ON THE OMNIPRESENCE OF SPURIOUS LOCAL MINIMA IN CERTAIN NEURAL NETWORK TRAINING PROBLEMS

CONSTANTIN CHRISTOF* AND JULIA KOWALCZYK*

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. The paper concludes with a numerical experiment which demonstrates that spurious local minima can indeed affect the convergence behavior of gradient-based solution algorithms in practice.

Key words. deep artificial neural network, spurious local minimum, training problem, loss landscape, Hadamard well-posedness, best approximation, stability analysis, local affine linearity

AMS subject classifications. 68T07, 49K40, 52A30, 90C31

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 [3, 6, 34] on this topic, with the contributions on stationary points and plateau phenomena in [1, 9, 15, 17, 50], with the results on suboptimal local minima and valleys in [11, 19, 24, 37, 41, 48, 52], and with the overview articles [5, 45, 46]. 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 [20, 31, 44, 52, 53]. Based on the findings of [30], it was hoped 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. [20, 44, 51]. 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, 47, 52]. On the analytical level, one of the first general negative results on the landscape properties of training problems for neural networks was proved by Yun et al. in [52, 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

*Technical University of Munich, Chair of Optimal Control, Center for Mathematical Sciences, Boltzmannstraße 3, 85748 Garching, Germany, christof@ma.tum.de, julia.kowalczyk@ma.tum.de

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 fit precisely with an affine function. This existence result was later also generalized in [24, Theorem 1] and [33, Theorem 1] to finite-dimensional training problems with arbitrary loss for deep networks with piecewise affine activation functions, in [19, 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, 19, 41, 47, 52].

The purpose of the present paper is to prove that the results of [52] 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, 19, 24, 52] can be relaxed significantly. More precisely, we show that [52, Theorem 1] can be extended straightforwardly 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 [52, 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 [24, 33, 52] 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 (SQNL, 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 a 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 4.3 and 4.4, and this construction has already been used in various papers on the landscape properties of training problems, e.g., [11, 19, 23, 24, 52]. What is typically considered as 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 [52, section 2.2], [11, section 1.4], and [24, 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 [52, Theorem 1], [33, Theorem 1], and [24, 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 [52, Theorem 1], [24,

Theorem 1], [11, Corollary 47], [33, Theorem 1], and [19, 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 proved by Cybenko, Hornik, and Pinkus in [16, 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 [24, 33, 52] that the activation functions are piecewise affine linear, the restriction to the one-hidden-layer case in [52], the restriction to the squared-loss function in [11, 19, 52], and the assumption of realizability in [19] 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 section 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 section 5 have essentially already been proved 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 section 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 [52, 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 to prove the spuriousness of local minima in a natural way. In section 5, we discuss further implications of the geometric properties of the images of neural networks exploited in section 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. In section 6, we conclude our analysis with a numerical experiment which demonstrates that the spurious local minima in Theorems 3.1 and 3.2 can indeed cause gradient-based algorithms to terminate with a suboptimal local minimum when the hyperparameters are chosen poorly. Here, we also investigate how robust this behavior is w.r.t. perturbations.

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 τ_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 $v\colon K \to \mathbb{R}$ equipped with the maximum norm $\|v\|_{C(K)} := \max\{|v(x)|\colon x \in K\}$ 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 [22, Corollary 7.18]. The corresponding dual pairing is denoted by $\langle\cdot,\cdot\rangle_{C(K)}\colon \mathcal{M}(K) \times C(K) \to \mathbb{R}$. 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 $L^p_\mu(K)$, $1 \le p \le \infty$, and equipped with the usual norms $\|\cdot\|_{L^p_\mu(K)}$, see [4, 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 $\mathrm{cl}_Z(E)$.

The neural networks that we study in this paper are standard fully-connected feedforward 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 $w_i \in \mathbb{N}$, $i = 1, ..., L$. 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 $b_i \in \mathbb{R}^{w_i}$, $i = 1, ..., L+1$, respectively, and the activation functions of the layers by $\sigma_i\colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L$. Here and in what follows, all vectors of real numbers are considered as column vectors. We will always assume that the functions σ_i are continuous, i.e., $\sigma_i \in C(\mathbb{R})$ for all $i = 1, ..., L$. To describe the action of the network layers, we define $\varphi_i^{A_i,b_i}\colon \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i}$, $i = 1, ..., L+1$, to be the functions

$$\varphi_i^{A_i,b_i}(z) := \sigma_i\left(A_i z + b_i\right) \ \forall i = 1, ..., L, \qquad \varphi_{L+1}^{A_{L+1},b_{L+1}}(z) := A_{L+1} z + b_{L+1},$$

with σ_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:

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

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 α, i.e., the parameter space of a network, we write

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

We equip the space D with the Euclidean norm $|\cdot|$ of the space $\mathbb{R}^m$, $m := \dim(D) = w_{L+1}(w_L + 1) + ... + w_1(w_0 + 1)$, that D can be transformed into by rearranging the entries of α. Note that, due to the continuity of the activation functions σ_i, the map $\psi\colon D \times \mathbb{R}^d \to \mathbb{R}$ in (2.1) gives rise to an operator from D into the space $C(K)$. We denote this operator by Ψ, i.e.,

$$\Psi\colon D \to C(K), \qquad \Psi(\alpha) := \psi(\alpha, \cdot)\colon K \to \mathbb{R}. \tag{2.2}$$

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

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

Here, $\mathcal{L}\colon C(K) \times C(K) \to \mathbb{R}$ denotes the loss function and $y_T \in C(K)$ the target function. Given a tuple $(v, y_T) \in C(K) \times C(K)$, we call $\mathcal{L}$ Gâteaux differentiable in its first argument at (v, y_T) if the limit

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

exists for all $h \in C(K)$ and if the map $\partial_1 \mathcal{L}(v, y_T; \cdot)\colon C(K) \to \mathbb{R}$, $h \mapsto \partial_1 \mathcal{L}(v, y_T; h)$, 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. [7, section 2.2.1]. As usual, a local minimum of (P) is a point $\bar{\alpha} \in D$ that satisfies

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

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

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

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 as precise as possible, 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 $\psi\colon D \times \mathbb{R}^d \to \mathbb{R}$ be a neural network with depth $L \in \mathbb{N}$, widths $w_i \in \mathbb{N}$, $i = 0, ..., L+1$, and nonpolynomial continuous activation functions $\sigma_i\colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L$, as in (2.1). Assume that:*

i) $w_i \geq 2$ holds for all $i = 1, ..., L$.
ii) σ_i is affine and nonconstant on an open interval $I_i \neq \emptyset$ for all $i = 1, ..., L$.
iii) $y_T \in C(K)$ is nonaffine, i.e., $\nexists (a, c) \in \mathbb{R}^d \times \mathbb{R}\colon y_T(x) = a^\top x + c \; \forall x \in K$.
iv) $\mathcal{L}\colon C(K) \times C(K) \to \mathbb{R}$ 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

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

Then there exists a set $E \subset D$ of Hausdorff dimension at least $\dim(D) - d - 1$ such that all elements of E are spurious local minima of the training problem

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

and such that

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

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 $\psi\colon D \times \mathbb{R}^d \to \mathbb{R}$ be a neural network with depth $L \in \mathbb{N}$, widths $w_i \in \mathbb{N}$, $i = 0, ..., L+1$, and nonpolynomial continuous activation functions $\sigma_i\colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L$, as in (2.1). Assume that:*

i) σ_j is constant on an open interval $I_j \neq \emptyset$ for some $j \in \{1, ..., L\}$.

ii) $y_T \in C(K)$ is nonconstant, i.e., $\nexists c \in \mathbb{R}\colon y_T(x) = c \; \forall x \in K$.

iii) $\mathcal{L}\colon C(K) \times C(K) \to \mathbb{R}$ 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

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

Then there exists a set $E \subset D$ of Hausdorff dimension at least $\dim(D) - 1$ such that all elements of E are spurious local minima of the training problem

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

and such that

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

The proofs of Theorems 3.1 and 3.2 rely on geometric properties of the image $\Psi(D)$ of the function $\Psi\colon D \to C(K)$ in (2.2) and are carried out in section 4, see Theorem 4.2 and Lemmas 4.3 to 4.6. 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 σ_i in Theorems 3.1 and 3.2 are optimal. The only continuous σ_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 proved in [30] that spurious local minima do not exist so that relaxing the assumptions on σ_i in Theorems 3.1 and 3.2 in this direction is provably impossible. Compare also with [20, 31, 44, 52, 53] in this context. Note that Theorem 3.1 covers in particular neural networks which 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 extended rather straightforwardly 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 [24] and the proof of [11, Corollary 47]. In particular, the arguments that we use in section 4 to establish that the local minima in E are indeed spurious carry over immediately. We omit this generalization 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 $\mathcal{L}(\cdot, y_T)\colon C(K) \to \mathbb{R}$ 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 $\mathcal{F}\colon C(K) \to [0, \infty)$ 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 $\mathcal{L}(\cdot, y_T)\colon C(K) \to \mathbb{R}$ 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*

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

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 [4, Theorem 3.3.2], it follows that $\mathcal{L}\colon C(K) \times C(K) \to [0, \infty)$ is Gâteaux differentiable everywhere with

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

Since $C(K)$ is dense in $L^q_\mu(K)$ for all $1 \leq q < \infty$ by [22, Proposition 7.9], (3.2) yields

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

Due to the continuity of the function $|v - y_T|^{p-1}$ and since the assumptions on μ 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 μ. This implies that restrictions of the map $\mathcal{L}(\cdot, y_T)\colon C(K) \to \mathbb{R}$ 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. ∎

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

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

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 [52, Theorem 1], where the assertion of Theorem 3.1 was proved for finite-dimensional squared-loss training problems for one-hidden-layer neural networks with activation functions of parameterized ReLU-type. Compare also with [11, 19, 24, 33] in this context. Another

natural choice for μ 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 $K = \mathrm{cl}_{\mathbb{R}^d}(\{x_k\}_{k=1}^\infty)$ 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 μ 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 relative and absolute terms and in terms of loss, see [11, Corollary 47]. It can also be proved that the appearance of spurious local minima in (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 α, see [11, Corollary 51]. The proofs used to establish these results in [11] make use of compactness arguments and homogeneity properties of $\mathcal{L}$ and thus do not carry over immediately to the general infinite-dimensional setting considered in Theorems 3.1 and 3.2, cf. the derivation of [11, Lemma 10].

4. Nonexistence of supporting half-spaces and proof 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]). *Suppose that a $d \in \mathbb{N}$ and a nonpolynomial continuous function $\sigma\colon \mathbb{R} \to \mathbb{R}$ are given. Consider the linear hull*

$$V := \mathrm{span}\left\{x \mapsto \sigma(a^\top x + c) \;\middle|\; a \in \mathbb{R}^d, c \in \mathbb{R}\right\} \subset C(\mathbb{R}^d). \tag{4.1}$$

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. [16, 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 ψ with the architecture in (2.1), depth $L = 1$, width $w_1 \geq \tilde{w}_1$, and activation function σ is able to approximate y_T in $(C(K), \|\cdot\|_{C(K)})$ up to the error ε. 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 $\Psi\colon D \to C(K)$ in (2.2) does not admit any supporting half-spaces when a neural network ψ with nonpolynomial continuous activations σ_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 $\psi\colon D \times \mathbb{R}^d \to \mathbb{R}$ be a neural network with depth $L \in \mathbb{N}$, widths $w_i \in \mathbb{N}$, $i = 0, ..., L+1$, and continuous nonpolynomial activation functions $\sigma_i\colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L$, as in (2.1). Denote with $\Psi\colon D \to C(K)$ the function in (2.2). Then a measure $\mu \in \mathcal{M}(K)$ and a constant $c \in \mathbb{R}$ satisfy*

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

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

Proof. The implication "$\Leftarrow$" is trivial. To prove "$\Rightarrow$", we assume that a $c \in \mathbb{R}$ and a $\mu \in \mathcal{M}(K)$ satisfying (4.2) are given. From the definition of Ψ, we obtain that $\beta\Psi(\alpha) \in \Psi(D)$ holds for all $\beta \in \mathbb{R}$ and all $\alpha \in D$. Using this information in (4.2) yields that c and μ have to satisfy $c \geq 0$ and

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

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

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

for some arbitrary but fixed $a_1 \in \mathbb{R}^d$, $a_i \in \mathbb{R}$, $i = 2, ..., L+1$, and $c_i \in \mathbb{R}$, $i = 1, ..., L+1$, 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 an $\tilde{\alpha}$, we obtain from (2.1) that

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

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

$$\begin{gathered} \theta_1^{a_1,c_1}(z) := \sigma_1\left(a_1^\top z + c_1\right), \qquad \theta_i^{a_i,c_i}(z) := \sigma_i\left(a_i z + c_i\right) \ \forall i = 2, ..., L, \\ \theta_{L+1}^{a_{L+1},c_{L+1}}(z) := a_{L+1} z + c_{L+1}. \end{gathered}$$

In combination with (4.3) and the definition of Ψ, this yields that

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

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

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

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

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

holds for all $k = 1, ..., n_l$ and all $l \in \mathbb{N}$. Due to the linearity of the integral, this yields

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

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

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

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

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

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. ∎

We remark that Theorem 4.2 is, in fact, equivalent to Theorem 4.1. Indeed, the implication "Theorem 4.1 ⇒ Theorem 4.2" has been proved 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 [16] 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.

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 4.3. *Consider a nonempty compact set $K \subset \mathbb{R}^d$ and a neural network $\psi \colon D \times \mathbb{R}^d \to \mathbb{R}$ with depth $L \in \mathbb{N}$, widths $w_i \in \mathbb{N}$, and continuous activation functions σ_i as in (2.1). Suppose that an $\mathcal{L} \colon C(K) \times C(K) \to \mathbb{R}$ and a $y_T \in C(K)$ are given such that $\mathcal{L}$, y_T, and the functions σ_i satisfy the conditions ii) and v) in Theorem 3.1. Then there exists a set $E \subset D$ of Hausdorff dimension at least $\dim(D) - d - 1$ such that all elements of E are local minima of (P) and such that*

$$\mathcal{L}(\Psi(\alpha), y_T) = \min_{(a,c) \in \mathbb{R}^d \times \mathbb{R}} \mathcal{L}(z_{a,c}, y_T) \tag{4.6}$$

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 $i = 1, ..., L$, and from v), we obtain that there exist $\bar{a} \in \mathbb{R}^d$ and $\bar{c} \in \mathbb{R}$ satisfying

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

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

$$
\begin{aligned}
\bar{A}_1 &:= \frac{\varepsilon_1}{2\max_{u\in K}|\bar{a}^\top u|+1}\begin{pmatrix}\bar{a}^\top\\ 0_{(w_1-1)\times w_0}\end{pmatrix} \in \mathbb{R}^{w_1\times w_0},\\
\bar{b}_1 &:= c_1 1_{w_1} \in \mathbb{R}^{w_1},\\
\bar{A}_i &:= \frac{\varepsilon_i}{\beta_{i-1}\varepsilon_{i-1}}\begin{pmatrix}1 & 0_{1\times(w_{i-1}-1)}\\ 0_{(w_i-1)\times w_{i-1}}\end{pmatrix} \in \mathbb{R}^{w_i\times w_{i-1}}, \qquad i=2,...,L,\\
\bar{b}_i &:= c_i 1_{w_i} - (c_{i-1}\beta_{i-1}+\gamma_{i-1})\bar{A}_i 1_{w_{i-1}} \in \mathbb{R}^{w_i}, \qquad i=2,...,L,\\
\bar{A}_{L+1} &:= \frac{2\max_{u\in K}|\bar{a}^\top u|+1}{\beta_L\varepsilon_L}\begin{pmatrix}1 & 0_{1\times(w_L-1)}\end{pmatrix} \in \mathbb{R}^{w_{L+1}\times w_L},\\
\bar{b}_{L+1} &:= \bar{c} - (c_L\beta_L+\gamma_L)\bar{A}_{L+1}1_{w_L} \in \mathbb{R}^{w_{L+1}},
\end{aligned}
\tag{4.7}
$$

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

$$
\begin{aligned}
\bar{A}_1 x+\bar{b}_1 &= \frac{\varepsilon_1}{2\max_{u\in K}|\bar{a}^\top u|+1}\begin{pmatrix}\bar{a}^\top x\\ 0_{w_1-1}\end{pmatrix} + c_1 1_{w_1} \in (c_1-\varepsilon_1, c_1+\varepsilon_1)^{w_1},\\
\bar{A}_i\left(\varphi_{i-1}^{\bar{A}_{i-1},\bar{b}_{i-1}}\circ ...\circ\varphi_1^{\bar{A}_1,\bar{b}_1}(x)\right)+\bar{b}_i &= \frac{\varepsilon_i}{2\max_{u\in K}|\bar{a}^\top u|+1}\begin{pmatrix}\bar{a}^\top x\\ 0_{w_i-1}\end{pmatrix} + c_i 1_{w_i}\\
&\in (c_i-\ \varepsilon_i, c_i+\varepsilon_i)^{w_i} \quad \forall i=2,...,L,
\end{aligned}
\tag{4.8}
$$

and

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

The parameter $\bar{\alpha}$ thus satisfies $\Psi(\bar{\alpha})=z_{\bar{a},\bar{c}}\in C(K)$. Since K is compact, since the sets $(c_i-\varepsilon_i,c_i+\varepsilon_i)^{w_i}$, $i=1,...,L$, are open, since $D\times\mathbb{R}^d\ni(\alpha,x)\mapsto A_1x+b_1\in\mathbb{R}^{w_1}$ and $D\times\mathbb{R}^d\ni(\alpha,x)\mapsto A_i(\varphi_{i-1}^{A_{i-1},b_{i-1}}\circ ...\circ\varphi_1^{A_1,b_1}(x))+b_i\in\mathbb{R}^{w_i}$, $i=2,...,L$, are continuous functions, and since σ_i is affine linear on $(c_i-\varepsilon_i,c_i+\varepsilon_i)$, it follows that there exists an $r>0$ such that all of the inclusions in (4.8) 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 (4.6).

To show that there are many such local minima, we require some additional notation. Henceforth, with $a_1,...,a_{w_1}\in\mathbb{R}^d$ we denote the row vectors in the weight matrix A_1 and with $e_1,...,e_{w_1}\in\mathbb{R}^{w_1}$ the standard basis vectors of $\mathbb{R}^{w_1}$. We further introduce the abbreviation α' for the collection of all parameters of ψ that belong to the degrees of freedom $A_{L+1},...,A_2,b_L,...,b_1$, and $a_2,...,a_{w_1}$. The space of all such α' 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 α'. As before, in what follows, a bar indicates that we refer to the parameter $\bar{\alpha}\in D$ constructed in (4.7), i.e., $\bar{a}_k$ refers to the k-th row of $\bar{A}_1$, $\bar{\alpha}'\in D'$ refers to $(\bar{A}_{L+1},...,\bar{A}_2,\bar{b}_L,...,\bar{b}_1,\bar{a}_2,...,\bar{a}_{w_1})$, etc.

To construct a set $E\subset D$ as in the assertion of the lemma, we first note that the local affine linearity of σ_i, the definition of $\bar{\alpha}$, our choice of $r>0$, and the architecture of ψ imply that there exists a continuous function $\Phi\colon D'\to\mathbb{R}$ which

satisfies $\Phi(\bar\alpha') + \bar b_{L+1} = \bar c$ and

$$\psi(\alpha, x) = \left(\prod_{i=1}^{L} \beta_i\right) (A_{L+1} A_L ... A_1)\, x + \Phi(\alpha') + b_{L+1} \tag{4.9}$$

for all $x \in K$ and all $\alpha = \{(A_i, b_i)\}_{i=1}^{L+1} \in B_r^D(\bar\alpha)$, cf. (4.8). Define

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

and

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

Then (4.9) can be recast as

$$\psi(\alpha, x) = \Theta(\alpha')^\top x + \Lambda(\alpha') a_1^\top x + \Phi(\alpha') + b_{L+1} \qquad \forall x \in K \qquad \forall \alpha \in B_r^D(\bar\alpha). \tag{4.10}$$

Note that, again by the construction of $\bar\alpha$ in (4.7), 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 $\Lambda\colon D' \to \mathbb{R}$, we can find an $r' > 0$ such that $\Lambda(\alpha') \neq 0$ holds for all $\alpha' \in B_{r'}^{D'}(\bar\alpha')$. This allows us to define

$$g_1\colon B_{r'}^{D'}(\bar\alpha') \to \mathbb{R}^d, \qquad g_1(\alpha') := \frac{\Lambda(\bar\alpha')}{\Lambda(\alpha')} \bar a_1 - \frac{\Theta(\alpha')}{\Lambda(\alpha')},$$

and

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

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

$$\Theta(\alpha')^\top x + \Lambda(\alpha') g_1(\alpha')^\top x + \Phi(\alpha') + g_2(\alpha') = \bar a^\top x + \bar c \tag{4.11}$$

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

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

For all elements $\tilde\alpha$ of the resulting set E, it now follows from (4.10) and (4.11) that

$$\begin{aligned} \psi(\tilde\alpha, x) &= \Theta(\tilde\alpha')^\top x + \Lambda(\tilde\alpha') \tilde a_1^\top x + \Phi(\tilde\alpha') + \tilde b_{L+1} \\ &= \Theta(\tilde\alpha')^\top x + \Lambda(\tilde\alpha') g_1(\tilde\alpha')^\top x + \Phi(\tilde\alpha') + g_2(\tilde\alpha') = \bar a^\top x + \bar c \qquad \forall x \in K. \end{aligned}$$

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 (4.6). Since E is, modulo reordering of the entries in α, 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 $m - d - 1$ follows immediately from the choice of the norm on D and classical results, see [18, Corollary 8.2c]. ∎

LEMMA 4.4. *Consider a nonempty compact set $K \subset \mathbb{R}^d$ and a neural network $\psi\colon D \times \mathbb{R}^d \to \mathbb{R}$ with depth $L \in \mathbb{N}$, widths $w_i \in \mathbb{N}$, and continuous activation functions σ_i as in (2.1). Suppose that an $\mathcal{L}\colon C(K) \times C(K) \to \mathbb{R}$ and a $y_T \in C(K)$ are given such that $\mathcal{L}$, y_T, and the functions σ_i satisfy the conditions i) and iv) in Theorem 3.2. Then there exists a set $E \subset D$ of Hausdorff dimension at least $\dim(D) - 1$ such that all elements of E are local minima of (P) and such that*

$$\mathcal{L}(\Psi(\alpha), y_T) = \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T) \tag{4.12}$$

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 4.4 is analogous to that of Lemma 4.3 but simpler: From i), we obtain that there exist a $j \in \{1, ..., L\}$ 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 $\bar{c} \in \mathbb{R}$ 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

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

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

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

for all $x \in K$. Let us denote the collection of all parameters of ψ belonging to the degrees of freedom $A_{L+1}, ..., A_1$ and $b_L, ..., b_1$ with α' and the space of all such α' with D' (again endowed with the Euclidean norm of the associated space $\mathbb{R}^{m-1}$ analogously to the proof of Lemma 4.3). Then the compactness of K, the openness of I_j, the fact that σ_j is constant on I_j, the definition of $\bar{\alpha}$, the architecture of ψ, and the continuity of the function $D \times \mathbb{R}^d \ni (\alpha, x) \mapsto A_j(\varphi_{j-1}^{A_{j-1}, b_{j-1}} \circ ... \circ \varphi_1^{A_1, b_1}(x)) + b_j \in \mathbb{R}^{w_j}$ imply that there exist an $r > 0$ and a continuous $\Phi\colon D' \to \mathbb{R}$ such that $\Phi(\bar{\alpha}') = 0$ holds and

$$\psi(\alpha, x) = \Phi(\alpha') + b_{L+1} \qquad \forall x \in K \qquad \forall \alpha \in B_r^D(\bar{\alpha}). \tag{4.13}$$

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

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

For all $\tilde{\alpha} \in E$, it now follows from (4.13) and the definition of g that

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

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 (4.12). That E has Hausdorff dimension at least $\dim(D) - 1$ follows completely analogously to the proof of Lemma 4.3. ∎

As already mentioned in the introduction, the approach that we have used in Lemmas 4.3 and 4.4 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 σ_i can also be found in various other contributions, e.g., [11, 19, 23, 24, 52]. The main challenge in the context of Theorems 3.1 and 3.2 is proving that the local minima in Lemmas 4.3 and 4.4 are indeed spurious for generic y_T and arbitrary σ_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 4.5. *Suppose that K, ψ, w_i, L, σ_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 4.3. Then it holds*

$$\inf_{\alpha \in D} \mathcal{L}(\Psi(\alpha), y_T) < \min_{(a,c)\in\mathbb{R}^d\times\mathbb{R}} \mathcal{L}(z_{a,c}, y_T). \tag{4.14}$$

Proof. We argue by contradiction. Suppose that the assumptions of Theorem 3.1 are satisfied and that (4.14) is false. Then it holds

$$\mathcal{L}(\Psi(\alpha), y_T) \geq \min_{(a,c)\in\mathbb{R}^d\times\mathbb{R}} \mathcal{L}(z_{a,c}, y_T) = \mathcal{L}(z_{\bar{a},\bar{c}}, y_T) \qquad \forall \alpha \in D, \tag{4.15}$$

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

$$\begin{aligned} A_1 &:= \begin{pmatrix} \bar{A}_1 \\ \tilde{A}_1 \end{pmatrix} \in \mathbb{R}^{w_1 \times w_0}, \\ A_i &:= \begin{pmatrix} \bar{A}_i & 0_{1\times(w_{i-1}-1)} \\ 0_{(w_i-1)\times 1} & \tilde{A}_i \end{pmatrix} \in \mathbb{R}^{w_i \times w_{i-1}},\ i = 2, ..., L, \\ A_{L+1} &:= (\bar{A}_{L+1} \quad \tilde{A}_{L+1}) \in \mathbb{R}^{w_{L+1}\times w_L}, \\ b_i &:= \begin{pmatrix} \bar{b}_i \\ \tilde{b}_i \end{pmatrix} \in \mathbb{R}^{w_i},\ i = 1, ..., L, \qquad b_{L+1} := \bar{b}_{L+1} + \tilde{b}_{L+1} \in \mathbb{R} \end{aligned} \tag{4.16}$$

with arbitrary but fixed $\bar{A}_1 \in \mathbb{R}^{1\times d}$, $\bar{A}_i \in \mathbb{R}$, $i = 2, ..., L+1$, $\bar{b}_i \in \mathbb{R}$, $i = 1, ..., L+1$, $\tilde{A}_1 \in \mathbb{R}^{(w_1-1)\times d}$, $\tilde{A}_i \in \mathbb{R}^{(w_i-1)\times(w_{i-1}-1)}$, $i = 2, ..., L$, $\tilde{A}_{L+1} \in \mathbb{R}^{1\times(w_L-1)}$, $\tilde{b}_i \in \mathbb{R}^{w_i-1}$, $i = 1, ..., L$, 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 parameter is possible due to the assumption $w_i \geq 2$, $i = 1, ..., L$, in i). Using (2.1), it is easy to check that every α of the type (4.16) 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 $\bar{w}_i = 1$, $i = 1, ..., L$, activation functions σ_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 $\tilde{w}_i = w_i - 1$, $i = 1, ..., L$, activation functions σ_i, and network parameter $\tilde{\alpha} = \{(\tilde{A}_i, \tilde{b}_i)\}_{i=1}^{L+1}$. In combination with (4.15), this implies

$$\mathcal{L}(\bar{\Psi}(\bar{\alpha}) + \tilde{\Psi}(\tilde{\alpha}), y_T) \geq \mathcal{L}(z_{\bar{a},\bar{c}}, y_T) \qquad \forall \bar{\alpha} \in \bar{D} \qquad \forall \tilde{\alpha} \in \tilde{D}. \tag{4.17}$$

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 4.3, we obtain that there exists an $\bar{\alpha} \in \bar{D}$ with $\bar{\Psi}(\bar{\alpha}) = z_{\bar{a},\bar{c}}$. Due to (4.17) and the fact that $\tilde{A}_{L+1}$ and $\tilde{b}_{L+1}$ can be rescaled at will, this yields

$$\mathcal{L}(z_{\bar{a},\bar{c}} + s\tilde{\Psi}(\tilde{\alpha}), y_T) \geq \mathcal{L}(z_{\bar{a},\bar{c}}, y_T) \qquad \forall \tilde{\alpha} \in \tilde{D} \qquad \forall s \in (0, \infty). \tag{4.18}$$

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 (4.18), divide by $s > 0$, and pass to the limit $s \to 0^+$ (for an arbitrary but fixed $\tilde{\alpha} \in \tilde{D}$) to obtain

$$\left\langle \partial_1 \mathcal{L}(z_{\bar{a},\bar{c}}, y_T), \tilde{\Psi}(\tilde{\alpha}) \right\rangle_{C(K)} \geq 0 \qquad \forall \tilde{\alpha} \in \tilde{D} \tag{4.19}$$

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 (4.19) can only be true if $\partial_1 \mathcal{L}(z_{\bar{a},\bar{c}}, y_T) = 0$ holds. Thus, we arrive at a contradiction, (4.15) cannot be correct, and the proof is complete. ∎

LEMMA 4.6. *Suppose that K, ψ, w_i, L, σ_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 4.4. Then it holds*

$$\inf_{\alpha \in D} \mathcal{L}(\Psi(\alpha), y_T) < \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T). \tag{4.20}$$

Proof. The proof of Lemma 4.6 is analogous to that of Lemma 4.5 but simpler. Suppose that (4.20) is false and that the assumptions of Theorem 3.2 are satisfied. Then it holds

$$\mathcal{L}(\Psi(\alpha), y_T) \geq \min_{c \in \mathbb{R}} \mathcal{L}(z_c, y_T) = \mathcal{L}(z_{\bar{c}}, y_T) \qquad \forall \alpha \in D, \tag{4.21}$$

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 (4.21), we obtain that

$$\mathcal{L}(z_{\bar{c}} + s\Psi(\alpha), y_T) \geq \mathcal{L}(z_{\bar{c}}, y_T) \qquad \forall \alpha \in D \qquad \forall s \in (0, \infty). \tag{4.22}$$

In combination with assumptions ii) and iii) of Theorem 3.2, (4.22) yields – completely analogously to (4.19) – that there exists a $\partial_1 \mathcal{L}(z_{\bar{c}}, y_T) \in \mathcal{M}(K) \setminus \{0\}$ satisfying

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

By invoking Theorem 4.2, we now again arrive at a contradiction. Thus, (4.21) cannot be true and the assertion of the lemma follows. ∎

To establish Theorems 3.1 and 3.2, it suffices to combine Lemmas 4.3 and 4.5 and Lemmas 4.4 and 4.6, respectively. This completes the proof of our main results on the existence of spurious local minima in training problems of the type (P).

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 $\psi \colon D \times \mathbb{R}^d \to \mathbb{R}$ 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 σ_i. Suppose that $(Z, \|\cdot\|_Z)$ is a real normed space and that $\iota \colon C(K) \to Z$ is a linear and continuous map with a dense image. Then the set $\mathrm{cl}_Z(\iota(\Psi(D)))$ is either nonconvex or equal to Z.*

Proof. Assume that $\mathrm{cl}_Z(\iota(\Psi(D)))$ is convex and that $\mathrm{cl}_Z(\iota(\Psi(D))) \neq Z$. Then there exists a $z \in Z \setminus \mathrm{cl}_Z(\iota(\Psi(D)))$ and it follows from the separation theorem for convex sets in normed spaces [21, Corollary I-1.2] that we can find a $\nu \in Z^* \setminus \{0\}$ and a $c \in \mathbb{R}$ such that

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

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

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

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

We remark that, for activation functions possessing a point of differentiability with a nonzero derivative, a version of Corollary 5.1 has already been proved 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 σ_i. In combination with classical results on the properties of Chebychev sets, see [49], the nonconvexity of the set $\mathrm{cl}_Z(\iota(\Psi(D)))$ in Corollary 5.1 immediately implies that the problem of determining a best approximating element for a given $u \in Z$ from the set $\mathrm{cl}_Z(\iota(\Psi(D)))$ 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, ψ, L, w_i, σ_i, $(Z, \|\cdot\|_Z)$, and ι 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 Π to be the best approximation map associated with the set $\mathrm{cl}_Z(\iota(\Psi(D)))$, i.e., the set-valued projection operator*

$$\Pi \colon Z \rightrightarrows Z, \qquad u \mapsto \operatorname*{arg\,min}_{z \in \mathrm{cl}_Z(\iota(\Psi(D)))} \|u - z\|_Z . \tag{5.2}$$

Then exactly one of the following is true:

i) $\mathrm{cl}_Z(\iota(\Psi(D)))$ is equal to Z and Π is the identity map.

ii) There does not exist a function $\pi \colon Z \to Z$ such that $\pi(z) \in \Pi(z)$ holds for all $z \in Z$ and such that π is continuous in an open neighborhood of the origin.

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

Note that there are two possible reasons for the nonexistence of a selection π with the properties in point ii) of Corollary 5.2. The first one is that there exists a $u \in Z$ 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 π taken from Π is discontinuous at some point u, i.e., that there exists a $u \in Z$ 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 $\mathrm{cl}_Z(\iota(\Psi(D))) \neq Z$ holds.

To make Corollary 5.2 more tangible, we next 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 $\iota\colon C(K) \to L^p_\mu(K)$ is the inclusion map. As usual, in the statement of the next corollary, we drop the inclusion map ι 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 ψ 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 σ_i. Assume that $\mu \in \mathcal{M}_+(K)$ and $p \in (1,\infty)$ are given and that the image $\Psi(D)$ of the map $\Psi\colon D \to C(K)$ in (2.2) is not dense in $L^p_\mu(K)$. Then there does not exist a function $\pi\colon L^p_\mu(K) \to L^p_\mu(K)$ such that*

$$\pi(u) \in \operatorname*{arg\,min}_{z \in \mathrm{cl}_{L^p_\mu(K)}(\Psi(D))} \|u - z\|_{L^p_\mu(K)} \qquad \forall u \in L^p_\mu(K) \tag{5.3}$$

holds and such that π is continuous in an open neighborhood of the origin.

Proof. From [36, Example 1.10.2, Theorem 5.2.11], it follows that $L^p_\mu(K)$ is uniformly convex with a uniformly convex dual, and from [22, Proposition 7.9], we obtain that the inclusion map $\iota\colon C(K) \to L^p_\mu(K)$ is linear and continuous with a dense image. The claim thus follows immediately from Corollary 5.2. ∎

As already mentioned in section 1, for neural networks with a single hidden layer, a variant of Corollary 5.3 has also been proved in [28, section 4]. For related results, see also [29, 40]. We obtain the discontinuity of $L^p_\mu(K)$-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 δ_x again denotes a Dirac measure supported at $x \in \mathbb{R}$. For these K and μ, we have $C(K) \cong L^2_\mu(K) \cong \mathbb{R}^3$ and the image $\Psi(D) \subset C(K)$ of the map Ψ in (2.2) can be identified with a subset of $\mathbb{R}^3$, namely,

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

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 $\mathrm{cl}_{\mathbb{R}^3}(\Psi(D))$, i.e.,

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

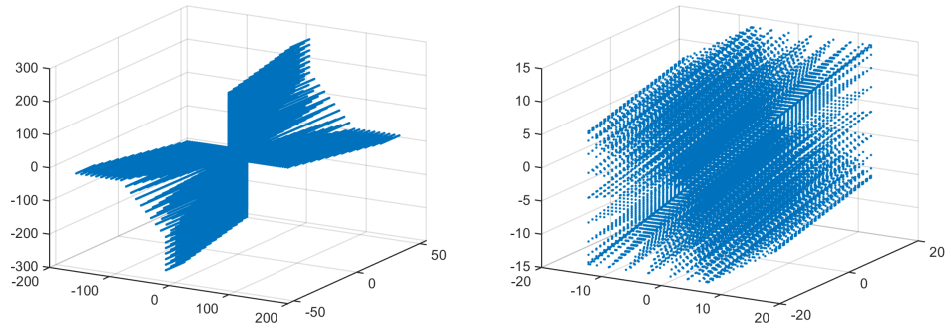
This makes it possible to visualize the image $\Psi(D)$ and to interpret the $L^2_\mu(K)$-best approximation operator associated with ψ geometrically. The sets $\Psi(D)$ that are obtained in the above situation for the ReLU-activation $\sigma_{\mathrm{relu}}(s) := \max(0, s)$ and the SQNL-activation

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

can be seen in Figure 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 $(0,1,0)^\top \notin \mathrm{cl}_{\mathbb{R}^3}(\Psi(D))$ holds. This shows that, for both the ReLU- and the SQNL-activation, the resulting network falls under the scope of Corollary 5.3. Since σ_{relu} and σ_{sqnl} possess constant segments, the training problems

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

associated with these activation functions are moreover covered by Theorem 3.2, cf. Lemma 3.3. As Figure 1 shows, the sets $\Psi(D)$ obtained for σ_{relu} and σ_{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 $\mathrm{cl}_{\mathbb{R}^3}(\Psi(D))$, possess discontinuities and give rise to training problems that contain various spurious local minima. We remark that the examples in Figure 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.



(a) $\Psi(D)$ for σ_{relu} (b) $\Psi(D)$ for σ_{sqnl}

Fig. 1: Scatter plot of the image $\Psi(D)$ of the function $\Psi\colon D \to C(K) \cong L^2_\mu(K) \cong \mathbb{R}^3$ 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.

Before we demonstrate that the effects discussed in Theorems 3.1 and 3.2 and Corollary 5.3 can indeed affect the behavior of gradient-based optimization algorithms in practice, we would like to point out that the "space-filling" cases $\mathrm{cl}_Z(\iota(\Psi(D))) = Z$ and $\mathrm{cl}_{L^p_\mu(K)}(\Psi(D)) = L^p_\mu(K)$ 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 ψ exceeds the number of training samples by far and in which ψ is able to fit arbitrary training data with zero error, see [2, 8, 15, 32, 39]. In the situation of Lemma 3.3, this means that a measure μ of the form $\mu = \frac{1}{n}\sum_{k=1}^n \delta_{x_k}$ supported on a finite set $K = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $n \in \mathbb{N}$, 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, 32, 39]. For $d = 1$, it can also 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 $\sigma\colon \mathbb{R} \to \mathbb{R}$ for which the density conditions $\mathrm{cl}_Z(\iota(\Psi(D))) = Z$ and $\mathrm{cl}_{L^p_\mu(K)}(\Psi(D)) = L^p_\mu(K)$ hold for arbitrary Z and μ. To be more precise, we have:

LEMMA 5.4. *Consider a nonempty compact set $K \subset \mathbb{R}$ and a neural network ψ 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 $i = 1, ..., L$ 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 σ with $\tilde{\sigma}$ in ψ satisfies $\mathrm{cl}_{C(K)}(\tilde{\Psi}(D)) = C(K)$.*

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 ψ 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 ψ have width one. Suppose that an $\varepsilon > 0$ and a nonempty open interval I are given. Using the continuity of σ, 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 $\phi\colon \mathbb{R} \to \mathbb{R}$ be the unique element of $C(\mathbb{R})$ with the following properties:

i) $\phi \equiv 0$ in $\mathbb{R} \setminus (a, a + \eta)$,
ii) ϕ 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 $\mathrm{cl}_{C([0,1])}(\tilde{\Psi}(D)) = C([0, 1])$. To prove this, we observe that, due to the choice of p_1 and the properties of $\bar{\sigma}$ and ϕ, we have

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

This equation allows us to turn the functions $\varphi_i^{A_i,b_i}\colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L - 1$, 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 4.1. For this one-hidden-layer case, we obtain analogously to (5.6) that

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

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 $\mathrm{cl}_{C([0,1])}(\tilde{\Psi}(D)) = C([0, 1])$ now follows immediately. This completes the proof. ∎

We remark that, under suitable assumptions on the depth and the widths of ψ, Lemma 5.4 can also be extended to the case $d > 1$, cf. [35, Theorem 4]. For some criteria ensuring that the image of Ψ is not dense, see [40, Appendix C3].

6. Numerical experiment. We conclude this paper with a simple numerical experiment which demonstrates that the spurious local minima in Theorem 3.1 can indeed affect the convergence behavior of gradient-based optimization algorithms in practice. For similar tests, see also [10, 23, 25]. As a prototypical example of a problem of the type (P), we consider a squared-loss training problem of the form

$$\text{Minimize} \quad \frac{1}{n}\sum_{k=1}^{n} |\psi(\alpha, x_k) - y_T(x_k)|^2 \qquad \text{w.r.t.} \quad \alpha \in D \tag{6.1}$$

for a neural network $\psi\colon D \times \mathbb{R}^d \to \mathbb{R}$ as in (2.1) with $d = 1$, $L = 2$, $w_0 = w_3 = 1$, $w_1 = w_2 = 2$, and ELU-activation functions, i.e.,

$$\sigma_1(s) := \sigma_2(s) := \begin{cases} s & \text{if } s > 0 \\ \mathrm{e}^s - 1 & \text{if } s \leq 0. \end{cases} \tag{6.2}$$

We choose the training samples $\{x_1, ..., x_n\}$ in (6.1) to be the $n := 501$ nodes of an equidistant partition of the interval $[-1, 1]$ with width $1/500$ and the target function y_T to be the map $y_T\colon \mathbb{R} \to \mathbb{R}$, $y_T(x) := \mathrm{e}^{x-1} - 1$ (or, more precisely, the restriction of this map to the set $\{x_1, ..., x_n\}$). Note that the function y_T is reproduced exactly by ψ on $\{x_1, ..., x_n\}$ if the network parameters are set to

$$A_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad b_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad b_3 = -1.$$

In fact, there are uncountably many choices of the parameter vector $\alpha = \{(A_i, b_i)\}_{i=1}^{L+1}$ that have this property. The optimal value of (6.1) is thus zero and (6.1) possesses infinitely many global minima. That (6.1) is covered by the analysis of sections 3 and 4 can be checked easily by means of Lemma 3.3 and (6.2). Indeed, with the definitions $K := \{x_1, ..., x_n\}$, $p := 2$, and $\mu := \frac{1}{n}\sum_{k=1}^{n} \delta_{x_k}$, we have

$$\int_K |\Psi(\alpha) - y_T|^2 \,\mathrm{d}\mu = \frac{1}{n}\sum_{k=1}^{n} |\psi(\alpha, x_k) - y_T(x_k)|^2$$

so that the objective function of (6.1) has precisely the form (3.1). In combination with Lemma 3.3, this yields that the loss function of (6.1) satisfies points iv) and v) of Theorem 3.1. Since the ELU-activation is nonpolynomial and equal to the identity map in $(0, \infty)$ and due to the choice of y_T and ψ, it is obvious that the remaining assumptions of Theorem 3.1 hold as well. In summary, this shows that (6.1) falls under the scope of Theorem 3.1 and, due to the statement of this theorem, that there exists a subset E of the parameter space of ψ with Hausdorff dimension at least eleven such that every element of E is a spurious local minimum of (6.1). To construct a starting value $\bar{\alpha}$ for the numerical solution of (6.1) that is bad in the sense that it causes algorithms to converge to an element of E, we use the formulas in (4.7) with $c_i = 1/2$, $\varepsilon_i = 1/2$, $\beta_i = 1$, and $\gamma_i = 0$ for $i = 1, 2$ and with $\bar{a}, \bar{c} \in \mathbb{R}$ as the coefficients of the affine least-squares best approximation of y_T with samples $\{x_1, ..., x_n\}$, i.e.,

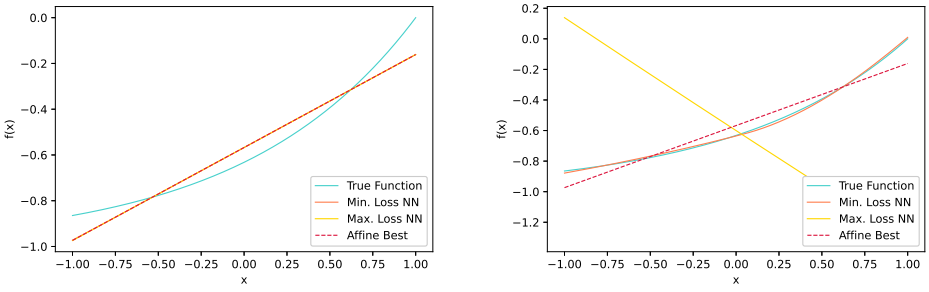
$$\begin{pmatrix} \bar{a} \\ \bar{c} \end{pmatrix} = (XX^\top)^{-1} X y \quad \text{with} \quad X := \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix} \quad \text{and} \quad y := \{y_T(x_k)\}_{k=1}^{n}.$$

The results that we have obtained in the above setting by applying the Adam optimizer implemented in Pytorch to (6.1) with learning rate 0.1, 1000 epochs, and various

starting values for α can be seen in Table 1 and Figure 2. In these experiments, we chose parameters of the form $\bar{\alpha}+r\xi$ with a uniformly distributed random perturbation $\xi \sim \mathcal{U}(-0.5, 0.5)^{\dim(D)}$ acting on each entry of $\bar{\alpha}$ and a scaling factor $r \geq 0$ to initialize the optimization algorithm. For each noise level r, 50 tests were conducted and the lowest and highest final loss value as well as the mean of all losses were determined. Note that the loss value of the affine best approximation of y_T can be computed to be ≈ 0.003594 in the situation of (6.1) and that the Adam algorithm can indeed be applied to (6.1) due to the C^1-regularity of the ELU-activations in (6.2).

Table 1: Maximal, minimal, and average final loss values obtained by applying the Adam optimizer with learning rate 0.1, 1000 epochs, and starting values of the form $\bar{\alpha}+r\xi$ to (6.1) for different values of $r \geq 0$. For each r, the minimization was performed 50 times. The line marks the critical value of r at which the optimization algorithm begins to escape the spurious local minima in Theorem 3.1.

Noise Level r	Maximal Loss	Minimal Loss	Average Loss
0.0	0.003594	0.003594	0.003594
0.1	0.003604	0.003571	0.003592
0.3	0.003634	0.003565	0.003594
0.6	0.003778	0.003105	0.003569
1.0	0.008520	0.001773	0.003647
1.3	0.023214	0.000080	0.004449
1.6	0.128047	0.001062	0.008064
2.0	0.436183	0.000086	0.036019



(a) $r = 0.1$ (b) $r = 2$

Fig. 2: Depiction of the graphs of the neural networks that achieve the minimal and the maximal loss in the situation of Table 1 for $r = 0.1$ (Figure 2a) and $r = 2$ (Figure 2b) on $[-1, 1]$. The dashed and the blue line show the affine best approximation of y_T and the target function $y_T(x) = \mathrm{e}^{x-1} - 1$, respectively.

As the first four rows of Table 1 and Figure 2a show, when starting with values too close to $\bar{\alpha}$, the Adam optimizer is unable to escape the spurious local minima from Theorem 3.1 and terminates with parameter vectors that reproduce the affine linear best approximation of y_T with high accuracy. This confirms the predictions of section 3 and also shows that the spurious local minima in Theorems 3.1 and 3.2 can indeed trap the iterates of a numerical solution algorithm. For sufficiently large values of r, Adam is able to leave the region of local optimality of $\bar{\alpha}$ and to produce loss values

that are significantly smaller than that of the affine linear best approximation of y_T, see the last four rows of Table 1 and Figure 2b. However, it can also be observed that the difference between the maximal and minimal losses increases with r. This shows that there is a large variance in the training results and highlights the importance of the choice of the starting value in network training.

REFERENCES

[1] M. Ainsworth and Y. Shin, *Plateau phenomenon in gradient descent training of RELU networks: Explanation, quantification, and avoidance*, SIAM J. Sci. Comput., 43 (2021), pp. 3438–3468.

[2] Z. Allen-Zhu, Y. Li, and Z. Song, *A convergence theory for deep learning via over-parameterization*, in Proc. 36th Int. Conf. Mach. Learn., K. Chaudhuri and R. Salakhutdinov, eds., vol. 97, PMLR, 2019, pp. 242–252.

[3] P. Auer, M. Herbster, and M. K. Warmuth, *Exponentially many local minima for single neurons*, in Adv. Neur. Inform. Proc. Sys., D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., vol. 8, Curran Associates, Inc., 1996, pp. 316–322.

[4] J. J. Benedetto and W. Czaja, *Integration and Modern Analysis*, Birkhäuser Advanced Texts, Birkhäuser, Boston, 2010.

[5] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, *The modern mathematics of deep learning*, arxiv:2105.04026v1, 2021.

[6] A. L. Blum and R. L. Rivest, *Training a 3-node neural network is NP-complete*, Neur. Netw., 5 (1992), pp. 117–127.

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

[8] Z. Chen, Y. Cao, D. Zou, and Q. Gu, *How much over-parameterization is sufficient to learn deep ReLU networks?*, arxiv:1911.12360v3, 2020. publ. as conf. paper, ICLR2021.

[9] P. Cheridito, A. Jentzen, and F. Rossmannek, *Landscape analysis for shallow neural networks: complete classification of critical points for affine target functions*, arxiv:2103.10922v2, 2021.

[10] P. Cheridito, A. Jentzen, and F. Rossmannek, *Non-convergence of stochastic gradient descent in the training of deep neural networks*, J. Complexity, 64 (2021), p. 101540.

[11] C. Christof, *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 (2021), pp. 1–77.

[12] C. Christof and D. Hafemeyer, *On the nonuniqueness and instability of solutions of tracking-type optimal control problems*, Math. Control Relat. Fields, (2021). in press.

[13] C. Clason, *Introduction to Functional Analysis*, Compact Textbooks in Mathematics, Birkhäuser, Cham, 2020.

[14] A. Cohen, R. DeVore, G. Petrova, and P. Wojtaszczyk, *Optimal stable nonlinear approximation*, Found. Comput. Math., (2021). published online.

[15] Y. Cooper, *The critical locus of overparameterized neural networks*, arxiv:2005.04210v2, 2020.

[16] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Systems, 2 (1989), pp. 303–314.

[17] Y. Dauphin, R. Pascanu, C. Gülçehre, K. Cho, S. Ganguli, and Y. Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, in Adv. Neur. Inform. Proc. Sys., Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds., vol. 27, Curran Associates, Inc., 2014, pp. 2933–2941.

[18] E. DiBenedetto, *Real Analysis*, Birkhäuser Advanced Texts, Birkhäuser, second ed., 2016.

[19] T. Ding, D. Li, and R. Sun, *Sub-optimal local minima exist for almost all over-parameterized neural networks*, arxiv:1911.01413v3, 2020.

[20] A. Eftekhari, *Training linear neural networks: non-local convergence and complexity results*, in Proc. 37th Int. Conf. Mach. Learn., H. Daumé and A. Singh, eds., vol. 119, PMLR, 2020, pp. 2836–2847.

[21] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland Publishing Company, 1976.

[22] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts, Wiley, second ed., 1999.

[23] M. Goldblum, J. Geiping, A. Schwarzschild, M. Moeller, and T. Goldstein, *Truth or backpropaganda? An empirical investigation of deep learning theory*, arxiv:1910.00359v3,

2020. publ. as conf. paper, ICLR2020.
[24] F. He, B. Wang, and D. Tao, *Piecewise linear activations substantially shape the loss surfaces of neural networks*, arxiv:2003.12236v1, 2020. publ. as conf. paper, ICLR2020.
[25] D. Holzmüller and I. Steinwart, *Training two-layer ReLU networks with gradient descent is inconsistent*, arxiv:2002.04861v2, 2021.
[26] K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neur. Netw., 4 (1991), pp. 251–257.
[27] S. I. Kabanikhin, *Inverse and Ill-posed Problems: Theory and Applications*, De Gruyter, 2012.
[28] P. C. Kainen, V. Kůrková, and A. Vogt, *Approximation by neural networks is not continuous*, Neurocomputing, 29 (1999), pp. 47–56.
[29] P. C. Kainen, V. Kůrková, and A. Vogt, *Continuity of approximation by neural networks in L^p-spaces*, Ann. Oper. Res., 101 (2001), pp. 143–147.
[30] K. Kawaguchi, *Deep learning without poor local minima*, in Adv. Neur. Inform. Proc. Sys., D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds., vol. 29, Curran Associates, Inc., 2016, pp. 586–594.
[31] T. Laurent and J. von Brecht, *Deep linear neural networks with arbitrary loss: All local minima are global*, in Proc. 35th Int. Conf. Mach. Learn., J. G. Dy and A. Krause, eds., vol. 80, PMLR, 2018, pp. 2908–2913.
[32] Y. Li and Y. Liang, *Learning overparameterized neural networks via stochastic gradient descent on structured data*, in Proc. 32nd Int. Conf. Neur. Inform. Proc. Sys., NIPS'18, Curran Associates Inc., 2018, pp. 8168–8177.
[33] B. Liu, *Spurious local minima are common for deep neural networks with piecewise linear activations*, arxiv:2102.13233v1, 2021.
[34] J. Šíma, *Training a single sigmoidal neuron is hard*, Neural Comput., 14 (2002), pp. 2709–2728.
[35] V. Maiorov and A. Pinkus, *Lower bounds for approximation by MLP neural networks*, Neurocomputing, 25 (1999), pp. 81–91.
[36] R. E. Megginson, *An Introduction to Banach Space Theory*, no. 183 in Graduate Texts in Mathematics, Springer, 1998.
[37] Q. Nguyen, M. C. Mukkamala, and M. Hein, *On the loss landscape of a class of deep neural networks with no bad local valleys*, arxiv:1809.10749v2, 2018.
[38] A. Nicolae, *PLU: The piecewise linear unit activation function*, arxiv:1809.09534, 2018.
[39] S. Oymak and M. Soltanolkotabi, *Towards moderate overparameterization: global convergence guarantees for training shallow neural networks*, IEEE J. Sel. Areas Inform. Theory, 1 (2020), pp. 84–105.
[40] P. Petersen, M. Raslan, and F. Voigtlaender, *Topological properties of the set of functions generated by neural networks of fixed size*, Found. Comput. Math, 21 (2021), pp. 375–444.
[41] H. Petzka and C. Sminchisescu, *Non-attracting regions of local minima in deep and wide neural networks*, J. Mach. Learn. Res., 22 (2021), pp. 1–34.
[42] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numer., 8 (1999), pp. 143–195.
[43] I. Safran and O. Shamir, *Spurious local minima are common in two-layer ReLU neural networks*, in Proc. 35th Int. Conf. Mach. Learn., J. G. Dy and A. Krause, eds., vol. 80, PMLR, 2018, pp. 4430–4438.
[44] A. M. Saxe, J. L. McClelland, and S. Ganguli, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, arxiv:1312.6120v3, 2014.
[45] R. Sun, *Optimization for deep learning: theory and algorithms*, arxiv:1912.08957v1, 2019.
[46] R. Sun, D. Li, S. Liang, T. Ding, and R. Srikant, *The global landscape of neural networks: an overview*, IEEE Signal Process. Mag., 37 (2020), pp. 95–108.
[47] G. Swirszcz, W. M. Czarnecki, and R. Pascanu, *Local minima in training of neural networks*, arxiv:1611.06310v2, 2016.
[48] L. Venturi, A. S. Bandeira, and J. Bruna, *Spurious valleys in one-hidden-layer neural network optimization landscapes*, J. Mach. Learn. Res., 20 (2019), pp. 1–34.
[49] L. P. Vlasov, *Almost convex and Chebyshev sets*, Math. Notes Acad. Sci. USSR, 8 (1970), pp. 776–779.
[50] Y. Yoshida and M. Okada, *Data-dependence of plateau phenomenon in learning with neural network—statistical mechanical analysis*, Adv. Neur. Inform. Proc. Sys., 32 (2019).
[51] X.-H. Yu and G.-A. Chen, *On the local minima free condition of backpropagation learning*, IEEE Trans. Neural Netw., 6 (1995), pp. 1300–1303.
[52] C. Yun, S. Sra, and A. Jadbabaie, *Small nonlinearities in activation functions create bad local minima in neural networks*, arxiv:1802.03487v4, 2019. publ. as conf. paper, ICLR2019.
[53] D. Zou, P. M. Long, and Q. Gu, *On the global convergence of training deep linear ResNets*, arxiv:2003.01094v1, 2020. publ. as conf. paper, ICLR2020.