  $h \in L^p(\Omega)$  given by the solution to  $(K_\epsilon)$  where  $N'_\epsilon$  is substituted by  $F'_\epsilon(y)(x) := \partial_y f_\epsilon(x, y(x))$ .

## 4 Existence and approximation results of optimal control

We return now to the optimal control problem  $(P_f)$ . To be consistent in our notation, we denote by  $(P_{\mathcal{N}_\epsilon})$  the optimal control problem corresponding to the state equation  $(E_{\mathcal{N}_\epsilon})$ . We also define the reduced problem for  $(P_f)$  as follows:

$$\min_{u \in \mathcal{C}_{ad}} \mathcal{J}(u) := J(S(u), u) := \frac{1}{2} \|S(u) - g\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2, \quad (\hat{P}_f)$$

and analogously we define the reduced objectives  $\mathcal{J}_N$  and  $\mathcal{J}_{\mathcal{N}_\epsilon}$ . Further, we consider  $\mathcal{C}_{ad}$  to be a nonempty, bounded, closed, convex subset of  $L^p(\Omega)$  with  $p \geq 2$  and  $p > d/2$ .



**Proposition 4.1.** *The optimal control problem*

$$\min_{u \in \mathcal{C}_{ad}} \mathcal{J}(u) \quad (4.1)$$

has a solution.

*Proof.* The proof is a standard application of the direct method of calculus of variations. Let  $(u_n)_{n \in \mathbb{N}}$  be a minimizing sequence for (4.1) in  $L^p(\Omega)$ . Since  $\mathcal{C}_{ad}$  is bounded in  $L^p(\Omega)$ , there exists a subsequence  $u_{n_k} \rightharpoonup u$  for some  $u \in L^p(\Omega)$ , which also belongs to  $\mathcal{C}_{ad}$  by Mazur's theorem. From Proposition 3.1 (ii), we have that  $S(u_{n_k}) \rightarrow S(u)$  in  $L^2(\Omega)$ . Then the results follows from the lower semicontinuity of  $\mathcal{J}$ .  $\square$

Observe that in the case  $p = 2$ , boundedness of  $\mathcal{C}_{ad}$  is not necessary to get existence for (4.1), due to the term  $\frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2$  appearing in the objective.

The following proposition summarizes approximation properties of solutions of the optimal control problem  $(P_f)$  and approximations  $(P_{f_n})$  with  $f_n$  close to  $f$ .

**Proposition 4.2.** *Consider the general optimal control problem  $(P_f)$  and its corresponding control-to-state map  $S : L^p(\Omega) \rightarrow Y \cap C^{0,a}(\overline{\Omega})$ . Let  $(P_{f_n})$  be an optimal control problem that corresponds to  $f_n$  which is regarded to approximate  $f$  in a certain sense, and denote by  $S_n : L^p(\Omega) \rightarrow Y \cap C^{0,a}(\overline{\Omega})$  the corresponding approximating control-to-state map. Suppose also that the corresponding Nemytskii operators satisfy*

$$\|F_n(y) - F(y)\|_{L^p(\Omega)} \leq \epsilon_n, \quad \text{for all } y \in L^\infty(\Omega) \text{ with } \|y\|_{L^\infty(\Omega)} \leq M \quad (4.2)$$

for some  $M > 0$  and some  $\epsilon_n \downarrow 0$  possibly depending on  $M$ . Then the following hold:

(i) *There exists  $c > 0$  such that*

$$\|S_n(u) - S(u)\|_Y \leq c\epsilon_n, \quad \text{for all } u \in \mathcal{C}_{ad}, \quad (4.3)$$

(ii) *For every sequence of minimizers  $(u_n)_{n \in \mathbb{N}}$  of  $(P_{f_n})$  there exists a minimizer  $u$  of  $(P_f)$  such that, up to a subsequence*

$$u_n \rightharpoonup u \text{ in } L^p(\Omega), \quad u_n \rightarrow u \text{ in } L^2(\Omega), \quad \text{and} \quad S_n(u_n) \rightarrow S(u) \text{ in } H_0^1(\Omega).$$

(iii) *If  $u_n \rightarrow u$  in  $L^p(\Omega)$ , then  $S_n(u_n) \rightarrow S(u)$  in  $Y \cap C^{0,a}(\overline{\Omega})$ .*

If  $\mathcal{C}_{ad}$  is of the form  $\mathcal{C}_{ad} = \{u \in L^p(\Omega) : u_a \leq u \leq u_b, \text{ a.e. in } \Omega\}$  for some  $u_a, u_b \in L^p(\Omega)$ , then for the above subsequence we have  $u_n \rightarrow u$  in  $L^p(\Omega)$  and (iii) holds.

*Proof.* (i) We fix  $u \in \mathcal{C}_{ad}$  and write  $y_n := S_n(u)$ ,  $y := S(u)$ . By subtracting the two state equations, adding and subtracting  $F(y_n)$ , testing with  $y_n - y$  and using the monotonicity of  $f$  we have

$$\begin{aligned} \|\nabla y_n - \nabla y\|_{L^2(\Omega)}^2 &= - \int_{\Omega} (F_n(y_n) - F(y_n))(y_n - y) \, dx - \int_{\Omega} (F(y_n) - F(y))(y_n - y) \, dx \\ &\leq \|F_n(y_n) - F(y_n)\|_{L^2(\Omega)} \|y_n - y\|_{L^2(\Omega)} \\ &\leq \|F_n(y_n) - F(y_n)\|_{L^2(\Omega)} \|y_n - y\|_{H_0^1(\Omega)}. \end{aligned}$$

Using the Poincaré inequality, dividing the inequality above by  $\|y_n - y\|_{H_0^1(\Omega)}$  and using the fact that  $\|y_n\|_{L^\infty(\Omega)} \leq c_\infty \|u - F_n(0)\|_{L^p(\Omega)}$  as well as (4.2), we get the existence of  $c > 0$  satisfying the an estimate analogous to (4.3) for the  $H_0^1$  norm. The required estimate for the  $Y$ -norm is shown by considering

$$\|\Delta y_n - \Delta y\|_{L^2} \leq \|F_n(y_n) - F(y_n)\|_{L^2(\Omega)} + \|F(y_n) - F(y)\|_{L^2(\Omega)},$$

and taking advantage of (4.2) and the local Lipschitz continuity of  $F$ .

(ii) The first two convergences follow from a direct application of [16, Theorem 2.3]. We only need to prove the last one. By setting now  $y_n := S_n(u_n)$  and performing a similar estimation as before, we get

$$\|y_n - y\|_{H_0^1(\Omega)} \leq c (\|F_n(y) - F(y)\|_{L^p(\Omega)} + \|u_n - u\|_{H^{-1}(\Omega)}) \leq c(\epsilon + \|u_n - u\|_{H^{-1}(\Omega)}).$$

We then use the fact that  $u_n \rightharpoonup u$  in  $L^p(\Omega)$  implies  $u_n \rightarrow u$  in  $H^{-1}(\Omega)$ .

(iii) Suppose that  $u_n \rightarrow u$  in  $L^p(\Omega)$ . We need to show  $y_n \rightarrow y$  both in  $Y$  and  $C^{0,a}(\overline{\Omega})$ . For the former one it suffices to show  $\Delta y_n \rightarrow \Delta y$  in  $L^2(\Omega)$  which follows from the estimate

$$\|\Delta y_n - \Delta y\|_{L^2(\Omega)} \leq \|F_n(y_n) - F(y_n)\|_{L^2(\Omega)} + \|F(y_n) - F(y)\|_{L^2(\Omega)} + \|u_n - u\|_{L^2(\Omega)}.$$

Finally, for the convergence  $y_n \rightarrow y$  in  $C^{0,a}(\overline{\Omega})$ , we work similarly to the proof of Lipschitz continuity of  $S : L^p(\Omega) \rightarrow C^{0,a}(\overline{\Omega})$  in Proposition 3.1.

Finally note that if  $\mathcal{C}_{ad}$  is of the form  $\mathcal{C}_{ad} = \{u \in L^p(\Omega) : u_a \leq u \leq u_b, \text{ a.e. in } \Omega\}$  for some  $u_a, u_b \in L^p(\Omega)$ , then for the same subsequence as in (ii), the convergence  $u_n \rightarrow u$  in  $L^p(\Omega)$  follows from an application of the dominated convergence theorem, as well as  $u_n \rightarrow u$  in  $L^2(\Omega)$  which implies almost everywhere convergence up to subsequence.  $\square$

**Remark 4.3.** *A few remarks about Proposition 4.2 are in order. The main requirement for the approximation  $f_n$  to  $f$  is to have a well-defined control-to-state map  $S_n : L^p(\Omega) \rightarrow Y \cap C^{0,a}(\overline{\Omega})$  that satisfies the continuity properties (i)–(iii) of Proposition 3.1 as well as the approximation property (4.2). The former will guarantee existence for the optimal control problem and the latter will imply the approximation results of Proposition 4.2. We note that these will be satisfied for the approximating sequences  $\mathcal{N}_n \rightarrow f$ ,  $\mathcal{N}_\epsilon \rightarrow \mathcal{N}$ ,  $f_\epsilon \rightarrow f$  of Propositions 3.3, 3.4 and Remark 3.5, respectively.*

## 5 Optimality systems (Stationarity conditions)

In this section we study first-order optimality respectively stationarity systems for the optimal control problem under consideration. For the sake of generality, we study problem  $(P_f)$  when the nonlinear function  $f$  satisfies Assumptions  $(A_1)$ – $(A_4)$ . Note that this setting covers the case  $f = \mathcal{N}$ . In what follows we specifically focus on the (reduced) minimization problem (4.1) and are interested in necessary conditions satisfied by its local minimizers. Recall here that  $\bar{u}$  is a local minimizer for (4.1) if there exists an  $r > 0$  such that for every  $u \in L^p(\Omega)$  satisfying  $u \in B(\bar{u}, r) \cap \mathcal{C}_{ad} \subset L^p(\Omega)$  we have  $\mathcal{J}(\bar{u}) \leq \mathcal{J}(u)$ .

Depending on whether we take a primal or primal-dual view on characterizing stationarity of a feasible control, but also in view of different prerequisites on the problem data, in the sequel we derive different types of first-order conditions for  $(P_f)$  respectively (4.1).

**B-stationarity.** We start with a primal first-order necessary optimality condition for  $\bar{u}$  which relies on Fermat's principle, corresponding to the directional derivative of  $\mathcal{J}$  at  $\bar{u}$  being non-negative along any feasible direction.

To identify these feasible directions, we define the contingent cone of  $\mathcal{C}_{ad}$  at  $\bar{u} \in \mathcal{C}_{ad}$  as

$$T_{\mathcal{C}_{ad}}(\bar{u}) := \{h \in L^p(\Omega) : \exists t_n \downarrow 0 \text{ and } h_n \rightarrow h \in L^p(\Omega), \text{ such that for all } n \in \mathbb{N}, \bar{u} + t_n h_n \in \mathcal{C}_{ad}\}. \quad (5.1)$$

It can be shown, see for instance [9, Lemma 6.34], that if  $\mathcal{C}_{ad}$  is of the form

$$\mathcal{C}_{ad} = \{u \in L^p(\Omega) : u_a(x) \leq u(x) \leq u_b(x), \text{ for almost every } x \in \Omega\} \quad (5.2)$$

with  $u_a, u_b \in L^\infty(\Omega)$ ,  $u_a < u_b$  almost everywhere, then  $T_{\mathcal{C}_{ad}}(\bar{u})$  can be characterized by

$$T_{\mathcal{C}_{ad}}(\bar{u}) = \left\{ h \in L^p(\Omega) : \begin{array}{l} h(x) \geq 0, \text{ almost everywhere in } \{x \in \Omega : \bar{u}(x) = u_a(x)\} \\ h(x) \leq 0, \text{ almost everywhere in } \{x \in \Omega : \bar{u}(x) = u_b(x)\} \end{array} \right\}. \quad (5.3)$$

**Proposition 5.1.** [*B-stationarity*] *Suppose that  $\bar{u}$  is a local minimizer for the problem (4.1) and let  $\bar{y} = S(\bar{u}) \in C^{0,\alpha}(\bar{\Omega}) \cap Y$  be the associated state. Then the pair  $(\bar{u}, \bar{y})$  satisfies the following variational inequality*

$$\mathcal{J}'(\bar{u}; h) = \langle \bar{y} - g, S'(\bar{u}; h) \rangle + \alpha \langle \bar{u}, h \rangle \geq 0, \quad \text{for all } h \in T_{\mathcal{C}_{ad}}(\bar{u}). \quad (5.4)$$

*Proof.* The proof closely follows [21, Theorem 1]. Note that the minimization problem (4.1) can be equivalently written as

$$\min_{u \in L^p(\Omega)} V(u) := \mathcal{J}(u) + \mathcal{X}_{\mathcal{C}_{ad}}(u), \quad (5.5)$$

where  $\mathcal{X}_{\mathcal{C}_{ad}}$  denotes the indicator function of  $\mathcal{C}_{ad}$ . Since  $\bar{u}$  is a local minimizer of (5.5), we have the following Fermat rule, see [4, Theorem 6.1.9]:

$$D_{\uparrow}V(\bar{u})(h) \geq 0, \quad \text{for all } h \in L^p(\Omega), \quad (5.6)$$

where for a  $u \in \text{dom}(V)$ ,  $D_{\uparrow}V(u)(h)$  denotes the contingent epiderivative of  $V$  at  $u$  in direction  $h \in L^p(\Omega)$ , given by

$$D_{\uparrow}V(u)(h) = \liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \frac{V(u + th') - V(u)}{t}. \quad (5.7)$$

Hence from (5.5), (5.6) and (5.7) we have

$$\liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \left( \frac{\mathcal{J}(\bar{u} + th') - \mathcal{J}(\bar{u})}{t} + \frac{\mathcal{X}_{\mathcal{C}_{ad}}(\bar{u} + th') - \mathcal{X}_{\mathcal{C}_{ad}}(\bar{u})}{t} \right) \geq 0 \quad \text{for all } h \in L^p(\Omega). \quad (5.8)$$

Note that for all  $h \in L^p(\Omega)$  we have

$$\begin{aligned} \liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \left( \frac{\mathcal{J}(\bar{u} + th') - \mathcal{J}(\bar{u})}{t} + \frac{\mathcal{X}_{\mathcal{C}_{ad}}(\bar{u} + th') - \mathcal{X}_{\mathcal{C}_{ad}}(\bar{u})}{t} \right) &\geq \liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \frac{\mathcal{J}(\bar{u} + th') - \mathcal{J}(\bar{u})}{t} \\ &\quad + \liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \frac{\mathcal{X}_{\mathcal{C}_{ad}}(\bar{u} + th') - \mathcal{X}_{\mathcal{C}_{ad}}(\bar{u})}{t}. \end{aligned} \quad (5.9)$$

Furthermore, it holds that

$$\liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \frac{\mathcal{X}_{\mathcal{C}_{ad}}(\bar{u} + th') - \mathcal{X}_{\mathcal{C}_{ad}}(\bar{u})}{t} = \liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \frac{\mathcal{X}_{\mathcal{C}_{ad}}(\bar{u} + th')}{t} = \begin{cases} 0, & \text{if } h \in T_{\mathcal{C}_{ad}}(\bar{u}) \\ \infty, & \text{otherwise.} \end{cases} \quad (5.10)$$

Thus, given that  $\mathcal{J}'(\bar{u}; h) < \infty$  for every  $h \in L^p(\Omega)$ , we have that the expression (5.8) is equivalent to the corresponding one, where  $h \in T_{\mathcal{C}_{ad}}(\bar{u})$  only. For those  $h$  we have

$$\liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \left( \frac{\mathcal{J}(\bar{u} + th') - \mathcal{J}(\bar{u})}{t} + \frac{\mathcal{X}_{\mathcal{C}_{ad}}(\bar{u} + th') - \mathcal{X}_{\mathcal{C}_{ad}}(\bar{u})}{t} \right) = \liminf_{\substack{t \downarrow 0 \\ h' \rightarrow h}} \frac{\mathcal{J}(\bar{u} + th') - \mathcal{J}(\bar{u})}{t} = \mathcal{J}'(\bar{u}; h), \quad (5.11)$$

and thus (5.8) is equivalent to

$$\mathcal{J}'(\bar{u}; h) \geq 0 \quad \text{for all } h \in T_{\mathcal{C}_{ad}}(\bar{u}). \quad (5.12)$$

By using the chain rule for Hadamard directional differentiability [9, Proposition 2.47], we have

$$\begin{aligned} \mathcal{J}'(\bar{u}; h) &= \partial_y J(\bar{y}, \bar{u}) S'(\bar{u}; h) + \partial_u J(\bar{y}, \bar{u}) h \\ &= \langle \bar{y} - g, S'(\bar{u}; h) \rangle + \alpha \langle \bar{u}, h \rangle, \end{aligned}$$

and the proof is complete.  $\square$

We will say that a point  $\bar{u} \in \mathcal{C}_{ad}$  with corresponding state  $\bar{y}$  is *B-stationary* if it satisfies (5.4).

**Optimality conditions via regularization.** Next, we are interested in optimality conditions that are derived as a limit of the optimality conditions of smooth problems, see the discussion at the end of Section 3. Since this approach is not new [5, 10, 26], we will rather sketch the proof of the following proposition and afterwards discuss some more special cases in more detail.

**Proposition 5.2.** [Weak stationarity] *Suppose that  $\bar{u}$  is a local minimizer for the problem (4.1) with  $\bar{y} = S(\bar{u}) \in Y \cap C^{0,\alpha}(\bar{\Omega})$  the associated state. Then there exists a nonnegative  $\zeta \in L^\infty(\Omega)$  and an adjoint state  $p \in Y$  such that the following hold*

$$-\Delta p + \zeta p = \bar{y} - g, \quad (5.13)$$

$$\langle p + \alpha \bar{u}, u - \bar{u} \rangle \geq 0 \quad \text{for all } u \in \mathcal{C}_{ad}. \quad (5.14)$$

If  $\mathcal{C}_{ad}$  is given by (5.2), then (5.14) can be equivalently formulated as follows: There exists a multiplier  $\mu = \mu_b - \mu_a$  with  $\mu_a, \mu_b \in L^2(\Omega)$  such that  $\bar{u}$  and  $p$  satisfy

$$p + \alpha \bar{u} + \mu = 0, \quad (5.15)$$

$$u_a \leq \bar{u} \leq u_b, \quad \mu_a \geq 0, \quad \mu_b \geq 0, \quad (5.16)$$

$$\mu_a(u_a - \bar{u}) = \mu_b(\bar{u} - u_b) = 0, \quad (5.17)$$

almost everywhere in  $\Omega$ .

*Proof.* We will state the proof in several steps:

*Step 1:* Consider  $(E_{f_\epsilon})$ , the regularized state equation that corresponds to the smoothing of  $f$  discussed in Remark 3.5. Then by slightly generalizing the standard arguments of [10, Section 4], one can show that there exist  $\epsilon > 0$  and local minimizers of  $(\tilde{P}_{f_\epsilon})$  such that  $u_\epsilon \rightarrow \bar{u}$  in  $L^p(\Omega)$  as  $\epsilon \rightarrow 0$ , where  $(\tilde{P}_{f_\epsilon})$  is the following modified regularized problem

$$\begin{aligned} &\underset{(y,u) \in H_0^1(\Omega) \times L^2(\Omega)}{\text{minimize}} && \tilde{J}(y, u) := \frac{1}{2} \|y - g\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 + \frac{1}{p} \|u - \bar{u}\|_{L^p(\Omega)}^p, \\ &\text{subject to} && \begin{cases} -\Delta y + F_\epsilon(y) = u, & \text{in } \Omega, \\ y = 0, & \text{on } \partial\Omega, \end{cases} \quad \text{and } u \in \mathcal{C}_{ad}. \end{aligned} \quad (\tilde{P}_{f_\epsilon})$$

From Proposition 4.2 we have that  $y_\epsilon = S_\epsilon(u_\epsilon) \rightarrow S(\bar{u}) = y$  in  $Y \cap C^{0,a}(\bar{\Omega})$  and from the Lipschitz continuity of  $S_\epsilon : L^p(\Omega) \rightarrow Y \cap C^{0,a}(\bar{\Omega})$  on bounded sets of  $L^p(\Omega)$  and with Lipschitz constant independent of  $\epsilon$  we get that  $\|y_\epsilon\|_{L^\infty(\Omega)} \leq c$  uniformly in  $\epsilon$  for some constant  $c > 0$ . Combining this with the fact that  $F'_\epsilon(y)(x) = \partial_y f_\epsilon(x, y(x))$  is a bounded operator from  $L^\infty(\Omega)$  to itself due to the uniform Lipschitz constant of  $f_\epsilon$  we get that

$$\|F'_\epsilon(y_\epsilon)\|_{L^\infty(\Omega)} \leq c, \quad (5.18)$$

with a constant  $c > 0$  independent of  $\epsilon$ . Hence, along a subsequence,  $F'_\epsilon(y_\epsilon) \overset{*}{\rightharpoonup} \zeta$  weakly\* in  $L^\infty(\Omega)$  for some  $\zeta$  with  $\|\zeta\|_{L^\infty(\Omega)} \leq c$ .

*Step 2:* It follows from standard arguments [30], that for every local minimizer  $u_\epsilon (\neq \bar{u})$  of  $(\tilde{P}_{f_\epsilon})$  there exists a unique adjoint state  $p_\epsilon \in Y$  such that

$$-\Delta p_\epsilon + F'_\epsilon(y_\epsilon)p_\epsilon = y_\epsilon - g, \quad (5.19)$$

$$\langle p_\epsilon + \alpha u_\epsilon, u - u_\epsilon \rangle + \langle |u_\epsilon - \bar{u}|^{p-2}(u_\epsilon - \bar{u}), u - u_\epsilon \rangle \geq 0 \quad \text{for all } u \in \mathcal{C}_{ad}. \quad (5.20)$$

*Step 3:* We now take the limit as  $\epsilon \rightarrow 0$  in (5.19)–(5.20) using the limiting behaviours of  $u_\epsilon$ ,  $y_\epsilon$  and  $F'_\epsilon(y_\epsilon)$ . Note first that by testing (5.19) with  $p_\epsilon$ , employing the Poincaré inequality and using that  $y_\epsilon \rightarrow y$  in  $L^2(\Omega)$ , we can easily show that  $p_\epsilon$  is bounded in  $Y$  and thus  $p_\epsilon \rightharpoonup p$  in  $Y$  and  $p_\epsilon \rightarrow p$  in  $H_0^1(\Omega)$  for some  $p \in Y$  along a subsequence. Since  $F'_\epsilon(y_\epsilon) \overset{*}{\rightharpoonup} \zeta$  in  $L^\infty(\Omega)$  and  $p_\epsilon \rightarrow p$  in  $L^2(\Omega)$  we deduce  $F'_\epsilon(y_\epsilon)p_\epsilon \rightharpoonup \zeta p$  in  $L^2(\Omega)$ . Since  $p_\epsilon \rightharpoonup p$  in  $Y$  we have  $\Delta p_\epsilon \rightharpoonup \Delta p$  in  $L^2(\Omega)$  and since weak convergence in  $L^2(\Omega)$  implies strong convergence in  $H^{-1}(\Omega)$ , by taking the limit  $\epsilon \rightarrow 0$  in (5.19) we have that  $p$  satisfies (5.13). Finally, we take the limits in (5.20). We can straightforwardly estimate that  $\| |u_\epsilon - \bar{u}|^{p-2}(u_\epsilon - \bar{u}) \|_{L^q(\Omega)}^q \rightarrow 0$  as  $\epsilon \rightarrow 0$  where  $q = p/(p-1)$ . Using this as well as  $p_\epsilon \rightarrow p$  in  $L^2(\Omega)$  and  $u_\epsilon \rightarrow \bar{u}$  in  $L^p(\Omega)$ , we pass to the limit in (5.20) and obtain (5.14).

Finally we refer to [30, Theorem 2.29] for a proof of equivalence of (5.14) and the conditions (5.15)–(5.17) in the case of a box constraint set. □

**C-stationarity.** In general we say that a  $\bar{u} \in \mathcal{C}_{ad}$  with state  $\bar{y}$  is *weakly stationary* if there exist a non-negative  $\zeta \in L^\infty(\Omega)$  and an adjoint state  $p \in Y$  such that (5.13)–(5.14) are satisfied. Proposition 5.2 states that every local mimimizer of (4.1) is weakly stationary. Note, however, that this weak stationarity condition (5.13)–(5.14) does not provide any information about the dual variable  $\zeta$ . In this context, we are in particular interested in conditions that guarantee

$$\zeta(x) \in \partial f(x, \bar{y}(x)) \quad \text{for almost every } x \in \Omega, \quad (C)$$

where the subdifferential operates on the  $y$ -component of  $f$ . This gives rise to the following stationarity notion: A point  $\bar{u} \in \mathcal{C}_{ad}$  with state  $\bar{y}$  is called *C-stationary* if it is weakly stationary and (C) holds in addition. Letting  $f_x := f(x, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$ , we write  $\partial f(x, y) =: \partial f_x(y)$  where the latter represents Clarke's generalized gradient of  $f_x$  at  $y$ . We recall here that for a locally Lipschitz  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ , the Clarke generalized gradient at a point  $y$  can be characterized as

$$\partial \phi(y) = \text{conv}(\{\xi \in \mathbb{R} : \exists y_n \text{ with } \phi \text{ differentiable at } y_n, \text{ such that } y_n \rightarrow y \text{ and } \phi'(y_n) \rightarrow \xi\}); \quad (5.21)$$

see [11] for a general definition. In order to show *C-stationarity* for a local minimizer  $\bar{u}$  of (4.1), one needs to guarantee that the weak\* limit  $\zeta$  of  $F'_\epsilon(y_\epsilon)$  of the regularized nonlinearity at  $y_\epsilon$  is (in a point-wise sense) an element of the generalized gradient  $\partial f(x, \bar{y}(x))$ . In the recent preprint [8] a similar

problem is considered where this difficulty is remedied by assuming that the nonlinearity  $f$  is either convex or concave locally around  $y_0 \in \mathbb{R}$  whenever  $f$  is nondifferentiable at  $y_0$  and the set  $\{x \in \Omega : \bar{y}(x) = y_0\}$  is of positive measure. This, however, excludes a variety of other nonsmooth monotone functions which might, for instance, be convex on an interval left of  $y_0$  and concave on an interval on the right. In order to include more general cases, here we will make use of alternative condition, namely piecewise continuous differentiability.

**Assumption 5.3** ( $PC^1(\mathbb{R})$ ). *We suppose that apart from assumptions  $(A_1)$ – $(A_4)$ , for almost every  $x \in \Omega$  it holds that  $f(x, \cdot)$  is piecewise continuously differentiable, and we write  $f(x, \cdot) \in PC^1(\mathbb{R})$ . This means that there exist finitely many points  $-\infty < y_1 < \dots < y_k < +\infty$  such that  $f(x, \cdot)$  is continuously differentiable on each of the intervals  $(-\infty, y_1]$ ,  $[y_k, \infty)$  and  $[y_{j-1}, y_j]$  for all  $j \in \{2, \dots, k\}$ , where the derivatives at  $y_j$  are understood as one-sided derivatives.*

**Remark 5.4.** *It follows immediately that every ReLU neural network  $\mathcal{N} : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$  satisfying  $(A_1)$ – $(A_4)$ , that is, every ReLU neural network which is restricted to  $\Omega \times \mathbb{R}$  and is monotone increasing in the second variable also satisfies Assumption 5.3, since for every  $x \in \mathbb{R}^d$ ,  $\mathcal{N}(x, \cdot)$  is continuous and piecewise affine.*

Note that in view of (5.21), it is easy to characterize the Clarke generalized gradient of a  $PC^1$  function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ . Indeed, denoting by  $\phi'(\bar{y}, h)$  the directional derivative at  $\bar{y}$  in direction  $h$ , and  $\phi'_\pm(\bar{y}) := \pm\phi'(\bar{y}, \pm 1)$  the left- and right-sided derivatives, respectively, we have

$$\phi'_-(\bar{y}) = \lim_{y \nearrow \bar{y}} \phi'(y) \quad \text{and} \quad \phi'_+(\bar{y}) = \lim_{y \searrow \bar{y}} \phi'(y).$$

Then, defining  $\underline{\partial}\phi = \min\{\phi'_-, \phi'_+\}$  and  $\bar{\partial}\phi = \max\{\phi'_-, \phi'_+\}$ , it can be checked that  $\partial\phi = [\underline{\partial}\phi, \bar{\partial}\phi]$ .

Before we proceed towards  $C$ -stationarity, we need the following lemma.

**Lemma 5.5.** *Let  $\phi \in PC^1(\mathbb{R})$  and let  $\phi_n := \rho_n * \phi$  be the standard mollification of  $\phi$  with  $\rho_n := \rho_{\epsilon_n}$ ,  $\epsilon_n \searrow 0$ . Suppose that  $(y_n)_{n \in \mathbb{N}}$  is a sequence in  $\mathbb{R}$  such that  $y_n \rightarrow \bar{y}$  for some  $\bar{y} \in \mathbb{R}$ . Then it holds that*

$$\limsup_{n \rightarrow \infty} \phi'_n(y_n) \leq \bar{\partial}\phi(\bar{y}) \quad \text{and} \quad \liminf_{n \rightarrow \infty} \phi'_n(y_n) \geq \underline{\partial}\phi(\bar{y}).$$

*Proof.* See Appendix A. □

**Proposition 5.6.** *Let  $f$  satisfy Assumption 5.3. If  $\bar{u}$  is a local minimizer for the for the problem (4.1) with associated state  $\bar{y}$ , then  $\bar{u}$  is  $C$ -stationary.*

*Proof.* Since  $\bar{u}$  is a local minimizer, by Proposition 5.2 we have that it is weakly stationary, satisfying (5.13) for a  $\zeta$  obtained via  $F'_\epsilon(y_\epsilon) \xrightarrow{*} \zeta$  in  $L^\infty(\Omega)$ . Since for almost every  $x \in \Omega$ ,  $f(x, \cdot) \in PC^1(\mathbb{R})$ , and  $y_\epsilon(x) \rightarrow \bar{y}(x)$  as  $\epsilon \rightarrow 0$ , we can apply Lemma 5.5 and get that for almost every  $x \in \Omega$  we have

$$\limsup_{\epsilon \searrow 0} f'_\epsilon(x, y_\epsilon(x)) \leq \bar{\partial}f(x, \bar{y}(x)) \quad \text{and} \quad \liminf_{\epsilon \searrow 0} f'_\epsilon(x, y_\epsilon(x)) \geq \underline{\partial}f(x, \bar{y}(x)). \quad (5.22)$$

From the  $PC^1$ -property of  $f$  we have that for almost every  $x \in \Omega$

$$\partial f(x, \bar{y}(x)) = [\underline{\partial}f(x, \bar{y}(x)), \bar{\partial}f(x, \bar{y}(x))].$$

In order to prove (C) for  $\zeta$  we aim for a contradiction and assume that there exists a set  $A \subset \Omega$  of positive measure such that  $\zeta(x) < \underline{\partial}f(x, \bar{y}(x))$  for all  $x \in A$ . Using (5.22), Fatou's lemma and the fact that  $F'_\epsilon(y_\epsilon) \xrightarrow{*} \zeta$  in  $L^\infty(\Omega)$  we have

$$\int_{\Omega} \zeta \mathbb{1}_A dx < \int_{\Omega} \liminf_{\epsilon \searrow 0} f'_\epsilon(x, y_\epsilon(x)) \mathbb{1}_A dx \leq \liminf_{\epsilon \searrow 0} \int_{\Omega} f'_\epsilon(x, y_\epsilon(x)) \mathbb{1}_A dx = \int_{\Omega} \zeta \mathbb{1}_A dx,$$

which is a contradiction. Similarly, via contradiction and using a reversed version of Fatou's lemma for nonnegative functions bounded from above,  $\zeta(x) > \bar{\partial}f(x, \bar{y}(x))$  cannot hold on a set of positive measure.  $\square$

**Strong stationarity.** Next we examine conditions that, additionally to  $C$ -stationarity, will guarantee a certain sign condition on the adjoint state  $p$ . We say that a point  $\bar{u} \in \mathcal{C}_{ad}$  with state  $\bar{y}$  satisfies the *sign condition*  $\sigma_{(A)}$  on a measurable set  $A \subset \Omega$  if it is  $C$ -stationary and

$$\zeta(x)p(x) \in [f'_+(\bar{y}(x))p(x), f'_-(\bar{y}(x))p(x)] \quad \text{for almost every } x \in A. \quad (\sigma_{(A)})$$

Then,  $\bar{u}$  is called *strongly stationary* if it satisfies the sign condition  $\sigma_{(\Omega)}$  on the whole of  $\Omega$ . Note that strong stationarity always implies  $C$ -stationarity even if  $f$  does not satisfy the  $PC^1$  condition. This is due to the fact that  $[\underline{\partial}f(x, \bar{y}(x)), \bar{\partial}f(x, \bar{y}(x))] \subset \partial f(x, \bar{y}(x))$  which is a consequence of the directional differentiability of  $f$ . Note also that on the set  $\{x \in \Omega : p(x) = 0\}$  one can change  $\zeta$  such that (C) is satisfied without affecting (5.13).

Here we examine conditions under which local minimizers of (4.1) are strongly stationary. We will also formulate a constraint qualification under which  $B$ -stationarity is equivalent to strong stationarity. The next proposition states that the latter always implies the former. We note that these results are inspired by the investigations in the preprint [8].

**Proposition 5.7.** *Suppose that  $\bar{u}$  is strongly stationary. Then  $\bar{u}$  is  $B$ -stationary.*

*Proof.* For  $z \in L^2(\Omega)$ , we have  $z = z^+ - z^-$  with  $z^+ := \max(0, z)$ ,  $z^- = \max(0, -z)$  and we write  $\Omega_{z^+} := \{x \in \Omega : z(x) \geq 0\}$ ,  $\Omega_{z^-} := \{x \in \Omega : z(x) \leq 0\}$ . From the positive homogeneity of the directional derivative with respect to the direction and from the sign condition for  $\zeta$  and  $p$  we have

$$\begin{aligned} \int_{\Omega} F'(\bar{y}; z)p dx &= \int_{\Omega_{z^+}} F'(\bar{y}; z^+)p dx + \int_{\Omega_{z^-}} F'(\bar{y}; -z^-)p dx \\ &= \int_{\Omega_{z^+}} f'_+(\bar{y})pz^+ dx + \int_{\Omega_{z^-}} f'_-(\bar{y})p(-z^-) dx \\ &\leq \int_{\Omega_{z^+}} \zeta pz^+ dx + \int_{\Omega_{z^-}} \zeta p(-z^-) dx = \int_{\Omega} \zeta pz dx \end{aligned}$$

and hence

$$\langle \zeta z, p \rangle - \langle F'(\bar{y}; z), p \rangle \geq 0. \quad (5.23)$$

In view of the adjoint equation (5.13) we have that  $p$  satisfies

$$\langle -\Delta z + \zeta z, p \rangle = \langle \nabla z, \nabla p \rangle + \langle \zeta z, p \rangle = \langle -\Delta p, z \rangle + \langle \zeta p, z \rangle = \langle \bar{y} - g, z \rangle, \quad \text{for all } z \in Y. \quad (5.24)$$

For  $h \in L^p(\Omega)$ , let  $z := S'(\bar{u}, h)$ . Then, testing the characterizing equation of  $z$  from Proposition 3.2 with  $p$  we get

$$\langle -\Delta z, p \rangle = -\langle F'(\bar{y}; z), p \rangle + \langle h, p \rangle. \quad (5.25)$$

Furthermore, we also have that

$$\langle p, h \rangle \geq -\alpha \langle \bar{u}, h \rangle, \quad \text{for all } h \in T_{\mathcal{C}_{ad}}(\bar{u}). \quad (5.26)$$

Indeed, using (5.14) we have that this is true for every  $h \in \mathcal{R}_{\mathcal{C}_{ad}} = \{t(u - \bar{u}) : u \in \mathcal{C}_{ad}, t > 0\}$ , and then (5.26) follows by using the fact that  $T_{\mathcal{C}_{ad}}(\bar{u}) = \overline{\mathcal{R}_{\mathcal{C}_{ad}}(\bar{u})}$  and the continuity of  $\langle p + \alpha \bar{u}, \cdot \rangle$ . We fix  $h \in T_{\mathcal{C}_{ad}}(\bar{u})$  and let  $z = S'(\bar{u}; h)$ . From (5.24), (5.25) as well as (5.23) and (5.26) we have

$$\langle \bar{y} - g, S'(\bar{u}; h) \rangle = \langle \bar{y} - g, z \rangle = \langle \zeta z, p \rangle + \langle -\Delta z, p \rangle = \langle \zeta z, p \rangle - \langle F'(\bar{y}; z), p \rangle + \langle h, p \rangle \geq -\alpha \langle \bar{u}, h \rangle. \quad (5.27)$$

Since  $h \in T_{\mathcal{C}_{ad}}(\bar{u})$  is arbitrary, the proof is complete.  $\square$

We now introduce a measurability assumption which will be relevant to the constraint qualification which we state afterwards.

**Assumption 5.8.** *The set  $\Omega_f := \{x \in \Omega : f(x, \cdot) \text{ is nondifferentiable at } \bar{y}(x)\}$  is Lebesgue measurable.*

Assumption 5.8 holds for instance when  $f$  satisfies  $(A_1)$ – $(A_4)$  and additionally is independent of  $x$ , or if it satisfies Assumption 5.3 and is jointly continuous on  $\Omega \times \mathbb{R}$ .

We now introduce a constraint qualification for  $\bar{u} \in \mathcal{C}_{ad}$  that lead to strong stationarity as a necessary condition in different settings:

$$T_{\mathcal{C}_{ad}}(\bar{u}) \text{ is dense in } L^2(\Omega). \quad (CQ)$$

Furthermore, if  $\mathcal{C}_{ad}$  is defined via box constraints as in (5.2), and Assumption 5.8 holds, then we define

$$\Omega_f \cap \bar{\Omega}_{a,b} \text{ has Lebesgue measure zero,} \quad (CQ_f)$$

where

$$\Omega_{a,b} := \{x \in \Omega : \bar{u}(x) = u_a(x) \text{ or } \bar{u}(x) = u_b(x)\}.$$

We note that if  $\mathcal{C}_{ad}$  is given as a box constraint set as in (5.2) then  $(CQ)$  is satisfied if and only if  $\Omega_{a,b}$  has Lebesgue measure zero. This can be seen directly from the characterization (5.3) of  $T_{\mathcal{C}_{ad}}(\bar{u})$  for a box constraint set. Thus if  $\Omega_{a,b}$  is closed, then  $(CQ)$  implies  $(CQ_f)$ . Note also that  $(CQ_f)$  is satisfied independently of the measure of  $\Omega_{a,b}$  in the case that  $f$  is differentiable in the second variable.

**Proposition 5.9.** *Let  $\bar{u}$  with corresponding state  $\bar{y}$  be a point which is both  $B$ -stationary and  $C$ -stationary. Moreover assume that Assumption 5.8 holds and that  $\mathcal{C}_{ad}$  is defined as in (5.2). Then  $\bar{u}$  satisfies the sign condition  $\sigma_{(\Omega \setminus (\Omega_f \cap \bar{\Omega}_{a,b}))}$ . In particular, if  $(CQ_f)$  is satisfied, then  $\bar{u}$  is strongly stationary.*

*Proof.* We fix an arbitrary direction  $h \in T_{\mathcal{C}_{ad}}(\bar{u})$  and let  $z := S'(\bar{u}; h) \in Y$ . Testing the equation that characterizes  $z$  with  $p$  and using (5.15) from weak stationarity, as well as (5.4) from  $B$ -stationarity and (5.24) we get

$$\langle -\Delta z + F'(\bar{y}; z), p \rangle = \langle h, p \rangle = -\alpha \langle \bar{u}, h \rangle - \langle \mu, h \rangle \leq \langle \bar{y} - g, z \rangle - \langle \mu, h \rangle = \langle -\Delta z + \zeta z, p \rangle - \langle \mu, h \rangle,$$

and thus

$$\begin{aligned} 0 &\leq \langle \zeta z - F'(\bar{y}; z), p \rangle - \langle \mu, h \rangle = \int_{\Omega} (\zeta z - F'(\bar{y}; z)) p \, dx - \int_{\Omega} \mu h \, dx \\ &= \int_{\Omega_f} (\zeta z - F'(\bar{y}; z)) p \, dx - \int_{\Omega_{a,b}} \mu h \, dx, \end{aligned} \quad (5.28)$$



where in the last inequality we used the fact that from  $C$ -stationarity we have  $\zeta(x)z(x) - F'(\bar{y}(x); z(x)) = 0$  for almost every  $x \in \Omega \setminus \Omega_f$  and  $\mu = 0$  on  $\Omega \setminus \Omega_{a,b}$  which can be easily verified.

According to Lemma A.1 we can find a non-negative  $\psi_m \in C^\infty(\mathbb{R}^d)$  such that for an arbitrary non-negative function  $v \in C_c^\infty(\Omega)$  the function

$$h_m := -\Delta(v\psi_m)(x) + F'(\bar{y}; v\psi_m)(x),$$

belongs to  $L^\infty(\Omega)$  and vanishes almost everywhere in  $\Omega_{a,b}$ , in particular  $h_m \in T_{\mathcal{C}_{ad}}(\bar{u})$  and it also holds  $v\psi_m = S'(\bar{u}; h_m)$ . Thus we can take  $z = v\psi_m$  in (5.28) and by using the positive homogeneity of the directional derivative, we get

$$0 \leq \int_{\Omega_f} (\zeta v\psi_m - F'(\bar{y}; v\psi_m))p \, dx = \int_{\Omega} \mathbb{1}_{\Omega_f} (\zeta - f'_+(\bar{y}))\psi_m p v \, dx.$$

Since  $v \in C_c^\infty(\Omega)$  is an arbitrary nonnegative function and from the fundamental lemma of calculus of variations, and the fact that  $\psi_m$  is also nonnegative, this implies that  $(\zeta(x) - f'_+(\bar{y}(x)))p(x) \geq 0$  for almost every  $x \in \Omega_f \cap \text{supp}\psi_m$ . Using the pointwise convergence  $\mathbb{1}_{\text{supp}\psi_m} \rightarrow \mathbb{1}_{\Omega \setminus \bar{\Omega}_{a,b}}$  from Lemma A.1 we conclude that

$$\zeta(x)p(x) \geq f'_+(\bar{y}(x))p(x) \quad \text{for almost every } x \in \Omega_f \setminus \bar{\Omega}_{a,b}.$$

By taking  $z = -v\psi_m$  in (5.28) we get similarly

$$0 \leq \int_{\Omega_f} (\zeta(-v\psi_m) - F'(\bar{y}; -v\psi_m))p \, dx = \int_{\Omega} \mathbb{1}_{\Omega_f} (f'_-(\bar{y}) - \zeta)\psi_m p v \, dx$$

and using the same argument we end up with

$$\zeta(x)p(x) \leq f'_-(\bar{y}(x))p(x) \quad \text{for almost every } x \in \Omega_f \setminus \bar{\Omega}_{a,b}.$$

Since for every  $x \in \Omega \setminus \Omega_f$  we have  $\zeta(x)p(x) = f'(\bar{y}(x))p(x)$  due to the  $C$ -stationarity, we deduce that  $\bar{u}$  satisfies the sign condition  $\sigma_{(\Omega \setminus (\Omega_f \cap \bar{\Omega}_{a,b}))}$ .  $\square$

Note that under Assumptions 5.3 and 5.8, the conditions of the previous proposition are satisfied for any local minimizer  $\bar{u}$  of (4.1). In the next proposition we show that the converse of Proposition 5.7 is also true provided that the constraint qualification (CQ) holds. The proof here closely follows [10, Proposition 4.13].

**Proposition 5.10.** *Suppose that one of the following is satisfied for  $\bar{u}$ :*

- (i) *The constraint qualification (CQ) holds.*
- (ii) *The set  $\Omega_f$  has zero Lebesgue measure and  $\mathcal{C}_{ad}$  is given as a box constraint set.*

*Then we have that if  $\bar{u}$  is  $B$ -stationary, it is also strongly stationary.*

*Proof.* We first assume that  $\bar{u}$  satisfies (CQ) and it is also  $B$ -stationary. We set  $p := -\alpha\bar{u} \in L^2(\Omega)$  as a candidate for the adjoint state and it clearly follows that (5.14) is satisfied. Let  $z \in Y$  such that there exists  $h \in T_{\mathcal{C}_{ad}}(\bar{u})$  with  $z = S'(\bar{u}; h)$ . Testing the equation that  $z$  satisfies with  $p$  and using the  $B$ -stationarity condition yields

$$\langle -\Delta z, p \rangle + \langle F'(\bar{y}; z), p \rangle = \langle p, h \rangle = -\alpha \langle \bar{u}, h \rangle \leq \langle \bar{y} - g, z \rangle$$

$$\Rightarrow \langle -\Delta z, p \rangle - \langle \bar{y} - g, z \rangle \leq -\langle F'(\bar{y}; z), p \rangle.$$

We can now interpret  $\xi := \Delta p + (\bar{y} - g) \in Y^*$  and estimate

$$\begin{aligned} \langle \xi, -z \rangle_Y &= -\langle \Delta z, p \rangle - \langle \bar{y} - g, z \rangle \leq -\langle F'(\bar{y}; z), p \rangle \leq \|F'(\bar{y}; z)\|_{L^2(\Omega)} \|p\|_{L^2(\Omega)} \\ &\leq c \|z\|_{L^2(\Omega)} \|p\|_{L^2(\Omega)} \end{aligned} \quad (5.29)$$

where we used that positive homogeneity of the directional derivative and the fact that  $F'(\bar{y}; \pm 1) \in L^\infty(\Omega)$ . Note that (5.29) is valid for all  $z \in \{S'(\bar{u}; h) \in Y : h \in T_{C_{ad}}(\bar{u})\}$ . We claim that the latter set is dense in  $Y$ . Indeed this follows from the fact that the operator  $h \mapsto S'(\bar{u}; h)$  derived from equation ( $K$ ) is surjective and Lipschitz continuous  $L^p(\Omega) \rightarrow Y$  from the density of  $T_{C_{ad}}(\bar{u})$  in  $L^2(\Omega)$ . Using this density we deduce from (5.29)

$$\langle \xi, z \rangle_Y \leq \tilde{c} \|z\|_{L^2(\Omega)}, \quad \text{for all } z \in Y.$$

From the Hahn-Banach theorem we can extend  $\xi$  from  $Y^*$  to  $L^2(\Omega)^* \simeq L^2(\Omega)$ . Denoting its Riesz representation by  $\xi$  again we have

$$\langle \xi, z \rangle_{L^2(\Omega)} = \langle \xi, z \rangle \leq -\langle F'(\bar{y}; -z), p \rangle, \quad \text{for all } z \in Y.$$

By density of  $Y$  in  $L^2(\Omega)$  and continuity of  $F'(\bar{y}; \cdot)$  on  $L^2(\Omega)$ , we can extend the previous estimate to

$$\langle \xi, z \rangle \leq -\langle F'(\bar{y}; -z), p \rangle, \quad \text{for all } z \in L^2(\Omega).$$

For arbitrary  $z \in C_c^\infty(\Omega)$  with  $z \geq 0$  we have

$$\begin{aligned} 0 &\leq \int_{\Omega} -F'(\bar{y}; -z)p - \xi z \, dx = \int_{\Omega} (f'_-(\bar{y})p - \xi)z \, dx, \\ 0 &\leq \int_{\Omega} -\xi(-z) - F'(\bar{y}; z)p = \int_{\Omega} (\xi - f'_+(\bar{y})p)z \, dx, \end{aligned}$$

and from the fundamental lemma of calculus of variations we get

$$f'_+(\bar{y}(x))p(x) \leq \xi(x) \leq f'_-(\bar{y}(x))p(x), \quad \text{for almost every } x \in \Omega. \quad (5.30)$$

By defining

$$\zeta(x) := \begin{cases} \frac{\xi(x)}{p(x)}, & \text{almost everywhere in } \{p(x) \neq 0\}, \\ f'_-(\bar{y}(x)), & \text{almost everywhere in } \{p(x) = 0\}, \end{cases} \quad (5.31)$$

we have that  $\zeta(x) \in [\partial f(\bar{y}(x)), \bar{\partial} f(\bar{y}(x))]$  and thus  $\zeta(x) \in \partial f(\bar{y}(x))$  almost everywhere in  $\Omega$ . Moreover  $\zeta \in L^\infty(\Omega)$ ,  $\zeta \geq 0$  and the sign condition for  $\zeta p$  holds almost everywhere in  $\Omega$ . Since  $\xi = \zeta p \in L^2(\Omega)$ , in view of  $\xi = \Delta p + (\bar{y} - g)$  we have

$$-\Delta p + \zeta p = (\bar{y} - g),$$

implying  $-\Delta p \in L^2(\Omega)$  and thus  $p \in Y$ . This concludes the strong stationarity of  $\bar{u}$ .

Now if we assume that  $\bar{u}$  is  $B$ -stationary and  $\Omega_f$  has zero Lebesgue measure, then it can be shown, see [24, Lemma 1.12], that there exists  $\mu = \mu_b - \mu_a$ , with  $\mu_a, \mu_b \in L^2(\Omega)$  such that

$$\begin{aligned} \mathcal{J}'(\bar{u}; h) + \langle \mu, h \rangle &= \langle \bar{y} - g, S'(\bar{u}; h) \rangle + \alpha \langle \bar{u}, h \rangle + \langle \mu, h \rangle = 0, \quad \text{for all } h \in L^p(\Omega), \\ u_a &\leq \bar{u} \leq u_b, \quad \mu_a \geq 0, \quad \mu_b \geq 0 \\ \mu_a(u_a - \bar{u}) &= \mu_b(\bar{u} - u_b) = 0 \end{aligned} \quad (5.32)$$

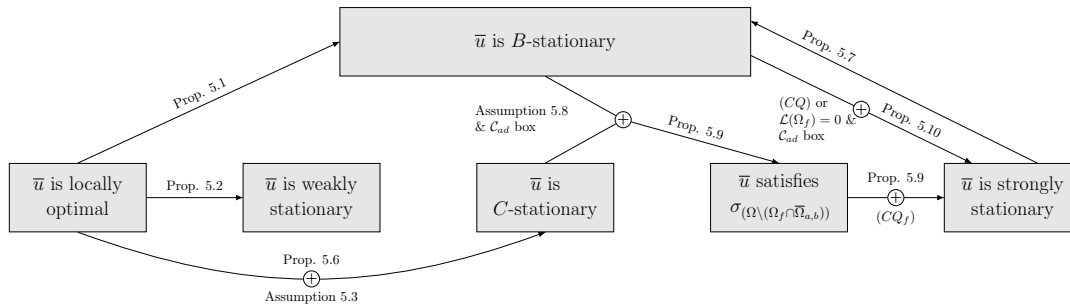


Figure 1: Summary of relations between the different stationarity conditions. Arrows represent direct implications which only require the assumptions they start from. Additional assumptions are marked with the “+” symbol.

almost everywhere in  $\Omega$ . In that case we can set  $p := -\alpha\bar{u} + \mu \in L^2(\Omega)$  and we proceed as before by using the equality  $\{S'(\bar{u}; h) \in Y : h \in L^p(\Omega)\} = Y$ .

□

The different stationarity conditions and their relations are summarized in Figure 1.

## 6 Conclusion

In this paper, we studied the optimal control of a family of nonsmooth neural network informed PDEs which approximate some semilinear PDE with unknown nonlinearity. We focused on theoretical investigations, such as well-posedness of the learning-informed PDEs and the optimal control problem, approximation properties, and established several first-order conditions based on purely primal concepts (B-stationarity) and primal-dual systems (e.g. C-stationarity and strong stationarity). Particular attention was paid to ReLU networks due to its wide application in practical learning. However, many of the results presented here fit to optimal control of more general nonsmooth semilinear PDEs. Different optimality/stationarity conditions require specific assumptions, and their relationships are clarified in Figure 1. We showed the equivalence between strong stationarity and B-stationarity, under some rather restrictive constraint qualifications. Whether this equivalence persists under milder constraint qualifications remains an open question.

The understanding of optimality conditions for the nonsmooth PDE constrained optimization problem is not only an important theoretical question, but also guides the development of robust nonsmooth numerical algorithms, which is the focus of the companion paper [15].

## A Appendix

*Proof of Lemma 5.5.* If  $\phi$  is differentiable  $\bar{y}$ , then  $\partial\phi(\bar{y}) = \bar{\partial}\phi(\bar{y}) = \phi'(\bar{y})$ . Since  $\phi \in PC^1(\mathbb{R})$ ,  $\phi'$  is continuous at a compact interval containing  $\bar{y}$ . Hence  $\phi'_n = \rho_n * \phi' \rightarrow \phi'$  uniformly in that interval and thus we have  $\lim_{n \rightarrow \infty} \phi'_n(y_n) = \phi'(\bar{y})$ .

Suppose now that  $\bar{y}$  is a nondifferentiable point of  $\phi$ . Since  $\phi \in PC^1(\mathbb{R})$ , there exists an  $M > 0$  such that  $\phi' \in C([\bar{y} - M, \bar{y}])$  and  $\phi' \in C([\bar{y}, \bar{y} + M])$ . From uniform continuity of  $\phi'$  on those

intervals we have that for every  $\epsilon > 0$ , there exists a  $\delta \leq M$  such that

$$0 < \bar{y} - y < \delta \Rightarrow |\phi'(y) - \phi'_-(\bar{y})| < \epsilon \quad \text{and} \quad 0 < y - \bar{y} < \delta \Rightarrow |\phi'(y) - \phi'_+(\bar{y})| < \epsilon. \quad (\text{A.1})$$

Since  $y_n \rightarrow \bar{y}$  and  $\epsilon_n \searrow 0$ , it follows that for all  $\epsilon > 0$  we can pick  $N \in \mathbb{N}$  such that for every  $n > N$  we have that for all  $z \in (-\epsilon_n, \epsilon_n)$

$$y_n - z < \bar{y} \Rightarrow |\phi'(y_n - z) - \phi'_-(\bar{y})| < \epsilon \quad \text{and} \quad y_n - z > \bar{y} \Rightarrow |\phi'(y_n - z) - \phi'_+(\bar{y})| < \epsilon. \quad (\text{A.2})$$

Note also that for every  $n$ , the set  $\{z \in (-\epsilon_n, \epsilon_n) : y_n - z = \bar{y}\}$  contains at most one point and hence it is a Lebesgue nullset. In order to prove that  $\limsup_{n \rightarrow \infty} \phi'_n(y_n) \leq \bar{\partial}\phi(\bar{y})$ , we show that for every  $\epsilon > 0$  we can find  $N$  large enough such that for all  $n > N$ , we have  $\phi'_n(y_n) < \bar{\partial}\phi(\bar{y}) + \epsilon$ . Choosing  $N$  as in (A.2) and using  $-\bar{\partial}\phi(\bar{y}) \leq -\phi'_\pm(\bar{y})$  we have

$$\begin{aligned} \phi'_n(y_n) - \bar{\partial}\phi(\bar{y}) &= \\ &= \int_{[-\epsilon_n, \epsilon_n]} \rho_n(z) (\phi'(y_n - z) - \bar{\partial}\phi(\bar{y})) \mathbb{1}_{\{y_n - z < \bar{y}\}}(z) \\ &\quad + \rho_n(z) (\phi'(y_n - z) - \bar{\partial}\phi(\bar{y})) \mathbb{1}_{\{y_n - z > \bar{y}\}}(z) dz \\ &\leq \int_{[-\epsilon_n, \epsilon_n]} \rho_n(z) (\phi'(y_n - z) - \phi'_-(\bar{y})) \mathbb{1}_{\{y_n - z < \bar{y}\}}(z) \\ &\quad + \rho_n(z) (\phi'(y_n - z) - \phi'_+(\bar{y})) \mathbb{1}_{\{y_n - z > \bar{y}\}}(z) dz \\ &< \epsilon \left( \int_{[-\epsilon_n, \epsilon_n]} \rho_n(z) \mathbb{1}_{\{y_n - z < \bar{y}\}}(z) dz + \int_{[-\epsilon_n, \epsilon_n]} \rho_n(z) \mathbb{1}_{\{y_n - z > \bar{y}\}}(z) dz \right) = \epsilon. \end{aligned}$$

The inequality  $\liminf_{n \rightarrow \infty} \phi'_n(y_n) \geq \underline{\partial}\phi(\bar{y})$  is proven analogously. □

The following lemma and its proof are slight adaptations from [8, Lemma 3.7] which we include here for completeness.

**Lemma A.1.** *For an arbitrary  $\mathcal{A} \subset \Omega$  there exists a sequence  $(\psi_m)_{m \in \mathbb{N}} \subset C^\infty(\mathbb{R}^d)$  with  $\psi_m \geq 0$  such that for every  $m \in \mathbb{N}$  and every  $v \in C^\infty(\mathbb{R}^d)$*

$$-\Delta(v\psi_m)(x) + F'(\bar{y}; v\psi_m)(x) = 0, \quad \text{for almost every } x \in \mathcal{A}, \quad (\text{A.3})$$

and

$$\mathbb{1}_{\text{supp}\psi_m}(x) \rightarrow \mathbb{1}_{\Omega \setminus \bar{\mathcal{A}}}(x), \quad \text{for every } x \in \Omega. \quad (\text{A.4})$$

*Proof.* For  $m \in \mathbb{N}$ , we set  $\mathcal{A}_m := \{x \in \mathbb{R}^d : \text{dist}(\mathcal{A}, x) < \frac{1}{m}\}$ , which are open sets, satisfying  $\mathcal{A} \subset \mathcal{A}_{m+1} \subset \mathcal{A}_m$  for all  $m$  and also  $\bigcap_{m \in \mathbb{N}} \mathcal{A}_m = \bar{\mathcal{A}}$ . According to a classical result by Whitney, see for instance [22, Theorem 2.1], for every  $m \in \mathbb{N}$  we can find  $\psi_m \in C^\infty(\mathbb{R}^d)$  with  $\psi_m > 0$  in  $\Omega \setminus \bar{\mathcal{A}}_m$  and  $\psi_m = 0$  in  $\mathbb{R}^d \setminus (\Omega \setminus \bar{\mathcal{A}}_m)$ . In particular,  $\psi_m = 0$  on the open set  $\Omega \cap \mathcal{A}_m$  and (A.3) is satisfied for almost every  $x \in \Omega \cap \mathcal{A}_m$  and hence for almost every  $x \in \mathcal{A}$ . In order to show (A.4), we note that  $\text{supp}\psi_m = (\Omega \setminus \bar{\mathcal{A}}_m) \cap \Omega = \bar{\Omega} \setminus \mathcal{A}_m$  and  $\text{supp}\psi_m \subset \text{supp}\psi_{m+1}$  for all  $m \in \mathbb{N}$ . Thus we have

$$\bigcup_{m \in \mathbb{N}} \text{supp}\psi_m = \bigcup_{m \in \mathbb{N}} (\bar{\Omega} \setminus \mathcal{A}_m) = \bar{\Omega} \setminus \left( \bigcap_{m \in \mathbb{N}} \mathcal{A}_m \right) = \bar{\Omega} \setminus \bar{\mathcal{A}},$$

which implies  $\mathbb{1}_{\text{supp}\psi_m} \rightarrow \mathbb{1}_{\bar{\Omega} \setminus \bar{\mathcal{A}}}$  pointwise in  $\mathbb{R}^d$ , and hence (A.4) follows. □

## References

- [1] R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Academic Press, 2nd edition, 2003.
- [2] C.D. Aliprantis, D. Harris, and R. Tourky. Continuous piecewise linear functions. *Macroeconomic Dynamics*, 10(1):77–99, 2006. <https://doi.org/10.1017/S1365100506050103>.
- [3] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018. [https://openreview.net/pdf?id=B1J\\_rgWRW](https://openreview.net/pdf?id=B1J_rgWRW).
- [4] J.-P. Aubin and H. Frankowska. *Set-valued analysis*. Springer Science & Business Media, 2009.
- [5] V. Barbu. Optimal control of variational inequalities. *Research Notes in Mathematics*, 100, 1984.
- [6] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- [7] J. Berner, D. Elbrächter, P. Grohs, and A. Jentzen. Towards a regularity theory for ReLU networks – chain rule and global error estimates. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5, 2019. <https://doi.org/10.1109/SampTA45681.2019.9031005>.
- [8] L. Betz. Strong stationarity for a highly nonsmooth optimization problem with control constraints. *preprint*, 2022. [https://www.researchgate.net/publication/360226766\\_STRONG\\_STATIONARITY\\_FOR\\_A\\_HIGHLY\\_NONSMOOTH\\_OPTIMIZATION\\_PROBLEM\\_WITH\\_CONTROL\\_CONSTRAINTS](https://www.researchgate.net/publication/360226766_STRONG_STATIONARITY_FOR_A_HIGHLY_NONSMOOTH_OPTIMIZATION_PROBLEM_WITH_CONTROL_CONSTRAINTS).
- [9] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.
- [10] C. Christof, C. Meyer, S. Walther, and C. Clason. Optimal control of a non-smooth semilinear elliptic equation. *Mathematical Control & Related Fields*, 8:247, 2018. <https://doi.org/10.3934/mcrf.2018011>.
- [11] F.H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Applied Mathematics, Philadelphia, 1987.
- [12] W.M. Czarnecki, S. Osindero, M. Jaderberg, G. Swirszcz, and R. Pascanu. Sobolev training for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4281–4290, 2017. <https://proceedings.neurips.cc/paper/2017/file/758a06618c69880a6cee5314ee42d52f-Paper.pdf>.
- [13] H. Daniels and M. Velikova. Monotone and partially monotone neural networks. *IEEE Transactions on Neural Networks*, 21(6):906–917, 2010. <https://doi.org/10.1109/TNN.2010.2044803>.
- [14] G. Dong, M. Hintermüller, and K. Papafitsoros. Quantitative magnetic resonance imaging: From fingerprinting to integrated physics-based models. *SIAM Journal on Imaging Sciences*, 12(2), 2019. <https://doi.org/10.1137/18M1222211>.

- [15] 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. *preprint*, 2022.
- [16] G. Dong, M. Hintermüller, and K. Papafitsoros. Optimization with learning-informed differential equation constraints and its applications. *ESAIM: COCV*, 28:1–44, 2022. <https://doi.org/10.1051/cocv/2021100>.
- [17] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 2015.
- [18] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.
- [19] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [20] I. Gühring, G. Kutyniok, and P. Petersen. Error bounds for approximations with deep ReLU neural networks in  $W^{s,p}$  norms. *Analysis and Applications*, 18(05):803–859, 2020. <https://doi.org/10.1142/S0219530519410021>.
- [21] M. Hintermüller, B.S. Mordukhovich, and T.M. Surowiec. Several approaches for the derivation of stationarity conditions for elliptic MPECs with upper-level control constraints. *Mathematical Programming*, 146(1):555–582, 2014. <https://doi.org/10.1007/s10107-013-0704-6>.
- [22] M. Hintermüller, K. Papafitsoros, and C.N. Rautenberg. Variable step mollifiers and applications. *Integral Equations and Operator Theory*, 92(53), 2020. <https://doi.org/10.1007/s00020-020-02608-2>.
- [23] M. Hintermüller and T.M. Surowiec. A bundle-free implicit programming approach for a class of elliptic mpecs in function space. *Mathematical Programming*, 160:271–305, 2016. <https://doi.org/10.1007/s10107-016-0983-9>.
- [24] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*, volume 23. Springer Science & Business Media, 2009. <https://doi.org/10.1007/978-1-4020-8839-1>.
- [25] X. Liu, X. Han, N. Zhang, and Q. Liu. Certified monotonic neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15427–15438, 2020. <https://proceedings.neurips.cc/paper/2020/file/b139aeda1c2914e3b579aafd3ceeb1bd-Paper.pdf>.
- [26] F. Mignot and J.P. Puel. Optimal control in some variational inequalities. *SIAM Journal on Control and Optimization*, 22(3):466–476, 1984. <https://doi.org/10.1137/0322028>.
- [27] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018. <https://doi.org/10.1016/j.neunet.2018.08.019>.
- [28] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999. <https://doi.org/10.1017/S0962492900002919>.

- [29] A. Sivaraman, G. Farnadi, T. Millstein, and G. Van den Broeck. Counterexample-guided learning of monotonic neural networks. *arXiv preprint arXiv:2006.08852*, 2020. <https://arxiv.org/abs/2006.08852>.
- [30] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.
- [31] S. Wang and X. Sun. Generalization of hinging hyperplanes. *IEEE Transactions on Information Theory*, 51(12):4425–4431, 2005. <https://doi.org/10.1109/TIT.2005.859246>.
- [32] D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649, 2018. <http://proceedings.mlr.press/v75/yarotsky18a.html>.
- [33] K. Yosida. *Functional Analysis*, volume 123 of *A Series of Comprehensive Studies in Mathematics*, 1978.