



# On the Omnipresence of Spurious Local Minima in Certain Neural Network Training Problems

Constantin Christof<sup>1</sup> · Julia Kowalczyk<sup>1</sup>

Received: 23 February 2022 / Revised: 7 March 2023 / Accepted: 25 May 2023 /

Published online: 14 June 2023

© The Author(s) 2023

## Abstract

We study the loss landscape of training problems for deep artificial neural networks with a one-dimensional real output whose activation functions contain an affine segment and whose hidden layers have width at least two. It is shown that such problems possess a continuum of spurious (i.e., not globally optimal) local minima for all target functions that are not affine. In contrast to previous works, our analysis covers all sampling and parameterization regimes, general differentiable loss functions, arbitrary continuous nonpolynomial activation functions, and both the finite- and infinite-dimensional setting. It is further shown that the appearance of the spurious local minima in the considered training problems is a direct consequence of the universal approximation theorem and that the underlying mechanisms also cause, e.g.,  $L^p$ -best approximation problems to be ill-posed in the sense of Hadamard for all networks that do not have a dense image. The latter result also holds without the assumption of local affine linearity and without any conditions on the hidden layers.

**Keywords** Deep artificial neural network · Spurious local minimum · Training problem · Loss landscape · Hadamard well-posedness · Best approximation · Stability analysis · Local affine linearity

**Mathematics Subject Classification** 68T07 · 49K40 · 52A30 · 90C31

---

Communicated by Gitta Kutyniok.

---

✉ Constantin Christof  
christof@ma.tum.de

Julia Kowalczyk  
julia.kowalczyk@ma.tum.de

<sup>1</sup> Chair of Optimal Control, Center for Mathematical Sciences, Technical University of Munich, Boltzmannstraße 3, 85748 Garching, Germany

## 1 Introduction

Due to its importance for the understanding of the behavior, performance, and limitations of machine learning algorithms, the study of the loss landscape of training problems for artificial neural networks has received considerable attention in the last years. Compare, for instance, with the early works [4, 7, 56] on this topic, with the contributions on stationary points and plateau phenomena in [1, 10, 15, 18, 52], with the results on suboptimal local minima and valleys in [3, 11, 20, 25, 37, 41, 49, 54], and with the overview articles [6, 46, 47]. For fully connected feedforward neural networks involving activation functions with an affine segment, much of the research on landscape properties was initially motivated by the observation of Kawaguchi [30] that networks with linear activation functions give rise to learning problems that do not possess spurious (i.e., not globally optimal) local minima and thus behave—at least as far as the notion of local optimality is concerned—like convex problems. For related work on this topic and generalizations of the results of [30], see also [21, 31, 45, 54, 55]. Based on the findings of [30], it was conjectured that “nice” landscape properties or even the complete absence of spurious local minima can also be established for nonlinear activation functions in many situations and that this behavior is one of the main reasons for the performance that machine learning algorithms achieve in practice, cf. [21, 45, 53]. It was quickly realized, however, that, in the nonlinear case, the situation is more complicated and that examples of training problems with spurious local minima can readily be constructed even when only “mild” nonlinearities are present or the activation functions are piecewise affine. Data sets illustrating this for certain activation functions can be found, for example, in [43, 48, 54]. On the analytical level, one of the first general negative results on the landscape properties of training problems for neural networks was proven by Yun et al. in [54, Theorem 1]. They showed that spurious local minima are indeed *always* present when a finite-dimensional squared loss training problem for a one-hidden-layer neural network with a one-dimensional real output, a hidden layer of width at least two, and a leaky ReLU-type activation function is considered and the training data cannot be precisely fit with an affine function. This existence result was later also generalized in [25, Theorem 1] and [34, Theorem 1] to finite-dimensional training problems with arbitrary loss for deep networks with piecewise affine activation functions, in [20, Corollary 1] to finite-dimensional squared loss problems for deep networks with locally affine activations under the assumption of realizability, and in [11, Corollary 47] to finite-dimensional squared loss problems for deep networks involving many commonly used activation functions. For contributions on spurious minima in the absence of local affine linearity, see [11, 20, 41, 48, 54].

The purpose of the present paper is to prove that the results of [54] on the existence of spurious local minima in training problems for neural networks with piecewise affine activation functions are also true in a far more general setting and that the various assumptions on the activations, the loss function, the network architecture, and the realizability of the data in [11, 20, 25, 54] can be significantly relaxed. More precisely, we show that [54, Theorem 1] can be straightforwardly extended to networks of arbitrary depth, to arbitrary continuous nonpolynomial activation functions with an affine segment, to all (sensible) loss functions, and to infinite dimensions. We moreover

establish that there is a whole continuum of spurious local minima in the situation of [54, Theorem 1] whose Hausdorff dimension can be estimated from below. For the main results of our analysis, we refer the reader to Theorems 3.1 and 3.2. Note that these theorems in particular imply that the observations made in [25, 34, 54] are not a consequence of the piecewise affine linearity of the activation functions considered in these papers but of general effects that apply to all nonpolynomial continuous activation functions with an affine segment (SQLN, PLU, ReLU, leaky/parametric ReLU, ISRLU, ELU, etc.), that network training without spurious local minima is impossible (except for the pathological situation of affine linear training data) when the simple affine structure of [30] is kept locally but a global nonlinearity is introduced to enhance the approximation capabilities of the network, and that there always exist choices of hyperparameters such that gradient-based solution algorithms terminate with a suboptimal point when applied to training problems of the considered type.

We would like to point out that establishing the existence of local minima in training problems for neural networks whose activation functions possess an affine segment is not the main difficulty in the context of Theorems 3.1 and 3.2. To see that such minima are present, it suffices to exploit that neural networks with locally affine activations can emulate linear neural networks, see Lemmas 4.4 and 4.5, and this construction has already been used in various papers on the landscape properties of training problems, e.g., [11, 20, 24, 25, 54]. What is typically considered difficult in the literature is proving that the local minima obtained from the affine linear segments of the activation functions are indeed *always* spurious—independently of the precise form of the activations, the loss function, the training data, and the network architecture. Compare, for instance, with the comments in [54, Section 2.2], [11, Section 2.1], and [25, Section 3.3] on this topic. In existing works on the loss surface of neural networks, the problem of rigorously proving the spuriousness of local minima is usually addressed by manually constructing network parameters that yield smaller values of the loss function, cf. the proofs of [54, Theorem 1], [34, Theorem 1], and [25, Theorem 1]. Such constructions “by hand” are naturally only possible when simple activation functions and network architectures are considered and not suitable to obtain general results. One of the main points that we would like to communicate with this paper is that the spuriousness of the local minima in [54, Theorem 1], [25, Theorem 1], [11, Corollary 47], [34, Theorem 1], and [20, Corollary 1] and also our more general Theorems 3.1 and 3.2 is, in fact, a straightforward consequence of the universal approximation theorem in the arbitrary width formulation as proven by Cybenko, Hornik, and Pinkus in [17, 26, 42], or, more precisely, the fact that the universal approximation theorem implies that the image of a neural network with a fixed architecture does not possess any supporting half-spaces in function space; see Theorem 4.2. By exploiting this observation, we can easily overcome the assumption of [25, 34, 54] that the activation functions are piecewise affine linear, the restriction to the one-hidden-layer case in [54], the restriction to the squared loss function in [11, 20, 54], and the assumption of realizability in [20] and are moreover able to extend the results of these papers to infinite dimensions.

Due to their connection to the universal approximation theorem, the proofs of Theorems 3.1 and 3.2 also highlight the direct relationship that exists between the approximation capabilities of neural networks and the optimization landscape and

well-posedness properties of the training problems that have to be solved in order to determine a neural network best approximation. For further results on this topic, we refer to [14] and [42, Section 6], where it is discussed that every approximation instrument that asymptotically achieves a certain rate of convergence for the approximation error in terms of its number of degrees of freedom necessarily gives rise to numerical algorithms that are unstable. In a spirit similar to that of [14], we show in Sect. 5 that the nonexistence of supporting half-spaces exploited in the proofs of Theorems 3.1 and 3.2 also immediately implies that best approximation problems for neural networks posed in strictly convex Banach spaces with strictly convex duals are always ill-posed in the sense of Hadamard when the considered network does not have a dense image. Note that this result holds regardless of whether the activation functions possess an affine segment or not and without any assumptions on the widths of the hidden layers. We remark that, for one-hidden-layer networks, the corollaries in Sect. 5 have essentially already been proven in [28, 29], see also [40]. Our analysis extends the considerations of [28, 29] to arbitrary depths.

We conclude this introduction with an overview of the content and the structure of the remainder of the paper:

Section 2 is concerned with preliminaries. Here, we introduce the notation, the functional analytic setting, and the standing assumptions that we use in this work. In Sect. 3, we present our main results on the existence of spurious local minima, see Theorems 3.1 and 3.2. This section also discusses the scope and possible extensions of our analysis and demonstrates that Theorems 3.1 and 3.2 cover the squared loss problem studied in [54, Theorem 1] as a special case. Section 4 contains the proofs of Theorems 3.1 and 3.2. In this section, we establish that the universal approximation theorem indeed implies that the image of a neural network in function space does not possess any supporting half-spaces and show that this property allows us to prove the spuriousness of local minima in a natural way. In Sect. 5, we discuss further implications of the geometric properties of the images of neural networks exploited in Sect. 4. This section contains the already mentioned results on the Hadamard ill-posedness of neural network best approximation problems posed in strictly convex Banach spaces with strictly convex duals. Note that tangible examples of such spaces are  $L^p$ -spaces with  $1 < p < \infty$ , see Corollary 5.3. The paper concludes with additional comments on the results derived in Sects. 3, 4 and 5 and remarks on open problems.

## 2 Notation, Preliminaries, and Basic Assumptions

Throughout this work,  $K \subset \mathbb{R}^d$ ,  $d \in \mathbb{N}$ , denotes a nonempty compact subset of the Euclidean space  $\mathbb{R}^d$ . We endow  $K$  with the subspace topology  $\tau_K$  induced by the standard topology on  $(\mathbb{R}^d, |\cdot|)$ , where  $|\cdot|$  denotes the Euclidean norm, and denote the associated Borel sigma-algebra on  $K$  with  $\mathcal{B}(K)$ . The space of continuous functions  $v: K \rightarrow \mathbb{R}$  equipped with the maximum norm  $\|v\|_{C(K)} := \max\{|v(x)|: x \in K\}$  is denoted by  $C(K)$ . As usual, we identify the topological dual space  $C(K)^*$  of  $(C(K), \|\cdot\|_{C(K)})$  with the space  $\mathcal{M}(K)$  of signed Radon measures on  $(K, \mathcal{B}(K))$  endowed with the total variation norm  $\|\cdot\|_{\mathcal{M}(K)}$ , see [23, Corollary 7.18]. The corresponding dual pairing is denoted by  $\langle \cdot, \cdot \rangle_{C(K)}: \mathcal{M}(K) \times C(K) \rightarrow \mathbb{R}$ . For the closed cone of

nonnegative measures in  $\mathcal{M}(K)$ , we use the notation  $\mathcal{M}_+(K)$ . The standard, real Lebesgue spaces associated with a measure space  $(K, \mathcal{B}(K), \mu)$ ,  $\mu \in \mathcal{M}_+(K)$ , are denoted by  $L^p_\mu(K)$ ,  $1 \leq p \leq \infty$ , and equipped with the usual norms  $\|\cdot\|_{L^p_\mu(K)}$ , see [5, Section 5.5]. For the open ball of radius  $r > 0$  in a normed space  $(Z, \|\cdot\|_Z)$  centered at a point  $z \in Z$ , we use the symbol  $B^Z_r(z)$ , and for the topological closure of a set  $E \subset Z$ , the symbol  $\text{cl}_Z(E)$ .

The neural networks that we study in this paper are standard fully connected feed-forward neural networks with a  $d$ -dimensional real input and a one-dimensional real output (with  $d$  being the dimension of the Euclidean space  $\mathbb{R}^d \supset K$ ). We denote the number of hidden layers of a network with  $L \in \mathbb{N}$  and the widths of the hidden layers with  $w_i \in \mathbb{N}$ ,  $i = 1, \dots, L$ . For the ease of notation, we also introduce the definitions  $w_0 := d$  and  $w_{L+1} := 1$  for the in- and output layer. The weights and biases are denoted by  $A_i \in \mathbb{R}^{w_i \times w_{i-1}}$  and  $b_i \in \mathbb{R}^{w_i}$ ,  $i = 1, \dots, L + 1$ , respectively, and the activation functions of the layers by  $\sigma_i : \mathbb{R} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, L$ . Here and in what follows, all vectors of real numbers are considered as column vectors. We will always assume that the functions  $\sigma_i$  are continuous, i.e.,  $\sigma_i \in C(\mathbb{R})$  for all  $i = 1, \dots, L$ . To describe the action of the network layers, we define  $\varphi_i^{A_i, b_i} : \mathbb{R}^{w_{i-1}} \rightarrow \mathbb{R}^{w_i}$ ,  $i = 1, \dots, L + 1$ , to be the functions

$$\varphi_i^{A_i, b_i}(z) := \sigma_i(A_i z + b_i), \quad \forall i = 1, \dots, L, \quad \varphi_{L+1}^{A_{L+1}, b_{L+1}}(z) := A_{L+1} z + b_{L+1},$$

with  $\sigma_i$  acting componentwise on the entries of  $A_i z + b_i \in \mathbb{R}^{w_i}$ . Overall, this notation allows us to denote a feedforward neural network in the following way:

$$\psi(\alpha, \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}, \quad \psi(\alpha, x) := \left( \varphi_{L+1}^{A_{L+1}, b_{L+1}} \circ \dots \circ \varphi_1^{A_1, b_1} \right)(x). \quad (2.1)$$

Here, we have introduced the variable  $\alpha := \{(A_i, b_i)\}_{i=1}^{L+1}$  as an abbreviation for the collection of all network parameters and the symbol “ $\circ$ ” to denote a composition. For the set of all possible  $\alpha$ , i.e., the parameter space of a network, we write

$$D := \left\{ \alpha = \{(A_i, b_i)\}_{i=1}^{L+1} \mid A_i \in \mathbb{R}^{w_i \times w_{i-1}}, b_i \in \mathbb{R}^{w_i}, \forall i = 1, \dots, L + 1 \right\}.$$

We equip the parameter space  $D$  with the Euclidean norm  $|\cdot|$  of the space  $\mathbb{R}^m$ ,  $m := w_{L+1}(w_L + 1) + \dots + w_1(w_0 + 1)$ , that  $D$  can be transformed into by rearranging the entries of  $\alpha$ . Note that this implies that  $m = \dim(D)$  holds, where  $\dim(\cdot)$  denotes the dimension of a vector space in the sense of linear algebra. Due to the continuity of the activation functions  $\sigma_i$ , the map  $\psi : D \times \mathbb{R}^d \rightarrow \mathbb{R}$  in (2.1) gives rise to an operator from  $D$  into the space  $C(K)$ . We denote this operator by  $\Psi$ , i.e.,

$$\Psi : D \rightarrow C(K), \quad \Psi(\alpha) := \psi(\alpha, \cdot) : K \rightarrow \mathbb{R}. \quad (2.2)$$

Using the function  $\Psi$ , we can formulate the training problems that we are interested in as follows:

$$\text{Minimize } \mathcal{L}(\Psi(\alpha), y_T) \quad \text{w.r.t. } \alpha \in D. \quad (\text{P})$$

Here,  $\mathcal{L}: C(K) \times C(K) \rightarrow \mathbb{R}$  denotes the loss function and  $y_T \in C(K)$  the target function. We call  $\mathcal{L}$  Gâteaux differentiable in its first argument at  $(v, y_T) \in C(K) \times C(K)$  if the limit

$$\partial_1 \mathcal{L}(v, y_T; h) := \lim_{s \rightarrow 0^+} \frac{\mathcal{L}(v + sh, y_T) - \mathcal{L}(v, y_T)}{s} \in \mathbb{R}$$

exists for all  $h \in C(K)$  and if the map  $\partial_1 \mathcal{L}(v, y_T; \cdot): C(K) \rightarrow \mathbb{R}, h \mapsto \partial_1 \mathcal{L}(v, y_T; h)$ , is linear and continuous, i.e., an element of the topological dual space of  $C(K)$ . In this case,  $\partial_1 \mathcal{L}(v, y_T) := \partial_1 \mathcal{L}(v, y_T; \cdot) \in \mathcal{M}(K)$  is called the partial Gâteaux derivative of  $\mathcal{L}$  at  $(v, y_T)$  w.r.t. the first argument, cf. [8, Section 2.2.1]. As usual, a local minimum of (P) is a point  $\bar{\alpha} \in D$  that satisfies

$$\mathcal{L}(\Psi(\alpha), y_T) \geq \mathcal{L}(\Psi(\bar{\alpha}), y_T), \quad \forall \alpha \in B_r^D(\bar{\alpha}),$$

for some  $r > 0$ . If  $r$  can be chosen as  $+\infty$ , then we call  $\bar{\alpha}$  a global minimum of (P). For a local minimum that is not a global minimum, we use the term *spurious local minimum*. We would like to point out that we will not discuss the existence of global minima of (P) in this paper. In fact, it is easy to construct examples in which (P) does not admit any global solutions, cf. [40]. We will focus entirely on the existence of spurious local minima that may prevent optimization algorithms from producing a minimizing sequence for (P), i.e., a sequence  $\{\alpha_k\}_{k=1}^\infty \subset D$  satisfying

$$\lim_{k \rightarrow \infty} \mathcal{L}(\Psi(\alpha_k), y_T) = \inf_{\alpha \in D} \mathcal{L}(\Psi(\alpha), y_T).$$

For later use, we recall that the Hausdorff dimension  $\dim_{\mathcal{H}}(E)$  of a set  $E \subset \mathbb{R}^m$  is defined by

$$\dim_{\mathcal{H}}(E) := \inf \{s \in [0, \infty) \mid \mathcal{H}_s(E) = 0\},$$

where  $\mathcal{H}_s(E)$  denotes the  $s$ -dimensional Hausdorff outer measure

$$\mathcal{H}_s(E) := \lim_{\epsilon \rightarrow 0^+} \left( \inf \left\{ \sum_{l=1}^{\infty} \text{diam}(E_l)^s \mid E \subset \bigcup_{l=1}^{\infty} E_l, \text{diam}(E_l) < \epsilon \right\} \right). \quad (2.3)$$

Here,  $\text{diam}(\cdot)$  denotes the diameter of a set and the infimum on the right-hand side of (2.3) is taken over the set of covers  $\{E_l\}_{l=1}^\infty$ ; see [19, Sections 3.5, 3.5.1c]. Note that the Hausdorff dimension of a subspace is identical to the “usual” dimension of the subspace in the sense of linear algebra. In particular, we have  $\dim(D) = \dim_{\mathcal{H}}(D) = m$ .

### 3 Main Results on the Existence of Spurious Local Minima

With the notation in place, we are in the position to formulate our main results on the existence of spurious local minima in training problems for neural networks whose

activation functions possess an affine segment. To be precise, we state our main observation in the form of two theorems—one for activation functions with a nonconstant affine segment and one for activation functions with a constant segment.

**Theorem 3.1** (Case I: activation functions with a nonconstant affine segment) *Let  $K \subset \mathbb{R}^d$ ,  $d \in \mathbb{N}$ , be a nonempty compact set and let  $\psi : D \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a neural network with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ ,  $i = 0, \dots, L + 1$ , and nonpolynomial continuous activation functions  $\sigma_i : \mathbb{R} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, L$ , as in (2.1). Assume that:*

- (i)  $w_i \geq 2$  holds for all  $i = 1, \dots, L$ .
- (ii)  $\sigma_i$  is affine and nonconstant on an open interval  $I_i \neq \emptyset$  for all  $i = 1, \dots, L$ .
- (iii)  $y_T \in C(K)$  is nonaffine, i.e.,  $\nexists(a, c) \in \mathbb{R}^d \times \mathbb{R} : y_T(x) = a^\top x + c, \forall x \in K$ .
- (iv)  $\mathcal{L} : C(K) \times C(K) \rightarrow \mathbb{R}$  is Gâteaux differentiable in its first argument with a nonzero partial derivative at all points  $(v, y_T) \in C(K) \times C(K)$  with  $v \neq y_T$ .
- (v)  $\mathcal{L}$  and  $y_T$  are such that there exists a global solution  $(\bar{a}, \bar{c})$  of the problem

$$\text{Minimize } \mathcal{L}(z_{a,c}, y_T) \text{ w.r.t. } (a, c) \in \mathbb{R}^d \times \mathbb{R} \text{ s.t. } z_{a,c}(x) = a^\top x + c.$$

Then there exists a set  $E \subset D$  of Hausdorff dimension  $\dim_{\mathcal{H}}(E) \geq m - d - 1$  such that all elements of  $E$  are spurious local minima of the training problem

$$\text{Minimize } \mathcal{L}(\Psi(\alpha), y_T) \text{ w.r.t. } \alpha \in D \tag{P}$$

and such that it holds

$$\mathcal{L}(\Psi(\alpha), y_T) = \min_{(a,c) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(z_{a,c}, y_T), \quad \forall \alpha \in E.$$

**Theorem 3.2** (Case II: activation functions with a constant segment) *Suppose that  $K \subset \mathbb{R}^d$ ,  $d \in \mathbb{N}$ , is a nonempty compact set and let  $\psi : D \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a neural network with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ ,  $i = 0, \dots, L + 1$ , and nonpolynomial continuous activation functions  $\sigma_i : \mathbb{R} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, L$ , as in (2.1). Assume that:*

- (i)  $\sigma_j$  is constant on an open interval  $I_j \neq \emptyset$  for some  $j \in \{1, \dots, L\}$ .
- (ii)  $y_T \in C(K)$  is nonconstant, i.e.,  $\nexists c \in \mathbb{R} : y_T(x) = c, \forall x \in K$ .
- (iii)  $\mathcal{L} : C(K) \times C(K) \rightarrow \mathbb{R}$  is Gâteaux differentiable in its first argument with a nonzero partial derivative at all points  $(v, y_T) \in C(K) \times C(K)$  with  $v \neq y_T$ .
- (iv)  $\mathcal{L}$  and  $y_T$  are such that there exists a global solution  $\bar{c}$  of the problem

$$\text{Minimize } \mathcal{L}(z_c, y_T) \text{ w.r.t. } c \in \mathbb{R} \text{ s.t. } z_c(x) = c.$$

Then there exists a set  $E \subset D$  of Hausdorff dimension  $\dim_{\mathcal{H}}(E) \geq m - 1$  such that all elements of  $E$  are spurious local minima of the training problem

$$\text{Minimize } \mathcal{L}(\Psi(\alpha), y_T) \text{ w.r.t. } \alpha \in D \tag{P}$$

and such that it holds

$$\mathcal{L}(\Psi(\alpha), y_T) = \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T), \quad \forall \alpha \in E.$$

The proofs of Theorems 3.1 and 3.2 rely on geometric properties of the image  $\Psi(D)$  of the function  $\Psi : D \rightarrow C(K)$  in (2.2) and are carried out in Sect. 4, see Theorem 4.2 and Lemmas 4.4 to 4.7. Before we discuss them in detail, we give some remarks on the applicability and scope of Theorems 3.1 and 3.2.

First of all, we would like to point out that—as far as continuous activation functions with an affine segment are concerned—the assumptions on the maps  $\sigma_i$  in Theorems 3.1 and 3.2 are optimal. The only continuous  $\sigma_i$  that are locally affine and not covered by Theorems 3.1 and 3.2 are globally affine functions and for those it has been proven in [30] that spurious local minima do not exist so that relaxing the assumptions on  $\sigma_i$  in Theorems 3.1 and 3.2 in this direction is provably impossible. Compare also with [21, 31, 45, 54, 55] in this context. Note that Theorem 3.1 covers in particular neural networks that involve an arbitrary mixture of PLU-, ISRLU-, ELU-, ReLU-, and leaky/parametric ReLU-activations and that Theorem 3.2 applies, for instance, to neural networks with a ReLU- or an SQNL-layer; see [38] and [11, Corollary 40] for the definitions of these functions. Because of this, the assertions of Theorems 3.1 and 3.2 hold in many situations arising in practice.

Second, we remark that Theorems 3.1 and 3.2 can be rather easily extended to neural networks with a vectorial output. For such networks, the assumptions on the widths  $w_i$  in point (i) of Theorem 3.1 have to be adapted depending on the in- and output dimension, but the basic ideas of the proofs remain the same, cf. the analysis of [25] and the proof of [11, Corollary 47]. In particular, the arguments that we use in Sect. 4 to establish that the local minima in  $E$  are indeed spurious carry over immediately. Similarly, it is also possible to extend the ideas presented in this paper to residual neural networks. To do so, one can exploit that networks with skip connections can emulate classical multilayer perceptron architectures of the type (2.1) on the training set  $K$  by saturation, cf. [11, proof of Corollary 52], and that skip connections do not impair the ability of a network with locally affine activation functions to emulate an affine linear mapping, cf. the proofs of Lemmas 4.4 and 4.5. We omit discussing these generalizations in detail here to simplify the presentation.

Regarding the assumptions on  $\mathcal{L}$ , it should be noted that the conditions in points (iv) and (v) of Theorem 3.1 and points (iii) and (iv) of Theorem 3.2 are not very restrictive. The assumption that the partial Gâteaux derivative  $\partial_1 \mathcal{L}(v, y_T)$  is nonzero for  $v \neq y_T$  simply expresses that the map  $\mathcal{L}(\cdot, y_T) : C(K) \rightarrow \mathbb{R}$  should not have any stationary points away from  $y_T$ . This is a reasonable thing to assume since the purpose of the loss function is to measure the deviation from  $y_T$  so that stationary points away from  $y_T$  are not sensible. In particular, this assumption is automatically satisfied if  $\mathcal{L}$  has the form  $\mathcal{L}(v, y_T) = \mathcal{F}(v - y_T)$  with a convex function  $\mathcal{F} : C(K) \rightarrow [0, \infty)$  that is Gâteaux differentiable in  $C(K) \setminus \{0\}$  and satisfies  $\mathcal{F}(v) = 0$  iff  $v = 0$ . Similarly, the assumptions on the existence of the minimizers  $(\bar{a}, \bar{c})$  and  $\bar{c}$  in Theorems 3.1 and 3.2 simply express that there should exist an affine linear/constant best approximation for  $y_T$  w.r.t. the notion of approximation quality encoded in  $\mathcal{L}$ . This condition is, for instance, satisfied when restrictions of the map  $\mathcal{L}(\cdot, y_T) : C(K) \rightarrow \mathbb{R}$  to finite-dimensional subspaces of  $C(K)$  are radially unbounded and lower semicontinuous. A prototypical class of functions  $\mathcal{L}$  that satisfy all of the above conditions are tracking-type functionals in reflexive Lebesgue spaces as the following lemma shows.



**Lemma 3.3** *Let  $K \subset \mathbb{R}^d$  be nonempty and compact, let  $\mu \in \mathcal{M}_+(K)$  be a measure whose support is equal to  $K$ , and let  $1 < p < \infty$  be given. Define*

$$\mathcal{L}: C(K) \times C(K) \rightarrow [0, \infty), \quad \mathcal{L}(v, y_T) := \int_K |v - y_T|^p \, d\mu. \tag{3.1}$$

*Then the function  $\mathcal{L}$  satisfies the assumptions (iv) and (v) of Theorem 3.1 and the assumptions (iii) and (iv) of Theorem 3.2 for all  $y_T \in C(K)$ .*

**Proof** From the dominated convergence theorem [5, Theorem 3.3.2], it follows that  $\mathcal{L}: C(K) \times C(K) \rightarrow [0, \infty)$  is Gâteaux differentiable everywhere with

$$\langle \partial_1 \mathcal{L}(v, y_T), z \rangle_{C(K)} = \int_K p \operatorname{sgn}(v - y_T) |v - y_T|^{p-1} z \, d\mu, \quad \forall v, y_T, z \in C(K). \tag{3.2}$$

Since  $C(K)$  is dense in  $L^q_\mu(K)$  for all  $1 \leq q < \infty$  by [23, Proposition 7.9], (3.2) yields

$$\partial_1 \mathcal{L}(v, y_T) = 0 \in \mathcal{M}(K) \iff \int_K p |v - y_T|^{p-1} \, d\mu = 0. \tag{3.3}$$

Due to the continuity of the function  $|v - y_T|^{p-1}$  and since the assumptions on  $\mu$  imply that  $\mu(O) > 0$  holds for all  $O \in \tau_K \setminus \{\emptyset\}$ , the right-hand side of (3.3) can only be true if  $v - y_T$  is the zero function in  $C(K)$ , i.e., if  $v = y_T$ . This shows that  $\mathcal{L}$  indeed satisfies condition (iv) in Theorem 3.1 and condition (iii) in Theorem 3.2 for all  $y_T \in C(K)$ . To see that  $\mathcal{L}$  also satisfies assumption (v) of Theorem 3.1 and assumption (iv) of Theorem 3.2, it suffices to note that  $\|\cdot\|_{L^p_\mu(K)}$  defines a norm on  $C(K)$  due to the assumptions on  $\mu$ . This implies that restrictions of the map  $\mathcal{L}(\cdot, y_T): C(K) \rightarrow \mathbb{R}$  to finite-dimensional subspaces of  $C(K)$  are continuous and radially unbounded for all arbitrary but fixed  $y_T \in C(K)$  and that the theorem of Weierstrass can be used to establish the existence of the minimizers  $(\bar{a}, \bar{c})$  and  $\bar{c}$  in points (v) and (iv) of Theorems 3.1 and 3.2, respectively.  $\square$

Note that, in the case  $\mu = \frac{1}{n} \sum_{k=1}^n \delta_{x_k}$ ,  $K = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ,  $d \in \mathbb{N}$ ,  $n \in \mathbb{N}$ , i.e., in the situation where  $\mu$  is the normalized sum of  $n$  Dirac measures supported at points  $x_k \in \mathbb{R}^d$ ,  $k = 1, \dots, n$ , a problem (P) with a loss function of the form (3.1) can be recast as

$$\text{Minimize } \frac{1}{n} \sum_{k=1}^n |\psi(\alpha, x_k) - y_T(x_k)|^p \quad \text{w.r.t. } \alpha \in D. \tag{3.4}$$

In particular, for  $p = 2$ , one recovers a classical squared loss problem with a finite number of data samples. This shows that our results indeed extend [54, Theorem 1], where the assertion of Theorem 3.1 was proven for finite-dimensional squared loss training problems for one-hidden-layer neural networks with activation functions of parameterized ReLU-type. Compare also with [11, 20, 25, 34] in this context. Another natural choice for  $\mu$  in (3.1) is the restriction of the Lebesgue measure to the Borel sigma-algebra of the closure  $K$  of a nonempty bounded open set  $\Omega \subset \mathbb{R}^d$ . For this

choice, (P) becomes a standard  $L^p$ -tracking-type problem as often considered in the field of optimal control, cf. [12] and the references therein. A further interesting example is the case  $K = \text{cl}_{\mathbb{R}^d}(\{x_k\}_{k=1}^{\infty})$  and  $\mu = \sum_{k=1}^{\infty} c_k \delta_{x_k}$  involving a bounded sequence of points  $\{x_k\}_{k=1}^{\infty} \subset \mathbb{R}^d$  and weights  $\{c_k\}_{k=1}^{\infty} \subset (0, \infty)$  with  $\sum_{k=1}^{\infty} c_k < \infty$ . Such a measure  $\mu$  gives rise to a training problem in an intermediate regime between the finite and continuous sampling case.

We remark that, for problems of the type (3.4) with  $p = 2$ , it can be shown that the spurious local minima in Theorems 3.1 and 3.2 can be arbitrarily bad in the sense that they may yield loss values that are arbitrarily far away from the optimal one and may give rise to realization vectors  $\{\psi(\alpha, x_k)\}_{k=1}^n$  that are arbitrarily far away in relative and absolute terms from every optimal realization vector of the network. For a precise statement of these results for finite-dimensional squared loss problems and the definitions of the related concepts, we refer the reader to [11, Corollary 47, Definition 3, and Estimates (39), (40)]. Similarly, it can also be proven that the appearance of spurious local minima in problems of the type (3.4) can, in general, not be avoided by adding a regularization term to the loss function that penalizes the size of the parameters in  $\alpha$ , see [11, Corollary 51]. We remark that the proofs used to establish these results in [11] all make use of compactness arguments and homogeneity properties of  $\mathcal{L}$  and thus do not carry over to the general infinite-dimensional setting considered in Theorems 3.1 and 3.2, cf. the derivation of [11, Lemma 10].

As a final remark, we would like to point out that, in the degenerate case  $n = 1$ , the training problem (3.4) *does not* possess any spurious local minima (as one may easily check by varying the bias  $b_{L+1}$  on the last layer of  $\psi$ ). This effect does not contradict our results since, for  $n = 1$ , the set  $K$  is a singleton, every  $y_T \in C(K) \cong \mathbb{R}$  can be precisely fit with a constant function, and condition (iii) in Theorem 3.1 and condition (ii) in Theorem 3.2 are always violated. Note that this highlights that the assumptions of Theorems 3.1 and 3.2 are sharp.

## 4 Nonexistence of Supporting Half-Spaces and Proofs of Main Results

In this section, we prove Theorems 3.1 and 3.2. The point of departure for our analysis is the following theorem of Pinkus.

**Theorem 4.1** [42, Theorem 3.1] *Let  $d \in \mathbb{N}$ . Let  $\sigma: \mathbb{R} \rightarrow \mathbb{R}$  be a nonpolynomial continuous function. Consider the linear hull*

$$V := \text{span} \left\{ x \mapsto \sigma(a^\top x + c) \mid a \in \mathbb{R}^d, c \in \mathbb{R} \right\} \subset C(\mathbb{R}^d). \quad (4.1)$$

*Then the set  $V$  is dense in  $C(\mathbb{R}^d)$  in the topology of uniform convergence on compacta.*

Note that, as  $V$  contains precisely those functions that can be represented by one-hidden-layer neural networks of the type (2.1) with  $\sigma_1 = \sigma$ , the last theorem is nothing else than the universal approximation theorem in the arbitrary width case, cf. [17, 26]. In other words, Theorem 4.1 simply expresses that, for every nonpolynomial  $\sigma \in C(\mathbb{R})$ , every nonempty compact set  $K \subset \mathbb{R}^d$ , every  $y_T \in C(K)$ , and every

$\varepsilon > 0$ , there exists a width  $\tilde{w}_1 \in \mathbb{N}$  such that a neural network  $\psi$  with the architecture in (2.1), depth  $L = 1$ , width  $w_1 \geq \tilde{w}_1$ , and activation function  $\sigma$  is able to approximate  $y_T$  in  $(C(K), \|\cdot\|_{C(K)})$  up to the error  $\varepsilon$ . In what follows, we will not explore what Theorem 4.1 implies for the approximation capabilities of neural networks when the widths go to infinity but rather which consequences the density of the space  $V$  in (4.1) has for a given neural network with a fixed architecture. More precisely, we will use Theorem 4.1 to prove that the image  $\Psi(D) \subset C(K)$  of the function  $\Psi: D \rightarrow C(K)$  in (2.2) does not admit any supporting half-spaces when a neural network  $\psi$  with nonpolynomial continuous activations  $\sigma_i$  and arbitrary fixed dimensions  $L, w_i \in \mathbb{N}$  is considered.

**Theorem 4.2** (Nonexistence of supporting half-spaces) *Let  $K \subset \mathbb{R}^d, d \in \mathbb{N}$ , be a nonempty compact set and let  $\psi: D \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a neural network with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}, i = 0, \dots, L + 1$ , and continuous nonpolynomial activation functions  $\sigma_i: \mathbb{R} \rightarrow \mathbb{R}, i = 1, \dots, L$ , as in (2.1). Denote with  $\Psi: D \rightarrow C(K)$  the function in (2.2). Then a measure  $\mu \in \mathcal{M}(K)$  and a constant  $c \in \mathbb{R}$  satisfy*

$$\langle \mu, z \rangle_{C(K)} \leq c, \quad \forall z \in \Psi(D), \tag{4.2}$$

if and only if  $\mu = 0$  and  $c \geq 0$ .

**Proof** The implication “ $\Leftarrow$ ” is trivial. To prove “ $\Rightarrow$ ”, we assume that  $c \in \mathbb{R}$  and  $\mu \in \mathcal{M}(K)$  satisfying (4.2) are given. From the definition of  $\Psi$ , we obtain that  $\beta\Psi(\alpha) \in \Psi(D)$  holds for all  $\beta \in \mathbb{R}$  and all  $\alpha \in D$ . If we exploit this property in (4.2), then we obtain that  $c$  and  $\mu$  have to satisfy  $c \geq 0$  and

$$\langle \mu, z \rangle_{C(K)} = 0, \quad \forall z \in \Psi(D). \tag{4.3}$$

It remains to prove that  $\mu$  vanishes. To this end, we first reduce the situation to the case  $w_1 = \dots = w_L = 1$ . Consider a parameter  $\tilde{\alpha} \in D$  whose weights and biases have the form

$$\begin{aligned} \tilde{A}_1 &:= \begin{pmatrix} a_1^\top \\ 0_{(w_1-1) \times d} \end{pmatrix}, \quad \tilde{A}_i := \begin{pmatrix} a_i & 0_{1 \times (w_{i-1}-1)} \\ 0_{(w_i-1) \times w_{i-1}} \end{pmatrix}, \quad i = 2, \dots, L + 1, \\ \tilde{b}_i &:= \begin{pmatrix} c_i \\ 0_{w_{i-1}} \end{pmatrix}, \quad i = 1, \dots, L + 1, \end{aligned} \tag{4.4}$$

for some arbitrary but fixed  $a_1 \in \mathbb{R}^d, a_i \in \mathbb{R}, i = 2, \dots, L + 1$ , and  $c_i \in \mathbb{R}, i = 1, \dots, L + 1$ , where  $0_{p \times q} \in \mathbb{R}^{p \times q}$  and  $0_p \in \mathbb{R}^p$  denote the zero matrix and zero vector in  $\mathbb{R}^{p \times q}$  and  $\mathbb{R}^p, p, q \in \mathbb{N}$ , respectively, with the convention that these zero entries are ignored in the case  $p = 0$  or  $q = 0$ . For such a parameter  $\tilde{\alpha}$ , we obtain from (2.1) that

$$\psi(\tilde{\alpha}, x) = (\theta_{L+1}^{a_{L+1}, c_{L+1}} \circ \dots \circ \theta_1^{a_1, c_1})(x), \quad \forall x \in \mathbb{R}^d,$$

holds with the functions  $\theta_1^{a_1, c_1} : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\theta_i^{a_i, c_i} : \mathbb{R} \rightarrow \mathbb{R}$ ,  $i = 2, \dots, L + 1$ , given by

$$\theta_1^{a_1, c_1}(z) := \sigma_1 \left( a_1^\top z + c_1 \right), \quad \theta_i^{a_i, c_i}(z) := \sigma_i \left( a_i z + c_i \right), \quad \forall i = 2, \dots, L,$$

$$\theta_{L+1}^{a_{L+1}, c_{L+1}}(z) := a_{L+1} z + c_{L+1}.$$

In combination with (4.3) and the definition of  $\Psi$ , this yields that

$$\langle \mu, \theta_{L+1}^{a_{L+1}, c_{L+1}} \circ \dots \circ \theta_1^{a_1, c_1} \rangle_{C(K)} = \int_K \left( \theta_{L+1}^{a_{L+1}, c_{L+1}} \circ \dots \circ \theta_1^{a_1, c_1} \right) (x) \, d\mu(x) = 0 \tag{4.5}$$

holds for all  $a_i, c_i, i = 1, \dots, L + 1$ . Next, we use Theorem 4.1 to reduce the number of layers in (4.5). Suppose that  $L > 1$  holds and let  $a_i, c_i, i \in \{1, \dots, L + 1\} \setminus \{L\}$ , be arbitrary but fixed parameters. From the compactness of  $K$  and the continuity of the function  $K \ni x \mapsto \left( \theta_{L-1}^{a_{L-1}, c_{L-1}} \circ \dots \circ \theta_1^{a_1, c_1} \right) (x) \in \mathbb{R}$ , we obtain that the image  $F := \left( \theta_{L-1}^{a_{L-1}, c_{L-1}} \circ \dots \circ \theta_1^{a_1, c_1} \right) (K) \subset \mathbb{R}$  is compact, and from Theorem 4.1, it follows that there exist numbers  $n_l \in \mathbb{N}$  and  $\beta_{k,l}, \gamma_{k,l}, \lambda_{k,l} \in \mathbb{R}, k = 1, \dots, n_l, l \in \mathbb{N}$ , such that the sequence of continuous functions

$$\zeta_l : F \rightarrow \mathbb{R}, \quad z \mapsto \sum_{k=1}^{n_l} \lambda_{k,l} \sigma_L(\beta_{k,l} z + \gamma_{k,l}),$$

converges uniformly on  $F$  to the identity map for  $l \rightarrow \infty$ . Since (4.5) holds for all choices of parameters, we further know that

$$\int_K a_{L+1} \lambda_{k,l} \sigma_L \left( \beta_{k,l} \left( \theta_{L-1}^{a_{L-1}, c_{L-1}} \circ \dots \circ \theta_1^{a_1, c_1} \right) (x) + \gamma_{k,l} \right) + \frac{1}{n_l} c_{L+1} \, d\mu(x) = 0$$

holds for all  $k = 1, \dots, n_l$  and all  $l \in \mathbb{N}$ . Due to the linearity of the integral, we can add all of the above equations to obtain that

$$\int_K a_{L+1} \zeta_l \left[ \left( \theta_{L-1}^{a_{L-1}, c_{L-1}} \circ \dots \circ \theta_1^{a_1, c_1} \right) (x) \right] + c_{L+1} \, d\mu(x) = 0, \quad \forall l \in \mathbb{N},$$

holds and, after passing to the limit  $l \rightarrow \infty$  by means of the dominated convergence theorem, that

$$\int_K a_{L+1} \left( \theta_{L-1}^{a_{L-1}, c_{L-1}} \circ \dots \circ \theta_1^{a_1, c_1} \right) (x) + c_{L+1} \, d\mu(x) = 0.$$

Since  $a_i, c_i, i \in \{1, \dots, L + 1\} \setminus \{L\}$ , were arbitrary, this is precisely (4.5) with the  $L$ -th layer removed. By proceeding iteratively along the above lines, it follows that  $\mu$  satisfies

$$\int_K a_{L+1} \sigma_1(a_1^\top x + c_1) + c_{L+1} \, d\mu(x) = 0$$

for all  $a_{L+1}, c_{L+1}, c_1 \in \mathbb{R}$  and all  $a_1 \in \mathbb{R}^d$ . Again by the density in (4.1) and the linearity of the integral, this identity can only be true if  $\langle \mu, z \rangle_{C(K)} = 0$  holds for all  $z \in C(K)$ . Thus,  $\mu = 0$  and the proof is complete.  $\square$

**Remark 4.3** • Theorem 4.2 is, in fact, equivalent to Theorem 4.1. Indeed, the implication “Theorem 4.1  $\Rightarrow$  Theorem 4.2” has been proven above. To see that Theorem 4.2 implies Theorem 4.1, one can argue by contradiction. If the space  $V$  in (4.1) is not dense in  $C(\mathbb{R}^d)$  in the topology of uniform convergence on compacta, then there exist a nonempty compact set  $K \subset \mathbb{R}^d$  and a nonzero  $\mu \in \mathcal{M}(K)$  such that  $\langle \mu, v \rangle_{C(K)} = 0$  holds for all  $v \in V$ , cf. the proof of [42, Proposition 3.10]. Since Theorem 4.2 applies to networks with  $L = 1$  and  $w_1 = 1$ , the variational identity  $\langle \mu, v \rangle_{C(K)} = 0$  for all  $v \in V$  can only be true if  $\mu = 0$ . Hence, one arrives at a contradiction and the density in Theorem 4.1 follows. Compare also with the classical proofs of the universal approximation theorem in [17] and [26] in this context which prove results similar to Theorem 4.2 as an intermediate step. In combination with the comments after Theorem 4.1, this shows that the arguments that we use in the following to establish the existence of spurious local minima in training problems of the form (P) are indeed closely related to the universal approximation property.

- It is easy to check that the nonexistence of supporting half-spaces in Theorem 4.2 implies that, for every finite training set  $K = \{x_1, \dots, x_n\}$  and every network  $\psi$  with associated function  $\Psi: D \rightarrow C(K) \cong \mathbb{R}^n$  satisfying the assumptions of Theorem 4.2, we have

$$\sup_{y_T \in \mathbb{R}^n: |y_T|=1} \inf_{y \in \Psi(D)} |y - y_T|^2 < 1. \tag{4.6}$$

This shows that Theorem 4.2 implies the “improved expressiveness”-condition in [11, Assumption 6-II)] and may be used to establish an alternative proof of [11, Theorem 39, Corollary 40]. We remark that, for infinite  $K$ , a condition analogous to (4.6) cannot be expected to hold for a neural network. In our analysis, Theorem 4.2 serves as a substitute for (4.6) that remains true in the infinite-dimensional setting and for arbitrary loss functions.

We are now in the position to prove Theorems 3.1 and 3.2. We begin by constructing the sets of local minima  $E \subset D$  that appear in these theorems. As before, we distinguish between activation functions with a nonconstant affine segment and activation functions with a constant segment.

**Lemma 4.4** Consider a nonempty compact set  $K \subset \mathbb{R}^d$  and a neural network  $\psi: D \times \mathbb{R}^d \rightarrow \mathbb{R}$  with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ , and continuous activation functions  $\sigma_i$  as in (2.1). Suppose that  $\mathcal{L}: C(K) \times C(K) \rightarrow \mathbb{R}$  and  $y_T \in C(K)$  are given such that  $\mathcal{L}$ ,  $y_T$ , and the functions  $\sigma_i$  satisfy the conditions (ii) and (v) in Theorem 3.1. Then there exists a set  $E \subset D$  of Hausdorff dimension  $\dim_{\mathcal{H}}(E) \geq m - d - 1$  such that all elements of  $E$  are local minima of (P) and such that

$$\mathcal{L}(\Psi(\alpha), y_T) = \min_{(a,c) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(z_{a,c}, y_T) \tag{4.7}$$

holds for all  $\alpha \in E$ , where  $z_{a,c}$  is defined by  $z_{a,c}(x) := a^\top x + c$  for all  $x \in \mathbb{R}^d$ .

**Proof** Due to (ii), we can find numbers  $c_i \in \mathbb{R}$ ,  $\varepsilon_i > 0$ ,  $\beta_i \in \mathbb{R} \setminus \{0\}$ , and  $\gamma_i \in \mathbb{R}$  such that  $\sigma_i(s) = \beta_i s + \gamma_i$  holds for all  $s \in I_i = (c_i - \varepsilon_i, c_i + \varepsilon_i)$  and all  $i = 1, \dots, L$ , and from (v), we obtain that there exist  $\bar{a} \in \mathbb{R}^d$  and  $\bar{c} \in \mathbb{R}$  satisfying

$$\mathcal{L}(z_{a,c}, y_T) \geq \mathcal{L}(z_{\bar{a},\bar{c}}, y_T), \quad \forall (a, c) \in \mathbb{R}^d \times \mathbb{R}.$$

Consider now the parameter  $\bar{\alpha} = \{(\bar{A}_i, \bar{b}_i)\}_{i=1}^{L+1} \in D$  whose weights and biases are given by

$$\begin{aligned} \bar{A}_1 &:= \frac{\varepsilon_1}{2 \max_{u \in K} |\bar{a}^\top u| + 1} \begin{pmatrix} \bar{a}^\top \\ 0_{(w_1-1) \times w_0} \end{pmatrix} \in \mathbb{R}^{w_1 \times w_0}, \\ \bar{b}_1 &:= c_1 1_{w_1} \in \mathbb{R}^{w_1}, \\ \bar{A}_i &:= \frac{\varepsilon_i}{\beta_{i-1} \varepsilon_{i-1}} \begin{pmatrix} 1 & 0_{1 \times (w_{i-1}-1)} \\ 0_{(w_{i-1}) \times w_{i-1}} \end{pmatrix} \in \mathbb{R}^{w_i \times w_{i-1}}, \quad i = 2, \dots, L, \\ \bar{b}_i &:= c_i 1_{w_i} - (c_{i-1} \beta_{i-1} + \gamma_{i-1}) \bar{A}_i 1_{w_{i-1}} \in \mathbb{R}^{w_i}, \quad i = 2, \dots, L, \\ \bar{A}_{L+1} &:= \frac{2 \max_{u \in K} |\bar{a}^\top u| + 1}{\beta_L \varepsilon_L} (1 \ 0_{1 \times (w_L-1)}) \in \mathbb{R}^{w_{L+1} \times w_L}, \\ \bar{b}_{L+1} &:= \bar{c} - (c_L \beta_L + \gamma_L) \bar{A}_{L+1} 1_{w_L} \in \mathbb{R}^{w_{L+1}}, \end{aligned} \tag{4.8}$$

where the symbols  $0_{p \times q} \in \mathbb{R}^{p \times q}$  and  $0_p \in \mathbb{R}^p$  again denote zero matrices and zero vectors, respectively, with the same conventions as before and where  $1_p \in \mathbb{R}^p$  denotes a vector whose entries are all one. Then it is easy to check by induction that, for all  $x \in K$ , we have

$$\begin{aligned} \bar{A}_1 x + \bar{b}_1 &= \frac{\varepsilon_1}{2 \max_{u \in K} |\bar{a}^\top u| + 1} \begin{pmatrix} \bar{a}^\top x \\ 0_{w_1-1} \end{pmatrix} + c_1 1_{w_1} \in (c_1 - \varepsilon_1, c_1 + \varepsilon_1)^{w_1}, \\ \bar{A}_i \left( \varphi_{i-1}^{\bar{A}_{i-1}, \bar{b}_{i-1}} \circ \dots \circ \varphi_1^{\bar{A}_1, \bar{b}_1} (x) \right) + \bar{b}_i &= \frac{\varepsilon_i}{2 \max_{u \in K} |\bar{a}^\top u| + 1} \begin{pmatrix} \bar{a}^\top x \\ 0_{w_{i-1}} \end{pmatrix} + c_i 1_{w_i} \\ &\in (c_i - \varepsilon_i, c_i + \varepsilon_i)^{w_i}, \quad \forall i = 2, \dots, L, \end{aligned} \tag{4.9}$$

and

$$\psi(\bar{\alpha}, x) = \left( \varphi_{L+1}^{\bar{A}_{L+1}, \bar{b}_{L+1}} \circ \dots \circ \varphi_1^{\bar{A}_1, \bar{b}_1} \right) (x) = \bar{a}^\top x + \bar{c}.$$

The parameter  $\bar{\alpha}$  thus satisfies  $\Psi(\bar{\alpha}) = z_{\bar{a},\bar{c}} \in C(K)$ . Because of the compactness of  $K$ , the openness of the sets  $(c_i - \varepsilon_i, c_i + \varepsilon_i)^{w_i}$ ,  $i = 1, \dots, L$ , the affine-linearity of  $\sigma_i$  on  $(c_i - \varepsilon_i, c_i + \varepsilon_i)$ , and the continuity of the functions  $D \times \mathbb{R}^d \ni (\alpha, x) \mapsto A_1 x + b_1 \in \mathbb{R}^{w_1}$  and  $D \times \mathbb{R}^d \ni (\alpha, x) \mapsto A_i(\varphi_{i-1}^{A_{i-1}, b_{i-1}} \circ \dots \circ \varphi_1^{A_1, b_1}(x)) + b_i \in \mathbb{R}^{w_i}$ ,  $i = 2, \dots, L$ , it follows that there exists  $r > 0$  such that all of the inclusions in (4.9) remain valid for  $x \in K$  and  $\alpha \in B_r^D(\bar{\alpha})$  and such that  $\Psi(\alpha) \in C(K)$  is affine (i.e., of the form  $z_{a,c}$ ) for all  $\alpha \in B_r^D(\bar{\alpha})$ . As  $z_{\bar{a},\bar{c}}$  is the global solution of the best approximation problem in (v), this shows that  $\bar{\alpha}$  is a local minimum of (P) that satisfies (4.7).

To show that there are many such local minima, we require some additional notation. Henceforth, with  $a_1, \dots, a_{w_1} \in \mathbb{R}^d$  we denote the row vectors in the weight matrix  $A_1$  and with  $e_1, \dots, e_{w_1} \in \mathbb{R}^{w_1}$  the standard basis vectors of  $\mathbb{R}^{w_1}$ . We further introduce the abbreviation  $\alpha'$  for the collection of all parameters of  $\psi$  that belong to the degrees of freedom  $A_{L+1}, \dots, A_2, b_L, \dots, b_1$ , and  $a_2, \dots, a_{w_1}$ . The space of all such  $\alpha'$  is denoted by  $D'$ . Note that this space has dimension  $\dim(D') = m - d - 1 > 0$ . We again endow  $D'$  with the Euclidean norm of the space  $\mathbb{R}^{m-d-1}$  that  $D'$  can be transformed into by reordering the entries in  $\alpha'$ . As before, in what follows, a bar indicates that we refer to the parameter  $\bar{\alpha} \in D$  constructed in (4.8), i.e.,  $\bar{a}_k$  refers to the  $k$ -th row of  $\bar{A}_1$ ,  $\bar{\alpha}' \in D'$  refers to  $(\bar{A}_{L+1}, \dots, \bar{A}_2, \bar{b}_L, \dots, \bar{b}_1, \bar{a}_2, \dots, \bar{a}_{w_1})$ , etc.

To construct a set  $E \subset D$  as in the assertion of the lemma, we first note that the local affine linearity of  $\sigma_i$ , the definition of  $\bar{\alpha}$ , our choice of  $r > 0$ , and the architecture of  $\psi$  imply that there exists a continuous function  $\Phi: D' \rightarrow \mathbb{R}$  which satisfies  $\Phi(\bar{\alpha}') + \bar{b}_{L+1} = \bar{c}$  and

$$\psi(\alpha, x) = \left( \prod_{i=1}^L \beta_i \right) (A_{L+1} A_L \dots A_1) x + \Phi(\alpha') + b_{L+1} \tag{4.10}$$

for all  $x \in K$  and all  $\alpha = \{(A_i, b_i)\}_{i=1}^{L+1} \in B_r^D(\bar{\alpha})$ , cf. (4.9). Define

$$\Theta: D' \rightarrow \mathbb{R}^d, \quad \Theta(\alpha') := \left( \prod_{i=1}^L \beta_i \right) \left[ \begin{array}{c} (A_{L+1} A_L \dots A_2) \begin{pmatrix} 0 \\ a_2^\top \\ \vdots \\ a_{w_1}^\top \end{pmatrix} \end{array} \right]^\top,$$

and

$$\Lambda: D' \rightarrow \mathbb{R}, \quad \Lambda(\alpha') := \left( \prod_{i=1}^L \beta_i \right) (A_{L+1} A_L \dots A_2) e_1.$$

Then (4.10) can be recast as

$$\psi(\alpha, x) = \Theta(\alpha')^\top x + \Lambda(\alpha') a_1^\top x + \Phi(\alpha') + b_{L+1}, \quad \forall x \in K, \quad \forall \alpha \in B_r^D(\bar{\alpha}). \tag{4.11}$$

Note that, again by the construction of  $\bar{\alpha}$  in (4.8), we have  $\Theta(\bar{\alpha}') = 0$ ,  $\Lambda(\bar{\alpha}') \neq 0$ , and  $\Lambda(\bar{\alpha}') \bar{a}_1 = \bar{a}$ . In particular, due to the continuity of  $\Lambda: D' \rightarrow \mathbb{R}$ , we can find  $r' > 0$  such that  $\Lambda(\alpha') \neq 0$  holds for all  $\alpha' \in B_{r'}^{D'}(\bar{\alpha}')$ . This allows us to define

$$g_1: B_{r'}^{D'}(\bar{\alpha}') \rightarrow \mathbb{R}^d, \quad g_1(\alpha') := \frac{\Lambda(\bar{\alpha}')}{\Lambda(\alpha')} \bar{a}_1 - \frac{\Theta(\alpha')}{\Lambda(\alpha')},$$

and

$$g_2: B_{r'}^{D'}(\bar{\alpha}') \rightarrow \mathbb{R}, \quad g_2(\alpha') := \bar{c} - \Phi(\alpha').$$

By construction, these functions  $g_1$  and  $g_2$  are continuous and satisfy  $g_1(\bar{\alpha}') = \bar{a}_1$ ,  $g_2(\bar{\alpha}') = \bar{b}_{L+1}$ , and

$$\Theta(\alpha')^\top x + \Lambda(\alpha')g_1(\alpha')^\top x + \Phi(\alpha') + g_2(\alpha') = \bar{a}^\top x + \bar{c} \tag{4.12}$$

for all  $\alpha' \in B_{r'}^{D'}(\bar{\alpha}')$  and all  $x \in \mathbb{R}^d$ . Again due to the continuity, this implies that, after possibly making  $r'$  smaller, we have

$$E := \left\{ \alpha \in D \mid \alpha' \in B_{r'}^{D'}(\bar{\alpha}'), a_1 = g_1(\alpha'), b_{L+1} = g_2(\alpha') \right\} \subset B_r^D(\bar{\alpha}).$$

For all elements  $\tilde{\alpha}$  of the resulting set  $E$ , it now follows from (4.11) and (4.12) that

$$\begin{aligned} \psi(\tilde{\alpha}, x) &= \Theta(\tilde{\alpha}')^\top x + \Lambda(\tilde{\alpha}')\tilde{a}_1^\top x + \Phi(\tilde{\alpha}') + \tilde{b}_{L+1} \\ &= \Theta(\tilde{\alpha}')^\top x + \Lambda(\tilde{\alpha}')g_1(\tilde{\alpha}')^\top x + \Phi(\tilde{\alpha}') + g_2(\tilde{\alpha}') = \bar{a}^\top x + \bar{c}, \quad \forall x \in K. \end{aligned}$$

Thus,  $\Psi(\tilde{\alpha}) = z_{\bar{a}, \bar{c}}$  and, due to the definitions of  $r$ ,  $\bar{a}$ , and  $\bar{c}$ ,

$$\mathcal{L}(\Psi(\tilde{\alpha}), y_T) = \mathcal{L}(z_{\bar{a}, \bar{c}}, y_T) = \min_{(a,c) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(z_{a,c}, y_T) = \min_{\alpha \in B_r^D(\bar{\alpha})} \mathcal{L}(\Psi(\alpha), y_T)$$

for all  $\tilde{\alpha} \in E \subset B_r^D(\bar{\alpha})$ . This shows that all elements of  $E$  are local minima of (P) that satisfy (4.7). Since  $E$  is, modulo reordering of the entries in  $\alpha$ , nothing else than the graph of a function defined on an open subset of  $\mathbb{R}^{m-d-1}$  with values in  $\mathbb{R}^{d+1}$ , the fact that the Hausdorff dimension of  $E$  in  $D$  is at least  $m - d - 1$  immediately follows from the choice of the norm on  $D$  and classical results, see [19, Corollary 8.2c].  $\square$

**Lemma 4.5** *Consider a nonempty compact set  $K \subset \mathbb{R}^d$  and a neural network  $\psi : D \times \mathbb{R}^d \rightarrow \mathbb{R}$  with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ , and continuous activation functions  $\sigma_i$  as in (2.1). Suppose that  $\mathcal{L} : C(K) \times C(K) \rightarrow \mathbb{R}$  and  $y_T \in C(K)$  are given such that  $\mathcal{L}$ ,  $y_T$ , and the functions  $\sigma_i$  satisfy the conditions (i) and (iv) in Theorem 3.2. Then there exists a set  $E \subset D$  of Hausdorff dimension  $\dim_{\mathcal{H}}(E) \geq m - 1$  such that all elements of  $E$  are local minima of (P) and such that*

$$\mathcal{L}(\Psi(\alpha), y_T) = \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T) \tag{4.13}$$

holds for all  $\alpha \in E$ , where  $z_c$  is defined by  $z_c(x) := c$  for all  $x \in \mathbb{R}^d$ .

**Proof** The proof of Lemma 4.5 is analogous to that of Lemma 4.4 but simpler: From (i), we obtain that there exist an index  $j \in \{1, \dots, L\}$  and numbers  $c_j \in \mathbb{R}$ ,  $\varepsilon_j > 0$ , and  $\gamma_j \in \mathbb{R}$  such that  $\sigma_j(s) = \gamma_j$  holds for all  $s \in I_j = (c_j - \varepsilon_j, c_j + \varepsilon_j)$ , and from (iv), it follows that we can find a number  $\bar{c} \in \mathbb{R}$  satisfying  $\mathcal{L}(z_c, y_T) \geq \mathcal{L}(z_{\bar{c}}, y_T)$  for all  $c \in \mathbb{R}$ . Define  $\bar{\alpha} = \{(\bar{A}_i, \bar{b}_i)\}_{i=1}^{L+1}$  to be the element of  $D$  whose weights and biases are given by

$$\begin{aligned} \bar{A}_i &:= 0 \in \mathbb{R}^{w_i \times w_{i-1}}, \quad \forall i \in \{1, \dots, L+1\}, \quad \bar{b}_i := 0 \in \mathbb{R}^{w_i}, \quad \forall i \in \{1, \dots, L\} \setminus \{j\}, \\ \bar{b}_j &:= c_j 1_{w_j} \in \mathbb{R}^{w_j}, \quad \text{and} \quad \bar{b}_{L+1} := \bar{c} \in \mathbb{R}^{w_{L+1}}, \end{aligned}$$



where  $1_p \in \mathbb{R}^p$ ,  $p \in \mathbb{N}$ , again denotes the vector whose entries are all one. For this  $\bar{\alpha}$ , it clearly holds  $\Psi(\bar{\alpha}) = z_{\bar{c}} \in C(K)$  and

$$\bar{A}_j \left( \varphi_{j-1}^{\bar{A}_{j-1}, \bar{b}_{j-1}} \circ \dots \circ \varphi_1^{\bar{A}_1, \bar{b}_1}(x) \right) + \bar{b}_j = c_j 1_{w_j} \in (c_j - \varepsilon_j, c_j + \varepsilon_j)^{w_j}$$

for all  $x \in K$ . Let us denote the collection of all parameters of  $\psi$  belonging to the degrees of freedom  $A_{L+1}, \dots, A_1$  and  $b_L, \dots, b_1$  with  $\alpha'$  and the space of all such  $\alpha'$  with  $D'$  (again endowed with the Euclidean norm of the associated space  $\mathbb{R}^{m-1}$  analogously to the proof of Lemma 4.4). Then the compactness of  $K$ , the openness of  $I_j$ , the fact that  $\sigma_j$  is constant on  $I_j$ , the definition of  $\bar{\alpha}$ , the architecture of  $\psi$ , and the continuity of the function  $D \times \mathbb{R}^d \ni (\alpha, x) \mapsto A_j(\varphi_{j-1}^{A_{j-1}, b_{j-1}} \circ \dots \circ \varphi_1^{A_1, b_1}(x)) + b_j \in \mathbb{R}^{w_j}$  imply that there exist  $r > 0$  and a continuous  $\Phi: D' \rightarrow \mathbb{R}$  such that  $\Phi(\bar{\alpha}') = 0$  holds and

$$\psi(\alpha, x) = \Phi(\alpha') + b_{L+1}, \quad \forall x \in K, \quad \forall \alpha \in B_r^D(\bar{\alpha}). \tag{4.14}$$

Define  $g: D' \rightarrow \mathbb{R}$ ,  $g(\alpha') := \bar{c} - \Phi(\alpha')$ . Then  $g$  is continuous, it holds  $g(\bar{\alpha}') = \bar{b}_{L+1}$ , and we can find a number  $r' > 0$  such that

$$E := \left\{ \alpha \in D \mid \alpha' \in B_{r'}^{D'}(\bar{\alpha}'), b_{L+1} = g(\alpha') \right\} \subset B_r^D(\bar{\alpha}).$$

For all  $\tilde{\alpha} \in E$ , it now follows from (4.14) and the definition of  $g$  that

$$\psi(\tilde{\alpha}, x) = \Phi(\tilde{\alpha}') + \tilde{b}_{L+1} = \Phi(\tilde{\alpha}') + g(\tilde{\alpha}') = \bar{c}, \quad \forall x \in K.$$

Due to the properties of  $\bar{c}$  and the definition of  $r$ , this yields

$$\mathcal{L}(\Psi(\tilde{\alpha}), y_T) = \mathcal{L}(z_{\bar{c}}, y_T) = \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T) = \min_{\alpha \in B_r^D(\bar{\alpha})} \mathcal{L}(\Psi(\alpha), y_T)$$

for all  $\tilde{\alpha} \in E \subset B_r^D(\bar{\alpha})$ . Thus, all elements of  $E$  are local minima of (P) satisfying (4.13). That  $E$  has Hausdorff dimension at least  $\dim(D) - 1$  follows completely analogously to the proof of Lemma 4.4.  $\square$

As already mentioned in the introduction, the approach that we have used in Lemmas 4.4 and 4.5 to construct the local minima in  $E$  is not new. The idea to choose biases and weights such that the network inputs only come into contact with the affine linear parts of the activation functions  $\sigma_i$  can also be found in various other contributions, e.g., [11, 20, 24, 25, 54]. The main challenge in the context of Theorems 3.1 and 3.2 is proving that the local minima in Lemmas 4.4 and 4.5 are indeed spurious for generic  $y_T$  and arbitrary  $\sigma_i, L, w_i$ , and  $\mathcal{L}$ . The following two lemmas show that this spuriousness can be established without lengthy computations and manual constructions by means of Theorem 4.2.

**Lemma 4.6** *Suppose that  $K, \psi, w_i, L, \sigma_i, y_T$ , and  $\mathcal{L}$  satisfy the assumptions of Theorem 3.1 and let  $z_{a,c} \in C(K)$  be defined as in Lemma 4.4. Then it holds*

$$\inf_{\alpha \in D} \mathcal{L}(\Psi(\alpha), y_T) < \min_{(a,c) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(z_{a,c}, y_T). \tag{4.15}$$

**Proof** We argue by contradiction. Suppose that the assumptions of Theorem 3.1 are satisfied and that (4.15) is false. Then it holds

$$\mathcal{L}(\Psi(\alpha), y_T) \geq \min_{(a,c) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(z_{a,c}, y_T) = \mathcal{L}(z_{\bar{a},\bar{c}}, y_T), \quad \forall \alpha \in D, \tag{4.16}$$

where  $(\bar{a}, \bar{c}) \in \mathbb{R}^d \times \mathbb{R}$  is the minimizer from assumption (v) of Theorem 3.1. To see that this inequality cannot be true, we consider network parameters  $\alpha = \{(A_i, b_i)\}_{i=1}^{L+1} \in D$  of the form

$$\begin{aligned} A_1 &:= \begin{pmatrix} \bar{A}_1 \\ \bar{A}_1 \end{pmatrix} \in \mathbb{R}^{w_1 \times w_0}, \\ A_i &:= \begin{pmatrix} \bar{A}_i & 0_{1 \times (w_{i-1}-1)} \\ 0_{(w_i-1) \times 1} & \bar{A}_i \end{pmatrix} \in \mathbb{R}^{w_i \times w_{i-1}}, \quad i = 2, \dots, L, \\ A_{L+1} &:= (\bar{A}_{L+1} \quad \bar{A}_{L+1}) \in \mathbb{R}^{w_{L+1} \times w_L}, \\ b_i &:= \begin{pmatrix} \bar{b}_i \\ \tilde{b}_i \end{pmatrix} \in \mathbb{R}^{w_i}, \quad i = 1, \dots, L, \quad b_{L+1} := \bar{b}_{L+1} + \tilde{b}_{L+1} \in \mathbb{R} \end{aligned} \tag{4.17}$$

with arbitrary but fixed  $\bar{A}_1 \in \mathbb{R}^{1 \times d}$ ,  $\bar{A}_i \in \mathbb{R}$ ,  $i = 2, \dots, L+1$ ,  $\bar{b}_i \in \mathbb{R}$ ,  $i = 1, \dots, L+1$ ,  $\tilde{A}_1 \in \mathbb{R}^{(w_1-1) \times d}$ ,  $\tilde{A}_i \in \mathbb{R}^{(w_i-1) \times (w_{i-1}-1)}$ ,  $i = 2, \dots, L$ ,  $\tilde{A}_{L+1} \in \mathbb{R}^{1 \times (w_L-1)}$ ,  $\tilde{b}_i \in \mathbb{R}^{w_i-1}$ ,  $i = 1, \dots, L$ , and  $\tilde{b}_{L+1} \in \mathbb{R}$ . Here,  $0_{p \times q} \in \mathbb{R}^{p \times q}$  again denotes a zero matrix. Note that such a structure of the network parameters is possible due to the assumption  $w_i \geq 2$ ,  $i = 1, \dots, L$ , in (i). Using (2.1), it is easy to check that every  $\alpha$  of the type (4.17) satisfies  $\psi(\alpha, x) = \bar{\psi}(\bar{\alpha}, x) + \tilde{\psi}(\tilde{\alpha}, x)$  for all  $x \in \mathbb{R}^d$ , where  $\bar{\psi}$  is a neural network as in (2.1) with depth  $\bar{L} = L$ , widths  $\bar{w}_i = 1$ ,  $i = 1, \dots, L$ , activation functions  $\sigma_i$ , and network parameter  $\bar{\alpha} = \{(\bar{A}_i, \bar{b}_i)\}_{i=1}^{L+1}$  and where  $\tilde{\psi}$  is a neural network as in (2.1) with depth  $\tilde{L} = L$ , widths  $\tilde{w}_i = w_i - 1$ ,  $i = 1, \dots, L$ , activation functions  $\sigma_i$ , and network parameter  $\tilde{\alpha} = \{(\tilde{A}_i, \tilde{b}_i)\}_{i=1}^{L+1}$ . In combination with (4.16), this implies

$$\mathcal{L}(\bar{\Psi}(\bar{\alpha}) + \tilde{\Psi}(\tilde{\alpha}), y_T) \geq \mathcal{L}(z_{\bar{a},\bar{c}}, y_T), \quad \forall \bar{\alpha} \in \bar{D}, \quad \forall \tilde{\alpha} \in \tilde{D}. \tag{4.18}$$

Here, we have used the symbols  $\bar{D}$  and  $\tilde{D}$  to denote the parameter spaces of  $\bar{\psi}$  and  $\tilde{\psi}$ , respectively, and the symbols  $\bar{\Psi}$  and  $\tilde{\Psi}$  to denote the functions into  $C(K)$  associated with  $\bar{\psi}$  and  $\tilde{\psi}$  defined in (2.2). Note that, by exactly the same arguments as in the proof of Lemma 4.4, we obtain that there exists  $\bar{\alpha} \in \bar{D}$  with  $\bar{\Psi}(\bar{\alpha}) = z_{\bar{a},\bar{c}}$ . Due to (4.18) and the fact that  $\tilde{A}_{L+1}$  and  $\tilde{b}_{L+1}$  can be rescaled at will, this yields

$$\mathcal{L}(z_{\bar{a},\bar{c}} + s\tilde{\Psi}(\tilde{\alpha}), y_T) \geq \mathcal{L}(z_{\bar{a},\bar{c}}, y_T), \quad \forall \tilde{\alpha} \in \tilde{D}, \quad \forall s \in (0, \infty). \tag{4.19}$$

Since  $z_{\bar{a},\bar{c}} \neq y_T$  holds by (iii) and since  $\mathcal{L}$  is Gâteaux differentiable in its first argument with a nonzero derivative  $\partial_1 \mathcal{L}(v, y_T)$  at all points  $(v, y_T) \in C(K) \times C(K)$  satisfying  $v \neq y_T$  by (iv), we can rearrange (4.19), divide by  $s > 0$ , and pass to the limit  $s \rightarrow 0^+$  (for an arbitrary but fixed  $\tilde{\alpha} \in \tilde{D}$ ) to obtain

$$\left\langle \partial_1 \mathcal{L}(z_{\bar{a}, \bar{c}}, y_T), \tilde{\Psi}(\tilde{\alpha}) \right\rangle_{C(K)} \geq 0, \quad \forall \tilde{\alpha} \in \tilde{D}, \tag{4.20}$$

with a measure  $\partial_1 \mathcal{L}(z_{\bar{a}, \bar{c}}, y_T) \in \mathcal{M}(K) \setminus \{0\}$ . From Theorem 4.2, we know that (4.20) can only be true if  $\partial_1 \mathcal{L}(z_{\bar{a}, \bar{c}}, y_T) = 0$  holds. Thus, we arrive at a contradiction, (4.16) cannot be correct, and the proof is complete.  $\square$

**Lemma 4.7** *Suppose that  $K, \psi, w_i, L, \sigma_i, y_T$ , and  $\mathcal{L}$  satisfy the assumptions of Theorem 3.2 and let  $z_c \in C(K)$  be defined as in Lemma 4.5. Then it holds*

$$\inf_{\alpha \in D} \mathcal{L}(\Psi(\alpha), y_T) < \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T). \tag{4.21}$$

**Proof** The proof of Lemma 4.7 is analogous to that of Lemma 4.6 but simpler. Suppose that (4.21) is false and that the assumptions of Theorem 3.2 are satisfied. Then it holds

$$\mathcal{L}(\Psi(\alpha), y_T) \geq \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T) = \mathcal{L}(z_{\bar{c}}, y_T), \quad \forall \alpha \in D, \tag{4.22}$$

where  $\bar{c} \in \mathbb{R}$  denotes the minimizer from point (iv) of Theorem 3.2. By exploiting that the parameter  $\alpha = \{(A_i, b_i)\}_{i=1}^{L+1} \in D$  is arbitrary, by shifting the bias  $b_{L+1}$  by  $\bar{c}$ , and by subsequently scaling  $A_{L+1}$  and  $b_{L+1}$  in (4.22), we obtain that

$$\mathcal{L}(z_{\bar{c}} + s\Psi(\alpha), y_T) \geq \mathcal{L}(z_{\bar{c}}, y_T), \quad \forall \alpha \in D, \quad \forall s \in (0, \infty). \tag{4.23}$$

In combination with assumptions (ii) and (iii) of Theorem 3.2, (4.23) yields—completely analogously to (4.20)—that there exists a measure  $\partial_1 \mathcal{L}(z_{\bar{c}}, y_T) \in \mathcal{M}(K) \setminus \{0\}$  satisfying

$$\langle \partial_1 \mathcal{L}(z_{\bar{c}}, y_T), \Psi(\alpha) \rangle_{C(K)} \geq 0, \quad \forall \alpha \in D.$$

By invoking Theorem 4.2, we now again arrive at a contradiction. Thus, (4.22) cannot be true and the assertion of the lemma follows.  $\square$

To establish Theorems 3.1 and 3.2, it suffices to combine Lemmas 4.4 and 4.6 and Lemmas 4.5 and 4.7, respectively. This completes the proof of our main results on the existence of spurious local minima in training problems of the type (P).

**Remark 4.8** • As it is irrelevant for our analysis whether the function  $\alpha \mapsto \mathcal{L}(\Psi(\alpha), y_T)$  appearing in (P) is used to train a network or to validate the generalization properties of a trained network, Theorems 3.1 and 3.2 also establish the existence of spurious local minima for the generalization error.

- We expect that it is possible to extend Theorems 3.1 and 3.2 to training problems defined on the whole of  $\mathbb{R}^d$  provided the activation functions  $\sigma_i$  and the loss function  $\mathcal{L}$  are sufficiently well-behaved. We omit a detailed discussion of this extension here since it requires nontrivial modifications of the functional analytic setting and leave this topic for future research.

- The distinction between the “nonconstant affine segment”-case and the “constant segment”-case in Theorems 3.1 and 3.2 and Lemmas 4.4 to 4.7 is necessary. This can be seen, e.g., in the formulas in (4.8) which degenerate when one of the slopes  $\beta_i$  is equal to zero.
- If a network of the type (2.1) with an additional activation function  $\sigma_{L+1}$  acting on the last layer is considered, then one can simply include  $\sigma_{L+1}$  into the loss function by defining  $\tilde{\mathcal{L}}(v, y_T) := \mathcal{L}(\sigma_{L+1} \circ v, y_T)$ . Along these lines, our results can be applied to networks with a nonaffine last layer as well (provided the function  $\tilde{\mathcal{L}}$  still satisfies the assumptions of Theorems 3.1 and 3.2).

## 5 Further Consequences of the Nonexistence of Supporting Half-Spaces

The aim of this section is to point out some further consequences of Theorem 4.2. Our main focus will be on the implications that this theorem has for the well-posedness properties of best approximation problems for neural networks in function space. We begin by noting that the nonexistence of supporting half-spaces for the image  $\Psi(D)$  in (4.2) implies that the closure of  $\Psi(D)$  can only be convex if it is equal to the whole of  $C(K)$ . More precisely, we have the following result:

**Corollary 5.1** *Let  $K \subset \mathbb{R}^d$ ,  $d \in \mathbb{N}$ , be a nonempty and compact set and let  $\psi: D \times \mathbb{R}^d \rightarrow \mathbb{R}$  be a neural network as in (2.1) with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ , and continuous nonpolynomial activation functions  $\sigma_i$ . Suppose that  $(Z, \|\cdot\|_Z)$  is a real normed space and that  $\iota: C(K) \rightarrow Z$  is a linear and continuous map with a dense image. Then the set  $\text{cl}_Z(\iota(\Psi(D)))$  is either nonconvex or equal to  $Z$ .*

**Proof** Assume that  $\text{cl}_Z(\iota(\Psi(D)))$  is convex and that  $\text{cl}_Z(\iota(\Psi(D))) \neq Z$ . Then there exists  $z \in Z \setminus \text{cl}_Z(\iota(\Psi(D)))$  and it follows from the separation theorem for convex sets in normed spaces [22, Corollary I–1.2] that we can find  $v \in Z^* \setminus \{0\}$  and  $c \in \mathbb{R}$  such that

$$\langle v, \iota(\Psi(\alpha)) \rangle_Z = \langle \iota^*(v), \Psi(\alpha) \rangle_{C(K)} \leq c, \quad \forall \alpha \in D. \quad (5.1)$$

Here,  $Z^*$  denotes the topological dual of  $Z$ ,  $\langle \cdot, \cdot \rangle_Z: Z^* \times Z \rightarrow \mathbb{R}$  denotes the dual pairing in  $Z$ , and  $\iota^*: Z^* \rightarrow C(K)^* = \mathcal{M}(K)$  denotes the adjoint of  $\iota$  as defined in [13, Section 9]. Due to Theorem 4.2, (5.1) is only possible if  $\iota^*(v) = 0$ , i.e., if

$$\langle \iota^*(v), v \rangle_{C(K)} = \langle v, \iota(v) \rangle_Z = 0, \quad \forall v \in C(K).$$

As  $\iota(C(K))$  is dense in  $Z$ , this yields  $v = 0$  which is a contradiction. Thus, the set  $\text{cl}_Z(\iota(\Psi(D)))$  is either nonconvex or equal to  $Z$  and the proof is complete.  $\square$

We remark that, for activation functions possessing a point of differentiability with a nonzero derivative, a version of Corollary 5.1 has already been proven in [40, Lemma C.9]. By using Theorem 4.2 and the separation theorem, we can avoid the assumption that such a point of differentiability exists and obtain Corollary 5.1 for all nonpolynomial continuous activations  $\sigma_i$ . In combination with classical results on

the properties of Chebychev sets, see [50], the nonconvexity of the set  $\text{cl}_Z(\iota(\Psi(D)))$  in Corollary 5.1 immediately implies that the problem of determining a best approximating element for a given  $u \in Z$  from the set  $\text{cl}_Z(\iota(\Psi(D)))$  of all elements of  $Z$  that can be approximated by points of the form  $\iota(\Psi(\alpha))$  is always ill-posed in the sense of Hadamard if  $Z$  is a strictly convex Banach space with a strictly convex dual and  $\iota(\Psi(D))$  is not dense.

**Corollary 5.2** *Let  $K, \psi, L, w_i, \sigma_i, (Z, \|\cdot\|_Z)$ , and  $\iota$  be as in Corollary 5.1. Assume additionally that  $(Z, \|\cdot\|_Z)$  is a Banach space and that  $(Z, \|\cdot\|_Z)$  and its topological dual  $(Z^*, \|\cdot\|_{Z^*})$  are strictly convex. Define  $\Pi$  to be the best approximation map associated with the set  $\text{cl}_Z(\iota(\Psi(D)))$ , i.e., the set-valued projection operator*

$$\Pi : Z \rightrightarrows Z, \quad u \mapsto \arg \min_{z \in \text{cl}_Z(\iota(\Psi(D)))} \|u - z\|_Z. \tag{5.2}$$

Then exactly one of the following is true:

- (i)  $\text{cl}_Z(\iota(\Psi(D)))$  is equal to  $Z$  and  $\Pi$  is the identity map.
- (ii) There does not exist a function  $\pi : Z \rightarrow Z$  such that  $\pi(z) \in \Pi(z)$  holds for all  $z \in Z$  and such that  $\pi$  is continuous in an open neighborhood of the origin.

**Proof** This immediately follows from Corollary 5.1, [28, Theorem 3.5], and the fact that the set  $\text{cl}_Z(\iota(\Psi(D)))$  is a cone. □

Note that there are two possible reasons for the nonexistence of a selection  $\pi$  with the properties in point (ii) of Corollary 5.2. The first one is that there exists an element  $u \in Z$  for which the set  $\Pi(u)$  is empty, i.e., for which the best approximation problem associated with the right-hand side of (5.2) does not possess a solution. The second one is that  $\Pi(u) \neq \emptyset$  holds for all  $u \in Z$  but that every selection  $\pi$  taken from  $\Pi$  is discontinuous at some point  $u$ , i.e., that there exists an element  $u \in Z$  for which the solution set of the best approximation problem associated with the right-hand side of (5.2) is unstable w.r.t. small perturbations of the problem data. In both of these cases, one of the conditions for Hadamard well-posedness is violated, see [27, Section 2.1], so that Corollary 5.2 indeed implies that the problem of determining best approximations is ill-posed when  $\text{cl}_Z(\iota(\Psi(D))) \neq Z$  holds.

To make Corollary 5.2 more tangible, we state its consequences for best approximation problems posed in reflexive Lebesgue spaces, cf. Lemma 3.3. Such problems arise when  $Z$  is equal to  $L^p_\mu(K)$  for some  $\mu \in \mathcal{M}_+(K)$  and  $p \in (1, \infty)$  and when  $\iota : C(K) \rightarrow L^p_\mu(K)$  is the inclusion map. In the statement of the next corollary, we drop the inclusion map  $\iota$  in the notation for the sake of readability.

**Corollary 5.3** *Suppose that  $K \subset \mathbb{R}^d, d \in \mathbb{N}$ , is a nonempty compact set and that  $\psi$  is a neural network as in (2.1) with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ , and continuous nonpolynomial activation functions  $\sigma_i$ . Assume that  $\mu \in \mathcal{M}_+(K)$  and  $p \in (1, \infty)$  are given and that the image  $\Psi(D)$  of the map  $\Psi : D \rightarrow C(K)$  in (2.2) is not dense in  $L^p_\mu(K)$ . Then there does not exist a function  $\pi : L^p_\mu(K) \rightarrow L^p_\mu(K)$  such that*

$$\pi(u) \in \arg \min_{z \in \text{cl}_{L^p_\mu(K)}(\Psi(D))} \|u - z\|_{L^p_\mu(K)}, \quad \forall u \in L^p_\mu(K), \tag{5.3}$$

holds and such that  $\pi$  is continuous in an open neighborhood of the origin.

**Proof** From [36, Example 1.10.2, Theorem 5.2.11], it follows that  $L_\mu^p(K)$  is uniformly convex with a uniformly convex dual, and from [23, Proposition 7.9], we obtain that the inclusion map  $\iota: C(K) \rightarrow L_\mu^p(K)$  is linear and continuous with a dense image. The claim thus immediately follows from Corollary 5.2.  $\square$

As already mentioned in Sect. 1, for neural networks with a single hidden layer, a variant of Corollary 5.3 has also been proven in [28, Section 4]. For related results, see also [29, 40]. We obtain the discontinuity of  $L_\mu^p(K)$ -best approximation operators for networks of arbitrary depth here as a consequence of Theorem 4.2 and thus, at the end of the day, as a corollary of the universal approximation theorem. This again highlights the connections that exist between the approximation capabilities of neural networks and the landscape/well-posedness properties of the optimization problems that have to be solved in order to determine neural network best approximations.

We remark that, to get an intuition for the geometric properties of the image  $\Psi(D) \subset C(K)$  that are responsible for the effects in Theorem 3.1, Theorem 3.2, and Corollary 5.3, one can indeed plot this set in simple situations. Consider, for example, the case  $d = 1$ ,  $K = \{-1, 0, 2\}$ ,  $\mu = \delta_{-1} + \delta_0 + \delta_2$ ,  $L = 1$ ,  $w_1 = 1$ , and  $p = 2$ , where  $\delta_x$  again denotes a Dirac measure supported at  $x \in \mathbb{R}$ . For these  $K$  and  $\mu$ , we have  $C(K) \cong L_\mu^2(K) \cong \mathbb{R}^3$  and the image  $\Psi(D) \subset C(K)$  of the map  $\Psi$  in (2.2) can be identified with a subset of  $\mathbb{R}^3$ , namely,

$$\Psi(D) = \left\{ z \in \mathbb{R}^3 \mid z = (\psi(\alpha, -1), \psi(\alpha, 0), \psi(\alpha, 2))^T \text{ for some } \alpha \in D \right\}.$$

Further, the best approximation problem associated with the right-hand side of (5.3) simply becomes the problem of determining the set-valued Euclidean projection of a point  $u \in \mathbb{R}^3$  onto  $\text{cl}_{\mathbb{R}^3}(\Psi(D))$ , i.e.,

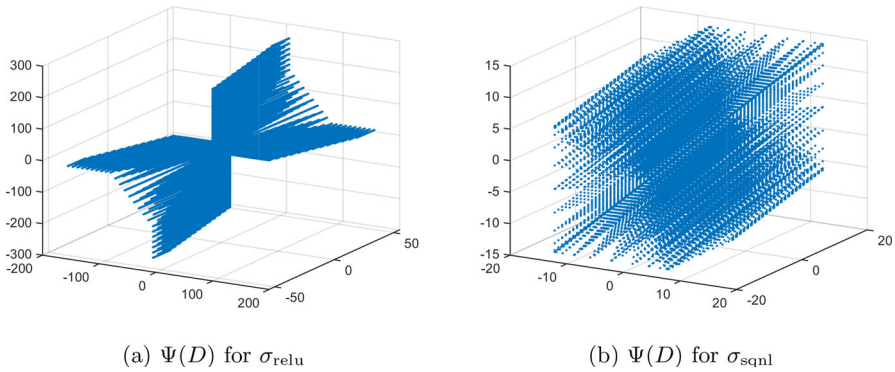
$$\text{Minimize } |u - z| \quad \text{w.r.t. } z \in \text{cl}_{\mathbb{R}^3}(\Psi(D)). \quad (5.4)$$

This makes it possible to visualize the image  $\Psi(D)$  and to interpret the  $L_\mu^2(K)$ -best approximation operator associated with  $\psi$  geometrically. The sets  $\Psi(D)$  that are obtained in the above situation for the ReLU-activation  $\sigma_{\text{relu}}(s) := \max(0, s)$  and the SQNL-activation

$$\sigma_{\text{sqnl}}(s) := \begin{cases} -1 & \text{if } s \leq -2 \\ s + s^2/4 & \text{if } -2 < s \leq 0 \\ s - s^2/4 & \text{if } 0 < s \leq 2 \\ 1 & \text{if } s > 2 \end{cases}$$

can be seen in Fig. 1. Note that, since both of these functions are monotonically increasing, the assumption  $L = w_1 = 1$  and the architecture in (2.1) imply that  $(0, 1, 0)^T \notin \text{cl}_{\mathbb{R}^3}(\Psi(D))$  holds. This shows that, for both the ReLU- and the SQNL-activation, the resulting network falls under the scope of Corollary 5.3. Since  $\sigma_{\text{relu}}$  and  $\sigma_{\text{sqnl}}$  possess constant segments, the training problems

$$\text{Minimize } |u - \Psi(\alpha)| \quad \text{w.r.t. } \alpha \in D \quad (5.5)$$



**Fig. 1** Scatter plot of the image  $\Psi(D)$  of the function  $\Psi : D \rightarrow C(K) \cong L^2_\mu(K) \cong \mathbb{R}^3$  in the case  $d = 1$ ,  $K = \{-1, 0, 2\}$ ,  $\mu = \delta_{-1} + \delta_0 + \delta_2$ ,  $L = 1$ , and  $w_1 = 1$  for the ReLU- and the SQNL-activation function. For the weights  $A_1, A_2 \in \mathbb{R}$ , we used samples from the interval  $[-10, 10]$ , and for the biases  $b_1, b_2 \in \mathbb{R}$ , from the interval  $[-5, 5]$ . Solving the problems (5.4) or (5.5) for a given  $u$  corresponds to calculating the set-valued Euclidean projection of  $u$  onto these sets

associated with these activation functions are moreover covered by Theorem 3.2, cf. Lemma 3.3. As Fig. 1 shows, the sets  $\Psi(D)$  obtained for  $\sigma_{\text{relu}}$  and  $\sigma_{\text{sqnl}}$  along the above lines are highly nonconvex and locally resemble two-dimensional subspaces of  $\mathbb{R}^3$  at many points. Because of these properties, it is only natural that the resulting  $L^2_\mu(K)$ -best approximation operators, i.e., the Euclidean projections onto  $\text{cl}_{\mathbb{R}^3}(\Psi(D))$ , possess discontinuities and give rise to training problems that contain various spurious local minima. We remark that the examples in Fig. 1 improve a construction in [11, Section 4], where a similar visualization for a more academic network was considered. We are able to overcome the restrictions of [11] here due to Theorems 4.1 and 4.2. Note that depicting the image  $\Psi(D)$  of a neural network along the lines of Fig. 1 only works well for very small architectures. Larger networks are too expressive to be properly visualized in three dimensions.

We would like to point out that the “space-filling” cases  $\text{cl}_Z(\iota(\Psi(D))) = Z$  and  $\text{cl}_{L^p_\mu(K)}(\Psi(D)) = L^p_\mu(K)$  in Corollaries 5.1 to 5.3 are not as pathological as one might think at first glance. In fact, in many applications, neural networks are trained in an “overparameterized” regime in which the number of degrees of freedom in  $\psi$  exceeds the number of training samples by far and in which  $\psi$  is able to fit arbitrary training data with zero error, see [2, 9, 15, 33, 39, 44]. In the situation of Lemma 3.3, this means that a measure  $\mu$  of the form  $\mu = \frac{1}{n} \sum_{k=1}^n \delta_{x_k}$  supported on a finite set  $K = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ ,  $n \in \mathbb{N}$ , is considered which satisfies  $n \ll m = \dim(D)$ . The absence of the ill-posedness effects in Corollary 5.3 is a possible explanation for the observation that overparameterized neural networks are far easier to train than their non-overparameterized counterparts, cf. [2, 33, 39, 44]. We remark that, for overparameterized finite-dimensional training problems, numerically determining a minimizer is also simplified by the fact that, in the case  $m = \dim(D) \gg n$ , the Jacobian of the map  $D \ni \alpha \mapsto \Psi(\alpha) \in C(K) \cong \mathbb{R}^n$  typically has full rank on a large subset of the parameter space  $D$ , cf. [32, 37]. Note that, for infinite  $K$ , there is no analogue to this effect since the Gâteaux derivative of the map  $D \ni \alpha \mapsto \Psi(\alpha) \in C(K)$  can

never be surjective if  $C(K)$  is infinite-dimensional. Theorems 3.1 and 3.2 further show that the Jacobian of the mapping  $D \ni \alpha \mapsto \Psi(\alpha) \in C(K) \cong \mathbb{R}^n$  can indeed only be expected to have full rank on a large set (e.g., a.e.) when an overparameterized finite-dimensional training problem is considered, but not everywhere.

Even though there is no sensible notion of “overparameterization” in the infinite-dimensional setting, it is still possible for a neural network to satisfy the conditions  $\text{cl}_Z(\iota(\Psi(D))) = Z$  and  $\text{cl}_{L_\mu^p(K)}(\Psi(D)) = L_\mu^p(K)$  in Corollaries 5.1 to 5.3 for a non-finite training set  $K$ . In fact, in the case  $d = 1$ , it can be shown that the set of activation functions that give rise to a “space-filling” network is dense in  $C(\mathbb{R})$  in the topology of uniform convergence on compacta. There thus indeed exist *many* choices of activation functions  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  for which the density conditions  $\text{cl}_Z(\iota(\Psi(D))) = Z$  and  $\text{cl}_{L_\mu^p(K)}(\Psi(D)) = L_\mu^p(K)$  in Corollaries 5.1 to 5.3 hold for arbitrary spaces  $Z$  and arbitrary measures  $\mu$ . To be more precise, we have:

**Lemma 5.4** *Consider a nonempty compact set  $K \subset \mathbb{R}$  and a neural network  $\psi$  as in (2.1) with depth  $L \in \mathbb{N}$ , widths  $w_i \in \mathbb{N}$ , and  $d = 1$ . Suppose that  $\sigma_i = \sigma$  holds for all  $i = 1, \dots, L$  with a function  $\sigma \in C(\mathbb{R})$ . Then, for all  $\varepsilon > 0$  and all nonempty open intervals  $I \subset \mathbb{R}$ , there exists a function  $\tilde{\sigma} \in C(\mathbb{R})$  such that  $\sigma \equiv \tilde{\sigma}$  holds in  $\mathbb{R} \setminus I$ , such that  $|\sigma(s) - \tilde{\sigma}(s)| < \varepsilon$  holds for all  $s \in \mathbb{R}$ , and such that the neural network  $\tilde{\psi}$  obtained by replacing  $\sigma$  with  $\tilde{\sigma}$  in  $\psi$  satisfies  $\text{cl}_{C(K)}(\tilde{\Psi}(D)) = C(K)$ .*

**Proof** The lemma is an easy consequence of the separability of  $(C(K), \|\cdot\|_{C(K)})$ , cf. [35]. Since we can replace  $K$  by a closed bounded interval that contains  $K$  to prove the claim, since we can rescale and translate the argument  $x$  of  $\psi$  by means of  $A_1$  and  $b_1$ , and since we can again consider parameters of the form (4.4), we may assume w.l.o.g. that  $K = [0, 1]$  holds and that all layers of  $\psi$  have width one. Suppose that a number  $\varepsilon > 0$  and a nonempty open interval  $I$  are given. Using the continuity of  $\sigma$ , it is easy to check that there exists a function  $\bar{\sigma} \in C(\mathbb{R})$  that satisfies  $\sigma \equiv \bar{\sigma}$  in  $\mathbb{R} \setminus I$ ,  $|\sigma(s) - \bar{\sigma}(s)| < \varepsilon/2$  for all  $s \in \mathbb{R}$ , and  $\bar{\sigma} = \text{const}$  in  $(a, a + \eta)$  for some  $a \in \mathbb{R}$  and  $\eta > 0$  with  $(a, a + \eta) \subset I$ . Let  $\{p_k\}_{k=1}^\infty \subset C([0, 1])$  denote the countable collection of all polynomials on  $[0, 1]$  that have rational coefficients and that are not identical zero, starting with  $p_1(x) = x$ , and let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be the unique element of  $C(\mathbb{R})$  with the following properties:

- (i)  $\phi \equiv 0$  in  $\mathbb{R} \setminus (a, a + \eta)$ ,
- (ii)  $\phi$  is affine on  $[a + \eta(1 - 2^{-2k+1}), a + \eta(1 - 2^{-2k})]$  for all  $k \in \mathbb{N}$ ,
- (iii)  $\phi(a + \eta(1 - 2^{-2k+2}) + \eta 2^{-2k+1}x) = p_k(x)\varepsilon/(2k\|p_k\|_{C([0,1])})$  for all  $x \in [0, 1]$  and all  $k \in \mathbb{N}$ .

We define  $\tilde{\sigma} := \bar{\sigma} + \phi$ . Note that, for this choice of  $\tilde{\sigma}$ , we clearly have  $\tilde{\sigma} \in C(\mathbb{R})$ ,  $\sigma \equiv \tilde{\sigma}$  in  $\mathbb{R} \setminus I$ , and  $|\sigma(s) - \tilde{\sigma}(s)| < \varepsilon$  for all  $s \in \mathbb{R}$ . It remains to show that the neural network  $\tilde{\psi}$  associated with  $\tilde{\sigma}$  satisfies  $\text{cl}_{C([0,1])}(\tilde{\Psi}(D)) = C([0, 1])$ . To prove this, we observe that, due to the choice of  $p_1$  and the properties of  $\bar{\sigma}$  and  $\phi$ , we have

$$\frac{2}{\varepsilon} \tilde{\sigma} \left( a + \frac{1}{2} \eta x \right) - \frac{2}{\varepsilon} \tilde{\sigma}(a) = p_1(x) = x, \quad \forall x \in [0, 1]. \quad (5.6)$$



This equation allows us to turn the functions  $\varphi_i^{A_i, b_i} : \mathbb{R} \rightarrow \mathbb{R}, i = 1, \dots, L - 1$ , into identity maps on  $[0, 1]$  by choosing the weights and biases appropriately and to consider w.l.o.g. the case  $L = 1$ , cf. the proof of Theorem 4.2. For this one-hidden-layer case, we obtain analogously to (5.6) that

$$\frac{2k \|p_k\|_{C([0,1])}}{\varepsilon} \tilde{\sigma} \left( a + \eta(1 - 2^{-2k+2}) + \eta 2^{-2k+1} x \right) - \frac{2k \|p_k\|_{C([0,1])}}{\varepsilon} \tilde{\sigma}(a) = p_k(x)$$

holds for all  $x \in [0, 1]$  and all  $k \in \mathbb{N}$ . For every  $k \in \mathbb{N}$ , there thus exists a parameter  $\alpha_k \in D$  satisfying  $\tilde{\Psi}(\alpha_k) = p_k \in C([0, 1])$ . Since  $\{p_k\}_{k=1}^\infty$  is dense in  $C([0, 1])$  by the Weierstrass approximation theorem, the identity  $\text{cl}_{C([0,1])}(\tilde{\Psi}(D)) = C([0, 1])$  now follows immediately. This completes the proof.  $\square$

Under suitable assumptions on the depth and the widths of  $\psi$ , Lemma 5.4 can also be extended to the case  $d > 1$ , cf. [35, Theorem 4]. For some criteria ensuring that the image of  $\Psi$  is not dense, see [40, Appendix C3]. We conclude this paper with some additional remarks on Theorems 3.1 and 3.2 and Corollaries 5.2 and 5.3:

**Remark 5.5** • As the proofs of Theorems 3.1 and 3.2 are constructive, they can be used to calculate explicit examples of spurious local minima for training problems of the type (P). To do so in the situation of Theorem 3.1, for example, one just has to calculate the slope  $\bar{a} \in \mathbb{R}^d$  and the offset  $\bar{c} \in \mathbb{R}$  of the affine linear best approximation of the target function  $y_T \in C(K)$  w.r.t.  $\mathcal{L}$  and then plug the resulting values into the formulas in (4.8). The resulting network parameter  $\bar{\alpha} = \{(\bar{A}_i, \bar{b}_i)\}_{i=1}^{L+1} \in D$  then yields an element of the set  $E$  of spurious local minima of (P) in Theorem 3.1 as desired. We remark that, along these lines, one can also easily construct “bad” choices of starting values for gradient-based training algorithms that cause the training process to terminate with a suboptimal point (as any reasonable gradient-based algorithm stalls when initialized directly in or near a local minimum). We omit including a numerical test of this type here since such experiments have been conducted in various previous works. We exemplarily mention the numerical investigations in [24, Section 2], in which a deep multilayer perceptron model is trained on CIFAR-10 by means of the logistic loss and in which the authors provoke gradient-based algorithms to fail by choosing an initialization near a spurious minimum of the type discussed above (albeit without the knowledge that this local minimum is indeed always spurious); the numerical experiments in [43] on the appearance, impact, and role of spurious minima in ReLU-networks; and the numerical tests in [48, Sections 3, 4], which are concerned with training problems for shallow networks on the XOR dataset.

- In contrast to the proofs of Theorems 3.1 and 3.2, the proofs of Corollaries 5.2 and 5.3 are *not* constructive. This significantly complicates finding explicit examples of points  $u \in L_\mu^p(K)$  at which the function  $\pi$  in Corollary 5.3 is necessarily discontinuous and at which the ill-posedness effects documented in Corollaries 5.2 and 5.3 become apparent—in particular as these points  $u$  can be expected to occupy a comparatively small set, cf. [51]. At least to our best knowledge, the construction of explicit data sets and test configurations which provably illustrate the effects in Corollaries 5.2 and 5.3 has not been accomplished so far in the literature (although

numerical experiments on instability effects are rather common, cf. [16]). We remark that, in [11], it has been shown that training data vectors, which give rise to ill-posedness effects, can be calculated for finite-dimensional squared loss training problems by solving a certain max-min-optimization problem, see [11, Lemma 12 and proof of Theorem 15]. This implicit characterization might provide a way for determining “worst case”-training data sets for problems of the type (P). We leave this topic for future research.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Ainsworth, M., Shin, Y.: Plateau phenomenon in gradient descent training of RELU networks: explanation, quantification, and avoidance. *SIAM J. Sci. Comput.* **43**, 3438–3468 (2021)
2. Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 242–252, PMLR (2019)
3. Arjevani, Y., Field, M.: Analytic study of families of spurious minima in two-layer ReLU neural networks: a tale of symmetry II. In: *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc. (2021)
4. Auer, P., Herbster, M., Warmuth, M.K.: Exponentially many local minima for single neurons. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) *Advances in Neural Information Processing Systems*, vol. 8, pp. 316–322. Curran Associates, Inc. (1996)
5. Benedetto, J.J., Czaja, W.: *Integration and Modern Analysis*. Birkhäuser Advanced Texts. Birkhäuser, Boston (2010)
6. Berner, J., Grohs, P., Kutyniok, G., Petersen, P.: The modern mathematics of deep learning. [arXiv:2105.04026v1](https://arxiv.org/abs/2105.04026v1) (2021)
7. Blum, A.L., Rivest, R.L.: Training a 3-node neural network is NP-complete. *Neural Netw.* **5**, 117–127 (1992)
8. Bonnans, J.F., Shapiro, A.: *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York (2000)
9. Chen, Z., Cao, Y., Zou, D., Gu, Q.: How much over-parameterization is sufficient to learn deep ReLU networks? [arXiv:1911.12360v3](https://arxiv.org/abs/1911.12360v3) (2020)
10. Cheridito, P., Jentzen, A., Rossmannek, F.: Landscape analysis for shallow neural networks: complete classification of critical points for affine target functions. [arXiv:2103.10922v2](https://arxiv.org/abs/2103.10922v2) (2021)
11. Christof, C.: On the stability properties and the optimization landscape of training problems with squared loss for neural networks and general nonlinear conic approximation schemes. *J. Mach. Learn. Res.* **22**, 1–77 (2021)
12. Christof, C., Hafemeyer, D.: On the nonuniqueness and instability of solutions of tracking-type optimal control problems. *Math. Control Relat. Fields* **12**, 421–431 (2022)
13. Clason, C.: *Introduction to Functional Analysis, Compact Textbooks in Mathematics*. Birkhäuser, Cham (2020)
14. Cohen, A., DeVore, R., Petrova, G., Wojtaszczyk, P.: Optimal stable nonlinear approximation. *Found. Comput. Math.* (2021)

15. Cooper, Y.: The critical locus of overparameterized neural networks. [arXiv:2005.04210v2](https://arxiv.org/abs/2005.04210v2) (2020)
16. Cunningham, P., Carney, J., Jacob, S.: Stability problems with artificial neural networks and the ensemble solution. *Art. Intell. Med.* **20**, 217–225 (2000)
17. Cybenko, G.: Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**, 303–314 (1989)
18. Dauphin, Y., Pascanu, R., Gülçehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 27, pp. 2933–2941. Curran Associates, Inc. (2014)
19. DiBenedetto, E.: *Real Analysis*, Birkhäuser Advanced Texts, 2nd edn. Birkhäuser, Basel (2016)
20. Ding, T., Li, D., Sun, R.: Sub-optimal local minima exist for almost all over-parameterized neural networks. [arXiv:1911.01413v3](https://arxiv.org/abs/1911.01413v3) (2020)
21. Eftekhari, A.: Training linear neural networks: non-local convergence and complexity results. In: Daumé, H., Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, pp. 2836–2847, PMLR (2020)
22. Ekeland, I., Temam, R.: *Convex Analysis and Variational Problems*. North-Holland, Amsterdam (1976)
23. Folland, G.B.: *Real Analysis: Modern Techniques and Their Applications*, Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts, 2nd edn. Wiley, New York (1999)
24. Goldblum, M., Geiping, J., Schwarzschild, A., Moeller, M., Goldstein, T.: Truth or backpropaganda? An empirical investigation of deep learning theory. [arXiv:1910.00359v3](https://arxiv.org/abs/1910.00359v3) (2020)
25. He, F., Wang, B., Tao, D.: Piecewise linear activations substantially shape the loss surfaces of neural networks. [arXiv:2003.12236v1](https://arxiv.org/abs/2003.12236v1) (2020)
26. Hornik, K.: Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991)
27. Kabanikhin, S.I.: *Inverse and Ill-posed Problems: Theory and Applications*. De Gruyter, Berlin (2012)
28. Kainen, P.C., Kurková, V., Vogt, A.: Approximation by neural networks is not continuous. *Neurocomputing* **29**, 47–56 (1999)
29. Kainen, P.C., Kurková, V., Vogt, A.: Continuity of approximation by neural networks in  $L^p$ -spaces. *Ann. Oper. Res.* **101**, 143–147 (2001)
30. Kawaguchi, K.: Deep learning without poor local minima. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29, pp. 586–594. Curran Associates, Inc. (2016)
31. Laurent, T., von Brecht, J.: Deep linear neural networks with arbitrary loss: all local minima are global. In: Dy, J.G., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 2908–2913, PMLR (2018)
32. Li, D., Ding, T., Sun, R.: On the benefit of width for neural networks: disappearance of basins. *SIAM J. Optim.* **32**, 1728–1758 (2022)
33. Li, Y., Liang, Y.: Learning overparameterized neural networks via stochastic gradient descent on structured data. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, pp. 8168–8177. Curran Associates Inc. (2018)
34. Liu, B.: Spurious local minima are common for deep neural networks with piecewise linear activations. [arXiv:2102.13233v1](https://arxiv.org/abs/2102.13233v1) (2021)
35. Maiorov, V., Pinkus, A.: Lower bounds for approximation by MLP neural networks. *Neurocomputing* **25**, 81–91 (1999)
36. Megginson, R.E.: *An Introduction to Banach Space Theory*, No. 183 in Graduate Texts in Mathematics. Springer, New York (1998)
37. Nguyen, Q., Mukkamala, M.C., Hein, M.: On the loss landscape of a class of deep neural networks with no bad local valleys. [arXiv:1809.10749v2](https://arxiv.org/abs/1809.10749v2) (2018)
38. Nicolae, A.: PLU: the piecewise linear unit activation function. [arXiv:1809.09534](https://arxiv.org/abs/1809.09534) (2018)
39. Oymak, S., Soltanolkotabi, M.: Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE J. Sel. Areas Inform. Theory* **1**, 84–105 (2020)
40. Petersen, P., Raslan, M., Voigtlaender, F.: Topological properties of the set of functions generated by neural networks of fixed size. *Found. Comput. Math* **21**, 375–444 (2021)
41. Petzka, H., Sminchisescu, C.: Non-attracting regions of local minima in deep and wide neural networks. *J. Mach. Learn. Res.* **22**, 1–34 (2021)
42. Pinkus, A.: Approximation theory of the MLP model in neural networks. *Acta Numer.* **8**, 143–195 (1999)

43. Safran, I., Shamir, O.: Spurious local minima are common in two-layer ReLU neural networks. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, vol. 80, pp. 4430–4438, PMLR (2018)
44. Safran, I., Yehudai, G., Shamir, O.: The effects of mild over-parameterization on the optimization landscape of shallow ReLU neural networks. In: Belkin, M., Kpotufe, S. (eds.) Proceedings of 34th Conference on Learning Theory, vol. 134, pp. 3889–3934, PMLR (2021)
45. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. [arXiv:1312.6120v3](https://arxiv.org/abs/1312.6120v3) (2014)
46. Sun, R.: Optimization for deep learning: theory and algorithms. [arXiv:1912.08957v1](https://arxiv.org/abs/1912.08957v1) (2019)
47. Sun, R., Li, D., Liang, S., Ding, T., Srikant, R.: The global landscape of neural networks: an overview. *IEEE Signal Process. Mag.* **37**, 95–108 (2020)
48. Swirszcz, G., Czarnecki, W.M., Pascanu, R.: Local minima in training of neural networks. [arXiv:1611.06310v2](https://arxiv.org/abs/1611.06310v2) (2016)
49. Venturi, L., Bandeira, A.S., Bruna, J.: Spurious valleys in one-hidden-layer neural network optimization landscapes. *J. Mach. Learn. Res.* **20**, 1–34 (2019)
50. Vlasov, L.P.: Almost convex and Chebyshev sets. *Math. Notes Acad. Sci. USSR* **8**, 776–779 (1970)
51. Westphal, U., Frerking, J.: On a property of metric projections onto closed subsets of Hilbert spaces. *Proc. Amer. Math. Soc.* **105**, 644–651 (1989)
52. Yoshida, Y., Okada, M.: Data-dependence of plateau phenomenon in learning with neural network—statistical mechanical analysis. *Adv. Neural Inform. Proc. Syst.* **32** (2019)
53. Yu, X.-H., Chen, G.-A.: On the local minima free condition of backpropagation learning. *IEEE Trans. Neural Netw.* **6**, 1300–1303 (1995)
54. Yun, C., Sra, S., Jadbabaie, A.: Small nonlinearities in activation functions create bad local minima in neural networks. [arXiv:1802.03487v4](https://arxiv.org/abs/1802.03487v4) (2019)
55. Zou, D., Long, P.M., Gu, Q.: On the global convergence of training deep linear ResNets. [arXiv:2003.01094v1](https://arxiv.org/abs/2003.01094v1) (2020)
56. Šíma, J.: Training a single sigmoidal neuron is hard. *Neural Comput.* **14**, 2709–2728 (2002)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.