On the Stability Properties and the Optimization Landscape of Training Problems with Squared Loss for Neural Networks and General Nonlinear Conic Approximation Schemes

Constantin Christof CHRISTOF@MA.TUM.DE
Technische Universität München
Chair of Optimal Control
Center for Mathematical Sciences, M17
Boltzmannstraße 3, 85748 Garching, Germany

Abstract

We study the optimization landscape and the stability properties of training problems with squared loss for neural networks and general nonlinear conic approximation schemes. It is demonstrated that, if a nonlinear conic approximation scheme is considered that is (in an appropriately defined sense) more expressive than a classical linear approximation approach and if there exist unrealizable label vectors, then a training problem with squared loss is necessarily unstable in the sense that its solution set depends discontinuously on the label vector in the training data. We further prove that the same effects that are responsible for these instability properties are also the reason for the emergence of saddle points and spurious local minima, which may be arbitrarily far away from global solutions, and that neither the instability of the training problem nor the existence of spurious local minima can, in general, be overcome by adding a regularization term to the objective function that penalizes the size of the parameters in the approximation scheme. The latter results are shown to be true regardless of whether the assumption of realizability is satisfied or not. We demonstrate that our analysis in particular applies to training problems for free-knot interpolation schemes and deep and shallow neural networks with variable widths that involve an arbitrary mixture of various activation functions (e.g., binary, sigmoid, tanh, arctan, soft-sign, ISRU, soft-clip, SQNL, ReLU, leaky ReLU, soft-plus, bent identity, SILU, ISRLU, and ELU). In summary, the findings of this paper illustrate that the improved approximation properties of neural networks and general nonlinear conic approximation instruments are linked in a direct and quantifiable way to undesirable properties of the optimization problems that have to be solved in order to train them.

Keywords: Loss Surface, Optimization Landscape, Stability Properties, Squared Loss, Sensitivity Analysis, Neural Networks, Nonlinear Approximation, Spurious Local Minima

1. Introduction and Summary of Results

The aim of this paper is to study the stability properties and the optimization landscape of training problems of the form

$$\min_{\alpha \in D} \frac{1}{2n} \sum_{k=1}^{n} \|\psi(\alpha, x_d^k) - y_d^k\|_{\mathcal{Y}}^2. \tag{1.1}$$

Here, $\mathcal{X}$ is supposed to be a nonempty set (the set of input elements), $\mathcal{Y}$ is supposed to be a finite-dimensional vector space over $\mathbb{R}$ that is endowed with an inner product $(\cdot,\cdot)_{\mathcal{Y}}$ and the associated norm $\|\cdot\|_{\mathcal{Y}}$ (the output space), $\psi\colon D \times \mathcal{X} \to \mathcal{Y}$, $(\alpha, \chi) \mapsto y$, is assumed to be an arbitrary but fixed approximation scheme (e.g., a neural network) that can be adjusted by selecting an m-dimensional vector α from a nonempty set of admissible parameters $D \subset \mathbb{R}^m$ (these may be weights, biases, coefficients, or something else), and (χ_d^k, y_d^k), $k = 1, ..., n$, $n \in \mathbb{N}$, $n \geq 2$, is a given training set consisting of a label vector $\{y_d^k\}_{k=1}^n \in \mathcal{Y}^n$ and an input vector $\{\chi_d^k\}_{k=1}^n \in \mathcal{X}^n$. Note that, by introducing the abbreviations

$$X := \mathcal{X}^n, \qquad Y := \mathcal{Y}^n, \qquad y_d := \{y_d^k\}_{k=1}^n \in Y, \qquad x_d := \{\chi_d^k\}_{k=1}^n \in X,$$

$$\left\|\{y_k\}_{k=1}^n\right\|_Y := \left(\frac{1}{2n}\sum_{k=1}^n \|y_k\|_{\mathcal{Y}}^2\right)^{1/2} \qquad \forall \{y_k\}_{k=1}^n \in Y, \tag{1.2}$$

$$\Psi\colon D \times X \to Y, \qquad \Psi\left(\alpha, \{\chi_k\}_{k=1}^n\right) := \{\psi(\alpha, \chi_k)\}_{k=1}^n\,,$$

the problem (1.1) can also be written in the more compact form

$$\min_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2. \tag{1.3}$$

1.1 Some Background

Due to the widespread use of the quadratic loss function, minimization problems of the type (1.1) (or (1.3), respectively) are nowadays almost omnipresent in machine learning and the field of approximation theory in general. One of the main reasons why problems of the form (1.1) are considered so frequently in the literature is that solving them (or, at least, solving them approximately) by means of classical first-order methods works very well in practical applications - in particular in the context of neural networks. This has led some authors to speculate that training problems of the type (1.1) are always very well behaved when neural networks are considered, e.g., in the sense that all local minima of (1.1) are also globally optimal or achieve a loss that is very close to the optimum. Compare, for instance, with the numerical results and discussions in LeCun et al. (2015), Nguyen et al. (2019), and Yu and Chen (1995) in this context. At least for neural networks with linear activation functions, the belief that training problems of the form (1.1) always possess very nice properties turns out to be not completely unfounded. Indeed, in Kawaguchi (2016) it could be proved that, for deep linear neural networks, local minima of (1.1) are always also globally optimal so that - as far as the notion of local optimality is concerned - (1.1) effectively behaves like a convex problem. This effect was later also discussed in more detail in Zhou and Liang (2017), Laurent and von Brecht (2018), Yun et al. (2018), and Yun et al. (2019), and, with view on the convergence properties of stochastic and ordinary gradient descent algorithms, in Eftekhari (2020) and Zou et al. (2020).

Unfortunately, for truly nonlinear approximation schemes, the picture turns out to be more bleak. Although there have been numerous attempts to establish, for instance, the "local minima = global minima"-property for networks with nonlinear activation functions (mostly based on the hope that the linear case gives a good enough impression of the nonlinear one, see Eftekhari (2020); Saxe et al. (2014)), the results that have been obtained in this context so far are typically only applicable in very special situations and under rather restrictive assumptions on the network architecture, the degree of overparameterization, and/or the considered training data. Compare, for instance, with the findings of Yu and

Chen (1995); Kazemipour et al. (2019); Li and Liang (2018); Li et al. (2018); Liang et al. (2018); Oymak and Soltanolkotabi (2020); Soudry and Carmon (2016); Wu et al. (2018); Cooper (2020) in this regard. For a critical discussion of this topic and further references, see also Goldblum et al. (2020) and Ding et al. (2020). The reason behind these deficits of the known positive results on the loss surface of general neural networks is that even slightest nonlinearities in the activation function can have a huge impact on the optimization landscape of training problems of the form (1.1) and may very well give rise to spurious (i.e., not globally optimal) local minima. Data sets illustrating this for two-layer ReLU neural networks can be found, for example, in Swirszcz et al. (2016), Zhou and Liang (2017), and Safran and Shamir (2018). The minima documented in the latter of these papers have recently also been studied in more detail in Arjevani and Field (2020). Further, in Yun et al. (2018, 2019) it was shown that, again for two-layer ReLU-like networks, spurious local minima emerge for almost all choices of the training data. This illustrates that local minima that are not globally optimal are not the exception but rather the rule when piecewise linear activation functions are considered. In Yun et al. (2018, 2019), the authors also provide explicit examples of training problems for non-ReLU neural networks with two layers which possess non-globally optimal local minima. For problems involving only a single neuron, an example with numerous local minima can also be found in the early work of Auer et al. (1996). Compare also with the results on spurious valleys of Nguyen et al. (2019); Venturi et al. (2019) in this context, and, for an overview of papers on the existence of spurious local minima, with Sun (2019); Sun et al. (2020). What all of the results on the existence of spurious local minima in the above contributions have in common is that they are only concerned with networks which are rather shallow (with depth not exceeding two). The reason for this is that, as soon as more layers are considered, the explicit construction of (nontrivial) spurious local minima - or, more precisely, proving that a constructed local minimum is indeed not a global one - becomes very cumbersome. Two of the few contributions that address the construction of spurious local minima for networks of arbitrary depth are the recent Goldblum et al. (2020) and Ding et al. (2020). In both of these papers, however, a detailed discussion of the neuralgic point of whether the constructed local minima are really spurious is largely avoided. In Goldblum et al. (2020), this issue is merely addressed by providing numerical evidence, and in Ding et al. (2020) the authors resort to the assumption of realizability to resolve this problem. What is further noteworthy is that the majority of contributions on the existence of spurious local minima currently found in the literature rely on the fact that neural networks with piecewise linear activation functions are able to locally emulate a linear neural network and thus inherit the solutions of training problems of the form (1.3) for linear approximation schemes as spurious local minima. Compare, for instance, with the methods of proof used in Yun et al. (2018, 2019); Goldblum et al. (2020); Ding et al. (2020) in this context.

1.2 Summary of Main Results on General Nonlinear Approximation Schemes

The purpose of the present paper is to demonstrate that the undesirable properties of the optimization landscape of training problems of the form (1.3) for neural networks with nonlinear activation functions are, in fact, not the result of a particular choice of network architecture but rather a necessary consequence of the improved approximation properties that neural networks enjoy in comparison with linear approaches. More precisely, in what follows, we will demonstrate that indeed *every* nonlinear approximation scheme that is conic and - in an appropriately defined sense - more expressive than a linear approximation instrument (regardless of whether it is a neural network or something different, e.g., an adaptive interpolation approach) gives rise to squared-loss training problems that suffer from stability and uniqueness issues and/or the existence of non-optimal stationary points.

The starting point of our analysis is the observation that the overwhelming majority of nonlinear approximation schemes $\psi\colon D \times \mathcal{X} \to \mathcal{Y}$ currently found in the literature possess the following two properties for all $n \geq 2$ and all training data vectors $x_d \in X$ with $\chi_d^k \neq \chi_d^j$ for all $k \neq j$:

I) **(Conicity)** The set $\Psi(D, x_d)$ (with Ψ etc. defined as in (1.2)) is a cone, i.e.,

$$y \in \Psi(D, x_d),\ s \in (0, \infty) \quad \Rightarrow \quad sy \in \Psi(D, x_d).$$

II) **(Improved Expressiveness)** The map $\Psi(\cdot, x_d)\colon D \to Y$ satisfies

$$\forall y_d \in Y \setminus \{0\} : \quad \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \|y_d\|_Y^2.$$

Note that the first of the above conditions is rather unremarkable. If, for example, a neural network is considered, then this assumption is automatically satisfied since the topmost layer is affine, see Lemma 5.6. Property II) is more interesting in this context. It expresses that, for the considered data vector x_d, the function $\Psi(\cdot, x_d)\colon D \to Y$ is able to provide an approximation of every nonzero training label vector y_d that is better than the trivial guess $y = 0 \in Y$. The main point here is that the map $\Psi(\cdot, x_d)$ can accomplish this regardless of the relationship between the number of parameters $m \in \mathbb{N}$ and the number of training samples $n \in \mathbb{N}$ (and in particular also in those situations with $m \ll n$). For further details on this topic, we refer to Section 3.

For every training problem of the type (1.3) that involves an approximation scheme $\psi\colon D \times \mathcal{X} \to \mathcal{Y}$ and a training data vector $x_d \in X$ satisfying I) and II), we are able to prove the following (see the theorems in brackets for the mathematical rigorous statements):

- **(Nonuniqueness and Instability of Best Approximations)** If there exist label vectors $y_d \in Y$ that are not realizable, then the map $\Psi(\cdot, x_d)\colon D \to Y$ is always unable to provide unique best approximations for all y_d. (See Definition 2.3 for the precise definition of what we mean with the term "best approximation" here.) Further, arbitrarily small perturbations in y_d can cause arbitrarily large changes in the set of best approximations. The degree of discontinuity of the best approximation map depends on the extent to which ψ and x_d satisfy condition II). (See Theorem 4.9.)

- **(Choice Between Excessive Nonuniqueness and Spurious Local Minima)** If there exist label vectors y_d that are not realizable and if the map $\Psi(\cdot, x_d)\colon D \to Y$ is continuous, then there exist uncountably many label vectors $y_d \in Y$ for which $\Psi(\cdot, x_d)\colon D \to Y$ provides infinitely many best approximations or there exists an open nonempty cone $K \subset Y$ such that, for each $y_d \in K$, (1.3) possesses spurious local minima and/or spurious valleys. (See Theorem 4.11.)

- **(Existence of Undesirable Stationary Points)** If the map $\Psi(\cdot, x_d)\colon D \to Y$ is differentiable at a point $\bar{\alpha} \in D$ and if the function value and partial derivatives of $\Psi(\cdot, x_d)$ at $\bar{\alpha}$ do not span the whole of Y, then there exist uncountably many $y_d \in Y$ such that $\bar{\alpha}$ is an arbitrarily bad saddle point or spurious local minimum of (1.3). In particular, in the case $m + 1 < \dim(Y)$, every point of differentiability of $\Psi(\cdot, x_d)$ is a saddle or spurious local minimum of (1.3) for uncountably many y_d. The position of these y_d depends on the extent to which ψ and x_d satisfy II). (See Theorem 4.13.)
- **(Existence of Spurious Local Minima)** If $\Psi(\cdot, x_d)$ is able to locally parameterize a proper subspace V of Y, then there exists an open nonempty cone $K \subset Y$ such that (1.3) possesses spurious local minima for all $y_d \in K$. These spurious minima satisfy a growth condition in Y and can be arbitrarily bad in relative and absolute terms and in terms of loss. The size of K depends on the extent to which ψ and x_d satisfy II). If every vector is realizable, then it holds $K = Y \setminus V$. (See Theorems 4.15 and 4.16.)
- **(Instability and Nonuniqueness in the Presence of Realizability)** If there exists an $\bar{\alpha} \in D$ such that $\Psi(\cdot, x_d)$ maps an open neighborhood of $\bar{\alpha}$ into a proper subspace of Y and if every $y_d \in Y$ is realizable, then the solution set of (1.3) is instable w.r.t. perturbations of the vector y_d and (1.3) is not uniquely solvable (in the sense of minimizing sequences) for certain choices of the vector y_d. (See Theorem 4.17.)
- **(Ineffectiveness of Regularization)** If a term of the form $\nu g(\alpha)$ with a $\nu > 0$ and a regularizer $g\colon D \to [0, \infty)$ is added to the objective function of (1.3), then the following is true (under appropriate assumptions on ψ and g, see Section 4.3):
  - i) There exist uncountably many combinations of vectors y_d and parameters ν such that the resulting regularized training problem possesses spurious local minima and these can be arbitrarily bad in terms of loss. (See Theorem 4.21.)
  - ii) There exist uncountably many combinations of y_d and ν such that the resulting regularized training problem is not uniquely solvable (in the sense of minimizing sequences) and possesses a discontinuous solution map. (See Theorem 4.23.)
  - iii) Regardless of the choice of ν, adding the term $\nu g(\alpha)$ to the objective function of (1.3) compromises the approximation property II). (See Theorem 4.22.)

Before we comment in more detail on how the above results are related to the literature and on the overall contribution of this paper, we briefly summarize the consequences that our analysis has for the study of neural networks.

1.3 Summary of Consequences for Deep and Shallow Neural Networks

Our first main result on neural networks establishes that these special instances of nonlinear approximation schemes are indeed covered by our abstract analysis:

- **(Conicity and Improved Expressiveness of Neural Networks)** Consider a fully connected feedforward neural network with input space $\mathcal{X} = \mathbb{R}^{d_x}$, output space $\mathcal{Y} = \mathbb{R}^{d_y}$, $d_\mathcal{X}, d_y \in \mathbb{N}$, depth $L \in \mathbb{N}$, widths $w_1, ..., w_L \in \mathbb{N}$, and activation functions $\sigma_i\colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L$. Suppose that an $n \in \mathbb{N}$ with $n \geq 2$ and an $x_d := \{\chi_d^k\}_{k=1}^n \in \mathcal{X}^n$ satisfying $\chi_d^j \neq \chi_d^k$ for all $j \neq k$ is given, and that one of the following is true:

i) The functions σ_i are of Heaviside type for all $i = 1, ..., L$ and it holds $w_1 \geq 2$.

ii) The set $\{1, ..., L\}$ can be decomposed into two (possibly empty) disjoint index sets I and J such that the function σ_i is of "sigmoid type" (e.g., sigmoid, tanh, arctan, soft-sign) for all $i \in I$, such that the function σ_i is of "ReLU type" (e.g., ReLU, soft-plus, swish) for all $i \in J$, such that $w_i \geq 2$ holds for all $i \in J$, and such that $w_1 \geq 2$ holds in the case $1 \in I$ and $w_1 \geq 4$ in the case $1 \in J$.

Then, the neural network and x_d satisfy I) and II). (See Lemma 5.8 and Theorem 5.9.)

We remark that the fact that neural networks indeed possess the approximation property II) under the above weak assumptions on the data, the network architecture, and the activation functions is also interesting on its own. (See Assumption 5.5 for the precise setting and Lemma 5.8 and Theorem 5.9 for an explanation of what we mean with the terms "sigmoid type", "ReLU type", and "Heaviside type" here.) We will comment in more detail on this topic in Section 1.4. For a result that shows that our analysis also covers ResNets, we refer the reader to Corollary 5.22. Given a training problem of the type (1.3) for a neural network and a vector x_d that satisfy the conditions in the last result, we obtain, for instance, the following corollaries from our abstract analysis (see again the results in brackets for the mathematical rigorous statements):

- **(Nonuniqueness and Instability of Best Approximations)** If there exist unrealizable label vectors y_d, then the neural network is unable to provide unique best approximations for all $y_d \in Y$. Further, arbitrarily small perturbations of the label vector y_d can affect the set of best approximations to an arbitrarily large extent. The degree of discontinuity of the best approximation map depends on the extent to which x_d and the considered network satisfy II). (See Corollary 5.13.)

- **(Choice Between Excessive Nonuniqueness and Spurious Local Minima)** If there exist unrealizable label vectors y_d and if the activation functions σ_i, $i = 1, ..., L$, are continuous, then there exist uncountably many $y_d \in Y$ for which the neural network provides infinitely many best approximations or there exists an open nonempty cone $K \subset Y$ such that, for each $y_d \in K$, the training problem (1.3) possesses (arbitrarily bad) spurious local minima and/or spurious valleys. (See Corollary 5.14.)

- **(Saddle Points and Spurious Minima in the Non-Overparameterized Case)** If the number of parameters m in the neural network is smaller than the product nd_y, then every point of differentiability of the neural network is a saddle point or a spurious local minimum of (1.3) for uncountably many y_d and, as a saddle point or spurious local minimum, can be made arbitrarily bad in relative and absolute terms and in terms of loss by choosing y_d appropriately. The position of these y_d depends on the extent to which x_d and the considered network satisfy II). (See Corollary 5.15.)

- **(Saddle Points and Spurious Minima for Arbitrary Problems)** If $d_\chi + 1 < n$ holds and if the functions σ_i are differentiable, then there exists an $(m - d_\chi w_1)$-dimensional subspace of the parameter space of the network such that each element of this subspace is a saddle point or a spurious local minimum of (1.3) for uncountably many y_d. Again, these points can be made arbitrarily bad in relative and absolute terms and in terms of loss by choosing appropriate y_d. (See Corollary 5.16.)

- **(Spurious Local Minima for Activation Functions with an Affine Segment)** If each σ_i is affine-linear on some open nonempty interval $I_i \subset \mathbb{R}$ of its domain of definition and if it holds $d_\mathcal{X} + 1 < n$ and $\min(d_\mathcal{X}, d_\mathcal{Y}) \leq \min(w_1, ..., w_L)$, then there exists an open, nonempty cone $K \subset Y$ such that (1.3) possesses a spurious local minimum for each $y_d \in K$. The size of this cone depends on the extent to which x_d and the neural network ψ satisfy II). If every vector is realizable, then the cone K is dense in Y and the solution map of (1.3) is discontinuous. Further, by choosing appropriate y_d, the spurious local minima can be made arbitrarily bad in relative and absolute terms and in terms of loss. (See Corollaries 5.17, 5.18 and 5.20.)

- **(Ineffectiveness of Regularization for Differentiable Activation Functions)** If the activation functions σ_i are twice differentiable, if $\frac{1}{2}(d_\mathcal{X}+2)(d_\mathcal{X}+1) < n$ holds, and if the training problem (1.3) is regularized by adding a term of the form $\nu\|\alpha\|_p^p$, $\nu > 0$, $p \in [1, 2]$, to the objective function, where $\|\cdot\|_p$ denotes the p-norm of the Euclidean space, then there exist uncountably many combinations of label vectors $y_d \in Y$ and regularization parameters $\nu > 0$ such that the resulting regularized training problem possesses an (arbitrarily bad) spurious local minimum, and there exist uncountably many values of $\nu > 0$ such that the regularized training problem is not uniquely solvable and possesses a discontinuous solution map. (See Corollary 5.21.)

Note that the set-valuedness and the instability of the best approximation map in points one and two above immediately carry over to the solution operator of the problem (1.3) w.r.t. α (just by taking preimages under the function $\alpha \mapsto \Psi(\alpha, x_d)$). For details on this topic, see the comments after Lemma 2.4 and Remark 4.10. We further would like to stress that the nonuniqueness of best approximations in, e.g., Corollary 5.13 has nothing to do with symmetries in the parameterization of a neural network. On the contrary, it expresses that there are different choices of the biases and weights (or, at least, minimizing sequences) which yield the same optimal loss in (1.3) but give rise to functions $\psi(\alpha, \cdot)\colon \mathcal{X} \to \mathcal{Y}$ that act differently not only on unseen data but even on the training data set (see Remark 4.10 for more details). Regarding Corollaries 5.15 and 5.16, we would like to point out that the fact that the saddle points and spurious local minima of (1.3) can be made arbitrarily bad is not merely a consequence of the conicity property I). Indeed, it is easy to check that simply scaling the involved vectors cannot affect how well a non-optimal point performs in relative terms, cf. Remark 4.14. Lastly, we would like to emphasize that our abstract analysis is not only applicable to deep and shallow neural networks but also to other nonlinear conic approximation instruments. For an example demonstrating this, we refer the reader to Section 5.1 where our results are applied to a free-knot spline interpolation scheme that has also been considered in Daubechies et al. (2019) and can be interpreted as a classical nonlinear dictionary approximation approach, cf. DeVore (1998). We only summarize the consequences of our abstract analysis for neural networks in this subsection because we expect that this is what the majority of readers are interested in.

1.4 Contribution of the Paper and Relations to Known Results

The main contribution of this paper is that it establishes a direct and quantifiable connection between the improved approximation properties that nonlinear approximation schemes like neural networks enjoy over their linear counterparts (i.e., property II)) and undesirable properties of the optimization problems that have to be solved in order to train a nonlinear approximation instrument on a given data set. Compare, for instance, with the estimates (4.14), (4.24), and (4.29) in this context, which show that the degree of discontinuity of the best approximation map of a given nonlinear conic approximation scheme ψ, the position of the label vectors y_d that cause a given point to be a saddle point or a spurious local minimum in Theorem 4.13, and the size of the cone of "bad" label vectors in Theorem 4.16 depend directly on the extent to which the considered approximation instrument satisfies condition II). At least to the best of the author's knowledge, this relationship between the expressiveness of an approximation scheme and the loss landscape of the associated training problems has not been explored systematically so far in the literature (although it is, of course, closely related to classical topics of nonlinear approximation theory and the study of nonlinear least-squares problems). Note that the results of this paper can be interpreted as an instance of the well-known fact that there is "no free lunch" as they show that the improved approximation properties of, e.g., neural networks come at the price that the associated training problems are always potentially ill-posed and possess spurious local minima or saddle points for certain choices of the training data. For further details on this topic and its relationship to the curse of dimensionality, see also Section 6.

We would like to emphasize that the connections that we draw in this paper are not only interesting for their own sake but also allow to improve and complement known results on the optimization landscape of training problems with squared loss found in the literature. By exploiting the approximation property II), for example, we are able to show that the assumption of realizability used in (Ding et al., 2020, Corollary 2) to establish that certain local minima are not globally optimal is unnecessary, that the conditions on the network widths in (Ding et al., 2020, Assumption 3) can be relaxed, and that the observations made in the numerical experiments of Goldblum et al. (2020) can also be backed up analytically, cf. Corollaries 5.14, 5.17 and 5.18. The main point in this context is that the approximation property II) allows to prove that a point $\bar{\alpha} \in D$ is a spurious local minimum of a problem of the type (1.3) without the explicit construction of a parameter $\tilde{\alpha} \in D$ that yields a smaller loss than $\bar{\alpha}$. This makes the rather cumbersome calculations that are normally used to establish that a local minimum is not globally optimal unnecessary, cf. the proofs of Theorems 4.13 and 4.15 and also the comments in the proof of (Yun et al., 2019, Theorem 1) where it is emphasized that proving the existence of points with smaller function values is precisely the hard part of constructing examples of spurious local minima. Since our approach does not require explicit constructions, we are also able to rigorously prove the existence of spurious local minima in situations in which the classical approach of manually checking the spuriousness of a local minimum becomes intractable due to the presence of additional regularization terms or the architecture of the considered nonlinear approximation scheme. Compare in particular with Corollaries 5.17, 5.18 and 5.21 in this context, which establish the existence of spurious local minima for both unregularized and regularized training problems and for neural networks with arbitrary depth and various activation functions. At least to the best of the author's knowledge, results on the existence of spurious local minima with a comparable generality can currently not be found in the literature. In particular, the existence of spurious local minima in Tikhonov-regularized problems for deep networks has apparently not been considered so far. Note that our approach additionally offers the advantage that it allows to establish that saddle points and spurious local minima of training problems with squared loss can be arbitrarily far away from global optima in relative and absolute terms and in terms of loss, see Lemma 4.12 and Theorems 4.13, 4.15 and 4.21 and the associated corollaries on neural networks in Section 5.

As already mentioned, by exploiting the approximation property II), we are also able to rigorously prove that solutions of training problems of the form (1.3) (or the associated best approximations, respectively) cannot be expected to be unique or stable with respect to perturbations of the training label vector y_d, see Theorems 4.9, 4.17 and 4.23. This gives an analytic explanation for the instability effects that are commonly observed in network training, cf. Cunningham et al. (2000) and also the comments on the nonuniqueness of global solutions in (Cooper, 2020, Section 1.1). We remark that, for neural networks with one hidden layer, instability results similar to that in our Theorem 4.9 have already been proved in the L^p-spaces in Kainen et al. (1999, 2001) by exploiting classical instruments of nonlinear approximation theory. The finite-dimensionality of the training problem (1.3) allows us to go further than the authors of these papers and to establish the nonuniqueness and instability of solutions and best approximations for neural networks of arbitrary depth. By exploiting the inequality of Jung (see (Burago and Zalgaller, 1988, Theorem 11.1.1)), we are further able to establish a quantifiable connection between the discontinuity properties of the best approximation map associated with (1.3) and the extent to which an approximation scheme satisfies II), cf. the estimate (4.14). The results that we prove in this context also seem to be new.

We would like to point out that, for deep and shallow neural networks whose activation functions are affine-linear on some open nonempty interval of their domain of definition, our results give a quite complete picture of how the optimization landscape of problems of the form (1.3) depends on the approximation property II) or, more precisely, on the number $\Theta(\Psi, x_d)$ defined in (4.2) that measures the extent to which property II) is satisfied. In the case $\Theta(\Psi, x_d) \in (0, 1)$ (which corresponds to the situation where there are unrealizable label vectors), one has to deal with both the instability of the set of best approximations of (1.3) and the existence of an open nonempty cone K of vectors y_d for which (1.3) possesses (potentially arbitrarily bad) spurious local minima (see Corollaries 5.13, 5.17 and 5.18). The closer $\Theta(\Psi, x_d)$ gets to zero (i.e., the more expressive the network becomes relative to Y, e.g., due to an increased number of network parameters or a smaller number of training pairs), the less pronounced the instability properties of the best approximation map of (1.3) are (see (4.14)) and the larger the cone K grows (see (4.29)). Finally, in the case $\Theta(\Psi, x_d) = 0$ (i.e., the case where every vector is realizable, cf. Definition 4.3), the instability properties of the best approximation map are not present anymore and the cone K is dense in Y so that, for almost all y_d, (1.3) possesses spurious local minima. In summary, the above shows that, when considering problems of the type (1.3) for a network satisfying the assumptions of Corollary 5.17 or Corollary 5.18 or, more generally, a nonlinear conic approximation instrument satisfying the conditions in Theorem 4.16, one can never get rid of both the discontinuity of the best approximation map and spurious local minima. The problem (1.3) always possesses at least one property that is undesirable (cf. also with Theorem 4.11 in this context). Note that the fact that the instability properties of the best approximation map associated with (1.3) are not present when every vector $y_d \in Y$ is realizable provides a possible explanation for the often made observation that overparameterization benefits the training of neural networks in practical applications. Compare, e.g., with the results in Chen et al. (2020); Cooper (2020); Li and Liang (2018); Oymak and Soltanolkotabi (2020); Allen-Zhu et al. (2019); Soudry and Carmon (2016) in this context. However, it seems to be unlikely that this is the only reason for the advantageous properties that overparameterized training problems typically enjoy. In fact, we will see in Theorems 4.17 and 4.23 that, even in the case where every vector $y_d \in Y$ is realizable and the objective contains additional regularization terms, there are still certain nonuniqueness and instability effects present in problems of the form (1.3). (These, however, are of a different quality than those arising from the nonuniqueness of best approximations in Theorem 4.9.) Note that the observation that neither by overparameterization nor by adding regularization terms to the objective function it is possible to completely remove the ill-posedness of training problems of the type (1.3) is also remarkable on its own. These results also seem to be new.

Regarding the application of our abstract analysis to neural networks, we would like to stress that the fact that these special instances of nonlinear approximation schemes indeed satisfy the condition II) under the weak assumptions of Lemma 5.8 and Theorem 5.9 is also interesting independently of the study of the loss landscape of training problems of the form (1.3). As we will see in Section 3, the property II) is a characteristic that distinguishes neural networks clearly from linear approximation schemes (e.g., polynomial approximation) and thus gives an idea of why these approximation instruments are able to outperform classical approaches. Compare also with Lemma 4.5 in this context which establishes that the property II) is directly related to worst-case estimates for the approximation error that nonlinear approximation schemes achieve for arbitrary training label vectors y_d. We also would like to emphasize at this point that II) is a global property of an approximation scheme and thus of a completely different flavor than, e.g., the local properties of activation functions (for instance, piecewise linearity) that are commonly worked with in the analysis of neural networks. This also becomes apparent in the proof of Theorem 5.9 which, in contrast to many classical approaches, is not based on concepts like linearization but on the observation that the overwhelming majority of neural networks used in practice are able to emulate networks with binary activation functions by saturation and that the property II) is inherited from these binary networks obtained in the saturation limit. Further details on this topic can be found in Section 5.

Before we begin with our analysis, we finally would like to emphasize that the theorems proved in this paper do not contradict the results on the absence of spurious local minima in training problems for neural networks with linear activation functions established, e.g., in Kawaguchi (2016); Zhou and Liang (2017); Laurent and von Brecht (2018). Since such networks parameterize a subspace, they only satisfy condition II) in pathological situations and thus do not fall under the scope of, e.g., Theorems 4.11, 4.15 and 4.16. Compare again with the example in Section 3 in this context. Similarly, our theorems also do not contradict the results on the absence of spurious valleys established in Nguyen et al. (2019) and Venturi et al. (2019) (simply because we are mainly concerned with classical spurious local minima in this work, cf. Definition 2.5). They are, however, in good accordance with the observations on the role and presence of saddle points in network training made, e.g., in Dauphin et al. (2014). For further details on this topic and additional remarks on the relationship between our results and the literature, we refer the reader to the comments after the respective theorems in the subsequent sections.

1.5 Overview of the Structure of the Paper

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

Section 2 is concerned with preliminaries. Here, we collect our standing assumptions on the quantities in the training problem (1.1) and introduce some notations and basic concepts that are needed for our analysis (e.g., the definition of the best approximation map appearing in points one and two of Section 1.2, see Definition 2.3).

In Section 3, we discuss a toy example that illustrates the basic ideas of our analysis and provides some intuition on how the approximation property II) is related to the loss landscape of training problems of the form (1.3). The example considered in this section also demonstrates that the property II) indeed distinguishes (truly) nonlinear approximation schemes from classical linear approximation approaches.

The subsequent Section 4 contains the bulk of our analysis of the optimization landscape and the stability properties of training problems with squared loss for general nonlinear approximation schemes satisfying the conditions I) and II). This section is subdivided into three parts. The first one, Section 4.1, is concerned with the discontinuity and set-valuedness of the best approximation map associated with (1.3) in the situation where there exist unrealizable vectors, i.e., in the case where the closure of the image $\Psi(D, x_d)$ of the function $\Psi(\cdot, x_d)\colon D \to Y$ is not the whole of Y. Here, we prove the first of the main results presented in Section 1.2, see Theorem 4.9. Note that this subsection also contains some more detailed comments on the relationship between our results and classical topics in nonlinear approximation theory, cf. Proposition 4.7. In the second part of Section 4, Section 4.2, we derive the results on the existence of spurious local minima and saddle points described in points two to five of Section 1.2, see Theorems 4.11, 4.13 and 4.15 to 4.17. We would like to emphasize that Theorems 4.11 and 4.17 are the only results in this subsection that require assumptions on the existence of unrealizable vectors. The remaining theorems in Section 4.2 are completely independent of whether the assumption of realizability is satisfied or not (and thus also on, e.g., the degree of overparameterization in (1.3)). The third part of Section 4, Section 4.3, addresses the well-posedness of training problems that involve additional regularization terms. Here, we prove the results on the existence of spurious local minima and the nonuniqueness and instability of solutions presented in point six of Section 1.2. Note that the analysis in this section requires additional assumptions on the properties of the nonlinear approximation scheme ψ, namely that it has a linear lowest level, see Assumption 4.18. This condition is, e.g., satisfied by neural networks.

Section 5 addresses the consequences that the analysis of Section 4 has for special instances of nonlinear conic approximation instruments. The first example that we consider in this section is a classical free-knot interpolation scheme that has also been studied in Daubechies et al. (2019), see Section 5.1. Second, we apply our abstract results to deep and shallow neural networks that involve variable widths and various activation functions, see Section 5.2 and Assumption 5.5. Note that this section in particular establishes that these special examples of approximation instruments indeed satisfy the conditions I) and II), see Lemma 5.2, Theorem 5.9, and Corollary 5.10. For the precise statements and proofs of the results on neural networks that we have presented in Section 1.3, we refer the reader to Corollaries 5.13 to 5.18, 5.20 and 5.21.

In Section 6, we conclude this paper with some additional remarks on the overall role that our results play in the study of neural networks and the field of nonlinear approximation theory in general. Here, we in particular comment on relations to the plateau phenomenon in gradient training and connections to the curse of dimensionality.

2. Notation, Setting, and Basic Concepts

Before we begin with our analysis, let us fix the notation and introduce some basic concepts. As already mentioned in the introduction, the main focus of this work will be on training problems of the form

$$\min_{\alpha \in D} \frac{1}{2n} \sum_{k=1}^{n} \|\psi(\alpha, \chi_d^k) - y_d^k\|_{\mathcal{Y}}^2. \tag{2.1}$$

For easy reference, we restate our assumptions on the quantities in (2.1) in:

Assumption 2.1 (Standing Assumptions and Notation for the Study of (2.1))

- *$\mathcal{X}$ is a nonempty set,*
- *$\mathcal{Y}$ is a finite-dimensional vector space over $\mathbb{R}$ that is endowed with an inner product $(\cdot,\cdot)_{\mathcal{Y}}$ and the associated norm $\|\cdot\|_{\mathcal{Y}}$ (i.e., $\|y\|_{\mathcal{Y}} := (y,y)_{\mathcal{Y}}^{1/2}$ for all $y \in \mathcal{Y}$),*
- *$m, n \in \mathbb{N}$ and $n \geq 2$,*
- *$D \subset \mathbb{R}^m$ is a nonempty set,*
- *$\psi\colon D \times \mathcal{X} \to \mathcal{Y}$ is a function (representing an approximation scheme),*
- *$\{\chi_d^k\}_{k=1}^n \in \mathcal{X}^n$, $\{y_d^k\}_{k=1}^n \in \mathcal{Y}^n$ is the training data.*

Similarly, we collect the abbreviations in (1.2) in:

Definition 2.2 (Some Abbreviations) *In the situation of Assumption 2.1, we define:*

- *X to be the Cartesian product $X := \mathcal{X}^n$,*
- *Y to be the Hilbert space $Y := \mathcal{Y}^n$ endowed with the product*

$$\left(\{y_k\}_{k=1}^n, \{z_k\}_{k=1}^n\right)_Y := \frac{1}{2n} \sum_{k=1}^{n} \left(y_k, z_k\right)_{\mathcal{Y}} \qquad \forall \{y_k\}_{k=1}^n, \{z_k\}_{k=1}^n \in Y$$

  and the associated norm $\|\cdot\|_Y$ (cf. (1.2)),
- *y_d to be the vector $y_d := \{y_d^k\}_{k=1}^n \in Y$,*
- *x_d to be the vector $x_d := \{\chi_d^k\}_{k=1}^n \in X$,*
- *Ψ to be the map*

$$\Psi\colon D \times X \to Y, \qquad (\alpha, \{\chi_k\}_{k=1}^n) \mapsto \{\psi(\alpha, \chi_k)\}_{k=1}^n. \tag{2.2}$$

We would like to emphasize that, here and in what follows, we always think of elements of the space $\mathbb{R}^m$ as column vectors. As already pointed out in Section 1, the abbreviations in Definition 2.2 allow us to restate the problem (2.1) in the more compact form

$$\min_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2. \tag{2.3}$$

Note that, since the objective function of (2.3) is not necessarily coercive w.r.t. α, it can, in general, not be expected that (2.3) possesses a global minimizer $\bar{\alpha} \in D$. One can only guarantee that there exists a minimizing sequence $\{\alpha_i\} \subset D$, i.e., a sequence satisfying

$$\|\Psi(\alpha_i, x_d) - y_d\|_Y^2 \to \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \tag{2.4}$$

as i tends to infinity. (This is, for example, the case when some of the activation functions in a neural network have to saturate to fit a training vector $y_d \in Y$ precisely.) To get a grip on these effects, it makes sense to not only study local and global minimizers $\bar{\alpha} \in D$ of (2.3) but also the set of all elements of Y that can be approximated by the function $\Psi(\cdot, x_d)$ for a given x_d and fit a training label vector y_d in an optimal manner. This gives rise to:

Definition 2.3 (Best Approximation Map) *Let $x_d \in X$ be arbitrary but fixed and let $\Psi\colon D \times X \to Y$ etc. be as before. Then, we define $P_\Psi^{x_d}$ to be the map*

$$P_\Psi^{x_d}\colon Y \rightrightarrows Y, \qquad y_d \mapsto \operatorname*{arg\,min}_{y \in \mathrm{cl}_Y(\Psi(D, x_d))} \|y - y_d\|_Y^2. \tag{2.5}$$

Here, the symbol $\rightrightarrows$ expresses that the function $P_\Psi^{x_d}$ may be set-valued and with $\mathrm{cl}_Y(\cdot)$ we denote the topological closure of a set in Y.

Note that the map $P_\Psi^{x_d}$ is precisely the set-valued metric projection in Y onto the closure of the image $\Psi(D, x_d)$ of D under the function $\Psi(\cdot, x_d)\colon D \to Y$ for the given vector x_d. Because of this, we in particular have:

Lemma 2.4 (Properties of the Set of Best Approximations) *Suppose that $x_d \in X$ and $\psi\colon D \times \mathcal{X} \to \mathcal{Y}$ are arbitrary but fixed. Then, the set $P_\Psi^{x_d}(y_d)$ is nonempty and compact for every training label vector $y_d \in Y$.*

Proof The nonemptyness and compactness of $P_\Psi^{x_d}(y_d)$ for all $y_d \in Y$ follow immediately from the fact that the minimization problem in the variable y associated with the right-hand side of (2.5) possesses a nonempty, closed, and bounded set of solutions for all $y_d \in Y$ due to the continuity and coercivity properties of the norm, the closedness and nonemptyness of the set $\mathrm{cl}_Y(\Psi(D, x_d))$, the finite-dimensionality of Y, and the theorem of Weierstrass. ■

We would like to point out that, by taking preimages and images under the function $\Psi(\cdot, x_d)\colon D \to Y$, properties of the map $P_\Psi^{x_d}$ directly translate into properties of the optimization landscape of (2.3) and vice versa. If, for example, x_d and y_d are vectors such that $P_\Psi^{x_d}(y_d) = \{\bar{y}_1, \bar{y}_2\}$ holds for some $\bar{y}_1 \neq \bar{y}_2$ and if we denote the closed balls in Y of radius ε around $\bar{y}_i$, $i = 1, 2$, with $B_\varepsilon^Y(\bar{y}_i)$, then, for every arbitrary but fixed $\varepsilon > 0$ with $B_\varepsilon^Y(\bar{y}_1) \cap B_\varepsilon^Y(\bar{y}_2) = \emptyset$, we trivially have that the preimages

$$D_1 := \Psi(\cdot, x_d)^{-1}\left(B_\varepsilon^Y(\bar{y}_1)\right) \subset D \quad \text{and} \quad D_2 := \Psi(\cdot, x_d)^{-1}\left(B_\varepsilon^Y(\bar{y}_2)\right) \subset D$$

satisfy $D_1 \neq \emptyset$, $D_2 \neq \emptyset$, $D_1 \cap D_2 = \emptyset$, and

$$\inf_{\alpha \in D_1} \|\Psi(\alpha, x_d) - y_d\|_Y^2 = \inf_{\alpha \in D_2} \|\Psi(\alpha, x_d) - y_d\|_Y^2 = \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2.$$

The above implies that each of the two disjoint subsets D_1 and D_2 of the parameter space D has to contain a global solution of the problem (2.3) or a sequence $\{\alpha_i\}$ satisfying (2.4). Note that the main advantage of considering the projection $P_\Psi^{x_d}$ instead of the objective $D \ni \alpha \mapsto \|\Psi(\alpha, x_d) - y_d\|_Y^2 \in \mathbb{R}$ of (2.3) is that the former function allows to also detect those cases where (2.3) possesses spurious local minima "at infinity" in the sense that the optimization landscape of (2.3) possesses basins which stretch to the boundary of the parameter space D and do not contain a local minimum in the classical sense. Compare, e.g., with the behavior of the function $\mathbb{R} \ni \alpha \mapsto \min(\mathrm{e}^\alpha, \mathrm{e}^{-\alpha} - 1) \in \mathbb{R}$ in this context and also with Theorem 4.11. Such cases should, of course, not be neglected as descent methods may very well get trapped in a non-optimal basin of this type and subsequently drive the parameter α to the boundary of the set D without approximating the optimal value of the loss function on the right-hand side of (2.3) in the limit. Completely analogously to the above, stability and instability properties of $P_\Psi^{x_d}$ carry over to (2.3), too. For further details on this topic, we refer the reader to Remark 4.10.

For the sake of clarity, let us finally make precise what we mean with the terms "global minimum", "local minimum", "spurious local minimum", etc. appearing in our analysis:

Definition 2.5 (Notions of Optimality) *Given a function $f\colon U \to \mathbb{R}$ that is defined on a subset U of a normed space $(V, \|\cdot\|_V)$, we call a point $\bar{v} \in U$ a:*

- *global minimum (or, more precisely, global minimizer) of the function f if $f(v) \geq f(\bar{v})$ holds for all $v \in U$.*
- *local minimum (or, more precisely, local minimizer) of the function f if there exists a closed ball $B_\varepsilon^V(\bar{v})$ of radius $\varepsilon > 0$ in V centered at $\bar{v}$ such that $f(v) \geq f(\bar{v})$ holds for all $v \in U \cap B_\varepsilon^V(\bar{v})$.*
- *spurious local minimum of f if $\bar{v}$ is a local minimum but not a global minimum of f.*
- *global (respectively, local, respectively, spurious local) maximum of f if $\bar{v}$ is a global (respectively, local, respectively, spurious local) minimum of the function $-f$.*
- *saddle point of f if $V = \mathbb{R}^l$ holds for some $l \in \mathbb{N}$, $\bar{v}$ is an element of the interior of U, f is differentiable at $\bar{v}$, it holds $\nabla f(\bar{v}) = 0$, and $\bar{v}$ is neither a local minimum nor a local maximum of f.*

We remark that some authors apparently go so far as to call every point with a vanishing gradient and a vanishing Hessian a spurious local minimum. We believe that the term "spurious local minimum" should be reserved for points that are local minima. Finally, we would like to emphasize that, throughout this work, the symbols min, arg min, etc. always refer to the global notion of optimality (e.g., in the definition of the map $P_\Psi^{x_d}$ in (2.5)).

3. A Toy Problem and Some Intuition

Having introduced the necessary notation, we now turn our attention to the optimization landscape and the stability properties of training problems of the form (2.3). We begin with a simple example that illustrates the main ideas of our analysis and gives some intuition on why nonlinear approximation schemes may possess better approximation properties than their linear counterparts and on how these properties are related to the behavior of the function $P_\Psi^{x_d}$ and the loss landscape of (2.3). To construct our example, let us suppose that

$$\mathcal{X} = \mathcal{Y} = \mathbb{R}, \quad n \in \mathbb{N}, \quad n \geq 3, \quad m = 2, \quad X = Y = \mathbb{R}^n, \quad \text{and} \quad D = \mathbb{R}^2, \tag{3.1}$$

and that $x_d = \{\chi_d^k\}_{k=1}^n \in X$ is an arbitrary but fixed training data vector which satisfies $\chi_d^1 < \chi_d^2 < ... < \chi_d^n$. Let us further assume, for a start, that we are given an approximation scheme $\psi\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, $(\alpha, \chi) \mapsto y$, that is linear in the sense that the function ψ is linear in the parameter vector α. Then, we trivially have

$$\Psi(\alpha, x_d) = \alpha_1 \Psi(e_1, x_d) + \alpha_2 \Psi(e_2, x_d),$$

where e_1, e_2 denote the standard basis vectors of $\mathbb{R}^2$, and the training problem (2.3) can also be written as

$$\min_{(\alpha_1,\alpha_2)\in\mathbb{R}^2} \|\alpha_1 \Phi_1 + \alpha_2 \Phi_2 - y_d\|_Y^2 \tag{3.2}$$

for every arbitrary but fixed $y_d = \{y_d^k\}_{k=1}^n \in Y$, where $\|\cdot\|_Y$ is the Euclidean norm on $Y = \mathbb{R}^n$ scaled with the factor $1/\sqrt{2n}$ and where $\Phi_j := \Psi(e_j, x_d) \in \mathbb{R}^n$, $j = 1, 2$. For a linear scheme ψ, (2.3) thus boils down to a standard approximation problem which aims to find a function $\phi\colon \mathbb{R} \to \mathbb{R}$ in the linear subspace spanned by the set $\{\chi \mapsto \psi(e_1, \chi),\ \chi \mapsto \psi(e_2, \chi)\}$ that fits the n given function values $y_d^k \in \mathbb{R}$ at the locations $\chi_d^k \in \mathbb{R}$ optimally in the least-squares sense. Note that the structure of (3.2) in particular implies that, regardless of which linear scheme $\psi\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ we consider here, there are always nontrivial choices of y_d for which the problem (3.2) possesses the optimal solution $\bar{\alpha} = (0, 0)^T$ so that ψ does not provide an approximation of y_d that is better than the trivial guess $\bar{y} = 0$. Indeed, for all y_d in the orthogonal complement of the space $\operatorname{span}\{\Phi_1, \Phi_2\} \subset \mathbb{R}^n$ w.r.t. the Euclidean scalar product, we clearly have

$$\operatorname*{arg\,min}_{(\alpha_1,\alpha_2)\in\mathbb{R}^2} \|\alpha_1 \Phi_1 + \alpha_2 \Phi_2 - y_d\|_Y^2 = \operatorname*{arg\,min}_{(\alpha_1,\alpha_2)\in\mathbb{R}^2} \|\alpha_1 \Phi_1 + \alpha_2 \Phi_2\|_Y^2 + \|y_d\|_Y^2 \supset \{0\}.$$

Using the notation in Definition 2.2, this observation can also be expressed in the more compact form

$$\exists y_d \in Y \setminus \{0\} : \qquad \inf_{\alpha\in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 = \|y_d\|_Y^2. \tag{3.3}$$

For comparison, let us now consider the nonlinear approximation scheme given by

$$\psi\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}, \quad (\alpha, \chi) \mapsto \min(0, |\alpha_1 \chi + \alpha_2 + 1| - 1) + \max(0, 1 - |\alpha_1 \chi + \alpha_2 - 1|) \tag{3.4}$$

and define $\bar{\alpha}_{l,\delta} \in \mathbb{R}^2$, $l \in \{1, ..., n\}$, $|\delta| \leq 1$, by

$$\bar{\alpha}_{l,\delta}^1 := 3 \left(\min_{k=2,...,n} \chi_d^k - \chi_d^{k-1} \right)^{-1}, \qquad \bar{\alpha}_{l,\delta}^2 := -\bar{\alpha}_{l,\delta}^1 \chi_d^l + \delta.$$

Then, from the properties of ψ, it follows straightforwardly that

$$\psi(\bar{\alpha}_{l,\delta}, x_d^k) = \begin{cases} \delta & \text{for } k = l \\ 0 & \text{for all } k \neq l \end{cases}$$

holds for all δ with $|\delta| \leq 1$. This implies in particular that, for every arbitrary but fixed label vector $y_d \in \mathbb{R}^n \setminus \{0\}$, there exists a parameter $\bar{\alpha} \in D$ such that the nonlinear approximation scheme (3.4) satisfies

$$\|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 < \|y_d\|_Y^2,$$

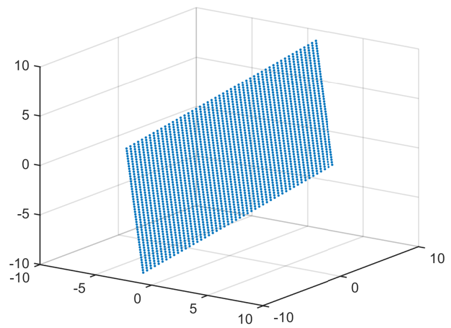
namely, in the case $y_d^l \neq 0$, the vector $\bar{\alpha}_{l,\delta}$ with $\delta := \operatorname{sgn}(y_d^l)\min(|y_d^l|, 1)$. In short,

$$\forall y_d \in Y \setminus \{0\} : \qquad \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \|y_d\|_Y^2. \tag{3.5}$$

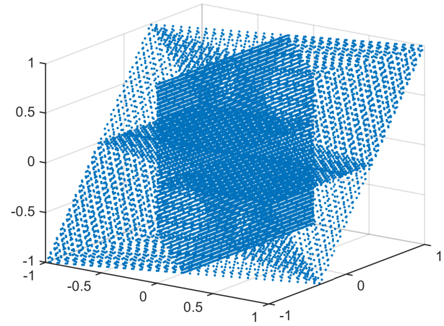
The above result shows that the nonlinearity of the function ψ in (3.4) indeed allows this map to possess better approximation properties than the linear schemes considered at the beginning of this section in the sense that, for every arbitrary but fixed nonzero $y_d \in \mathbb{R}^n$, we can find an $\alpha \in D$ such that $\Psi(\alpha, x_d)$ provides a loss that is smaller than that of the trivial guess $\alpha = 0$ and the associated vector $\Psi(0, x_d) = 0$. The map ψ in (3.4) is thus able to approximate every given label vector y_d at least to a small extent even in those situations where the problem (2.3) is grossly underparameterized, i.e., satisfies $m \ll n$ - a property that is not obtainable with a scheme that is linear in α and possesses the parameter space $D = \mathbb{R}^2$ as we have seen in (3.3).

However, the example (3.4) also immediately shows that the property (3.5) does not come for free. If we consider, for instance, the image of the parameter space $D = \mathbb{R}^2$ under the function $\Psi(\cdot, x_d)$ associated with the nonlinear approximation scheme in (3.4) in the case $n = 3$ for the training data vector $x_d = (-1/2, 1/2, 1)^T$, then it is readily seen that this set is a nontrivial union of numerous segments of two-dimensional subspaces, cf. Fig. 1. This implies in particular that the projection $P_\Psi^{x_d}$ onto $\operatorname{cl}_Y(\Psi(D, x_d))$ is not single-valued at all points and, as a consequence, that the best approximating element provided by $\operatorname{cl}_Y(\Psi(D, x_d))$ is not uniquely determined for all possible choices of the training label vector y_d. It is moreover easy to check that the locally affine-linear structure of the set $\operatorname{cl}_Y(\Psi(D, x_d))$ entails that the optimization landscape of the training problem (2.3) for the approximation scheme ψ in (3.4) possesses spurious local minima and saddle points for various choices of y_d, cf. Theorems 4.15 and 4.16 below. The intuitive reason behind all these effects is that the same geometric properties of the image $\Psi(D, x_d)$, that allow ψ to satisfy (3.5), also imply that this set is folded in a way that causes the normal cones of various points on $\operatorname{cl}_Y(\Psi(D, x_d))$ to intersect.

In the remainder of this paper, we will prove that the above undesirable properties of the function $P_\Psi^{x_d}$ and the optimization problem (2.3) indeed inevitably appear when the considered approximation scheme ψ satisfies (3.5) and is conic in the sense that the set $\Psi(D, x_d)$ is a cone. We will moreover demonstrate that nearly all commonly used nonlinear approximation instruments (and in particular neural networks) are covered by this setting and are thus subject to the above effects. Note that this also shows that (3.5) is, in fact, a quite fundamental property.



(a) $\Psi(D, x_d)$ *for a linear scheme* ψ



(b) $\Psi(D, x_d)$ *for the scheme* ψ *in* (3.4)

Figure 1: *Scatter plot of the image* $\Psi(D, x_d)$ *for the linear, polynomial approximation scheme* $\psi\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, $\psi(\alpha, \chi) = \alpha_1 \chi + \alpha_2$, *(Figure a)) and the nonlinear function* ψ *in* (3.4) *(Figure b)) for* $n = 3$ *and the training data vector* $x_d = (-1/2, 1/2, 1)^T$.

Before we demonstrate that the above observations indeed carry over to a far more general setting, we would like to point out that the example that we have studied in this section is a rather academic one. It is easy to check that the approximation scheme (3.4) possesses various properties that are highly undesirable and thus would never be a sensible choice for a practical application. Moreover, the scheme ψ in (3.4) is clearly not conic and thus violates one of the main assumptions of the subsequent analysis. We have considered this function here since, on the one hand, it possesses the property (3.5) and, on the other hand, satisfies $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ for $n = 3$ - thus enabling the visualization in Fig. 1. As already mentioned, we will see in Section 5 that various commonly used approximation schemes exhibit a behavior that is very similar to that of the function ψ in (3.4). In Section 4.2, we will moreover see that, at least as far as the existence of saddle points and spurious local minima is concerned, it is not essential that it holds $n > m$ as in (3.1).

4. Analysis in the Abstract Setting

The aim of this section is to study the behavior of the function $P_\Psi^{x_d}$ and the loss landscape of the training problem (2.3) for a general, nonlinear, conic approximation scheme satisfying (3.5). Motivated by the observations made in Section 3 and by what is encountered in practical applications, we will consider the following setting:

Assumption 4.1 (Standing Assumptions for the Analysis of Section 4) *Let* $\mathcal{X}$, $\mathcal{Y}$, *etc. be defined as in Section 2. We assume that an approximation scheme* $\psi\colon D \times \mathcal{X} \to \mathcal{Y}$ *and an arbitrary but fixed training data vector* $x_d \in X$ *are given such that the following two conditions are satisfied:*

I) (Conicity) *The set* $\Psi(D, x_d)$ *is a cone in the sense that*

$$y \in \Psi(D, x_d),\ s \in (0, \infty) \quad \Rightarrow \quad sy \in \Psi(D, x_d).$$

II) (Improved Expressiveness) *The map* $\Psi(\cdot, x_d)$ *satisfies*

$$\forall y_d \in Y \setminus \{0\} : \quad \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \|y_d\|_Y^2.$$

As already mentioned, various examples of approximation schemes $\psi\colon D \times \mathcal{X} \to \mathcal{Y}$ satisfying the conditions in Assumption 4.1 will be presented in Section 5. Henceforth, the basic idea of our analysis will be to prove that the properties I) and II) - although very desirable from the approximation point of view - also automatically imply that training problems of the form (2.3) possess various disadvantageous properties. We begin with some basic observations:

Lemma 4.2 (Reformulation of Property II)) *In the situation of Assumption 4.1, the property in* II) *is equivalent to the condition*

$$\min_{y \in \mathrm{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2 < \|y_d\|_Y^2 \qquad \forall y_d \in Y \setminus \{0\}. \tag{4.1}$$

Proof The implication "II) $\Rightarrow$ (4.1)" is trivial. To prove "(4.1) $\Rightarrow$ II)", it suffices to note that, for every arbitrary but fixed $y_d \in Y \setminus \{0\}$, there exists a $\bar{y} \in \mathrm{cl}_Y(\Psi(D,x_d))$ with

$$\|\bar{y} - y_d\|_Y^2 = \min_{y \in \mathrm{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2$$

by Lemma 2.4 and to subsequently exploit the definition of the closure and the continuity of the norm $\|\cdot\|_Y$. This also shows that it indeed makes sense to write "min" on the left-hand side of (4.1) instead of "inf". ■

Note that Lemma 4.2 implies that II) is a property of the closure of the image $\Psi(D,x_d)$ of the map $\Psi(\cdot,x_d)\colon D \to Y$ and completely independent of how this image is parameterized by the variable $\alpha \in D$. To measure the extent to which condition II) is satisfied by a given approximation scheme, we introduce:

Definition 4.3 (Number $\boldsymbol{\Theta(\Psi, x_d)}$) *In the situation of Assumption 4.1, we define*

$$\Theta(\Psi, x_d) := \sup_{y_d \in Y,\, \|y_d\|_Y = 1} \left(\inf_{y \in \mathrm{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2 \right). \tag{4.2}$$

Remark 4.4 *The number $\Theta(\Psi, x_d)$ is precisely the square of the deviation of the unit sphere in $(Y, \|\cdot\|_Y)$ from the closure of the image of the function $\Psi(\cdot, x_d)\colon D \to Y$ in the sense of nonlinear approximation theory, see (Kůrková and Sanguineti, 2002, Section II).*

Using our assumptions I) and II) and the closedness of the set $\mathrm{cl}_Y(\Psi(D,x_d))$, it is easy to establish the following:

Lemma 4.5 (Properties of $\boldsymbol{\Theta(\Psi, x_d)}$) *It holds $\Theta(\Psi, x_d) \in [0,1)$. Further, for all label vectors $y_d \in Y$, the optimal value of the loss function in* (2.3) *satisfies*

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \leq \Theta(\Psi, x_d) \|y_d\|_Y^2, \tag{4.3}$$

and there exists at least one $\bar{y}_d \in Y$ with the properties

$$\|\bar{y}_d\|_Y = 1 \qquad \text{and} \qquad \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - \bar{y}_d\|_Y^2 = \Theta(\Psi, x_d). \tag{4.4}$$

Proof Using the distance function $\operatorname{dist}(\cdot, \operatorname{cl}_Y(\Psi(D, x_d)))$ to the set $\operatorname{cl}_Y(\Psi(D, x_d))$ w.r.t. the norm $\|\cdot\|_Y$, the identity in (4.2) can also be written as

$$\Theta(\Psi, x_d) = \sup_{y_d \in Y,\, \|y_d\|_Y = 1} \operatorname{dist}(y_d, \operatorname{cl}_Y(\Psi(D, x_d)))^2.$$

Since the map $Y \ni y \mapsto \operatorname{dist}(y, \operatorname{cl}_Y(\Psi(D, x_d))) \in \mathbb{R}$ is continuous, since the unit sphere $\{y \in Y \mid \|y\|_Y = 1\}$ is compact due to the finite-dimensionality of Y, and since

$$\operatorname{dist}(y_d, \operatorname{cl}_Y(\Psi(D, x_d)))^2 = \min_{y \in \operatorname{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2 = \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \tag{4.5}$$

holds for all $y_d \in Y$ by exactly the same arguments as in the proof of Lemma 4.2, it now follows immediately that there exists at least one $\bar{y}_d \in Y$ with the properties in (4.4). Note that, in combination with II), this also yields

$$0 \le \Theta(\Psi, x_d) = \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - \bar{y}_d\|_Y^2 < \|\bar{y}_d\|_Y^2 = 1$$

so that $\Theta(\Psi, x_d)$ is an element of the interval $[0, 1)$ as claimed. It remains to prove (4.3). To this end, we note that (4.5) and the cone property of the set $\operatorname{cl}_Y(\Psi(D, x_d))$ (which follows immediately from I)) imply that $y_d = 0$ is an element of $\operatorname{cl}_Y(\Psi(D, x_d))$ and that

$$\begin{aligned} \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 &= \min_{y \in \operatorname{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2 = \min_{\tilde{y} \in \operatorname{cl}_Y(\Psi(D,x_d))} \big\| \|y_d\|_Y \tilde{y} - y_d \big\|_Y^2 \\ &= \|y_d\|_Y^2 \min_{\tilde{y} \in \operatorname{cl}_Y(\Psi(D,x_d))} \left\| \tilde{y} - \frac{y_d}{\|y_d\|_Y} \right\|_Y^2 \le \Theta(\Psi, x_d) \|y_d\|_Y^2 \end{aligned}$$

holds for all $y_d \in Y \setminus \{0\}$. Combining the last two observations gives the desired estimate (4.3). This completes the proof. ∎

As Lemma 4.5 shows, the smaller the number $\Theta(\Psi, x_d)$, the better the ability of the function $\Psi(\cdot, x_d)$ to fit arbitrarily chosen label vectors $y_d \in Y$. However, since $\Theta(\Psi, x_d)$ is also a measure for the nonlinearity of the considered approximation scheme (at least in the case $\operatorname{cl}_Y(\Psi(D, x_d)) \neq Y$), one also has to expect that the optimization landscape of the problem (2.3) worsens as $\Theta(\Psi, x_d)$ tends to zero, cf. the observations made in Section 3. In Section 4.2, we will see that such an effect is indeed present and that the value of $\Theta(\Psi, x_d)$ also gives an estimate on how likely it is to encounter vectors y_d for which the problem (2.3) possesses spurious local minima and saddle points, cf. Theorems 4.13, 4.15 and 4.16.

Before we turn our attention to this topic, we study the:

4.1 Set-Valuedness and Discontinuity of the Best Approximation Map $P_\Psi^{x_d}$

Recall that we have defined $P_\Psi^{x_d}$ to be the function that maps a label vector $y_d \in Y$ to the set of elements of the closure $\operatorname{cl}_Y(\Psi(D, x_d))$ that attain the minimal loss in (2.3), i.e.,

$$P_\Psi^{x_d} \colon Y \rightrightarrows Y, \qquad y_d \mapsto \operatorname*{arg\,min}_{y \in \operatorname{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2.$$

The purpose of this subsection is to analyze which consequences the properties I) and II) in Assumption 4.1 have for this metric projection onto the set $\operatorname{cl}_Y(\Psi(D, x_d))$ and the stability properties of the training problem (2.3). As talking about the map $P_\Psi^{x_d}$ is only really sensible when the closure of the image of $\Psi(\cdot, x_d)\colon D \to Y$ is not the whole of Y (otherwise $P_\Psi^{x_d}$ is just the identity map), throughout this subsection, we will always assume that $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ holds. This corresponds to the situation where there exist label vectors y_d that are unrealizable in the sense that they cannot be approximated by the map $\Psi(\cdot, x_d)\colon D \to Y$ up to an arbitrary tolerance and thus yield a positive optimal value of the loss in (2.3). We would like to emphasize that the assumption $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ is only needed for the analysis of this subsection and Theorem 4.11 in Section 4.2. For the derivation of our other results on stationary points and spurious local minima, it is sufficient to assume that a local approximation of the function $\Psi(\cdot, x_d)\colon D \to Y$ is unable to fit arbitrary label vectors y_d precisely, cf. Theorems 4.15 to 4.17. The starting point for our study of the properties of the function $P_\Psi^{x_d}$ is the following observation:

Lemma 4.6 *Suppose that $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ holds. Then, the number $\Theta(\Psi, x_d)$ satisfies $\Theta(\Psi, x_d) \in (0, 1)$ and, for every $\bar{y}_d \in Y$ with the properties in (4.4), the following is true:*

i) The set $P_\Psi^{x_d}(\bar{y}_d)$ contains more than one element.

ii) The set $P_\Psi^{x_d}(\bar{y}_d)$ is a subset of the affine-linear space

$$H := (1 - \Theta(\Psi, x_d))\bar{y}_d + \bar{y}_d^\perp. \tag{4.6}$$

Here, $\bar{y}_d^\perp$ denotes the orthogonal complement $\bar{y}_d^\perp := \{z \in Y \mid (\bar{y}_d, z)_Y = 0\}$.

iii) It holds $(1 - \Theta(\Psi, x_d))\,\bar{y}_d \in \mathrm{conv}(P_\Psi^{x_d}(\bar{y}_d))$, where $\mathrm{conv}(\cdot)$ denotes the convex hull.

Proof Since $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ holds and since the set $\mathrm{cl}_Y(\Psi(D, x_d))$ is a closed cone, there exists at least one $y \in Y$ with $\|y\|_Y = 1$ and $\mathrm{dist}(y, \mathrm{cl}_Y(\Psi(D, x_d))) > 0$. This shows that the number $\Theta(\Psi, x_d)$ has to be positive in the situation of the lemma and, in combination with Lemma 4.5, that $\Theta(\Psi, x_d) \in (0, 1)$ holds as claimed. To prove the remaining assertions i), ii), and iii), let us assume that an arbitrary but fixed $\bar{y}_d \in Y$ satisfying (4.4) is given. (Recall that the existence of such a $\bar{y}_d$ is guaranteed by Lemma 4.5.) Then, we obtain from Lemma 2.4 that the set $P_\Psi^{x_d}(\bar{y}_d)$ contains at least one element $\bar{y} \in Y$ and it follows from the second equality in (4.4), the fact that $\Theta(\Psi, x_d)$ is smaller than one, and the definition of $P_\Psi^{x_d}(\bar{y}_d)$ that $\bar{y} \neq 0$ has to hold. Since $\bar{y}$ is a solution of the problem

$$\min_{y \in \mathrm{cl}_Y(\Psi(D, x_d))} \|y - \bar{y}_d\|_Y^2$$

and again due to the cone property of the set $\mathrm{cl}_Y(\Psi(D, x_d))$, we moreover have

$$\|\bar{y} - \bar{y}_d\|_Y^2 \leq \|s\bar{y} - \bar{y}_d\|_Y^2 \qquad \forall s > 0. \tag{4.7}$$

Choosing parameters of the form $s = 1 - \delta$ with $0 < |\delta| < 1$ in (4.7), using the binomial identities, dividing by δ, and passing to the limit $\delta \to 0$ yields that $(\bar{y} - \bar{y}_d, \bar{y})_Y = 0$ has to hold and, as a consequence,

$$(\bar{y} - (1 - \Theta(\Psi, x_d))\bar{y}_d, \bar{y}_d)_Y = (\bar{y} - \bar{y}_d, \bar{y}_d - \bar{y})_Y + \Theta(\Psi, x_d)\,\|\bar{y}_d\|_Y^2 = 0.$$

The above shows that $\bar{y}$ is contained in the affine subspace H in (4.6) and, since $\bar{y}$ was an arbitrary element of $P_\Psi^{x_d}(\bar{y}_d)$, that $P_\Psi^{x_d}(\bar{y}_d) \subset H$. This establishes ii). To prove iii), we use a contradiction argument: Suppose that the vector $(1-\Theta(\Psi, x_d))\bar{y}_d \in H$ is not an element of the convex hull $\operatorname{conv}(P_\Psi^{x_d}(\bar{y}_d)) \subset H$. Then, by noting that the set $\operatorname{conv}(P_\Psi^{x_d}(\bar{y}_d))$ is compact due to the finite-dimensionality of Y and the compactness of $P_\Psi^{x_d}(\bar{y}_d)$, see Lemma 2.4, and by applying the strong hyperplane separation theorem to the sets $(1-\Theta(\Psi, x_d))\bar{y}_d + \mathbb{R}\bar{y}_d$ and $\operatorname{conv}(P_\Psi^{x_d}(\bar{y}_d)) + \mathbb{R}\bar{y}_d$, see (Rockafellar and Wets, 1998, Theorem 2.39), we obtain that there exist a nonzero $z \in Y$ and constants $c \in \mathbb{R}$ and $\varepsilon > 0$ such that

$$(z, v_1)_Y \leq c - \varepsilon < c \leq (z, v_2)_Y$$

holds for all $v_1 \in (1-\Theta(\Psi, x_d))\bar{y}_d + \mathbb{R}\bar{y}_d$ and $v_2 \in \operatorname{conv}(P_\Psi^{x_d}(\bar{y}_d)) + \mathbb{R}\bar{y}_d$. Note that the above is only possible if $z \in \bar{y}_d^\perp$ and $c \geq \varepsilon$. We may thus conclude that z, c, and ε satisfy

$$\varepsilon \leq c \leq (z, \bar{y})_Y \qquad \forall \bar{y} \in \operatorname{conv}(P_\Psi^{x_d}(\bar{y}_d)). \tag{4.8}$$

Consider now for all sufficiently small $\tau > 0$ the vectors $y_d^\tau := (\bar{y}_d - \tau z)/\|\bar{y}_d - \tau z\|_Y$ and select arbitrary but fixed $w_\tau \in P_\Psi^{x_d}(y_d^\tau)$. (Recall that the sets $P_\Psi^{x_d}(y_d^\tau)$ are nonempty by Lemma 2.4.) Then, it follows from (4.3), the definition of $P_\Psi^{x_d}$, and the second property in (4.4) that w_τ, y_d^τ, and $\bar{y}_d$ satisfy

$$\|y_d^\tau - w_\tau\|_Y^2 \leq \Theta(\Psi, x_d) \leq \|\bar{y}_d - w_\tau\|_Y^2. \tag{4.9}$$

Using the binomial identities, the properties $\|y_d^\tau\|_Y = \|\bar{y}_d\|_Y = 1$ and $z \in \bar{y}_d^\perp$, and the definition of y_d^τ in (4.9) yields

$$\begin{aligned}
0 &\geq (y_d^\tau - \bar{y}_d, -w_\tau)_Y \\
&= \left(\frac{\bar{y}_d - \tau z}{\|\bar{y}_d - \tau z\|_Y} - \bar{y}_d, -w_\tau \right)_Y \\
&= \frac{\tau}{\|\bar{y}_d - \tau z\|_Y} (z, w_\tau)_Y + \left(\frac{\|\bar{y}_d - \tau z\|_Y - 1}{\|\bar{y}_d - \tau z\|_Y} \right) (\bar{y}_d, w_\tau)_Y \\
&= \frac{\tau}{\|\bar{y}_d - \tau z\|_Y} \left((z, w_\tau)_Y + \frac{\tau \|z\|_Y^2}{1 + \|\bar{y}_d - \tau z\|_Y} (\bar{y}_d, w_\tau)_Y \right).
\end{aligned}$$

Since the family $\{w_\tau\}$ is necessarily bounded (see the first inequality in (4.9)), the above implies that there exists a $\tau_0 > 0$ such that $(z, w_\tau)_Y \leq \varepsilon/2$ holds for all $0 < \tau < \tau_0$, where ε is the constant in (4.8), and, again by (4.8), that there exists an $\tilde{\varepsilon} > 0$ with $\operatorname{dist}(w_\tau, P_\Psi^{x_d}(\bar{y}_d)) \geq \tilde{\varepsilon}$ for all $0 < \tau < \tau_0$. However, from the boundedness of $\{w_\tau\}$ and the closedness of the set $\operatorname{cl}_Y(\Psi(D, x_d))$, we also obtain that we can find a sequence $\{w_{\tau_i}\}$ with $(0, \tau_0) \ni \tau_i \to 0$ and $w_{\tau_i} \to w$ for some $w \in \operatorname{cl}_Y(\Psi(D, x_d))$, and, due to the convergence $y_d^\tau \to \bar{y}_d$ for $\tau \to 0$ and (4.9), such a w clearly has to satisfy

$$\|\bar{y}_d - w\|_Y^2 = \Theta(\Psi, x_d) = \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - \bar{y}_d\|_Y^2.$$

The above yields $w \in P_\Psi^{x_d}(\bar{y}_d)$ and, as a consequence,

$$0 < \tilde{\varepsilon} \leq \operatorname{dist}(w, P_\Psi^{x_d}(\bar{y}_d)) = 0$$

which is not possible. The vector $(1-\Theta(\Psi,x_d))\bar{y}_d$ thus has to be an element of the set $\operatorname{conv}(P_\Psi^{x_d}(\bar{y}_d))$ and the proof of iii) is complete. Since the assertion in i) is a trivial consequence of ii), iii), and the fact that $\Theta(\Psi,x_d)$ is smaller than one by Lemma 4.5, this concludes the proof of the lemma. ■

Note that Lemma 4.6 implies that the cone $\mathrm{cl}_Y(\Psi(D,x_d))$ can only be convex if it is equal to the whole space Y. By exploiting the properties I) and II) directly, we can also establish the following, stronger result on the geometry of this set:

Proposition 4.7 (Nonexistence of Solar Points) *Consider the situation in Assumption 4.1 and suppose that* $\mathrm{cl}_Y(\Psi(D,x_d))\neq Y$ *holds. Then, for every* $y_d\in Y\setminus \mathrm{cl}_Y(\Psi(D,x_d))$ *and every arbitrary but fixed* $\bar{y}\in P_\Psi^{x_d}(y_d)$*, it is true that*

$$\bar{y}\notin P_\Psi^{x_d}\left(\bar{y}+s\frac{(y_d-\bar{y})}{\|y_d-\bar{y}\|_Y}\right)\qquad \forall s\in\mathbb{R}\text{ with }|s|>\left(\frac{\Theta(\Psi,x_d)}{1-\Theta(\Psi,x_d)}\right)^{1/2}\|\bar{y}\|_Y. \tag{4.10}$$

In particular, the set $\mathrm{cl}_Y(\Psi(D,x_d))$ *does not admit any solar points, i.e., there do not exist any* $y_d\in Y\setminus\mathrm{cl}_Y(\Psi(D,x_d))$ *such that there is a* $\bar{y}\in P_\Psi^{x_d}(y_d)$ *with*

$$\bar{y}\in P_\Psi^{x_d}(\bar{y}+s(y_d-\bar{y}))\qquad\forall s\in(0,\infty).$$

Proof Consider an arbitrary but fixed $y_d\in Y\setminus\mathrm{cl}_Y(\Psi(D,x_d))$ and some $\bar{y}\in P_\Psi^{x_d}(y_d)$. Then, it necessarily holds $y_d\neq\bar{y}$, and we obtain from the same arguments as in the proof of Lemma 4.6 that $(\bar{y}-y_d,\bar{y})_Y=0$ has to hold. Define $v:=(y_d-\bar{y})/\|y_d-\bar{y}\|_Y$ and $y_d^s:=\bar{y}+sv\in Y$ for all $s\in\mathbb{R}$, and let $\bar{y}_s\in Y$ be arbitrary but fixed elements of the sets $P_\Psi^{x_d}(y_d^s)$ for all $s\in\mathbb{R}$. Then, from (4.3), the definition of y_d^s, the orthogonality between v and $\bar{y}$, and the equation $\|v\|_Y=1$, we obtain that

$$\|\bar{y}_s-y_d^s\|_Y^2-\|\bar{y}-y_d^s\|_Y^2\leq\Theta(\Psi,x_d)\|y_d^s\|_Y^2-\|sv\|_Y^2=\Theta(\Psi,x_d)\|\bar{y}\|_Y^2+(\Theta(\Psi,x_d)-1)s^2 \tag{4.11}$$

holds for all $s\in\mathbb{R}$. Since the set $\Psi(D,x_d)$ is dense in $\mathrm{cl}_Y(\Psi(D,x_d))$ and since $\Theta(\Psi,x_d)$ is an element of the interval $(0,1)$ by Lemma 4.6, the above shows that, for all $s\in\mathbb{R}$ with

$$|s|>\left(\frac{\Theta(\Psi,x_d)}{1-\Theta(\Psi,x_d)}\right)^{1/2}\|\bar{y}\|_Y,$$

we have

$$\inf_{\alpha\in D}\|\Psi(\alpha,x_d)-y_d^s\|_Y^2<\|\bar{y}-y_d^s\|_Y^2$$

and, as a consequence, $\bar{y}\notin P_\Psi^{x_d}(\bar{y}+sv)$. This establishes the first claim of the proposition. The second one is an immediate consequence. ■

It is easy to check that the property (4.10) implies that the set $\mathrm{cl}_Y(\Psi(D,x_d))$ does not admit *any* supporting hyperplanes in the situation of Proposition 4.7. This shows that the cone $\mathrm{cl}_Y(\Psi(D,x_d))$ indeed has to be highly nonconvex if the map $\Psi(\cdot,x_d)\colon D\to Y$ satisfies I) and II) and there exist unrealizable vectors. Compare also with the geometry of the set in Fig. 1b) in this context. For further details on solar points and their role in the field of nonlinear approximation theory, we refer the reader to Braess (1986). We remark that arguments very similar to those in the proof of Proposition 4.7 will also be used in Section 4.2 for the derivation of our results on saddle points and spurious minima.

To study which consequences the inclusion in point iii) of Lemma 4.6 has for the continuity properties of the map $P_\Psi^{x_d}$, we need:

Lemma 4.8 (A Variant of Jung's Inequality) *Suppose that H is an affine-linear subspace of Y with dimension $d \in \{1, 2, ..., \dim(Y)\}$. Assume further that a point $\bar{y} \in H$, a compact set $E \subset H$, and a number $r > 0$ satisfying*

$$\bar{y} \in \operatorname{conv}(E) \qquad \text{and} \qquad \|\bar{y} - z\|_Y = r \quad \forall z \in E$$

are given. Then, it is true that

$$\sup_{z_1, z_2 \in E} \|z_1 - z_2\|_Y \geq \left(\frac{2d+2}{d}\right)^{1/2} r. \tag{4.12}$$

Proof Note that, by introducing a suitably defined orthonormal basis and by restricting the attention to the space of directions of the affine subspace H, we can always transform the situation considered in the lemma into that with $Y = H = \mathbb{R}^d$ and $\|\cdot\|_Y = \|\cdot\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm. It thus suffices to prove (4.12) in the space $(\mathbb{R}^d, \|\cdot\|_2)$ for all $\bar{y} \in \mathbb{R}^d$, compact sets $E \subset \mathbb{R}^d$, and constants $r > 0$ that satisfy $\bar{y} \in \operatorname{conv}(E)$ and $\|\bar{y} - z\|_2 = r$ for all $z \in E$. So let us assume that such $\bar{y}$, E, and r are given, and suppose further that $B_R(v)$ is a closed ball in $(\mathbb{R}^d, \|\cdot\|_2)$ with center v and radius R that covers the set E. Then, in the case $v = \bar{y}$, our assumption $\|\bar{y} - z\|_2 = r$ for all $z \in E$ immediately yields that $R \geq r$ has to hold. In what follows, we will show that this inequality is also true for $v \neq \bar{y}$. To this end, we note that the inclusion $\bar{y} \in \operatorname{conv}(E)$ and Carathéodory's theorem, see (Borwein and Vanderwerff, 2010, Theorem 1.2.5), imply that there exist $\lambda_1, ..., \lambda_{d+1} \in [0, 1]$ and $z_1, ..., z_{d+1} \in E$ satisfying $\sum_{i=1}^{d+1} \lambda_i = 1$ and $\sum_{i=1}^{d+1} \lambda_i z_i = \bar{y}$, and, as a consequence,

$$0 = (\bar{y} - v, \bar{y} - \bar{y})_2 = \sum_{i=1}^{d+1} \lambda_i \left(\bar{y} - v, z_i - \bar{y}\right)_2.$$

Here, $(\cdot, \cdot)_2$ denotes the Euclidean scalar product. The above implies in particular that there has to be at least one $j \in \{1, ..., d+1\}$ with $(\bar{y} - v, z_j - \bar{y})_2 \geq 0$, and from this inequality and the inclusion $E \subset B_R(v)$, it follows straightforwardly that

$$R^2 \geq \|z_j - v\|_2^2 = \|z_j - \bar{y} + \bar{y} - v\|_2^2 = \|z_j - \bar{y}\|_2^2 + 2\left(\bar{y} - v, z_j - \bar{y}\right)_2 + \|\bar{y} - v\|_2^2 \geq r^2.$$

Thus, $R \geq r$ as claimed. In summary, we have now proved that every closed ball $B \subset \mathbb{R}^d$ with $E \subset B$ has to have radius at least r. In combination with the classical inequality of Jung, see (Burago and Zalgaller, 1988, Theorem 11.1.1), this yields

$$\sup_{z_1, z_2 \in E} \|z_1 - z_2\|_2 \left(\frac{d}{2d+2}\right)^{1/2} \geq r.$$

Rearranging the above establishes (4.12) and completes the proof. ■

By combining Lemmas 4.6 and 4.8 and by using elementary properties of the map $P_\Psi^{x_d}$, we now arrive at the following main result of this subsection:

Theorem 4.9 (Nonuniqueness and Instability of Best Approximations) *Consider the situation in Assumption 4.1 and suppose that* $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ *holds. Then, the best approximation map*

$$P_\Psi^{x_d} \colon Y \rightrightarrows Y, \qquad y_d \mapsto \operatorname*{arg\,min}_{y \in \mathrm{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2,$$

associated with the training problem (2.3) *has the following properties:*

i) There are uncountably many y_d *such that* $P_\Psi^{x_d}(y_d)$ *contains more than one element.*

ii) The function $P_\Psi^{x_d}$ *is discontinuous in the following sense: For every arbitrary but fixed* $C > 0$*, there exists an uncountable set* $\mathcal{M}_C \subset Y$ *such that, for every label vector* $y_d \in \mathcal{M}_C$*, there exist sequences* $\{y_d^l\}, \{\tilde{y}_d^l\} \subset Y$ *with*

$$\begin{gathered} y_d^l \to y_d \text{ for } l \to \infty, \qquad \tilde{y}_d^l \to y_d \text{ for } l \to \infty, \\ |P_\Psi^{x_d}(y_d^l)| = |P_\Psi^{x_d}(\tilde{y}_d^l)| = 1 \quad \forall l, \quad \text{and} \quad \|P_\Psi^{x_d}(y_d^l) - P_\Psi^{x_d}(\tilde{y}_d^l)\|_Y \geq C \quad \forall l. \end{gathered} \tag{4.13}$$

Here, $|\cdot|$ *denotes the cardinality of a set and with* $\|P_\Psi^{x_d}(y_d^l) - P_\Psi^{x_d}(\tilde{y}_d^l)\|_Y$ *we mean the distance between the elements of the singletons* $P_\Psi^{x_d}(y_d^l)$ *and* $P_\Psi^{x_d}(\tilde{y}_d^l)$*. Further, for every* $C > 0$*, there exists at least one* $y_d \in Y$ *with the properties*

$$[1, \infty) y_d \subset \mathcal{M}_C \quad \text{and} \quad \|y_d\|_Y = C \left(\frac{\dim(Y) - 1}{2 \dim(Y)(\Theta(\Psi, x_d) - \Theta(\Psi, x_d)^2)} \right)^{1/2}. \tag{4.14}$$

Proof Let $\bar{y}_d \in Y$ be an arbitrary but fixed vector with the properties in (4.4). Then, from Lemma 4.6, it follows that $|P_\Psi^{x_d}(\bar{y}_d)| > 1$ holds, and we obtain from the conicity of the set $\mathrm{cl}_Y(\Psi(D, x_d))$ that $P_\Psi^{x_d}$ satisfies $P_\Psi^{x_d}(s y_d) = s P_\Psi^{x_d}(y_d)$ for all $y_d \in Y$ and all $s > 0$. Combining these two observations yields $|P_\Psi^{x_d}(s\bar{y}_d)| > 1$ for all $s > 0$ which proves the assertion of i). To establish ii), we recall that, by Lemma 4.6, the compact set $E := P_\Psi^{x_d}(\bar{y}_d)$ has to satisfy $(1 - \Theta(\Psi, x_d))\, \bar{y}_d \in \mathrm{conv}(E) \subset H$, where H again denotes the affine subspace in (4.6), and that the definition of $P_\Psi^{x_d}(\bar{y}_d)$ yields $(z - \bar{y}_d, z)_Y = 0$ for all $z \in E$ (see the first part of the proof of Lemma 4.6). The latter implies, in combination with the properties of $\bar{y}_d$, that

$$\begin{aligned} \|z - (1 - \Theta(\Psi, x_d))\, \bar{y}_d\|_Y^2 &= \|z - \bar{y}_d\|_Y^2 + 2\Theta(\Psi, x_d)\, (z - \bar{y}_d, \bar{y}_d)_Y + \Theta(\Psi, x_d)^2 \\ &= \Theta(\Psi, x_d) - \Theta(\Psi, x_d)^2 \qquad \forall z \in E. \end{aligned}$$

The vector $\bar{y} := (1 - \Theta(\Psi, x_d))\, \bar{y}_d$ and the number $r := (\Theta(\Psi, x_d) - \Theta(\Psi, x_d)^2)^{1/2} > 0$ thus satisfy

$$\bar{y} \in \mathrm{conv}(E) \subset H \quad \text{and} \quad \|\bar{y} - z\|_Y = r \quad \forall z \in E,$$

and we may invoke Lemma 4.8 to deduce that there exist $z_1, z_2 \in P_\Psi^{x_d}(\bar{y}_d)$ with

$$\|z_1 - z_2\|_Y \geq \left(\frac{2 \dim(Y)(\Theta(\Psi, x_d) - \Theta(\Psi, x_d)^2)}{\dim(Y) - 1} \right)^{1/2}.$$

Consider now the sequence $y_d^l := (1-1/l)\bar{y}_d + (1/l)z_1$, $l \in \mathbb{N}$. Then, we clearly have $y_d^l \to \bar{y}_d$ for $l \to \infty$ and it holds

$$\begin{aligned} \|y_d^l - z\|_Y &= \|(1-1/l)\bar{y}_d + (1/l)z_1 - z\|_Y \\ &\geq \|\bar{y}_d - z\|_Y - (1/l)\|\bar{y}_d - z_1\|_Y \\ &\geq (1-1/l)\Theta(\Psi, x_d)^{1/2} \qquad \forall z \in \mathrm{cl}_Y(\Psi(D, x_d)) \end{aligned} \tag{4.15}$$

with equality everywhere if and only if $z = z_1$. In combination with the definition of $P_\Psi^{x_d}$, this implies in particular that $P_\Psi^{x_d}(y_d^l) = \{z_1\}$ holds for all $l \in \mathbb{N}$. Completely analogously, we also obtain that the vectors $\tilde{y}_d^l := (1-1/l)\bar{y}_d + (1/l)z_2$, $l \in \mathbb{N}$, satisfy $\tilde{y}_d^l \to \bar{y}_d$ for $l \to \infty$ and $P_\Psi^{x_d}(\tilde{y}_d^l) = \{z_2\}$ for all $l \in \mathbb{N}$. By again exploiting the positive homogeneity of the map $P_\Psi^{x_d} \colon Y \rightrightarrows Y$ and by combining all of the above, it now follows immediately that, for every arbitrary but fixed $C > 0$ and all

$$s \geq C \left(\frac{\dim(Y) - 1}{2\dim(Y)(\Theta(\Psi, x_d) - \Theta(\Psi, x_d)^2)} \right)^{1/2},$$

we have

$$s y_d^l \to s\bar{y}_d \text{ for } l \to \infty, \quad s\tilde{y}_d^l \to s\bar{y}_d \text{ for } l \to \infty,$$

and

$$P_\Psi^{x_d}(s y_d^l) = \{s z_1\}, \qquad P_\Psi^{x_d}(s\tilde{y}_d^l) = \{s z_2\}, \qquad \|s z_1 - s z_2\|_Y \geq C \quad \forall l \in \mathbb{N}.$$

Since $\|\bar{y}_d\|_Y = 1$ holds by (4.4), this establishes ii) and completes the proof. ■

Several remarks are in order regarding the last result:

Remark 4.10

- *Theorem 4.9 shows that, if there exist label vectors y_d that cannot be approximated up to arbitrary tolerances and if* I) *and* II) *hold, then the approximation scheme ψ is always unable to provide unique best approximations for all possible choices of y_d (see point i)) and arbitrarily small perturbations in y_d can change the set of best approximations to an arbitrarily large extent (see point ii)). This implies in particular that, in the situation of Theorem 4.9, the problem of finding best approximations for a given y_d is always ill-posed in the sense of Hadamard for certain choices of y_d.*
- *As already mentioned in the introduction, for neural networks with one hidden layer, instability results similar to those in Theorem 4.9 have already been proved in the L^p-spaces in Kainen et al. (1999, 2001) by exploiting classical instruments from nonlinear approximation theory. The finite-dimensionality of the training problem* (2.3) *allows us to show - not only for one-hidden-layer networks but for all approximation schemes satisfying the conditions* I) *and* II) *and* $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ *- that the instability of the best approximation map $P_\Psi^{x_d}$ associated with* (2.3) *is directly linked to the number* $\Theta(\Psi, x_d)$ *in* (4.2) *which also measures the worst-case approximation error achievable with the function $\Psi(\cdot, x_d) \colon D \to Y$, see* (4.3) *and* (4.14). *(Note that the arguments that we have used to establish* (4.14) *indeed only work in the finite-dimensional setting, cf. the proofs of Lemmas 4.6 and 4.8.)*

- *The instability properties in Theorem 4.9 are of a different type than those arising, e.g., in a least-squares problem of the form*

$$\min_{\alpha \in \mathbb{R}^m} \|A\alpha - y_d\|_2^2$$

with given $y_d \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$, *and* $n \geq m$, *when the matrix* $A^T A \in \mathbb{R}^{m \times m}$ *(i.e., the matrix in the normal equation) is ill-conditioned or singular. Indeed, as we have seen in Section 3, for approximation schemes that depend linearly on* α, *the map* $P_\Psi^{x_d}$ *is always a metric projection onto a linear subspace of* Y *and thus necessarily single-valued and globally one-Lipschitz. The set-valuedness and the discontinuity of the function* $P_\Psi^{x_d}$ *in Theorem 4.9 are effects that can only be encountered in the nonlinear setting as they stem from curvature properties of the set* $\mathrm{cl}_Y(\Psi(D, x_d))$. *Instability properties arising from a particular choice of the parameterization of the set* $\mathrm{cl}_Y(\Psi(D, x_d))$ *via the parameter* α *come on top of the effects documented in Theorem 4.9.*

- *It is easy to check (e.g., by means of the examples* $\mathrm{cl}_Y(\Psi(D, x_d)) = \mathbb{R}\bar{y}$, $\bar{y} \in Y$ *arbitrary but fixed, and* $\mathrm{cl}_Y(\Psi(D, x_d)) = B_1^Y(0)$, *and by observing that* $P_\Psi^{x_d}$ *is the identity map when* $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$ *holds) that neither the conditions in Assumption 4.1 nor the assumption* $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ *can be dropped for Theorem 4.9 to be true.*

- *Note that the right-hand side of the estimate* (4.14) *tends to infinity when* $\Theta(\Psi, x_d)$ *goes to zero or one, respectively. This makes sense as the function* ψ *behaves more and more like a linear approximation scheme when* $\Theta(\Psi, x_d)$ *converges to one (at least as far as the worst-case approximation error is concerned, cf. Section 3), and since, in the limit* $\Theta(\Psi, x_d) \to 0$, *one recovers the case with* $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$, *so that, for both* $\Theta(\Psi, x_d) \to 0$ *and* $\Theta(\Psi, x_d) \to 1$, *the setting considered in Theorem 4.9 approximates a situation in which the map* $P_\Psi^{x_d}$ *is single-valued and continuous.*

- *The nonuniqueness in point i) of Theorem 4.9 has nothing to do with, e.g., a non-injective parameterization of the set* $\Psi(D, x_d)$ *via the variable* α *as present, for instance, in neural networks due to symmetries. On the contrary, as* $P_\Psi^{x_d}(y_d)$ *is defined as the set of best approximations for a given* y_d *in the space* Y, *the set-valuedness of* $P_\Psi^{x_d}$ *implies (just by taking preimages) that for some choices of* y_d *there are different parameters* α *(or, at least, minimizing sequences) which yield the same optimal loss in* (2.3) *but give rise to maps* $\psi(\alpha, \cdot)\colon \mathcal{X} \to \mathcal{Y}$ *that behave differently not only on unseen data but even on the data in* x_d *that the approximation scheme is trained on.*

- *The discontinuity properties of the map* $P_\Psi^{x_d}$ *in point ii) of Theorem 4.9 imply that, if we solve the training problem* (2.3) *with a descent method and, by doing so, obtain a sequence of parameters* $\{\alpha_i\}$ *satisfying* (2.4) *and* $\Psi(\alpha_i, x_d) \to \bar{y}$ *for some* $\bar{y} \in Y$, *then an arbitrarily small perturbation of the training label vector* y_d *can cause the solution algorithm to produce a different sequence* $\{\tilde{\alpha}_i\}$, *which again satisfies* (2.4) *and, in the limit* $i \to \infty$, *yields a loss that is arbitrarily close to that obtained with* $\{\alpha_i\}$, *but satisfies* $\Psi(\tilde{\alpha}_i, x_d) \to \tilde{y}$ *with a vector* $\tilde{y} \in Y$ *that is arbitrarily far away from* $\bar{y}$. *Note that the latter again implies that the functions* $\psi(\alpha_i, \cdot)\colon \mathcal{X} \to \mathcal{Y}$ *and* $\psi(\tilde{\alpha}_i, \cdot)\colon \mathcal{X} \to \mathcal{Y}$ *behave differently on the training data as* i *tends to infinity. We remark that these instability effects predicted by our analysis can also be observed in practice. Compare, for instance, with the results of Cunningham et al. (2000) in this context.*

4.2 Existence of Spurious Local Minima and Saddle Points

Having discussed the properties of the map $P_\Psi^{x_d}$, we now turn our attention to the question of whether the problem (2.3) possesses saddle points and spurious local minima. We begin with a result that builds upon the findings of Theorem 4.9 and shows that, in the presence of unrealizable vectors, the training problem (2.3) can only lack spurious local minima and non-optimal basins stretching to the boundary of the parameter set D for all $y_d \in Y$ if the image of the function $Y \ni y_d \mapsto |P_\Psi^{x_d}(y_d)| \in \mathbb{N} \cup \{\infty\}$ is equal to $\{1, \infty\}$, i.e., if the space Y can be decomposed into two nonempty disjoint sets Y_1 and Y_2 such that every $y_d \in Y_1$ possesses exactly one best approximation in $\mathrm{cl}_Y(\Psi(D, x_d))$ and such that, for every $y_d \in Y_2$, there are infinitely many best approximations in $\mathrm{cl}_Y(\Psi(D, x_d))$.

Theorem 4.11 (Relation Between Set-Valuedness and Spurious Local Minima) *Consider the situation in Assumption 4.1 and suppose that $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ holds. Assume further that the function $\Psi(\cdot, x_d)\colon D \to Y$ is continuous and that the image of the map $Y \ni y_d \mapsto |P_\Psi^{x_d}(y_d)| \in \mathbb{N} \cup \{\infty\}$ is not equal to $\{1, \infty\}$ (where the symbol $|\cdot|$ again denotes the cardinality of a set). Then, there exist an open nonempty cone $K \subset Y$ and a number $M \in \mathbb{N}$ with $M \geq 2$ such that, for every $y_d \in K$, there are nonempty, disjoint, relatively closed subsets $D_1, ..., D_M$ of the parameter set $D \subset \mathbb{R}^m$ with*

$$\inf_{\alpha \in D_1} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \inf_{\alpha \in D_i} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \quad \forall i = 2, ..., M \tag{4.16}$$

and

$$\sup_{\alpha \in D_1 \cup ... \cup D_M} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \|\Psi(\tilde{\alpha}, x_d) - y_d\|_Y^2 \quad \forall \tilde{\alpha} \in D \setminus (D_1 \cup ... \cup D_M). \tag{4.17}$$

Proof If the image of the map $Y \ni y_d \mapsto |P_\Psi^{x_d}(y_d)| \in \mathbb{N} \cup \{\infty\}$ is not equal to $\{1, \infty\}$, then it follows from Lemma 2.4 and Theorem 4.9i) that there has to exist at least one $\bar{y}_d \in Y \setminus \{0\}$ with $1 < |P_\Psi^{x_d}(\bar{y}_d)| < \infty$. Define $M := |P_\Psi^{x_d}(\bar{y}_d)|$ and $r := \mathrm{dist}(\bar{y}_d, P_\Psi^{x_d}(\bar{y}_d)) > 0$, and let us denote the M distinct elements of the set $P_\Psi^{x_d}(\bar{y}_d)$ with $\bar{y}_i$, $i = 1, ..., M$. Consider further an $\varepsilon > 0$ such that the closed balls $B_\varepsilon^Y(\bar{y}_i)$, $i = 1, ..., M$, satisfy $\mathrm{dist}(B_\varepsilon^Y(\bar{y}_i), B_\varepsilon^Y(\bar{y}_j)) > 2\varepsilon$ for all $i \neq j$. Then, it follows from the definition of $P_\Psi^{x_d}(\bar{y}_d)$ that there exists a number $\delta \in (0, \varepsilon]$ with $\mathrm{dist}(B_r^Y(\bar{y}_d) \setminus (B_\varepsilon^Y(\bar{y}_1) \cup ... \cup B_\varepsilon^Y(\bar{y}_M)), \mathrm{cl}_Y(\Psi(D, x_d))) \geq \delta$. Using this δ, we define $E_i := B_{\varepsilon+\delta}^Y(\bar{y}_i) \cap B_{r+\delta/2}^Y(\bar{y}_d) \cap \mathrm{cl}_Y(\Psi(D, x_d))$. Note that this construction ensures that the sets E_i, $i = 1, ..., M$, are nonempty, compact, and disjoint. From the choice of δ and ε, we further obtain that the sets E_i, $i = 1, ..., M$, satisfy

$$B_{r+\delta/2}^Y(\bar{y}_d) \cap \mathrm{cl}_Y(\Psi(D, x_d)) = \bigcup_{i=1}^M E_i. \tag{4.18}$$

Indeed, the inclusion "$\supset$" in the equality (4.18) follows immediately from the definition of the sets E_i, and if there was a $\bar{y} \in B_{r+\delta/2}^Y(\bar{y}_d) \cap \mathrm{cl}_Y(\Psi(D, x_d)) \setminus \bigcup_{i=1}^M E_i$, then this vector $\bar{y}$ would satisfy $\mathrm{dist}(\bar{y}, B_r^Y(\bar{y}_d)) \leq \delta/2$ and $\|\bar{y} - \bar{y}_i\|_Y > \varepsilon + \delta$ for all $i = 1, ..., M$ which would imply the existence of a $\tilde{y} \in B_r^Y(\bar{y}_d) \setminus (B_\varepsilon^Y(\bar{y}_1) \cup ... \cup B_\varepsilon^Y(\bar{y}_M))$ with $\|\bar{y} - \tilde{y}\|_Y \leq \delta/2$ and thus contradict the definition of δ.

To prove the claim of the theorem, we now consider the vector

$$\tilde{y}_d := \bar{y}_d + \frac{\delta}{8}\frac{\bar{y}_1 - \bar{y}_d}{\|\bar{y}_1 - \bar{y}_d\|_Y}.$$

Note that, by exactly the same arguments as in (4.15), we obtain that this $\tilde{y}_d$ satisfies $P_\Psi^{x_d}(\tilde{y}_d) = \{\bar{y}_1\}$ and, as a consequence,

$$\|\bar{y}_1 - \tilde{y}_d\|_Y = \min_{y \in E_1} \|y - \tilde{y}_d\|_Y = \min_{y \in \mathrm{cl}_Y(\Psi(D,x_d))} \|y - \tilde{y}_d\|_Y = r - \frac{\delta}{8}.$$

The above implies in particular that

$$\begin{aligned} \mathrm{dist}(\tilde{y}_d, E_1) = \mathrm{dist}(\tilde{y}_d, \mathrm{cl}_Y\left(\Psi(D,x_d)\right)) < r,\\ \mathrm{dist}(\tilde{y}_d, E_1) < \mathrm{dist}(\tilde{y}_d, E_i) \quad \forall i = 2, ..., M. \end{aligned} \tag{4.19}$$

Due to the Lipschitz continuity of the distance functions in (4.19), the definitions of $\tilde{y}_d$ and E_i, and (4.18), the estimates in (4.19) remain valid for all y_d that are sufficiently close to $\tilde{y}_d$. We can thus find a $\tau \in (0, \delta/8)$ such that, for every $y_d \in Y$ with $\|y_d - \tilde{y}_d\|_Y < \tau$, we have

$$\begin{aligned} \mathrm{dist}(y_d, E_1) = \mathrm{dist}(y_d, \mathrm{cl}_Y\left(\Psi(D,x_d)\right)) < r,\\ \mathrm{dist}(y_d, E_1) < \mathrm{dist}(y_d, E_i) \quad \forall i = 2, ..., M. \end{aligned} \tag{4.20}$$

Note that the choice $\tau \in (0, \delta/8)$ ensures that the closed ball $B^Y_{r+\delta/4}(y_d)$ is contained in the interior of $B^Y_{r+\delta/2}(\bar{y}_d)$ for every y_d with $\|y_d - \tilde{y}_d\|_Y < \tau$, and that the intersection of the interior of the ball $B^Y_{r+\delta/4}(y_d)$ with each E_i is nonempty. The latter property implies, in combination with the fact that every vector in E_i can be approximated by elements of the image $\Psi(D, x_d)$, the compactness and disjointness of the sets E_i, and (4.18), that the estimates in (4.20) remain true when we intersect the sets E_i with $B^Y_{r+\delta/4}(y_d) \cap \Psi(D, x_d)$, i.e., it holds

$$\mathrm{dist}(y_d, E_1 \cap B^Y_{r+\delta/4}(y_d) \cap \Psi(D,x_d)) = \mathrm{dist}(y_d, \mathrm{cl}_Y\left(\Psi(D,x_d)\right)) < r \tag{4.21}$$

and

$$\mathrm{dist}(y_d, E_1 \cap B^Y_{r+\delta/4}(y_d) \cap \Psi(D,x_d)) < \mathrm{dist}(y_d, E_i \cap B^Y_{r+\delta/4}(y_d) \cap \Psi(D,x_d)) \tag{4.22}$$

for all $i = 2, ..., M$. Consider now an arbitrary but fixed y_d with $\|y_d - \tilde{y}_d\|_Y < \tau$ and define $D_i := \Psi(\cdot, x_d)^{-1}(E_i \cap B^Y_{r+\delta/4}(y_d))$. Then, the continuity of the map $\Psi(\cdot, x_d)$ and the properties discussed above imply that the sets D_i, $i = 1, ..., M$, are relatively closed, disjoint, and nonempty subsets of D which satisfy

$$\inf_{\alpha \in D_1} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \inf_{\alpha \in D_i} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \quad \forall i = 2, ..., M.$$

From the definition of the sets D_i, we further obtain that, for every arbitrary but fixed $\tilde{\alpha} \in D \setminus (D_1 \cup ... \cup D_M)$, we have $\Psi(\tilde{\alpha}, x_d) \notin (E_1 \cup ... \cup E_M) \cap B^Y_{r+\delta/4}(y_d)$. Since (4.18) and the inclusion $B^Y_{r+\delta/4}(y_d) \subset B^Y_{r+\delta/2}(\bar{y}_d)$ yield

$$B^Y_{r+\delta/4}(y_d) \cap \mathrm{cl}_Y(\Psi(D,x_d)) = \bigcup_{i=1}^M B^Y_{r+\delta/4}(y_d) \cap E_i,$$

this implies in particular that $\Psi(\tilde{\alpha}, x_d) \notin B^Y_{r+\delta/4}(y_d)$ holds for all $\tilde{\alpha} \in D \setminus (D_1 \cup ... \cup D_M)$ and, as a consequence, that

$$\sup_{\alpha \in D_1 \cup ... \cup D_M} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \leq \left(r + \frac{\delta}{4}\right)^2 < \|\Psi(\tilde{\alpha}, x_d) - y_d\|_Y^2$$

for all $\tilde{\alpha} \in D \setminus (D_1 \cup ... \cup D_M)$. The vector y_d and the sets D_i thus indeed satisfy (4.16) and (4.17). As y_d was an arbitrary vector with $\|y_d - \tilde{y}_d\|_Y < \tau$, the existence of an open set K with the properties in Theorem 4.11 now follows immediately. To see that the set K can be chosen to be an open cone, it suffices to note that, since all of the above arguments up to the estimates (4.21) and (4.22) only rely on geometric properties of the set $\mathrm{cl}_Y(\Psi(D, x_d))$ and since the set $\mathrm{cl}_Y(\Psi(D, x_d))$ is a cone by I), by rescaling, we also obtain the claim for all $y_d \in Y$ which satisfy $\|y_d - s\tilde{y}_d\|_Y < s\tau$ for some $s > 0$. This completes the proof. ∎

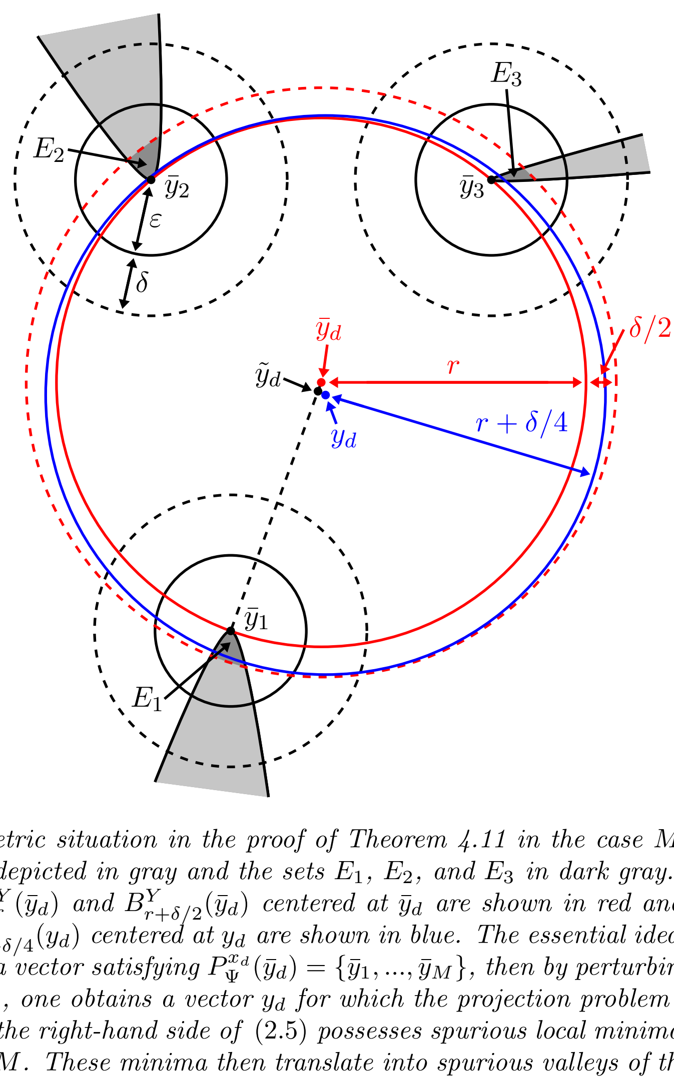


Figure 2: *Geometric situation in the proof of Theorem 4.11 in the case $M = 3$. The set $\mathrm{cl}_Y(\Psi(D, x_d))$ is depicted in gray and the sets E_1, E_2, and E_3 in dark gray. The vector $\bar{y}_d$ and the circles $B^Y_r(\bar{y}_d)$ and $B^Y_{r+\delta/2}(\bar{y}_d)$ centered at $\bar{y}_d$ are shown in red and the vector y_d and the circle $B^Y_{r+\delta/4}(y_d)$ centered at y_d are shown in blue. The essential idea of the proof is that, if $\bar{y}_d \in Y$ is a vector satisfying $P^{x_d}_\Psi(\bar{y}_d) = \{\bar{y}_1, ..., \bar{y}_M\}$, then by perturbing $\bar{y}_d$ slightly in the direction of $\bar{y}_1$, one obtains a vector y_d for which the projection problem in the variable y associated with the right-hand side of (2.5) possesses spurious local minima in each of the sets E_i, $i = 2, ..., M$. These minima then translate into spurious valleys of the optimization landscape of the problem (2.3) by taking preimages under the function $\Psi(\cdot, x_d)\colon D \to Y$.*

Note that the result in Theorem 4.11 is, in fact, slightly stronger than that stated in point two of Section 1.2 as it not only expresses that, in the presence of unrealizability, one cannot simultaneously get rid of both vectors y_d with infinitely many best approximations and vectors y_d for which (2.3) possesses spurious basins, but even that the only situation, in which spurious basins can be completely absent in (2.3) for all $y_d \in Y$ in the presence of unrealizable vectors, is that where the image of the function $y_d \mapsto |P_\Psi^{x_d}(y_d)|$ is equal to $\{1, \infty\}$. (Recall that, if there exists a $y_d \in Y$ with $|P_\Psi^{x_d}(y_d)| = \infty$, then there are automatically uncountably many such vectors by the conicity in I).) We remark that, as the cone $\mathrm{cl}_Y(\Psi(D, x_d))$ has to have very special geometric properties for the map $y_d \mapsto |P_\Psi^{x_d}(y_d)|$ to only take the values one and infinity, cases without spurious local minima and/or basins seem to be very rare in the above context. (An example of a cone $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ that satisfies $|P_\Psi^{x_d}(y_d)| \in \{1, \infty\}$ for all $y_d \in Y$ is the complement of a Lorentz cone in $\mathbb{R}^3$.) This impression is also confirmed by the results on the existence of spurious valleys in one-hidden-layer neural networks with non-polynomial non-negative activation functions proved in Venturi et al. (2019). We would like to point out that the properties (4.16) and (4.17) immediately imply that the problem (2.3) possesses spurious valleys in the sense of (Venturi et al., 2019, Definition 1). Theorem 4.11 thus complements the findings of these authors. For related work on the existence and role of bad basins in the loss landscape, see also Nguyen et al. (2019); Cooper (2020); Li et al. (2018).

Checking whether Theorem 4.11 is applicable in a certain situation or not is, of course, typically far from trivial. Because of this and since Theorem 4.11 does not yield any information about how far away spurious local minima can be from global solutions of (2.3) (should they exist), in what follows, we will prove criteria for the existence of non-optimal stationary points that do not rely on the geometric properties of the set $\mathrm{cl}_Y(\Psi(D, x_d))$ but rather exploit the condition II) directly. As we will see below, this approach has the additional advantage that it does not require the condition $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ and is thus also applicable in the overparameterized regime. The starting point of our analysis is the following lemma whose proof follows the lines of that of Proposition 4.7:

Lemma 4.12 *Suppose that Assumption 4.1 holds and that $\bar\alpha \in D$ is arbitrary but fixed. Assume further that a vector $v \in Y$ satisfying $\|v\|_Y = 1$ and $(v, \Psi(\bar\alpha, x_d))_Y = 0$ is given, and define $y_d^s := \Psi(\bar\alpha, x_d) + sv$ for all $s \in \mathbb{R}$. Then, it holds*

$$\bar\alpha \notin \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 \quad \forall s \in \mathbb{R} \text{ with } |s| > \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)} \right)^{1/2} \|\Psi(\bar\alpha, x_d)\|_Y, \quad (4.23)$$

i.e., $\bar\alpha$ is not a global minimum of (2.3) *for all $s \in \mathbb{R}$ that satisfy the condition in* (4.23).

Proof Let $\bar y_s$ denote an arbitrary but fixed element of $P_\Psi^{x_d}(y_d^s)$ for all $s \in \mathbb{R}$. Then, we may use (4.3) and the properties of v to compute (completely analogously to (4.11)) that

$$\|\bar y_s - y_d^s\|_Y^2 - \|\Psi(\bar\alpha, x_d) - y_d^s\|_Y^2 \leq \Theta(\Psi, x_d)\|\Psi(\bar\alpha, x_d)\|_Y^2 + (\Theta(\Psi, x_d) - 1)s^2$$

holds for all $s \in \mathbb{R}$. The above implies

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 < \|\Psi(\bar\alpha, x_d) - y_d^s\|_Y^2$$

for all $s \in \mathbb{R}$ that satisfy the condition in (4.23). This proves the claim. ■

By exploiting the observation in Lemma 4.12, we readily obtain:

Theorem 4.13 (Criterion for the Existence of Non-Optimal Stationary Points) *Suppose that Assumption 4.1 holds, that $\bar{\alpha} \in D$ is an arbitrary but fixed element of the interior of the set D, and that the function $\Psi(\cdot, x_d)\colon D \to Y$ is differentiable at $\bar{\alpha}$. Assume further that the linear hull $V := \operatorname{span}(\Psi(\bar{\alpha}, x_d), \partial_1 \Psi(\bar{\alpha}, x_d), ..., \partial_m \Psi(\bar{\alpha}, x_d)) \subset Y$ of the vector $\Psi(\bar{\alpha}, x_d)$ and the partial derivatives $\partial_i \Psi(\bar{\alpha}, x_d)$, $i = 1, ..., m$, of the map $\Psi(\cdot, x_d)\colon D \to Y$ at $\bar{\alpha}$ is not equal to Y. Then, for every arbitrary but fixed element v of the $(\cdot,\cdot)_Y$-orthogonal complement of V with $\|v\|_Y = 1$ and every $s \in \mathbb{R}$ satisfying*

$$|s| > \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)} \right)^{1/2} \|\Psi(\bar{\alpha}, x_d)\|_Y, \tag{4.24}$$

there exists a $\tau \in \{-1, 1\}$ such that $\bar{\alpha}$ is a spurious local minimum or a saddle point of the training problem (2.3) with label vector $y_d^{\tau s} := \Psi(\bar{\alpha}, x_d) + \tau s v$. Moreover, for every arbitrary but fixed $C > 0$, there exist uncountably many label vectors $y_d \in Y$ such that $\bar{\alpha}$ is a saddle point or a spurious local minimum of (2.3), such that

$$\inf_{\bar{y} \in P_\Psi^{x_d}(y_d)} \|\Psi(\bar{\alpha}, x_d) - \bar{y}\|_Y \geq C \qquad \text{and} \qquad \inf_{\bar{y} \in P_\Psi^{x_d}(y_d)} \frac{\|\Psi(\bar{\alpha}, x_d) - \bar{y}\|_Y}{\|\bar{y}\|_Y} \geq 1 - \frac{1}{C}, \tag{4.25}$$

and such that

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + C \leq \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 \tag{4.26}$$

holds. The absolute error between $\Psi(\bar{\alpha}, x_d)$ and every true best approximation of y_d can thus be made arbitrarily large, the relative error between $\Psi(\bar{\alpha}, x_d)$ and every true best approximation of y_d can be made larger than $1 - \varepsilon$ for all $\varepsilon > 0$, and the difference between the value of the loss function at $\bar{\alpha}$ and the optimal loss can be made arbitrarily large.

Proof Suppose that an $\bar{\alpha} \in D$ with the properties in the theorem is given and that $v \in Y$ is an arbitrary but fixed element of the orthogonal complement of V satisfying $\|v\|_Y = 1$. Define $y_d^s := \Psi(\bar{\alpha}, x_d) + sv$ for all $s \in \mathbb{R}$. Then, the differentiability of the map $\Psi(\cdot, x_d)\colon D \to Y$ at $\bar{\alpha}$, the properties of v, and the definition of y_d^s imply that

$$\begin{aligned} &\|\Psi(\bar{\alpha} + h, x_d) - y_d^s\|_Y^2 - \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2 \\ &\quad = 2 \left(\Psi(\bar{\alpha} + h, x_d) - \Psi(\bar{\alpha}, x_d), \frac{\Psi(\bar{\alpha} + h, x_d) - \Psi(\bar{\alpha}, x_d)}{2} - sv \right)_Y \\ &\quad = 2 \left(\sum_{i=1}^m h_i \partial_i \Psi(\bar{\alpha}, x_d), \frac{\Psi(\bar{\alpha} + h, x_d) - \Psi(\bar{\alpha}, x_d)}{2} - sv \right)_Y + o(\|h\|_2) = o(\|h\|_2) \end{aligned} \tag{4.27}$$

holds for all $s \in \mathbb{R}$ and all sufficiently small $h \in \mathbb{R}^m$, where the Landau symbol refers to the limit $\|h\|_2 \to 0$. Dividing by $\|h\|_2$ and passing to the limit in (4.27) yields that $\bar{\alpha}$ is a stationary point of (2.3) for every y_d^s, $s \in \mathbb{R}$, i.e., the gradient of the loss function of (2.3) vanishes at $\bar{\alpha}$. In combination with Lemma 4.12, it now follows immediately that $\bar{\alpha}$ has to be a spurious local minimum, a local maximum, or a saddle point of the training problem (2.3) with label vector y_d^s for all $s \in \mathbb{R}$ satisfying (4.24). Next, we show that, for every $s \in \mathbb{R}$

with (4.24), the point $\bar{\alpha}$ is a saddle point or a spurious local minimum of (2.3) for one of the vectors y_d^s and y_d^{-s}. To see this, let us assume that there exists an $s \in \mathbb{R}$ with (4.24) such that the latter is not the case. Then, $\bar{\alpha}$ has to be a local maximum of (2.3) for both y_d^s and y_d^{-s} and we obtain from the same calculation as in (4.27) that

$$\begin{aligned}&\|\Psi(\bar{\alpha}+h,x_d)-y_d^s\|_Y^2-\|\Psi(\bar{\alpha},x_d)-y_d^s\|_Y^2\\&\qquad=\|\Psi(\bar{\alpha}+h,x_d)-\Psi(\bar{\alpha},x_d)\|_Y^2+2\left(\Psi(\bar{\alpha}+h,x_d)-\Psi(\bar{\alpha},x_d),-sv\right)_Y\le 0\end{aligned}$$

and

$$\begin{aligned}&\|\Psi(\bar{\alpha}+h,x_d)-y_d^{-s}\|_Y^2-\|\Psi(\bar{\alpha},x_d)-y_d^{-s}\|_Y^2\\&\qquad=\|\Psi(\bar{\alpha}+h,x_d)-\Psi(\bar{\alpha},x_d)\|_Y^2+2\left(\Psi(\bar{\alpha}+h,x_d)-\Psi(\bar{\alpha},x_d),sv\right)_Y\le 0\end{aligned}$$

for all $h \in \mathbb{R}^m$ in a sufficiently small open ball around zero. Adding the above yields

$$2\|\Psi(\bar{\alpha}+h,x_d)-\Psi(\bar{\alpha},x_d)\|_Y^2\le 0$$

for all small h which can only be true if the function $\alpha \mapsto \Psi(\alpha, x_d)$ is constant in a small open neighborhood of $\bar{\alpha}$. But if this is the case, then $\bar{\alpha}$ is trivially also a local minimum of (2.3) (for both y_d^s and y_d^{-s}). The point $\bar{\alpha}$ is thus indeed always a spurious local minimum or a saddle point of (2.3) for at least one of the label vectors y_d^s and y_d^{-s} for all $s \in \mathbb{R}$ with (4.24). This proves the first claim of the theorem. To see that we can also achieve (4.25) and (4.26) for all arbitrary but fixed $C > 0$, we note that (4.3) implies that

$$\|y_d^s\|_Y-\|\bar{y}_s\|_Y\le\|\bar{y}_s-y_d^s\|_Y\le\Theta(\Psi,x_d)^{1/2}\|y_d^s\|_Y$$

holds for all $\bar{y}_s \in P_\Psi^{x_d}(y_d^s)$ and that, as a consequence,

$$\inf_{\bar{y}\in P_\Psi^{x_d}(y_d^s)}\|\bar{y}\|_Y\ge\left(1-\Theta(\Psi,x_d)^{1/2}\right)\left(\|\Psi(\bar{\alpha},x_d)\|_Y^2+s^2\right)^{1/2}\qquad\forall s\in\mathbb{R}.$$

We thus have

$$\inf_{\bar{y}\in P_\Psi^{x_d}(y_d^s)}\|\Psi(\bar{\alpha},x_d)-\bar{y}\|_Y\ge\left(1-\Theta(\Psi,x_d)^{1/2}\right)\left(\|\Psi(\bar{\alpha},x_d)\|_Y^2+s^2\right)^{1/2}-\|\Psi(\bar{\alpha},x_d)\|_Y\to\infty$$

as well as

$$\inf_{\bar{y}\in P_\Psi^{x_d}(y_d^s)}\frac{\|\Psi(\bar{\alpha},x_d)-\bar{y}\|_Y}{\|\bar{y}\|_Y}\ge 1-\frac{\|\Psi(\bar{\alpha},x_d)\|_Y}{\left(1-\Theta(\Psi,x_d)^{1/2}\right)\left(\|\Psi(\bar{\alpha},x_d)\|_Y^2+s^2\right)^{1/2}}\to 1$$

and (again by (4.3))

$$\begin{aligned}&\inf_{\alpha\in D}\|\Psi(\alpha,x_d)-y_d^s\|_Y^2-\|\Psi(\bar{\alpha},x_d)-y_d^s\|_Y^2\\&\qquad\le\Theta(\Psi,x_d)\|y_d^s\|_Y^2-\|sv\|_Y^2=\Theta(\Psi,x_d)\|\Psi(\bar{\alpha},x_d)\|_Y^2+(\Theta(\Psi,x_d)-1)s^2\to-\infty\end{aligned}$$

for $|s| \to \infty$. This shows that (4.25) and (4.26) hold for every arbitrary but fixed constant $C > 0$ provided $|s|$ is large enough. In combination with what we already know about the vectors y_d^s, this establishes the second assertion of the theorem and completes the proof. ■

Remark 4.14 *As already pointed out in the introduction, the fact that the saddle points and spurious local minima in Theorem 4.13 can be made arbitrarily bad is not a mere consequence of the conicity condition* I). *Indeed, if we naively scale the vectors appearing in Theorem 4.13 by a factor $\gamma > 0$, then this factor clearly cancels out in the second estimate of* (4.25) *and it is not possible to ensure that the relative error in* (4.25) *becomes larger than $1 - \varepsilon$ for every $\varepsilon > 0$ by passing to the limit $\gamma \to \infty$.*

Note that the assumptions on the linear hull $\operatorname{span}(\Psi(\bar{\alpha}, x_d), \partial_1\Psi(\bar{\alpha}, x_d), ..., \partial_m\Psi(\bar{\alpha}, x_d))$ in Theorem 4.13 are trivially satisfied if $m + 1 < \dim(Y) = n \dim(\mathcal{Y})$ holds, i.e., if the product of the number of training pairs in (1.1) and the dimension of the output space $\mathcal{Y}$ exceeds the number of parameters in the considered approximation scheme by more than one. In this non-overparameterized case, Theorem 4.13 yields that *every* point of differentiability of the function $\Psi(\cdot, x_d)\colon D \to Y$ is an arbitrarily bad saddle point or spurious local minimum of (2.3) for uncountably many choices of the label vector y_d in the situation of Assumption 4.1. Compare also with Corollary 5.15 in Section 5.2 in this context. However, as already mentioned, such a non-overparameterization is not necessary to be able to apply the last theorem. To do so, it suffices to show that a local affine-linear approximation of the function $\Psi(\cdot, x_d)\colon D \to Y$ is not surjective (in contrast to, e.g., Theorem 4.11 which requires that the image of the map $\Psi(\cdot, x_d)\colon D \to Y$ itself is not dense in Y). In Lemma 4.19, we will prove that, for approximation schemes on an Euclidean space $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$, that, after reordering the entries of the vector $\alpha \in \mathbb{R}^m$ as a tuple $(\beta, A) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_\mathcal{X}}$ with some $p, q \in \mathbb{N}$ satisfying $m = p + qd_\mathcal{X}$, can be written in the form $\psi(\alpha, \chi) = \phi(\beta, A\chi)$ with a differentiable function ϕ (and thus in particular for neural networks with differentiable activations), points with the property $\operatorname{span}(\Psi(\bar{\alpha}, x_d), \partial_1\Psi(\bar{\alpha}, x_d), ..., \partial_m\Psi(\bar{\alpha}, x_d)) \neq Y$ always exist if the number $d_\mathcal{X} + 1$ is smaller than n. This shows that Theorem 4.13 can be used to establish the existence of spurious local minima or saddle points that are arbitrarily far away from global optima in many situations arising in practice.

If we not only know that the linear hull $\operatorname{span}(\Psi(\bar{\alpha}, x_d), \partial_1\Psi(\bar{\alpha}, x_d), ..., \partial_m\Psi(\bar{\alpha}, x_d))$ is not equal to Y, but even that $\bar{\alpha}$ possesses an open neighborhood $U \subset D$ such that the image of U under $\Psi(\cdot, x_d)\colon D \to Y$ is contained in a proper subspace V of Y, then we can prove that the construction in Theorem 4.13 always produces spurious local minima:

Theorem 4.15 (Criterion for the Existence of Spurious Local Minima) *Consider the situation in Assumption 4.1 and suppose that a point $\bar{\alpha} \in D$ is given such that there exist an open set $U \subset D$ with $\bar{\alpha} \in U$ and a subspace V of Y satisfying $\Psi(U, x_d) \subset V \neq Y$. Then, for every arbitrary but fixed element v of the $(\cdot,\cdot)_Y$-orthogonal complement of V with $\|v\|_Y = 1$ and every $s \in \mathbb{R}$ satisfying* (4.24), *the point $\bar{\alpha}$ is a spurious local minimum of the training problem* (2.3) *with label vector $y_d^s := \Psi(\bar{\alpha}, x_d) + sv$ that satisfies a quadratic growth condition of the form*

$$\|\Psi(\alpha, x_d) - y_d^s\|_Y^2 - \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2 \geq \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d)\|_Y^2 \qquad \forall \alpha \in U. \tag{4.28}$$

Further, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in Y$ such that $\bar{\alpha}$ is a spurious local minimum of (2.3) *and such that* (4.25), (4.26), *and* (4.28) *hold.*

Proof Suppose that an $\bar{\alpha} \in D$ satisfying the assumptions of the theorem is given and that $v \in Y$ is an arbitrary but fixed vector with $\|v\|_Y = 1$ that is $(\cdot,\cdot)_Y$-orthogonal to V. Then, it follows from the inclusion $\Psi(\bar{\alpha}, x_d) \in \Psi(U, x_d) \subset V$ that v is orthogonal to $\Psi(\bar{\alpha}, x_d)$ and we may invoke Lemma 4.12 to deduce that $\bar{\alpha}$ is not a global minimum of (2.3) when the training label vector is chosen as $y_d^s = \Psi(\bar{\alpha}, x_d) + sv$ with an $s \in \mathbb{R}$ satisfying (4.24). From exactly the same calculation as in (4.27), we further obtain that

$$\begin{aligned}\|\Psi(\alpha, x_d) - y_d^s\|_Y^2 - \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2 &= \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d) - sv\|_Y^2 - \|sv\|_Y^2 \\ &= \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d)\|_Y^2\end{aligned}$$

holds for all $\alpha \in U$. This shows that $\bar{\alpha}$ satisfies the growth condition (4.28) for all y_d^s and, in combination with our first observation, that $\bar{\alpha}$ is indeed a spurious local minimum of (2.3) for all y_d^s with an $s \in \mathbb{R}$ satisfying (4.24). To complete the proof, it remains to show that, for every $C > 0$, there exist uncountably many vectors y_d such that $\bar{\alpha}$ is a spurious local minimum of (2.3) and such that the estimates (4.25), (4.26), and (4.28) hold. This, however, follows completely analogously to the proof of Theorem 4.13. ■

We would like to emphasize that the last result does not require any form of differentiability. Under slightly stronger assumptions on the function $\Psi(\cdot, x_d)\colon D \to Y$, we can also analyze the size of the set of label vectors y_d that give rise to training problems (2.3) with spurious local minima in the situation of Theorem 4.15:

Theorem 4.16 (An Open Cone of Label Vectors with Spurious Local Minima) *Consider the situation in Assumption 4.1 and suppose that there exists a subspace V of Y with $V \neq Y$ such that, for every $z \in V$, there exist an $\bar{\alpha} \in D$ and an open set $U \subset D$ satisfying $\bar{\alpha} \in U$, $z = \Psi(\bar{\alpha}, x_d)$, and $\Psi(U, x_d) \subset V$. Denote the orthogonal complement of V in Y with $V^\perp$ (so that $Y = V \oplus V^\perp$). Then, the training problem* (2.3) *possesses at least one spurious local minimum satisfying a growth condition of the form* (4.28) *for all label vectors $y_d \in Y$ that are elements of the open cone*

$$K := \left\{ y_d^1 + y_d^2 \in Y \;\middle|\; y_d^1 \in V,\, y_d^2 \in V^\perp,\, \|y_d^2\|_Y > \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|y_d^1\|_Y \right\}. \tag{4.29}$$

Further, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in K$ such that at least one of the spurious local minima of (2.3) *satisfies* (4.25), (4.26), *and* (4.28), *and if $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$ holds, then the cone K in* (4.29) *is equal to $Y \setminus V$ and* (2.3) *possesses spurious local minima for all y_d that are not elements of V.*

Proof Consider an arbitrary but fixed element $y_d = y_d^1 + y_d^2 \in V \oplus V^\perp$ of the cone K. Then, our assumptions on the map $\Psi(\cdot, x_d)\colon D \to Y$ imply that we can find an $\bar{\alpha} \in D$ and an open set $U \subset D$ satisfying $\bar{\alpha} \in U$, $y_d^1 = \Psi(\bar{\alpha}, x_d)$, and $\Psi(U, x_d) \subset V \neq Y$. Define $v := y_d^2/\|y_d^2\|_Y$ and $s := \|y_d^2\|_Y$. (Note that y_d^2 cannot be zero by the definition of K.) Then, it clearly holds $v \in V^\perp$, $\|v\|_Y = 1$, $y_d = \Psi(\bar{\alpha}, x_d) + sv$, and

$$\left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|\Psi(\bar{\alpha}, x_d)\|_Y = \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|y_d^1\|_Y < \|y_d^2\|_Y = s.$$

By applying Theorem 4.15, it now follows immediately that $\bar{\alpha}$ is a spurious local minimum of the problem (2.3) with label vector y_d that satisfies a growth condition of the form (4.28).

This proves the first part of the theorem. To establish that there exist uncountably many vectors $y_d \in K$ with arbitrarily bad spurious local minima and that the identity $K = Y \setminus V$ holds in the case $\text{cl}_Y(\Psi(D, x_d)) = Y$, it suffices to invoke Theorem 4.15 and the definition of the number $\Theta(\Psi, x_d)$ in (4.2). This completes the proof. ■

Note that the inequalities in (4.24) and (4.29) again link the approximation properties of the map $\Psi(\cdot, x_d)\colon D \to Y$ to properties of the loss landscape of the training problem (2.3) (cf. Definition 4.3). If $\Theta(\Psi, x_d) \to 1$ holds, i.e., if the behavior of $\Psi(\cdot, x_d)\colon D \to Y$ approximates that of a linear approximation scheme, then the distance between the point $\Psi(\bar{\alpha}, x_d)$ and the label vectors y_d that cause the parameter $\bar{\alpha}$ to be a saddle point or a spurious local minimum of (2.3) in the situation of Theorem 4.13 tends to infinity and the cone K in Theorem 4.16 degenerates. If, on the other hand, $\Theta(\Psi, x_d)$ tends to zero, i.e., if the expressiveness of the considered nonlinear approximation scheme relative to Y increases and we approach the case $\text{cl}_Y(\Psi(D, x_d)) = Y$, then the vectors y_d in Theorem 4.13 can be chosen arbitrarily close to the point $\Psi(\bar{\alpha}, x_d)$ and the cone K in Theorem 4.16 exhausts the set $Y \setminus V$. We remark that, in combination with the results on the stability properties of the best approximation map $P_\Psi^{x_d}$ in Theorem 4.9, the above observations give a quite good impression of the issues that one has to deal with when considering training problems with squared loss for nonlinear approximation schemes satisfying the conditions in Assumption 4.1 and of how these issues are related to the number $\Theta(\Psi, x_d)$ in Definition 4.3. We will get back to this topic in Section 5.2, where we demonstrate that Theorems 4.9 and 4.16 in particular apply to neural networks that involve activation functions with an affine segment.

We conclude this subsection with a result that demonstrates that the subspace property in Theorem 4.15 is not only relevant for the existence of spurious local minima but also for the stability and uniqueness of global solutions of (2.3) in the case $\text{cl}_Y(\Psi(D, x_d)) = Y$, i.e., in the situation where every vector $y_d \in Y$ is realizable.

Theorem 4.17 (Instability and Nonuniqueness in the Realizable Case) *Consider the situation in Assumption 4.1 and suppose that a point $\bar{\alpha} \in D$ is given such that there exist an open set $U \subset D$ with $\bar{\alpha} \in U$ and a subspace V of Y satisfying $\Psi(U, x_d) \subset V \neq Y$. Assume further that $\text{cl}_Y(\Psi(D, x_d)) = Y$ holds and that v is an arbitrary but fixed element of the orthogonal complement of V in Y with $\|v\|_Y = 1$. Define $\bar{y}_d := \Psi(\bar{\alpha}, x_d)$ and $y_d^s := \bar{y}_d + sv$ for all $s \in \mathbb{R} \setminus \{0\}$. Then, the solution map*

$$Y \ni y_d \mapsto \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \subset D$$

of the problem (2.3) is discontinuous at $\bar{y}_d$ in the sense that the following is true:

$$\bar{\alpha} \in \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - \bar{y}_d\|_Y^2, \quad y_d^s \xrightarrow{s \to 0} \bar{y}_d, \quad U \cap \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 = \emptyset \;\; \forall s \in \mathbb{R} \setminus \{0\}. \tag{4.30}$$

Further, in this situation, there exists a family of parameters $\{\alpha_s\}_{s \in \mathbb{R} \setminus \{0\}}$ satisfying

$$\{\alpha_s\}_{s \in \mathbb{R} \setminus \{0\}} \subset D \setminus U \qquad \text{and} \qquad \lim_{s \to 0} \|\Psi(\alpha_s, x_d) - \bar{y}_d\|_Y^2 = 0 = \|\Psi(\bar{\alpha}, x_d) - \bar{y}_d\|_Y^2.$$

The problem (2.3) with label vector $\bar{y}_d$ is thus not uniquely solvable in the generalized sense that it possesses at least one minimizing sequence that does not converge to $\bar{\alpha}$.

Proof In the situation of the theorem, it follows from Definition 4.3 and the assumption $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$ that $\Theta(\Psi, x_d) = 0$ holds, and we obtain from Theorem 4.15 that $\bar{\alpha}$ is a local minimum of the problem (2.3) with label vector y_d^s for all $s \in \mathbb{R} \setminus \{0\}$ that satisfies

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 = 0 < s^2 = \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2 \leq \|\Psi(\tilde{\alpha}, x_d) - y_d^s\|_Y^2 \tag{4.31}$$

for all $\tilde{\alpha} \in U$ and all $s \in \mathbb{R} \setminus \{0\}$. The above implies in particular that none of the points in U can be a global minimizer of (2.3) with label vector y_d^s for all $s \in \mathbb{R} \setminus \{0\}$. This establishes (4.30). On the other hand, (4.31) also yields that, for every $s \in \mathbb{R} \setminus \{0\}$, we can find an $\alpha_s \in D \setminus U$ with

$$|s| > \|\Psi(\alpha_s, x_d) - y_d^s\|_Y \geq \|\Psi(\alpha_s, x_d) - \bar{y}_d\|_Y - |s|.$$

The existence of a family $\{\alpha_s\}_{s \in \mathbb{R} \setminus \{0\}}$ with the properties in the second part of the theorem now follows immediately. This completes the proof. ∎

4.3 Spurious Local Minima in the Presence of a Regularization Term

A standard technique to overcome the ill-posedness of an inverse problem (i.e., the nonexistence or instability of solutions) is to add a regularization term to the objective function that penalizes the size of the involved parameters. In the context of training problems of the type (2.3), this approach has the additional advantage that it allows to promote desirable sparsity properties of the vectors $\alpha \in D$ that are obtained from the optimization procedure, cf. Pörner (2018); Hofmann (2013); Pieper and Petrosyan (2020); Wen et al. (2016); Yoon and Hwang (2017) and the references therein. The aim of this subsection is to demonstrate that, as far as the existence of spurious local minima and the expressiveness of nonlinear approximation schemes are concerned, adding a regularization term to the loss function in (2.3) can also have detrimental effects. The main idea in the following is to exploit that many commonly used approximation instruments possess a "linear" lowest level in the sense that they depend on the product of the input variable χ and a matrix A whose entries are part of the parameter vector α. In the situation of our standing Assumption 4.1, this structural property can be expressed as follows:

Assumption 4.18 (Linearity of the Lowest Level)

- *It holds $D = \mathbb{R}^m$, $\mathcal{X} = \mathbb{R}^{d_x}$, and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}}) = (\mathbb{R}^{d_y}, \|\cdot\|_2)$ with some $d_\chi, d_y \in \mathbb{N}$.*
- *There exists a function $\phi\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ such that, after reordering the vector α and reshaping it into a tuple $(\beta, A) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_x}$ with $m = p + qd_\chi$, we can write $\psi(\alpha, \chi) = \phi(\beta, A\chi)$ for all $\alpha \in \mathbb{R}^m$ and all $\chi \in \mathbb{R}^{d_x}$.*

Note that standard neural networks trivially satisfy the above conditions as they involve an affine-linear transformation on the lowest level, see Section 5.2. A main feature of approximation schemes satisfying Assumption 4.18 is that they behave polynomially when linearized at points $\alpha \in D$ that, after reordering, yield the matrix $A = 0$. More precisely, we have the following result:

Lemma 4.19 (Polynomial First- and Second-Order Approximations) *Suppose that Assumptions 4.1 and 4.18 hold. Consider further an arbitrary but fixed $\bar{\alpha} \in \mathbb{R}^m$ which, after the reshaping procedure in Assumption 4.18, takes the form $(\bar{\beta}, 0) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_\chi}$. Then, the following is true:*

i) If the map $\phi\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ in Assumption 4.18 is differentiable at $(\bar{\beta}, 0)$, then there exists a subspace $V_1 \subset Y$ of dimension at most $d_y(d_\chi + 1)$ satisfying

$$\Psi(\bar{\alpha}, x_d) + \partial_\alpha \Psi(\bar{\alpha}, x_d) \langle h \rangle \in V_1 \quad \forall h \in \mathbb{R}^m.$$

ii) If the map $\phi\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ in Assumption 4.18 is continuously differentiable in an open neighborhood of the point $(\bar{\beta}, 0)$ and twice differentiable at $(\bar{\beta}, 0)$, then there exist a subspace $V_1 \subset Y$ of dimension at most $d_y(d_\chi + 1)$ and a subspace $V_2 \subset Y$ of dimension at most $\frac{1}{2} d_y (d_\chi + 2)(d_\chi + 1)$ satisfying

$$V_1 \subset V_2, \qquad \Psi(\bar{\alpha}, x_d) + \partial_\alpha \Psi(\bar{\alpha}, x_d) \langle h \rangle \in V_1 \quad \forall h \in \mathbb{R}^m, \tag{4.32}$$

and

$$\Psi(\bar{\alpha}, x_d) + \partial_\alpha \Psi(\bar{\alpha}, x_d) \langle h \rangle + \frac{1}{2} \partial_\alpha^2 \Psi(\bar{\alpha}, x_d) \langle h, h \rangle \in V_2 \quad \forall h \in \mathbb{R}^m. \tag{4.33}$$

Here, $\partial_\alpha \Psi(\bar{\alpha}, x_d) \langle h \rangle$ and $\partial_\alpha^2 \Psi(\bar{\alpha}, x_d) \langle h, h \rangle$ denote the first and the second derivative of the function Ψ w.r.t. the variable α at $\bar{\alpha}$ evaluated at h and (h, h), respectively.

Proof Suppose that an $\bar{\alpha} \in \mathbb{R}^m$ satisfying the assumptions in point i) of the lemma is given. Then, it holds

$$\begin{aligned} \psi(\bar{\alpha} + h, \chi) &= \phi(\bar{\beta} + \tilde{h}, 0 + H\chi) = \phi(\bar{\beta}, 0) + \phi'(\bar{\beta}, 0)\langle \tilde{h}, H\chi \rangle + o\big(\|(\tilde{h}, H\chi)\|_2\big) \\ &= \phi(\bar{\beta}, 0) + \phi'(\bar{\beta}, 0)\langle \tilde{h}, 0 \rangle + \sum_{i=1}^{d_\chi} \phi'(\bar{\beta}, 0)\langle 0, He_i \rangle \chi_i + o\big(\|(\tilde{h}, H\chi)\|_2\big) \end{aligned}$$

for all arbitrary but fixed $\chi \in \mathbb{R}^{d_\chi}$ and all $h \in \mathbb{R}^m$ which, after the reshaping procedure in Assumption 4.18, take the form $(\tilde{h}, H) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_\chi}$. Here, with $\phi'(\bar{\beta}, 0) \langle \cdot \rangle : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ we mean the first derivative of the function ϕ at $(\bar{\beta}, 0)$, the Landau symbol refers to the limit $\|(\tilde{h}, H\chi)\|_2 \to 0$, and e_i, $i = 1, ..., d_\chi$, are the unit vectors in $\mathbb{R}^{d_\chi}$. Note that the above implies in particular that

$$\psi(\bar{\alpha}, \chi) + \partial_\alpha \psi(\bar{\alpha}, \chi) \langle h \rangle = \phi(\bar{\beta}, 0) + \phi'(\bar{\beta}, 0)\langle \tilde{h}, 0 \rangle + \sum_{i=1}^{d_\chi} \phi'(\bar{\beta}, 0)\langle 0, He_i \rangle \chi_i = P_{\bar{\beta}, h}(\chi)$$

holds for all $h \in \mathbb{R}^m$ and all arbitrary but fixed $\chi \in \mathbb{R}^{d_\chi}$, where $P_{\bar{\beta}, h}\colon \mathbb{R}^{d_\chi} \to \mathbb{R}^{d_y}$ is an affine map that depends only on $\bar{\beta}$ and h. The function $\chi \mapsto \psi(\bar{\alpha}, \chi) + \partial_\alpha \psi(\bar{\alpha}, \chi) \langle h \rangle$ is thus contained in a subspace of dimension $d_y(d_\chi + 1)$ that is independent of h, namely the space of vector-valued polynomials of degree at most one which map $\mathbb{R}^{d_\chi}$ to $\mathbb{R}^{d_y}$. Since the function Ψ is defined by $\Psi(\alpha, x_d) := \big\{\psi(\alpha, \chi_d^k)\big\}_{k=1}^n$, i.e., by plugging in certain values of χ, the first claim of the lemma now follows immediately.

To establish the assertion in ii), we can proceed completely along the same lines as in the first part of the proof. By Taylor's formula, we obtain that

$$\begin{aligned}
&\psi(\bar{\alpha} + h, \chi) = \phi(\bar{\beta} + \tilde{h}, 0 + H\chi) \\
&\quad = \phi(\bar{\beta}, 0) + \phi'(\bar{\beta}, 0)\langle \tilde{h}, H\chi\rangle + \frac{1}{2}\phi''(\bar{\beta}, 0)\langle (\tilde{h}, H\chi), (\tilde{h}, H\chi)\rangle + o\big(\|(\tilde{h}, H\chi)\|_2^2\big) \\
&\quad = \phi(\bar{\beta}, 0) + \phi'(\bar{\beta}, 0)\langle \tilde{h}, 0\rangle + \frac{1}{2}\phi''(\bar{\beta}, 0)\langle (\tilde{h}, 0), (\tilde{h}, 0)\rangle \\
&\qquad + \sum_{i=1}^{d_\chi} \left(\phi'(\bar{\beta}, 0)\langle 0, He_i\rangle + \frac{1}{2}\phi''(\bar{\beta}, 0)\langle (0, He_i), (\tilde{h}, 0)\rangle + \frac{1}{2}\phi''(\bar{\beta}, 0)\langle (\tilde{h}, 0), (0, He_i)\rangle\right)\chi_i \\
&\qquad + \frac{1}{2}\sum_{i=1}^{d_\chi}\sum_{j=1}^{d_\chi} \phi''(\bar{\beta}, 0)\langle (0, He_i), (0, He_j)\rangle \chi_i\chi_j + o\big(\|(\tilde{h}, H\chi)\|_2^2\big)
\end{aligned}$$

holds for all arbitrary but fixed $\chi \in \mathbb{R}^{d_\chi}$ and all vectors $h \in \mathbb{R}^m$ which, after reshaping, take the form $(\tilde{h}, H) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_\chi}$. Here, $\phi'(\bar{\beta}, 0)\langle\cdot\rangle : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ again denotes the first and $\phi''(\bar{\beta}, 0)\langle\cdot,\cdot\rangle : (\mathbb{R}^p \times \mathbb{R}^q)^2 \to \mathbb{R}^{d_y}$ the second derivative of ϕ at $(\bar{\beta}, 0)$. The above yields

$$\psi(\bar{\alpha}, \chi) + \partial_\alpha \psi(\bar{\alpha}, \chi)\langle h\rangle = P_{\bar{\beta}, h}(\chi) \quad \forall \chi \in \mathbb{R}^{d_\chi}$$

and

$$\psi(\bar{\alpha}, \chi) + \partial_\alpha \psi(\bar{\alpha}, \chi)\langle h\rangle + \frac{1}{2}\partial_\alpha^2 \psi(\bar{\alpha}, \chi)\langle h, h\rangle = Q_{\bar{\beta}, h}(\chi) \quad \forall \chi \in \mathbb{R}^{d_\chi}$$

for all $h \in \mathbb{R}^m$, where $P_{\bar{\beta},h} \colon \mathbb{R}^{d_\chi} \to \mathbb{R}^{d_y}$ and $Q_{\bar{\beta},h} \colon \mathbb{R}^{d_\chi} \to \mathbb{R}^{d_y}$ are vector-valued polynomials of degree at most one and two, respectively, that depend only on $\bar{\beta}$ and h. The functions $\chi \mapsto \psi(\bar{\alpha}, \chi) + \partial_\alpha \psi(\bar{\alpha}, \chi)\langle h\rangle$ and $\chi \mapsto \psi(\bar{\alpha}, \chi) + \partial_\alpha \psi(\bar{\alpha}, \chi)\langle h\rangle + \frac{1}{2}\partial_\alpha^2 \psi(\bar{\alpha}, \chi)\langle h, h\rangle$ are thus contained in subspaces of dimension $d_y(d_\chi + 1)$ and $\frac{1}{2}d_y(d_\chi + 2)(d_\chi + 1)$, respectively, that are independent of h, namely the spaces of polynomials of degree at most one and two, respectively, which map $\mathbb{R}^{d_\chi}$ to $\mathbb{R}^{d_y}$. By again exploiting the definition of Ψ, the assertion of ii) now follows immediately. This completes the proof. ■

We would like to point out that Lemma 4.19 is not only interesting for the study of regularized training problems but also for the results on the existence of non-optimal critical points that we have derived in Section 4.2. Indeed, as a straightforward consequence of Theorem 4.13 and Lemma 4.19, we obtain:

Corollary 4.20 (Critical Points in the Presence of a Linear Lowest Level) *Suppose that Assumptions 4.1 and 4.18 are satisfied and that $d_\chi + 1 < n$ holds. Consider further an arbitrary but fixed $\bar{\alpha} \in \mathbb{R}^m$ which, after reshaping, takes the form $(\bar{\beta}, 0) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_\chi}$, and assume that the map $\phi \colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ in Assumption 4.18 is differentiable at $(\bar{\beta}, 0)$. Then, for every $\varepsilon > 0$, there exist uncountably many label vectors $y_d \in Y$ satisfying*

$$\left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|\Psi(\bar{\alpha}, x_d)\|_Y < \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y < \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|\Psi(\bar{\alpha}, x_d)\|_Y + \varepsilon$$

such that $\bar{\alpha}$ is a spurious local minimum or a saddle point of (2.3)*. Further, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in Y$ such that $\bar{\alpha}$ is a spurious local minimum or a saddle point of* (2.3) *and such that* (4.25) *and* (4.26) *hold.*

Proof From Lemma 4.19, we obtain that, in the considered situation, the linear hull of the vectors $\Psi(\bar{\alpha}, x_d)$ and $\partial_i \Psi(\bar{\alpha}, x_d)$, $i = 1, ..., m$, is contained in a subspace of dimension $d_y(d_\chi + 1) < d_y n = \dim(Y) = \dim(\mathcal{Y}^n)$. This shows that $\bar{\alpha}$ satisfies the assumptions of Theorem 4.13. By invoking this theorem, the claim of the corollary follows immediately. ∎

Note that, in the case $d_\chi + 1 \geq n$, for every training set (χ_d^k, y_d^k), $k = 1, ..., n$, with $\dim(\operatorname{span}(\chi_d^2 - \chi_d^1, ..., \chi_d^n - \chi_d^1)) = n - 1$, we can find an affine-linear $T\colon \mathbb{R}^{d_\chi} \to \mathbb{R}^{d_y}$ with $T(\chi_d^k) = y_d^k$ for all $k = 1, ..., n$. The condition $d_\chi + 1 < n$ in Corollary 4.20 is thus directly related to the approximation properties of affine functions. In Section 5.2, Corollary 4.20 will in particular allow us to show that, for neural networks with differentiable activations, there is always a subspace of parameters α in $\mathbb{R}^m$ that can be turned into arbitrarily bad saddle points or spurious local minima of (2.3) by choosing appropriate label vectors $y_d \in Y$, see Corollary 5.16.

To show that regularized versions of the training problem (2.3) can indeed possess spurious local minima, we will use that, by adding a regularization term to the loss function of (2.3), the stationary points in Corollary 4.20 can be transformed into local minimizers. This leads to:

Theorem 4.21 (Spurious Local Minima in the Presence of Regularization Terms) *Suppose that Assumptions 4.1 and 4.18 are satisfied and that* $\frac{1}{2}(d_\chi + 2)(d_\chi + 1) < n$ *holds. Consider an arbitrary but fixed vector* $\bar{\alpha} \in \mathbb{R}^m$ *which, after the reshaping procedure in Assumption 4.18, takes the form* $(\bar{\beta}, 0) \in \mathbb{R}^p \times \mathbb{R}^{q \times d_\chi}$*, and assume that the function* ϕ *in Assumption 4.18 is continuously differentiable in an open neighborhood of* $(\bar{\beta}, 0)$ *and twice differentiable at* $(\bar{\beta}, 0)$*. Assume further that a function* $g\colon \mathbb{R}^m \to [0, \infty)$ *with* $g(0) = 0$ *is given such that there exist a constant* $c > 0$ *and an open neighborhood* $U \subset \mathbb{R}^m$ *of the origin with* $g(z) \geq c\|z\|_2^2$ *for all* $z \in U$*. Then, for every arbitrary but fixed* $C > 0$*, there exist uncountably many combinations of training label vectors* $y_d \in Y$ *and regularization parameters* $\nu > 0$ *such that* $\bar{\alpha}$ *is a spurious local minimum of the regularized training problem*

$$\min_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu g(\alpha - \bar{\alpha}) \tag{4.34}$$

that satisfies a local quadratic growth condition of the form

$$\|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu g(\alpha - \bar{\alpha}) \geq \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 + \nu g(0) + \varepsilon\|\alpha - \bar{\alpha}\|_2^2 \qquad \forall \alpha \in B_r(\bar{\alpha}) \tag{4.35}$$

for some $\varepsilon, r > 0$ *and*

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu g(\alpha - \bar{\alpha}) + C \leq \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 + \nu g(0). \tag{4.36}$$

Proof Suppose that a point $\bar{\alpha}$ satisfying the assumptions of the theorem is given and let V_2 denote the subspace from part ii) of Lemma 4.19. Then, it follows from the inequality $\frac{1}{2}(d_\chi + 2)(d_\chi + 1) < n$ that $\dim(V_2) < \dim(Y)$ holds and that there exists a $v \in Y$ that is orthogonal to V_2 and satisfies $\|v\|_Y = 1$. Let us again define $y_d^s := \Psi(\bar{\alpha}, x_d) + sv \in Y$, $s \in \mathbb{R}$, and assume that $C > 0$ is an arbitrary but fixed constant. Then, we obtain completely analogously to the proof of Theorem 4.13 that there exists an $M > 0$ with

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 + C + 2 \leq \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2$$

for all $s \in \mathbb{R}$ with $|s| > M$. Note that the above in particular implies that, for all $s \in \mathbb{R}$ with $|s| > M$, we can find an $\tilde{\alpha}_s \in D$ with

$$\|\Psi(\tilde{\alpha}_s, x_d) - y_d^s\|_Y^2 + C + 1 \leq \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2.$$

If we now choose a $\nu_s > 0$ for all $s \in \mathbb{R}$ with $|s| > M$ such that $\nu_s g(\tilde{\alpha}_s - \bar{\alpha}) < 1$ holds and exploit the identity $g(0) = 0$, then it readily follows that

$$\|\Psi(\tilde{\alpha}_s, x_d) - y_d^s\|_Y^2 + \nu_s g(\tilde{\alpha}_s - \bar{\alpha}) + C \leq \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2 + \nu_s g(0).$$

This establishes (4.36). It remains to prove that, for all of the above y_d^s, ν_s, and s, the vector $\bar{\alpha}$ is indeed a spurious local minimum of (4.34) that satisfies a local quadratic growth condition of the form (4.35). To this end, we note that the binomial identities, the definition of y_d^s, the choice of v, and our assumptions on g and $\bar{\alpha}$ imply that

$$\begin{aligned}
&\|\Psi(\alpha, x_d) - y_d^s\|_Y^2 + \nu_s g(\alpha - \bar{\alpha}) - \|\Psi(\bar{\alpha}, x_d) - y_d^s\|_Y^2 - \nu_s g(0) \\
&\quad \geq \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d) - sv\|_Y^2 - \|sv\|_Y^2 + \nu_s c \|\alpha - \bar{\alpha}\|_2^2 \\
&\quad = \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d)\|_Y^2 - 2\left(\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d), sv\right)_Y + \nu_s c \|\alpha - \bar{\alpha}\|_2^2 \\
&\quad = \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d)\|_Y^2 + \nu_s c \|\alpha - \bar{\alpha}\|_2^2 \\
&\qquad - 2\left(\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d) - \partial_\alpha \Psi(\bar{\alpha}, x_d)\langle \alpha - \bar{\alpha}\rangle - \frac{1}{2}\partial_\alpha^2 \Psi(\bar{\alpha}, x_d)\langle \alpha - \bar{\alpha}, \alpha - \bar{\alpha}\rangle, sv\right)_Y \\
&\quad = \|\Psi(\alpha, x_d) - \Psi(\bar{\alpha}, x_d)\|_Y^2 + \nu_s c \|\alpha - \bar{\alpha}\|_2^2 + o(\|\alpha - \bar{\alpha}\|_2^2)
\end{aligned} \tag{4.37}$$

holds for all $\alpha \in D$ with $\alpha - \bar{\alpha} \in U$, where the Landau symbol refers to the limit $\|\alpha - \bar{\alpha}\|_2 \to 0$. By choosing sufficiently small $\varepsilon, r > 0$, it now follows immediately that $\bar{\alpha}$ is a spurious local minimum of (4.34) that satisfies (4.35). This completes the proof. ■

As the last result demonstrates, the addition of a regularization term to the objective function of (2.3) may create spurious local minima by introducing a bias towards certain values of the parameter vector α. The proof of Theorem 4.21 further shows that these effects are a direct consequence of the approximation property II) and will typically appear when the regularization parameter ν is too small relative to the size of the training label vector y_d (see the conditions $|s| > M$ and $\nu_s g(\tilde{\alpha}_s - \bar{\alpha}) < 1$). Choosing ν too large, however, is also not a good idea as the following result demonstrates:

Theorem 4.22 (Loss of Approximation Property II) by Regularization) *Suppose that Assumptions 4.1 and 4.18 hold, that the function* $\phi\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^{d_y}$ *in Assumption 4.18 is continuously differentiable in an open neighborhood of the origin and twice differentiable at the origin, and that* $\frac{1}{2}(d_\chi + 2)(d_\chi + 1) < n$ *and* $\phi(0, 0) = 0$. *Assume further that a function* $g\colon \mathbb{R}^m \to [0, \infty)$ *with* $g(0) = 0$ *is given such that there exist constants* $c_1, c_2 > 0$ *and an open neighborhood* $U \subset \mathbb{R}^m$ *of the origin satisfying* $g(z) \geq c_1 \|z\|_2^2$ *for all* $z \in U$ *and* $g(z) \geq c_2$ *for all* $z \in \mathbb{R}^m \setminus U$. *Then, for every arbitrary but fixed regularization parameter* $\nu > 0$, *there exist uncountably many label vectors* $y_d \in Y \setminus \{0\}$ *such that* $\bar{\alpha} = 0$ *is the unique global solution of the problem*

$$\min_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu g(\alpha). \tag{4.38}$$

Proof The assumptions of the theorem imply that $\bar\alpha = 0 \cong (0,0) \in \mathbb{R}^p \times \mathbb{R}^{q\times d_\chi}$ satisfies the conditions in part ii) of Lemma 4.19 with $\frac{1}{2}d_y(d_\chi+2)(d_\chi+1) < \dim(Y)$. We can thus again find a proper subspace V_2 of Y such that the first and second derivatives of the function $\Psi(\cdot,x_d)\colon D \to Y$ at $\bar\alpha = 0$ are contained in V_2 in the sense of (4.32) and (4.33), choose a vector $v \in Y$ with $\|v\|_Y = 1$ that is orthogonal to V_2, and define $y_d^s := \Psi(0,x_d) + sv \in Y$ for all $s \in \mathbb{R}$. Since ϕ is twice differentiable at the origin and since U is open, we further obtain that there exists an $r > 0$ with $\alpha \in U$ and

$$\left\|\Psi(\alpha,x_d) - \Psi(0,x_d) - \partial_\alpha\Psi(0,x_d)\langle\alpha\rangle - \frac{1}{2}\partial_\alpha^2\Psi(0,x_d)\langle\alpha,\alpha\rangle\right\|_Y \le \frac{1}{2}\|\alpha\|_2^2$$

for all $\alpha \in \mathbb{R}^m$ with $\|\alpha\|_2 \le r$. Note that this estimate and exactly the same calculation as in (4.37) yield

$$\begin{aligned}
&\|\Psi(\alpha,x_d) - y_d^s\|_Y^2 + \nu g(\alpha) - \|\Psi(0,x_d) - y_d^s\|_Y^2 - \nu g(0)\\
&\ge \nu c_1\|\alpha\|_2^2 - 2\left(\Psi(\alpha,x_d) - \Psi(0,x_d) - \partial_\alpha\Psi(0,x_d)\langle\alpha\rangle - \frac{1}{2}\partial_\alpha^2\Psi(0,x_d)\langle\alpha,\alpha\rangle, sv\right)_Y\\
&\ge (\nu c_1 - |s|)\|\alpha\|_2^2
\end{aligned}$$

for all α with $\|\alpha\|_2 \le r$ and all arbitrary but fixed $\nu > 0$. For all $\alpha \in \mathbb{R}^m$ with $\|\alpha\|_2 > r$, on the other hand, we have

$$\|\Psi(\alpha,x_d) - y_d^s\|_Y^2 + \nu g(\alpha) - \|\Psi(0,x_d) - y_d^s\|_Y^2 - \nu g(0) \ge \nu\min(c_2, c_1 r^2) - s^2.$$

By combining the last two estimates, it follows that $\bar\alpha = 0$ is the unique global solution of (4.38) with label vector y_d^s for all $s \in \mathbb{R}$ with $|s| < \min(\nu c_1, (\nu c_2)^{1/2}, (\nu c_1)^{1/2} r)$. ■

Theorem 4.22 shows that, although the function $\Psi(\cdot,x_d)\colon D \to Y$ is able to provide a best approximation for every $y_d \neq 0$ that is better than the origin by II), the regularized problem (4.38) may very well possess the optimal solution $\bar\alpha = 0$ with the associated vector $\Psi(0,x_d) = 0$ for nonzero label vectors y_d. (Recall that $\phi(0,0) = 0$ implies $\psi(0,\chi) = 0$ for all $\chi \in \mathbb{R}^{d_\chi}$ so that we indeed have $\Psi(0,x_d) = 0$ here.) Adding a regularization term to the objective function of (2.3) thus impairs the approximation properties of the map $\Psi(\cdot,x_d)\colon D \to Y$ and compromises the property II) that distinguishes the function ψ from a linear approximation scheme in the first place, cf. the discussion in Section 3. Note that the estimate $|s| < \min(\nu c_1, (\nu c_2)^{1/2}, (\nu c_1)^{1/2} r)$ in the proof of Theorem 4.22 suggests that these effects get worse when ν increases as in this case the solution $\bar\alpha = 0$ is obtained from (4.38) for vectors y_d with larger norms. We would like to point out that studying how the approximation properties of the global solutions of the problem (4.38) are affected by the choice of the tuple (ν, g) is in general far from straightforward. The main reason for this is that, in a nonlinear approximation scheme, there is typically no immediate connection between, e.g., the norm of the parameter vector α and the size of the output $\Psi(\alpha,x_d)$ so that it is a-priori often completely unclear which features of the elements of the image $\Psi(D,x_d)$ are penalized by a term of the form $\nu g(\alpha)$.

We conclude this section with a result that shows that the addition of a regularization term to the objective function of (2.3) does not necessarily remove the instability and nonuniqueness of solutions, either:

Theorem 4.23 (Instability and Nonuniqueness in the Regularized Case) *Suppose that Assumptions 4.1 and 4.18 are satisfied and that* $\frac{1}{2}(d_\chi+2)(d_\chi+1) < n$ *holds. Consider an arbitrary but fixed* $\bar\alpha \in \mathbb{R}^m$ *which, after the reshaping procedure in Assumption 4.18, takes the form* $(\bar\beta, 0) \in \mathbb{R}^p \times \mathbb{R}^{q\times d_\chi}$*, and assume that the function* ϕ *in Assumption 4.18 is continuously differentiable in an open neighborhood of* $(\bar\beta,0)$ *and twice differentiable at* $(\bar\beta,0)$*. Assume further that a function* $g\colon \mathbb{R}^m \to [0,\infty)$ *with* $g(0)=0$ *is given such that there exist a constant* $c>0$ *and an open neighborhood* $U\subset\mathbb{R}^m$ *of the origin with* $g(z)\geq c\|z\|_2^2$ *for all* $z\in U$*. Then, there exist uncountably many combinations of regularization parameters* $\nu>0$ *and label vectors* $y_d\in Y$ *such that there exist an* $s_0\geq 0$*, a sequence* $\{y_d^s\}_{s>s_0}\subset Y$*, and an open neighborhood* $\tilde U\subset\mathbb{R}^m$ *of* $\bar\alpha$ *satisfying* $y_d^s\to y_d$ *for* $s\to s_0$*,*

$$\tilde U \cap \operatorname*{arg\,min}_{\alpha\in D} \|\Psi(\alpha,x_d) - y_d^s\|_Y^2 + \nu g(\alpha-\bar\alpha) = \emptyset \qquad \forall s>s_0, \tag{4.39}$$

and

$$\bar\alpha \in \operatorname*{arg\,min}_{\alpha\in D} \|\Psi(\alpha,x_d) - y_d\|_Y^2 + \nu g(\alpha-\bar\alpha). \tag{4.40}$$

Further, there are uncountably many tuples $(y_d,\nu)\in Y\times(0,\infty)$ *with the above properties such that there exists a family* $\{\alpha_s\}_{s>s_0}\subset D\setminus\tilde U$ *satisfying*

$$\lim_{s\to s_0} \|\Psi(\alpha_s,x_d)-y_d\|_Y^2 + \nu g(\alpha_s-\bar\alpha) = \|\Psi(\bar\alpha,x_d)-y_d\|_Y^2 + \nu g(0).$$

The solutions of the regularized training problem (4.34) *thus possess the same nonuniqueness and instability properties as the overparameterized problem in Theorem 4.17.*

Proof Suppose that $\bar\alpha$ is an arbitrary but fixed point satisfying the assumptions of the theorem and let $v\in Y$ and y_d^s, $s\in\mathbb{R}$, be defined as in the proof of Theorem 4.21. Then, from exactly the same construction as in the proof of Theorem 4.21, we obtain that there exist uncountably many tuples $(\nu, y_d^{\bar s})$, $\nu>0$, $\bar s>0$, with w.l.o.g. different ν such that

$$\inf_{\alpha\in D} \|\Psi(\alpha,x_d)-y_d^{\bar s}\|_Y^2 + \nu g(\alpha-\bar\alpha) < \|\Psi(\bar\alpha,x_d)-y_d^{\bar s}\|_Y^2 + \nu g(0) \tag{4.41}$$

holds. Let us fix such a tuple $(\nu, y_d^{\bar s})$ and consider the auxiliary function

$$F\colon [0,\bar s]\to[0,\infty), \qquad s\mapsto \inf_{\alpha\in D}\|\Psi(\alpha,x_d)-y_d^s\|_Y^2 + \nu g(\alpha-\bar\alpha).$$

We claim that this function is Lipschitz continuous. Indeed, for all $s_1,s_2\in[0,\bar s]$ and every sequence $\{\alpha_i\}\subset D$ with

$$\lim_{i\to\infty}\|\Psi(\alpha_i,x_d)-y_d^{s_1}\|_Y^2 + \nu g(\alpha_i-\bar\alpha) = F(s_1),$$

we obtain from the non-negativity of g and the definitions of F, $y_d^{s_1}$, and $\{\alpha_i\}$ that

$$\begin{aligned}
0 &\leq \limsup_{i\to\infty}\|\Psi(\alpha_i,x_d)-\Psi(\bar\alpha,x_d)\|_Y \\
&\leq \limsup_{i\to\infty}\left(\|\Psi(\alpha_i,x_d)-y_d^{s_1}\|_Y^2 + \nu g(\alpha_i-\bar\alpha)\right)^{1/2} + s_1\|v\|_Y \\
&= \left(\inf_{\alpha\in D}\|\Psi(\alpha,x_d)-y_d^{s_1}\|_Y^2 + \nu g(\alpha-\bar\alpha)\right)^{1/2} + s_1 \\
&\leq \left(\|\Psi(\bar\alpha,x_d)-y_d^{s_1}\|_Y^2 + \nu g(0)\right)^{1/2} + s_1 \\
&\leq 2\bar s,
\end{aligned}$$

and, as a consequence,

$$
\begin{aligned}
F(s_1) &= \lim_{i\to\infty} \|\Psi(\alpha_i, x_d) - y_d^{s_1}\|_Y^2 + \nu g(\alpha_i - \bar\alpha) \\
&= \lim_{i\to\infty} \|\Psi(\alpha_i, x_d) - \Psi(\bar\alpha, x_d) - s_2 v - (s_1 - s_2)v\|_Y^2 + \nu g(\alpha_i - \bar\alpha) \\
&\geq \limsup_{i\to\infty} \|\Psi(\alpha_i, x_d) - y_d^{s_2}\|_Y^2 + \nu g(\alpha_i - \bar\alpha) - 2\|\Psi(\alpha_i, x_d) - \Psi(\bar\alpha, x_d) - s_2 v\|_Y |s_1 - s_2| \\
&\geq F(s_2) - 6\bar s\,|s_1 - s_2|.
\end{aligned}
$$

After exchanging the roles of s_1 and s_2, we thus have $|F(s_1) - F(s_2)| \leq 6\bar s\,|s_1 - s_2|$ and F is Lipschitz continuous on $[0, \bar s]$ as claimed. Consider now the value

$$
s_0 := \inf\left\{\tilde s \in [0, \bar s] \,\middle|\, F(s) < \|\Psi(\bar\alpha, x_d) - y_d^s\|_Y^2 + \nu g(0) \quad \forall s \in (\tilde s, \bar s]\right\}. \tag{4.42}
$$

Then, it follows from the continuity of F, the definition of y_d^s, (4.41), and the trivial identity $F(0) = 0 = \|\Psi(\bar\alpha, x_d) - y_d^0\|_Y^2 + \nu g(0)$ that s_0 satisfies $0 \leq s_0 < \bar s$ and

$$
\inf_{\alpha\in D} \|\Psi(\alpha, x_d) - y_d^{s_0}\|_Y^2 + \nu g(\alpha - \bar\alpha) = \|\Psi(\bar\alpha, x_d) - y_d^{s_0}\|_Y^2 + \nu g(0).
$$

This shows that $\bar\alpha$ satisfies (4.40) for $y_d := y_d^{s_0}$. To see that the above construction also yields (4.39), we note that the same estimates as in (4.37) imply that

$$
\begin{aligned}
&\|\Psi(\alpha, x_d) - y_d^s\|_Y^2 + \nu g(\alpha - \bar\alpha) - \|\Psi(\bar\alpha, x_d) - y_d^s\|_Y^2 - \nu g(0) \\
&\quad\geq \|\Psi(\alpha, x_d) - \Psi(\bar\alpha, x_d)\|_Y^2 + \nu c\|\alpha - \bar\alpha\|_2^2 \\
&\qquad - 2\bar s\left\|\Psi(\alpha, x_d) - \Psi(\bar\alpha, x_d) - \partial_\alpha\Psi(\bar\alpha, x_d)\langle\alpha - \bar\alpha\rangle - \frac{1}{2}\partial_\alpha^2\Psi(\bar\alpha, x_d)\langle\alpha - \bar\alpha, \alpha - \bar\alpha\rangle\right\|_Y \\
&\quad\geq 0
\end{aligned}
$$

holds for all $0 \leq s < \bar s$ and all $\alpha \in \mathbb{R}^m$ in a sufficiently small open neighborhood $\tilde U \subset \mathbb{R}^m$ of $\bar\alpha$ that depends only on Ψ, g, ν, and the bound $\bar s$. As $\bar\alpha$ is not a global minimizer of the problem (4.34) for all label vectors y_d^s with $s_0 < s < \bar s$ by the definition of s_0 in (4.42), this shows that $\bar\alpha$, $\tilde U$, and the vectors y_d^s, $s \in (s_0, \bar s)$, indeed satisfy (4.39). Since the convergence $y_d^s \to y_d^{s_0}$ for $s \to s_0$ is trivial, this proves the first part of the theorem. (Note that we indeed end up with uncountably many different tuples (ν, y_d) with the desired properties here since, although we have modified the label vector during the course of the proof, we have not altered the regularization parameter ν.) To establish the second assertion of the theorem, it suffices to note that the above considerations and the triangle inequality imply that, for all $s \in (s_0, \bar s)$, there has to exist an $\alpha_s \in D \setminus \tilde U$ with

$$
\begin{aligned}
\|\Psi(\bar\alpha, x_d) - y_d^{s_0}\|_Y^2 + \nu g(0) &= \|\Psi(\bar\alpha, x_d) - y_d^s\|_Y^2 + \nu g(0) + o(1) \\
&> \|\Psi(\alpha_s, x_d) - y_d^s\|_Y^2 + \nu g(\alpha_s - \bar\alpha) + o(1) \\
&= \|\Psi(\alpha_s, x_d) - y_d^{s_0}\|_Y^2 + \nu g(\alpha_s - \bar\alpha) + o(1) \\
&\geq \inf_{\alpha\in D} \|\Psi(\alpha, x_d) - y_d^{s_0}\|_Y^2 + \nu g(\alpha - \bar\alpha) + o(1) \\
&= \|\Psi(\bar\alpha, x_d) - y_d^{s_0}\|_Y^2 + \nu g(0) + o(1),
\end{aligned}
$$

where the Landau symbol refers to the limit $(s_0, \bar s) \ni s \to s_0$. This proves the claim. ∎

5. Tangible Examples

With the abstract results of Section 4 in place, we are in the position to turn our attention to tangible examples and applications. In what follows, we will first consider a classical free-knot spline interpolation scheme that is closely related to neural networks with ReLU activation functions, see Section 5.1. After this, we then turn our attention to training problems for neural networks with various architectures, see Section 5.2.

5.1 Free-Knot Linear Spline Interpolation

The first example that we consider in this section is a special instance of a dictionary approximation approach that generalizes classical piecewise linear interpolation - the so-called free-knot linear spline interpolation method. This technique is based on the idea to not only adapt the function values at the nodes of a linear spline to the function that is to be approximated, but also to vary the nodes of the underlying mesh. For details on this topic and its background, we refer to DeVore (1998) and Daubechies et al. (2019). The setting that we consider in this subsection is the following:

Assumption 5.1 (Assumptions for the Discussion of Free-Knot Interpolation)

- $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ *and the norm of* $\mathcal{Y}$ *is just the absolute value function,*
- $n \in \mathbb{N}$ *and* $n \geq 2$*,*
- $x_d := \{\chi_d^k\}_{k=1}^n \in \mathcal{X}^n$ *is an arbitrary but fixed vector satisfying* $\chi_d^1 < \chi_d^2 < ... < \chi_d^n$*,*
- $m = 2p$*,* $p \in \mathbb{N}$*,* $p \geq 3$*, and* $D \subset \mathbb{R}^m \cong \mathbb{R}^p \times \mathbb{R}^p$ *is defined by*

$$D := \left\{ \alpha = (\beta, \gamma) \in \mathbb{R}^p \times \mathbb{R}^p \,\middle|\, \gamma_1 < \gamma_2 < ... < \gamma_p \right\}, \tag{5.1}$$

- $\psi \colon D \times \mathcal{X} \to \mathcal{Y}$ *is defined by*

$$\psi((\beta,\gamma),\chi) := \begin{cases} \beta_1 & \text{if } \chi \leq \gamma_1 \\ \dfrac{\gamma_{j+1} - \chi}{\gamma_{j+1} - \gamma_j}\beta_j + \dfrac{\chi - \gamma_j}{\gamma_{j+1} - \gamma_j}\beta_{j+1} & \text{if } \gamma_j < \chi \leq \gamma_{j+1},\ j \in \{1, ..., p-1\} \\ \beta_p & \text{if } \chi > \gamma_p. \end{cases} \tag{5.2}$$

Note that the above situation is trivially covered by our general Assumption 2.1. To simplify the notation, in what follows, we will again use the abbreviations collected in Definition 2.2. For every arbitrary but fixed training label vector $y_d = \{y_d^k\}_{k=1}^n \in Y = \mathcal{Y}^n$, our squared-loss training problem (1.1) thus takes the form

$$\min_{\alpha = (\beta,\gamma) \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 = \frac{1}{2n} \sum_{k=1}^n \left(\psi(\alpha, \chi_d^k) - y_d^k\right)^2. \tag{5.3}$$

Similarly to the example from Section 3, the problem (5.3) models the task of finding a vector of breakpoints $\gamma \in \mathbb{R}^p$ and a coefficient vector $\beta \in \mathbb{R}^p$ such that the map in (5.2) possesses function values at the points χ_d^k, $k = 1, ..., n$, that fit the given data vector $y_d := \{y_d^k\}_{k=1}^n$ optimally in the least-squares sense. Solving (5.3) for the function (5.2) is thus a problem of nonlinear regression. As already mentioned, the free-knot interpolation scheme (5.2) is closely related to neural networks involving the ReLU activation function. In fact, it has been shown in Daubechies et al. (2019) that the image of a ReLU-based network with a real in- and output is always contained in the image of the scheme (5.2) for a sufficiently large p and that the image of (5.2) is always contained in the image of a ReLU-network of sufficient depth and width. See (Daubechies et al., 2019, Sections 3 and 4) for precise results on this topic. We will see below that, as far as the optimization landscape and the stability of the training problem (5.3) are concerned, ReLU-networks and the scheme (5.2) share many common properties as well.

To be able to apply the abstract results of Section 4 to the scheme (5.2) and the problem (5.3), we have to check if the conditions in Assumption 4.1 are satisfied. This, however, is an easy task:

Lemma 5.2 (Conicity and Improved Expressiveness of (5.2)) *In the situation of Assumption 5.1, the function $\Psi(\cdot, x_d)\colon D \to Y$ associated with x_d and the free-knot spline interpolation scheme ψ possesses the properties* I) *and* II)*. Moreover, the number $\Theta(\Psi, x_d)$ in* (4.2) *associated with x_d and the function ψ in* (5.2) *satisfies $\Theta(\Psi, x_d) \leq 1 - 1/n$.*

Proof If we consider an arbitrary but fixed $y \in \Psi(D, x_d)$, then there exists a tuple $(\beta, \gamma) \in D$ with $\Psi((\beta, \gamma), x_d) = y$ and it follows immediately from (5.1) and (5.2) that we also have $(s\beta, \gamma) \in D$ and $sy \in \Psi(D, x_d)$ for all $s \in \mathbb{R}$. This establishes I). To prove II), let us assume that an arbitrary but fixed $y_d \in Y \setminus \{0\}$ is given. Then, there exists at least one $l \in \{1, ..., n\}$ with $y_d^l \neq 0$. Consider now a vector $(\beta, \gamma) \in D$ with

$$\begin{gathered}\beta_1 = \beta_p = 0, \qquad \beta_j = y_d^l \quad \forall j = 2, ..., p-1, \qquad \gamma_1 < ... < \gamma_p, \\ \chi_d^l = \gamma_2, \qquad \chi_d^k \notin [\gamma_1, \gamma_p] \quad \forall k \neq l.\end{gathered} \tag{5.4}$$

(Note that such a vector always exists by our assumptions on the entries of x_d.) Then, (5.2) and the definition of Ψ yield that $\Psi((\beta, \gamma), x_d) = y_d^l\, e_l$ holds, where e_l denotes the l-th unit vector of $Y = \mathbb{R}^n$, and we obtain from (5.3) that

$$\|\Psi((\beta, \gamma), x_d) - y_d\|_Y^2 = \frac{1}{2n} \sum_{k \neq l} \left|y_d^k\right|^2 < \|y_d\|_Y^2.$$

This proves II). To finally establish the inequality $\Theta(\Psi, x_d) \leq 1 - 1/n$, it suffices to note that every $y_d \in Y$ with $\|y_d\|_Y = 1$ possesses at least one entry that has an absolute value greater than or equal to $\sqrt{2}$. In combination with the construction in (5.4), this yields

$$\Theta(\Psi, x_d) = \sup_{y_d \in Y, \|y_d\|_Y = 1} \left(\inf_{y \in \mathrm{cl}_Y(\Psi(D, x_d))} \|y - y_d\|_Y^2 \right) \leq 1 - \frac{1}{n}$$

and completes the proof. ∎

Note that the estimate $\Theta(\Psi, x_d) \leq 1 - 1/n$ in Lemma 5.2 is very pessimistic. (We have, after all, only used one node to establish it.) Deriving better estimates for this quantity not only for the scheme (5.2) but also for the neural networks discussed in the next subsection is an interesting topic and more precise results on the number $\Theta(\Psi, x_d)$ and its dependence on x_d and n would certainly improve the understanding of the expressiveness of nonlinear approximation schemes - in particular in view of the inequality (4.3). We leave this topic for future research.

Next, we collect some results on the mapping properties of the function $\Psi(\cdot, x_d)\colon D \to Y$ that make it possible to decide which theorems of Section 4 are applicable to (5.2):

Lemma 5.3 (Mapping Properties of (5.2)) *In the situation of Assumption 5.1, the following is true:*

i) If $n \leq p$ holds, then we have $\Psi(D, x_d) = Y$.

ii) If $n > p$ holds, then we have $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$.

iii) Define

$$V := \Big\{ \{z_k\}_{k=1}^n \in Y \,\Big|\, \exists a, b \in \mathbb{R} \text{ such that } z_k = a\chi_d^k + b \;\; \forall k = 1, ..., n \Big\}. \tag{5.5}$$

Then, for every element z of the subspace V, there exist a point $\bar{\alpha} \in D$ and an open set $U \subset D$ satisfying $\bar{\alpha} \in U$, $z = \Psi(\bar{\alpha}, x_d)$, and $\Psi(U, x_d) \subset V$.

iv) If $n > 3p$ holds, then, for every $\bar{\alpha} \in D$, there exist an open set $U \subset D$ with $\bar{\alpha} \in U$ and a subspace V of Y with $V \neq Y$ such that $\Psi(U, x_d) \subset V$ holds.

Proof If we suppose that $n \leq p$ holds and that $y = \{y_k\}_{k=1}^n$ is an arbitrary but fixed element of the space $Y = \mathbb{R}^n$, then every $\alpha = (\beta, \gamma) \in D$ with $\gamma_k := \chi_d^k$ and $\beta_k := y_k$ for all $k = 1, ..., n$ satisfies $\Psi(\alpha, x_d) = y$. This establishes the equality in i).

To prove ii), let us assume that there is a situation with $p < n$ and $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$. Then, the density of the set $\Psi(D, x_d)$ in Y implies that we can find an $\bar{\alpha} = (\bar{\beta}, \bar{\gamma}) \in D$ which satisfies $|\psi(\bar{\alpha}, \chi_d^k) - (-1)^k| < 0.1$ for all $k = 1, ..., n$. Consider now an interval of the form $(\chi_d^{2l}, \chi_d^{2l+2})$, $l = 1, ..., \lfloor (n-2)/2 \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Then, by the properties of $\bar{\alpha}$, we have $\psi(\bar{\alpha}, \chi_d^{2l}) > 0.9$, $\psi(\bar{\alpha}, \chi_d^{2l+1}) < -0.9$, and $\psi(\bar{\alpha}, \chi_d^{2l+2}) > 0.9$, and it follows that the function $\mathbb{R} \ni \chi \mapsto \psi(\bar{\alpha}, \chi) \in \mathbb{R}$ attains its minimum on $[\chi_d^{2l}, \chi_d^{2l+2}]$ in the open interval $(\chi_d^{2l}, \chi_d^{2l+2})$. As the map $\chi \mapsto \psi(\bar{\alpha}, \chi)$ is piecewise linear, we also know that this minimum has to be attained at one of the breakpoints $\bar{\gamma}_j$, $j = 1, ..., p$. Note that the cases $j = 1$ and $j = p$ are impossible here since the map $\chi \mapsto \psi(\bar{\alpha}, \chi)$ is constant on the left of $\bar{\gamma}_1$ and on the right of $\bar{\gamma}_p$, since we know that the minimal function value in $[\chi_d^{2l}, \chi_d^{2l+2}]$ is smaller than -0.9, and since $\psi(\bar{\alpha}, \chi_d^{2l}) > 0.9$ and $\psi(\bar{\alpha}, \chi_d^{2l+2}) > 0.9$. In summary, we may thus conclude that each of the open intervals $(\chi_d^{2l}, \chi_d^{2l+2})$, $l = 1, ..., \lfloor (n-2)/2 \rfloor$, has to contain (at least) one $\bar{\gamma}_j$ with $j \in \{2, ..., p-1\}$ and $\psi(\bar{\alpha}, \bar{\gamma}_j) = \bar{\beta}_j < -0.9$. Using exactly the same arguments (with maxima instead of minima), we also obtain that each of the intervals $(\chi_d^{2l-1}, \chi_d^{2l+1})$, $l = 1, ..., \lfloor (n-1)/2 \rfloor$, has to contain (at least) one $\bar{\gamma}_j$ with $j \in \{2, ..., p-1\}$ and $\psi(\bar{\alpha}, \bar{\gamma}_j) = \bar{\beta}_j > 0.9$. Since the intervals in both of these groups are mutually disjoint and due to the different conditions on the function values, it now follows immediately that there have to be at least $\lfloor (n-2)/2 \rfloor + \lfloor (n-1)/2 \rfloor + 2 = n$ breakpoints in (5.2). Thus, $p \geq n$ which contradicts our assumption $p < n$. This establishes ii).

Next, we prove iii): Let $z = \{z_k\}_{k=1}^n$ be an arbitrary but fixed element of the space V in (5.5) with associated $a, b \in \mathbb{R}$, i.e., $z_k = a\chi_d^k + b$ for all $k = 1, ..., n$. Then, we can clearly find a point $\bar{\alpha} = (\bar{\beta}, \bar{\gamma}) \in D$ with the properties

$$\bar{\gamma}_1 < \chi_d^1 < \chi_d^2 < ... < \chi_d^n < \bar{\gamma}_2 < ... < \bar{\gamma}_p, \qquad \bar{\beta}_1 = a\bar{\gamma}_1 + b, \quad \text{and} \quad \bar{\beta}_2 = a\bar{\gamma}_2 + b. \tag{5.6}$$

Due to (5.2), such an $\bar{\alpha} \in D$ trivially satisfies $z = \Psi(\bar{\alpha}, x_d)$, and from the strictness of the inequalities in (5.6) we obtain that there exists an open neighborhood $U \subset D$ of $\bar{\alpha}$ such that $\gamma_1 < \chi_d^1 < \chi_d^2 < ... < \chi_d^n < \gamma_2$ holds for all $\alpha = (\beta, \gamma) \in U$. Since the latter property implies that the map $\chi \mapsto \psi(\alpha, \chi)$ is affine-linear on the open interval (γ_1, γ_2) and that $\chi_d^k \in (\gamma_1, \gamma_2)$ holds for all $k = 1, ..., n$, it follows immediately that $\Psi(U, x_d) \subset V$. In summary, we thus have $\bar{\alpha} \in U$, $z = \Psi(\bar{\alpha}, x_d)$, and $\Psi(U, x_d) \subset V$ and the proof of iii) is complete.

To finally obtain iv), we note that, in the case $n > 3p$, every $\bar{\alpha} = (\bar{\beta}, \bar{\gamma}) \in D$ has to satisfy (at least) one of the following three conditions (as one may easily check by contradiction):

$$\chi_d^1 < \chi_d^2 < \bar{\gamma}_1, \qquad \bar{\gamma}_p < \chi_d^{n-1} < \chi_d^n,$$
$$\exists j \in \{1, ..., p-1\} \text{ and } l \in \{1, ..., n-2\}\colon \; \bar{\gamma}_j < \chi_d^l < \chi_d^{l+1} < \chi_d^{l+2} < \bar{\gamma}_{j+1}.$$

Due to the definition of ψ in (5.2) and the strictness of the involved inequalities, the above implies that every $\bar{\alpha} \in D$ admits an open neighborhood $U \subset D$ such that (at least) one of the following is true:

$$\Psi(U, x_d) \subset V_1 := \left\{ z \in Y = \mathbb{R}^n \,\middle|\, z_1 = z_2 \right\},$$
$$\Psi(U, x_d) \subset V_2 := \left\{ z \in Y = \mathbb{R}^n \,\middle|\, z_{n-1} = z_n \right\},$$
$$\Psi(U, x_d) \subset V_3 := \left\{ z \in Y = \mathbb{R}^n \,\middle|\, \frac{\chi_d^{l+2} - \chi_d^{l+1}}{\chi_d^{l+2} - \chi_d^l} z_l + \frac{\chi_d^{l+1} - \chi_d^l}{\chi_d^{l+2} - \chi_d^l} z_{l+2} = z_{l+1} \right\}$$

for some $l \in \{1, ..., n-2\}$. The assertion of iv) now follows immediately. This completes the proof of the lemma. ■

By invoking the results of Section 4, we now obtain (for example) the following for the free-knot linear spline interpolation scheme (5.2):

Corollary 5.4 (Properties of Squared-Loss Training Problems for (5.2)) *In the situation of Assumption 5.1, the following is true:*

i) **(Nonuniqueness and Instability of Best Approximations)** *If $n > p$ holds, then the best approximation map*

$$P_\Psi^{x_d}\colon Y \rightrightarrows Y, \qquad y_d \mapsto \underset{y \in \mathrm{cl}_Y(\Psi(D, x_d))}{\arg\min} \|y - y_d\|_Y^2,$$

associated with the free-knot linear spline interpolation scheme (5.2) *is set-valued and there exist uncountably many training label vectors $y_d \in Y$ satisfying $|P_\Psi^{x_d}(y_d)| > 1$. Moreover, the map $P_\Psi^{x_d}\colon Y \rightrightarrows Y$ is discontinuous in the sense that, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in Y$ such that there are sequences $\{y_d^l\}, \{\tilde{y}_d^l\} \subset Y$ with the properties in* (4.13).

ii) **(Existence of Spurious Local Minima)** *If $n > 2$ holds, then there exists an open nonempty cone $K \subset Y$ such that the training problem* (5.3) *possesses at least one spurious local minimum satisfying a growth condition of the form* (4.28) *for all $y_d \in K$. These spurious local minima can be arbitrarily bad in the sense that, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in K$ such that at least one of the spurious local minima of* (5.3) *satisfies* (4.25), (4.26), *and* (4.28). *The size of the cone K depends on the extent to which condition* II) *is satisfied, cf.* (4.29). *If $n \leq p$ holds, then the cone K consists of all vectors in Y that are not affine-linearly fittable, i.e., it holds $K = Y \setminus V$ with the subspace V in* (5.5).

iii) **(Every Point is a Potential Spurious Local Minimum in the Case $\mathbf{n > 3p}$)** *If $n > 3p$ holds, then, for every $\bar{\alpha} \in D$ and every arbitrary but fixed $C > 0$, there exist uncountably many label vectors y_d such that $\bar{\alpha}$ is a spurious local minimum of the training problem* (5.3) *that satisfies* (4.25), (4.26), *and a quadratic growth condition of the form* (4.28).

iv) **(Instability of Solutions in the Case $\mathbf{2 < n \leq p}$)** *If $2 < n \leq p$ holds, then the solution operator*

$$Y \ni y_d \mapsto \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \subset D$$

of the problem (5.3) *is discontinuous in the sense that there exist points $\bar{\alpha} \in D$ and vectors $y_d \in Y$ such that there are a family $\{y_d^s\}_{s>0} \subset Y$ and an open neighborhood $U \subset D$ of $\bar{\alpha}$ satisfying $y_d^s \to y_d$ for $s \to 0$,*

$$\bar{\alpha} \in \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2,$$

and

$$U \cap \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 = \emptyset \quad \forall s > 0.$$

v) **(Nonuniqueness of Solutions in the Case $\mathbf{2 < n \leq p}$)** *If $2 < n \leq p$ holds, then there exist choices of the training label vector y_d such that the problem* (5.3) *is not uniquely solvable in the sense of minimizing sequences. More precisely, there exist vectors $y_d \in Y$ such that there are an $\bar{\alpha} \in D$, an open set $U \subset D$, and a family $\{\alpha_s\}_{s>0}$ satisfying $\bar{\alpha} \in U$, $\{\alpha_s\}_{s>0} \subset D \setminus U$, and*

$$\lim_{s \to 0} \|\Psi(\alpha_s, x_d) - y_d\|_Y^2 = \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 = \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2.$$

Proof To prove the various claims of the corollary, it suffices to combine Lemmas 5.2 and 5.3 with Theorems 4.9 and 4.15 to 4.17 in Section 4. ■

We remark that, using the same ideas as in the proofs of Lemma 5.2, Lemma 5.3, and Corollary 5.4, one can also show that the results of Section 4 can be applied to other nonlinear approximation schemes that are based on the idea to not only optimize the coefficient vector w.r.t. a certain basis but also the choice of the basis itself. We omit a detailed discussion of this topic to avoid overloading this paper and will focus on the consequences that our results have for neural networks instead.

5.2 Neural Networks

Next, we apply our abstract results to neural networks with vector-valued in- and output. The setting that we consider in this subsection is as follows:

Assumption 5.5 (Standing Assumptions for the Study of Neural Networks)

- $\mathcal{X} := \mathbb{R}^{d_\chi}$, $\mathcal{Y} := \mathbb{R}^{d_y}$, $d_\chi, d_y \in \mathbb{N}$, *and* $\mathcal{Y}$ *is endowed with the Euclidean norm* $\|\cdot\|_2$*,*
- $n \in \mathbb{N}$, $n \geq 2$, *and* $x_d := \{\chi_d^k\}_{k=1}^n \in \mathcal{X}^n$ *is an arbitrary but fixed training data vector satisfying* $\chi_d^j \neq \chi_d^k$ *for all* $j \neq k$*,*
- $w_i \in \mathbb{N}$, $i = 1, ..., L$, $L \in \mathbb{N}$, *are given numbers,* $w_0 := d_\chi$, $w_{L+1} := d_y$, *and the set* D *is defined by*

$$D := \left\{ \alpha = \{(A_i, b_i)\}_{i=1}^{L+1} \,\middle|\, A_i \in \mathbb{R}^{w_i \times w_{i-1}},\, b_i \in \mathbb{R}^{w_i} \ \forall i = 1, ..., L+1 \right\}, \tag{5.7}$$

- $\sigma_i \colon \mathbb{R} \to \mathbb{R}$, $i = 1, ..., L$, *are given activation functions,*
- $\varphi_i^{A_i,b_i} \colon \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i}$, $i = 1, ..., L+1$, *are the functions defined by*

$$\varphi_i^{A_i,b_i}(z) := \sigma_i\left(A_i z + b_i\right) \ \forall i = 1, ..., L, \qquad \varphi_{L+1}^{A_{L+1},b_{L+1}}(z) := A_{L+1} z + b_{L+1}, \tag{5.8}$$

where σ_i *acts componentwise on the entries of the vectors* $A_i z + b_i$*,*

- $\psi \colon D \times \mathcal{X} \to \mathcal{Y}$ *is defined by*

$$\psi(\alpha, \chi) := \left(\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ ... \circ \varphi_1^{A_1,b_1}\right)(\chi) \tag{5.9}$$

for all $\chi \in \mathcal{X}$ *and all* $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$*.*

Note that, for the sake of brevity and readability, in the remainder of this section, we will not reorder elements $\alpha = \{(A_i, b_i)\}_{i=1}^{L+1} = (A_{L+1}, b_{L+1}, ..., A_1, b_1)$ of the parameter space D in (5.7) as column vectors to conform with the notation of Section 4, i.e., we will not always explicitly state that we use the isomorphism

$$\mathbb{R}^{w_{L+1} \times w_L} \times \mathbb{R}^{w_{L+1}} \times ... \times \mathbb{R}^{w_1 \times w_0} \times \mathbb{R}^{w_1} \cong \mathbb{R}^m, \quad m := w_{L+1}(w_L+1) + ... + w_1(w_0+1), \tag{5.10}$$

when referring to the results of the previous sections. We will further again employ the abbreviations introduced in Definition 2.2 so that the squared-loss training problem for the neural network (5.9) reads as follows:

$$\min_{\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 = \frac{1}{2n} \sum_{k=1}^n \|\psi(\alpha, \chi_d^k) - y_d^k\|_2^2. \tag{5.11}$$

We would like to point out that the setting in Assumption 5.5 is a very general one as it not only allows for different widths of the L hidden layers of the network but also for the use of different activation functions. Compare, e.g., with the architectures considered in Daubechies et al. (2019) and Ding et al. (2020) in this context. As in Section 5.1, we begin our analysis of the approximation scheme (5.9) by checking whether the conditions I) and II) are satisfied. For I), we obtain:

Lemma 5.6 (Conicity of Neural Networks) *In the situation of Assumption 5.5, the function $\Psi(\cdot, x_d)\colon D \to Y$ associated with x_d and the network ψ in* (5.9) *satisfies* I).

Proof Since $\varphi_{L+1}^{sA_{L+1}, sb_{L+1}} \circ ... \circ \varphi_1^{A_1,b_1} = s\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ ... \circ \varphi_1^{A_1,b_1}$ holds for all $s \in \mathbb{R}$, the cone property in I) follows immediately. ■

Verifying condition II) is more delicate. In what follows, the main idea to establish this approximation property will be to first prove II) for neural networks with Heaviside-type activations of the form

$$\sigma_i(s) = \begin{cases} 0 & \text{if } s < 0 \\ c_i & \text{if } s = 0 \, , \\ 1 & \text{if } s > 0 \end{cases} \qquad c_i \in \mathbb{R}, \tag{5.12}$$

and to subsequently exploit that almost all activation functions that are currently used in the literature can emulate step functions of the form (5.12) by saturation. In combination with the observation in Lemma 4.2 that II) is a property of the closure $\mathrm{cl}_Y\left(\Psi(D, x_d)\right)$ and completely independent of how the elements in this set are realized or approximated by the parameters in D, this then immediately yields the desired condition II). We would like to point out that the above approach to the analysis of neural networks and, in view of the results of Section 4, the existence of saddle points and spurious local minima is conceptually very different from the techniques used, e.g., in Yun et al. (2019) and Goldblum et al. (2020) which primarily rely on the observation that many neural networks can locally imitate linear approximation schemes. In our analysis, the main step is not to exploit such a local linearity but, on the contrary, to reduce the problem to the situation where the activation functions are essentially binary and thus to the most nonlinear case possible. The starting point of our proof of II) is the following lemma:

Lemma 5.7 (A Separation Lemma) *Consider the situation in Assumption 5.5. Then, for every arbitrary but fixed $l \in \{1, ..., n\}$, there exist a matrix $A \in \mathbb{R}^{2\times d_x}$ and a vector $b \in \mathbb{R}^2$ satisfying*

$$A\chi_d^k + b \in (-\infty, 0)^2 \cup (0, \infty)^2 \quad \forall k \neq l \qquad \textit{and} \qquad A\chi_d^l + b \in (0, \infty) \times (-\infty, 0).$$

Proof To establish the assertion of the lemma, we first prove by induction w.r.t. $p \in \mathbb{N}$ that, for every collection of vectors $z_1, ..., z_p \in \mathbb{R}^{d_x} \setminus \{0\}$, there exists an $a \in \mathbb{R}^{d_x}$ with $a^T z_j \neq 0$ for all $j = 1, ..., p$. For $p = 1$, the existence of such an a is trivial as we can simply choose $a := z_1$. So let us assume that $p > 1$ holds. Then, the induction hypothesis yields that there exists an $a \in \mathbb{R}^{d_x}$ with $a^T z_j \neq 0$ for all $j = 1, ..., p-1$. If this a also satisfies $a^T z_p \neq 0$, then there is nothing left to show. If, on the other hand, $a^T z_p = 0$ holds, then we can find a small $\varepsilon > 0$ with $(a + \varepsilon z_p)^T z_j \neq 0$ for all $j = 1, ..., p-1$, and it follows immediately that the vector $\tilde{a} := a + \varepsilon z_p$ has all of the desired properties. This concludes the induction step.

Consider now an arbitrary but fixed $l \in \{1, ..., n\}$. Then, the above result and our assumption $\chi_d^k \neq \chi_d^j$ for all $k \neq j$ imply that there exists an $a \in \mathbb{R}^{d_x}$ with $a^T(\chi_d^k - \chi_d^l) \neq 0$

for all $k \neq l$, and we can find an $\varepsilon > 0$ with $a^T(\chi_d^k - \chi_d^l) \pm \varepsilon \neq 0$ and $\operatorname{sgn}(a^T(\chi_d^k - \chi_d^l) \pm \varepsilon) = \operatorname{sgn}(a^T(\chi_d^k - \chi_d^l))$ for all $k \neq l$. If we use this ε to define

$$A := \begin{pmatrix} a^T \\ a^T \end{pmatrix} \in \mathbb{R}^{2 \times d_x}, \qquad b := \begin{pmatrix} \varepsilon - a^T \chi_d^l \\ -\varepsilon - a^T \chi_d^l \end{pmatrix} \in \mathbb{R}^2,$$

then it holds

$$A\chi_d^k + b = \begin{pmatrix} a^T(\chi_d^k - \chi_d^l) + \varepsilon \\ a^T(\chi_d^k - \chi_d^l) - \varepsilon \end{pmatrix} \in \begin{cases} (-\infty, 0)^2 \cup (0, \infty)^2 & \text{if } k \neq l \\ (0, \infty) \times (-\infty, 0) & \text{if } k = l \end{cases}$$

as desired. This completes the proof. ■

Using Lemma 5.7, we can prove:

Lemma 5.8 (Approximation Property II) for Heaviside-Type Activations) *Consider the situation in Assumption 5.5. Suppose that $w_1 \geq 2$ holds and that there exist constants $c_i \in \mathbb{R}$ with*

$$\sigma_i(s) = \begin{cases} 0 & \text{if } s < 0 \\ c_i & \text{if } s = 0 \\ 1 & \text{if } s > 0 \end{cases} \qquad \forall i = 1, ..., L. \tag{5.13}$$

Then, the function $\Psi(\cdot, x_d) \colon D \to Y$ associated with x_d and the neural network ψ in (5.9) *possesses the property* II) *and the number $\Theta(\Psi, x_d)$ in* (4.2) *satisfies $\Theta(\Psi, x_d) \leq 1 - 1/n$.*

Proof Suppose that a $y_d \in Y \setminus \{0\}$ is given. Then, there exists at least one $l \in \{1, ..., n\}$ with $y_d^l \in \mathbb{R}^{d_y} \setminus \{0\}$, and it follows from our assumption $w_1 \geq 2$ and Lemma 5.7 that we can find a matrix $A \in \mathbb{R}^{2 \times d_x}$ and a vector $b \in \mathbb{R}^2$ such that the parameters

$$A_1 := \begin{pmatrix} A \\ 0_{(w_1 - 2) \times w_0} \end{pmatrix} \in \mathbb{R}^{w_1 \times w_0} \qquad \text{and} \qquad b_1 := \begin{pmatrix} b \\ 0_{(w_1 - 2)} \end{pmatrix} \in \mathbb{R}^{w_1} \tag{5.14}$$

satisfy

$$A_1 \chi_d^k + b_1 \in \begin{cases} \left[(-\infty, 0)^2 \times \{0_{(w_1-2)}\}\right] \cup \left[(0, \infty)^2 \times \{0_{(w_1-2)}\}\right] & \text{if } k \neq l \\ (0, \infty) \times (-\infty, 0) \times \{0_{(w_1-2)}\} & \text{if } k = l. \end{cases}$$

Here and in what follows, the symbols $0_{p \times q}$ and 0_p denote the zero matrix in $\mathbb{R}^{p \times q}$ and the zero (column) vector in $\mathbb{R}^p$, respectively, with the convention that these zero-blocks are ignored when p or q vanishes. In combination with (5.13), the above implies in particular that

$$\underbrace{(1, -1, 0_{1 \times (w_1 - 2)})}_{\in \mathbb{R}^{1 \times w_1}} \sigma_1 \left(A_1 \chi_d^k + b_1 \right) = \begin{cases} 0 & \text{if } k \neq l \\ 1 & \text{if } k = l. \end{cases}$$

Let us now first consider the case $L = 1$, i.e., the situation where the neural network (5.9) possesses precisely one hidden layer. Then, the properties of A_1, b_1, and σ_1 and the definitions

$$A_{L+1} = A_2 := \left(y_d^l, 0_{d_y \times (d_y-1)}\right) \begin{pmatrix} 1, -1, 0_{1\times(w_1-2)} \\ 0_{(w_2-1)\times w_1} \end{pmatrix} \in \mathbb{R}^{w_2 \times w_1} \quad \text{and} \quad b_2 := 0_{w_2}$$

yield

$$A_2 \sigma_1 \left(A_1 \chi_d^k + b_1\right) + b_2 = \begin{cases} 0_{d_y} & \text{if } k \neq l \\ y_d^l & \text{if } k = l. \end{cases}$$

The parameter $\bar{\alpha} := (A_2, b_2, A_1, b_1) \in D$ thus satisfies

$$\|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 = \frac{1}{2n} \sum_{k \neq l} \|y_d^k\|_2^2 < \|y_d\|_Y^2 \tag{5.15}$$

which establishes II) as desired. If, on the other hand, L is bigger than one, then by defining A_1 and b_1 as in (5.14) and by setting

$$\begin{aligned} A_2 &:= \begin{pmatrix} 1, -1, 0_{1\times(w_1-2)} \\ 0_{(w_2-1)\times w_1} \end{pmatrix}, \qquad & b_2 &:= \begin{pmatrix} -1/2 \\ \dots \\ -1/2 \end{pmatrix} \in \mathbb{R}^{w_2}, \\ A_i &:= \begin{pmatrix} 1, 0_{1\times(w_{i-1}-1)} \\ 0_{(w_i-1)\times w_{i-1}} \end{pmatrix}, \qquad & b_i &:= \begin{pmatrix} -1/2 \\ \dots \\ -1/2 \end{pmatrix} \in \mathbb{R}^{w_i} \qquad i = 3, ..., L, \\ A_{L+1} &:= \left(y_d^l, 0_{d_y\times(w_L-1)}\right), \qquad & b_{L+1} &:= 0_{d_y}, \end{aligned}$$

we obtain a parameter $\bar{\alpha} := (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$ with

$$\left(\varphi_i^{A_i,b_i} \circ ... \circ \varphi_1^{A_1,b_1}\right)(\chi_d^k) = \begin{cases} 0_{w_i} & \text{if } k \neq l \\ \begin{pmatrix} 1 \\ 0_{w_i-1} \end{pmatrix} & \text{if } k = l \end{cases} \qquad \forall i = 2, ..., L$$

and, analogously to the case $L = 1$,

$$\psi(\bar{\alpha}, \chi_d^k) = \left(\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ ... \circ \varphi_1^{A_1,b_1}\right)(\chi_d^k) = \begin{cases} 0_{d_y} & \text{if } k \neq l \\ y_d^l & \text{if } k = l. \end{cases}$$

Using the same calculation as in (5.15), II) now follows immediately. To finally see that $\Theta(\Psi, x_d) \leq 1 - 1/n$ holds, it suffices to note that every $y_d \in Y$ with $\|y_d\|_Y = 1$ has to possess at least one component with $\|y_d^k\|_2 \geq \sqrt{2}$ and to use the same arguments as in Lemma 5.2. This completes the proof. ■

We would like to mention that the estimate $\Theta(\Psi, x_d) \leq 1 - 1/n$ established above is again very pessimistic. We leave the derivation of better bounds for the number $\Theta(\Psi, x_d)$ in (4.2) for future research. Further, we would like to point out that, in the (essentially) binary case studied in Lemma 5.8, the task of solving a squared-loss problem of the form (5.11) is closely related to classical mixed integer programming. For further details on this topic, we refer to Kurtz and Bah (2020). By combining Lemmas 4.2 and 5.8, we arrive at:

Theorem 5.9 (Approximation Property II) for General Neural Networks) *Consider the situation in Assumption 5.5 and suppose that the index set $\{1, ..., L\}$ can be split into two (possibly empty) disjoint subsets I and J such that the following is true:*

i) For each $i \in I$, the activation function $\sigma_i \colon \mathbb{R} \to \mathbb{R}$ is continuous, the limits

$$\sigma_i(-\infty) := \lim_{s \to -\infty} \sigma_i(s) \qquad \text{and} \qquad \sigma_i(\infty) := \lim_{s \to \infty} \sigma_i(s)$$

exist in $\mathbb{R}$, and it holds $\sigma_i(-\infty) \neq \sigma_i(\infty)$.

ii) For each $i \in J$, the function $\tilde{\sigma}_i(s) := \sigma_i(s) - \sigma_i(s-1)$ satisfies the conditions in i) and it holds $w_i \geq 2$ for all $i \in J$.

iii) The width $w_1 \in \mathbb{N}$ of the lowest hidden layer satisfies $w_1 \geq 2$ in the case $1 \in I$ and $w_1 \geq 4$ in the case $1 \in J$.

Then, the map $\Psi(\cdot, x_d) \colon D \to Y$ associated with x_d and the neural network ψ in (5.9) with the activation functions σ_i, $i = 1, ..., L$, the depth L, and the widths w_i, $i = 1, ..., L$, satisfies condition II) *and the number $\Theta(\Psi, x_d)$ in (4.2) is at most $1 - 1/n$.*

Proof We first consider the case $I = \{1, ..., L\}$ and $J = \emptyset$. In this situation, it follows from the properties of the activation functions σ_i, $i = 1, ..., L$, that

$$\frac{\sigma_i(\gamma s) - \sigma_i(-\infty)}{\sigma_i(\infty) - \sigma_i(-\infty)} \to \bar{\sigma}_i(s) := \begin{cases} 0 & \text{if } s < 0 \\ \dfrac{\sigma_i(0) - \sigma_i(-\infty)}{\sigma_i(\infty) - \sigma_i(-\infty)} & \text{if } s = 0 \\ 1 & \text{if } s > 0 \end{cases} \tag{5.16}$$

holds for all $s \in \mathbb{R}$ and all $i = 1, ..., L$ for $0 < \gamma \to \infty$ and, as a consequence, that

$$\frac{1}{\sigma_i(\infty) - \sigma_i(-\infty)} \left[\varphi_i^{\gamma A_i, \gamma b_i}(z) - \sigma_i(-\infty) 1_{w_i} \right] \to \bar{\sigma}_i(A_i z + b_i) =: \bar{\varphi}_i^{A_i, b_i}(z)$$

holds for all $z \in \mathbb{R}^{w_{i-1}}$ and all $i = 1, ..., L$ for $0 < \gamma \to \infty$, where 1_{w_i} denotes the (column) vector in $\mathbb{R}^{w_i}$ that contains the entry one in each component. Due to the definitions of the map $\Psi(\cdot, x_d) \colon D \to Y$ and D, we further know that, for all $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$ and all $\gamma > 0$, we have

$$\begin{aligned} &\left\{ \psi \left(A_{L+1}, b_{L+1}, ..., \frac{A_2}{\sigma_1(\infty) - \sigma_1(-\infty)}, b_2 - \frac{\sigma_1(-\infty) A_2 1_{w_1}}{\sigma_1(\infty) - \sigma_1(-\infty)}, \gamma A_1, \gamma b_1, \chi_d^k \right) \right\}_{k=1}^n \\ &= \Bigg\{ \left(\varphi_{L+1}^{A_{L+1}, b_{L+1}} \circ ... \circ \varphi_3^{A_3, b_3} \circ \sigma_2 \right) \\ &\qquad \left(A_2 \left(\frac{1}{\sigma_1(\infty) - \sigma_1(-\infty)} \left[\varphi_1^{\gamma A_1, \gamma b_1}(\chi_d^k) - \sigma_1(-\infty) 1_{w_1} \right] \right) + b_2 \right) \Bigg\}_{k=1}^n \in \mathrm{cl}_Y \left(\Psi(D, x_d) \right). \end{aligned}$$

(Here, in the borderline case $L = 1$, the "empty" composition $\varphi_{L+1}^{A_{L+1}, b_{L+1}} \circ ... \circ \varphi_3^{A_3, b_3} \circ \sigma_2$ has to be interpreted as the identity map.) Combining the last two results, exploiting that the functions σ_i are continuous, and using that the set $\mathrm{cl}_Y \left(\Psi(D, x_d) \right)$ is closed yields that

$$\left\{ \left(\varphi_{L+1}^{A_{L+1}, b_{L+1}} \circ ... \circ \varphi_2^{A_2, b_2} \circ \bar{\varphi}_1^{A_1, b_1} \right) (\chi_d^k) \right\}_{k=1}^n \in \mathrm{cl}_Y \left(\Psi(D, x_d) \right)$$

holds for all $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$. Note that, here and in what follows, with $\Psi(\cdot, x_d)\colon D \to Y$ we still mean the function (2.2) associated with the original neural network ψ involving the activation functions σ_i, $i = 1, ..., L$. By proceeding along exactly the same lines as above for the layers $i = 2, ..., L$ (in that order), we obtain that, for all $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$, we have

$$\left\{ \left(\varphi_{L+1}^{A_{L+1}, b_{L+1}} \circ \bar{\varphi}_L^{A_L, b_L} \circ ... \circ \bar{\varphi}_1^{A_1, b_1} \right) (x_d^k) \right\}_{k=1}^n \in \mathrm{cl}_Y \left(\Psi(D, x_d) \right).$$

As the map $\varphi_{L+1}^{A_{L+1}, b_{L+1}}$ does not depend on any activation function, this shows that the set $\mathrm{cl}_Y \left(\Psi(D, x_d) \right)$ associated with the neural network ψ involving the activation functions σ_i, $i = 1, ..., L$, is at least as big as the set $\mathrm{cl}_Y \left(\bar{\Psi}(D, x_d) \right)$ that is associated with the neural network $\bar{\psi}$ that has the same depth L and widths w_i as ψ and involves the Heaviside-type activation functions on the right-hand side of (5.16). Since the latter set satisfies (4.1) by Lemmas 4.2 and 5.8 and due to our assumptions on w_1, it now follows immediately that

$$\min_{y \in \mathrm{cl}_Y(\Psi(D, x_d))} \|y - y_d\|_Y^2 \leq \min_{y \in \mathrm{cl}_Y\left(\bar{\Psi}(D, x_d)\right)} \|y - y_d\|_Y^2 < \|y_d\|_Y^2 \qquad \forall y_d \in Y \setminus \{0\},$$

and, due to (4.2) and again Lemma 5.8, that $\Theta(\Psi, x_d) \leq \Theta(\bar{\Psi}, x_d) \leq 1 - 1/n$. This completes the proof for $I = \{1, ..., L\}$ and $J = \emptyset$ (see Lemma 4.2).

Let us now suppose that J is not empty, and let us first assume that an $i \in J$ with $i \geq 2$ and $w_i = 2\tilde{w}_i$, $\tilde{w}_i \in \mathbb{N}$, is given. Then, for all parameters of the form

$$A_{i+1} := \tilde{A}_{i+1} \left(\mathrm{Id}_{\tilde{w}_i \times \tilde{w}_i}, -\mathrm{Id}_{\tilde{w}_i \times \tilde{w}_i} \right), \qquad A_i := \begin{pmatrix} \tilde{A}_i \\ \tilde{A}_i \end{pmatrix}, \qquad b_i := \begin{pmatrix} \tilde{b}_i \\ \tilde{b}_i - 1_{\tilde{w}_i} \end{pmatrix}, \tag{5.17}$$

with $\tilde{A}_{i+1} \in \mathbb{R}^{w_{i+1} \times \tilde{w}_i}$, $\tilde{A}_i \in \mathbb{R}^{\tilde{w}_i \times w_{i-1}}$, and $\tilde{b}_i \in \mathbb{R}^{\tilde{w}_i}$, we have

$$A_{i+1}\sigma_i(A_i z + b_i) = \tilde{A}_{i+1}\tilde{\sigma}_i(\tilde{A}_i z + \tilde{b}_i) \qquad \forall z \in \mathbb{R}^{w_{i-1}}.$$

Here, $\mathrm{Id}_{\tilde{w}_i \times \tilde{w}_i} \in \mathbb{R}^{\tilde{w}_i \times \tilde{w}_i}$ is the identity matrix, $1_{\tilde{w}_i}$ again denotes the vector in $\mathbb{R}^{\tilde{w}_i}$ that contains the entry one in every component, and $\tilde{\sigma}_i$ is defined as in point ii) of the theorem, i.e., $\tilde{\sigma}_i(s) := \sigma_i(s) - \sigma_i(s-1)$ for all $s \in \mathbb{R}$. In combination with (5.9) and our assumptions on the activation functions σ_i, $i \in J$, the above shows that every layer of ψ, that is associated with an index $2 \leq i \in J$ and possesses an even width, can emulate a neural network layer of a smaller width that involves an activation function of the type studied in point i) of the theorem. Note that, in the case $i \in J$ with $i \geq 2$ and $w_i = 2\tilde{w}_i + 1$, $\tilde{w}_i \in \mathbb{N}$, and in the case $i = 1 \in J$, we can proceed completely analogously to the above by adding suitable rows/columns of zeros in (5.17), and that, in the case $i = 1 \in J$, we can always achieve that $\tilde{w}_1 \geq 2$ holds by our assumptions on w_1. In summary, we may thus conclude that, for arbitrary I and J, we can always find a neural network of the type (5.9), $\tilde{\psi}$ lets say, which satisfies the assumptions of the first part of this proof and $\mathrm{cl}_Y(\tilde{\Psi}(D, x_d)) \subset \mathrm{cl}_Y(\Psi(D, x_d))$, where $\tilde{\Psi}$ and Ψ are the functions in (2.2) associated with $\tilde{\psi}$ and the original network ψ, respectively. The claim of the theorem now follows immediately from Lemma 4.2 and the definition of the number $\Theta(\Psi, x_d)$ (cf. the first part of this proof). ■

Note that the above proof shows that a network ψ with the properties in Theorem 5.9 is always at least as expressive as a network $\bar{\psi}$ that involves the Heaviside-type activation functions on the right-hand side of (5.16) and possesses the widths w_i for all $i \in I$ and $\lfloor w_i/2 \rfloor$ for all $i \in J$. (Here, with "at least as expressive", we mean that the inclusion $\mathrm{cl}_Y(\bar{\Psi}(D, x_d)) \subset \mathrm{cl}_Y(\Psi(D, x_d))$ holds for all data vectors x_d that satisfy the conditions in Assumption 5.5.) As an immediate consequence of Theorem 5.9, we obtain (cf. (Calin, 2020, Chapter 2)):

Corollary 5.10 (Approximation Property II) for Popular Activation Functions) *Consider the situation in Assumption 5.5 and suppose that the set $\{1, ..., L\}$ can be split into two (possibly empty) disjoint subsets I and J such that the following is true:*

1. *For each $i \in I$, the activation function σ_i is of one of the following types:*

   i) $\sigma_{\mathrm{sig}}(s) := 1/(1 + \mathrm{e}^{-s})$ *(sigmoid/logistic/soft-step activation),*

   ii) $\sigma_{\tanh}(s) := \tanh(s)$ *(tanh-activation),*

   iii) $\sigma_{\arctan}(s) := \arctan(s)$ *(arctan-activation),*

   iv) $\sigma_{\mathrm{es}}(s) := s/(1 + |s|)$ *(soft-sign/Elliot-sign activation),*

   v) $\sigma_{\mathrm{isru}}(s) := s/(1 + cs^2)^{1/2}$ *with some* $c > 0$ *(inverse square root unit),*

   vi) $\sigma_{\mathrm{sc}}(s) := c^{-1} \log\left((1 + \mathrm{e}^{cs})/(1 + \mathrm{e}^{c(s-1)})\right)$ *with some* $c > 0$ *(soft-clip activation),*

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

2. *It holds $w_i \geq 2$ for all $i \in J$, and, for every $i \in J$, σ_i is of one of the following types:*

   i) $\sigma_{\mathrm{relu}}(s) := \max(0, s)$ *(rectified linear unit),*

   ii) $\sigma_{\mathrm{prelu}}(s) := \max(0, s) + \min(0, cs)$, $|c| \neq 1$ *(leaky/parametric ReLU),*

   iii) $\sigma_{\mathrm{soft}+}(s) := \ln(1 + \mathrm{e}^s)$ *(soft-plus activation),*

   iv) $\sigma_{\mathrm{bentid}}(s) := \frac{1}{2}(s^2 + 1)^{1/2} - \frac{1}{2} + s$ *(bent-identity activation),*

   v) $\sigma_{\mathrm{silu}}(s) := s/(1 + \mathrm{e}^{-s})$ *(sigmoid linear unit a.k.a. swish-1),*

   vi)
   $$\sigma_{\mathrm{isrlu}}(s) := \begin{cases} s/(1 + cs^2)^{1/2} & \text{if } s < 0 \\ s & \text{if } s \geq 0 \end{cases}, \quad c > 0 \qquad \text{(ISRL-unit),}$$

   vii)
   $$\sigma_{\mathrm{elu}}(s) := \begin{cases} c\,(\mathrm{e}^s - 1) & \text{if } s < 0 \\ s & \text{if } s \geq 0 \end{cases}, \quad c \in \mathbb{R} \qquad \text{(exponential linear unit).}$$

3. *It holds $w_1 \geq 2$ in the case $1 \in I$ and $w_1 \geq 4$ in the case $1 \in J$.*

Then, the function $\Psi(\cdot, x_d) \colon D \to Y$ associated with x_d and the neural network ψ in (5.9) *possesses the property* II) *and the number $\Theta(\Psi, x_d)$ in* (4.2) *is at most $1 - 1/n$.*

Proof This follows immediately by checking the assumptions of Theorem 5.9. ■

We remark that the list of activation functions in Corollary 5.10 is far from exhaustive and that, even when the assumptions of Theorem 5.9 are not satisfied, it is still often possible to establish II) by hand. Compare, e.g., with the calculation in Section 3 in this context, where we have done precisely that. Having checked that the properties I) and II) hold under reasonable assumptions on the activation functions and widths in (5.9), we can now again apply the abstract results of Section 4. Before we collect the numerous corollaries that we obtain in this way, we prove two lemmas that simplify the application of Theorem 4.16.

Lemma 5.11 *Consider the situation in Assumption 5.5. Suppose further that the inequality* $\min(d_\chi, d_y) \leq \min(w_1, ..., w_L)$ *holds and that, for each* $i \in \{1, ..., L\}$*, there exists an open nonempty interval* $I_i \subset \mathbb{R}$ *such that* σ_i *is affine with a non-vanishing derivative on* I_i*. Define*

$$V := \left\{ \{z_k\}_{k=1}^n \in Y \,\middle|\, \exists A \in \mathbb{R}^{d_y \times d_\chi}, b \in \mathbb{R}^{d_y} \text{ such that } z_k = A\chi_d^k + b \quad \forall k = 1, ..., n \right\}.$$

Then, for every element z *of the subspace* V*, there exist an* $\bar{\alpha} \in D$ *and an open set* $U \subset D$ *satisfying* $\bar{\alpha} \in U$*,* $z = \Psi(\bar{\alpha}, x_d)$*, and* $\Psi(U, x_d) \subset V$*.*

Proof Suppose that an arbitrary but fixed element z of the subspace V with associated $A \in \mathbb{R}^{d_y \times d_\chi}$ and $b \in \mathbb{R}^{d_y}$ is given and let $a_i \in \mathbb{R}$, $\varepsilon_i > 0$, $\beta_i \in \mathbb{R} \setminus \{0\}$, and $\gamma_i \in \mathbb{R}$ satisfy $I_i = (a_i - \varepsilon_i, a_i + \varepsilon_i)$ and $\sigma_i(s) = \beta_i s + \gamma_i$ for all $s \in I_i$ and all $i = 1, ..., L$. Let us further again use the symbols $0_{p\times q}$ and 0_p to denote the zero matrix in $\mathbb{R}^{p\times q}$ and the zero (column) vector in $\mathbb{R}^p$, respectively, with the convention that these blocks are ignored in the cases $p = 0$ and $q = 0$, let $\mathrm{Id}_{p\times p}$ and 1_p denote the identity matrix in $\mathbb{R}^{p\times p}$ and the vector in $\mathbb{R}^p$ whose entries are all equal to one, respectively, and let $\|\cdot\|_\infty$ be the maximum norm on the Euclidean space. Then, in the case $d_\chi \leq \min(w_1, ..., w_L)$, it is easy to check that the matrices and vectors

$$\begin{aligned}
A_1 &:= \frac{\varepsilon_1}{2\max_{k=1,...,n} \|\chi_d^k\|_\infty} \begin{pmatrix} \mathrm{Id}_{d_\chi \times d_\chi} \\ 0_{(w_1 - d_\chi)\times d_\chi} \end{pmatrix} \in \mathbb{R}^{w_1 \times w_0}, \qquad b_1 := a_1 1_{w_1} \in \mathbb{R}^{w_1}, \\
A_i &:= \frac{\varepsilon_i}{\beta_{i-1}\varepsilon_{i-1}} \begin{pmatrix} \mathrm{Id}_{d_\chi \times d_\chi}, 0_{d_\chi \times (w_{i-1} - d_\chi)} \\ 0_{(w_i - d_\chi)\times w_{i-1}} \end{pmatrix} \in \mathbb{R}^{w_i \times w_{i-1}}, \qquad i = 2, ..., L, \\
b_i &:= a_i 1_{w_i} - (a_{i-1}\beta_{i-1} + \gamma_{i-1}) A_i 1_{w_{i-1}} \in \mathbb{R}^{w_i}, \qquad i = 2, ..., L, \\
A_{L+1} &:= \frac{2\max_{k=1,...,n} \|\chi_d^k\|_\infty}{\beta_L \varepsilon_L} \left(A, 0_{d_y \times (w_L - d_\chi)}\right) \in \mathbb{R}^{w_{L+1} \times w_L}, \\
b_{L+1} &:= b - (a_L\beta_L + \gamma_L) A_{L+1} 1_{w_L} \in \mathbb{R}^{w_{L+1}}
\end{aligned}$$

satisfy

$$\begin{aligned}
A_1\chi_d^j + b_1 &= \frac{\varepsilon_1}{2\max_{k=1,...,n}\|\chi_d^k\|_\infty} \begin{pmatrix} \chi_d^j \\ 0_{w_1 - d_\chi} \end{pmatrix} + a_1 1_{w_1} \in (a_1 - \varepsilon_1, a_1 + \varepsilon_1)^{w_1}, \\
A_i\left(\varphi_{i-1}^{A_{i-1}, b_{i-1}} \circ ... \circ \varphi_1^{A_1, b_1}(\chi_d^j)\right) + b_i &= \frac{\varepsilon_i}{2\max_{k=1,...,n}\|\chi_d^k\|_\infty} \begin{pmatrix} \chi_d^j \\ 0_{w_i - d_\chi} \end{pmatrix} + a_i 1_{w_i} \\
&\qquad \in (a_i - \varepsilon_i, a_i + \varepsilon_i)^{w_i} \quad \forall i = 2, ..., L,
\end{aligned} \tag{5.18}$$

and

$$\left(\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ \dots \circ \varphi_1^{A_1,b_1}\right)(\chi_d^j) = A\chi_d^j + b$$

for all $j = 1, ..., n$. The parameter $\bar{\alpha} := (A_{L+1}, b_{L+1}, ..., A_1, b_1)$ thus satisfies $\Psi(\bar{\alpha}, x_d) = z$ as desired. Since the inclusions in (5.18) are stable w.r.t. small perturbations in the matrices A_i and the vectors b_i, we further obtain that the function $\psi(\alpha, \cdot)\colon \mathcal{X} \to \mathcal{Y}$ also behaves affine-linearly on the training data χ_d^k, $k = 1, ..., n$, for all $\alpha \in D$ in a small open neighborhood U of $\bar{\alpha}$. This shows that there exists an open set $U \subset D$ satisfying $\bar{\alpha} \in U$, $\Psi(\bar{\alpha}, x_d) = z$, and $\Psi(U, x_d) \subset V$ and proves the claim of the lemma in the case $d_\chi \le \min(w_1, ..., w_L)$.

In the situation $d_y \le \min(w_1, ..., w_L)$, we can proceed along similar lines as above. Given an arbitrary but fixed $z \in V$ with associated $A \in \mathbb{R}^{d_y \times d_x}$ and $b \in \mathbb{R}^{d_y}$, we define

$$\begin{aligned}
A_1 &:= \frac{\varepsilon_1}{2\max_{k=1,\dots,n} \|A\chi_d^k\|_\infty + 1} \begin{pmatrix} A \\ 0_{(w_1 - d_y)\times d_\chi} \end{pmatrix} \in \mathbb{R}^{w_1 \times w_0}, \\
b_1 &:= a_1 1_{w_1} \in \mathbb{R}^{w_1}, \\
A_i &:= \frac{\varepsilon_i}{\beta_{i-1}\varepsilon_{i-1}} \begin{pmatrix} \mathrm{Id}_{d_y\times d_y}, 0_{d_y \times (w_{i-1}-d_y)} \\ 0_{(w_i - d_y)\times w_{i-1}} \end{pmatrix} \in \mathbb{R}^{w_i \times w_{i-1}}, \qquad i = 2, ..., L, \\
b_i &:= a_i 1_{w_i} - (a_{i-1}\beta_{i-1} + \gamma_{i-1}) A_i 1_{w_{i-1}} \in \mathbb{R}^{w_i}, \qquad i = 2, ..., L, \\
A_{L+1} &:= \frac{2\max_{k=1,\dots,n} \|A\chi_d^k\|_\infty + 1}{\beta_L \varepsilon_L} \left(\mathrm{Id}_{d_y\times d_y}, 0_{d_y\times(w_L - d_y)}\right) \in \mathbb{R}^{w_{L+1}\times w_L}, \\
b_{L+1} &:= b - (a_L\beta_L + \gamma_L) A_{L+1} 1_{w_L} \in \mathbb{R}^{w_{L+1}}.
\end{aligned}$$

Then, it is easy to check that it holds

$$\begin{aligned}
A_1\chi_d^j + b_1 &= \frac{\varepsilon_1}{2\max_{k=1,\dots,n} \|A\chi_d^k\|_\infty + 1} \begin{pmatrix} A\chi_d^j \\ 0_{w_1 - d_y} \end{pmatrix} + a_1 1_{w_1} \in (a_1 - \varepsilon_1, a_1 + \varepsilon_1)^{w_1}, \\
A_i\left(\varphi_{i-1}^{A_{i-1},b_{i-1}} \circ \dots \circ \varphi_1^{A_1,b_1}(\chi_d^j)\right) + b_i &= \frac{\varepsilon_i}{2\max_{k=1,\dots,n} \|A\chi_d^k\|_\infty + 1} \begin{pmatrix} A\chi_d^j \\ 0_{w_i - d_y} \end{pmatrix} + a_i 1_{w_i} \\
&\in (a_i - \ \varepsilon_i, a_i + \varepsilon_i)^{w_i} \quad \forall i = 2, ..., L,
\end{aligned}$$

and

$$\left(\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ \dots \circ \varphi_1^{A_1,b_1}\right)(\chi_d^j) = A\chi_d^j + b$$

for all $j = 1, ..., n$. The existence of a parameter $\bar{\alpha} \in D$ and an open set $U \subset D$ with the desired properties now follows completely analogously to the first part of the proof. ■

Lemma 5.12 *Consider the situation in Assumption 5.5, assume that the functions σ_i, $i = 1, ..., L$, are bounded on bounded sets, and suppose that there exists an index $j \in \{1, ..., L\}$ such that the function σ_j is constant on an open nonempty interval $I_j \subset \mathbb{R}$. Define*

$$V := \left\{ \{z_k\}_{k=1}^n \in Y \,\middle|\, z_k = z_l \quad \forall k, l \in \{1, ..., n\} \right\}.$$

Then, for every element z of the subspace V, there exist an $\bar{\alpha} \in D$ and an open set $U \subset D$ satisfying $\bar{\alpha} \in U$, $z = \Psi(\bar{\alpha}, x_d)$, and $\Psi(U, x_d) \subset V$.

Proof Consider an arbitrary but fixed $z = \{z_k\}_{k=1}^n \in V$, let $b \in \mathbb{R}^{d_y}$ denote the unique vector with $z_k = b$ for all $k = 1, ..., n$, and let $a_j \in \mathbb{R}$ and $\varepsilon_j > 0$ satisfy $I_j = (a_j - \varepsilon_j, a_j + \varepsilon_j)$. Then, by defining

$$A_i := 0 \in \mathbb{R}^{w_i \times w_{i-1}} \quad \forall i \in \{1, ..., L+1\}, \qquad b_i := 0 \in \mathbb{R}^{w_i} \quad \forall i \in \{1, ..., L\} \setminus \{j\},$$
$$b_j := a_j 1_{w_j} \in \mathbb{R}^{w_j}, \quad \text{and} \quad b_{L+1} := b \in \mathbb{R}^{w_{L+1}},$$

where 1_{w_j} again denotes the vector which contains the number one in each of its entries, we obtain a parameter $\bar{\alpha} := (A_{L+1}, b_{L+1}, ..., A_1, b_1)$ which trivially satisfies $\psi(\bar{\alpha}, \chi_d^k) = b = z_k$ for all $k = 1, ..., n$ and

$$A_j \left(\varphi_{j-1}^{A_{j-1}, b_{j-1}} \circ ... \circ \varphi_1^{A_1, b_1}(\chi_d^k) \right) + b_j = b_j \in (a_j - \ \varepsilon_j, a_j + \varepsilon_j)^{w_j} \quad \forall k = 1, ..., n. \tag{5.19}$$

(Here, the expression $\varphi_{j-1}^{A_{j-1}, b_{j-1}} \circ ... \circ \varphi_1^{A_1, b_1}$ again has to be interpreted as the identity map in the borderline case $j = 1$.) As the inclusion in (5.19) remains true for all parameters α in a small neighborhood $U \subset D$ of $\bar{\alpha}$ and since the activation function σ_j is constant on I_j, it now follows immediately that there exist an $\bar{\alpha} \in D$ and an open set $U \subset D$ with the desired properties $\bar{\alpha} \in U$, $z = \Psi(\bar{\alpha}, x_d)$, and $\Psi(U, x_d) \subset V$. This completes the proof. ■

Note that the condition $\min(d_\chi, d_y) \leq \min(w_1, ..., w_L)$ in Lemma 5.11 is vacuous in the case $d_y = 1$, i.e., in the situation where the neural network has a scalar output. With Lemmas 5.11 and 5.12 in place, we are in the position to study the consequences that the analysis of Section 4 has for the optimization landscape and the stability properties of training problems of the form (5.11). Following the structure of Section 4, we begin with a result on the set-valuedness and the stability of the best approximation map $P_\Psi^{x_d}$ associated with (5.11) in the case where there exist unrealizable vectors:

Corollary 5.13 (Nonuniqueness and Instability of Best Approximations) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i are such that Lemma 5.8, Theorem 5.9, or Corollary 5.10 can be applied to ψ. Assume further that there exist unrealizable vectors $y_d \in Y$, i.e., that $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ holds. Then, the best approximation map*

$$P_\Psi^{x_d} \colon Y \rightrightarrows Y, \qquad y_d \mapsto \underset{y \in \mathrm{cl}_Y(\Psi(D, x_d))}{\arg\min} \|y - y_d\|_Y^2, \tag{5.20}$$

associated with the training problem (5.11) *is set-valued and there exist uncountably many training label vectors $y_d \in Y$ such that $|P_\Psi^{x_d}(y_d)| > 1$ holds. Moreover, the map $P_\Psi^{x_d} \colon Y \rightrightarrows Y$ is discontinuous in the sense that, for every arbitrary but fixed $C > 0$, there exists an uncountable set $\mathcal{M}_C \subset Y$ such that, for every $y_d \in \mathcal{M}_C$, there are sequences $\{y_d^l\}, \{\tilde{y}_d^l\} \subset Y$ satisfying*

$$y_d^l \to y_d \text{ for } l \to \infty, \qquad \tilde{y}_d^l \to y_d \text{ for } l \to \infty,$$
$$|P_\Psi^{x_d}(y_d^l)| = |P_\Psi^{x_d}(\tilde{y}_d^l)| = 1 \quad \forall l, \quad \text{and} \quad \|P_\Psi^{x_d}(y_d^l) - P_\Psi^{x_d}(\tilde{y}_d^l)\|_Y \geq C \quad \forall l.$$

Further, for every $C > 0$, there exists at least one $y_d \in \mathcal{M}_C$ satisfying (4.14) *(with the number $\Theta(\Psi, x_d) \in (0, 1)$ associated with ψ and x_d defined in* (4.2)*).*

Proof As Lemma 5.8, Theorem 5.9, or Corollary 5.10 can be applied to ψ by assumption and due to Lemma 5.6, we know that the map $\Psi(\cdot, x_d)\colon D \to Y$ associated with x_d and ψ satisfies I) and II). By invoking Theorem 4.9, the claim now follows immediately. ■

Note that, in the situation of Corollary 5.13, the comments made in Remark 4.10 (e.g., on the potential consequences for the convergence behavior of descent methods) apply to the training problem (5.11) as well. As a corollary of Theorem 4.11, we next obtain:

Corollary 5.14 (Excessive Nonuniqueness or Spurious Local Minima) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i satisfy the conditions in Theorem 5.9 (or Corollary 5.10, respectively). Assume further that the functions σ_i are continuous and that there exist unrealizable label vectors $y_d \in Y$, i.e., that $\mathrm{cl}_Y(\Psi(D, x_d)) \neq Y$ holds. Then, there exist uncountably many $y_d \in Y$ such that the best approximation map $P_\Psi^{x_d}\colon Y \rightrightarrows Y$ defined in (5.20) satisfies $|P_\Psi^{x_d}(y_d)| = \infty$ or there exist an open nonempty cone $K \subset Y$ and a number $M \in \mathbb{N}$ with $M \geq 2$ such that, for every $y_d \in K$, there are nonempty disjoint closed subsets $D_1, ..., D_M$ of the parameter space D in (5.7) satisfying*

$$\inf_{\alpha \in D_1} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \inf_{\alpha \in D_i} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \quad \forall i = 2, ..., M \tag{5.21}$$

and

$$\sup_{\alpha \in D_1 \cup ... \cup D_M} \|\Psi(\alpha, x_d) - y_d\|_Y^2 < \|\Psi(\tilde{\alpha}, x_d) - y_d\|_Y^2 \quad \forall \tilde{\alpha} \in D \setminus (D_1 \cup ... \cup D_M). \tag{5.22}$$

Further, if the second of the above cases applies, then the spurious local minima/valleys in (5.21) can be arbitrarily bad in the sense that, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in K$ which not only satisfy (5.21) but even

$$\inf_{\alpha \in D_1} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + C < \inf_{\alpha \in D_i} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \quad \forall i = 2, ..., M.$$

Proof The first half of the corollary follows immediately from Lemma 5.6, Theorem 5.9, Corollary 5.10, and Theorem 4.11. To see that the spurious minima/valleys in (5.21) can get arbitrarily bad, it suffices to note that, for every $\gamma > 0$ and every $y_d \in K$ that satisfies (5.21) and (5.22) for some $M \in \mathbb{N}$ and $D_1, ..., D_M$, the vector $\widehat{y}_d := \gamma y_d$ and the (trivially nonempty, disjoint, and closed) sets

$$\widehat{D}_i := \left\{ \widehat{\alpha} = (\gamma A_{L+1}, \gamma b_{L+1}, A_L, b_L, ..., A_1, b_1) \,\middle|\, (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D_i \right\}, \; i = 1, ..., M,$$

satisfy

$$\begin{aligned} &\inf_{\widehat{\alpha} \in \widehat{D}_1} \|\Psi(\widehat{\alpha}, x_d) - \widehat{y}_d\|_Y^2 - \inf_{\widehat{\alpha} \in \widehat{D}_i} \|\Psi(\widehat{\alpha}, x_d) - \widehat{y}_d\|_Y^2 \\ &\qquad = \gamma^2 \left(\inf_{\alpha \in D_1} \|\Psi(\alpha, x_d) - y_d\|_Y^2 - \inf_{\alpha \in D_i} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \right) \end{aligned}$$

for all $i = 1, ..., M$ and

$$\sup_{\widehat{\alpha} \in \widehat{D}_1 \cup ... \cup \widehat{D}_M} \|\Psi(\widehat{\alpha}, x_d) - \widehat{y}_d\|_Y^2 < \|\Psi(\tilde{\alpha}, x_d) - \widehat{y}_d\|_Y^2 \quad \forall \tilde{\alpha} \in D \setminus (\widehat{D}_1 \cup ... \cup \widehat{D}_M)$$

by the architecture of ψ. This completes the proof. ■

As already mentioned in Section 4, the last result complements the findings of Venturi et al. (2019) on the existence of spurious valleys in training problems for one-hidden-layer neural networks with non-polynomial non-negative activation functions in the sense that it shows that the existence of such valleys also is to be expected in the multi-layer case provided that there are unrealizable label vectors, cf. the comments after Theorem 4.11. Let us now turn our attention to non-optimal stationary points:

Corollary 5.15 (Non-Optimal Stationary Points in the Case $\mathbf{m} < \mathbf{nd}_y$) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i are such that Lemma 5.8, Theorem 5.9, or Corollary 5.10 can be applied to ψ. Assume further that the number of parameters $m := w_{L+1}(w_L + 1) + ... + w_1(w_0 + 1)$ in the neural network ψ is smaller than the product nd_y. Then, for every point $\bar\alpha \in D$ at which the map $\Psi(\cdot, x_d)\colon D \to Y$ is differentiable and every arbitrary but fixed $\varepsilon > 0$, there exist uncountably many label vectors $y_d \in Y$ satisfying*

$$\left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|\Psi(\bar\alpha, x_d)\|_Y < \|\Psi(\bar\alpha, x_d) - y_d\|_Y < \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)}\right)^{1/2} \|\Psi(\bar\alpha, x_d)\|_Y + \varepsilon \tag{5.23}$$

such that $\bar\alpha$ is a spurious local minimum or a saddle point of the training problem (5.11)*. Here, $\Theta(\Psi, x_d) \in [0, 1)$ again denotes the number in* (4.2) *associated with ψ and x_d that measures the extent to which* II) *is satisfied and the worst-case approximation error in* (4.3)*. Further, for every $C > 0$, there exist uncountably many $y_d \in Y$ such that $\bar\alpha$ is a spurious local minimum or a saddle point of* (5.11) *and such that* (4.25) *and* (4.26) *hold. As a saddle point or spurious local minimum, $\bar\alpha$ can thus be arbitrarily bad.*

Proof The assertion follows straightforwardly from Lemma 5.6, Lemma 5.8, Theorem 5.9, Corollary 5.10 and Theorem 4.13 and by noting that the affine-linearity of the topmost layer of ψ implies that $\Psi(\bar\alpha, x_d)$ is an element of the linear hull $\operatorname{span}(\partial_1\Psi(\bar\alpha, x_d), ..., \partial_m\Psi(\bar\alpha, x_d))$ for all points of differentiability $\bar\alpha$ of the map $\Psi(\cdot, x_d)\colon D \to Y$. ■

Note that the last three results provide strong arguments for the overparameterization of training problems of the form (5.11) (or, more precisely, for choosing problem setups that ensure realizability). However, as we have already seen in Section 4, overparameterization cannot resolve all of the difficulties that arise when training problems of the type (5.11) are considered. This is also illustrated by the following corollary on the existence of non-optimal stationary points that also covers the case $nd_y \leq m$.

Corollary 5.16 (Non-Optimal Stationary Points in the Case $\mathbf{d}_\chi + \mathbf{1} < \mathbf{n}$) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i satisfy the conditions in Theorem 5.9 (or Corollary 5.10, respectively). Assume further that $d_\chi + 1 < n$ holds and that the activation functions σ_i are differentiable. Then, for every $\bar\alpha \in D$ of the form $\bar\alpha = (A_{L+1}, b_{L+1}, ..., A_2, b_2, 0, b_1)$ (and thus for all elements of an $(m - d_\chi w_1)$-dimensional subspace of D) and every arbitrary but fixed $\varepsilon > 0$, there exist uncountably many label vectors $y_d \in Y$ satisfying* (5.23) *such that $\bar\alpha$ is a spurious local minimum or a saddle point of* (5.11)*. Further, for every $\bar\alpha$ of the above type and every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in Y$ such that $\bar\alpha$ is a spurious local minimum or a saddle point of* (5.11) *and such that* (4.25) *and* (4.26) *hold.*

Proof This follows from Lemma 5.6, Theorem 5.9, Corollary 5.10, and Corollary 4.20. Note that Assumption 4.18 is trivially satisfied in the situation of Assumption 5.5 (up to the isomorphism in (5.10)) so that Corollary 4.20 is indeed applicable here. ∎

Next, we consider neural networks with activation functions that are affine-linear on some open nonempty subset of their domain of definition. We begin with two results on the existence of spurious local minima:

Corollary 5.17 (Spurious Minima for Activations with an Affine Segment) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i are such that Theorem 5.9 (or Corollary 5.10) can be applied to ψ. Assume further that, for every $i \in \{1, ..., L\}$, there exists an open nonempty interval $I_i \subset \mathbb{R}$ such that σ_i is affine-linear with a non-vanishing derivative on I_i and that the inequalities $\min(d_\chi, d_y) \leq \min(w_1, ..., w_L)$ and $n > d_\chi + 1$ hold. Define*

$$V := \left\{ \{z_k\}_{k=1}^n \in Y \,\middle|\, \exists A \in \mathbb{R}^{d_y \times d_x}, b \in \mathbb{R}^{d_y} \text{ such that } z_k = A\chi_d^k + b \quad \forall k = 1, ..., n \right\}$$

and denote the $(\cdot,\cdot)_Y$-orthogonal complement of the space V in Y with $V^\perp$. Then, the training problem (5.11) possesses at least one spurious local minimum satisfying a growth condition of the form (4.28) for all label vectors $y_d \in Y$ that are elements of the open cone

$$K := \left\{ y_d^1 + y_d^2 \in Y \,\middle|\, y_d^1 \in V,\, y_d^2 \in V^\perp,\, \|y_d^2\|_Y > \left(\frac{\Theta(\Psi, x_d)}{1 - \Theta(\Psi, x_d)} \right)^{1/2} \|y_d^1\|_Y \right\}. \tag{5.24}$$

Here, $\Theta(\Psi, x_d) \in [0, 1)$ again denotes the number in (4.2) associated with ψ and x_d that measures the extent to which II) *is satisfied and the worst-case approximation error in* (4.3)*. Further, for every arbitrary but fixed $C > 0$, there exist uncountably many $y_d \in K$ such that at least one of the spurious local minima of* (5.11) *satisfies* (4.25)*,* (4.26)*, and* (4.28)*, and if $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$ holds, then the cone K in* (5.24) *is equal to $Y \setminus V$ and* (5.11) *possesses spurious local minima for all y_d that are not affine-linearly fittable.*

Proof To establish the assertions of the corollary, it suffices to combine Lemma 5.6, Theorem 5.9, and Corollary 5.10 with Lemma 5.11 and Theorem 4.16. ∎

Corollary 5.18 (Spurious Minima for Activations with a Constant Segment) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i are such that Lemma 5.8, Theorem 5.9, or Corollary 5.10 can be applied to ψ. Assume that the functions σ_i, $i = 1, ..., L$, are bounded on bounded sets and that there exists a $j \in \{1, ..., L\}$ such that σ_j is constant on an open nonempty interval $I_j \subset \mathbb{R}$. Define*

$$V := \left\{ \{z_k\}_{k=1}^n \in Y \,\middle|\, z_k = z_l \quad \forall k, l \in \{1, ..., n\} \right\}$$

and let K be defined as in (5.24) *(with the above V). Then,* (5.11) *possesses at least one spurious local minimum satisfying a growth condition of the form* (4.28) *for all $y_d \in K$ and, for every $C > 0$, there exist uncountably many $y_d \in K$ such that at least one of the spurious local minima of* (5.11) *satisfies* (4.25)*,* (4.26)*, and* (4.28)*. In particular, in the case $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$, the cone K is equal to $Y \setminus V$ and* (5.11) *possesses spurious local minima for all y_d that cannot be fitted precisely with a constant function.*

Proof This follows completely analogously to the proof of Corollary 5.17 with Lemma 5.11 replaced by Lemma 5.12. ■

Some remarks regarding the last two results are in order:

Remark 5.19

- *Corollary 5.17 covers in particular neural networks with ReLU-, leaky ReLU-, ISRL-, and ELU-activation functions. Corollary 5.18 applies, for instance, to networks that involve a binary, ReLU-, or SQNL-layer.*
- *As the proofs of Corollaries 5.17 and 5.18 (or Theorems 4.15 and 4.16, respectively) are constructive, they can also be used to find explicit examples of data sets that give rise to spurious local minima in* (5.11)*. We do not pursue this approach here to avoid overloading the paper.*
- *In the case* $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$*, i.e., in the situation where all vectors are realizable, Corollary 5.17 yields the same result as (Ding et al., 2020, Corollary 2) (albeit under weaker assumptions on the network widths* w_i*,* $i = 1, ..., L$*). Corollaries 5.17 and 5.18 are further similar in nature to (Yun et al., 2019, Theorem 1) where the existence of spurious local minima in squared-loss training problems for one-hidden-layer neural networks with leaky ReLU activation functions is proved for all label vectors that are not affine-linearly fittable. Note that, in Corollaries 5.17 and 5.18, we only obtain a result of comparable strength in the case* $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$*. If the assumption of realizability is violated, then our analysis only yields that there exists an open nonempty cone* $K \subset Y$ *of label vectors for which the problem* (5.11) *possesses spurious local minima. However, in contrast to (Yun et al., 2019, Theorem 1), Corollaries 5.17 and 5.18 also cover neural networks with output dimension* $d_y > 1$*, depth* $L > 1$*, and activation functions* σ_i *that are not positively homogeneous and additionally also show that the spurious local minima of* (5.11) *can be arbitrarily bad in relative and absolute terms and in terms of loss. The statements on the size of the set of label vectors with spurious local minima in Corollaries 5.17 and 5.18 are thus weaker than that of (Yun et al., 2019, Theorem 1) but our results are also far more general. In particular, they also cover the analytically very challenging and in practice due to mild overparameterization frequently appearing situation where the considered network is deep and the assumption of realizability is violated (or, alternatively, simply not verifiable). At least to the best of the author's knowledge, results on the existence of spurious local minima of a similar strength and generality can currently not be found in the literature.*
- *Recall that the number* $\Theta(\Psi, x_d)$ *is a measure for the worst-case approximation error in the situation of* (5.11) *and the extent to which the approximation property* II) *is satisfied, see* (4.3) *and Definition 4.3. Corollaries 5.17 and 5.18 thus imply that, for neural networks with activation functions that are affine on some open nonempty subset of their domain of definition, an improved expressiveness of the map* $\Psi(\cdot, x_d) \colon D \to Y$ *necessarily comes at the price of a larger cone* K *of label vectors* y_d *that give rise to spurious local minima in* (5.11)*. This shows that there is indeed "no free lunch" in the situation of Corollaries 5.17 and 5.18.*

- *If it can be shown that a neural network parameterizes multiple subspaces in the sense of Theorem 4.16, then one can, of course, also invoke this result multiple times. This then allows to prove that certain y_d give rise to training problems of the form* (5.11) *that possess several different spurious local minima. Using Theorem 4.15, it is further easy to also establish results on the existence of spurious local minima for training problems of the form* (5.11) *that involve neural networks whose activation functions are polynomial on an open subset of their domain of definition. We omit discussing these extensions of our analysis in detail in this paper.*

For networks satisfying the assumptions of Corollaries 5.17 and 5.18, we also have:

Corollary 5.20 (Nonuniqueness and Instability in the Presence of Realizability) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i are such that Lemma 5.8, Theorem 5.9, or Corollary 5.10 can be applied to ψ. Assume further that $\mathrm{cl}_Y(\Psi(D, x_d)) = Y$ holds and that one of the following is true:*

i) For every $i \in \{1, ..., L\}$, there exists an open nonempty interval I_i such that σ_i is affine and non-constant on I_i and it holds $\min(d_\chi, d_y) \leq \min(w_1, ..., w_L)$ and $n > d_\chi + 1$.

ii) The functions σ_i, $i = 1, ..., L$, are bounded on bounded sets and there exists an index $j \in \{1, ..., L\}$ such that σ_j is constant on an open nonempty interval I_j.

Then, the solution map

$$Y \ni y_d \mapsto \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 \subset D$$

of the training problem (5.11) *is discontinuous in the sense that there exist uncountably many $y_d \in Y$ such that there are an open set $U \subset D$, an $\bar{\alpha} \in D$, and a family $\{y_d^s\}_{s>0} \subset Y$ satisfying $\bar{\alpha} \in U$, $y_d^s \to y_d$ for $s \to 0$,*

$$\bar{\alpha} \in \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2,$$

and

$$U \cap \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 = \emptyset \qquad \forall s > 0.$$

Further, in the above situation, there exist uncountably many $y_d \in Y$ such that (5.11) *is not uniquely solvable in the sense that there are an $\bar{\alpha} \in D$, an open set $U \subset D$, and a family $\{\alpha_s\}_{s>0}$ satisfying $\bar{\alpha} \in U$, $\{\alpha_s\}_{s>0} \subset D \setminus U$, and*

$$\lim_{s \to 0} \|\Psi(\alpha_s, x_d) - y_d\|_Y^2 = \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 = \inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2.$$

Proof To establish this corollary, it suffices to combine Corollary 5.10, Lemmas 5.6, 5.8, 5.11 and 5.12, and Theorems 4.17 and 5.9. ∎

It remains to study the consequences that the abstract results on regularized training problems in Section 4.3 have for the neural networks in Assumption 5.5. For the sake of simplicity, in what follows, we will restrict our attention to regularization terms of the form $\nu \| \cdot \|_p^p$, $p \in [1, 2]$, $\nu > 0$, where $\| \cdot \|_p$ denotes the usual p-norm on the Euclidean space $\mathbb{R}^m \cong \mathbb{R}^{w_{L+1} \times w_L} \times \mathbb{R}^{w_{L+1}} \times ... \times \mathbb{R}^{w_1 \times w_0} \times \mathbb{R}^{w_1}$. Other regularizers can be treated completely analogously, cf. the more general setting considered in Theorems 4.21 to 4.23.

Corollary 5.21 (Regularized Training Problems for Neural Networks) *Consider the situation in Assumption 5.5 and suppose that the widths w_i and the activation functions σ_i satisfy the conditions in Theorem 5.9 (or Corollary 5.10, respectively). Assume further that $\frac{1}{2}(d_\chi+2)(d_\chi+1) < n$ holds and that the functions σ_i, $i = 1, ..., L$, are twice differentiable, and consider for an arbitrary but fixed $p \in [1, 2]$ the regularized squared-loss training problem given by*

$$\min_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu\|\alpha\|_p^p = \frac{1}{2n}\sum_{k=1}^{n} \|\psi(\alpha, \chi_d^k) - y_d^k\|_2^2 + \nu\|\alpha\|_p^p. \tag{5.25}$$

Then, the following is true:

i) For every arbitrary but fixed $C > 0$, there exist uncountably many combinations of training label vectors $y_d \in Y$ and regularization parameters $\nu > 0$ such that the origin $\bar{\alpha} = 0 \in \mathbb{R}^m \cong \mathbb{R}^{w_{L+1}\times w_L} \times \mathbb{R}^{w_{L+1}} \times ... \times \mathbb{R}^{w_1 \times w_0} \times \mathbb{R}^{w_1}$ is a spurious local minimum of (5.25) that satisfies a local quadratic growth condition of the form

$$\|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu\|\alpha\|_p^p \geq \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 + \nu\|\bar{\alpha}\|_p^p + \varepsilon\|\alpha - \bar{\alpha}\|_2^2 \quad \forall \alpha \in B_r(\bar{\alpha})$$

for some $\varepsilon, r > 0$ and

$$\inf_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu\|\alpha\|_p^p + C \leq \|\Psi(\bar{\alpha}, x_d) - y_d\|_Y^2 + \nu\|\bar{\alpha}\|_p^p.$$

The problem (5.25) can thus possess arbitrarily bad spurious local minima.

ii) For every arbitrary but fixed regularization parameter $\nu > 0$, there exist uncountably many label vectors $y_d \in Y \setminus \{0\}$ such that $\bar{\alpha} = 0$ is the unique global solution of the problem (5.25). Adding the regularization term $\nu\|\alpha\|_p^p$ to the objective function of the problem (5.11) thus necessarily compromises the approximation property II).

iii) There exist uncountably many combinations of regularization parameters $\nu > 0$ and label vectors $y_d \in Y$ such that there exist an $s_0 \geq 0$, a family $\{y_d^s\}_{s > s_0} \subset Y$, and an open neighborhood $U \subset D$ of the origin $\bar{\alpha} = 0$ satisfying $y_d^s \to y_d$ for $s \to s_0$,

$$U \cap \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d^s\|_Y^2 + \nu\|\alpha\|_p^p = \emptyset \qquad \forall s > s_0,$$

and

$$\bar{\alpha} \in \operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu\|\alpha\|_p^p.$$

Further, there exist uncountably many tuples $(y_d, \nu) \in Y \times (0, \infty)$ such that the set of solutions

$$\operatorname*{arg\,min}_{\alpha \in D} \|\Psi(\alpha, x_d) - y_d\|_Y^2 + \nu\|\alpha\|_p^p \tag{5.26}$$

of the problem (5.25) contains more than one element. The regularized training problem (5.25) thus possesses the same nonuniqueness and instability properties as the overparameterized problem in Corollary 5.20.

Proof From Lemma 5.6 and Theorem 5.9 (or Corollary 5.10, respectively), we obtain that I) and II) hold. Further, the conditions in Assumptions 2.1 and 4.18 are trivially satisfied in the considered situation (up to the isomorphism $\mathbb{R}^m \cong \mathbb{R}^{w_{L+1}\times w_L} \times \mathbb{R}^{w_{L+1}} \times ... \times \mathbb{R}^{w_1 \times w_0} \times \mathbb{R}^{w_1}$) with a twice differentiable function ϕ. The various claims of the corollary thus follow immediately from Theorems 4.21 to 4.23. Note that, as the regularization term in (5.25) is coercive, given a sequence $\{\alpha_s\}_{s>s_0}$ with the properties in the second part of Theorem 4.23, we can pass over to a convergent subsequence to obtain that the solution set in (5.26) indeed contains more than one element. This completes the proof. ∎

Note that Corollary 5.21 does not require any assumptions on the existence of unrealizable vectors or the relationship between m and n. We conclude this section with a result that illustrates that our analysis can also be applied to neural networks which possess an architecture different from that in Assumption 5.5:

Corollary 5.22 (Properties I) and II) for Residual Neural Networks) *Suppose that $\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}$, $\mathcal{Y} = \mathbb{R}^{d_y}$, $d_\mathcal{X}, d_y \in \mathbb{N}$, $n \in \mathbb{N}$, $x_d = \{\chi_d^k\}_{k=1}^n$, $L \in \mathbb{N}$, the numbers $w_i \in \mathbb{N}$, the set D, and the functions $\sigma_i \colon \mathbb{R} \to \mathbb{R}$ are as in Assumption 5.5. Suppose further that arbitrary but fixed matrices $E_i \in \mathbb{R}^{w_i \times w_{i-1}}$, $i = 1, ..., L$, are given, let $\xi_i^{A_i,b_i} \colon \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i}$ be the functions defined by*

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

where σ_i again acts componentwise on the entries of the vectors $A_i z + b_i$, and consider the residual neural network $\psi \colon D \times \mathcal{X} \to \mathcal{Y}$ defined by

$$\psi(\alpha, \chi) := \left(\xi_{L+1}^{A_{L+1},b_{L+1}} \circ ... \circ \xi_1^{A_1,b_1}\right)(\chi) \tag{5.27}$$

for all $\chi \in \mathcal{X}$ and all $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$. Assume that the activation functions σ_i satisfy

$$\lim_{0<\gamma\to\infty} \frac{1}{\gamma}\sigma_i(\gamma s) = \sigma_i^- \min(0, s) + \sigma_i^+ \max(0, s) \qquad \forall i = 1, ..., L \qquad \forall s \in \mathbb{R} \tag{5.28}$$

for some numbers $\sigma_i^-, \sigma_i^+ \in \mathbb{R}$, $i = 1, ..., L$, with $\sigma_i^- \neq \sigma_i^+$, and that it holds $w_i \geq 2$ for all $i = 2, ..., L$ and $w_1 \geq 4$. Then, the function $\Psi(\cdot, x_d) \colon D \to Y$, $\alpha \mapsto \{\psi(\alpha, \chi_d^k)\}_{k=1}^n$, associated with x_d and the neural network ψ in (5.27) possesses the properties I) *and* II).

Proof The proof of I) is trivial. To establish II), we can proceed along similar lines as in the first half of the proof of Theorem 5.9: From the definitions of the set D and the function $\Psi(\cdot, x_d) \colon D \to Y$, it follows straightforwardly that, for every arbitrary but fixed parameter vector $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$ and all $\gamma > 0$, we have

$$\begin{aligned}
&\left\{\psi\left(\frac{1}{\gamma}A_{L+1}, b_{L+1}, \gamma A_L, \gamma b_L, A_{L-1}, b_{L-1}, ..., A_1, b_1, \chi_d^k\right)\right\}_{k=1}^n \\
&= \Bigg\{\frac{1}{\gamma}A_{L+1}\Big[E_L\big(\xi_{L-1}^{A_{L-1},b_{L-1}} \circ ... \circ \xi_1^{A_1,b_1}\big)(\chi_d^k) \\
&\qquad + \sigma_L\left(\gamma A_L\big(\xi_{L-1}^{A_{L-1},b_{L-1}} \circ ... \circ \xi_1^{A_1,b_1}\big)(\chi_d^k) + \gamma b_L\right)\Big] + b_{L+1}\Bigg\}_{k=1}^n \in \mathrm{cl}_Y\left(\Psi(D, x_d)\right).
\end{aligned}$$

Here, in the borderline case $L = 1$, the "empty" composition $\xi_{L-1}^{A_{L-1},b_{L-1}} \circ ... \circ \xi_1^{A_1,b_1}$ again has to be interpreted as the identity map. By passing to the limit $0 < \gamma \to \infty$ in the above and by exploiting (5.28), we obtain that

$$\left\{ A_{L+1}\Big[\tilde\sigma_L\left(A_L\big(\xi_{L-1}^{A_{L-1},b_{L-1}} \circ ... \circ \xi_1^{A_1,b_1}\big)(\chi_d^k) + b_L\right)\Big] + b_{L+1}\right\}_{k=1}^n$$
$$= \left\{ \left(\tilde\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ \tilde\varphi_L^{A_L,b_L} \circ \xi_{L-1}^{A_{L-1},b_{L-1}} \circ ... \circ \xi_1^{A_1,b_1}\right)(\chi_d^k)\right\}_{k=1}^n \in \mathrm{cl}_Y\left(\Psi(D, x_d)\right)$$

holds for all $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$, where $\tilde\sigma_L$ denotes the ReLU-type activation function on the right-hand side of (5.28) for $i = L$, i.e., $\tilde\sigma_L(s) := \sigma_L^- \min(0, s) + \sigma_L^+ \max(0, s)$, and where $\tilde\varphi_L^{A_L,b_L}$ and $\tilde\varphi_{L+1}^{A_{L+1},b_{L+1}}$ are defined as in (5.8), i.e.,

$$\tilde\varphi_L^{A_L,b_L}(z) := \tilde\sigma_L\left(A_L z + b_L\right), \qquad \tilde\varphi_{L+1}^{A_{L+1},b_{L+1}}(z) := A_{L+1} z + b_{L+1}.$$

Using exactly the same saturation argument as above for the remaining layers of the network (starting with the topmost unsaturated layer and then proceeding downwards and exploiting the continuity of the functions $\tilde\sigma_i(s) := \sigma_i^- \min(0, s) + \sigma_i^+ \max(0, s)$, $i = 1, ..., L$) yields that

$$\left\{ \left(\tilde\varphi_{L+1}^{A_{L+1},b_{L+1}} \circ \tilde\varphi_L^{A_L,b_L} \circ ... \circ \tilde\varphi_1^{A_1,b_1}\right)(\chi_d^k)\right\}_{k=1}^n \in \mathrm{cl}_Y\left(\Psi(D, x_d)\right)$$

holds for all $\alpha = (A_{L+1}, b_{L+1}, ..., A_1, b_1) \in D$, where $\tilde\varphi_i^{A_i,b_i}$, $i = 1, ..., L+1$, are the functions in (5.8) associated with the activations $\tilde\sigma_i$ and where the set $\mathrm{cl}_Y\left(\Psi(D, x_d)\right)$ still refers to the original network ψ in (5.27). The above shows that the closure $\mathrm{cl}_Y(\Psi(D, x_d))$ has to be at least as big as the set $\mathrm{cl}_Y(\tilde\Psi(D, x_d))$ that is obtained from the function $\tilde\Psi(\cdot, x_d)\colon D \to Y$, $\alpha \mapsto \{\tilde\psi(\alpha, \chi_d^k)\}_{k=1}^n$, associated with a neural network $\tilde\psi$ that possesses the architecture in Assumption 5.5 and involves the ReLU-type activation functions $\tilde\sigma_i$, $i = 1, ..., L$. Since this network $\tilde\psi$ satisfies II) by our assumptions on the widths w_i, $i = 1, ..., L$, and Theorem 5.9, it now follows immediately that

$$\min_{y \in \mathrm{cl}_Y(\Psi(D,x_d))} \|y - y_d\|_Y^2 \le \min_{y \in \mathrm{cl}_Y\left(\tilde\Psi(D,x_d)\right)} \|y - y_d\|_Y^2 < \|y_d\|_Y^2 \qquad \forall y_d \in Y \setminus \{0\},$$

and, by Lemma 4.2, that the map $\Psi(\cdot, x_d)\colon D \to Y$ indeed possesses the property II). This completes the proof. ∎

It is easy to check that the last result applies in particular to residual neural networks of the type (5.27) that involve an arbitrary mixture of the activation functions in point 2 of Corollary 5.10. For more details on ResNets, see He et al. (2016). We remark that, with Corollary 5.22 at hand, one can again use the abstract analysis of Section 4 to obtain results analogous to Corollaries 5.13 to 5.18, 5.20 and 5.21 for the networks in (5.27). We do not state these here for the sake of brevity. Note further that the technique used in the proof of Corollary 5.22 (i.e., the idea to establish II) by saturating the activation functions and by subsequently invoking Lemma 5.8 or Theorem 5.9) also works for other architectures. Once the properties I) and II) are established, one can then again apply the theoretical machinery developed in Section 4 to the network under consideration. This flexibility is the main advantage of the general, axiomatic approach that we have taken in Section 4.

6. Conclusions and Some Further Comments

We conclude this paper with some additional remarks:

First, we would like to stress that, although the results proved in the previous sections paint a somewhat bleak picture of the optimization landscape and the stability properties of squared-loss training problems for neural networks and general nonlinear conic approximation schemes, one should keep in mind that, even when applying an optimization algorithm to a problem of the form (1.1) only allows to determine a spurious local minimum or a saddle point (which may very well happen as we have seen, for instance, in Corollary 5.17), this resulting point may still perform far better, e.g., in terms of loss than anything that is obtainable with a classical approximation approach. The fact that driving the value of the objective function of (1.1) to the global optimum may, in practice, not be possible due to spurious local minima or the instability effects discussed in Sections 4 and 5 thus does not mean that trying to solve problems of the type (1.1) is not sensible (in particular as the results obtained, for instance, with stochastic gradient descent methods often turn out to be remarkably good in applications). The main issue that arises from the observations made in Sections 4 and 5 is more one of reliability and robustness. As solving problems of the type (1.1) numerically may only provide good or locally optimal choices of the parameter $\alpha \in D$ but not globally optimal ones and since points with similar optimal or nearly optimal loss values may perform very differently even on the training data (see point ii) of Theorem 4.9), theoretical guarantees on, for example, the generalization behavior or approximation properties of global minimizers of problems of the form (1.1) may simply not apply to the points that are determined with optimization algorithms in reality. Further, due to the instability and nonuniqueness effects documented, e.g., in Theorems 4.9, 4.17 and 4.23, small perturbations of the training data or the hyper-parameters of the considered numerical solution method and/or a different behavior of stochastic components of the used optimization algorithm may affect the performance of the obtained solutions significantly. As already mentioned, these predictions of our analysis can also be confirmed in numerical experiments, cf. the results of Cunningham et al. (2000).

We would like to point out that the observation that undesirable properties of the optimization landscape may prevent a proper identification of those parameters α for which, e.g., a neural network provides the best approximation properties in a particular situation also suggests that one should be very careful with claims that nonlinear approximation instruments are able to break the curse of dimensionality. The main point here is that this curse may not only manifest itself in the fact that the number of operations or degrees of freedom in an approximation scheme has to grow exponentially with, for instance, the spatial dimension of an underlying PDE to achieve a certain prescribed precision, but also in the loss surface of the minimization problems that have to be solved in order to adapt an approximation instrument to a given function. Compare, e.g., with Corollaries 5.17 and 5.18 in this context which demonstrate that improved approximation properties are necessarily paid for in the form of a larger set of label vectors for which (1.1) possesses spurious local minima when ReLU-type neural networks are considered. To see the essential problem, one can also consider the extreme case of a continuous function $\Psi\colon \mathbb{R} \to Y$ from the real line into a (not necessarily finite-dimensional) Hilbert space $(Y, \|\cdot\|_Y)$ whose image $\Psi(\mathbb{R})$ is dense in Y (i.e., a space-filling curve). Such an approximation scheme only requires one parameter to approximate arbitrary elements of Y to an arbitrary precision and thus clearly does not suffer from the scaling behavior that classically characterizes the curse of dimensionality. However, this construction certainly does not break this curse, either, simply because, for a high- or infinite-dimensional space Y, the optimization landscape of the problem $\min_{\alpha\in\mathbb{R}} \|\Psi(\alpha) - y_d\|_Y^2$ for a given $y_d \in Y$ typically contains countably many spurious local minima and can thus not be effectively navigated with classical optimization algorithms so that identifying parameters $\alpha \in \mathbb{R}$ for which the error $\|\Psi(\alpha) - y_d\|_Y$ becomes small is in practice impossible. The results proved in Sections 4 and 5 suggest that it makes sense to interpret nonlinear approximation instruments like neural networks as elements of a spectrum which, at the one end, has linear approximation schemes (which suffer from the usual scaling problems related to the curse of dimensionality but also give rise to, e.g., squared-loss problems with the best possible optimization landscape) and, at the other end, has space-filling curves (which only require a single parameter to achieve an arbitrary approximation accuracy but also give rise to optimization problems which typically have the worst properties possible). Considering only the scaling behavior of the degrees of freedom w.r.t. an underlying dimension without taking into account the effort necessary to determine best approximating elements does not seem to be sensible when studying how neural networks and nonlinear approximation schemes in general behave in view of the curse of dimensionality.

Regarding the optimization landscape of the squared-loss training problems in (1.1), we finally would like to point out that, if y_d is not a label vector that gives rise to a problem with a spurious local minimum or a saddle point, but close to a vector that does, then the objective function of (1.1) will still possess points which are almost stationary since the gradients (or subgradients, respectively) of the function $\alpha \mapsto \|\Psi(\alpha, x_d) - y_d\|_Y^2$ depend continuously on y_d. The presence of flat regions in the optimization landscape that slow down gradient descent or may falsely indicate convergence is thus also to be expected for label vectors that are not directly covered by, e.g., Theorems 4.11, 4.13, 4.15 and 4.16. We remark that these predictions of our analysis again agree very well with what is observed in the numerical practice, cf. Dauphin et al. (2014).

Acknowledgments

This research was conducted within the International Research Training Group IGDK 1754, funded by the German Science Foundation (DFG) and the Austrian Science Fund (FWF) under project number 188264188/GRK1754.

References

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252, Long Beach, California, USA, 2019. PMLR.

Y. Arjevani and M. Field. Symmetry & critical points for a model shallow neural network. arxiv:2003.10576, 2020.

P. Auer, M. Herbster, and M. K. K. Warmuth. Exponentially many local minima for single neurons. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 316–322. MIT Press, 1996.

J. M. Borwein and J. D. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*. Cambridge University Press, Cambridge, 2010.

D. Braess. *Nonlinear Approximation Theory*. Number 7 in Springer Series in Computational Mathematics. Springer, Berlin/Heidelberg, 1986.

Y. D. Burago and V. A. Zalgaller. *Geometric Inequalities*. Springer, Berlin/Heidelberg, 1988.

O. Calin. *Deep Learning Architectures: A Mathematical Approach*. Springer Series in the Data Sciences. Springer, 2020.

Z. Chen, Y. Cao, D. Zou, and Q. Gu. How much over-parameterization is sufficient to learn deep ReLU networks? arxiv:1911.12360, 2020.

Y. Cooper. The critical locus of overparameterized neural networks. arxiv:2005.04210, 2020.

P. Cunningham, J. Carney, and S. Jacob. Stability problems with artificial neural networks and the ensemble solution. *Art. Intell. Med.*, 20(3):217–225, 2000.

I. Daubechies, R. A. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear approximation and (deep) ReLU networks. arxiv:1905.021992, 2019.

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 *NIPS*, 2014.

R. A. DeVore. Nonlinear approximation. *Acta Numer.*, pages 51–150, 1998.

T. Ding, D. Li, and R. Sun. Sub-optimal local minima exist for almost all over-parameterized neural networks. arxiv:1911.01413v2, 2020.

A. Eftekhari. Training linear neural networks: Non-local convergence and complexity results. arxiv:2002.09852, 2020.

M. Goldblum, J. Geiping, A. Schwarzschild, M. Moeller, and T. Goldstein. Truth or backpropaganda? An empirical investigation of deep learning theory. In *International Conference on Learning Representations*, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

B. Hofmann. *Regularization for Applied Inverse and Ill-Posed Problems: A Numerical Approach*. Teubner-Texte zur Mathematik. Vieweg+Teubner, 2013.

P. C. Kainen, V. Kůrková, and A. Vogt. Approximation by neural networks is not continuous. *Neurocomputing*, 29(1):47–56, 1999.

P. C. Kainen, V. Kůrková, and A. Vogt. Continuity of approximation by neural networks in L^p-spaces. *Ann. Oper. Res.*, 101:143–147, 2001.

K. Kawaguchi. Deep learning without poor local minima. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 586–594. Curran Associates, Inc., 2016.

A. Kazemipour, B. Larsen, and S. Druckmann. No spurious local minima in deep quadratic networks. arxiv:2001.00098, 2019.

V. Kůrková and M. Sanguineti. Comparison of worst case errors in linear and neural network approximation. *IEEE Trans. Inform. Theory*, 48(1):264–275, 2002.

J. Kurtz and B. Bah. An integer programming approach to deep neural networks with binary activation functions. arxiv:2007.03326v1, 2020.

T. Laurent and J. von Brecht. Deep linear neural networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning*, pages 2908–2913, 2018.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

D. Li, T. Ding, and R. Sun. On the benefit of width for neural networks: Disappearance of bad basins. arxiv:1812.11039, 2018.

Y. Li and Y. Liang. 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, pages 8168–8177, Red Hook, NY, USA, 2018. Curran Associates Inc.

S. Liang, R. Sun, J. D. Lee, and R. Srikant. Adding one neuron can eliminate all bad local minima. arxiv:1805.08671, 2018.

Q. Nguyen, M. C. Mukkamala, and M. Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*, 2019.

S. Oymak and M. Soltanolkotabi. Towards moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE J. Sel. Areas Inform. Theory*, 1(1):84–105, 2020.

K. Pieper and A. Petrosyan. Nonconvex penalization for sparse neural networks. arxiv:2004.11515, 2020.

F. Pörner. *Regularization Methods for Ill-Posed Optimal Control Problems.* Würzburg University Press, 2018.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 1998.

I. Safran and O. Shamir. Spurious local minima are common in two-layer ReLU neural networks. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80, pages 4430–4438. PMLR, 2018.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120, 2014.

D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. arxiv:1605.08361, 2016.

R. Sun. Optimization for deep learning: theory and algorithms. arxiv:1912.08957, 2019.

R. Sun, D. Li, S. Liang, T. Ding, and R. Srikant. The global landscape of neural networks: An overview. arxiv:2007.01429v1, 2020.

G. Swirszcz, W. M. Czarnecki, and R. Pascanu. Local minima in training of neural networks. arxiv:1611.06310, 2016.

L. Venturi, A. S. Bandeira, and J. Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *J. Mach. Learn. Res.*, 20:1–34, 2019.

W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2082–2090. Curran Associates Inc., 2016.

C. Wu, J. Luo, and J. D. Lee. No spurious local minima in a two hidden unit ReLU network. In *International Conference on Learning Representations Workshop*, 2018.

J. Yoon and S. J. Hwang. Combined group and exclusive sparsity for deep neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3958–3966. PMLR, 2017.

X.-H. Yu and G.-A. Chen. On the local minima free condition of backpropagation learning. *IEEE Trans. Neural Netw.*, 6(5):1300–1303, 1995.

C. Yun, S. Sra, and A. Jadbabaie. Global optimality conditions for deep neural networks. In *International Conference on Learning Representations*, 2018.

C. Yun, S. Sra, and A. Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations*, 2019.

Y. Zhou and Y. Liang. Critical points of neural networks: Analytical forms and landscape properties. arxiv:1710.11205, 2017.

D. Zou, P. M. Long, and Q. Gu. On the global convergence of training deep linear ResNets. In *International Conference on Learning Representations*, 2020.