Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

On the Existence of Optimal Shallow Feedforward Networks with ReLU Activation

Steffen Dereich * [1] and Sebastian Kassing † [2]

[1]Institute for Mathematical Stochastics, Faculty of Mathematics and Computer Science, University of Münster, Germany.
[2]Faculty of Mathematics, University of Bielefeld, Germany.

Abstract. We prove existence of global minima in the loss landscape for the approximation of continuous target functions using shallow feedforward artificial neural networks with ReLU activation. This property is one of the fundamental artifacts separating ReLU from other commonly used activation functions. We propose a kind of closure of the search space so that in the extended space minimizers exist. In a second step, we show under mild assumptions that the newly added functions in the extension perform worse than appropriate representable ReLU networks. This then implies that the optimal response in the extended target space is indeed the response of a ReLU network.

Keywords:
Neural Networks,
Shallow Networks,
Best Approximation,
ReLU Activation,
Approximatively Compact.

Article Info.:
Volume: 3
Number: 1
Pages: 1 - 22
Date: March/2024
doi.org/10.4208/jml.230903

Article History:
Received: 03/09/2023
Accepted: 24/01/2024

Communicated by:
Arnulf Jentzen

1 Introduction

Modern machine learning algorithms are commonly based on the optimization of artificial neural networks (ANNs) through gradient based algorithms. The overwhelming success of these methods in practical applications has encouraged many scientists to build the mathematical foundations of machine learning and, in particular, to identify universal structures in the training dynamics that might provide an explanation for the mind-blowing observations practitioners make. One key component of ANNs is the activation function. Among the various activation functions that have been proposed, the rectified linear unit (ReLU), which is defined as the maximum between zero and the input value, has emerged as the most widely used and most effective activation function. There are several reasons why ReLU has become such a popular choice, e.g. it is easy to implement, computational efficient and overcomes the vanishing gradient problem, which is a common issue with other activation functions when training ANNs. In this work, we point out and prove a more subtle feature of the ReLU function that separates ReLU from several other common activation functions and might be one of the key reasons for its popularity in practice: the existence of global minima in the optimization landscape.

*Corresponding author. steffen.dereich@uni-muenster.de
†skassing@math.uni-bielefeld.de

A popular line of research studies the optimization procedure (also called training) for ANNs using gradient descent (GD) type methods. Since the error function in a typical machine learning optimization task is non-linear, non-convex and even non-coercive it remains an open problem to rigorously prove (or disprove) convergence of GD even in the simple scenario of optimizing a shallow ANN, i.e. an ANN with only one hidden layer. Existing theoretical convergence results often assume the process to stay bounded, i.e. for every realization there exists a compact set such that the process does not leave this set during training, see, e.g. [3, 11, 15] for results concerning gradient flows, [1, 2] for results concerning deterministic gradient methods, [5, 7, 23, 28] for results concerning stochastic gradient methods and [8] for results concerning gradient based diffusion processes. Many results go back to classical works by Łojasiewicz concerning gradient inequalities for analytic target functions and direct consequences for the convergence of gradient flow trajectories under the assumption of staying bounded [20–22].

In this context, it seems natural to ask for the existence of ANNs that solve the minimization task within the search space. More explicitly, if there does not exist a global minimum in the optimization landscape then every sequence that approaches the minimal loss value diverges to infinity. This might lead to slow convergence or even rule out convergence of the loss value, which is the property that practitioners are most interested in. Therefore, it seems reasonable to choose a network architecture, activation function and loss function such that there exist global optima in the optimization landscape.

Overparametrized networks in the setting of empirical risk minimization (more ReLU neurons than data points to fit) are able to perfectly interpolate the data (see, e.g. [12, Lemma 27.3]) such that there exists a network configuration achieving zero error and, thus, a global minimum in the search space. For shallow feedforward ANNs using ReLU activation it has been shown that also in the underparametrized regime there exists a global minimum if the ANN has a one-dimensional output [18], whereas there are pathological counterexamples in higher dimensions [19]. However, for general measures μ not necessarily consisting of a finite number of Dirac measures, the literature on the existence of global minima is very limited. There exist positive results for the approximation of functions in the space $L^p([0,1]^d)$ with shallow feedforward ANNs using heavyside activation [16], the approximation of Lipschitz continuous target functions with shallow feedforward ANNs using ReLU activation and the standard mean square error in the case where the input and output dimension is one-dimensional [15], and the approximation of multi-dimensional, real-valued continuous target functions with shallow residual ANNs using ReLU activation [6]. On the other hand, for several common (smooth) activations such as the standard logistic activation, softplus, arctan, hyperbolic tangent and softsign there, generally, do not exist minimizers in the optimization landscape for smooth target functions (or even polynomials), see [13, 24]. This phenomenon can also be observed in empirical risk minimization for the hyperbolic tangent activation. As shown in [19], in the underparametrized setting, there exist input data such that for all output data from a set of positive Lebesgue measure there does not exist minimizers in the optimization landscape.

In this article, we prove, for the first time, existence results for shallow feedforward ReLU ANNs with multi-dimensional input space for the population loss. Interestingly, minimizers exist under very mild assumptions on the optimization problem. This exis-

tence property indicates the robustness of ReLU activation and may be a reason for its success in practical applications. For the proof we proceed as follows. First, we show existence of minimizers in an extended target space that comprises of the representable responses of ANNs and additional discontinuous generalized responses. Note that for many activation functions (including ReLU) the set of realization/response functions is not closed for an appropriate metric, see also [14]. Second, we show that the additional discontinuous responses perform worse than representable ones under mild conditions on the optimization problem. Compared to [6], where residual networks are treated, the situation is more complex for classical feedforward ReLU networks as treated here. This is caused by the more sophisticated structure of the extended search space, see Definition 2.1.

We present a special case of our main result in the situation where we focus on the approximation of continuous target functions with shallow ANNs using ReLU activation under L^p-loss.

Theorem 1.1. *Let $d_{\text{in}}, d \in \mathbb{N}$, $p>1$ and $\mathfrak{d}=(d_{\text{in}}+2)d+1$. Let $f\colon \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ and $h\colon \mathbb{R}^{d_{\text{in}}} \to [0,\infty)$ be continuous functions and assume that $h^{-1}((0,\infty))$ is a bounded convex set. For every $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ let $\operatorname{err}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be given by*

$$\operatorname{err}(\theta) = \int_{\mathbb{R}^{d_{\text{in}}}} |f(x) - \mathfrak{N}_\theta(x)|^p h(x)\, dx,$$

where

$$\mathfrak{N}_\theta(x) = \theta_{\mathfrak{d}} + \sum_{j=1}^{d} \theta_{(d_{\text{in}}+1)d+j} \max\left(\theta_{d_{\text{in}}d+j} + \sum_{i=1}^{d_{\text{in}}} \theta_{(j-1)d_{\text{in}}+i} x_i, 0\right).$$

Then there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ such that $\operatorname{err}(\theta) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \operatorname{err}(\vartheta)$.

Let us explain the statement of Theorem 1.1 in more detail. We consider the regression problem of fitting the parameters (i.e. weights and biases) $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ of a shallow neural network with input dimension d_{in}, d neurons on the hidden layer and one-dimensional output such that its response $\mathfrak{N}_\theta : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ is a good approximation of the continuous function $f : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$. If we measure the quality of the approximation in terms of the L^p-loss, where the data distribution of the input data is assumed to have continuous Lebesgue density h and a compact and convex support, then there exists a global minimum of the error function $\operatorname{err} : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ inside the search space. Theorem 1.1 is a special case of the more general Theorem 1.2, which treats a broader class of loss functions and measures.

Next, we introduce the central objects and notations of this article. In the following, we represent ANNs in a more structurized way. We consider networks with d_{in}-dimensional input space and one hidden layer consisting of d neurons that apply ReLU activation, i.e. $(x)^+ = \max(x, 0)$. We describe the weights of the ANN by a matrix $W^1 = (w^1_{j,i})_{j=1,\ldots,d, i=1,\ldots,d_{\text{in}}}$ and a row vector $W^2 = (w^2_1, \ldots, w^2_d)$, and the biases by a column vector $b^1 = (b^1_i)_{i=1,\ldots,d}$ and a scalar b^2. Moreover, for $j = 1, \ldots, d$, we write $w^1_j = (w^1_{j,1}, \ldots, w^1_{j,d_{\text{in}}})^\intercal$, where $a^\intercal$ denotes the transpose of a vector or a matrix a. We let

$$\mathbb{W} = (W^1, b^1, W^2, b^2) \in \mathbb{R}^{d \times d_{\text{in}}} \times \mathbb{R}^d \times \mathbb{R}^{1 \times d} \times \mathbb{R} =: \mathcal{W}_d,$$

and call $\mathbb{W}$ a network configuration and $\mathcal{W}_d$ the parametrization class. We often refer to a configuration of a neural network as the (neural) network $\mathbb{W}$. A configuration $\mathbb{W} \in \mathcal{W}_d$ describes a function $\mathfrak{N}^{\mathbb{W}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ via

$$\mathfrak{N}^{\mathbb{W}}(x) = \sum_{j=1}^{d} w_j^2 \left(w_j^1 \cdot x + b_j^1\right)^+ + b^2, \tag{1.1}$$

where $\cdot$ denotes the scalar product on $\mathbb{R}^{d_{\text{in}}}$. We call $\mathfrak{N}^{\mathbb{W}}$ realization function or response of the network $\mathbb{W}$. We allow as parameter d all values from $\mathbb{N}_0 := \{0, 1, \ldots\}$, where a response of a network with zero neurons is a constant function (by definition). For an introduction into general neural networks with possibly multiple hidden layers see, e.g. [24].

Note that, in general, the response of a network is a continuous, piecewise affine function from $\mathbb{R}^{d_{\text{in}}}$ to $\mathbb{R}$. We conceive $\mathbb{W} \mapsto \mathfrak{N}^{\mathbb{W}}$ as a parametrization of a class of potential response functions $\{\mathfrak{N}^{\mathbb{W}} : \mathbb{W} \in \mathcal{W}_d\}$ in a minimization problem. More explicitly, let μ be a finite measure on the Borel sets of $\mathbb{R}^{d_{\text{in}}}$, let $\mathbb{D} = \operatorname{supp}(\mu)$ and $\mathcal{L} : \mathbb{D} \times \mathbb{R} \to \mathbb{R}_+$ be a product-measurable function, the loss function. We aim to minimize the error

$$\operatorname{err}^{\mathcal{L}}(\mathbb{W}) = \int_{\mathbb{D}} \mathcal{L}\left(x, \mathfrak{N}^{\mathbb{W}}(x)\right) \mathrm{d}\mu(x)$$

over all $\mathbb{W} \in \mathcal{W}_d$ for a given $d \in \mathbb{N}_0$ and let

$$\operatorname{err}_d^{\mathcal{L}} = \inf_{\mathbb{W} \in \mathcal{W}_d} \operatorname{err}^{\mathcal{L}}(\mathbb{W}) \tag{1.2}$$

be the minimal error for the optimization task when using a neural network with d neurons on the hidden layer.

The aim of this work is to give sufficient conditions on the loss function $\mathcal{L}$ and the measure μ that guarantee existence of a network $\mathbb{W} \in \mathcal{W}_d$ with $\operatorname{err}^{\mathcal{L}}(\mathbb{W}) = \operatorname{err}_d^{\mathcal{L}}$. We stress that if there does not exist a neural network $\mathbb{W} \in \mathcal{W}_d$ satisfying $\operatorname{err}^{\mathcal{L}}(\mathbb{W}) = \operatorname{err}_d^{\mathcal{L}}$ then every sequence $(\mathbb{W}_n)_{n \in \mathbb{N}} \subset \mathcal{W}_d$ of networks satisfying $\lim_{n\to\infty} \operatorname{err}^{\mathcal{L}}(\mathbb{W}_n) = \operatorname{err}_d^{\mathcal{L}}$ diverges to infinity.

We state the main result of this article.

Theorem 1.2. *Suppose that $\mathbb{D} = \operatorname{supp}(\mu)$ is compact and that μ has a continuous Lebesgue density $h : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}_+$. Assume that for every hyperplane H that intersects the interior of the convex hull of $\mathbb{D}$, there exists an $x \in H$ with $h(x) > 0$. Moreover, assume that the loss function $\mathcal{L} : \mathbb{D} \times \mathbb{R} \to \mathbb{R}_+$ satisfies the following assumptions:*

(i) (Continuity in the First Argument) For every $y \in \mathbb{R}$, $\mathbb{D} \ni x \mapsto \mathcal{L}(x, y)$ is continuous.

(ii) (Strict Convexity in the Second Argument) For all $x \in \mathbb{D}$, $y \mapsto \mathcal{L}(x, y)$ is strictly convex and attains its minimum.

Then, for every $d \in \mathbb{N}_0$, there exists an optimal network $\mathbb{W} \in \mathcal{W}_d$ with $\operatorname{err}^{\mathcal{L}}(\mathbb{W}) = \operatorname{err}_d^{\mathcal{L}}$.

Theorem 1.2 is an immediate consequence of Proposition 3.1 below. We stress that the statement of Proposition 3.1 is stronger in the sense that it even shows that in many situations the newly added functions to the extended target space perform strictly worse than the representable responses. We get the statement of Theorem 1.1 as a corollary of Theorem 1.2 as explained in the following example. Note that if μ has a continuous Lebesgue density and a compact and convex support then μ satisfies the assumptions in Theorem 1.2.

Example 1.1 (Regression Problem)**.** Let μ be as in Theorem 1.2 and suppose that $f\colon\mathbb{R}^{d_{\text{in}}}\to\mathbb{R}$ is a continuous and $L : \mathbb{R} \to \mathbb{R}_+$ a strictly convex function that attains its minimum. Then $\mathcal{L} : \mathbb{R}^{d_{\text{in}}} \times \mathbb{R} \to \mathbb{R}_+$ given by

$$\mathcal{L}(x, y) = L\big(y - f(x)\big),$$

satisfies the assumptions of the latter theorem and, thus, the infimum

$$\inf_{\mathbb{W}\in\mathcal{W}_d} \int L\big(\mathfrak{N}^{\mathbb{W}}(x) - f(x)\big)\,\mathrm{d}\mu(x)$$

is attained for a network $\mathbb{W} \in \mathcal{W}_d$.

For a general introduction into best approximators in normed spaces we refer the reader to [26]. A good literature review regarding the loss landscape in neural network training can be found in [10]. For statements about the existence of non-optimal local minima in the training of (shallow) networks we refer the reader to [4, 25, 27, 29]. Lastly, we note that weight regularization can also be used to ensure the existence of a global optimum. In particular, consider the error function

$$\mathrm{err}^{\mathcal{L},\mathcal{P}}(\mathbb{W}) := \int \mathcal{L}\big(x, \mathfrak{N}^{\mathbb{W}}(x)\big)\,\mathrm{d}\mu(x) + \mathcal{P}(\mathbb{W}),$$

where $\mathcal{P}$ is a penalty term that satisfies $\mathcal{P}(\mathbb{W}) \to \infty$ as $|\mathbb{W}| \to \infty$. Assuming continuity of $\mathrm{err}^{\mathcal{L},\mathcal{P}}$ one can use compactness arguments to show that there exists an ANN minimizing the error function. In that case, there exist results proving boundedness of the SGD paths, see, e.g., [23, Theorem 1], [17, Proposition 1] and [9, Lemma D.1].

2 Generalized response of neural networks

We will work with more intuitive geometric descriptions of realization functions of networks $\mathbb{W} \in \mathcal{W}_d$ as introduced in [6]. We call a network $\mathbb{W} \in \mathcal{W}_d$ non-degenerate if for all $j = 1, \dots, d$ we have $w_j^1 \neq 0$. For a non-degenerate network $\mathbb{W}$, we say that the neuron $j \in \{1, \dots, d\}$ has

- normal $\mathfrak{n}_j = (1/|w_j^1|)w_j^1 \in \mathbb{S}^{d_{\text{in}}-1} := \{x \in \mathbb{R}^{d_{\text{in}}} : |x| = 1\}$,
- offset $o_j = -(1/|w_j^1|)b_j^1 \in \mathbb{R}$,
- kink $\Delta_j = |w_j^1|w_j^2 \in \mathbb{R}$.

Moreover, we call $\mathfrak{b} = b^2$ the bias of $\mathbb{W}$. We call $(\mathfrak{n}, o, \Delta, \mathfrak{b})$ with

$$\mathfrak{n} = (\mathfrak{n}_1, \dots, \mathfrak{n}_d) \in (\mathbb{S}^{d_{\text{in}}-1})^d, \quad o = (o_1, \dots, o_d) \in \mathbb{R}^d, \quad \Delta = (\Delta_1, \dots, \Delta_d) \in \mathbb{R}^d,$$

and $\mathfrak{b} \in \mathbb{R}$ the effective tuple of $\mathbb{W}$ and write $\mathcal{E}_d$ for the set of all effective tuples using d ReLU neurons.

First we note that the response of a non-degenerate network $\mathbb{W}$ can be represented in terms of its effective tuple: One has, for $x \in \mathbb{R}^{d_{\text{in}}}$,

$$\begin{aligned}\mathfrak{N}^{\mathbb{W}}(x) &= \mathfrak{b} + \sum_{j=1}^{d} w_j^2 \left(w_j^1 \cdot x + b_j^1\right)^+ = \mathfrak{b} + \sum_{j=1}^{d} \Delta_j \left(\frac{1}{|w_j^1|} w_j^1 \cdot x + \frac{1}{|w_j^1|} b_j^1\right)^+ \\ &= \mathfrak{b} + \sum_{j=1}^{d} \Delta_j (\mathfrak{n}_j \cdot x - o_j)^+.\end{aligned}$$

With slight misuse of notation we also write

$$\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}, \quad x \mapsto \mathfrak{b} + \sum_{j=1}^{d} \Delta_j (\mathfrak{n}_j \cdot x - o_j)^+,$$

and

$$\text{err}^{\mathcal{L}}(\mathfrak{n}, o, \Delta, \mathfrak{b}) = \int \mathcal{L}\left(x, \mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}(x)\right) \mathrm{d}\mu(x).$$

Although the tuple $(\mathfrak{n}, o, \Delta, \mathfrak{b})$ does not uniquely describe a neural network, it describes a response function uniquely and thus we will speak of the neural network with effective tuple $(\mathfrak{n}, o, \Delta, \mathfrak{b})$.

We stress that the response of a degenerate network $\mathbb{W}$ can also be described as response associated to an effective tuple. Indeed, for every $j \in \{1, \dots, d\}$ with $w_j^1 = 0$ the respective neuron has a constant contribution $w_j^2 (b_j^1)^+$. Now, one can choose an arbitrary normal $\mathfrak{n}_j$ and offset o_j, set the kink equal to zero ($\Delta_j = 0$) and add the constant $w_j^2 (b_j^1)^+$ to the bias $\mathfrak{b}$. Repeating this procedure for every such neuron we get an effective tuple $(\mathfrak{n}, o, \Delta, \mathfrak{b}) \in \mathcal{E}_d$ that satisfies $\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}} = \mathfrak{N}^{\mathbb{W}}$. Conversely, for every effective tuple $(\mathfrak{n}, o, \Delta, \mathfrak{b}) \in \mathcal{E}_d$, $\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}$ is the response of an appropriate network $\mathbb{W} \in \mathcal{W}_d$. In fact one can choose $b^2 = \mathfrak{b}$ and, for $j = 1, \dots, d$, $w_j^1 = \mathfrak{n}_j$, $b_j^1 = -o_j$ and $w_j^2 = \Delta_j$ such that for all $x \in \mathbb{R}^{d_{\text{in}}}$,

$$w_j^2 \left(w_j^1 \cdot x + b_j^1\right)^+ = \Delta_j (\mathfrak{n}_j \cdot x - o_j)^+.$$

This entails that

$$\text{err}_d^{\mathcal{L}} = \inf_{(\mathfrak{n},o,\Delta,\mathfrak{b}) \in \mathcal{E}_d} \int \mathcal{L}\left(x, \mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}(x)\right) \mathrm{d}\mu(x),$$

and the infimum is attained if there is a network $\mathbb{W} \in \mathcal{W}_d$ for which the infimum in (1.2) is attained. For an effective tuple $(\mathfrak{n}, o, \Delta, \mathfrak{b}) \in \mathcal{E}_d$, we say that the j-th ReLU neuron has the breakline

$$H_j = \left\{x \in \mathbb{R}^{d_{\text{in}}} : \mathfrak{n}_j \cdot x = o_j\right\},$$

and we call

$$A_j = \{x \in \mathbb{R}^{d_{\text{in}}} : \mathfrak{n}_j \cdot x > o_j\} \tag{2.1}$$

the domain of activity of the j-th ReLU neuron. By construction, we have

$$\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}(x) = \mathfrak{b} + \sum_{j=1}^{d} \mathbb{1}_{A_j}(x)\big(\Delta_j(\mathfrak{n}_j \cdot x - o_j)\big).$$

Outside the breaklines, the function $\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{b}}$ is differentiable with

$$D\mathfrak{N}^{\mathfrak{n},o,\Delta,\mathfrak{a}}(x) = \sum_{j=1}^{d} \mathbb{1}_{A_j}(x)\Delta_j\mathfrak{n}_j.$$

Note that for each summand $j = 1, \dots, d$ along the breakline the difference of the differential on A_j and $\overline{A_j}^c$ equals $\Delta_j\mathfrak{n}_j$ (which is also true for the response function $\mathfrak{N}^{\mathbb{W}}$ provided that it is differentiable in the reference points and there does not exist a second neuron having the same breakline H_j).

In empirical risk minimization, one can deduce existence of a global minimum for the error function $\sum_{i=1}^{n} L(y_i, \mathfrak{N}^{\mathbb{W}}(x_i))$, where $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}$, by showing closedness of the set $\{(\mathfrak{N}^{\mathbb{W}}(x_1), \dots, \mathfrak{N}^{\mathbb{W}}(x_n)) : \mathbb{W} \in \mathcal{W}_d\}$, see [18, Proposition 3.1]. However, [24, Theorem 3.1] shows that the set $\{\mathfrak{N}^{\mathbb{W}} : \mathbb{W} \in \mathcal{W}_d\}$ is not closed in $L^p(\mu)$ for any $p > 0$ and measure μ that has a continuous Lebesgue density and compact support. The main task of this article is to show that the additional limiting functions provide larger errors than network responses.

We introduce the class of generalized network responses. This extension of the search space, consisting of network responses and the additional limiting functions, has the advantage that under quite mild assumptions minimizers can be found by applying compactness arguments.

Definition 2.1. *We call a function $\mathcal{R} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ generalized response if it admits the following representation: There are $K \in \mathbb{N}_0$, a tuple of open half-spaces $\mathbf{A} = (A_1, \dots, A_K)$ of $\mathbb{R}^{d_{\text{in}}}$, an affine mapping $\mathfrak{a} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$, vectors $\delta_1, \dots, \delta_K \in \mathbb{R}^{d_{\text{in}}}$ and reals $\mathfrak{b}_1, \dots, \mathfrak{b}_K \in \mathbb{R}$ such that for all $x \in \mathbb{R}^{d_{\text{in}}}$,*

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{k=1}^{K} \mathbb{1}_{A_k}(x)(\delta_k \cdot x + \mathfrak{b}_k). \tag{2.2}$$

We assign a representation (2.2) *a multiplicity as follows: For every $k = 1, \dots, K$, we assign the k-th term the multiplicity $m_k = 1$, if $x \mapsto \mathbb{1}_{A_k}(x)(\delta_k \cdot x + \mathfrak{b}_k)$ is continuous or, equivalently, $\partial A_k \subset \{x \in \mathbb{R}^{d_{\text{in}}} : \delta_k \cdot x + \mathfrak{b}_k = 0\}$ and otherwise multiplicity $m_k = 2$. Moreover, we assign the affine term $\mathfrak{a}$ multiplicity $m_0 = 0$, if*

(a) *$\mathfrak{a}$ is a constant function or*

(b) *$(\mathfrak{n}_k : k \in \{1, \dots, K\}$ with $m_k = 2)$ is linearly dependent, where each $\mathfrak{n}_k$ is a normal of the hyperplane ∂A_k as in* (2.1)*,*

and otherwise multiplicity $m_0 = 1$. Then $m_0 + \cdots + m_K$ is said to be the multiplicity of the representation (2.2).

A generalized response admits various representations and the minimal multiplicity $m_0 + \cdots + m_K$ that can be achieved is called dimension of the generalized response.

We call $A_1, \ldots, A_K$ active half-spaces of the response, $m_1, \ldots, m_K$ the multiplicities of the half-spaces $A_1, \ldots, A_K$ or summands. For $d \in \mathbb{N}_0$, we denote by $\mathfrak{R}_d$ the space of all generalized responses of dimension at most d. Moreover, we call a response $\mathcal{R} \in \mathfrak{R}_d$ strict at dimension d if it is of dimension $d-1$ or lower or if it is discontinuous. Denote by $\mathfrak{R}_d^{\text{strict}}$ the set of strict responses at dimension d. Moreover, we call a response representable if it is continuous or, equivalently, if it admits a representation with all multiplicities $m_1, \ldots, m_K$ being one.

We conceive the space of generalized responses of dimension d as an extension of the space $\{\mathfrak{N}^{\mathbb{W}} : \mathbb{W} \in \mathcal{W}_d\}$, with the representable responses being responses for neural networks with d neurons on the hidden layer and the discontinuous generalized responses being additional limits in $L^p(\mu)$. Strictly speaking, a representable response of dimension d is not necessarily the response of a network with d neurons on the hidden layer since in the case where $m_0 = 1$ we might need two ReLU neurons (instead of one) to generate the linear component of $\mathfrak{a}$. However, for every compact set $K \subset \mathbb{R}^{d_{\text{in}}}$ and every representable response $\mathcal{R} \in \mathfrak{R}_d$ we can find a network with d neurons on the hidden layer whose response agrees on K with $\mathcal{R}$. Consequently, for compactly supported measures μ, the subset of representable generalized responses can all be realized on the relevant domain by appropriate shallow networks.

For a better understanding of the space of generalized responses, we give the following lemma.

Lemma 2.1. *Let μ be a finite measure on the Borel sets of $\mathbb{R}^{d_{\text{in}}}$ with a continuous Lebesgue density such that $\mathbb{D} = \operatorname{supp}(\mu)$ is compact. Let $d \in \mathbb{N}$ and $\mathcal{R} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ be a generalized response of dimension d. Then there exists a sequence $(\mathcal{R}_n)_{n\in\mathbb{N}} \subset \{\mathfrak{N}^{\mathbb{W}} : \mathbb{W} \in \mathcal{W}_d\}$ of network responses with d neurons on the hidden layer such that for all $p > 0$,*

$$\mathcal{R}_n \;\to\; \mathcal{R} \quad \text{in } L^p(\mu).$$

Proof. Since μ has a continuous Lebesgue density and compact support it suffices to show that $\mathcal{R}_n(x) \to \mathcal{R}(x)$ for Lebesgue-almost all $x \in \mathbb{D}$. [6, Remark 3.3] shows that for every open half-space A, $\delta \in \mathbb{R}^{d_{\text{in}}}$ and $\mathfrak{b} \in \mathbb{R}$ the function $\tilde{\mathcal{R}}(x) = \mathbb{1}_A(\delta \cdot x + \mathfrak{b})$ is on $\mathbb{R}^{d_{\text{in}}} \setminus (\partial A)$ the limit of the response of two ReLU neurons. Moreover, if $\tilde{\mathcal{R}}(x)$ is continuous then $\tilde{\mathcal{R}}(x)$ is clearly the response of a single ReLU neuron. This implies the statement in the case where $\mathfrak{a}$ is a constant function. If $m_0 = 1$ the affine function $\mathfrak{a}$ can be realized on $\mathbb{D}$ as the response of one neuron such that $\mathbb{D}$ completely lies within its domain of activity. It remains to show the statement for a generalized response $\mathcal{R}$ of the form

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{k=1}^{K} \mathbb{1}_{A_k}(x)(\delta_k \cdot x + \mathfrak{b}_k),$$

where $K \in \mathbb{N}, m_0 = 0, m_1, \dots, m_K = 2$ and there exists $(\alpha_1, \dots, \alpha_K) \in \mathbb{R}^K$ with $\alpha_k \neq 0$ for all $k = 1, \dots, K$ such that

$$\sum_{k=1}^{K} \alpha_k \mathfrak{n}_k = 0$$

with $\mathfrak{n}_1, \dots, \mathfrak{n}_K$ being as in (2.1). By switching the sides of the active areas we can assume without loss of generality that all $\alpha_k > 0$ (indeed this will change the definition of $\mathcal{R}$ only on the respective hyperplanes). Set $\delta_1^+ = \mathfrak{a}' + \delta_1, \delta_1^- = \mathfrak{a}'$ (with $\mathfrak{a}'$ being the derivative of $\mathfrak{a}$) as well as $\delta_k^+ = \delta_k$ and $\delta_k^- = 0$ for all $k = 2, \dots, K$ so that for almost all $x \in \mathbb{D}$,

$$\mathcal{R}(x) = \mathfrak{a}(0) + \sum_{k=1}^{K} \mathbb{1}_{A_k}(x)\big(\delta_k^+ \cdot x + \mathfrak{b}_k\big) + \mathbb{1}_{A_k^c}(x)\big(\delta_k^- \cdot x\big).$$

Now, for $\kappa > 0$ the function

$$\mathcal{R}^\kappa(x) := \mathfrak{b}^\kappa + \sum_{k=1}^{K} \mathcal{N}_k^\kappa(x),$$

where $\mathfrak{b}^\kappa := \mathfrak{a}(0) + \sum_{k=1}^{K} \kappa \alpha_k o_k$ with $o_1, \dots, o_K$ being as in (2.1) and

$$\mathcal{N}_k^\kappa(x) := \big(\delta_k^+ \cdot x + \mathfrak{b}_k + \kappa\alpha_k(\mathfrak{n}_k \cdot x - o_k)\big)^+ - \big(-\delta_k^- \cdot x - \kappa\alpha_k(\mathfrak{n}_k \cdot x - o_k)\big)^+$$

is the response of a neural network having $2K$ neurons on the hidden layer. Moreover, $\mathcal{R}^\kappa(x) \to \mathcal{R}$ as $\kappa \to \infty$ for all $x \notin \bigcup_{k=1}^{K} \partial A_k$. □

When analyzing the minimization problem over the class of generalized responses, we can impose weaker assumptions than in Theorem 1.2. We will use the following concepts.

Definition 2.2. *(i) An element x of a hyperplane $H \subset \mathbb{R}^{d_{\text{in}}}$ is called H-regular if $x \in \operatorname{supp} \mu|_A$ and $x \in \operatorname{supp} \mu|_{\overline{A}^c}$, where A is an open half-space with $\partial A = H$.*

(ii) A measure μ on $\mathbb{R}^{d_{\text{in}}}$ is called nice if all hyperplanes have μ-measure zero and if for every open half-space A with $\mu(A), \mu(\overline{A}^c) > 0$ the set of ∂A-regular points cannot be covered by finitely many hyperplanes different from ∂A.

Proposition 2.1. *Assume that μ is a nice measure on $\mathbb{R}^{d_{\text{in}}}$ and that the loss function $\mathcal{L} : \mathbb{D} \times \mathbb{R} \to \mathbb{R}_+$ is measurable and satisfies the following assumptions:*

(i) (Lower-Semicontinuity in the Second Argument) For all $x \in \mathbb{D}$ and $y \in \mathbb{R}$, we have

$$\liminf_{y' \to y} \mathcal{L}(x, y') \geq \mathcal{L}(x, y).$$

(ii) (Unbounded in the Second Argument) For all $x \in \mathbb{D}$, we have

$$\lim_{|y| \to \infty} \mathcal{L}(x, y) = \infty.$$

Let $d \in \mathbb{N}_0$ with $\mathrm{err}_d^{\mathcal{L}} < \infty$. Then there exists an $\mathcal{R} \in \mathfrak{R}_d$ with

$$\int \mathcal{L}\big(x, \mathcal{R}(x)\big)\,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_d^{\mathcal{L}} := \inf_{\tilde{\mathcal{R}} \in \mathfrak{R}_d} \int \mathcal{L}\big(x, \tilde{\mathcal{R}}(x)\big)\,\mathrm{d}\mu(x).$$

Furthermore, for $d \geq 1$ the infimum $\inf_{\tilde{\mathcal{R}} \in \mathfrak{R}_d^{\mathrm{strict}}} \int \mathcal{L}(x, \tilde{\mathcal{R}}(x))\,\mathrm{d}\mu(x)$ is attained on $\mathfrak{R}_d^{\mathrm{strict}}$.

Proof. Let $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ be a sequence of generalized responses of at most dimension d that satisfy

$$\lim_{n\to\infty} \int \mathcal{L}\big(x, \mathcal{R}^{(n)}(x)\big)\,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_d^{\mathcal{L}}.$$

We use the representations as in (2.2) and write

$$\mathcal{R}^{(n)}(x) = \mathfrak{a}^{(n)}(x) + \sum_{k=1}^{K_n} \mathbb{1}_{A_k^{(n)}}\big(\delta_k^{(n)} \cdot x + \mathfrak{b}_k^{(n)}\big).$$

Moreover, for $k = 1, \ldots, K_n$ we denote by $\mathfrak{n}_k^{(n)} \in \mathbb{S}^{d_{\mathrm{in}}-1}$ and $o_k^{(n)} \in \mathbb{R}$ the quantities with

$$A_k^{(n)} = \big\{x \in \mathbb{R}^{d_{\mathrm{in}}} : \mathfrak{n}_k^{(n)} \cdot x > o_k^{(n)}\big\},$$

and by $m_k^{(n)}$ the multiplicity of the k-th term. We also denote by $m_0^{(n)}$ the multiplicity of the affine term and assume that the representation satisfies $m_0^{(n)} + \cdots + m_{K_n}^{(n)} \leq d$.

Step 1. Deriving a limit admitting a representation (2.2).

We choose a subsequence $(n_l)_{l\in\mathbb{N}}$ along which always the K-number and the multiplicities are the same and so that, in the case that $m_0^{(n_l)} \equiv 0$, always the same case (a), (b) enters. Moreover, we assume that for each $k = 1, \ldots, K$, $(\mathfrak{n}_k^{(n_l)})_{l\in\mathbb{N}}$ converges in $\mathbb{S}^{d_{\mathrm{in}}-1}$ to $\mathfrak{n}_k$ and $(o_k^{(n_l)})_{l\in\mathbb{N}}$ in $\mathbb{R} \cup \{\pm\infty\}$ to o_k. For ease of notation we will assume that this is the case for the full sequence.

We call

$$A_k = \big\{x \in \mathbb{R}^{d_{\mathrm{in}}} : \mathfrak{n}_k \cdot x > o_k\big\}$$

the asymptotic active area of the k-th term and let $H_k = \partial A_k$. Let $\mathbb{J}$ denote the collection of all subsets $J \subset \{1, \ldots, K\}$ for which the set

$$A_J = \bigcap_{j\in J} A_j \cap \bigcap_{j\in J^c} \overline{A}_j^c$$

satisfies $\mu(A_J) > 0$. We note that the sets $(A_J : J \in \mathbb{J})$ are non-empty, open and pairwise disjoint and their union has full μ-measure since

$$\mu\Bigg(\mathbb{R}^{d_{\mathrm{in}}} \setminus \bigcup_{J\subset\{1,\ldots,K\}} A_J\Bigg) \leq \sum_{j=1}^{K} \mu(H_j) = 0.$$

Moreover, for every $J \in \mathbb{J}$ and every compact set B with $B \subset A_J$ one has from a B-dependent n onwards that the generalized response $\mathcal{R}^{(n)}$ satisfies for all $x \in B$ that

$$\mathcal{R}^{(n)}(x) = \mathcal{D}_J^{(n)} \cdot x + \beta_J^{(n)},$$

where

$$\mathcal{D}_J^{(n)} := \mathfrak{a}'^{(n)} + \sum_{j \in J} \delta_j^{(n)}, \quad \beta_J^{(n)} := \mathfrak{a}^{(n)}(0) + \sum_{j \in J} \mathfrak{b}_j^{(n)}.$$

Let $J \in \mathbb{J}$. Next, we show that along an appropriate subsequence, we have convergence of $(\mathcal{D}_J^{(n)})_{n \in \mathbb{N}}$ in $\mathbb{R}^{d_{\text{in}}}$. First assume that along a subsequence one has that $(|\mathcal{D}_J^{(n)}|)_{n \in \mathbb{N}}$ converges to ∞. For ease of notation we assume without loss of generality that one has $|\mathcal{D}_J^{(n)}| \to \infty$. We let

$$\mathcal{H}_J^{(n)} = \{x \in \mathbb{R}^{d_{\text{in}}} : \mathcal{D}_J^{(n)} \cdot x + \beta_J = 0\}.$$

For every n with $\mathcal{D}_J^{(n)} \neq 0, \mathcal{H}_J^{(n)}$ is a hyperplane which can be parametrized by taking a normal and the respective offset. As above we can argue that along an appropriate subsequence (which is again assumed to be the whole sequence) one has convergence of the normals in $\mathbb{S}^{d_{\text{in}}-1}$ and of the offsets in $\bar{\mathbb{R}}$. We denote by $\mathcal{H}_J$ the hyperplane being associated to the limiting normal and offset (which is assumed to be the empty set in the case where the offsets do not converge in $\mathbb{R}$). Since the norm of the gradient $\mathcal{D}_J^{(n)}$ tends to infinity we get that for every $x \in A_J \backslash \mathcal{H}_J$ one has $|\mathcal{R}^{(n)}(x)| \to \infty$ and, hence, $\mathcal{L}(x, \mathcal{R}^{(n)}(x)) \to \infty$. Consequently, Fatou's lemma implies that

$$\liminf_{n \to \infty} \int_{A_J \backslash \mathcal{H}_J} \mathcal{L}\big(x, \mathcal{R}^{(n)}(x)\big) \, \mathrm{d}\mu(x) \geq \int_{A_J \backslash \mathcal{H}_J} \liminf_{n \to \infty} \mathcal{L}\big(x, \mathcal{R}^{(n)}(x)\big) \, \mathrm{d}\mu(x) = \infty$$

contradicting the asymptotic optimality of $(\mathcal{R}^{(n)})_{n \in \mathbb{N}}$. We showed that the sequence $(\mathcal{D}_J^{(n)})_{n \in \mathbb{N}}$ is precompact and by switching to an appropriate subsequence we can guarantee that the limit $\mathcal{D}_J = \lim_{n \to \infty} \mathcal{D}_J^{(n)}$ exists.

Similarly, we show that along an appropriate subsequence, $(\beta_J^{(n)})_{n \in \mathbb{N}}$ converges to a value $\beta_J \in \mathbb{R}$. Suppose this were not the case, then there were a subsequence along which $|\beta_J^{(n)}| \to \infty$. Again we assume for ease of notation that this were the case along the full sequence. Then, for every $x \in A_J$, one has that $|\mathcal{R}^{(n)}(x)| \to \infty$ and we argue as above to show that this would contradict the optimality of $(\mathcal{R}^{(n)})_{n \in \mathbb{N}}$. Consequently, we have on a compact set $B \subset A_J$ uniform convergence

$$\lim_{n \to \infty} \mathcal{R}^{(n)}(x) = \mathcal{D}_J \cdot x + \beta_J. \tag{2.3}$$

Since $\bigcup_{J \in \mathbb{J}} A_J$ has full μ-measure we get with the lower semicontinuity of $\mathcal{L}$ in the second argument and Fatou's lemma that for every measurable function $\mathcal{R} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$

satisfying for each $J \in \mathbb{J}$ and $x \in A_J$ that

$$\mathcal{R}(x) = \mathcal{D}_J \cdot x + \beta_J, \tag{2.4}$$

we have

$$\begin{aligned}\int \mathcal{L}\big(x, \mathcal{R}(x)\big)\,\mathrm{d}\mu(x) &\le \int \liminf_{n\to\infty} \mathcal{L}\big(x, \mathcal{R}^{(n)}(x)\big)\,\mathrm{d}\mu(x)\\ &\le \liminf_{n\to\infty} \int \mathcal{L}\big(x, \mathcal{R}^{(n)}(x)\big)\,\mathrm{d}\mu(x) = \overline{\mathrm{err}}_d^{\mathcal{L}}.\end{aligned}$$

We call a summand $j \in \{1, \dots, K\}$ degenerate if A_j or $\overline{A}_j^c$ has μ-measure zero. Now, let j be a non-degenerate summand. Since μ is nice there exists a ∂A_j-regular point x that is not in $\bigcup_{A\in\mathbb{A}:\partial A\neq\partial A_j} \partial A$, where $\mathbb{A} := \{A_i : i \text{ is non-degenerate}\}$. We let

$$J_-^x = \{i : x \in A_i\} \cup \big\{i : \overline{A}_i^c = A_j\big\}, \quad J_+^x = \{i : x \in A_i\} \cup \{i : A_i = A_j\}.$$

Since $x \in \mathrm{supp}(\mu|_{\overline{A}_j^c})$ we get that the cell $A_{J_-^x}$ has strictly positive μ-measure so that $J_-^x \in \mathbb{J}$. Analogously, $x \in \mathrm{supp}(\mu|_{A_j})$ entails that $J_+^x \in \mathbb{J}$. (Note that J_+^x and J_-^x are just the cells that lie on the opposite sides of the hyperplane ∂A_j at x.) We thus get that

$$\delta_{A_j}^{(n)} := \sum_{i:A_i=A_j} \delta_i^{(n)} - \sum_{i:\overline{A}_i^c=A_j} \delta_i^{(n)} = \mathcal{D}_{J_+^x}^{(n)} - \mathcal{D}_{J_-^x}^{(n)} \;\to\; \mathcal{D}_{J_+^x} - \mathcal{D}_{J_-^x} =: \delta_{A_j},$$

where the definitions of $\delta_{A_j}^{(n)}$ and δ_{A_j} do not depend on the choice of x. Analogously,

$$\mathfrak{b}_{A_j}^{(n)} := \sum_{i:A_i=A_j} \mathfrak{b}_i^{(n)} - \sum_{i:\overline{A}_i^c=A_j} \mathfrak{b}_i^{(n)} = \beta_{J_+^x}^{(n)} - \beta_{J_-^x}^{(n)} \;\to\; \beta_{J_+^x} - \beta_{J_-^x} =: \mathfrak{b}_{A_j}.$$

We form the set $\mathbb{A}_0$ by thinning $\mathbb{A}$ in such a way that for two active areas that share the same hyperplane as boundary only one is kept (meaning that for two active areas on opposite sides of a hyperplane only one is kept). Then there exists an affine function $\mathfrak{a} : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}$ such that for $x \in \bigcup_{J\in\mathbb{J}} A_J$,

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{A\in\mathbb{A}_0} \mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A). \tag{2.5}$$

We use the latter identity to define $\mathcal{R}$ on the whole space $\mathbb{R}^{d_{\mathrm{in}}}$.

Step 2. Analyzing the multiplicities of representation (2.5).

We assign the active areas in $A \in \mathbb{A}_0$ multiplicities. If

$$\partial A \subset \big\{x \in \mathbb{R}^{d_{\mathrm{in}}} : \delta_A \cdot x + \mathfrak{b}_A = 0\big\},$$

or, equivalently, $x \mapsto \mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A)$ is continuous, then we assign A the multiplicity one and otherwise two.

Next, we show that an active area $A \in \mathbb{A}_0$ whose breakline is only served by one summand k of multiplicity one is again assigned multiplicity one. For this it remains to show continuity of $x \mapsto \mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A)$ for such an A. Suppose that the k-th summand is the unique summand that contributes to A. Then $\delta_k^{(n)} = \delta_A^{(n)} \to \delta_A$ and $\mathfrak{b}_k^{(n)} = \mathfrak{b}_A^{(n)} \to \mathfrak{b}_A$. Moreover, one has

$$\big\{x \in \mathbb{R}^{d_{\text{in}}} : \mathfrak{n}_k^{(n)} \cdot x - o_k^{(n)} = 0\big\} \subset \big\{x \in \mathbb{R}^{d_{\text{in}}} : \delta_k^{(n)} \cdot x + \mathfrak{b}_k^{(n)} = 0\big\},$$

which entails that, in particular, $\delta_k^{(n)}$ is a multiple of $\mathfrak{n}_k^{(n)}$. Both latter vectors converge and $|\mathfrak{n}_k| = 1$ which also entails that the limit δ_A is a multiple of $\mathfrak{n}_k$. To show that

$$\partial A \subset \big\{x \in \mathbb{R}^{d_{\text{in}}} : \delta_A \cdot x + \mathfrak{b}_A = 0\big\},$$

it thus suffices to verify that one point of the hyperplane on the left-hand side lies also in the set on the right-hand side. Indeed, this is the case for $x = o_k\mathfrak{n}_k$ since for $n \to \infty$,

$$\delta_A \cdot (o_k\mathfrak{n}_k) + \mathfrak{b}_A = \lim_{n\to\infty} \delta_k^{(n)} \cdot \big(o_k^{(n)}\mathfrak{n}_k^{(n)}\big) + \mathfrak{b}_k^{(n)} = 0.$$

This entails contains continuity and we also showed that for such a k,

$$\mathbb{1}_{A_k^{(n)}}(x)\big(\delta_k^{(n)} \cdot x + \mathfrak{b}_k^{(n)}\big) \;\to\; \mathbb{1}_A(\delta_A \cdot x + \mathfrak{b}_A)$$

pointwise in x.

Note that the sum over all multiplicities assigned to the active areas $A \in \mathbb{A}_0$ is strictly smaller than d if

(1) for all $n \in \mathbb{N}$, $m_0^{(n)} = 1$,

(2) at least one of the summands $j \in \{1, \dots, K\}$ is degenerate, or

(3) there is an active area $A \in \mathbb{A}_0$ whose contributing terms have a cumulated multiplicity that is strictly larger than the one assigned to A.

In the latter cases we can choose $m_0 = 1$ and $\mathcal{R}$ as in (2.5) is of dimension at most d.

Step 3. Separate treatment of the cases where the $\mathcal{R}^{(n)}$ are of type (a), (b).

If $m_0^{(n)} \equiv 1$, we are done since property (1) above is satisfied.

Now suppose that the $\mathcal{R}^{(n)}$ are all of type (b) and let $\mathbb{I} \subset \{1, \dots, K\}$ denote the indices of the summands with multiplicity two. Then linear dependence of $(\mathfrak{n}_j^{(n)} : j \in \mathbb{I})$ implies linear dependence of the limits $(\mathfrak{n}_j : j \in \mathbb{I})$. If there is no degenerate asymptotic active area and the entries of $(\partial A_j : j \in \mathbb{I})$ are pairwise different, then the representation (2.5) satisfies property (b) and we verified that $\mathcal{R}$ is in $\mathfrak{R}_d$. On the other hand, in the case that the entries in $(\partial A_j : j \in \mathbb{I})$ are not pairwise different, then there is an active area $A \in \mathbb{A}_0$ whose contributing summands contribute multiplicity at least four and thus property (3) above holds and we are done.

It remains to consider the case (a). We can assume that there are no degenerate active areas and that every $A \in \mathbb{A}_0$ is served by terms of total multiplicity at most two since otherwise property (2) or (3) from above holds and we are done. Then every $A \in \mathbb{A}_0$ is served by

(i) a single summand,

(ii) two summands of multiplicity one that have the same asymptotic active area or

(iii) two summands of multiplicity one that have their asymptotic areas on opposite sides of the related hyperplane.

For an $A \in \mathbb{A}_0$ of type (i) or (ii) the asymptotic contribution of the related summands satisfies outside the hyperplane ∂A

$$\lim_{n\to\infty} \sum_{j:A_j=A} \mathbb{1}_{A_j^{(n)}}(x)\big(\delta_j^{(n)} \cdot x + \mathfrak{b}_j^{(n)}\big) = \mathbb{1}_A(x)(\delta_A \cdot x + \mathfrak{b}_A).$$

If there is a single summand of multiplicity one contributing the limit will be continuous for the same reason as above. Hence, if there exists no $A \in \mathbb{A}_0$ of type (iii) we use that $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ converges on $\bigcup_{J\in\mathbb{J}} A_J$ in order to deduce that $\lim_{n\to\infty} \mathfrak{b}^{(n)} =: \mathfrak{b}$ exists and obtain a representation (2.2) with $\mathfrak{a} \equiv \mathfrak{b}$ and are in case (a).

Now let $\mathbb{A}_0^*$ denote the subset of all $A \in \mathbb{A}_0$ that are of type (iii) and assume that $\mathbb{A}_0^* \neq \emptyset$. We say that $A \in \mathbb{A}_0^*$ is served by the pair of twins (i, j) if $A_i = A$ and $A_j = \overline{A}^c$. Moreover, we call a summand k to be of type (iii) if it contributes to one active area of type (iii). By switching to an appropriate subsequence we can ensure that there exists a summand k of type (iii) such that $|\delta_k^{(n)}|$ is maximal over all summands of type (iii) for all $n \in \mathbb{N}$.

We distinguish two cases. If there is a subsequence along which $(|\delta_k^{(n)}|)_{n\in\mathbb{N}}$ is uniformly bounded, then again by switching to appropriate subsequences we get that for every summand j of type (iii), $\lim_{n\to\infty} \delta_j^{(n)} =: \delta_j \in \mathbb{R}^{d_{\text{in}}}$ exists. Since $\mathbb{1}_{A_j^{(n)}(x)}(\delta_j^{(n)} \cdot x + \mathfrak{b}_j^{(n)})$ is continuous we get for a sequence $(x_n)_{n\in\mathbb{N}}$ that satisfies $x_n \in \partial A_j^{(n)}$ for all $n \in \mathbb{N}$, and that converges to an $x \in \partial A_j$ that

$$\mathfrak{b}_j^{(n)} = -\delta_j^{(n)} \cdot x_n \;\; \to \;\; -\delta_j \cdot x =: \mathfrak{b}_j,$$

where the left-hand side does not depend on the choice of $(x_n)_{n\in\mathbb{N}}$ or x. Moreover, the limit $\mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)$ is continuous. Thus, for a pair of twins (i, j) the contributing terms i and j to A_i have in total multiplicity two but the respective term

$$\mathbb{1}_{A_i}(x)(\delta_{A_i} \cdot x + \mathfrak{b}_{A_i}) = \mathbb{1}_{A_i}(x)\big((\delta_i - \delta_j) \cdot x + \mathfrak{b}_i - \mathfrak{b}_j\big)$$

is continuous and thus has multiplicity one. Hence, we are in case (3) above and are done.

It remains to consider the case where $(|\delta_k^{(n)}|)_{n\in\mathbb{N}}$ tends to infinity. For every twin (i, j) and $n \in \mathbb{N}$ we can choose $\alpha_i^{(n)} \in [-1, 1]$ and $\alpha_j^{(n)} \in [-1, 1]$ with

$$\frac{1}{\left|\delta_k^{(n)}\right|}\delta_i^{(n)} = \alpha_i^{(n)}\mathfrak{n}_i^{(n)}, \quad \frac{1}{\left|\delta_k^{(n)}\right|}\delta_j^{(n)} = -\alpha_j^{(n)}\mathfrak{n}_j^{(n)}$$

and along an appropriate subsequence we have convergence of all $(\alpha_i^{(n)})_{n\in\mathbb{N}}$, $(\alpha_j^{(n)})_{n\in\mathbb{N}}$ to limits $\alpha_i \in [-1,1]$ and $\alpha_j \in [-1,1]$. Since $(\delta_i^{(n)} - \delta_j^{(n)})_{n\in\mathbb{N}}$ converges to δ_{A_i}, $|\delta_k^{(n)}| \to \infty$ and

$$\lim_{n\to\infty} \mathfrak{n}_i^{(n)} = \mathfrak{n}_i = -\mathfrak{n}_j = \lim_{n\to\infty} -\mathfrak{n}_j^{(n)},$$

we get that $\alpha_i = \alpha_j$. Consequently, for $x \in \bigcup_{J\in\mathbb{J}} A_J$, one has

$$\lim_{n\to\infty} \frac{1}{|\delta_k^{(n)}|}\big(\mathcal{R}^{(n)}(x) - \mathfrak{b}^{(n)}\big) = \sum_{i:A_i\in\mathbb{A}_0^*} \alpha_i(\mathfrak{n}_i \cdot x - o_i).$$

If $\sum_{i:A_i\in\mathbb{A}_0^*} \alpha_i\mathfrak{n}_i$ is not equal to zero, then the linear term on the right-hand side does not vanish. This contradicts convergence of $\mathcal{R}^{(n)}$ on $\bigcup_{J\in\mathbb{J}} A_J$. Hence, we have $\sum_{i:A_i\in\mathbb{A}_0^*} \alpha_i\mathfrak{n}_i = 0$. Note that not all α_i's are equal to zero since $\alpha_k \in \{\pm 1\}$ and either k or its twin appears in the sum. Thus we showed that the normals belonging to the active areas in $\mathbb{A}_0^*$ are linearly dependent. Hence, we are in case (b) and the proof is achieved.

Step 4. Discussion of the minimization problem for strict responses at dimension d.

Now we choose a sequence of responses $(\mathcal{R}^{(n)})_{n\in\mathbb{N}}$ from $\mathfrak{R}_d^{\text{strict}}$ with

$$\lim_{n\to\infty} \int \mathcal{L}\big(x, \mathcal{R}^{(n)}(x)\big)\,\mathrm{d}\mu(x) = \inf_{\tilde{\mathcal{R}}\in\mathfrak{R}_d^{\text{strict}}} \int \mathcal{L}\big(x, \tilde{\mathcal{R}}(x)\big)\,\mathrm{d}\mu(x).$$

If infinitely many of the responses $\mathcal{R}^{(n)}$ are in $\mathfrak{R}_{d-1}$ then the response constructed above is a minimizer and in $\mathfrak{R}_{d-1}$. On the other hand, if all but finitely many responses are discontinuous, then the construction above yields a limit $\mathcal{R} \in \mathfrak{R}_d$ that minimizes the error. Note that there is at least one summand that contributes multiplicity two to one of the active areas of the limit $\mathcal{R}$. If $\mathcal{R}$ is continuous along the respective hyperplane, then it is of dimension at most $d-1$ and otherwise the response is discontinuous. In both cases we have $\mathcal{R} \in \mathfrak{R}_d^{\text{strict}}$. □

3 Strict generalized responses are not better than representable ones

In this section, we finish the proof of Theorem 1.2. Recall that

$$\operatorname{err}_d^{\mathcal{L}} := \inf_{\mathbb{W}\in\mathcal{W}_d} \int \mathcal{L}\big(x, \mathfrak{N}^{\mathbb{W}}(x)\big)\,\mathrm{d}\mu(x) \quad \text{and} \quad \overline{\operatorname{err}}_d^{\mathcal{L}} := \inf_{\mathcal{R}\in\mathfrak{R}_d} \int \mathcal{L}\big(x, \mathcal{R}(x)\big)\,\mathrm{d}\mu(x),$$

i.e. $\operatorname{err}_d^{\mathcal{L}}$ denotes the infimum of the error function taken over all neural networks with d neurons on the hidden layer and $\overline{\operatorname{err}}_d^{\mathcal{L}}$ denotes the infimum over all generalized responses

of at most dimension d. Note that the latter infimum is attained due to Proposition 2.1. We will show that, in the setting of Theorem 1.2, for every $d \in \{2, 3, \dots\}$ with $\operatorname{err}_d^{\mathcal{L}} < \operatorname{err}_{d-1}^{\mathcal{L}}$, the infimum taken over the strict generalized responses produces a larger error than the best representable response. This will entail Theorem 1.2. For a discussion on sufficient and necessary assumptions on the loss function $\mathcal{L}$ that implies $\operatorname{err}_d^{\mathcal{L}} < \operatorname{err}_{d-1}^{\mathcal{L}}$, see [6, Proposition 3.5, Example 3.7].

Proposition 3.1. *Suppose that the assumptions of Theorem* 1.2 *are satisfied. Let $d \in \mathbb{N}_0$. Then there exists an optimal network $\mathbb{W} \in \mathcal{W}_d$ with*

$$\operatorname{err}^{\mathcal{L}}(\mathbb{W}) = \operatorname{err}_d^{\mathcal{L}} = \overline{\operatorname{err}}_d^{\mathcal{L}}.$$

If, additionally, $d \geq 2$ and $\operatorname{err}_d^{\mathcal{L}} < \operatorname{err}_{d-1}^{\mathcal{L}}$, then one has that

$$\inf_{\mathcal{R} \in \mathfrak{R}_d^{\text{strict}}} \int \mathcal{L}\big(x, \mathcal{R}(x)\big) \, \mathrm{d}\mu(x) > \operatorname{err}_d^{\mathcal{L}}. \tag{3.1}$$

Proof. We can assume without loss of generality that $\mu \neq 0$. First we verify the assumptions of Proposition 2.1 in order to conclude that there are generalized responses $\mathcal{R} \in \mathfrak{R}_d$ with

$$\int \mathcal{L}\big(x, \mathcal{R}(x)\big) \, \mathrm{d}\mu(x) = \overline{\operatorname{err}}_d^{\mathcal{L}}.$$

We verify that μ is a nice measure: In fact, since μ has Lebesgue density h, we have $\mu(H)=0$ for all hyperplanes $H \subset \mathbb{R}^{d_{\text{in}}}$. Moreover, for every half-space A with $\mu(A), \mu(\overline{A}^c) > 0$ we have that ∂A intersects the interior of the convex hull of $\mathbb{D}$ so that there exists a point $x \in \partial A$ with $h(x) > 0$. Since $\{x \in \mathbb{R}^{d_{\text{in}}} : h(x) > 0\}$ is an open set, $\{x \in \partial A : h(x) > 0\}$ cannot be covered by finitely many hyperplanes different from ∂A. Moreover, since for all $x \in \mathbb{R}^{d_{\text{in}}}$ the function $y \mapsto \mathcal{L}(x, y)$ is strictly convex and attains its minimum we clearly have for fixed $x \in \mathbb{R}^{d_{\text{in}}}$ continuity of $y \mapsto \mathcal{L}(x, y)$ and

$$\lim_{|y| \to \infty} \mathcal{L}(x, y) = \infty.$$

We prove the statement via induction over the dimension d. If $d \leq 1$, all generalized responses of dimension d are, on the compact set $\mathbb{D}$, representable by a neural network and we are done. Now let $d \geq 2$ and suppose that $\mathcal{R}$ is a strict generalized response at dimension d that satisfies

$$\int \mathcal{L}\big(x, \mathcal{R}(x)\big) \, \mathrm{d}\mu(x) = \inf_{\tilde{\mathcal{R}} \in \mathfrak{R}_d^{\text{strict}}} \int \mathcal{L}\big(x, \tilde{\mathcal{R}}(x)\big) \, \mathrm{d}\mu(x).$$

It suffices to show that one of the following two cases enters: One has

$$\int \mathcal{L}\big(x, \mathcal{R}(x)\big) \, \mathrm{d}\mu(x) \geq \overline{\operatorname{err}}_{d-1}^{\mathcal{L}} \tag{3.2}$$

or

$$\int \mathcal{L}\big(x, \mathcal{R}(x)\big) \, \mathrm{d}\mu(x) > \overline{\operatorname{err}}_d^{\mathcal{L}}. \tag{3.3}$$

Indeed, then in the case that (3.3) does not hold, we have as consequence of (3.2) that

$$\overline{\mathrm{err}}_{d-1}^{\mathcal{L}} \leq \int \mathcal{L}\big(x, \mathcal{R}(x)\big)\, \mathrm{d}\mu(x) = \overline{\mathrm{err}}_{d}^{\mathcal{L}},$$

and the induction hypothesis entails that $\mathrm{err}_{d-1}^{\mathcal{L}} = \overline{\mathrm{err}}_{d-1}^{\mathcal{L}} \leq \overline{\mathrm{err}}_{d}^{\mathcal{L}} \leq \mathrm{err}_{d}^{\mathcal{L}} \leq \mathrm{err}_{d-1}^{\mathcal{L}}$ so that $\mathrm{err}_{d}^{\mathcal{L}} = \overline{\mathrm{err}}_{d}^{\mathcal{L}}$ and $\mathrm{err}_{d}^{\mathcal{L}} = \mathrm{err}_{d-1}^{\mathcal{L}}$. Thus, an optimal representable response $\mathcal{R}$ of dimension at most $d-1$ (which exists by induction hypothesis) is also optimal when taking the minimum over all generalized responses of dimension d or smaller. Conversely, if (3.3) holds, an optimal generalized response (which exists by Proposition 2.1) is representable so that, in particular, $\mathrm{err}_{d}^{\mathcal{L}} = \overline{\mathrm{err}}_{d}^{\mathcal{L}}$. This shows that there always exists an optimal representable response. Moreover, it also follows that in the case where $\mathrm{err}_{d}^{\mathcal{L}} < \mathrm{err}_{d-1}^{\mathcal{L}}$, either of the properties (3.2) and (3.3) entail property (3.1).

Suppose that the optimal strict generalized response $\mathcal{R}$ at dimension d is given in the standard representation (2.2)

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{j=1}^{K} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)$$

with $A_1, \ldots, A_K$ being the half-spaces with pairwise distinct boundaries and $m_0, m_1, \ldots, m_K$ being the respective multiplicities. Suppose that $\mathcal{R}$ is an optimal response with the minimal number of terms of multiplicity two.

First note that for every $k = 1, \ldots, K$ for which ∂A_k does not intersect the interior of the convex hull of $\mathbb{D}$, A_k has either zero or full μ-measure. In both cases we can remove the k-th summand, set $m_0 = 1$ and adapt the affine function $\mathfrak{a}$ appropriately to get to a response that agrees μ-almost everywhere with the former response and is again of dimension at most d. Thus we can without loss of generality assume that for every $k = 1, \ldots, K$, ∂A_k intersects the interior of the convex hull of $\mathbb{D}$.

We distinguish cases. If one has $m_0 = 1$ or $\mathfrak{a} \equiv \mathfrak{b} \in \mathbb{R}$ (case (a)), then the proof of [6, Proposition 3.3] shows that an appropriate replacement of a summand of multiplicity two by two summands of multiplicity one reduces the error which shows (3.3).

It remains to treat the case (b). Let $\mathbb{I}$ be the set of indices with multiplicity two. Now we have that the vectors $(\mathfrak{n}_j : j \in \mathbb{I})$ are linear dependent. If the set is not minimal in the sense that we can remove one of the vectors and still have a linearly dependent set, then we can argue as above and apply an appropriate replacement of this particular summand by two summands of multiplicity one that still satisfies (b) and has strictly smaller error. Hence, we can assume without loss of generality that the set $(\mathfrak{n}_j : j \in \mathbb{I})$ is minimal in the sense that for a nontrivial linear combination

$$\sum_{j \in \mathbb{I}} \alpha_j \mathfrak{n}_j = 0, \tag{3.4}$$

one has that $\alpha_j \neq 0$ for all $j \in \mathbb{I}$. For $x \in \mathbb{D}$ and $y \in \mathbb{R}$ let

$$\tilde{\mathcal{L}}(x, y) = \mathcal{L}\bigg(x, y + \sum_{j \in \mathbb{I}^c} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)\bigg)$$

and note that due to continuity of $x \mapsto \sum_{j\in\mathbb{I}^c} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)$ the function $\tilde{\mathcal{L}}$ satisfies the same assumptions as imposed on $\mathcal{L}$ in the proposition. If we can find a representable response $\hat{\mathcal{R}}$ of dimension $2\#\mathbb{I}$ with

$$\int \tilde{\mathcal{L}}\big(x, \hat{\mathcal{R}}(x)\big)\,\mathrm{d}\mu(x) < \int \tilde{\mathcal{L}}\big(x, \tilde{\mathcal{R}}(x)\big)\,\mathrm{d}\mu(x),$$

where

$$\tilde{\mathcal{R}}(x) = \mathfrak{a}(x) + \sum_{j\in\mathbb{I}} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j),$$

then

$$\int \mathcal{L}\bigg(x, \hat{\mathcal{R}}(x) + \sum_{j\in\mathbb{I}^c} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j)\bigg)\,\mathrm{d}\mu(x) < \int \mathcal{L}\big(x, \mathcal{R}(x)\big)\,\mathrm{d}\mu(x)$$

and it follows validity of (3.3).

Therefore, we can assume without loss of generality that $\mathbb{I} = \{1, \dots, K\}$ and the optimal strict response is given by

$$\mathcal{R}(x) = \mathfrak{a}(x) + \sum_{j=1}^{K} \mathbb{1}_{A_j}(x)(\delta_j \cdot x + \mathfrak{b}_j).$$

We fix a non-trivial vector $(\alpha_j)_{j\in\mathbb{I}}$ satisfying (3.4) and choose $(\delta_j^+, \delta_j^-) \in \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{in}}}$, $(\mathfrak{b}_j^+, \mathfrak{b}_j^-) \in \mathbb{R}^2$ and $\mathfrak{b} \in \mathbb{R}$ with $\delta_j = \delta_j^+ - \delta_j^-$ and $\mathfrak{b}_j = \mathfrak{b}_j^+ - \mathfrak{b}_j^-$ such that for all $x \in \mathbb{R}^{d_{\text{in}}}$,

$$\mathcal{R}(x) = \mathfrak{b} + \sum_{j=1}^{K} \mathbb{1}_{A_j}(x)\big(\delta_j^+ \cdot x + \mathfrak{b}_j^+\big) + \mathbb{1}_{A_j^c}(x)\big(\delta_j^- \cdot x + \mathfrak{b}_j^-\big).$$

By switching the sides of the active areas we can assume without loss of generality that all $\alpha_j > 0$ (indeed this will change the definition of $\mathcal{R}$ only on the respective hyperplanes). We will replace $\mathcal{N}_j(x) := \mathbb{1}_{A_j}(x)(\delta_j^+ \cdot x + \mathfrak{b}_j^+) + \mathbb{1}_{A_j^c}(x)(\delta_j^- \cdot x + \mathfrak{b}_j^-)$ by

$$\mathcal{N}_j^\kappa(x) := \big(\delta_j^+ \cdot x + \mathfrak{b}_j^+ + \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j)\big)^+ - \big(-\delta_j^- \cdot x - \mathfrak{b}_j^- - \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j)\big)^+,$$

where $\kappa > 0$. Let

$$Q_j^\kappa := \big\{x \in \mathbb{R}^{d_{\text{in}}} : \mathcal{N}_j^\kappa(x) \neq \mathcal{N}_j(x) + \kappa\alpha_j(\mathfrak{n}_j \cdot x - o_j)\big\},$$

and compare $\mathcal{R}$ with

$$\mathcal{R}^\kappa(x) := \mathfrak{b}^\kappa + \sum_{j=1}^{K} \mathcal{N}_j^\kappa(x),$$

where $\mathfrak{b}^\kappa := \mathfrak{b} + \sum_{j=1}^{K} \kappa\alpha_j o_j$. Since $\sum_{j=1}^{K} \alpha_j \mathfrak{n}_j = 0$, we have $\mathcal{R}^\kappa = \mathcal{R}$ on $\mathbb{R}^{d_{\text{in}}} \setminus (\bigcap_{j=1,\dots,K} Q_j^\kappa)$. Furthermore, the set $\{x \in \mathbb{R}^{d_{\text{in}}} : x \text{ is in two } Q_j^\kappa\} \cap \mathbb{D}$ is of size $\mathcal{O}(\kappa^{-2})$. Hence,

$$\int \mathcal{L}\big(x, \mathcal{R}^\kappa(x)\big) - \mathcal{L}\big(x, \mathcal{R}(x)\big)\,\mathrm{d}\mu(x)$$

$$= \sum_{j=1}^{K} \int_{Q_j^\kappa} h(x)\big(\mathcal{L}\big(x, \mathcal{R}^\kappa(x)\big) - \mathcal{L}\big(x, \mathcal{R}(x)\big)\big)\,\mathrm{d}x + \mathcal{O}(\kappa^{-2}). \tag{3.5}$$

Moreover, using the uniform continuity of h on the compact set $\mathbb{D}$ and the uniform boundedness of $|\mathcal{L}(x, \mathcal{R}^\kappa(x)) - \mathcal{L}(x, \mathcal{R}(x))|$ over all $x \in \mathbb{D}$ and $\kappa \geq 1$ we conclude that as $\kappa \to \infty$,

$$\begin{aligned}
&\int_{Q_j^\kappa} h(x)\big(\mathcal{L}\big(x, \mathcal{R}^\kappa(x)\big) - \mathcal{L}\big(x, \mathcal{R}(x)\big)\big)\,\mathrm{d}x \\
&= \int_{H_j} \int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^\kappa} h(z)\big(\mathcal{L}\big(z, \mathcal{R}^\kappa(z)\big) - \mathcal{L}\big(z, \mathcal{R}(z)\big)\big)\,\mathrm{d}z\,\mathrm{d}x' \\
&= \int_{H_j} h(x') \int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^\kappa} \big(\mathcal{L}\big(x', \mathcal{R}^\kappa(z)\big) - \mathcal{L}\big(x', \mathcal{R}(z)\big)\big)\,\mathrm{d}z\,\mathrm{d}x' + o(\kappa^{-1}),
\end{aligned} \tag{3.6}$$

where $H_j = \partial A_j$. Now note that for a fixed $x' \in H_j$ for which $(x' + \mathfrak{n}_j\mathbb{R}) \cap Q_j^\kappa$ does not intersect one of the Q_i^κ with $i \neq j$, one has

$$\begin{aligned}
\int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^\kappa} \mathcal{L}\big(x', \mathcal{R}(z)\big)\,\mathrm{d}z = {}& \big|\big(x' + \mathbb{R}\mathfrak{n}_j\big) \cap Q_j^\kappa \cap A_j\big|\,\big(L_j^+(x') + o(1)\big) \\
&+ \big|\big(x' + \mathbb{R}\mathfrak{n}_j\big) \cap Q_j^\kappa \cap A_j^c\big|\big(L_j^-(x') + o(1)\big),
\end{aligned}$$

where $|\cdot|$ denotes the one-dimensional Hausdorff measure (i.e. the length of the segment),

$$\begin{aligned}
L_j^+(x') &= \mathcal{L}\big(x', \delta_j^+ \cdot x' + \mathfrak{b}_j^+ + \hat{\mathcal{R}}_j(x')\big), \\
L_j^-(x') &= \mathcal{L}\big(x', \delta_j^- \cdot x' + \mathfrak{b}_j^- + \hat{\mathcal{R}}_j(x')\big)
\end{aligned}$$

with

$$\hat{\mathcal{R}}_j(x') = \mathfrak{b} + \sum_{i\neq j}\big(\mathbb{1}_{A_i}(x')\big(\delta_i^+ \cdot x' + \mathfrak{b}_i^+\big) + \mathbb{1}_{A_i^c}(x')\big(\delta_i^- \cdot x' + \mathfrak{b}_i^-\big)\big).$$

Moreover, for the same x'

$$\int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^\kappa} \mathcal{L}\big(x', \mathcal{R}^\kappa(z)\big)\,\mathrm{d}z = \big|\big(x' + \mathfrak{n}_j\mathbb{R}\big) \cap Q_j^\kappa\big|\big(\overline{L}_j(x') + o(1)\big),$$

where $\overline{L}_j(x')$ is the average of $\mathcal{L}(x', \cdot)$ on the segment $[\delta_j^- \cdot x' + \mathfrak{b}_j^- + \hat{\mathcal{R}}_j(x'), \delta_j^+ \cdot x' + \mathfrak{b}_j^+ + \hat{\mathcal{R}}_j(x')]$.

We calculate the Hausdorff measure of the segments $(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j$ and $(x' + \mathbb{R}\mathfrak{n}_j) \cap Q_j^\kappa \cap A_j^c$. We note that for $t \in \mathbb{R}$, $x' + t\mathfrak{n}_j$ lies in Q_j^κ if t lies between the solutions $t_{+/-}^\kappa$ of

$$\delta_j^{+/-} \cdot \big(x' + t_{+/-}^\kappa \mathfrak{n}_j\big) + \mathfrak{b}_j^{+/-} + \kappa\alpha_j t_{+/-}^\kappa = 0,$$

so that

$$\big|\big(x' + \mathbb{R}\mathfrak{n}_j\big) \cap Q_j^\kappa \cap A_j\big| = \big|[t_-^\kappa, t_+^\kappa] \cap [0, \infty)\big|.$$

Since

$$\lim_{\kappa\to\infty} \kappa\, t^{\kappa}_{+/-} = -\frac{1}{\alpha_j}\big(\delta_j^{+/-}\cdot x' + \mathfrak{b}_j^{+/-}\big) =: t_{+/-}(x'),$$

we get that

$$q_j^+(x') := \lim_{\kappa\to\infty} \kappa\big|\big(x'+\mathbb{R}\mathfrak{n}_j\big)\cap Q_j^{\kappa}\cap A_j\big| = \big|[t_-(x'),t_+(x')]\cap[0,\infty)\big|.$$

Analogously, it follows that

$$q_j^-(x') := \lim_{\kappa\to\infty} \kappa\big|\big(x'+\mathbb{R}\mathfrak{n}_j\big)\cap Q_j^{\kappa}\cap A_j^c\big| = \big|[t_-(x'),t_+(x')]\cap(-\infty,0]\big|.$$

Combining the estimates gives that for every $x'\in H_j\backslash\bigcup_{i\neq j}H_i$ one has

$$\begin{aligned}&\lim_{\kappa\to\infty}\kappa\int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^{\kappa}}\big(\mathcal{L}\big(x',\mathcal{R}^{\kappa}(z)\big)-\mathcal{L}\big(x',\mathcal{R}(z)\big)\big)\,\mathrm{d}z\\ &=\big(q_j^+(x')+q_j^-(x')\big)\overline{L}_j(x')-\big(q_j^-(x')L_j^+(x')+q_j^+(x')L_j^-(x')\big)\end{aligned}$$

By (3.5), (3.6) and dominated convergence, we get that

$$\begin{aligned}&\lim_{\kappa\to\infty}\kappa\int\big(\mathcal{L}\big(x,\mathcal{R}^{\kappa}(x)\big)-\mathcal{L}\big(x,\mathcal{R}(x)\big)\big)\,\mathrm{d}\mu(x)\\ &=\sum_{j=1}^{K}\int_{H_j}h(x')\big(\big(q_j^+(x')+q_j^-(x')\big)\overline{L}_j(x')-\big(q_j^+(x')L_j^+(x')+q_j^-(x')L_j^-(x')\big)\big)\,\mathrm{d}x',\end{aligned}$$

where we used that $\kappa\int_{(x'+\mathbb{R}\mathfrak{n}_j)\cap Q_j^{\kappa}}(\mathcal{L}(x',\mathcal{R}^{\kappa}(z))-\mathcal{L}(x',\mathcal{R}(z)))\mathrm{d}z$ is uniformly bounded over all $j=1,\dots,K, x'\in H_j\cap\mathbb{D}$ and $\kappa\geq 1$.

Now consider $\mathcal{R}^{-\kappa}$ given by

$$\mathcal{R}^{-\kappa}(x)=\mathfrak{b}^{-\kappa}+\sum_{j=1}^{K}-\big(-\delta_j^+\cdot x-\mathfrak{b}_j^++\kappa\alpha_j(\mathfrak{n}_j\cdot x-o_j)\big)^+ +\big(\delta_j^-\cdot x+\mathfrak{b}_j^--\kappa\alpha_j(\mathfrak{n}_j\cdot x-o_j)\big)^+,$$

where $\mathfrak{b}^{-\kappa}:=\mathfrak{b}-\sum_{j=1}^{K}\kappa\alpha_j o_j$. Following the same arguments as above we get that

$$\begin{aligned}&\lim_{\kappa\to\infty}\kappa\int\mathcal{L}\big(x,\mathcal{R}^{-\kappa}(x)\big)-\mathcal{L}\big(x,\mathcal{R}(x)\big)\,\mathrm{d}\mu(x)\\ &=\sum_{j=1}^{K}\int_{H_j}h(x')\big(\big(q_j^+(x')+q_j^-(x')\big)\overline{L}_j(x')-\big(q_j^-(x')\,L_j^+(x')+q_j^+(x')\,L_j^-(x')\big)\big)\,\mathrm{d}x'.\end{aligned}$$

Adding the estimates we get with $q_j(x')=q_j^+(x')+q_j^-(x')$ that

$$\begin{aligned}&\lim_{\kappa\to\infty}\kappa\int\big(\mathcal{L}\big(x,\mathcal{R}^{-\kappa}(x)\big)+\mathcal{L}\big(x,\mathcal{R}^{\kappa}(x)\big)-2\mathcal{L}\big(x,\mathcal{R}(x)\big)\big)\,\mathrm{d}\mu(x)\\ &=\sum_{j=1}^{K}\int_{H_j}h(x')q_j(x')\big(2\overline{L}_j(x')-\big(L_j^+(x')+L_j^-(x')\big)\big)\,\mathrm{d}x'.\end{aligned}\tag{3.7}$$

By strict convexity of $\mathcal{L}(x', \cdot)$, one has $2\overline{L}_j(x') \leq L_j^+(x') + L_j^-(x')$ with strict inequality whenever $\delta_j^- \cdot x' + \mathfrak{b}_j^- \neq \delta_j^+ \cdot x' + \mathfrak{b}_j^+$. Since $\partial A_1 \not\subset \{x \in \mathbb{R}^{d_{\text{in}}} : \delta_1 \cdot x + \mathfrak{b}_1 = 0\}$ we have that the set H_1' consisting of all $x' \in H_1$ such that $h(x') > 0$, $q_1(x') > 0$ and $\delta_j^- \cdot x' + \mathfrak{b}_j^- \neq \delta_j^+ \cdot x' + \mathfrak{b}_j^+$ has strictly positive $(d_{\text{in}} - 1)$-dimensional Hausdorff measure. Consequently, the limit in (3.7) is strictly negative and there exists $\kappa > 0$ for which either $\mathcal{R}^\kappa$ or $\mathcal{R}^{-\kappa}$ is a better response than $\mathcal{R}$ which contradicts optimality of $\mathcal{R}$. □

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044–390685587, Mathematics Münster: Dynamics–Geometry–Structure. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1283/2 2021 – 317210226.

References

[1] P. A. Absil, R. Mahony, and B. Andrews, Convergence of the iterates of descent methods for analytic cost functions, *SIAM J. Optim.*, 16(2):531–547, 2005.

[2] H. Attouch and J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Math. Program.*, 116(1):5–16, 2009.

[3] J. Bolte, A. Daniilidis, and A. Lewis, The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM J. Optim.*, 17(4):1205–1223, 2007.

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

[5] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, Stochastic subgradient method converges on tame functions, *Found. Comput. Math.*, 20(1):119–154, 2020.

[6] S. Dereich, A. Jentzen, and S. Kassing, On the existence of minimizers in shallow residual ReLU neural network optimization landscapes, *arXiv:2302.14690*, 2023.

[7] S. Dereich and S. Kassing, Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes, *arXiv:2102.09385*, 2021.

[8] S. Dereich and S. Kassing, Cooling down stochastic differential equations: Almost sure convergence, *Stochastic Process. Appl.*, 152:289–311, 2022.

[9] S. Dereich and S. Kassing, Central limit theorems for stochastic gradient descent with averaging for stable manifolds, *Electron. J. Probab.*, 28:1–48, 2023.

[10] W. E, C. Ma, L. Wu, and S. Wojtowytsch, Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't, *CSIAM Trans. Appl. Math.*, 1(4):561–615, 2020.

[11] S. Eberle, A. Jentzen, A. Riekert, and G. S. Weiss, Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation, *Electron. Res. Arch.*, 31(5):2519–2554, 2023.

[12] S. Foucart, *Mathematical Pictures at a Data Science Exhibition*, Cambridge University Press, 2022.

[13] D. Gallon, A. Jentzen, and F. Lindner, Blow up phenomena for gradient descent optimization methods in the training of artificial neural networks, *arXiv:2211.15641*, 2022.

[14] F. Girosi and T. Poggio, Networks and the best approximation property, *Biol. Cybern.*, 63(3):169–176, 1990.

[15] A. Jentzen and A. Riekert, On the existence of global minima and convergence analyses for gradient descent methods in the training of deep neural networks, *J. Mach. Learn.*, 1(2):141–246, 2022.

[16] P. C. Kainen, V. Kurková, and A. Vogt, Best approximation by linear combinations of characteristic functions of half-spaces, *J. Approx. Theory*, 122(2):151–159, 2003.
[17] M. R. Karimi, Y.-P. Hsieh, P. Mertikopoulos, and A. Krause, Riemannian stochastic approximation algorithms, *arXiv:2206.06795*, 2022.
[18] Q.-T. Le, E. Riccietti, and R. Gribonval, Does a sparse ReLU network training problem always admit an optimum? *arXiv:2306.02666*, 2023.
[19] L.-H. Lim, M. Michałek, and Y. Qi, Best k-layer neural network approximations, *Constr. Approx.*, 55(1):583–604, 2022.
[20] S. Łojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, *Les équations aux dérivées partielles*, 117:87–89, 1963.
[21] S. Łojasiewicz, *Ensembles Semi-Analytiques*, in: *Lectures Notes IHES*, 1965.
[22] S. Łojasiewicz, Sur les trajectoires du gradient d'une fonction analytique, *Seminari di geometria*, Universita' degli Studi di Bologna, 115–117, 1984.
[23] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher, On the almost sure convergence of stochastic gradient descent in non-convex problems, *Adv. Neural Inf. Process. Syst.*, 33:1117–1128, 2020.
[24] P. Petersen, M. Raslan, and F. Voigtlaender, Topological properties of the set of functions generated by neural networks of fixed size, *Found. Comput. Math.*, 21(2):375–444, 2021.
[25] I. Safran and O. Shamir, Spurious local minima are common in two-layer ReLU neural networks, in: *International Conference on Machine Learning*, 4433–4441, PMLR, 2018.
[26] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer, 1970.
[27] G. Swirszcz, W. M. Czarnecki, and R. Pascanu, Local minima in training of neural networks, *arXiv:1611.06310*, 2016.
[28] V. B. Tadić, Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema, *Stochastic Process. Appl.*, 125(5):1715–1755, 2015.
[29] L. Venturi, A. S. Bandeira, and J. Bruna, Spurious valleys in one-hidden-layer neural network optimization landscapes, *J. Mach. Learn. Res.*, 20(133):1–34, 2019.