Topological properties of the set of functions generated by neural networks of fixed size

Philipp Petersen* Mones Raslan* Felix Voigtlaender*

June 25, 2018

Abstract

In this paper, we analyze the topological properties of the set of functions that can be implemented by neural networks of a fixed size. Surprisingly, this set has many undesirable properties: It is highly non-convex, except possibly for a few exotic activation functions. Moreover, the set is not closed with respect to $L^p$-norms, $0 < p < \infty$, for all practically-used activation functions, and also not closed with respect to the $L^\infty$-norm for all practically-used activation functions except for the ReLU and the parametric ReLU. Finally, the function that maps a family of weights to an associated network is not inverse stable, for every practically used activation function. In other words, if $f_1, f_2$ are two functions realized by neural networks that are very close in the sense that $\|f_1 - f_2\|_{L^\infty} \leq \varepsilon$, it is usually not possible to find weights $w_1, w_2$ close together such that each $f_i$ is realized by a neural network with weights $w_i$. These observations identify a couple of potential causes for problems in the optimization of neural networks such as no guaranteed convergence, explosion of parameters, and very slow convergence.

1 Introduction

Neural networks, introduced in 1943 by McCulloch and Pitts [34], are the basis of every modern machine learning algorithm based on deep learning [18, 28, 43]. The term deep learning describes a variety of methods that are based on the data-driven manipulation of the weights of a neural network. Since these methods work so well in practice, they have become the state-of-the-art technology for a host of applications including image classification [24, 45, 26], speech recognition [22, 13, 49], game intelligence [44, 46, 50], and many more.

This success of deep learning caused many scientists to pick up research in the area of neural networks after the field had gone dormant for decades. In particular, quite a few mathematicians have recently investigated the properties of different neural network architectures, hoping that this can explain the effectiveness of deep learning techniques.

Many results in the area are based on approximation theory where one analyzes the expressiveness of deep neural network architectures. The universal approximation theorem [12, 23, 29] demonstrates that neural networks can approximate any continuous function as long as one uses ever more complex networks for the approximation. Considering more specific function classes than all continuous functions, one can often quantify more precisely how large the networks have to be to achieve a given approximation accuracy for functions from the restricted class. Examples of such results are [4, 35, 36, 51, 6, 39]. Some papers [37, 10, 52, 39] study in particular in which sense deep networks have a larger expressiveness than their shallow counterparts, thereby partially explaining the efficiency of deep networks in deep learning.

Even though all mentioned results offer some insight into the expressiveness of certain neural network architectures, the practical relevance of these approximation theoretical results is limited. Indeed, all approaches mentioned above that yield quantitative error estimates reduce the problem of approximation by neural networks to a classical approximation problem using polynomial-, local Taylor-, wavelet-, or spline-approximation. Because of that, the functions to be approximated are assumed to belong to classical spaces

* All three authors contributed equally to this work.
such as the spaces of smooth functions or Sobolev functions. For applications such as image classification, however, it is unclear if these function classes are a suitable model for the regularity of the true classifier functions. Moreover, approximation theoretical results usually only offer asymptotic estimates, which have limited meaning in applications where properties of a fixed network architecture of finite size need to be understood.

Apart from the papers focusing on approximation theory or the expressiveness of neural networks, several authors have studied other aspects of neural network architectures, such as the invariance properties of functions implemented by deep convolutional neural networks [7, 48], or the effect of the network architecture on the optimization procedure [17, 29]. However, to the best of our knowledge, the (topological) structure of the class of all neural networks of a fixed size has not been studied at all.

With this paper, we aim to close this gap of understanding. Contrary to approximation theoretical results, we focus not on the expressive capacity of the class of neural networks but on its structure. The precise mathematical results will be explained in the next subsections. On a conceptual level, our results have two implications: First, the derived topological properties pinpoint the reason for several issues that are observed when optimizing neural networks, namely: no guaranteed convergence, very slow convergence, or diverging network weights. These issues are highly undesirable in practice. We hope that any knowledge of a potential cause of these problems can be helpful for a practitioner. Second, our results show that the structure of the set of functions implemented by neural networks differs considerably from that of classical spaces used for function approximation, such as spaces of polynomials, splines, or wavelets [33, 14]. These structural differences suggest that to explain the efficiency of neural networks more accurately, a different paradigm than the usual reduction to classical approximation problems will be necessary.

In order to state our results precisely, we first introduce a precise distinction between a neural network as a set of weights and the associated function, referred to as its realization.

Let us fix a number of layers \( L \in \mathbb{N} \) and an input dimension \( d = N_0 \in \mathbb{N} \). For \( N_1, \ldots, N_L \in \mathbb{N} \), we say that a family \( \Phi = (W_\ell)_{\ell=1}^L \) of affine linear maps \( W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell} \) is a neural network. We call \( S := (d, N_1, \ldots, N_L) \) the architecture of \( \Phi \); furthermore \( N(S) := \sum_{\ell=0}^L N_\ell \) is called the number of neurons of \( S \) and \( L = L(S) \) is the number of layers of \( S \). In this introduction we will always assume that the output dimension \( N_L \) of the networks is equal to one.

Defining the realization of such a network \( \Phi = (W_\ell)_{\ell=1}^L \) requires two additional ingredients: a so-called activation function \( \varrho : \mathbb{R} \to \mathbb{R} \), and a domain of definition \( \Omega \subset \mathbb{R}^d \). Given these, the realization of the network \( \Phi = (W_\ell)_{\ell=1}^L \) is the function

\[
R^\Omega_{\varrho}(\Phi) : \Omega \to \mathbb{R}, x \mapsto W_L(\varrho(W_{L-1}(\varrho(\ldots \varrho(W_1(x) \ldots)))]) \tag{1.1},
\]

where \( \varrho \) is applied component-wise. In the remainder of the introduction, we will always assume \( \Omega \subset \mathbb{R}^d \) to be bounded and measurable.

In what follows, we study topological properties of sets of realizations of neural networks with a fixed size. Naturally, there are multiple conventions to specify the size of a network. We will study the following two: First, we denote by \( \mathcal{RNN}_\varrho(S) \) the set of realizations of networks with a given architecture \( S \) and activation function \( \varrho \). In the context of machine learning, this point of view is natural, since one usually prescribes the network architecture, and during training only adapts the weights of the network. The second point of view, which is more common in approximation theory, is to prescribe only the total number of neurons and the number of layers. This leads to the set of realizations of networks with \( N \) neurons, \( L \) layers and activation function \( \varrho \), which is given by

\[
\mathcal{RNN}_\varrho(N, L) = \bigcup \{ \mathcal{RNN}_\varrho(S) : \text{S architecture with } N(S) = N, \text{ and } L(S) = L \}
\]

for \( N, L \in \mathbb{N} \).

The definition of networks and realizations from above is precise enough so that we can state our results, while omitting non-essential technicalities. For proofs and calculations in the main part of the paper, however, the more technical Definition 2.1 will be used.
In the remainder of this introduction, we discuss our results concerning the topological properties of the two sets of realizations of neural networks that we just introduced. Finally, we give an overview over the structure of the paper, and fix the notations and conventions that will be used in the remainder of the paper.

1.1 Shape of the set of realizations

We will show that for a given architecture $S$, there is some $N_\ast \in \mathbb{N}$ such that the set $\mathcal{RNN}_\varrho(S)$ of neural network realizations with architecture $S$ and activation function $\varrho$ is star-shaped, that is, there exists a center $f \in \mathcal{RNN}_\varrho(S)$, which means that for all $g \in \mathcal{RNN}_\varrho(S)$, also

$$\{\lambda f + (1 - \lambda)g : \lambda \in [0, 1]\} \subset \mathcal{RNN}_\varrho(S).$$

Moreover, we show that $\mathcal{RNN}_\varrho(S)$ does not have more than $N_\ast$ linearly independent centers, except when the activation function is of a special form, which does not include any of the activation functions that are commonly used in practice.

In particular, for every commonly-used activation function $\varrho$, the set $\mathcal{RNN}_\varrho(S)$ is not convex. Even more, this set is highly non-convex in the sense that for every $r \in \mathbb{R}^+$, the set of functions having uniform distance less than $r$ to any function in $\mathcal{RNN}_\varrho(S)$ is not convex. The same results hold for the set $\mathcal{RNN}_\varrho(N, L)$ of realizations of networks with $N$ neurons, $L$ layers and activation function $\varrho$.

This nonconvexity is undesirable, since in classical statistical learning theory [11], the hypothesis space is often assumed to be convex, and because for non-convex hypothesis spaces, the learning problem is significantly harder; see [11] Chapter 7. Furthermore, in applications where the realization of a network is the quantity of interest—for example when a network is used as an Ansatz for the solution of a PDE, as in [27, 15]—our results show that the solution space is non-convex. This is undesirable if one aims for a convergence proof of the underlying optimization algorithm.

As a further result, we show for all commonly used activation functions $\varrho$ that the sets $\mathcal{RNN}_\varrho(S)$ and $\mathcal{RMN}_\varrho(N, L)$ have empty interior in $L^p(K)$ and in $C(K)$, and we examine conditions under which these sets are even nowhere dense in $L^p(K)$ and in $C(K)$.

1.2 (Non-)closedness of the set of realizations

For any fixed architecture $S$ and any fixed number of neurons $N \in \mathbb{N}$ and layers $L \in \mathbb{N}$, we show that neither $\mathcal{RNN}_\varrho(S)$ nor $\mathcal{RMN}_\varrho(N, L)$ are closed subsets of $L^p(\Omega)$ for $0 < p < \infty$, under very general assumptions on the activation function $\varrho$ which are satisfied for all activation functions commonly used in practice.

For the case $p = \infty$, the situation is more involved: For all activation functions that are commonly used in practice—except for the (parametric) ReLU—the results from above remain true also for $p = \infty$. But for the (parametric) ReLU, we do not know in general whether this is the case. However, for network architectures with only two layers, we prove that the associated sets of realizations of (parametric) ReLU networks with a given architecture, or with a given number of neurons, are closed with respect to the $L^\infty(\Omega)$ norm.

The established non-closedness of $\mathcal{RNN}_\varrho(S)$ and $\mathcal{RMN}_\varrho(N, L)$ is delicate in the sense that we additionally show that the set

$$\{R_\varrho^S(\Phi) : \Phi = (W_\ell)_{\ell=1} \text{ has architecture } S \text{ and } W_\ell = A_\ell(\cdot) + b_\ell \text{ with } \|A_\ell\| + \|b_\ell\| \leq C\}$$

of realizations of neural networks with a fixed architecture and all affine linear maps bounded in a suitable norm, is closed.

As a consequence of the preceding observations, we see that if a function $f$ lies in the $L^p$-closure of $\mathcal{RNN}_\varrho(S)$, but not in $\mathcal{RNN}_\varrho(S)$ itself, then for any sequence of networks $(\Phi_n)_{n \in \mathbb{N}}$ of architecture $S$ with $\|f - R_\varrho^S(\Phi_n)\|_{L^p} \to 0$, the weights of the networks $\Phi_n$ are not uniformly bounded as $n \to \infty$. The same observation holds when approximating with network realizations in $\mathcal{RMN}_\varrho(N, L)$. Clearly, such a behavior is undesirable in applications.

Finally, these results indicate a certain advantage of choosing the (parametric) ReLU as the underlying activation function, since—at least for two-layer networks—the problem just described does not occur.
1.3 Failure of inverse stability of the realization map

As our final result, we study the stability of the realization mapping $R^\Omega_\varphi$ from equation (1.1), which maps a family of weights to its realization. Even though this mapping will turn out to be continuous from the finite dimensional parameter space to $L^p(\Omega)$ for any $p \in (0, \infty]$, we will show that it is not inverse stable. In other words, for two realizations that are very close in the uniform norm there do not always exist networks (or rather, network weights) associated with these realizations that have a small distance. For both of these results—continuity and no inverse stability—we only need to assume that the activation function $\varphi$ is Lipschitz continuous and also differentiable at some point $x_0 \in \Omega$ with $\varphi'(x_0) \neq 0$.

These properties of the realization map pinpoint a potential problem that can occur when training a neural network: Let us consider a regression problem, where a network is iteratively updated by a (stochastic) gradient descent algorithm trying to minimize a loss function. It is then possible that at some iterate the loss function exhibits a very small error, even though the associated network weights are far away from the optimal solution. This is especially severe since a small error term leads to small steps if a gradient descent method is used in the optimization. Consequently, convergence to the very distant optimal weights will be slow even if the energy landscape of the optimization problem happens to be free of spurious local minima.

1.4 Outline of the paper

After this general discussion of our results, we will properly start our investigation of neural networks in Section 2 where we introduce our network model, as well as a variety of commonly used activation functions that we will refer to throughout this paper. Afterwards, we define a number of operations on networks that will be used frequently: First, we show how one can build networks that compute the cartesian product and the composition of two given network realizations. Second, we show how the identity function can be locally approximated arbitrarily well by networks of fixed complexity.

In Section 3, we study the shape of the set of realizations of networks. Theorem 3.3 shows that this set is star-shaped, but cannot have more than a certain number of linearly independent centers, if the activation function is locally Lipschitz continuous. As a consequence, we will see in Corollary 3.4 and Theorem 3.5 that the set of realizations of networks of a fixed architecture is never convex, unless the activation function is of a special form, which is never true for the commonly used activation functions. Proposition 3.8 shows that even a considerably weaker form of convexity does not hold.

After this analysis of the convexity properties, we study the interior and the density of the class of realizations of neural networks with a given architecture. More precisely, in Subsection 3.3 we will see that $\mathcal{RNN}_{\varphi}(S)$ and $\mathcal{RNN}_{\varphi}(N,L)$ both have empty interior in any infinite dimensional topological vector space, provided that the activation function $\varphi$ is locally Lipschitz. We also provide conditions under which these sets are nowhere dense.

The closedness of the set or realizations will be studied in Section 4. Theorem 4.1 shows that all common activation functions $\varphi$ yield network sets $\mathcal{RNN}_{\varphi}(S)$ and $\mathcal{RNN}_{\varphi}(N,L)$ that are not closed in $L^p$, for $0 < p < \infty$. The same holds in $L^\infty(\Omega)$ for a large variety of common activation functions except for the (parametric) ReLU, as we will see in Theorem 4.2, Theorem 4.4 and Theorem 4.5. The consequences of this non-closedness, such as exploding weights in regression problems, are discussed in Subsection 4.4. In Subsection 4.4, we consider networks that use the (parametric) ReLU as their activation function; in particular, we study whether the realization sets of such networks are closed in $L^\infty(\Omega)$. Even though we were unable to answer this question in full generality, we prove the closedness for networks with only two layers and for networks with uniformly bounded weights, but potentially unbounded biases.

Finally, in Section 5, we study properties of the function $R^\Omega_\varphi$ that maps a family of weights to the associated realization. We will see that $R^\Omega_\varphi$ is continuous as a map into $L^p(\Omega)$ ($0 < p \leq \infty$) if the activation function $\varphi$ is continuous. Likewise, if $\varphi$ is locally Lipschitz continuous, then so is $R^\Omega_\varphi$. This last property will be especially useful, since the mapping properties of locally Lipschitz functions are well understood; for example they map $n$-dimensional null sets to $n$-dimensional null sets. However, although $R^\Omega_\varphi$ is continuous, Theorem 5.2 and Corollary 5.3 will show under mild assumptions on the activation function $\varphi$ that $R^\Omega_\varphi$ is not inverse stable.
To not interrupt the flow of reading, almost all proofs are deferred to the appendix.

1.5 Notation

The symbol \( \mathbb{N} \) will denote the natural numbers \( \mathbb{N} = \{1, 2, 3, \ldots \} \), whereas \( \mathbb{N}_0 \) stands for the natural numbers including zero. Moreover, we denote by \( \mathbb{N}_{\geq d} \) the set of all natural numbers which are greater or equal to \( d \in \mathbb{N} \). The number of elements of a set \( M \) will be denoted by \( |M| \in \mathbb{N}_0 \cup \{\infty\} \). Furthermore, we write \( \mathbb{N}_n := \{k \in \mathbb{N} : k \leq n\} \) for \( n \in \mathbb{N}_0 \).

For two sets \( A, B \), a map \( f : A \to B \), and \( C \subset A \), we write \( f|_C \) for the restriction of \( f \) to \( C \). For a set \( A \), we denote by \( \chi_A = 1_A \) the indicator function of \( A \), so that \( \chi_A(x) = 1 \) if \( x \in A \) and \( \chi_A(x) = 0 \) otherwise. For any \( \mathbb{R} \)-vector space \( \mathcal{Y} \) we write \( A + B := \{a + b : a \in A, b \in B\} \) and \( \lambda A := \{\lambda a : a \in A\} \), for subsets \( A, B \subset \mathcal{Y} \) and \( \lambda \in \mathbb{R} \).

The algebraic dual space of a \( \mathbb{K} \)-vector space \( \mathcal{Y} \), (with \( \mathbb{K} = \mathbb{R} \) or \( \mathbb{K} = \mathbb{C} \)), that is the space of all linear functions \( \varphi : \mathcal{Y} \to \mathbb{K} \), will be denoted by \( \mathcal{Y}^* \). In contrast, if \( \mathcal{Y} \) is a topological vector space, we denote by \( \mathcal{Y}' \) the topological dual space of \( \mathcal{Y} \), which consists of all functions \( \varphi \in \mathcal{Y}^* \) that are continuous. Given functions \( f : \mathcal{X} \to \mathcal{C} \) and \( g : \mathcal{B} \to \mathcal{C} \), we write \( f \otimes g : \mathcal{X} \times \mathcal{B} \to \mathcal{C} \), \((x, y) \mapsto f(x) \cdot g(y)\) for the tensor product of \( f, g \). The tensor product of more than two functions is defined similarly.

The closure of a subset \( A \) of a topological space \( \mathcal{Z} \) will be denoted by \( \overline{A} \), while the interior of \( A \) is denoted by \( A^\circ \). For a metric space \( (\mathcal{U}, d) \), we write \( B_\varepsilon(x) := \{y \in \mathcal{U} : d(x, y) < \varepsilon\} \) for the \( \varepsilon \)-ball around \( x \), where \( x \in \mathcal{U} \) and \( \varepsilon > 0 \). Furthermore, for a Lipschitz continuous function \( f : \mathcal{U}_1 \to \mathcal{U}_2 \) between two metric spaces \( \mathcal{U}_1 \) and \( \mathcal{U}_2 \), we denote by \( \text{Lip}(f) \) the smallest possible Lipschitz constant for \( f \).

For \( d \in \mathbb{N} \) and a function \( f : A \to \mathbb{R}^d \) or a vector \( v \in \mathbb{R}^d \) we denote for \( j \in \{1, \ldots, d\} \) the \( j \)-th component of \( f \) or \( v \) by \( (f)_j \) or \( v_j \), respectively. As an example, the Euclidean scalar product on \( \mathbb{R}^d \) is given by \( \langle x, y \rangle = \sum_{i=1}^d x_i y_i \). For a matrix \( A \in \mathbb{R}^{n \times d} \), let \( \|A\|_{\max} := \max_{i=1,\ldots,n} \max_{j=1,\ldots,d} |A_{ij}| \). The transpose of a matrix \( A \in \mathbb{R}^{n \times d} \) will be denoted by \( A^\top \in \mathbb{R}^{d \times n} \). The euclidean unit sphere in \( \mathbb{R}^d \) will be denoted by \( S^{d-1} \subset \mathbb{R}^d \).

For a compact set \( K \subset \mathbb{R}^d \), we denote by \((C(K), \|\cdot\|_{\max})\) the Banach space of all real-valued, continuous functions defined on \( K \), equipped with the supremum norm. We note that on \((C(K), \|\cdot\|_{\max})\), the supremum norm coincides with the \( L^\infty(K) \)-norm, if for all \( x \in K \) and for all \( \varepsilon > 0 \) we have that \( \lambda(K \cap B_\varepsilon(x)) > 0 \), where \( \lambda \) denotes the Lebesgue measure on \( \mathbb{R}^d \). For any nonempty set \( U \subset \mathbb{R} \), we say that a function \( f : U \to \mathbb{R} \) is increasing if \( f(x) \leq f(y) \) for every \( x, y \in U \) with \( x < y \). If even \( f(x) < f(y) \) for all such \( x, y \), we say that \( f \) is strictly increasing. The terms “decreasing” and “strictly decreasing” are defined similarly.

The Schwartz space will be denoted by \( \mathcal{S}(\mathbb{R}^d) \) and the space of tempered distributions by \( \mathcal{S}'(\mathbb{R}^d) \). The corresponding bilinear dual pairing will be denoted by \( \langle \cdot, \cdot \rangle_{\mathcal{S}, \mathcal{S}'} \). We refer to [10] Sections 8.1–8.3 and 9.2] for more details on the spaces \( \mathcal{S}(\mathbb{R}^d) \) and \( \mathcal{S}'(\mathbb{R}^d) \). Finally, the Dirac delta distribution \( \delta_x \) at \( x \in \mathbb{R}^d \) is given by \( \delta_x : C(\mathbb{R}^d) \to \mathbb{R}, f \mapsto f(x) \).

2 Neural networks and basic operations

In this section, we introduce our basic framework of neural networks, explain some basic operations on networks, and state all required definitions for the remainder of the paper.

First, we introduce a terminology for neural networks allowing us to differentiate between a network as a family of weights and the function implemented by the network. This implemented function will be called the realization of the network.

**Definition 2.1.** Let \( d, L \in \mathbb{N} \). A neural network \( \Phi \) with input dimension \( d \) and \( L \) layers is a sequence of matrix-vector tuples

\[ \Phi = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L)) \]

where \( N_0 = d \) and \( N_1, \ldots, N_L \in \mathbb{N} \), and where each \( A_j \) is an \( N_j \times N_{j-1} \) matrix, and \( b_j \in \mathbb{R}^{N_j} \).
If \( \Phi \) is a neural network as above, \( K \subset \mathbb{R}^d \), and if \( \varrho : \mathbb{R} \to \mathbb{R} \) is arbitrary, then we define the associated \( \varrho \)-realization of \( \Phi \) with activation function \( \varrho \) over \( K \) as the map \( R^K_\varrho(\Phi) : K \to \mathbb{R}^{N_L} \) such that
\[
R^K_\varrho(\Phi)(x) = x_L,
\]
where \( x_L \) results from the following scheme:
\[
x_0 := x,
\]
\[
x_\ell := \varrho(A_\ell x_{\ell-1} + b_\ell), \quad \text{for } \ell = 1, \ldots, L - 1,
\]
\[
x_L := A_L x_{L-1} + b_L,
\]
where \( \varrho \) acts componentwise, that is, \( \varrho(y) = (\varrho(y_1), \ldots, \varrho(y_m)) \) for \( y = (y_1, \ldots, y_m) \in \mathbb{R}^m \).

We call the tuple \( (d, N_1, \ldots, N_L) \) the architecture of \( \Phi \). We call \( N(\Phi) := d + \sum_{j=1}^L N_j \) the number of neurons of the network \( \Phi \) and \( L = L(\Phi) \) the number of layers. Moreover, we refer to \( N_L \) as the dimension of the output layer of \( \Phi \), or simply as the output dimension of \( \Phi \).

While the activation function can theoretically be arbitrarily chosen, a couple of particularly useful activation functions have been established in the literature. We proceed by listing some of the most common activation functions, some of their properties, as well as some references to articles using these functions in the context of deep learning. We note that the following list only includes non-constant, monotonically increasing, globally Lipschitz continuous functions, whereas some results of this paper also hold for locally Lipschitz continuous functions. Furthermore, all functions listed below belong to the class \( C^\infty(\mathbb{R} \setminus \{0\}) \).

| Name                      | Given by                                                                 | Smoothness/ Boundedness | Cit. |
|---------------------------|--------------------------------------------------------------------------|-------------------------|------|
| rectified linear unit (ReLU) | \( \max\{0, x\} \)                                                     | \( C(\mathbb{R}) / \text{Unbounded} \) | 38   |
| parametric ReLU           | \( \max\{ax, x\} \) for some \( a \geq 0 \)                           | \( C(\mathbb{R}) / \text{Unbounded} \) | 21   |
| parametric exponential linear unit | \( x \cdot \chi_{x \geq 0}(x) + a \cdot (\exp(x) - 1) \cdot \chi_{x < 0}(x) \), for \( a > 0, a \neq 1 \) | \( C(\mathbb{R}) / \text{Unbounded} \) | 9    |
| exponential linear unit   | \( x \cdot \chi_{x \geq 0}(x) + (\exp(x) - 1) \cdot \chi_{x < 0}(x) \) | \( C^1(\mathbb{R}) / \text{Unbounded} \) | 9    |
| softsign                  | \( \frac{x}{1 + |x|} \)                                                | \( C^1(\mathbb{R}) / \text{Bounded} \) | 5    |
| inverse square root linear unit | \( x \cdot \chi_{x \geq 0}(x) + \frac{x}{\sqrt{1 + ax^2}} \cdot \chi_{x < 0}(x) \) for \( a > 0 \) | \( C^2(\mathbb{R}) / \text{Unbounded} \) | 8    |
| inverse square root unit  | \( \frac{x}{\sqrt{1 + ax^2}} \) for some \( a > 0 \)                  | \( C^\infty(\mathbb{R}) / \text{Bounded} \) | 8    |
| sigmoid / logistic        | \( \frac{1}{1 + \exp(-x)} \)                                          | Analytic / Bounded       | 20   |
| tanh                      | \( \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \)                  | Analytic / Bounded       | 31   |
| arctan                    | \( \arctan(x) \)                                                      | Analytic / Bounded       | 30   |
| softplus                  | \( \ln(1 + \exp(x)) \)                                               | Analytic / Unbounded     | 17   |

Table 1: Commonly-used activation functions and their properties
2.1 Sets of networks with fixed size

We will mainly be interested in the sets of neural networks of fixed size. Here we consider two ways to fix the size of networks. First, we prescribe the architecture of a network. Second, we only fix the number of neurons and layers. Fixing the architecture has the advantage that the set of underlying networks (that is, the set of weights) forms a normed vector space. Moreover, in applications one typically defines an architecture before optimizing the weights. Fixing the number of neurons on the other hand is more common in approximation theoretical results.

We start by fixing our notation for sets of networks with fixed architecture.

**Definition 2.2.** For \( d, N_1, \ldots, N_L \in \mathbb{N} \), we define the space \( \mathcal{N}(d, N_1, \ldots, N_L) \) as the set of all networks with architecture \((d, N_1, \ldots, N_L)\). This space, equipped with

\[
\|\Phi\|_{\mathcal{N}(d, N_1, \ldots, N_L)} := \|\Phi\|_{\text{scaling}} + \max_{\ell=1,\ldots,L} \|b_\ell\|_{\text{max}},
\]

where, \( \|\Phi\|_{\text{scaling}} := \max_{\ell=1,\ldots,L} \|A_\ell\|_{\text{max}} \), is a finite-dimensional normed vector space.

For the sake of brevity, and unless the underlying space matters, we write \( \|\Phi\|_{\text{total}} := \|\Phi\|_{\mathcal{N}(d, N_1, \ldots, N_L)} \) for \( \Phi \in \mathcal{N}(d, N_1, \ldots, N_L) \).

For a compact set \( K \subset \mathbb{R}^d \), and a continuous activation function \( \varphi : \mathbb{R} \to \mathbb{R} \), the map

\[
R^K_\varphi : \mathcal{N}(d, N_1, \ldots, N_L) \to C(K; \mathbb{R}^{N_L}), \quad \Phi \mapsto R^K_\varphi(\Phi)
\]

will from now on be called the realization map. Now we can state the definition of the set of realizations of networks with fixed architecture.

**Definition 2.3.** For \( d, N_1, \ldots, N_L \in \mathbb{N} \), \( K \subset \mathbb{R}^d \), and \( \varphi : \mathbb{R} \to \mathbb{R} \) continuous, we define

\[
\mathcal{RN}(d, N_1, \ldots, N_L, \varphi) := R^K_\varphi(\mathcal{N}(d, N_1, \ldots, N_L)).
\]

We call \( \mathcal{RN}(d, N_1, \ldots, N_L, \varphi) \) the set of \( \varphi \)-realizations of networks with architecture \((d, N_1, \ldots, N_L)\).

Additionally, we will also deal with neural networks having a fixed number of neurons, since this model is used frequently in approximation theoretical results.

**Definition 2.4.** For given \( d, L, N, N_L \in \mathbb{N} \) we denote by \( \mathcal{N}_{d,L,N,N_L} \) the set of all neural networks with \( d \)-dimensional input, \( L \) layers, \( N \) neurons, and \( N_L \)-dimensional output. We denote by \( \mathcal{RN}(d, L, N, N_L, \varphi) \) the set of all \( \varphi \)-realizations of neural networks in \( \mathcal{N}(d, L, N, N_L) \). If \( N_L = 1 \), we define \( \mathcal{N}_{d,L,N} := \mathcal{N}_{d,L,N,1} \), and \( \mathcal{RN}(d, L, N, \varphi) := \mathcal{RN}(d, L, N, 1, \varphi) \).

It is clear that

\[
\mathcal{N}_{d,L,N,N_L} = \bigcup_{\sum_{\ell=1}^{L-1} N_{\ell-1} = N-L, N_L \in \mathbb{N}} \mathcal{N}(d, N_1, \ldots, N_L),
\]

\[
\mathcal{RN}(d, L, N, N_L, \varphi) = \bigcup_{\sum_{\ell=1}^{L-1} N_{\ell-1} = N-L, N_L \in \mathbb{N}} \mathcal{RN}(d, L, N, 1, \varphi).
\]

2.2 Operations on networks

We will now see that it is possible to enlarge a given neural network in such a way that the realizations of the original network and the enlarged network coincide. To be more precise, the following holds:
Lemma 2.5. Let $d, L \in \mathbb{N}$, $K \subset \mathbb{R}^d$, and $\varphi : \mathbb{R} \to \mathbb{R}$. Moreover, let $\Phi = ((A_1, b_1), \ldots, (A_L, b_L))$ be a neural network with architecture $(d, N_1, \ldots, N_L)$ and let $\tilde{N}_1, \ldots, \tilde{N}_{L-1} \in \mathbb{N}$ such that $\tilde{N}_\ell \geq N_\ell$ for all $\ell = 1, \ldots, L-1$. Then there exists a neural network $\hat{\Phi}$ with architecture $(d, \tilde{N}_1, \ldots, \tilde{N}_{L-1}, N_L)$ and $R^K_\varphi(\hat{\Phi}) = R^K_\varphi(\Phi)$. In particular,

$$R^N_{K}(d, N_1, \ldots, N_{L-1}, N_L) \subset R^N_{K}(d, \tilde{N}_1, \ldots, \tilde{N}_{L-1}, N_L).$$

Moreover, for every $\tilde{N} \in \mathbb{N}$ with $N < \tilde{N}$, we have

$$R^N_{K}(d, N_1, \ldots, N_{L-1}, N_L) \subset R^N_{K}(d, \tilde{N}_1, \ldots, \tilde{N}_{L-1}, N_L).$$

Proof. For the proof of this statement we refer to Appendix A.1.

In addition to increasing the size of a network without changing its realization, one can also manipulate networks. We can put two networks in parallel—which means that the realization of the resulting network is the cartesian product of the realizations of the two original networks—by using the following procedure:

Definition 2.6. Let $d, L \in \mathbb{N}$ and let $\Phi^1 = ((A^1_1, b^1_1), \ldots, (A^1_L, b^1_L))$, $\Phi^2 = ((A^2_1, b^2_1), \ldots, (A^2_L, b^2_L))$ be two neural networks with $L$ layers and with $d$-dimensional input. We define

$$P(\Phi^1, \Phi^2) := \left( (\tilde{A}_1, \tilde{b}_1), \ldots, (\tilde{A}_L, \tilde{b}_L) \right),$$

where

$$\tilde{A}_1 := \left( \begin{array}{c} A^1_1 \\ A^2_1 \end{array} \right), \quad \tilde{b}_1 := \left( \begin{array}{c} b^1_1 \\ b^2_1 \end{array} \right), \quad \text{and} \quad \tilde{A}_\ell := \left( \begin{array}{cc} A^1_\ell & 0 \\ 0 & A^2_\ell \end{array} \right), \quad \tilde{b}_\ell := \left( \begin{array}{c} b^1_\ell \\ b^2_\ell \end{array} \right), \quad \text{for } 1 < \ell \leq L.$$

Then $P(\Phi^1, \Phi^2)$ is a neural network with $d$-dimensional input and $L$ layers, called the parallelization of $\Phi^1$ and $\Phi^2$.

One readily verifies that if the network $\Phi^1$ has architecture $(d, N_1, \ldots, N_L)$ and the network $\Phi^2$ has architecture $(d, \tilde{N}_1, \ldots, \tilde{N}_L)$, then $P(\Phi^1, \Phi^2)$ has architecture $(d, N_1 + \tilde{N}_1, \ldots, N_L + \tilde{N}_L)$. Consequently, $N(P(\Phi^1, \Phi^2)) = N(\Phi^1) + N(\Phi^2) - d$, where $d$ is the input dimension of $\Phi^1$ and $\Phi^2$. Furthermore, it is not hard to see for every $K \subset \mathbb{R}^d$ that

$$R^K_\varphi(P(\Phi^1, \Phi^2))(x) = (R^K_\varphi(\Phi^1)(x), R^K_\varphi(\Phi^2)(x)), \quad \text{for all } x \in K.$$

A consequence of being able to put two networks in parallel is that sums of realizations of neural networks are themselves realizations of neural networks, but with an increased number of neurons.

Lemma 2.7. Let $d, L, N, N_1, \ldots, N_{L-1} \in \mathbb{N}$, let $K \subset \mathbb{R}^d$, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function. Then for all $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\lambda_1 R^N_{K}(d, L, N, \varphi) + \lambda_2 R^N_{K}(d, L, N, \varphi) \subset R^{2N}_{K}(d, L, 2N_{L-1}, \varphi)$$

and

$$\lambda_1 R^N_{K}(d, N_1, \ldots, N_{L-1}, 1) + \lambda_2 R^N_{K}(d, N_1, \ldots, N_{L-1}, 1) \subset R^{2N}_{K}(d, 2N_1, \ldots, 2N_{L-1}, 1).$$

Proof. For the proof of this statement we refer to Appendix A.2.

The second operation we can perform with networks is concatenation, as given in the following definition.

Definition 2.8. Let $L_1, L_2 \in \mathbb{N}$ and let $\Phi^1 = ((A^1_1, b^1_1), \ldots, (A^1_{L_1}, b^1_{L_1}))$, $\Phi^2 = ((A^2_1, b^2_1), \ldots, (A^2_{L_2}, b^2_{L_2}))$ be two neural networks such that the input layer of $\Phi^1$ has the same dimension as the output layer of $\Phi^2$. Then, $\Phi^1 \cdot \Phi^2$ denotes the following $L_1 + L_2 - 1$ layer network:

$$\Phi^1 \cdot \Phi^2 := ((A^1_1, b^1_1), \ldots, (A^1_{L_2}, b^1_{L_2}), (A^1_{L_2+1} A^2_1 b^1_{L_2} + b^1_1), (A^2_2, b^2_2), \ldots, (A^1_{L_1}, b^1_{L_1})).$$

We call $\Phi^1 \cdot \Phi^2$ the concatenation of $\Phi^1$ and $\Phi^2$. 

8
One directly verifies that for every $\varrho: \mathbb{R} \to \mathbb{R}$ the definition of concatenation is reasonable, that is, if $d_i$ is the dimension of the input layer of $\Phi^i$, $i = 1, 2$, and if $K \subset \mathbb{R}^{d_2}$, then $R^K_\varrho(\Phi^1 \bullet \Phi^2) = R^K_\varrho(\Phi^1) \circ R^K_\varrho(\Phi^2)$. If $\Phi^2$ has architecture $(d, N_1, \ldots, N_{L_2})$ and $\Phi^1$ has architecture $(N_{L_2}, \bar{N}_2, \ldots, \bar{N}_{L_{L_1}-1}, \bar{N}_{L_{L_1}})$, then $\Phi^1 \bullet \Phi^2$ has architecture $(d, N_1, \ldots, N_{L_2}, \bar{N}_2, \ldots, \bar{N}_{L_{L_1}})$. Therefore, $N(\Phi^1 \bullet \Phi^2) = N(\Phi^1) + N(\Phi^2) - N_{L_{L_1}}$.

Before we continue, we will show that under some very mild assumptions on $\varrho$, which are almost always satisfied in practice, one can construct a neural network with a limited number of non-zero weights which locally approximates the identity mapping $\text{id}_{\mathbb{R}^d}$ to every given accuracy. Similarly, one can obtain a neural network the realization of which approximates the projection onto the $i$-th coordinate.

**Proposition 2.9.** Let $\varrho: \mathbb{R} \to \mathbb{R}$ be continuous, and assume that there exists $x_0 \in \mathbb{R}$ such that $\varrho$ is continuously differentiable in a neighborhood of $x_0$ and that $\varrho'(x_0) \neq 0$. Then, for every $\varepsilon > 0, d \in \mathbb{N}, B > 0$ and every $L \in \mathbb{N}$ there exists a neural network $\Phi^B_{\varepsilon} \in \mathcal{N}\mathcal{N}(d, d, \ldots, d) \subset \mathcal{N}\mathcal{N}_{d,L,(L+1)d,d}$ such that

- $|R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon})(x) - x| \leq \varepsilon$, for all $x \in [-B, B]^d$;
- $R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon})(0) = 0$;
- $R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon})$ is continuously differentiable;
- $\left(\frac{\partial}{\partial x_i}\right)_{x=0} \left(R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon})\right)(x) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{else}; \end{cases}$
- for $j \in \{1, \ldots, d\}$, $\left(R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon})\right)_j$ is constant in all but the $j$-th coordinate.

Furthermore, for every $d \in \mathbb{N}$, $\varepsilon > 0$, $B > 0$ and every $i \in \{1, \ldots, d\}$, one can construct a neural network $\Phi^B_{\varepsilon,i} \in \mathcal{N}\mathcal{N}(d, 1, \ldots, 1) \subset \mathcal{N}\mathcal{N}_{d,L,L+d}$ such that

- $|R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon,i})(x) - x_i| \leq \varepsilon$ for all $x \in [-B, B]^d$;
- $R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon,i})(0) = 0$;
- $R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon,i})$ is continuously differentiable;
- $\frac{\partial}{\partial x_i}\left|_{x=0} \right. R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon,i})(x) = 1$; and
- $R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon,i})$ is constant in all but the $i$-th coordinate.

Finally, if $\varrho$ is increasing, then $\left(R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon})\right)_j$ and $\left(R^\varrho_{-B,B}^d(\Phi^B_{\varepsilon,i})\right)_j$ are monotonically increasing in every coordinate and for all $j \in \{1, \ldots, d\}$.

**Proof.** For the proof of this statement we refer to Appendix A.3.

---

### 3 Shape of the set of realizations

In this section, we analyze several algebraic and geometric properties of the set of realizations of neural networks. These results give an insight into the shape of this set. We start by analyzing to what extent the set of neural network realizations is star-shaped. Afterwards, we will show for a large class of activation functions that the corresponding set of neural network realizations is highly non-convex. We finish the section by analyzing the interior of the set of neural network realizations.
3.1 Star-shapedness of the set of neural network realizations

Before we study the star-shapedness of the set of all realizations of neural networks with a fixed architecture or with a certain number of neurons and layers, we first investigate under which conditions this set is nontrivial:

Lemma 3.1. Let $d, L, N \in \mathbb{N}$, let $\emptyset \neq K \subset \mathbb{R}^d$, and let $\varrho : \mathbb{R} \to \mathbb{R}$. The set $\mathcal{RNN}_{d,L,N,\varrho}^K$ is nonempty if and only if $N \geq d + L$.

Proof. For the proof of this statement we refer to Appendix B.1.

Now, the star-shapedness of the set of all realizations of neural networks is a direct consequence of the fact that the set is invariant under scalar multiplication. The following proposition provides the details.

Proposition 3.2. Let $d, L, N \in \mathbb{N}$ with $N \geq d + L$, let $K \subset \mathbb{R}^d$, and let $\varrho : \mathbb{R} \to \mathbb{R}$. Then, $\mathcal{RNN}_{d,L,N,\varrho}^K$ is closed under scalar multiplication and is star-shaped with respect to the origin (that is, the zero function).

If $N_1, \ldots, N_{L-1} \in \mathbb{N}$ then $\mathcal{RNN}_{d,L,N,\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ is closed under scalar multiplication and is star-shaped with respect to the origin.

Proof. For the proof of this statement we refer to Appendix B.2.

As a direct consequence of Proposition 3.2, we see that $\mathcal{RNN}_{d,L,N,\varrho}^K$ is path-connected. Next, we will see that $\mathcal{RNN}_{d,L,N,\varrho}^K$ cannot contain infinitely many linearly independent centers. A center of a set $A$ is a point $x_0 \in A$ such that for all $x \in A$ and $\lambda \in [0, 1]$ also $\lambda x_0 + (1 - \lambda)x \in A$.

Theorem 3.3. Let $d, L, N, N_1, \ldots, N_{L-1} \in \mathbb{N}$, let $K \subset \mathbb{R}^d$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz continuous. Then the number of linearly independent centers of $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1})$ is at most $\sum_{\ell=1}^L (N_{\ell-1} + 1)N_{\ell}$, where $N_0 = d$. Moreover, $\mathcal{RNN}_{d,L,N,\varrho}^K$ has at most $N^2 \cdot \max\{0, L-2\} + (d + L) \cdot N + 1$ linearly independent centers.

Proof. For the proof of the statement we refer to Appendix B.3.

3.2 Non-convexity of the set of neural network realizations

Next, we analyze the convexity of $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ and of $\mathcal{RNN}_{d,L,N,\varrho}^K$. As a direct consequence of Theorem 3.3, we see that $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ and $\mathcal{RNN}_{d,L,N,\varrho}^K$ are never convex if the respective set contains a set with more than a certain number of linearly independent functions.

Corollary 3.4. Let $d, L, N, N_1, \ldots, N_{L-1} \in \mathbb{N}$, $N_0 = d$, $K \subset \mathbb{R}^d$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz continuous.

- If $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ contains more than $\sum_{\ell=1}^L (N_{\ell-1} + 1)N_{\ell}$ linearly independent functions, then $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ is not convex.
- If $\mathcal{RNN}_{d,L,N,\varrho}^K$ contains more than $N^2 \cdot \max\{0, L-2\} + (d + L) \cdot N + 1$ linearly independent functions, then $\mathcal{RNN}_{d,L,N,\varrho}^K$ is not convex.

Proof. Every element of a convex set is a center. Thus the result follows directly from Theorem 3.3.

Corollary 3.3 claims that if a set of realizations of neural networks with fixed size contains more than a fixed number of linearly independent functions, then it cannot be convex. Since $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ and $\mathcal{RNN}_{d,L,N,\varrho}^K$ are by construction translation invariant sets, it is very likely that they contain infinitely many linearly independent functions. In fact, our next result shows, under minor regularity assumptions on $\varrho$, that if the set $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1)$ does not contain infinitely many linearly independent functions, then $\varrho$ can only be a sum of finitely many polynomials multiplied by exponential functions. Since $\mathcal{RNN}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1) \subset \mathcal{RNN}_{d,L,N,\varrho}^K$ for $N = d + \sum_{\ell=1}^L N_{\ell}$, the same statement holds if $\mathcal{RNN}_{d,L,N,\varrho}^K$ contains only finitely many linearly independent functions and $N \geq d + L$. 

10
Theorem 3.5. Let $d \in \mathbb{N}$, let $L \in \mathbb{N}_{\geq 2}$, and let $N_1, \ldots, N_{L-1} \in \mathbb{N}$. Moreover, let $g : \mathbb{R} \to \mathbb{R}$ be continuous. Assume that there exists some $x_0 \in \mathbb{R}$ such that $g$ is continuously differentiable on a neighborhood of $x_0$ and $g'(x_0) \neq 0$.

Assume further that $\mathcal{RNN}^{K}_{g}(d, N_1, \ldots, N_{L-1}, 1)$ does not contain infinitely many linearly independent functions. Then there exist some $r \in \mathbb{N}$, and $a_i, \lambda_i \in \mathbb{R}$, $k_i \in \mathbb{N}_0$ for $i = 1, \ldots, r$ such that

$$\varphi(x) = \sum_{i=1}^{r} a_i x^{k_i} e^{\lambda_i x} \quad \text{for all } x \in \mathbb{R}.$$ 

Proof. For the proof of the statement we refer to Appendix B.4. \hfill \Box

We have seen in Corollary 3.4 and Theorem 3.5 that sets of realizations of neural networks with fixed size are only rarely convex. In applications where the approximation capabilities of networks are studied, we might not necessarily care about the convexity of $\mathcal{RNN}^{K}_{g}(d, N_1, \ldots, N_{L-1}, 1)$ or $\mathcal{RNN}^{K}_{d,L,N,\varphi}$ but rather about the convexity of the respective closures with respect to the sup-norm.

Towards this goal, we first recall the definition of a discriminatory function from [12].

Definition 3.6 ([12]). Let $d \in \mathbb{N}$, and let $K \subset \mathbb{R}^d$ be compact. A measurable function $f : \mathbb{R} \to \mathbb{R}$ is discriminatory with respect to $K$ if for every finite, signed, regular Borel measure $\mu$ on $K$ we have that

$$\left( \int_K f(Ax + b) d\mu(x) = 0 \quad \text{for all } A \in \mathbb{R}^{1 \times d} \text{ and } b \in \mathbb{R} \right) \implies \mu = 0.$$ 

Next, we demonstrate for a compact set $K \subset \mathbb{R}^d$ that if $\overline{\mathcal{RNN}^{K}_{1,L,N,\varphi}}$ contains a discriminatory function for $K$, then for all $d \in \mathbb{N}$ the set $\mathcal{RNN}^{K}_{d,L,N,\varphi}$ is either dense in $C(K)$ or its closure is not convex. The same result holds when replacing $\mathcal{RNN}^{K}_{d,L,N,\varphi}$ by $\mathcal{RNN}^{K}_{\varphi}(d, N_1, \ldots, N_{L-1}, 1)$. For a discussion which of the commonly used activation functions of Table [1] lead to realizations containing a discriminatory function, we refer to Subsection 3.2.2.

Proposition 3.7. Let $d, L, N, N_1, \ldots, N_{L-1} \in \mathbb{N}$ such that $N \geq L + d$, let $K \subset \mathbb{R}^d$ be compact, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be continuous.

- If there exists a discriminatory function $f \in \overline{\mathcal{RNN}^{K}_{1,L,N,\varphi}}$ with respect to $K$ and if $\overline{\mathcal{RNN}^{K}_{d,L,N,\varphi}}$ is convex, then $\mathcal{RNN}^{K}_{d,L,N,\varphi} = C(K)$.

- If there exists a discriminatory function $f \in \overline{\mathcal{RNN}^{K}_{\varphi}(1, N_1, \ldots, N_{L-1}, 1)}$ with respect to $K$ and if $\mathcal{RNN}^{K}_{\varphi}(d, N_1, \ldots, N_{L-1}, 1)$ is convex, then $\overline{\mathcal{RNN}^{K}_{\varphi}(d, N_1, \ldots, N_{L-1}, 1)} = C(K)$.

Proof. For the proof of the statement we refer to Appendix B.5. \hfill \Box

Next, we extend Proposition 3.7 to a relaxed notion of convexity. To this end, for a subset $A$ of a vector space $\mathcal{Y}$, we denote the convex hull of $A$ by

$$\text{co}(A) := \bigcap_{\mathcal{Y} \ni B \supset A, B \text{ convex}} B.$$ 

For $\varepsilon > 0$, we say that a subset $A$ of a normed vector space is $\varepsilon$-convex in $(\mathcal{Y}, \| \cdot \|_\mathcal{Y})$, if

$$\text{co}(A) \subset A + B_\varepsilon(0).$$

Hence, the notion of $\varepsilon$-convexity simply asks whether the convex hull of a set is contained in an enlargement of this set. However, we have the following negative result concerning the $\varepsilon$-convexity of the set of all neural network realizations with a given number of layers and neurons, if the underlying normed vector space is given by $(C(K), \| \cdot \|_{\text{sup}})$.  

11
Proposition 3.8. Let \( d, L, N, N_1, \ldots, N_{L-1} \in \mathbb{N} \), let \( K \subset \mathbb{R}^d \) be compact, and let \( \varrho : \mathbb{R} \to \mathbb{R} \) be continuous.

- If there exists a function \( f \in \overline{\mathcal{RN}}_{d,L,N}^K \) that is discriminatory with respect to \( K \) and if we have \( \overline{\mathcal{RN}}_{d,L,N}^K \neq C(K) \), then there does not exist any \( \varepsilon > 0 \) such that \( \overline{\mathcal{RN}}_{d,L,N}^K \) is \( \varepsilon \)-convex in \( (C(K), \| \cdot \|_{\sup}) \).

- If there exists a function \( f \in \overline{\mathcal{RN}}_{\varrho}^K(1, N_1, \ldots, N_{L-1}, 1) \) that is discriminatory with respect to \( K \) and if we have \( \overline{\mathcal{RN}}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1) \neq C(K) \), then there does not exist any \( \varepsilon > 0 \) such that \( \overline{\mathcal{RN}}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1) \) is \( \varepsilon \)-convex in \( (C(K), \| \cdot \|_{\sup}) \).

Proof. For the proof of the statement we refer to Appendix B.6.

Before we proceed, we note that Proposition 3.7 and Proposition 3.8 yield the non-convexity of the closure of the set of network realizations of a given size, under the assumption that the network sets are not dense in \( C(K) \) and that a discriminatory function is a realization of a network with a prescribed size. In the following two subsections, we present a couple of conditions guaranteeing that the aforementioned assumptions are satisfied.

3.2.1 Non-dense network sets

We briefly review criteria on \( \varrho \) which ensure that \( \overline{\mathcal{RN}}_{d,L,N}^K \neq C(K) \), as well as criteria ensuring \( \overline{\mathcal{RN}}_{d,L,N}^K = C(K) \). The arguments in this subsection remain valid when replacing \( \overline{\mathcal{RN}}_{d,L,N}^K \) by \( \overline{\mathcal{RN}}_{\varrho}^K(d, N_1, \ldots, N_{L-1}, 1) \).

While it is certainly natural to think that \( \overline{\mathcal{RN}}_{d,L,N}^K \neq C(K) \) should hold for most activation functions, giving a reference including large classes of activation functions such that the claim holds is not straightforward. In fact, such a result does not hold for all activation functions. Indeed, [32, Theorem 4] gives a construction of an activation function such that \( \overline{\mathcal{RN}}_{d,L,N}^K \) is dense in \( C(K) \) and thus in \( L^p(K) \), \( p \in [1, \infty) \).

On the other hand, if \( \varrho \) is a piecewise polynomial with finitely many pieces, it is not hard to see that any element of \( \overline{\mathcal{RN}}_{d,L,N}^K \) must be a piecewise polynomial, and the number of pieces is bounded depending only on \( N \) and \( L \). This set is certainly not dense in \( C(K) \) or \( L^p(K) \), \( p \in [1, \infty) \). Moreover, under various conditions on the activation function, it is shown in [33, Section 14] that the pseudo-dimension of the set of realizations of neural networks is bounded. But it is not hard to see that every set of continuous functions that is dense in \( L^p(K) \) or \( C(K) \) needs to have infinite pseudo-dimension. Therefore, if \( \varrho \) is such that \( \overline{\mathcal{RN}}_{d,L,N}^K \) has finite pseudo-dimension, then \( \overline{\mathcal{RN}}_{d,L,N}^K \) is neither dense in \( C(K) \), nor in \( L^p(K) \).

3.2.2 Discriminatory functions

A wide class of discriminatory functions has been given in [12]. There it has been shown that every bounded, measurable and sigmoidal function \( \varrho : \mathbb{R} \to \mathbb{R} \), that is, every bounded measurable function \( \varrho \) satisfying \( \lim_{x \to -\infty} \varrho(x) = 1 \) and \( \lim_{x \to -\infty} \varrho(x) = 0 \), is discriminatory with respect to every compact rectangle \([-B, B]^d \) for \( B > 0 \). In fact, this includes a variety of commonly used activation functions such as the softsign function or the sigmoid function. Furthermore, there exist bounded, sigmoidal functions which are, up to possible shifts, linear combinations of the tanh function (consider for instance \( \text{tanh}(x/2 + 1/2) \)), the arctan function (consider for instance \( \arctan(x/\pi + 1/2) \)), and the ReLU function (consider for instance \( \text{ReLU}(x) - \text{ReLU}(-1) \)).

All of these functions have in common that they are strictly monotonically increasing on \([0, 1]\). This shows that for widely used activation functions \( \varrho : \mathbb{R} \to \mathbb{R} \) and a comparatively small number of neurons \( N \), the set \( \overline{\mathcal{RN}}_{[-B, B]^d}^{1,2,N} \) contains a bounded, continuous, sigmoidal function \( f \), which is thus discriminatory with respect to every \([-B, B]^d \). This function can be used to obtain bounded, continuous, sigmoidal functions \( f_L \) for every prescribed number of layers \( L \geq 2 \).
Proposition 3.9. Let \( L \in \mathbb{N}_{\geq 3} \). Furthermore, let \( \varrho : \mathbb{R} \to \mathbb{R} \) be continuous. Assume that for some \( \tilde{N} \in \mathbb{N} \) there exists a sigmoidal function \( f \in \mathcal{RN}_\varrho^R(1,\tilde{N},1) \) such that \( f(\mathbb{R}) \subset [0,1] \) and such that \( f \) is strictly monotonically increasing on \([0,1]\). Then, setting \( \tilde{N}_\ell := \tilde{N} \) for \( \ell = 1, \ldots, L - 1 \), there exists a sigmoidal function \( \tilde{f} \in \mathcal{RN}_\varrho^R(1,\tilde{N}_1,\tilde{N}_2,\ldots,\tilde{N}_{L-1},1) \subset \mathcal{RN}_\varrho^R_{1:L,L,N',\varrho} \), where \( N' \geq (L-1)\tilde{N} + 2 \) is arbitrary.

Proof. For the proof of this statement we refer to Appendix B.8.

3.3 Empty interior in \( C(K) \) and \( L^p(K) \)

As a final step of our study of the shape of the set of neural network realizations, we study its interior. It turns out that the interiors of \( \mathcal{RN}^K_{d,L,N,\varrho} \) and \( \mathcal{RN}^K_{d,L,N,\varrho} \) are empty for all practically relevant activation functions and all reasonable topologies. In fact, under mild assumptions on \( \varrho \), we will see that \( \mathcal{RN}^K_{d,L,N,\varrho} \) and \( \mathcal{RN}^K_{d,L,N,\varrho} \) have empty interior with respect to any reasonable topology. Since

\[
\mathcal{RN}^K_{\varrho}(d,N_1,\ldots,N_{L-1},1) \subset \mathcal{RN}^K_{d,L,N',\varrho}
\]

with \( N' = d + 1 + \sum_{\ell=1}^{L-1} N_\ell \) we see that empty interior or nowhere density of \( \mathcal{RN}^K_{d,L,N,\varrho} \) implies empty interior or nowhere density of \( \mathcal{RN}^K_{d,L,N,\varrho} \). Thus we state the following theorems only for the sets of realizations of networks with a fixed number of neurons.

Corollary 3.10. Let \( d, L, N \in \mathbb{N} \), let \( K \subset \mathbb{R}^d \), and let \( \varrho : \mathbb{R} \to \mathbb{R} \) be locally Lipschitz continuous. Let \( \mathcal{Y} \) be an infinite dimensional topological vector space such that \( C(K) \subset \mathcal{Y} \). Then \( \mathcal{RN}^K_{d,L,N,\varrho} \) has empty interior in \( \mathcal{Y} \).

In particular, if \( K \) is an infinite compact set or a compact set of positive measure, then this holds for \( \mathcal{Y} = C(K) \) and \( \mathcal{Y} = L^p(K) \) \( (p \in (0,\infty) \) arbitrary), respectively.

Proof. For the proof of the statement we refer to Appendix B.9.

Now we will see under an additional mild assumption that in every topological vector space \( \mathcal{Y} \) containing \( C(K) \), we have the even stronger statement that either \( \mathcal{RN}^K_{d,L,N,\varrho} \) is nowhere dense in \( \mathcal{Y} \), or \( \mathcal{RN}^K_{d,L,2N-d-1,\varrho} \) is dense in \( \mathcal{Y} \). Conditions ensuring that \( \mathcal{RN}^K_{d,L,2N-d-1,\varrho} \) is not dense in \( C(K) \) or \( L^p(K) \) were discussed in Subsection 3.2.1

Proposition 3.11. Let \( d, L, N \in \mathbb{N} \), let \( K \subset \mathbb{R}^d \), and let \( \varrho : \mathbb{R} \to \mathbb{R} \) be locally Lipschitz continuous. Let \( \mathcal{Y} \) be an infinite dimensional topological vector space such that \( C(K) \subset \mathcal{Y} \). If \( \mathcal{RN}^K_{d,L,2N-d-1,\varrho} \) is not dense in \( \mathcal{Y} \), then \( \mathcal{RN}^K_{d,L,N,\varrho} \) is nowhere dense in \( \mathcal{Y} \).

Proof. For the proof of the statement we refer to Appendix B.9.

4 Closedness of the set of realizations

In this section, for a compact set \( \varnothing \neq K \subset \mathbb{R}^d \) and for all practically-used activation functions \( \varrho \) of Table 1, we analyze if \( \mathcal{RN}^K_{d,L,N,\varrho} \) and \( \mathcal{RN}^K_{\varrho}(d,N_1,\ldots,N_{L-1},1) \) are closed sets with respect to the topologies of \( C(K) \) or \( L^p(K) \) for \( p \in (0,\infty) \). We will show that under very mild assumptions on the activation functions, the sets \( \mathcal{RN}^K_{d,L,N,\varrho} \) and \( \mathcal{RN}^K_{\varrho}(d,N_1,\ldots,N_{L-1},1) \) are not closed in \( L^p(K) \) for every \( p \in (0,\infty) \). However, concerning the closedness in \( C(K) \) we need to distinguish between different types of activation functions. In fact, we can show that for functions such as the sigmoid, the softplus, the arctan, the tanh, the inverse square root linear unit and the exponential linear unit, the sets \( \mathcal{RN}^K_{d,L,N,\varrho} \) as well as \( \mathcal{RN}^K_{\varrho}(d,N_1,\ldots,N_{L-1},1) \) are not closed in \( C(K) \). After a discussion of the consequences of the non-closedness, we conclude this section by proving that the set \( \mathcal{RN}^K_{d,2,N,\varrho} \) is closed in \( C(K) \), where \( \varrho \) is the ReLU or the parametric ReLU. We note that in the case of two layers, the set of neural networks with a fixed size coincides with the set of neural networks with a fixed architecture, that is, \( \mathcal{RN}^K_{d,2,N,\varrho} = \mathcal{RN}^K_{\varrho}(d,(N-d-1),1) \) so that in this case we do not need to distinguish between the two notions. Moreover, we note that most of the results to
come in this section are only formulated for compact rectangles of the form \( K = [-B, B]^d \) for \( B > 0 \); but these results can easily be generalized to hold for any compact set \( K \subset \mathbb{R}^d \) with non-empty interior.

### 4.1 Non-closedness in \( L^p([-B, B]^d) \)

We will see in this subsection that for \( B > 0 \) and all suitable, widely used activation functions (including all activation functions presented in Table 1), the classes \( \mathcal{RNN}_{d,L,N,\hat{\varphi}}^{[-B, B]^d} \) and \( \mathcal{RNN}_{\hat{\varphi}}^{[-B, B]^d}(d, N_1, \ldots, N_{L-1}, 1) \) are not closed in \( L^p([-B, B]^d) \), for any \( p \in (0, \infty) \). To be more precise, the following is true:

**Theorem 4.1.** Let \( d \in \mathbb{N} \), let \( p \in (0, \infty) \), and let \( \varphi : \mathbb{R} \rightarrow \mathbb{R} \) be a function such that

(i) \( \varphi \) is continuous and increasing.

(ii) There is some \( r > 0 \) such that \( \varphi \in C^1((-\infty, -r) \cup (r, \infty)) \) and some \( x_0 \in (-\infty, -r) \cup (r, \infty) \) such that \( \varphi'(x_0) > 0 \).

(iii) At least one of the following two conditions is satisfied:

(a) There are \( \lambda, \lambda' \geq 0 \) with \( \lambda \neq \lambda' \) such that \( \varphi'(x) \rightarrow \lambda \) as \( x \rightarrow \infty \), and \( \varphi'(x) \rightarrow \lambda' \) as \( x \rightarrow -\infty \).

(b) \( \varphi \) is bounded.

Then, for every \( L \in \mathbb{N}_{\geq 2} \), \( N > L+d \) and for all \( B > 0 \), the set \( \mathcal{RNN}_{d,L,N,\hat{\varphi}}^{[-B, B]^d} \) is not closed in \( L^p([-B, B]^d) \). Moreover, for all possible neural network architectures \((d, N_1, \ldots, N_{L-1}, 1)\) such that \( N_{L-1} \geq 2 \), the set \( \mathcal{RNN}_{\hat{\varphi}}^{K}(d, N_1, \ldots, N_{L-1}, 1) \) is not closed in \( L^p([-B, B]^d) \). Finally, also \( \bigcup_{N=L+d+1}^{\infty} \mathcal{RNN}_{d,L,N,\hat{\varphi}}^{[-B, B]^d} \) is not closed in \( L^p([-B, B]^d) \).

**Proof.** For the proof of this statement we refer to Appendix C.1.

### 4.2 Non-closedness in \( C([-B, B]^d) \) for many widely used activation functions

We have seen in Theorem 4.1 that for \( N > L + d \) and under reasonably mild assumptions on the activation function \( \varphi \), including all common activation functions, the set \( \mathcal{RNN}_{d,L,N,\hat{\varphi}}^{[-B, B]^d} \) as well as the set \( \mathcal{RNN}_{\hat{\varphi}}^{[-B, B]^d}(d, N_1, \ldots, N_{L-1}, 1) \) are not-closed in any \( L^p \)-space where \( p \) is positive and finite. However, the argument of the proof of Theorem 4.1 breaks down if one considers closedness with respect to the \( \| \cdot \|_{\sup} \) norm. Therefore, we will analyze this setting more closely in the present section.

We start by demonstrating that the sets \( \mathcal{RNN}_{d,L,N,\hat{\varphi}}^{[-B, B]^d} \) and \( \mathcal{RNN}_{\hat{\varphi}}^{[-B, B]^d}(d, N_1, \ldots, N_{L-1}, 1) \) are never closed in \( C([-B, B]^d) \), if the activation function \( \varphi \) satisfies \( \varphi \in C^1(\mathbb{R}) \setminus C^\infty(\mathbb{R}) \). This setting includes as common activation functions the integer powers of ReLUs, i.e., \( x \mapsto \max\{0, x^k\}, k \geq 2 \), the softsign function, the inverse square root linear unit and the exponential linear unit.

**Theorem 4.2.** Let \( d \in \mathbb{N} \), let \( B > 0 \), let \( L \in \mathbb{N}_{\geq 2} \) with \( N > L + d \), and let \( \varphi \in C^1(\mathbb{R}) \setminus C^\infty(\mathbb{R}) \). Then, \( \mathcal{RNN}_{d,L,N,\hat{\varphi}}^{[-B, B]^d} \) is not closed in \( C([-B, B]^d) \). Moreover, for any fixed neural network architecture \((d, N_1, \ldots, N_{L-1}, 1)\), satisfying \( N_{L-1} \geq 2 \), the set \( \mathcal{RNN}_{\hat{\varphi}}^{[-B, B]^d}(d, N_1, \ldots, N_{L-1}, 1) \) is not closed in the space \( C([-B, B]^d) \).

**Proof.** For the proof of the statement we refer to Appendix C.2.

**Remark 4.3.** Theorem 4.2 cannot be applied to some of the most frequently-used activation functions such as the ReLU, the leaky ReLU, the sigmoid function, the tanh function, the arctan function, and the softplus function, so that these cases need to be handled separately.

Another result concerning the non-closedness of the set of neural network realizations can be given for bounded analytic activation functions. This includes the sigmoid function, the tanh function, and the arctan function.
**Theorem 4.4.** Let $d \in \mathbb{N}$, let $B > 0$, let $L \in \mathbb{N}_{>2}$ with $N > L + d$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be bounded, analytic, and not constant. Then, $\mathcal{RNN}^{[-B,B]^d}_{d,L,N,\varrho}$ is not closed in $C([-B,B]^d)$. Moreover, for any fixed neural network architecture $(d,N_1,\ldots,N_{L-1},1)$ satisfying $N_{L-1} \geq 2$, the set $\mathcal{RNN}^{[-B,B]^d}_{\varrho}(d,N_1,\ldots,N_{L-1},1)$ is not closed in $C([-B,B]^d)$.

**Proof.** For the proof of the statement we refer to Appendix C.3.

Finally, we prove the non-closedness of the set of neural network realizations for functions that are essentially homogeneous, such as the soft-plus function and other smooth functions which uniformly approximate the ReLU function. We call a function $f$ **approximately homogeneous of order** $(p,q) \in \mathbb{N}^2$ if there exists $r > 0$ such that $|f(x) - x^p| \leq r$ for all $x \geq 0$ and $|f(x) - x^q| \leq r$ for all $x \leq 0$. Clearly, this holds for the soft-plus function, for $q = 1$ and $p = 0$.

**Theorem 4.5.** Let $d \in \mathbb{N}$, let $p,q \in \mathbb{N}_0$ with $p \neq q$, let $B > 0$ and $L \in \mathbb{N}_{>2}$ with $N > L + d$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be approximately homogeneous of order $(p,q)$ and such that $\varrho \in C^{\text{max}(p,q)}(\mathbb{R})$. Then, $\mathcal{RNN}^{[-B,B]^d}_{d,L,N,\varrho}$ is not closed in $C([-B,B]^d)$. Moreover, for any fixed neural network architecture $(d,N_1,\ldots,N_{L-1},1)$, the set $\mathcal{RNN}^{[-B,B]^d}_{\varrho}(d,N_1,\ldots,N_{L-1},1)$ is not closed in $C([-B,B]^d)$.

**Proof.** For the proof of the statement we refer to Appendix C.3.

### 4.3 Consequences of non-closedness

We have just seen that the sets $\mathcal{RNN}^{[-B,B]^d}_{d,L,N,\varrho}$ and $\mathcal{RNN}^{[-B,B]^d}_{\varrho}(d,N_1,\ldots,N_{L})$ are not closed in $L^p([-B,B]^d)$ for any $p \in (0,\infty)$ and basically every practically relevant activation function. Furthermore, for a variety of activation functions, we have seen that $\mathcal{RNN}^{[-B,B]^d}_{d,L,N,\varrho}$ and $\mathcal{RNN}^{[-B,B]^d}_{\varrho}(d,N_1,\ldots,N_{L-1},1)$ are not closed in $C([-B,B]^d)$. In this subsection, we briefly discuss the consequences of these results. To this end, we first analyze the closedness of the set of realizations if we only allow for bounded scaling weights and biases.

**Proposition 4.6.** Let $d,L,N \in \mathbb{N}$, let $K \subset \mathbb{R}^d$ be compact, let furthermore $p \in (0,\infty)$, and let $\varrho : \mathbb{R} \to \mathbb{R}$ be continuous. For $C > 0$, let

$$\Theta_C := \{ \Phi \in \mathcal{NN}_{d,L,N} : \| \Phi \|_{\text{total}} \leq C \}.$$ 

Then the set $\mathcal{R}^K_{\varrho}(\Theta_C)$ is compact in $C(K)$ as well as in $L^p(K)$. Moreover, for any fixed neural network architecture $(d,N_1,\ldots,N_{L})$, the set $\mathcal{R}^K_{\varrho}(\Theta_C \cap \mathcal{NN}(d,N_1,\ldots,N_{L-1},1))$ is compact in $C(K)$ as well as in $L^p(K)$.

**Proof.** For the proof of the statement we refer to Appendix C.3.

We now discuss the results of the preceding subsections in combination with the statement of Proposition 4.6. The non-closedness of $\mathcal{R}$, where $\mathcal{R}$ is either $\mathcal{RNN}^{[-B,B]^d}_{d,L,N,\varrho}$ or $\mathcal{RNN}^{[-B,B]^d}_{\varrho}(d,N_1,\ldots,N_{L-1},1)$, implies that for certain functions $f \in \mathcal{Y}$, where $\mathcal{Y}$ is either $L^p([-B,B]^d)$ or $C([-B,B]^d)$, there does not exist a best approximation in $\mathcal{R}$, that is, there does not exist a function $g \in \mathcal{R}$ such that

$$\| f - g \|_{\mathcal{Y}} = \inf_{h \in \mathcal{R}} \| f - h \|_{\mathcal{Y}}.$$ 

Additionally, Proposition 4.6 shows that the subset of $\mathcal{R}$ that contains only realizations of networks with uniformly bounded weights is compact. Hence we conclude that whenever $f \in \overline{\mathcal{R}} \setminus \mathcal{R}$, where the closure is taken in $\mathcal{Y}$, then for every sequence $(\Phi_n)_{n \in \mathbb{N}} \subset \mathcal{R}$ of networks with $\| \mathcal{R}^{[-B,B]^d}_{\varrho}(\Phi_n) - f \|_{\mathcal{Y}} \to 0$, we must have $\| \Phi_n \|_{\text{total}} \to \infty$, which means that the weights of the networks $\Phi_n$ explode.
4.4 Closedness of ReLU networks in $C(K)$

In this subsection we analyze the closedness of sets of realizations of neural networks with respect to the ReLU or the parametric ReLU activation function in $C(K)$. We conjecture that the set of ReLU networks of a fixed complexity is closed in $C(K)$, but were not able to prove such a result in full generality. In two special cases, namely when the number of layers is bounded by 2 or when at least the scaling weights are

Proposition 5.1.

We first analyze the set of realizations with uniformly bounded scaling weights and possibly unbounded biases, before proceeding with the analysis of the set $\mathcal{RNN}^K_{d,L,N,N,L}$.

For $\Phi = ((A_1, b_1), \ldots, (A_L, b_L)) \in \mathcal{NN}_{d,L,N,N,L}$ such that $\|\Phi\|_{\text{scaling}} \leq C$ for some $C > 0$, we say that the network $\Phi$ has $C$-bounded scaling weights. Note that this does not require the biases $b_{\ell}$ of the network to satisfy $|b_{\ell}| \leq C$.

Our first goal in this subsection is to show that if $\varrho$ denotes the ReLU, and if $\mathcal{K} \subset \mathbb{R}^d$ is measurable and bounded, then the set

$$\mathcal{RNN}^K_{d,L,N,N,L,\varrho} := \{ R_\varrho^K(\Phi) : \Phi \in \mathcal{NN}_{d,L,N,N,L} \text{ with } \|\Phi\|_{\text{scaling}} \leq C \}$$

is closed in $C(K; \mathbb{R}^N)$ and in $L^p(K; \mathbb{R}^N)$ for arbitrary $p \in [1, \infty)$. Here, and in the remainder of the paper, we use the norm $\|f\|_{L^p(K; \mathbb{R}^N)} = \|f\|_{L^p}$ for vector-valued $L^p$-spaces. The norm on $C(K; \mathbb{R}^N)$ is defined similarly. The difference between the following proposition and Proposition 4.6 is that in the following proposition, the “shift weights” (the biases) of the networks can be potentially unbounded.

Proposition 4.7. Let $\mathcal{K} \subset \mathbb{R}^d$ be measurable and bounded and of positive measure. Let $d, N_L, L, N \in \mathbb{N}$, and let $C > 0$. Finally, let $\varrho : \mathbb{R} \to \mathbb{R}$, $x \mapsto \max\{0, x\}$ denote the ReLU.

Then the set $\mathcal{RNN}^K_{d,L,N,N,L,\varrho}$ is closed in $L^p(K; \mathbb{R}^N)$ for every $p \in [1, \infty]$, and also in $C(K; \mathbb{R}^N)$.

Proof. For the proof of the statement we refer to Appendix C.6 \hfill \Box

Remark. In fact, the proof shows that each subset of $\mathcal{RNN}^K_{d,L,N,N,L,\varrho}$ which is bounded in the $L^1$-norm is compact in $L^p(K; \mathbb{R}^N)$ and in $C(K; \mathbb{R}^N)$.

Now we will see that the set of realizations of two-layer neural networks with arbitrary scaling weights and biases is closed in $C([-B, B]^d)$, if the activation is the parametric ReLU.

Theorem 4.8. Let $d, N \in \mathbb{N}$ and $B > 0$, and let $0 \leq a \leq 1$. Let $\varrho_a : \mathbb{R} \to \mathbb{R}, x \mapsto \max\{x, ax\}$ be the parametric ReLU. Then $\mathcal{RNN}^{[aB,B]^d}_{a^2N,a\varrho_a}$ is closed in $C([-B, B]^d)$.

Proof. For the proof of the statement we refer to Appendix C.7 \hfill \Box

5 Failure of inverse instability of the realization map

Let us fix a neural network architecture $(N_0, \ldots, N_{L-1}, 1)$ with $N_0 = d$, a compact set $\mathcal{K} \subset \mathbb{R}^d$, and a continuous activation function $\varrho : \mathbb{R} \to \mathbb{R}$. We study the properties of the realization map $R_\varrho^K$. First of all, we observe that the realization map is a continuous function.

Proposition 5.1. Let $d \in \mathbb{N}$, and let $\mathcal{K} \subset \mathbb{R}^d$ be compact. If the activation function $\varrho : \mathbb{R} \to \mathbb{R}$ is continuous, then the realization map from Equation 2.1 is continuous. If $\varrho$ is locally Lipschitz continuous, then so is $R_\varrho^K$.

Finally, if $\varrho$ is Lipschitz continuous, then there is a constant $C = C(\varrho, d, N_1, \ldots, N_L)$ with

$$\text{Lip}(R^K_\varrho(\Phi)) \leq C \cdot \|\Phi\|_{\text{scaling}}$$

for all $\Phi \in \mathcal{NN}(d, N_1, \ldots, N_L)$.

Proof. For the proof of this statement we refer to Appendix D.4 \hfill \Box
In general, the realization map is not injective, that is, there can be networks \( \Phi \neq \Psi \) but such that \( R^K_\varrho(\Phi) = R^K_\varrho(\Psi) \); in fact, if for instance

\[
\Phi = ((A_1, b_1), \ldots, (A_{L-1}, b_{L-1}), (0, 0)) \quad \text{and} \quad \Psi = ((B_1, c_1), \ldots, (B_{L-1}, c_{L-1}), (0, 0)),
\]

then the realizations of \( \Phi, \Psi \) are identical.

In this section, our main goal is to determine whether, up to the failure of injectivity, the realization map is a homeomorphism onto its range. We will see that this is not the case.

To this end, we will prove that even if \( R^K_\varrho(\Phi) \) is very close to \( R^K_\varrho(\Psi) \), it is not true in general that \( R^K_\varrho(\Psi) = R^K_\varrho(\Phi) \) for a network \( \tilde{\Psi} \) that is close to \( \Phi \). Precisely, the following holds:

**Theorem 5.2.** Let \( \varrho : \mathbb{R} \to \mathbb{R} \) be Lipschitz continuous and continuously differentiable in some \( x_0 \in \mathbb{R} \) with \( \varrho'(x_0) \neq 0 \), but not affine-linear. Let \( S = (N_0, \ldots, N_{L-1}, 1) \) be a network architecture with \( L \geq 2 \), and with \( N_0 = d, \) and \( N_1 \geq 3 \). Let \( K \subset \mathbb{R}^d \) be bounded with nonempty interior.

Then, there is a sequence \( (\Phi_n)_{n \in \mathbb{N}} \) of networks with architecture \( S \) and the following properties:

1. We have \( R^K_\varrho(\Phi_n) \to 0 \) uniformly on \( K \).
2. We have \( \text{Lip}(R^K_\varrho(\Phi_n)) \to \infty \) as \( n \to \infty \).

Finally, if \( (\Phi_n)_{n \in \mathbb{N}} \) is a sequence of networks with architecture \( S \) and the preceding two properties, then the following holds: For each sequence of networks \( (\Psi_n)_{n \in \mathbb{N}} \) with architecture \( S \) and \( R^K_\varrho(\Psi_n) = R^K_\varrho(\Phi_n) \), we have \( \|\Psi_n\|_{\text{scaling}} \to \infty \).

**Proof.** For the proof of the statement we refer to Appendix D.2.

We finally rephrase the preceding result in more topological terms:

**Corollary 5.3.** Under the assumptions of Theorem 5.2, the realization map \( R^K_\varrho \) from Equation (2.1) is not a quotient map when considered as a map onto its range.

**Proof.** For the proof of the statement we refer to Appendix D.3.

## Acknowledgements

P.P. and M.R. are supported by the DFG Collaborative Research Center TRR 109 “Discretization in Geometry and Dynamics”. M.R. is supported by the Berlin Mathematical School. FV acknowledges support from the European Commission through DEDALE (contract no. 665044) within the H2020 Framework Program.

## A Proofs of the results in Section 2

### A.1 Proof of Lemma 2.5

Setting \( \tilde{N}_0 := d \) and \( \tilde{N}_L := N_L \), we define \( \tilde{\Phi} := ((\tilde{A}_1, \tilde{b}_1), \ldots, (\tilde{A}_L, \tilde{b}_L)) \) by

\[
\tilde{A}_\ell := \begin{pmatrix}
A_\ell \\
0_{\tilde{N}_\ell \times (\tilde{N}_\ell - 1)}
\end{pmatrix}
\quad \text{and} \quad
\tilde{b}_\ell := \begin{pmatrix}
b_\ell \\
0_{\tilde{N}_\ell - 1}
\end{pmatrix}
\in \mathbb{R}^{\tilde{N}_\ell \times \tilde{N}_{\ell-1}}
\quad \text{for} \quad \ell = 1, \ldots, L,
\]

where \( 0_{m \times n} \) denotes the zero-matrix in \( \mathbb{R}^{m \times n} \). Clearly, \( R^K_\varrho(\tilde{\Phi}) = R^K_\varrho(\Phi) \). This yields the first part of the lemma and (2.2).

For the proof of (2.3), let \( \tilde{N} \in \mathbb{N} \) such that \( N < \tilde{N} \) and \( \Phi \in \mathcal{N}\mathcal{N}_{d, L, N, N_L} \). Then it is clear that there exist \( N_1, \ldots, N_{L-1} \) summing up to \( N - d - N_L \) such that \( \Phi \in \mathcal{N}\mathcal{N}(d, N_1, \ldots, N_L) \). By what we have shown before, there exists some \( \tilde{\Phi} \in \mathcal{N}\mathcal{N}(d, N_1, N_2, \ldots, N_{L-1} + (\tilde{N} - N), N_L) \subset \mathcal{N}\mathcal{N}_{d, L, \tilde{N}, N_L} \) with \( R^K_\varrho(\Phi) = R^K_\varrho(\tilde{\Phi}) \). This yields the claim.

\[\square\]
A.2 Proof of Lemma 2.7

Let \( \Phi^1, \Phi^2 \in \mathcal{N}_{d,L,N} \) with architecture \((d, N, \ldots, N)\), and consider the network

\[
P(\Phi^1, \Phi^2) := ((A_1, b_1), \ldots, (A_L, b_L)).
\]

Then, for the network

\[
\tilde{\Phi} := ((A_1, b_1), \ldots, (A_{L-1}, b_{L-1}), ((\lambda_1 \lambda_2) A_L, (\lambda_1 \lambda_2) b_L))
\]

it holds that

\[
R^K_\varrho(\tilde{\Phi})(x) = \lambda_1 R^K_\varrho(\Phi^1)(x) + \lambda_2 R^K_\varrho(\Phi^2)(x) \text{ for all } x \in K.
\]

Moreover, since \( P(\Phi^1, \Phi^2) \in \mathcal{N}_{d,2N,\ldots,2N-L+1,2} \subset \mathcal{N}_{d,L,2N-d,2} \), we conclude that

\[
\tilde{\Phi} \in \mathcal{N}_{d,2N,\ldots,2N-L+1} \subset \mathcal{N}_{d,L,2N-d-1}.
\]

A.3 Proof of Proposition 2.9

Without loss of generality, we only consider the case \( \varepsilon \leq 1 \). Define \( \varepsilon' := \varepsilon/(dL) \). Let \( x_0 \in \mathbb{R} \) be such that \( \varrho \) is continuously differentiable in a neighborhood of \( x_0 \) and such that \( \varrho'(x_0) \neq 0 \). We set \( r_0 := \varrho(x_0) \) and \( s_0 := \varrho'(x_0) \). Next, for every \( C > 0 \), we define

\[
\varrho_C : [-B - L \varepsilon, B + L \varepsilon] \to \mathbb{R}, \quad x \mapsto C s_0 \varrho \left( \frac{x}{C} + x_0 \right) - C r_0.
\]

We claim that there is some \( C_0 > 0 \) such that \( |\varrho_C(x) - x| \leq \varepsilon' \) for all \( x \in [-B - L \varepsilon, B + L \varepsilon] \) and all \( C \geq C_0 \). To see this, first note by definition of the derivative that there is some \( \delta > 0 \) with

\[
|\varrho(t + x_0) - r_0 - s_0 t| \leq \frac{|s_0| \cdot \varepsilon'}{1 + B + L} \cdot |t| \quad \text{for all } t \in \mathbb{R} \text{ with } |t| \leq \delta.
\]

Here we implicitly used that \( s_0 = \varrho'(x_0) \neq 0 \) to ensure that the right-hand side is positive. Now, set \( C_0 := (B + L)/\delta \), and let \( C \geq C_0 \) be arbitrary. Note because of \( \varepsilon' \leq \varepsilon \leq 1 \) that every \( x \in [-B - L \varepsilon, B + L \varepsilon] \) satisfies \( |x| \leq B + L \). Hence, if we set \( t := x/C \), then \( |t| \leq \delta \). Therefore,

\[
|\varrho_C(x) - x| = \left| \frac{C}{s_0} \right| \cdot |\varrho(t + x_0) - r_0 - s_0 t| \leq \left| \frac{C}{s_0} \right| \cdot \frac{|s_0| \cdot \varepsilon'}{1 + B + L} \cdot \left| \frac{x}{C} \right| \leq \varepsilon'.
\]

Since \( \varrho \) is continuously differentiable in a neighborhood of \( x_0 \), it is not hard to see that \( \varrho_C \) is continuously differentiable on \([-B - L \varepsilon, B + L \varepsilon] \) for \( C \geq C_0 \) sufficiently large, which we will always assume in the following.

Let \( \Phi^C_C := ((A_1, b_1), (A_2, b_2)) \), where

\[
A_1 := \frac{1}{C} \cdot \text{id}_{\mathbb{R}^d} \in \mathbb{R}^{d \times d}, \quad b_1 := x_0 \cdot (1, \ldots, 1)^T \in \mathbb{R}^d,
\]

and where

\[
A_2 := \frac{C}{s_0} \cdot \text{id}_{\mathbb{R}^d} \in \mathbb{R}^{d \times d}, \quad b_2 := -\frac{C r_0}{s_0} \cdot (1, \ldots, 1)^T \in \mathbb{R}^d.
\]

Note \( \Phi^C_C \in \mathcal{N}_{d,2,3d,d} \). To shorten the notation, define \( K := [-B, B]^d \). Then, \( R^K_\varrho(\Phi^C_C) = \varrho_C|_K \otimes \cdots \otimes \varrho_C|_K \), where the tensor product has \( d \) factors. We define \( \Phi_C := \Phi^C_C \bullet \Phi^C_C \bullet \cdots \bullet \Phi^C_C \), where we take \( L - 2 \) concatenations. We obtain \( \Phi_C \in \mathcal{N}_{d,L,(L+1)d,d} \) and

\[
R^K_\varrho(\Phi_C)(x) = (\varrho_C \circ \varrho_C \circ \cdots \circ \varrho_C(x_i))_{i=1,...,d}, \quad \text{for all } x \in K.
\]
Since $|g_C(x) - x| \leq \varepsilon' \leq \varepsilon$ for all $x \in [-B - L\varepsilon, B + L\varepsilon]$, it is not hard to see by induction that
\[
|(g_C \circ \cdots \circ g_C)(x) - x| \leq t \cdot \varepsilon' \leq t \cdot \varepsilon,
\]
where the composition has $t \leq L$ factors. Therefore, since $\varepsilon' = \varepsilon/(dL)$, we conclude for $C \geq C_0$ that
\[
|R^K_\varrho(\Phi_C)(x) - x| \leq \varepsilon, \quad \text{for all } x \in K.
\]

By construction, $R^K_\varrho(\Phi_C)$ is continuously differentiable. Moreover, $g_C(0) = 0$ and $g_C'(0) = 1$. By induction, we thus get $\frac{d}{dx} |(g_C \circ \cdots \circ g_C)(x)| = 1$, where the composition has at most $L$ factors. Thanks to Equation (A.1), this yields $D(R^K_\varrho(\Phi_C))(0) = \text{id}_{\mathbb{R}^d}$, as claimed.

Also by Equation (A.1), we see that for every $j \in \{1, \ldots, d\}$, $(R^K_\varrho(\Phi_C)_j)(x)$ is constant in all but the $j$-th coordinate. Additionally, if $\varrho$ is increasing, then $g_C$ and hence $(R^K_\varrho(\Phi_C)_j)$ are increasing in the $j$-th coordinate. Hence, $\Phi^B_\varrho := \Phi_C$ gives the desired estimate.

We proceed with the second part of the proposition. We first prove the statement for $i = 1$. Let $\Phi^C_1 := ((A'_1, b'_1), (A''_1, b''_1))$, where
\[
A'_1 := \left(\frac{1}{C} \quad 0 \quad \cdots \quad 0\right) \in \mathbb{R}^{1 \times d}, \quad b'_1 := x_0 \in \mathbb{R}^1, \quad A''_1 := \frac{C}{s_0} \in \mathbb{R}^{1 \times 1}, \quad b''_1 := -\frac{Cr_0}{s_0} \in \mathbb{R}^1.
\]
We have that $\Phi^C_1 \in \mathcal{NN}_{d,2,d+2}$. Further let $\Phi^C_2 := ((A'_2, b'_2), (A''_2, b''_2))$, where
\[
A'_2 := \frac{1}{C} \in \mathbb{R}^{1 \times 1}, \quad b'_2 := x_0 \in \mathbb{R}^1, \quad A''_2 := \frac{C}{s_0} \in \mathbb{R}^{1 \times 1}, \quad b''_2 := -\frac{Cr_0}{s_0} \in \mathbb{R}^1.
\]
We have that $\Phi^C_2 \in \mathcal{NN}_{1,2,3}$. Setting $\Phi_C := \Phi^C_2 \circ \cdots \circ \Phi^C_1$, where we take $L - 2$ concatenations, yields a neural network $\Phi_C \in \mathcal{NN}_{d, L, L + d}$ such that
\[
R^K_\varrho(\Phi_C)(x) := g_C \circ g_C \circ \cdots \circ g_C(x_1), \quad \text{for all } x \in K.
\]
Exactly as in the proof of the first part, this implies for $C \geq C_0$ that
\[
|R^K_\varrho(\Phi_C)(x) - x| \leq \varepsilon, \quad \text{for all } x \in K.
\]
Setting $\Phi^B_\varrho := \Phi_C$ and repeating the previous arguments yields the claim for $i = 1$. Permuting the columns of $A'_i$ yields the result for arbitrary $i \in \{1, \ldots, d\}$.

Now, let $\varrho$ be increasing. Then also $g_C$ is increasing for every $C > 0$. Since $R^K_\varrho(\Phi_C)$ is the composition of componentwise monotonically increasing functions, the claim follows. \hfill \Box

## B Proofs of the results in Section 3

### B.1 Proof of Lemma 3.1

\textbf{”⇒”}: The set $\mathcal{NN}_{d,L,N,\varrho}$ is nonempty if and only if the set $\mathcal{NN}_{d,L,N,1}$ is. But for arbitrary $\Phi \in \mathcal{NN}_{d,L,N,1}$, we have $\Phi = ((A_1, b_1), \ldots, (A_L, b_L))$, where each $A_\ell \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$ for certain $N_1, \ldots, N_{L-1} \in \mathbb{N}$ and $N_0 = 0, N_L = 1$. Thus, $N = N(\Phi) = d + \sum_{\ell=1}^L N_\ell \geq d + L$.

\textbf{”⇐”}: Assume $N \geq d + L$, and define $N_0 := d \in \mathbb{N}, N_1 := N - d - L + 1 \in \mathbb{N}$, and $N_\ell := 1$ for $\ell \in \{2, \ldots, L\}$. Finally, choose $A_\ell := 0 \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$ and $b_\ell := 0 \in \mathbb{R}^d$ for $\ell \in \{1, \ldots, L\}$. Then $\Phi := ((A_1, b_1), \ldots, (A_L, b_L)) \in \mathcal{NN}_{d,L,N,\varrho} \neq \emptyset$, so that also $\mathcal{NN}_{d,L,N,\varrho}^K \neq \emptyset$. \hfill \Box
B.2 Proof of Proposition 3.2

Let $f \in \mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$ and $\Phi := ((A_1, b_1), \ldots, (A_L, b_L)) \in \mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1)$ be such that $f = R^K_\theta(\Phi)$. For $\lambda \in \mathbb{R}$, we define

$$\bar{\Phi} := ((A_1, b_1), \ldots, (A_{L-1}, b_{L-1}), (\lambda A_L, \lambda b_L))$$

and observe that $\bar{\Phi} \in \mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1)$ and $\lambda f = R^K(\bar{\Phi}) \in \mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$. This yields the closedness of $\mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$ under scalar multiplication and thus also $\mathcal{RN}_d^K (d, L, N, \theta)$ is closed under scalar multiplication.

Since $\mathcal{RN}_d^K (d, L, N, \theta) \neq \emptyset$ by Lemma 3.1, we can choose $\lambda = 0$ in the argument above to get $0 \in \mathcal{RN}_d^K (d, L, N, \theta)$. For every $f \in \mathcal{RN}_d^K (d, L, N, \theta)$, the line between 0 and $f$ is given by \{ $\lambda f : \lambda \in [0,1]$ \} $\subset \mathcal{RN}_d^K (d, L, N, \theta)$, where the inclusion follows by the closedness of $\mathcal{RN}_d^K (d, L, N, \theta)$ under scalar multiplication. We conclude that $\mathcal{RN}_d^K (d, L, N, \theta)$ is star-shaped with respect to the origin. Similarly, $\mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$ is seen to be star-shaped with center 0.

B.3 Proof of Theorem 3.3

We first handle the case $L = 1$. Here, it is not hard to see that every function in $\mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$ is the restriction of an affine linear function $W : \mathbb{R}^d \rightarrow \mathbb{R}$ to $K$. The space of affine-linear functions $W : \mathbb{R}^d \rightarrow \mathbb{R}$ has dimension $d + 1$, so that we easily get the claim in this case. Thus, we will assume for the remainder of the proof that $L \geq 2$.

It is easy to see that $\dim \mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1) = D$ for $D := \sum_{l=1}^L (N_{l-1} + 1)N_l$.

Let us assume towards a contradiction that $\mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$ contains $M := D + 1$ linearly independent centers $R^K_\theta(\Phi_1), \ldots, R^K_\theta(\Phi_M)$. Since $\mathcal{RN}_d^K (d, N_1, \ldots, N_{L-1}, 1)$ is closed under multiplication with scalars, this easily implies

$$V := \text{span} \{ R^K_\theta(\Phi_1), \ldots, R^K_\theta(\Phi_M) \} \subset R^K_\theta(d, N_1, \ldots, N_{L-1}, 1). \quad \text{(B.1)}$$

Set $W := \text{span} \{ \delta_x : x \in K \} \subset V^*$, where $\delta_x$ is the linear functional $\delta_x : V \rightarrow \mathbb{R}$, $f \mapsto f(x)$. Since $W$ is a finite-dimensional vector space (because of $\dim W \leq \dim V^* = \dim V = M$), there are $x_1, \ldots, x_M \in K$ with $W = \text{span} \{ \delta_{x_1}, \ldots, \delta_{x_M} \}$. Set $K' := \{ x_1, \ldots, x_M \}$. We claim that the restriction map $V \rightarrow C(K')$, $f \mapsto f|_{K'}$ is injective. Indeed, if $f \in V \subset C(K)$ with $f|_{K'} \equiv 0$, and if $x \in K$ is arbitrary, then $\delta_x = \sum_{i=1}^M a_i \delta_{x_i}$, for certain $a_1, \ldots, a_M \in \mathbb{R}$, and thus $f(x) = \delta_x(f) = \sum_{i=1}^M a_i \cdot f(x_i) = 0$, so that $f \equiv 0$. Overall, the considerations in the preceding paragraph show that $R^K_\theta(\Phi_1), \ldots, R^K_\theta(\Phi_M)$ are not only linearly independent as elements of $C(K)$, but also as elements of $C(K')$. Recalling $K' = \{ x_1, \ldots, x_M \}$, this entails that the $x_1, \ldots, x_M$ are pairwise distinct, since otherwise $\dim C(K') \leq |\{ x_1, \ldots, x_M \}| < M$. Furthermore, we see $C(K') = V = \text{span} \{ R^K_\theta(\Phi_1), \ldots, R^K_\theta(\Phi_M) \}$. In combination with Equation \text{(B.1)} this shows that $R^K_\theta : \mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1) \rightarrow C(K')$ is surjective.

Let us fix isomorphisms $J_1 : C(K') \rightarrow \mathbb{R}^M$ and $J_2 : \mathbb{R}^D \rightarrow \mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1)$, and recall $M = D + 1$.

Since $\theta$ is locally Lipschitz continuous, by Proposition 5.1 the realization map

$$R^K_\theta : \mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1), \| \cdot \|_{\mathcal{NN}(d, N_1, \ldots, N_{L-1}, 1)} \rightarrow (C(K'), \| \cdot \|_{\text{sup}})$$

\[\text{This follows by induction on } M, \text{ using the following observation: If } V \text{ is a vector space contained in a set } A, \text{ if } A \text{ is closed under multiplication with scalars, and if } x_0 \in A \text{ is a center for } A, \text{ then } V + \text{span} \{ x_0 \} \subset A. \text{ Indeed, let } \mu \in \mathbb{R} \text{ and } v \in V. \text{ There is some } \varepsilon \in \{ 1, -1 \} \text{ such that } \varepsilon \mu = |\mu|. \text{ Now set } x := \varepsilon v \in V \subset A \text{ and } \lambda := |\mu|/(1 + |\mu|) \in [0,1]. \text{ Then } v + \varepsilon x_0 = \varepsilon (v + |\mu| x_0) = \varepsilon (1 + |\mu|) (\frac{1}{1 + |\mu|} v + \frac{|\mu|}{1 + |\mu|} x_0) = \varepsilon (1 + |\mu|) (\lambda x_0 + (1 - \lambda) x) \in A.\]
is locally Lipschitz continuous. As we saw above, it is also surjective. Overall, we see that

\[ H : \mathbb{R}^{D+1} \to \mathbb{R}^{D+1}, \]

\[ (x_1, \ldots, x_D, x_{D+1}) \mapsto J_1 \circ R_{\phi}^{K'} \circ J_2(x_1, \ldots, x_D) \]

is locally Lipschitz continuous. Moreover, \( H \) is surjective since \( J_1 \) and \( J_2 \) are surjective, and since we saw above that the realization map \( R_{\phi}^{K'} : \mathcal{N}(d, N_1, \ldots, N_{L-1}) \to C(K') \) is also surjective. By construction \( H(\mathbb{R}^D \times \{0\}) = H(\mathbb{R}^{D+1}) = \mathbb{R}^{D+1} \). But it is well known (see for instance [1] Theorem 5.9), that a locally Lipschitz function between Euclidean spaces of the same dimension maps sets of Lebesgue measure zero to sets of Lebesgue measure zero. This yields the desired contradiction and thus establishes that \( \mathcal{N} \mathcal{N}_{\varepsilon}^K(d, N_1, \ldots, N_{L-1}, 1) \) can have at most \( D \) linearly independent centers.

Finally, we also address the number of centers of \( \mathcal{N} \mathcal{N}_{d,L,N,\varepsilon}^K \). By Lemma 2.3, for each neural network \( \Phi \in \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1) \) such that \( d + 1 + \sum_{k=1}^{L-1} = N \), there is a network \( \tilde{\Phi} \in \mathcal{N}(d, N, N, \ldots, N, 1) \) such that \( R_{\varepsilon}^K(\tilde{\Phi}) = R_{\varepsilon}^K(\Phi) \), which implies that

\[ \mathcal{N} \mathcal{N}_{d,L,N,\varepsilon}^K \subset \mathcal{N} \mathcal{N}_{d,L,N,\varepsilon}^K(\mathcal{N}(d, N, \ldots, N, 1)). \]

Therefore, \( \mathcal{N} \mathcal{N}_{d,L,N,\varepsilon}^K \) having more than \( (L-2) \cdot N^2 + Nd + N + 1 + (L-1) \cdot N \) centers contradicts the first part of the proof. □

B.4 Proof of Theorem 3.5

Let \( N_1, \ldots, N_{L-1} \in \mathbb{N} \) and define

\[ \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* := \left\{ f : \mathbb{R} \to \mathbb{R} \mid \text{there is some } g \in \mathcal{N} \mathcal{N}_{d}^*(d, N_1, \ldots, N_{L-1}, 1) \text{ with } f(x) = g(x, 0, \ldots, 0) \text{ for all } x \in \mathbb{R} \right\}. \]

Clearly, \( \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* \) is translation invariant, i.e., if \( f \in \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* \), then also \( f(-x) \in \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* \) for all \( x \in \mathbb{R} \). Furthermore, by our assumption, \( \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* \) only contains finitely many linearly independent functions. Therefore, \( V := \text{span} \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* \) is a finite-dimensional translation invariant subspace of \( C(\mathbb{R}) \). Thus, as shown in [2], there exists some \( r \in \mathbb{N} \), and certain \( \lambda_i \in \mathbb{R} \), \( k_i \in \mathbb{N}_0 \) for \( i = 1, \ldots, r \) such that

\[ \mathcal{N} \mathcal{N}_{d,L,\varepsilon}^* \subset V \subset \text{span} \left\{ x \mapsto x^{k_i} e^{\lambda_i x} : i = 1, \ldots, r \right\}. \]

Let \( \varepsilon, B > 0 \) be arbitrary. Since \( \phi \) is continuous, it is uniformly continuous on \( [-B - 1, B + 1] \), i.e., there is some \( \delta \in (0, 1) \) such that \( |\phi(x) - \phi(y)| \leq \varepsilon \) for all \( x, y \in [-B - 1, B + 1] \) with \( |x - y| \leq \delta \). Since \( L \geq 2 \), Proposition 2.9 and Lemma 2.5 yield a neural network \( \Phi_{\varepsilon,B} \in \mathcal{N}(d, N_1, \ldots, N_{L-1}) \) such that

\[ |R_{\varepsilon}^{[-B,B]^d}(\Phi_{\varepsilon,B})(x) - x_1| \leq \delta, \text{ for all } x \in [-B, B]^d. \]

In particular, this implies because of \( \delta \leq 1 \) that \( R_{\varepsilon}^{[-B,B]^d}(\Phi_{\varepsilon,B})(x) \in [-B - 1, B + 1] \) for all \( x \in [-B, B]^d \). We conclude that

\[ |\phi(R_{\varepsilon}^{[-B,B]^d}(\Phi_{\varepsilon,B})(x)) - \phi(x_1)| \leq \varepsilon, \text{ for all } x \in [-B, B]^d. \]  \hspace{1cm} (B.2)

Clearly, (B.2) implies that there exists a neural network \( \Phi_{\varepsilon,B} \in \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1) \) such that

\[ |R_{\varepsilon}^{[-B,B]^d}(\Phi_{\varepsilon,B})(x_1, 0, \ldots, 0) - \phi(x_1)| \leq \varepsilon, \text{ for all } x_1 \in [-B, B]. \]  \hspace{1cm} (B.3)

From (B.3) we conclude

\[ \{ f \in [-B, B] : f \in \text{span} \left\{ [-B, B] \ni x \mapsto x^{k_i} e^{\lambda_i x} : i = 1, \ldots, r \right\}, \]

21
where the closure is taken with respect to the sup norm, and where we implicitly used that the space on the right-hand side is a closed subspace of $C([-B,B])$, since it is a finite dimensional subspace.

We have thus shown for every $B > 0$ that there are coefficients $a_i^{(B)} \in \mathbb{R}$ for $i \in \{1,\ldots,r\}$ such that $g(x) = \sum_{i=1}^{r} a_i^{(B)} x^k_i e^{\lambda_i x}$ for all $x \in [-B,B]$. We claim that this implies $g(x) = \sum_{i=1}^{r} a_i^{(1)} x^k_i e^{\lambda_i x}$ for all $x \in \mathbb{R}$, which then completes the proof. To see this, let $B > 1$. Then, we have

$$
\sum_{i=1}^{r} a_i^{(1)} x^k_i e^{\lambda_i x} = g(x) = \sum_{i=1}^{r} a_i^{(B)} x^k_i e^{\lambda_i x}
$$

for all $x \in [-1,1]$. But since both sides of this identity are analytic functions on all of $\mathbb{R}$, they coincide on the whole real line. In particular for $|x| \leq B$, $g(x) = \sum_{i=1}^{r} a_i^{(B)} x^k_i e^{\lambda_i x} = \sum_{i=1}^{r} a_i^{(1)} x^k_i e^{\lambda_i x}$ and thus $g(x) = \sum_{i=1}^{r} a_i^{(1)} x^k_i e^{\lambda_i x}$ for all $x \in \mathbb{R}$.

\section{Proof of Proposition 3.7}

We only prove the first part of the proposition, since the second part follows analogously by replacing $\mathcal{RNN}_{d,L,N,\varnothing}^K$ by $\mathcal{RNN}_{d,L,N,\varnothing}^K(d,N_1,\ldots,N_{L-1},1)$ and some straightforward adaptations.

Since $\mathcal{RNN}_{d,L,N,\varnothing}^K$ is convex and $\mathcal{RNN}_{d,L,N,\varnothing}^K$ is closed under scalar multiplication, we conclude that $\mathcal{RNN}_{d,L,N,\varnothing}^K$ forms a closed linear subspace of $C(K)$. Assume towards a contradiction that we have $\mathcal{RNN}_{d,L,N,\varnothing}^K \subseteq C(K)$. Then, we conclude by the theorem of Hahn-Banach that there exists $h \in C(K)'$ such that $h \neq 0$ but $h$ vanishes on $\mathcal{RNN}_{d,L,N,\varnothing}^K$.

By the representation theorem of Riesz (see \cite[Theorem 6.19]{40}), we conclude that there exists a finite, signed, regular Borel measure $\mu$ such that

$$
h(f) = \int_K f(x)d\mu \quad \text{for all } f \in C(K).
$$

By assumption, $\mathcal{RNN}_{d,L,N,\varnothing}^K$ contains a discriminatory function $g$ with respect to $K$. This easily implies $g(A \cdot + b)|_K \in \mathcal{RNN}_{d,L,N,\varnothing}^K$ for all $A \in \mathbb{R}^{1 \times d}, b \in \mathbb{R}$. We conclude with Definition 3.6 and the fact that $h|_{\mathcal{RNN}_{d,L,N,\varnothing}^K} = 0$, that $\mu = 0$, and hence $h = 0$ on $C(K)$, which is a contradiction.

\section{Proof of Proposition 3.8}

We only prove the first part of the proposition, since the second part follows analogously by replacing $\mathcal{RNN}_{d,L,N,\varnothing}^K$ by $\mathcal{RNN}_{d,L,N,\varnothing}^K(d,N_1,\ldots,N_{L-1},1)$ and some straightforward adaptations.

It is well known, that the intersection of convex sets is convex. Thus, if we assume that $\mathcal{RNN}_{d,L,N,\varnothing}^K$ is $\varepsilon$-convex for all $\varepsilon > 0$ we conclude that

$$
\text{co} \left( \mathcal{RNN}_{d,L,N,\varnothing}^K \right) \subset \bigcap_{\varepsilon > 0} \left( \mathcal{RNN}_{d,L,N,\varnothing}^K + B_{\varepsilon}(0) \right) = \mathcal{RNN}_{d,L,N,\varnothing}^K
$$

where the last identity holds true, since if $\bar{f} \notin \mathcal{RNN}_{d,L,N,\varnothing}^K$, there exists $\varepsilon' > 0$ such that $\| \bar{f} - f \|_{\sup} > \varepsilon'$ for all $f \in \mathcal{RNN}_{d,L,N,\varnothing}^K$. Proposition 3.2 yields a contradiction to (B.4), which shows that $\mathcal{RNN}_{d,L,N,\varnothing}^K$ cannot be $\varepsilon$-convex for all $\varepsilon > 0$.

We conclude that there exists $\varepsilon_0$ such that $\mathcal{RNN}_{d,L,N,\varnothing}^K$ is not $\varepsilon_0$-convex. As a result, there exist $g \in \text{co}(\mathcal{RNN}_{d,L,N,\varnothing}^K)$ such that $\| g - f \|_{\sup} \geq \varepsilon_0$ for all $f \in \mathcal{RNN}_{d,L,N,\varnothing}^K$. Let $\varepsilon > 0$, then we have that $\frac{\varepsilon}{\varepsilon_0} g \in \text{co}(\mathcal{RNN}_{d,L,N,\varnothing}^K)$ since $\mathcal{RNN}_{d,L,N,\varnothing}^K$ is closed under scalar multiplication. Moreover,

$$
\left\| \frac{\varepsilon}{\varepsilon_0} g - f \right\|_{\sup} \geq \varepsilon, \quad \text{for all } f \in \mathcal{RNN}_{d,L,N,\varnothing}^K.
$$
B.7 Proof of Proposition 3.9

For $\ell = 1, \ldots, L - 1$, let $f_\ell := f \circ f \circ \cdots \circ f$, where we take $\ell$ compositions. Since $f$ is strictly monotonically increasing on $[0,1]$ and since we have $f(x) \in [0,1]$ for all $x \in \mathbb{R}$, it follows by a simple induction that $f_\ell(1) - f_\ell(0) > 0$, for each $\ell = 1, \ldots, L - 1$.

Let $\Phi \in \mathcal{NN}(1, \tilde{N}, 1)$ such that $\mathbb{R}^R_\Phi(\Phi) = f$. Then $\tilde{\Phi} := \Phi \circ \cdots \circ \Phi \in \mathcal{NN}(1, \tilde{N}, \ldots, \tilde{N}, 1)$ where we take $L - 1$ concatenations. We have that $\mathbb{R}^R_\Phi(\Phi) = f_{L-1}$. Consider the neural network

$$\Phi^1 = (A, b) \in \mathcal{NN}_{1,1,2},$$

where $A = \frac{1}{f_{L-2}(1) - f_{L-2}(0)}$ and $b = \frac{-f_{L-2}(0)}{f_{L-2}(1) - f_{L-2}(0)}$. Then the network $\Phi^1 \circ \tilde{\Phi} \in \mathcal{NN}(1, \tilde{N}, \ldots, \tilde{N}, 1)$ satisfies that

$$\tilde{f} := \mathbb{R}^{\varphi_2}_\phi(\Phi^1 \circ \tilde{\Phi}) = \frac{f_{L-1}(\cdot) - f_{L-2}(0)}{f_{L-2}(1) - f_{L-2}(0)}.$$

Since $f$ is continuous and sigmoidal, $\tilde{f}$ is sigmoidal. Furthermore, by Lemma 2.5, we have that

$$\mathbb{R}^{\varphi_2}_\phi(1, \tilde{N}, \ldots, \tilde{N}, 1) \subset \mathbb{R}^{\varphi_2}_\phi(1, \tilde{N}, \ldots, \tilde{N}, 1)$$

for every $N' \geq (L-1)\tilde{N} + 2$. □

B.8 Proof of Corollary 3.10

Let $\mathcal{Y}$ be an infinite dimensional topological vector space such that $C(K) \subset \mathcal{Y}$. Assume towards a contradiction that there exists a nonempty open set $U \subset \mathcal{Y}$ such that $U \subset \mathbb{R}^{\varphi_2}_{\Phi_{d,L,N,\vartheta}}$

As a consequence of Lemma 2.7, we have

$$U - U := \{ x - y : x, y \in U \} \subset \mathbb{R}^{\varphi_2}_{\Phi_{d,L,N,\vartheta}}.$$

Moreover, $U - U$ is an open neighborhood of 0 and therefore $U - U$ is absorbing (see [11] Definition 1.33). By Proposition 3.2 we see that $\mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$ is closed under scalar multiplication, and since $U - U$ is absorbing this yields $C(K) \subset \mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}} \subset C(K)$. Therefore, $\mathcal{Y} = C(K) = \mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$.

Since $\mathcal{Y}$ is infinite dimensional, we arrive at a contradiction after invoking Theorem 3.3. □

B.9 Proof of Proposition 3.11

Assume towards a contradiction that $\mathbb{R}^{\varphi_2}_{\Phi_{d,L,N,\vartheta}}$ is not nowhere dense in $\mathcal{Y}$, i.e., that there exists a nonempty open set $U \subset \mathcal{Y}$ such that $U \subset \mathbb{R}^{\varphi_2}_{\Phi_{d,L,N,\vartheta}}$, where the closure is taken in $\mathcal{Y}$. By Lemma 2.7, we have that $U - U \subset \mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$. Moreover, $U - U$ is an open neighborhood of 0 and thus absorbing. By the closedness of $\mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$ under scalar multiplication we also conclude that the set $\mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$ is closed under scalar multiplication, and thus $C(K) \subset \mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$. Hence $\mathbb{R}^{\varphi_2}_{\Phi_{d,L,2N-d-1,\vartheta}}$ is dense in $\mathcal{Y}$. □

C Proofs of the results in Section 4

C.1 Proof of Theorem 4.1

We set $K := [-B, B]^d$. We show that for $g$ as in the statement of the theorem there exists a sequence of networks in $\mathbb{R}^{\varphi_2}_{\Phi_{d,N_1,\ldots,N_l-2,2,1}}$, where $N_i = 1$ for all $i = 1, \ldots, L - 2$, converging to a discontinuous
limit point. Since
\[ \bigcup_{N \geq L+d} R^K_{d,L,N,\varepsilon} \subset C(K) \]
this yields with Lemma 2.9 that for no \( N_1, \ldots, N_{L-1} \in \mathbb{N} \), \( N_{L-1} \geq 2 \) we have that \( R^K_{d,L,N,\varepsilon} \) is closed. Additionally, for no \( N > d + L \) we have that \( R^K_{d,L,N,\varepsilon} \) is closed. Finally, we also get that
\[ \bigcup_{N \geq L+d} R^K_{d,L,N,\varepsilon} \]
is never closed.

For \( \varepsilon > 0 \), choose—according to Proposition 2.9—a neural network \( \Phi \in \mathcal{NN}(d,L-1,L+d-1) \) such that

1. \( |R^K_{\varepsilon}(\Phi)(x) - x_1| \leq \varepsilon \) for all \( x \in K \),
2. \( R^K_{\varepsilon}(\Phi)(0) = 0 \),
3. \( R^K_{\varepsilon}(\Phi) \) is differentiable and \( \frac{\partial R_{\varepsilon}(\Phi)}{\partial x_1}(0) = 1 \), and
4. \( R^K_{\varepsilon}(\Phi) \) is constant in all but the \( x_1 \)-direction.

Let \( J := R^K_{\varepsilon}(\Phi) \). Since \( g \) is continuous and monotonically increasing, Proposition 2.9 yields that \( J \) is continuous and monotonically increasing with respect to the first coordinate. Combining this estimate with (2), (3), and (4) implies
\[ J(x) < 0 \quad \text{for all } x \in K \text{ with } x_1 < 0, \quad \text{and } J(x) > 0 \quad \text{for all } x \in K \text{ with } x_1 > 0. \] (C.1)

We now distinguish the cases given in Assumption (iii)(a) and (b). First we assume that \( \lambda \neq \lambda' \). For \( n \in \mathbb{N} \) let \( \Phi_n = ((A^n_1, b^n_1), (A^n_2, b^n_2)) \in \mathcal{NN}(1,2,1) \) be given by
\[ A^n_1 = \begin{pmatrix} n \\ n \end{pmatrix} \in \mathbb{R}^{2 \times 1}, \quad b^n_1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \in \mathbb{R}^2, \quad A^n_2 = (1 \quad -1) \in \mathbb{R}^{1 \times 2}, \quad b^n_2 = 0 \in \mathbb{R}^1. \]

Then \( \Phi_n \Phi \in \mathcal{NN}(d,N_1,\ldots,N_{L-2},2,1) \) for \( N_i = 1 \) for \( i = 1, \ldots, L-2 \) and we define
\[ h_n(x) := R^K_{\varepsilon}(\Phi_n \Phi)(x) = g(nJ(x)) - g(nJ(x) - 1), \quad \text{for } x \in K. \]
Then, since \( h_n \) is continuous, we have that \( h_n \in L^p(K) \) for every \( n \in \mathbb{N} \) and all \( p \in (0, \infty] \).

Let \( x \in K \) such that \( x_1 > 0 \). Since by (C.1) we have \( J(x) > 0 \), there exists some \( N_x \in \mathbb{N} \) such that for all \( n \geq N_x \), there holds \( nJ(x) > 1 > r \). Hence, by the mean value theorem, there exists some \( \xi^x_n \in [nJ(x) - 1, nJ(x)] \) such that
\[ \lim_{n \to \infty} h_n(x) = \lim_{n \to \infty} g'(\xi^x_n) = \lambda, \]
since \( \xi^x_n \to \infty \) as \( n \to \infty \), \( n \geq N_x \). Analogously, it follows for \( x \in K \) with \( x_1 < 0 \) that
\[ \lim_{n \to \infty} h_n(x) = \lambda', \quad \text{as } n \to \infty. \]

Hence, we get for each \( x \in K \) that
\[ \lim_{n \to \infty} h_n(x) = \left( \lambda \cdot \chi([0,B] \times [-B,B]^{d-1}) + (g(0) - g(1)) \cdot \chi([-B,B]^{d-1}) + \lambda' \cdot \chi([-B,0) \times [-B,B]^{d-1}) \right)(x) =: h(x). \]

We now claim that there is some \( M > 0 \) with \( |g(x) - g(x-1)| \leq M \) for all \( x \in \mathbb{R} \). To see this, note because of \( g'(x) \to \lambda \) as \( x \to \infty \) and because of \( g'(x) \to \lambda' \) as \( x \to -\infty \) that there are \( R, M_0 > 0 \) with \( |g'(x)| \leq M_0 \) for all \( x \in \mathbb{R} \) with \( |x| \geq R \). Hence, \( g \) is \( M_0 \)-Lipschitz on \(( -\infty, -R] \cup [R, \infty) \), so that \( |g(x) - g(x-1)| \leq M_0 \).
for all \( x \in \mathbb{R} \) with \( |x| \geq R + 1 \). But by continuity and compactness, we also have \( |\varrho(x) - \varrho(x - 1)| \leq M_1 \) for all \( |x| \leq R + 1 \) and some constant \( M_1 > 0 \). Thus, we can simply choose \( M := \max\{M_0, M_1\} \).

By what was shown in the preceding paragraph, we get \( |h_n| \leq M \) for all \( n \in \mathbb{N} \). Hence, by the Lebesgue Dominated Convergence Theorem, for all \( p \in (0, \infty) \) we have

\[
\lim_{n \to \infty} \|h_n - h\|_{L^p(K)} = 0.
\]

But since \( \lambda \neq \lambda' \), it is not hard to see that \( h \) has no continuous representative (with respect to equality almost everywhere). This yields the required non-continuity of a limit point as discussed at the beginning of the proof.

We now consider the case that \( \varrho \) is bounded. Since \( \varrho \) is also increasing, there exist \( c, c' \in \mathbb{R} \) such that

\[
\lim_{x \to \infty} \varrho(x) = c, \quad \text{and} \quad \lim_{x \to -\infty} \varrho(x) = c'.
\]

Since \( \varrho \) is monotonically increasing and not constant, we have \( c > c' \).

For each \( n \in \mathbb{N} \), we now consider the neural network \( \tilde{\varphi}_n \in \mathcal{N} \), given by \( \tilde{\varphi}_n = \left( (A_1^n, b_1^n), (A_2^n, b_2^n) \right) \), where

\[
A_1^n = n \in \mathbb{R}^{1 \times 1}, \quad b_1^n = 0 \in \mathbb{R}^1, \quad A_2^n = 1 \in \mathbb{R}^{1 \times 1}, \quad b_2^n = 0 \in \mathbb{R}^1.
\]

Then \( \tilde{\varphi}_n \in \mathcal{N} \), where \( \mathcal{N} = \{ (d, N_1, \ldots, N_{L-2}, N_{L-1}, 1) \mid N_i = 1 \text{ for } i = 1, \ldots, L - 1 \} \), and we define

\[
\tilde{\varphi}_n(x) := R^K_{\varrho} \left( \tilde{\varphi}_n \right)(x) = \varrho(nJ(x)), \quad \text{for } x \in K.
\]

Since each of the \( \tilde{\varphi}_n \) is continuous, we have \( \tilde{\varphi}_n \in L^p(K) \) for all \( p \in (0, \infty] \). Equation (C.1) implies that \( J(x) > 0 \) for all \( x \in K \) with \( x_1 > 0 \). This in turn yields that

\[
\lim_{n \to \infty} \tilde{\varphi}_n(x) = c, \quad \text{for all } x \in K \text{ such that } x_1 > 0.
\]

Similarly, the fact that \( J(x) < 0 \) for all \( x \in K \) with \( x_1 < 0 \) yields

\[
\lim_{n \to \infty} \tilde{\varphi}_n(x) = c', \quad \text{for all } x \in K \text{ such that } x_1 < 0.
\]

Combining (C.2) with (C.3) yields for all \( x \in K \) that

\[
\lim_{n \to \infty} \tilde{\varphi}_n(x) = (c \cdot \chi_{\{0, B\} \times [-B, B]}^{d-1} + \varrho(0) \cdot \chi_{\{0\} \times [-B, B]}^{d-1} + c' \cdot \chi_{[-B, 0) \times [-B, B]}^{d-1}) (x) =: \tilde{h}(x).
\]

By the boundedness of \( \varrho \), we get \( |\tilde{\varphi}_n(x)| \leq C \) for all \( n \in \mathbb{N} \) and \( x \in K \) and a suitable \( C > 0 \). Together with the Lebesgue Dominated Convergence Theorem, this implies for all \( p \in (0, \infty) \)

\[
\lim_{n \to \infty} \left\| \tilde{\varphi}_n - \tilde{\varphi} \right\|_{L^p(K)} = 0.
\]

Since \( c \neq c' \), \( \tilde{\varphi} \) does not have a continuous representative (with respect to equality almost everywhere). This yields the required non-continuity of a limit point as discussed at the beginning of the proof.

\[\square\]

### C.2 Proof of Theorem 4.2

Set \( K := [-B, B]^d \). Let \( m \in \mathbb{N} \) be such that \( \varrho \in C^m(\mathbb{R}) \setminus C^{m+1}(\mathbb{R}) \). Then

\[
\mathcal{RN}\mathcal{N}_d \cap C^m(K) \subset C^m(K).
\]

We now consider the smallest possible architecture \( (d, N_1, \ldots, N_{L-2}, 2, 1) \), where \( N_i = 1 \) for all \( i = 1, \ldots, L - 2 \) that the set \( \mathcal{RN}\mathcal{N}_d (d, N_1, \ldots, N_{L-2}, 2, 1) \) is not closed in \( C(K) \). By assumption, there exists some \( \lambda > 0 \)
with \( g(\lambda) \in C^m([-B, B]) \setminus C^{m+1}([-B, B]) \). Moreover, since the continuous derivative \( g' \) is bounded on the compact set \([-B, B]\), we see that \( g(\lambda) \) is Lipschitz continuous on \([-B, B]\), and we denote \( \text{Lip}(g(\lambda)) := M_1 \).

Next, by the uniform continuity of \( \lambda \cdot g'(\lambda) \) on \((-B + 1, B + 1)\), if we set

\[
\varepsilon_n := \sup_{x, y \in (-B + 1, B + 1)} \frac{|\lambda \cdot g'(\lambda x) - \lambda \cdot g'(\lambda y)|}{|x - y|^{1/n}}
\]

then \( \varepsilon_n \to 0 \) as \( n \to \infty \).

For \( n \in \mathbb{N} \) let \( \Phi_n = ((A_1^n, b_1^n), (A_2^n, b_2^n)) \in \mathcal{N}\mathcal{N}(1, 2, 1) \) be given by

\[
A_1^n = \left( \frac{\lambda}{n} \right) \in \mathbb{R}^{2 \times 1}, \quad b_1^n = \left( \frac{\lambda}{n} \right) \in \mathbb{R}^2, \quad A_2^n = (n - n) \in \mathbb{R}^{1 \times 2}, \quad b_2^n = 0 \in \mathbb{R}^1.
\]

For \( n \in \mathbb{N} \) choose, according to Proposition 2.3, a neural network \( \Phi_n = \mathcal{N}\mathcal{N}_{d,L-1,L+d-1} \) such that

\[
|\mathcal{R}_\theta^K(\Phi_n^2)(x) - x_1| \leq \frac{1}{2n^2}, \quad \text{for all } x \in K.
\]  

(C.4)

We set \( \Phi_n = \Phi_n^1 \bullet \Phi_n^2 \in \mathcal{N}\mathcal{N}(d, 1, \ldots, 1, 2, 1) \). For \( x \in K \), we then have

\[
|\mathcal{R}_\theta^K(\Phi_n)(x) - \lambda g'(\lambda x_1)| = |n \left( g(\lambda \mathcal{R}_\theta^K(\Phi_n^1)(x) + \lambda \cdot n^{-1}) - g(\lambda \mathcal{R}_\theta^K(\Phi_n^2)(x)) \right) - \lambda g'(\lambda x_1)|.
\]

By the Lipschitz continuity of \( g(\lambda) \) and Equation (C.4) we conclude that

\[
|n \left( g(\lambda \mathcal{R}_\theta^K(\Phi_n^1)(x) + \lambda \cdot n^{-1}) - g(\lambda \mathcal{R}_\theta^K(\Phi_n^2)(x)) \right) - n \left( g(\lambda (x_1 + n^{-1})) - g(\lambda x_1) \right)| \leq \frac{M_1}{n}.
\]

This implies for every \( x \in K \) that

\[
|\mathcal{R}_\theta^K(\Phi_n)(x) - \lambda g'(\lambda x_1)| \leq \frac{M_1}{n}. \tag{by the mean value theorem, \( \xi_n \in (x_1 - n^{-1}, x_1) \) and \( x_1, \xi_n \in [-B + 1, B + 1] \)}
\]

Here, the last step used that \( |\xi_n - x_1| \leq n^{-1} \leq 1 \), so that \( x_1, \xi_n \in [-B + 1, B + 1] \).

Overall, we thus showed existence of a sequence in \( \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, 1, \ldots, 1, 2, 1) \) which converges uniformly to the function \( K \to \mathbb{R}, \ x \mapsto \rho(\lambda) := \lambda g'(\lambda x_1) \). By assumption we have that \( \rho \lambda \notin C^m(K) \). Since \( \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, 1, \ldots, 1, 2, 1) \subset C^m(K) \) we can conclude that \( \rho \lambda \notin \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, 1, \ldots, 1, 2, 1) \) and hence the set \( \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, 1, \ldots, 1, 2, 1) \) is not closed in \( C(K) \). Now, let \( N_1, \ldots, N_{L-1} \in \mathbb{N} \) be arbitrary such that \( N_{L-1} \geq 2 \).

Since by Lemma 2.3 we have that

\[
\mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, 1, \ldots, 1, 2, 1) \subset \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, N_1, \ldots, N_L, 1),
\]

also \( \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, N_1, \ldots, N_L, 1) \) is not closed in \( C(K) \). Finally, using a similar argument, for any \( N > L + d + 1 \) the set \( \mathcal{R}\mathcal{N}\mathcal{N}_{d,L,N,\theta}^K \) is not closed in \( C(K) \).

\[\square\]

C.3 Proof of Theorem 4.4

Let \( K := [-B, B]^d \). We first show that for \( (d, N_1, \ldots, N_{L-2}, 2, 1) \), where \( N_i = 1 \) for all \( i = 1, \ldots, L - 2 \), there exists a limit point of \( \mathcal{R}\mathcal{N}\mathcal{N}_{\theta}^K(d, N_1, \ldots, N_{L-2}, 2, 1) \) that is unbounded.

Since \( \rho \) is not constant, there is some \( x_0 \in \mathbb{R} \) such that \( \rho'(x_0) \neq 0 \). For \( n \in \mathbb{N} \), let us define \( \Phi_n^1 := ((A_1^n, b_1^n), (A_2^n, b_2^n)) \in \mathcal{N}\mathcal{N}(1, 2, 1) \) by

\[
A_1^n := \left( \frac{1}{n} \right) \in \mathbb{R}^{2 \times 1}, \quad b_1^n := \left( \frac{0}{x_0} \right) \in \mathbb{R}^2, \quad A_2^n := (1 \ n) \in \mathbb{R}^{1 \times 2}, \quad b_2^n := -\rho(x_0)n \in \mathbb{R}.
\]
With this choice we have
\[ R^p_n(\Phi^1_n)(x) = \varrho(x) + n \cdot (\varrho(x/n + x_0) - \varrho(x_0)), \quad \text{for all } x \in \mathbb{R}. \tag{C.5} \]

By the mean-value theorem, it is not hard to see for \( B > 0 \) and \( x \in [-B, B] \) that
\[ |R^p_n(\Phi^1_n)(x) - (\varrho(x) + \varrho'(x_0)x)| \leq B|\varrho'(\bar{x}) - \varrho'(x_0)|, \quad \text{for some } \bar{x} \in [x_0 - B/n, x_0 + B/n]. \]

Since \( \varrho' \) is continuous in \( x_0 \) we conclude that
\[ \sup_{x \in [-B, B]} R^p_n(\Phi^1_n)(x) - (\varrho(x) + \varrho'(x_0)x) \xrightarrow{n \to \infty} 0. \tag{C.6} \]

Moreover, note that \( \frac{d}{dx} R^p_n(\Phi^1_n)(x) = \varrho'(x) + \varrho'(x_0 + n^{-1}x) \) is bounded on \([-B + 1, B + 1]\), uniformly with respect to \( n \in \mathbb{N} \). Hence, \( R^p_n(\Phi^1_n) \) is Lipschitz continuous on \([-B - 1, B + 1]\), with Lipschitz constant \( C' > 0 \) independent of \( n \in \mathbb{N} \).

For \( n \in \mathbb{N} \) we choose with Proposition 2.9 a neural network \( \Phi^2_n \in \mathcal{N}_{d,L-1,L+d-1} \) such that
\[ |R^K_n(\Phi^2_n)(x) - x_1| \leq \frac{1}{n}, \quad \text{for all } x \in K. \tag{C.7} \]

We set \( \Phi_n = \Phi^1_n \cdot \Phi^2_n \in \mathcal{N}(d, 1, \ldots, 2, 1) \) and conclude with (C.5) for all \( x \in K \) that
\[ |R^p_n(\Phi_n)(x) - (\varrho(x_1) + \varrho'(x_0)x_1)| = |R^p_n(\Phi^1_n)(R^K_n(\Phi^2_n)(x)) - (\varrho(x_1) + \varrho'(x_0)x_1)|. \]

By the Lipschitz continuity of \( R^p_n(\Phi^1_n) \) on \([-B - 1, B + 1]\) and (C.7) we conclude that
\[ |R^K_n(\Phi_n)(x) - (\varrho(x_1) + \varrho'(x_0)x_1)| \leq |R^K_n(\Phi_n)(x_1) - (\varrho(x_1) + \varrho'(x_0)x_1)| + C', \]

An application of (C.6) yields that
\[ \sup_{x \in K} |R^K_n(\Phi_n)(x) - (\varrho(x_1) + \varrho'(x_0)x_1)| \xrightarrow{n \to \infty} 0. \]

To show that \( \mathcal{R}\mathcal{N}_{d,L-1}(d, N_1, \ldots, N_{L-1}, 1) \) is not closed in \( C(K) \), for all \( N_1, \ldots, N_{L-2} \in \mathbb{N}, N_{L-1} \in \mathbb{N}_{\geq 2} \) it suffices to show that with
\[ F: \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \varrho(x_1) + \varrho'(x_0)x_1, \]
\( F|_K \) is not an element of \( \mathcal{R}\mathcal{N}_{d,L-1}(d, N_1, \ldots, N_{L-1}, 1) \). Similarly, to show that \( \mathcal{R}\mathcal{N}_{d,L,N}_d \) is not closed for any \( N > d + L \) it suffices to show that \( F|_K \) is not an element of \( \mathcal{R}\mathcal{N}_{d,L,N}_d \). Both claims are accomplished, when we show that there do not exist \( N_1, \ldots, N_{L-1} \in \mathbb{N} \) such that \( F|_K \) is an element of \( \mathcal{R}\mathcal{N}_{d,L,N}_d \).

Towards a contradiction, we assume that there exist \( N_1, \ldots, N_{L-1} \in \mathbb{N} \) such that \( F|_K = R^K_n(\Phi^n) \) for a network \( \Phi^n \in \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1) \). Since \( F \) and \( R^K_n(\Phi^n) \) are both analytic functions that coincide on \( K \), they must be equal. However, \( F \) is unbounded (since \( \varrho \) is bounded, and since \( \varrho'(x_0) \neq 0 \)), while \( R^K_n(\Phi^n) \) has to be bounded by the boundedness of \( \varrho \). This produces the desired contradiction and shows that \( \mathcal{R}\mathcal{N}_{d,L,N}_d \) is not closed in \( C(K) \).

\[ \square \]

### C.4 Proof of Theorem 4.5

Let \( \varrho \in C^{\max\{p,q\}}(\mathbb{R}) \) be approximately homogeneous of order \((p,q)\). For simplicity, let us assume that \( p > q \); the case \( q > p \) can be handled similarly. Let \( (x)_+ := \max\{x, 0\} \). We start by showing that
\[ k^{-p} \varrho(k) \to (\cdot)^p_q \tag{C.8} \]
uniformly on $[-B, B]$, as $k \to \infty$. Let $r > 0$ such that $|g(x) - x^p| \leq r$ for all $x > 0$ and $|g(x) - x^q| \leq r$ for all $x < 0$. We have for any $k \in \mathbb{N}$ that
\[
\sup_{x \in [-B, B]} |k^{-p}g(kx) - (x)^p_k| = \max \left\{ \sup_{x \in [-B, 0]} |k^{-p}g(kx)|, \sup_{x \in [0, B]} |k^{-p}g(kx) - x^p| \right\}.
\]
By assumption, we have that
\[
\sup_{x \in [-B, 0]} |k^{-p}g(kx)| \leq k^{-p}((kB)^q + r) \leq c_0k^{-1}
\]
for a constant $c_0 = c_0(B, r) > 0$. Moreover,
\[
\sup_{x \in [0, B]} |k^{-p}g(kx) - x^p| = \sup_{x \in [0, B]} |k^{-p}(g(kx) - (kx)^p)| \leq rk^{-p}.
\]
Overall, we conclude that
\[
\sup_{x \in [-B, B]} |k^{-p}g(kx) - (x)^p_k| \leq \max\{c_0, r\}k^{-1},
\]
which implies (C.8). We observe that $(x)^p_k \notin C^p([-B, B])$. Additionally, for $d \in \mathbb{N}$ and $L \in \mathbb{N}_{\geq 2}$ with $N > L + d$, we have that $\mathcal{R}\mathcal{N}^d_{d, L, N, \varrho} \subset C^p([-B, B]^d)$. Moreover, for any fixed neural network architecture $(d, N_1, \ldots, N_{L-1}, 1)$, the set $\mathcal{R}\mathcal{N}_{d, L, N, \varrho}^d(d, N_1, \ldots, N_{L-1}, 1) \subset C^p([-B, B]^d)$. Hence the proof is complete if we can construct a sequence of neural networks, the $\varrho$ realizations of which uniformly converge to the function $[-B, B]^d \to \mathbb{R}, x \mapsto (x)^p_k$. By the preceding considerations, this is clearly possible, as can be seen by the same arguments used in the proofs of the previous results. □

## C.5 Proof of Proposition 4.6

The set $\Theta_C \cap \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1)$ is bounded and closed in the normed space
\[
(\mathcal{N}(d, N_1, \ldots, N_{L-1}, 1), \| \cdot \|_{\mathcal{N}(d, N_1, \ldots, N_{L-1})}).
\]

The Heine-Borel Theorem implies the compactness of $\Theta_C$. By Proposition 5.1, the map
\[
R^K_{\varrho} : (\mathcal{N}(d, N_1, \ldots, N_{L-1}, 1), \| \cdot \|_{\mathcal{N}(d, N_1, \ldots, N_{L-1}, 1)}) \to (C(K), \| \cdot \|_{\sup})
\]
is continuous. As a consequence, the set $R^K_{\varrho}(\Theta_C \cap \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1))$ is compact in $C(K)$. Since
\[
\Theta_C = \bigcup_{N_1, \ldots, N_{L-1} \in \mathbb{N}}^{N_1, \ldots, N_{L-1} \in \mathbb{N}} \Theta_C \cap \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1),
\]
we conclude that the set $R^K_{\varrho}(\Theta_C)$ is the finite union of compact sets in $C(K)$ and therefore compact itself. Because of the compactness of $K$, $C(K)$ is continuously embedded into $L^p(K)$ for every $p \in (0, \infty)$. This implies that the sets $R^K_{\varrho}(\Theta_C \cap \mathcal{N}(d, N_1, \ldots, N_{L-1}, 1))$ and also $R^K_{\varrho}(\Theta_C)$ are compact in $L^p(K)$. □

## C.6 Proof of Proposition 4.7

The main trick in the proof will be to show that one can replace a given sequence of networks with $C$-bounded scaling weights by another sequence with $C$-bounded scaling weights that also has bounded biases. Then one can apply Proposition 4.6.
Lemma C.1. Let $K \subset \mathbb{R}^d$ be measurable and bounded and of positive measure. Let $d, N_L, L, N \in \mathbb{N}$, and let $C > 0$. Finally, let $\varrho : \mathbb{R} \to \mathbb{R}$, $x \mapsto \max\{0, x\}$ denote the ReLU.

Let $(\Phi_n)_{n \in \mathbb{N}}$ be a sequence of networks in $\mathcal{N}_d, L, N, N_L$ with C-bounded scaling weights and such that $\|R^\varrho(\Phi_n)\|_{L^1(K)} \leq M$ for all $n \in \mathbb{N}$ and some $M > 0$.

Then there is an infinite set $I' \subset \mathbb{N}$ and a family of networks $(\Psi_n)_{n \in I'}$ with C-bounded weights which satisfies $R^\varrho(\Phi_n) = R^\varrho(\Psi_n)$ for $n \in I$ and such that $\|\Psi_n\|_{\text{total}} \leq C'$ for all $n \in I$ and a suitable constant $C' > 0$.

Proof. Since $K$ is bounded, there is some $R > 0$ with $\|x\|_\infty \leq R$ for all $x \in K$. Below, we will use without further comment the estimate $\|Ax\|_\infty \leq k \cdot \|A\|_{\text{max}} \cdot \|x\|_\infty$ which is valid for $A \in \mathbb{R}^{n \times k}$ and $x \in \mathbb{R}^k$. Let $\Phi = ((A_1, b_1), \ldots, (A_L, b_L)) \in \mathcal{N}_d, L, N, N_L$ be a neural network.

Note that $N_j \leq \sum_{\ell=0}^{L} N_\ell = N(\Phi) = N$ for all $j \in \{0, \ldots, L\}$, so that the total number of possible architectures of networks in $\mathcal{N}_d, L, N, N_L$ is finite. Thus, there are suitable $\Phi_0, \ldots, \Phi_L \in \mathbb{N}$ and an infinite subset $I_{L-1} \subset \mathbb{N}$ such that each $\Phi_n$ with $n \in I_{L-1}$ has architecture $S := (N_0, \ldots, N_L)$.

Below, we will show by induction on $m \in \{0, \ldots, L - 1\}$ that for each $m \in \{0, \ldots, L - 1\}$, there is an infinite subset $I_m \subset I_{L-1}$, and a family of networks $(\Psi_n^{(m)})_{n \in I_m} \subset \mathcal{N}_d, L, N, N_L$ with architecture $S$ and

$$\Psi_n^{(m)} := \left( (B_{1}^{(n,m)}, c_1^{(n,m)}), \ldots, (B_{L}^{(n,m)}, c_L^{(n,m)}) \right)$$

with the following properties:

(A) We have $R^\varrho(\Psi_n^{(m)}) = R^\varrho(\Phi_n)$ for all $n \in I_m$.

(B) Each $\Psi_n^{(m)}$ has C-bounded weights.

(C) There is a constant $C_m > 0$ with $\|c^{(n,m)}\|_\infty \leq C_m$ for all $n \in I_m$ and all $\ell \in \{1, \ldots, m\}$.

Once this is shown, we set $I := I_{L-1}$ and $\Psi_n := \Psi_n^{(L-1)}$ for $n \in I$. Clearly, $\Psi_n$ has C-bounded scaling weights and satisfies $R^\varrho(\Phi_n) = R^\varrho(\Psi_n)$, so that it remains to show $\|\Psi_n\|_{\text{total}} \leq C'$, for which it suffices to show $|c^{(n,L-1)}_L| \leq C''$ for some $C'' > 0$ and all $n \in I$, since we have $|c^{(n,L-1)}_L| \leq C_m$ for all $\ell \in \{1, \ldots, L-1\}$.

Now, note for $\ell \in \{1, \ldots, L - 1\}$ and $x \in \mathbb{R}^{\ell-1}$ that $T^{(n,L-1)}_\ell x := B^{(n,L-1)}_{\ell} x + c^{(n,L-1)}_\ell$ satisfies

$$\|T^{(n,L-1)}_\ell x\|_{\infty} \leq N_{\ell-1} \cdot C \cdot \|x\|_{\infty} + C_L - 1.$$

Since $K$ is bounded, and since $|\varrho(x)| \leq |x|$ for all $x \in \mathbb{R}$, there is thus a constant $C'_{L-1} > 0$ such that if we set

$$\beta^{(n)}(x) := (\varrho \circ T^{(n,L-1)}_{L-1} \circ \cdots \circ \varrho \circ T^{(n,L-1)}_{1})(x) \quad \text{for} \quad x \in K,$$

then $\|\beta^{(n)}\|_{\infty} \leq C'_{L-1}$ for all $x \in K$ and all $n \in I_{L-1}$.

For arbitrary $i \in \{1, \ldots, N_L\}$ and $x \in K$, this implies

$$\|R^\varrho(\Phi_n)(x)i\| = \|R^\varrho(\Psi_n^{(L-1)})(x)i\| = \|c^{(n,L-1)}_L i\| = \|c^{(n,L-1)}_L\| + \|\beta^{(n)}(x)i\| \geq \|c^{(n,L-1)}_L\| \geq |c^{(n,L-1)}_L| - N_{L-1} \cdot C \cdot \|\beta^{(n)}(x)i\| \geq |c^{(n,L-1)}_L| - N_{L-1} \cdot C \cdot C'_{L-1}.$$

Since by assumption $\|R^\varrho(\Phi_n)\|_{L^1(K)} \leq M$, we see that $(c^{(n,L-1)}_L)_{n \in I_{L-1}}$ must be a bounded sequence.

Thus, it remains to construct the networks $\Psi_n^{(m)}$ for $n \in I_m$ (and the sets $I_m$) for $m \in \{0, \ldots, L - 1\}$ with the properties (A)–(C) from above.

For the start of the induction ($m = 0$), we can simply take $I_0 := I_{L-1}$, $\Psi_n^{(0)} := \Phi_n$, and $C_0 > 0$ arbitrary, since condition (C) is void in this case.
Now, assume that a family of networks $\{\Psi_n^m\}_{n \in I_m}$, as in equation (C.9), with an infinite subset $I_m \subset \mathbb{N}$ and satisfying conditions (A)–(C) have been constructed for some $m \in \{0, \ldots, L - 2\}$. In particular, $L \geq 2$.

For brevity, set $T_{\ell}^{(n)} : \mathbb{R}^{N_{\ell - 1}} \to \mathbb{R}^{N_{\ell}}$, $x \mapsto B_{\ell}^{(n,m)} x + c_{\ell}^{(n,m)}$ for $\ell \in \{1, \ldots, L\}$, and $\varphi_L := id_{\mathbb{R}^L}$, as well as $\varphi_\ell := \varphi^{(n)}_\ell$ for $\ell \in \{1, \ldots, L - 1\}$. Furthermore, let us define $\beta_n := \varphi_m \circ T_{\ell}^{(n)} \circ \cdots \circ \varphi_1 \circ T_1^{(n)} : \mathbb{R}^d \to \mathbb{R}^{N_m}$. Note $|\varphi_\ell(x)| \leq |x|$ for all $x \in \mathbb{R}^{N_\ell}$. Furthermore, observe for $n \in I_m$, $\ell \in \{1, \ldots, m\}$ and $x \in \mathbb{R}^{N_{\ell - 1}}$ that

$$
\left\| \phi_{\ell}^{(n)}(x) \right\|_{\ell \to \infty} = \left\| B_{\ell}^{(n,m)} x + c_{\ell}^{(n,m)} \right\|_{\ell \to \infty} \leq N_{L - 1} \cdot C \cdot |x| + C_m.
$$

Combining these observations, we easily see that there is some $R' > 0$ with $|\beta_n(x)| \leq R'$ for all $x \in K$ and $n \in I_m$.

Next, since $(c^{(n,m)}_{m+1})_{n \in I_m}$ is an infinite family in $\mathbb{R}^{N_{m+1}} \subset [-\infty, \infty]^{N_{m+1}}$, we can find (by compactness) an infinite subset $I_m^{(0)} \subset I_m$ such that $c^{(n,m)}_{m+1} \to c_{m+1} \in [-\infty, \infty]^{N_{m+1}}$ as $n \to \infty$ in the set $I_m^{(0)}$.

Our goal is to construct vectors $d^{(n)}, e^{(n)} \in \mathbb{R}^{N_m}$, matrices $C^{(n)} \in \mathbb{R}^{N_{m+1} \times N_m}$, and an infinite subset $I_{m+1}^{(0)} \subset I_m^{(0)}$ such that $\|C^{(n)}\|_{\max} \leq C$ for all $n \in I_{m+1}$, such that $(d^{(n)})_{n \in I_{m+1}}$ is a bounded family, and such that we have

$$
\varphi_{m+1} \left( T_{m+1}^{(n)} x \right) = \varphi_{m+1} \left( C^{(n)} x + d^{(n)} \right) + e^{(n)} \quad \text{for all } x \in \mathbb{R}^{N_m} \text{ with } |x| \leq R',
$$

for all $n \in I_{m+1}$.

Once $d^{(n)}, e^{(n)}, C^{(n)}$ are constructed, we can simply choose $\Psi_{m+1}^{(n)}$ as in equation (C.9), where we define $B_{\ell}^{(n,m+1)} := B_{\ell}^{(n,m)}$ and $c_{\ell}^{(n,m+1)} := c_{\ell}^{(n,m)}$ for $\ell \in \{1, \ldots, L\} \setminus \{m + 1, m + 2\}$, and finally

$$
B_{m+2}^{(n,m+1)} := C^{(n)}, \quad B_{m+2}^{(n,m)} := B_{m+2}^{(n,m+1)}, \quad c_{m+1}^{(n,m+1)} := d^{(n)}, \quad \text{and } c_{m+2}^{(n,m+1)} := e^{(n)}
$$

for $n \in I_{m+1}$. Indeed, these choices clearly ensure $\|B_{\ell}^{(n,m+1)}\|_{\max} \leq C$ as well as $\|c_{\ell}^{(n,m+1)}\|_{\ell \to \infty} \leq C_{m+1}$ for all $\ell \in \{1, \ldots, L\}$ and $n \in I_{m+1}$, for a suitable constant $C_{m+1} > 0$.

Finally, since $|\beta_n(x)| \leq R'$ for all $x \in K$ and $n \in I_m$, equation (C.10) implies

$$
T_{m+2}^{(n)} \left( \varphi_{m+1} \left( T_{m+1}^{(n)} (\beta_n(x)) \right) \right) = T_{m+2}^{(n)} \left( \varphi_{m+1} \left( C^{(n)} \beta_n(x) + d^{(n)} \right) + e^{(n)} \right)
$$

for $n \in I_{m+1}$. By recalling the definition of $\beta_n$, and by noting that $B_{\ell}^{(n,m+1)}, c_{\ell}^{(n,m+1)}$ are identical to $B_{\ell}^{(n,m)}, c_{\ell}^{(n,m)}$ for $\ell \in \{1, \ldots, L\} \setminus \{m + 1, m + 2\}$, this easily yields

$$
R^K_{\ell} \left( \Psi_{m+1}^{(n)} \right) = R^K_{\ell} \left( \Psi_{m}^{(n)} \right) = R^K_{\ell} \left( \Phi_n \right) \quad \text{for all } n \in I_{m+1}.
$$

Thus, it remains to construct $d^{(n)}, e^{(n)}, C^{(n)}$ for $n \in I_{m+1}$ (and the set $I_{m+1}$ itself) as described around equation (C.10). For $n \in I^{(0)}_m$ and $k \in \{1, \ldots, N_{m+1}\}$, define

$$
d_k^{(n)} := \begin{cases} 
R' \cdot C'm, & \text{if } (c_{m+1})_k = \infty, \\
0, & \text{if } (c_{m+1})_k = -\infty, \\
\left( \left( c_{m+1}^{(n,m)} \right)_k \right), & \text{if } (c_{m+1})_k \in \mathbb{R},
\end{cases}
$$

as well as

$$
C_{k,-}^{(n)} := \begin{cases} 
\left( B_{m+1}^{(n,m)} \right)_{k,-}, & \text{if } (c_{m+1})_k = \infty, \\
0 \in \mathbb{R}^{N_m}, & \text{if } (c_{m+1})_k = -\infty, \\
\left( B_{m+1}^{(n,m)} \right)_{k,-}, & \text{if } (c_{m+1})_k \in \mathbb{R},
\end{cases}
$$

for $n \in I^{(0)}_m$ and $k \in \{1, \ldots, N_{m+1}\}$.
To see that these choices indeed fulfill the conditions outlined around equation (C.10) for a suitable choice of $I_{m+1} \subset I_{m}^{(0)}$, first note that $(d^{(n)})_{n \in I_{m}^{(0)}}$ is indeed a bounded family. Furthermore, $|C_{k}^{(n)}| \leq \|B_{m+1}^{(n,m)}\|_{k,i}$ for all $k \in \{1, \ldots, N_{m+1}\}$ and $i \in \{1, \ldots, N_{m}\}$, which easily implies $\|C^{(n)}\|_{\max} \leq \|B_{m+1}^{(n,m)}\|_{\max} \leq C$ for all $n \in I_{m}^{(0)}$. Thus, it remains to verify equation (C.10) itself. But the estimate $\|B_{m+1}^{(n,m)}\| \leq C$ also implies

$$\left(\left|B_{m+1}^{(n,m)}x\right|\right)_{k} \leq N_{m} \cdot C \cdot \|x\|_{\ell^{\infty}} \leq N_{m} \cdot C \cdot R'$$

for all $k \in \{1, \ldots, N_{m+1}\}$ and all $x \in \mathbb{R}^{N_{m}}$ with $|x| \leq R'$. (C.11)

As a final preparation, note that $\vartheta_{m+1} = \vartheta_{m+1}^{\otimes N_{m+1}}$ is a tensor product of ReLU functions, since $m \leq L - 2$. Now, for $k \in \{1, \ldots, N_{m+1}\}$ there are three cases:

**Case 1:** We have $(c_{m+1})_{k} = \infty$. This implies that there is some $n_{k} \in \mathbb{N}$ such that $(c_{m+1}^{(n,m)})_{k} \geq R' \cdot N_{m} C$ for all $n \in I_{m}^{(0)}$ with $n \geq n_{k}$. In view of equation (C.11), this implies $\left(\left|T_{m+1}^{(n)}(x)\right|\right)_{k} = \left(\left|B_{m+1}^{(n,m)}x + c_{m+1}^{(n,m)}\right|\right)_{k} \geq 0$, and hence

$$\left[\vartheta_{m+1} \left(\left|T_{m+1}^{(n)}(x)\right|\right)\right]_{k} = \left(\left|B_{m+1}^{(n,m)}x + c_{m+1}^{(n,m)}\right|\right)_{k},$$

where the last step used our choice of $d^{(n)}, e^{(n)}, C^{(n)}$, and the fact that $(C^{(n)} x + d^{(n)})_{k} \geq 0$ by equation (C.11).

**Case 2:** We have $(c_{m+1})_{k} = -\infty$. This implies that there is some $n_{k} \in \mathbb{N}$ with $(c_{m+1}^{(n,m)})_{k} \leq -R' \cdot N_{m} C$ for all $n \in I_{m}^{(0)}$ with $n \geq n_{k}$. Because of equation (C.11), this yields $\left(\left|T_{m+1}^{(n)}(x)\right|\right)_{k} = \left(\left|B_{m+1}^{(n,m)}x + c_{m+1}^{(n,m)}\right|\right)_{k} \leq 0$, and hence

$$\left[\vartheta_{m+1} \left(\left|T_{m+1}^{(n)}(x)\right|\right)\right]_{k} = 0 = \left[\vartheta_{m+1} \left(C^{(n)} x + d^{(n)}\right) + e^{(n)}\right]_{k},$$

where the last step used our choice of $d^{(n)}, e^{(n)}, C^{(n)}$.

**Case 3:** We have $(c_{m+1})_{k} \in \mathbb{R}$. In this case, set $n_{k} := 1$, and note by our choice of $d^{(n)}, e^{(n)}, C^{(n)}$ for $n \in I_{m}^{(0)}$ with $n \geq 1 = n_{k}$ that

$$\left[\vartheta_{m+1} \left(C^{(n)} x + d^{(n)}\right) + e^{(n)}\right]_{k} = \left[\vartheta_{m+1} \left(B_{m+1}^{(n,m)}x + c_{m+1}^{(n,m)}\right)\right]_{k} = \left[\vartheta_{m+1} \left(T_{m+1}^{(n)}(x)\right)\right]_{k}.$$

Overall, we have thus shown that equation (C.10) is satisfied for all $n \in I_{m+1}$, where

$$I_{m+1} := \left\{n \in I_{m}^{(0)} : n \geq \max \{n_{k} : k \in \{1, \ldots, N_{m+1}\}\}\right\}$$

is clearly an infinite set, since $I_{m}^{(0)}$ is.

Using Lemma C.1, we can now easily show that the set $\mathcal{R}_{N_{d} L, N_{L} \varrho}^{\mathcal{N}^{K,C}_{d,L,N,L_{\varrho}}} \subset L^{p}(K)$ is closed:

Let $\mathcal{Y}$ denote either $L^{p}(K)$ for some $p \in [1, \infty]$, or $C(K)$. Let $(f_{n})_{n \in \mathbb{N}}$ be a sequence in $\mathcal{R}_{N_{d} L, N_{L} \varrho}^{\mathcal{N}^{K,C}_{d,L,N,L_{\varrho}}}$ which satisfies $f_{n} \to f$ for some $f \in \mathcal{Y}$. Thus, $f_{n} = R_{\varrho}(\Psi_{n})$ for a suitable sequence $(\Phi_{n})_{n \in \mathbb{N}}$ in $\mathcal{N}_{d,L,N,L_{\varrho}}^{K,C}$.

Since $(f_{n})_{n \in \mathbb{N}} = (\Phi_{n}(\Phi_{n}))_{n \in \mathbb{N}}$ is convergent in $\mathcal{Y}$, it is also bounded in $\mathcal{Y}$. But since $K$ is bounded, it is not hard to see $\mathcal{Y} \to L^{1}(K)$, so that we get $\|R_{\varrho}(\Phi_{n})\|_{L^{1}(K)} \leq M$ for all $n \in \mathbb{N}$ and a suitable constant $M > 0$.

Therefore, Lemma C.1 yields an infinite set $I \subset \mathbb{N}$ and a family of networks $(\Psi_{n})_{n \in I}$ with $C$-bounded scaling weights such that $f_{n} = R_{\varrho}(\Psi_{n})$ and $\|\Psi_{n}\|_{\text{total}} \leq C$ for all $n \in I$ and a suitable $C' > 0$.

As in the proof of Lemma C.1, we can then find an infinite subset $I_{0} \subset I$ such that all $\Psi_{n}$ for $n \in I_{0}$ have the same architecture $(N_{0}, N_{1}, \ldots, N_{L})$. Therefore, $(\Psi_{n})_{n \in I_{0}}$ is a bounded, infinite family in the finite dimensional vector space $\mathcal{N}_{N}(N_{0}, \ldots, N_{L})$. Thus, there is a further infinite set $I_{1} \subset I_{0}$ such that $\Psi_{n} \to \Psi \in \mathcal{N}_{N}(N_{0}, \ldots, N_{L})$ as $n \to \infty$ in $I_{1}$.

But since $K$ is bounded, the realization map $R_{\varrho}^{K} : \mathcal{N}_{N}(N_{0}, \ldots, N_{L}) \to C(K; \mathbb{R}^{N_{L}}), \Phi \mapsto R_{\varrho}^{K}(\Phi)$ is continuous (even locally Lipschitz continuous). Since $C(K)$ is continuously embedded into $\mathcal{Y}$, we thus get $f_{n} = R_{\varrho}^{K}(\Psi_{n}) \to R_{\varrho}^{K}(\Psi)$ with convergence in $\mathcal{Y}$ as $n \to \infty$ in $I_{1}$. Hence, $f = R_{\varrho}^{K}(\Psi) \in \mathcal{R}_{N_{d} L, N_{L} \varrho}^{\mathcal{N}^{K,C}_{d,L,N,L_{\varrho}}}$. □
C.7 Proof of Theorem 4.8

For the proof of Theorem 4.8 we will use a careful analysis of the singularity hyperplanes of functions of the form \( x \mapsto g_\alpha(\langle \alpha, x \rangle + \beta) \), that is, the hyperplane on which this function is not differentiable. To simplify this analysis, we first introduce a convenient terminology and discuss quite a few auxiliary results.

**Definition C.2.** For \( \alpha, \beta, \beta \in \mathbb{R} \), we write \( (\alpha, \beta) \sim (\tilde{\alpha}, \tilde{\beta}) \) iff there is some \( \varepsilon \in \{ \pm 1 \} \) such that \( (\alpha, \beta) = \varepsilon \cdot (\tilde{\alpha}, \tilde{\beta}) \).

Furthermore, for \( \alpha \geq 0 \) and with \( g_\alpha : \mathbb{R} \to \mathbb{R}, x \mapsto \max \{ x, ax \} \) denoting the parametric ReLU, we set
\[
S_{\alpha, \beta} := \{ x \in \mathbb{R}^d : \langle \alpha, x \rangle + \beta = 0 \}
\]
and
\[
h^{(\alpha)}_{\alpha, \beta} : \mathbb{R}^d \to \mathbb{R}, x \mapsto g_\alpha(\langle \alpha, x \rangle + \beta).
\]

Furthermore, we define
\[
W^+_{\alpha, \beta} := \{ x \in \mathbb{R}^d : \langle \alpha, x \rangle + \beta > 0 \}
\]
and \( W^-_{\alpha, \beta} := \{ x \in \mathbb{R}^d : \langle \alpha, x \rangle + \beta < 0 \} \),

and finally
\[
U^{(c)}_{\alpha, \beta} := \{ x \in \mathbb{R}^d : |\langle \alpha, x \rangle + \beta| \geq \varepsilon \}, \quad U^{(\varepsilon, +)}_{\alpha, \beta} := U^{(c)}_{\alpha, \beta} \cap W^+_{\alpha, \beta} \quad \text{and} \quad U^{(\varepsilon, -)}_{\alpha, \beta} := U^{(c)}_{\alpha, \beta} \cap W^-_{\alpha, \beta} \quad \text{for} \ \varepsilon > 0.
\]

**Lemma C.3.** Let \( (\alpha, \beta) \in S^{d-1} \times \mathbb{R} \) and \( x_0 \in S_{\alpha, \beta} \). Furthermore, let \( (\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N) \in S^{d-1} \times \mathbb{R} \) with \( (\alpha_\ell, \beta_\ell) \neq (\alpha, \beta) \) for all \( \ell \in \mathbb{N} \). Then there is some \( z \in \mathbb{R}^d \) satisfying
\[
\langle z, \alpha \rangle = 0 \quad \text{and} \quad \langle z, \alpha_j \rangle \neq 0 \ \forall j \in \mathbb{N} \quad \text{with} \ x_0 \in S_{\alpha_j, \beta_j}.
\]

**Proof.** By discarding those \( (\alpha_j, \beta_j) \) for which \( x_0 \notin S_{\alpha_j, \beta_j} \), we can assume that \( x_0 \in S_{\alpha_j, \beta_j} \) for all \( j \in \mathbb{N} \).

Assume towards a contradiction that the claim of the lemma is false; that is,
\[
\alpha^\perp = \bigcup_{j=1}^N \{ z \in \alpha^\perp : \langle z, \alpha_j \rangle = 0 \},
\]
where \( \alpha^\perp = \{ z \in \mathbb{R}^d : \langle z, \alpha \rangle = 0 \} \). Since \( \alpha^\perp \) is a closed subset of \( \mathbb{R}^d \) and thus a complete metric space, and since the right-hand side of (C.12) is a countable (in fact, finite) union of closed sets, the Baire category theorem (see [10] Theorem 5.9) shows that there is some \( j \in \mathbb{N} \) such that
\[
V := \{ z \in \alpha^\perp : \langle z, \alpha_j \rangle = 0 \} \supset B_\varepsilon(x) \cap \alpha^\perp \quad \text{for some} \ x \in V.
\]

But since \( V \) is a vector space, this easily implies \( V = \alpha^\perp \), that is, \( \langle z, \alpha_j \rangle = 0 \) for all \( z \in \alpha^\perp \). In other words, \( \alpha^\perp \subset \alpha_j^\perp \), and then \( \alpha^\perp = \alpha_j^\perp \) by a dimension argument, since \( \alpha, \alpha_j \neq 0 \).

Hence, \( \text{span} \alpha = (\alpha^\perp)^\perp = (\alpha_j^\perp)^\perp = \text{span} \alpha_j \). Because of \( |\alpha| = |\alpha_j| = 1 \), we thus see \( \alpha = \varepsilon \alpha_j \) for some \( \varepsilon \in \{ \pm 1 \} \). Finally, since \( x_0 \in S_{\alpha, \beta} \cap S_{\alpha_j, \beta_j} \), we see
\[
\beta = -\langle \alpha, x_0 \rangle = -\varepsilon \langle \alpha_j, x_0 \rangle = \varepsilon \beta_j,
\]
and thus \( (\alpha, \beta) = \varepsilon (\alpha_j, \beta_j) \), in contradiction to \( (\alpha, \beta) \not\sim (\alpha_j, \beta_j) \).

**Lemma C.4.** Let \( (\alpha, \beta) \in S^{d-1} \times \mathbb{R} \) and \( (\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N) \in S^{d-1} \times \mathbb{R} \) with \( (\alpha_i, \beta_i) \not\sim (\alpha, \beta) \) for all \( i \in \mathbb{N} \). Furthermore, let \( U \subset \mathbb{R}^d \) be open with \( S_{\alpha, \beta} \cap U \neq \emptyset \).

Then there is some \( \varepsilon > 0 \) satisfying
\[
U \cap S_{\alpha, \beta} \cap \bigcap_{j=1}^N U^{(\varepsilon)}_{\alpha_j, \beta_j} \neq \emptyset.
\]
Proof. By assumption, there is some \( x_0 \in U \cap S_{a,\beta} \). Next, Lemma C.3 yields some \( z \in \mathbb{R}^d \) such that \( \langle z, \alpha \rangle \neq 0 \) and \( \langle z, \alpha_j \rangle \neq 0 \) for all \( j \in \mathbb{N} \) with \( x_0 \in S_{a_j,\beta_j} \). Note that this implies \( \langle \alpha, x_0 + tz \rangle + \beta = \langle \alpha, x_0 \rangle + \beta = 0 \) and hence \( x_0 + tz \in S_{a,\beta} \) for all \( t \in \mathbb{R} \).

Next, let \( J := \{ j \in \mathbb{N} : x_0 \notin S_{a_j,\beta_j} \} \), so that \( \langle \alpha_j, x_0 \rangle + \beta_j \neq 0 \) for all \( j \in J \). Thus, there are \( \varepsilon_1, \delta > 0 \) with \( |\langle \alpha_j, x_0 + tz \rangle + \beta_j | \geq \varepsilon_1 \) (that is, with \( x_0 + tz \in U^{(\varepsilon_1)} ) \) for all \( t \in \mathbb{R} \) with \( |t| \leq \delta \). Since \( U \) is open with \( x_0 \in U \), we can shrink \( \delta \) so that \( x_0 + tz \in U \) for all \( |t| \leq \delta \). Let \( t := \delta \).

We claim that there is some \( \varepsilon > 0 \) such that \( x := x_0 + tz \in U \cap S_{a,\beta} \cap \bigcap_{j=1}^N U^{(\varepsilon)}_{a_j,\beta_j} \). To see this, note for \( j \in \mathbb{N} \setminus J \) that \( x_0 \in S_{a_j,\beta_j} \), and hence \( |\langle x_0 + tz, \alpha_j \rangle + \beta_j | = |t| \cdot |\langle z, \alpha_j \rangle| \geq \delta \cdot \min_{j \in \mathbb{N} \setminus J} |\langle z, \alpha_j \rangle| =: \varepsilon_2 > 0 \), since \( \langle z, \alpha \rangle \neq 0 \) for all \( j \in \mathbb{N} \setminus J \), by choice of \( z \). By combining all our observations, we see that \( x_0 + tz \in U \cap S_{a,\beta} \cap \bigcap_{j=1}^N U^{(\varepsilon)}_{a_j,\beta_j} \) for \( \varepsilon := \min\{\varepsilon_1, \varepsilon_2\} > 0 \).

Lemma C.5. If \( 0 < a < 1 \) and \( (\alpha, \beta) \in S^{d-1} \times \mathbb{R} \), then \( h_{a,\beta} \) is not differentiable at any \( x_0 \in S_{a,\beta} \).

Proof. Assume towards a contradiction that \( h_{a,\beta} \) is differentiable at some \( x_0 \in S_{a,\beta} \). Then
\[
f : \mathbb{R} \to \mathbb{R}, t \mapsto h_{a,\beta}(x_0 + ta)
\]
is differentiable at \( t = 0 \). But since \( x_0 \in S_{a,\beta} \) and \( |\alpha| = 1 \), we have
\[
f(t) = g_0(\langle \alpha, x_0 + ta \rangle + \beta) = g_0(t) = \begin{cases} t, & \text{if } t \geq 0, \\ at, & \text{if } t < 0, \end{cases}
\]
for all \( t \in \mathbb{R} \). This easily shows that \( f \) is not differentiable at \( t = 0 \), since the right-sided derivative is 1, while the left-sided derivative is \( a \neq 1 \). This is the desired contradiction.

Lemma C.6. Let \( 0 < a < 1 \), and let \( (\alpha_1, \beta_1), \ldots, (\alpha_N, \beta_N) \in S^{d-1} \times \mathbb{R} \) with \( (\alpha_i, \beta_i) \neq (\alpha_j, \beta_j) \) for \( j \neq i \). Furthermore, let \( U \subset \mathbb{R}^d \) be open with \( U \cap S_{a_i,\beta_i} \neq \emptyset \) for all \( i \in \mathbb{N} \). Finally, set \( h_i := h_{a_i,\beta_i} \big|_U \) for \( i \in \mathbb{N} \) with \( h_{a_i,\beta_i} \) as in Definition C.3 and let \( h_{N+1} : U \to \mathbb{R}, x \mapsto 1 \).

Then the family \( (h_i)_{i=1}^{N+1} \) is linearly independent.

Proof. Assume towards a contradiction that \( 0 = \sum_{i=1}^{N+1} \gamma_i h_i \) for certain \( \gamma_1, \ldots, \gamma_{N+1} \in \mathbb{R} \) with \( \gamma_t \neq 0 \) for some \( i \in \mathbb{N} \). Note that if we had \( \gamma_i = 0 \) for all \( i \in \mathbb{N} \), we would get \( 0 = \gamma_{N+1} h_{N+1} \equiv \gamma_{N+1} \), and thus \( \gamma_1 = 0 \) for all \( i \in \mathbb{N} \), a contradiction. Hence, there is some \( j \in \mathbb{N} \) with \( \gamma_j \neq 0 \).

By Lemma C.4 there is some \( x_0 \in U \cap S_{a_j,\beta_j} \cap \bigcap_{i \in \mathbb{N} \setminus \{j\}} U^{(\varepsilon)}_{a_i,\beta_i} \). Therefore, \( x_0 \in U \cap S_{a_j,\beta_j} \cap V \) for the open set \( V := \bigcap_{i \in \mathbb{N} \setminus \{j\}} (\mathbb{R}^d \setminus S_{a_i,\beta_i}) \).

Because of \( x_0 \in U \cap S_{a_j,\beta_j} \), Lemma C.5 shows that \( h_{a_j,\beta_j} \big|_U \) is not differentiable at \( x_0 \). On the other hand, we have
\[
h_{a_j,\beta_j} \big|_U = h_j = -\gamma_j^{-1} \cdot \left( \gamma_{N+1} h_{N+1} + \sum_{i \in \mathbb{N} \setminus \{j\}} \gamma_i h_{a_i,\beta_i} \big|_U \right),
\]
where the right-hand side is differentiable at \( x_0 \), since each summand is easily seen to be differentiable on the open set \( V \), with \( x_0 \in V \cap U \).

Lemma C.7. Let \( (\alpha, \beta) \in S^{d-1} \times \mathbb{R} \). If \( K \subset \mathbb{R}^d \) is compact with \( K \cap S_{a,\beta} = \emptyset \), then there is some \( \varepsilon > 0 \) such that \( K \subset U^{(\varepsilon)}_{a,\beta} \).

Proof. The continuous function \( K \to (0, \infty), x \mapsto |\langle \alpha, x \rangle + \beta| \), which is well-defined by assumption, attains a minimum \( \varepsilon = \min_{x \in K} |\langle \alpha, x \rangle + \beta| > 0 \).
Lemma C.8. Let $0 \leq a < 1$, let $(\alpha, \beta) \in S^{d-1} \times \mathbb{R}$, and let $U \subset \mathbb{R}^d$ be open with $U \cap S_{\alpha, \beta} \neq \emptyset$. Finally, let $f : U \to \mathbb{R}$ be continuous, and assume that $f$ is affine-linear on $U \cap W^+_{\alpha, \beta}$ and on $U \cap W^-_{\alpha, \beta}$.

Then there are $c, \kappa \in \mathbb{R}$ and $\zeta \in \mathbb{R}^d$ such that

$$f(x) = c \cdot \varrho_a((\alpha, x) + \beta) + (\zeta, x) + \kappa \quad \forall x \in U.$$ 

Proof. By assumption, there are $\xi_1, \xi_2 \in \mathbb{R}^d$ and $\omega_1, \omega_2 \in \mathbb{R}$ satisfying

$$f(x) = \langle \xi_1, x \rangle + \omega_1 \quad \text{for } x \in U \cap W^+_{\alpha, \beta} \quad \text{and} \quad f(x) = \langle \xi_2, x \rangle + \omega_2 \quad \text{for } x \in U \cap W^-_{\alpha, \beta}. $$

Step 1: We claim that $U \cap S_{\alpha, \beta} \subset U \cap \overline{W^+_{\alpha, \beta}}$. Indeed, for arbitrary $x \in U \cap S_{\alpha, \beta}$, we have $x + t\alpha \in U$ for $t \in (-\varepsilon, \varepsilon)$ for a suitable $\varepsilon > 0$, since $U$ is open. But since $x \in S_{\alpha, \beta}$ and $|\alpha| = 1$, we have $(x + t\alpha, \alpha) + \beta = t$. Hence, $x + t\alpha \in U \cap W^+_{\alpha, \beta}$ for $t \in (0, \varepsilon)$ and $x + t\alpha \in U \cap W^-_{\alpha, \beta}$ for $t \in (-\varepsilon, 0)$. This easily implies the claim of this step.

Step 2: We claim that $\xi_1 - \xi_2 \in \text{span } \alpha$. To see this, consider the modified function

$$\tilde{f} : U \to \mathbb{R}, x \mapsto f(x) - \langle \xi_2, x \rangle + \omega_2,$$

which is continuous and satisfies $\tilde{f} \equiv 0$ on $U \cap W^+_{\alpha, \beta}$ and $\tilde{f}(x) = \langle \theta, x \rangle + \omega$ on $U \cap W^+_{\alpha, \beta}$, where we defined $\theta := \xi_1 - \xi_2$ and $\omega := \omega_1 - \omega_2$.

Since we saw in Step 1 that $U \cap S_{\alpha, \beta} \subset U \cap \overline{W^+_{\alpha, \beta}}$, we thus get by continuity of $\tilde{f}$ that

$$0 = \tilde{f}(x) = \langle \theta, x \rangle + \omega \quad \forall x \in U \cap S_{\alpha, \beta}.$$ 

But by assumption on $U$, there is some $x_0 \in U \cap S_{\alpha, \beta}$. For arbitrary $v \in \alpha^\perp$, we then have $x_0 + tv \in U \cap S_{\alpha, \beta}$ for all $t \in (-\varepsilon, \varepsilon)$ and a suitable $\varepsilon = \varepsilon(v) > 0$, since $U$ is open. Hence, $0 = \langle \theta, x_0 + tv \rangle + \omega = \langle \theta, v \rangle$ for all $t \in (-\varepsilon, \varepsilon)$, and thus $v \in \theta^\perp$. In other words, $\alpha^\perp \subset \theta^\perp$, and thus span $\alpha = (\alpha^\perp) \supset (\theta^\perp) \supset \theta = \xi_1 - \xi_2$, as claimed in this step.

Step 3: In this step, we complete the proof. As seen in the previous step, there is some $c \in \mathbb{R}$ satisfying $c\alpha = (\xi_1 - \xi_2)/(1 - a)$. Now, set $\zeta := (\xi_2 - a\xi_1)/(1 - a)$ and $\kappa := f(x_0) - \langle \zeta, x_0 \rangle$, where $x_0 \in U \cap S_{\alpha, \beta}$ is arbitrary. Finally, define

$$g : \mathbb{R}^d \to \mathbb{R}, x \mapsto c \cdot \varrho_a((\alpha, x) + \beta) + \langle \zeta, x \rangle + \kappa.$$ 

Because of $x_0 \in S_{\alpha, \beta}$, we then have $g(x_0) = \langle \zeta, x_0 \rangle + \kappa = f(x_0)$. Furthermore, since $\varrho_a(x) = x$ for $x \geq 0$, we see for all $x \in U \cap W^+_{\alpha, \beta}$ that

$$g(x) - f(x_0) = g(x) - g(x_0) = c \cdot ((\alpha, x) + \beta) + \langle \zeta, x - x_0 \rangle$$

(since $x_0 \in S_{\alpha, \beta}$, i.e., $(\alpha, x_0) + \beta = 0$) = $c \cdot ((\alpha, x - x_0) + \langle \zeta, x - x_0 \rangle = \left(\frac{\xi_1 - \xi_2}{1 - a} + \frac{\xi_2 - a\xi_1}{1 - a}, x - x_0\right)$ \hspace{1cm} (C.13)

$$= \langle \xi_1, x - x_0 \rangle = f(x) - f(x_0).$$

Here, the last step used that $f(x) = \langle \xi_1, x \rangle + \omega_1$ for $x \in U \cap W^+_{\alpha, \beta}$, and that $x_0 \in U \cap S_{\alpha, \beta} \subset \overline{U \cap W^+_{\alpha, \beta}}$ by Step 1.

Likewise, since $\varrho_a(x) = ax$ for $x < 0$, we see for $x \in U \cap W^-_{\alpha, \beta}$ that

$$g(x) - f(x_0) = g(x) - g(x_0) = ac \cdot ((\alpha, x) + \beta) + \langle \zeta, x - x_0 \rangle$$

(since $x_0 \in S_{\alpha, \beta}$, i.e., $(\alpha, x_0) + \beta = 0$) = $ac((\alpha, x - x_0) + \langle \zeta, x - x_0 \rangle = \left(a \frac{\xi_1 - \xi_2}{1 - a} + \frac{\xi_2 - a\xi_1}{1 - a}, x - x_0\right)$ \hspace{1cm} (C.14)

$$= \langle \xi_2, x - x_0 \rangle = f(x) - f(x_0).$$

In combination, Equations (C.13) and (C.14) show $f(x) = g(x)$ for all $x \in U \cap (W^+_\alpha \cup W^-_{\alpha, \beta})$. Since this set is dense in $U$ by Step 1, we are done. \hfill \Box
With all of these preparations, we can finally prove Theorem 4.8.

**Proof of Theorem 4.8.** Since \( g_1 = \text{id}_R \), the result is trivial for \( a = 1 \), since \( \mathcal{R} \mathcal{N} \mathcal{N}_{d,2,N}^{[−B,B]^d} \) is just the set of all affine-linear maps \( [−B,B]^d \to \mathbb{R} \). Therefore, we assume \( a < 1 \) in the sequel. For brevity, let \( K := [−B,B]^d \) and \( N_0 := N − d − 1 \). Then, each \( \Phi \in \mathcal{N} \mathcal{N}_{d,2,N} \) is of the form \( \Phi = ((A_1, b_1), (A_2, b_2)) \) with \( A_1 \in \mathbb{R}^{N_a \times d}, A_2 \in \mathbb{R}^{1 \times N_0} \), and \( b_1 \in \mathbb{R}^{N_0}, b_2 \in \mathbb{R}^1 \).

Let \( (\Phi_n)_{n \in \mathbb{N}} \subset \mathcal{N} \mathcal{N}_{d,2,N} \) with \( \Phi_n = (\tilde{A}_n, \tilde{b}_n) \) be such that \( f_n := R^K_{g_a}(\Phi_n) \) converges uniformly to some \( f \in C(K) \). Our goal is to prove \( f \in \mathcal{R} \mathcal{N} \mathcal{N}^{K}_d(N_{d,2,N},g_a) \).

**Step 1 (Normalizing the rows of the first layer):** Our first goal is to normalize the rows of the matrices \( \tilde{A}_1 \), that is, to change the parametrization of the network such that \( |(\tilde{A}_1^n)_{i,−}| = 1 \) for all \( i \in N_a \). To see that this is possible, consider arbitrary \( A \in \mathbb{R}^{M_1 \times M_2} \neq 0 \) and \( b \in \mathbb{R}^{M_2} \); then we obtain by the positive homogeneity of \( g_a \) for all \( C > 0 \) that

\[
g_a(Ax + b) = C \cdot g_a \left( \frac{A x + b}{C} \right) \quad \text{for all } x \in \mathbb{R}^{M_2}.
\]

This identity shows that for each \( n \in \mathbb{N} \), we can find a network

\[
\tilde{\Phi}_n = \left( (A_1^n, b_1^n), (A_2^n, b_2^n) \right) \in \mathcal{N} \mathcal{N}_{d,2,N} \,
\]

such that the rows of \( A_1^n \) are normalized, that is, \( |(A_1^n)_{i,−}| = 1 \) for all \( i \in N_a \), and such that

\[
R^K_{g_a}(\tilde{\Phi}_n) = R^K_{g_a}(\Phi_n) = f_n \quad \text{for all } n \in \mathbb{N}.
\]

**Step 2 (Extracting a partially convergent subsequence):** By the Theorem of Bolzano-Weierstraß, there is a common subsequence of \( (A_1^n)_{n \in \mathbb{N}} \) and \( (b_1^n)_{n \in \mathbb{N}} \), denoted by \((A_1^{n_k})_{k \in \mathbb{N}} \) and \((b_1^{n_k})_{k \in \mathbb{N}} \), converging to \( A_1 \in \mathbb{R}^{N_0 \times d} \) and \( b_1 \in [−\infty, \infty)^{N_0} \), respectively.

For \( j \in N_0 \), let \( a_{k,j} \in \mathbb{R}^d \) denote the \( j \)-th row of \( A_1^{n_k} \), and let \( a_j \in \mathbb{R}^d \) denote the \( j \)-th row of \( A_1 \). Note that \( |a_{k,j}| = |a_j| = 1 \) for all \( j \in N_0 \) and \( k \in \mathbb{N} \). Next, let

\[
J := \left\{ j \in N_0 : (b_1)_j \in \{±\infty\} \right\} \cup \left\{ (b_1)_j \in \mathbb{R} \quad \text{and} \quad S_{a_j,(b_1)_j} \cap K^\circ = \emptyset \right\},
\]

where \( K^\circ = (−B,B)^d \) denotes the interior of \( K \). Additionally, let \( J^c := N_0 \setminus J \), and for \( j, \ell \in J^c \) write \( j \simeq \ell \) if \( (a_j, (b_1)_j) \sim (a_\ell, (b_1)_\ell) \), with the relation ~ introduced in Definition C.2. Let \( (J_i)_{i=1,...,r} \) denote the equivalence classes of the relation ~. For each \( i \in \mathbb{N} \), choose \( \alpha^{(i)} \in S^{d−1} \) and \( \beta^{(i)} \in \mathbb{R} \) such that for each \( j \in J_i \), there is a (unique) \( \varepsilon_j \in \{±1\} \) with \( (a_j, (b_1)_j) = \varepsilon_j \cdot (\alpha^{(i)}, \beta^{(i)}) \).

**Step 3 (Handling the case of distinct singularity hyperplanes):** Note that \( r \leq |J^c| \leq N_0 \). Before we continue with the general case, let us consider the special case where equality occurs, that is, where \( r = N_0 \). This means that \( J = \emptyset \), and that each equivalence class \( J_i \) has precisely one element, that is, \( (a_j, (b_1)_j) \neq (a_\ell, (b_1)_\ell) \) for \( j, \ell \in N_0 \) with \( j \neq \ell \).

Therefore, Lemma C.6 shows that the functions \( (h_j)_{K^\circ} = h_j : N_0 + 1 \to \mathbb{R} \) for \( j \in N_0 \) and \( h_{N+1} : K \to \mathbb{R}, x \mapsto 1 \) are linearly independent. In particular, these functions are linearly independent when considered on all of \( K \). Thus, we can define a norm \( \| \cdot \|_s \) on \( \mathbb{R}^{N_0+1} \) by virtue of

\[
\|c\|_s := \left\| c_{N_0+1} + \sum_{j=1}^{N_0} c_j h_{a_j,(b_1)_j} \right\|_{L^\infty(K)} \quad \text{for } c = (c_j)_{j=1,...,N_0+1} \in \mathbb{R}^{N_0+1}.
\]

Since all norms on the finite dimensional vector space \( \mathbb{R}^{N_0+1} \) are equivalent, there is some \( \tau > 0 \) with \( \|c\|_\tau \geq \tau \cdot \|c\|_s \) for all \( c \in \mathbb{R}^{N_0+1} \).
Case 1: We have $(b_{1,k})_j \to (b_{1,k})_j$ as $k \to \infty$. This easily implies for arbitrary $j \in N_0$ and $h_j^{(k)} := h_{a_j,(b_{1,k})_j}^{(a)}$ that $h_j^{(k)} \to h_{a_j,(b_{1,k})_j}^{(a)}$, with uniform convergence on $K$. Thus, there is some $N_0 \in \mathbb{N}$ such that $\|h_j^{(k)} - h_{a_j,(b_{1,k})_j}^{(a)}\|_{L^\infty(K)} \leq \tau/2$ for all $k \geq N_0$ and $j \in N_0$. Therefore,

$$\left\| c_{N_0+1} + \sum_{j=1}^{N_0} c_j h_j^{(k)} \right\|_{L^\infty(K)} \geq \left\| c_{N_0+1} + \sum_{j=1}^{N_0} c_j h_{a_j,(b_{1,k})_j}^{(a)} \right\|_{L^\infty(K)} - \sum_{j=1}^{N_0} c_j \left( h_{a_j,(b_{1,k})_j}^{(a)} - h_j^{(k)} \right) \right\|_{L^\infty(K)} \geq \tau \cdot \|c\| \cdot \left( \sum_{j=1}^{N_0} |c_j| \cdot \|h_{a_j,(b_{1,k})_j}^{(a)} - h_j^{(k)}\|_{L^\infty(K)} \right) \geq \left( \tau - \frac{\tau}{2} \right) \cdot \|c\| \|\varepsilon\| = \frac{\tau}{2} \cdot \|c\| \|\varepsilon\| \quad \forall \varepsilon = (c_j)_{j=1,...,N_0+1} \in \mathbb{R}^{N_0+1}.$$

Since $f_{nk} = \mathbb{R}^{K} \left( \Phi_{nk} \right) = b_{2,k}^{nk} + \sum_{i=1}^{N_0} \left( A_2^{nk} \right)_{i,j} h_j^{(k)}$ converges uniformly on $K$, we thus see that the sequence consisting of $(A_2^{nk}, b_{2,k}^{nk}) \in \mathbb{R}^{1 \times N_0} \times \mathbb{R} \cong \mathbb{R}^{N_0+1}$ is bounded. Thus, there is a further subsequence $(n_{k_\ell})_{\ell \in \mathbb{N}}$ such that $A_2^{nk_\ell} \to A_2 \in \mathbb{R}^{1 \times N_0}$ and $b_{2,k}^{nk_\ell} \to b_2 \in \mathbb{R}$ as $\ell \to \infty$. But this implies as desired that

$$f = \lim_{\ell \to \infty} f_{nk_\ell} = \lim_{\ell \to \infty} \left[ b_{2,k}^{nk_\ell} + \sum_{j=1}^{N_0} \left( A_2^{nk_\ell} \right)_{i,j} h_j^{(k)} \right] = b_2 + \sum_{j=1}^{N_0} \left( A_2 \right)_{i,j} h_{a_j,(b_{1,k})_j}^{(a)} K \in \mathcal{R} \mathcal{N} \mathcal{N}_d^{K} \mathcal{N}_{N_0+1} = \mathcal{R} \mathcal{N} \mathcal{N}_d^{K} \mathcal{N}_{N_0+1,0}.$$

Step 4 (Showing that the $j$-th neuron is eventually affine-linear, for $j \in J_j$): From now on, we consider the case where $r < N_0$.

For $j \in J$, there are two cases: In case of $(b_{1,j})_j \in [0,\infty]$, define

$$\phi_j^{(k)} : \mathbb{R}^d \to \mathbb{R}, \ x \mapsto (A_2^{nk})_{i,j} \cdot (a_{k,j}, x) + (b_{1,k})_j \quad \text{for all } k \in \mathbb{N}.$$

If otherwise $(b_{1,j})_j \in [-\infty,0)$, define

$$\phi_j^{(k)} : \mathbb{R}^d \to \mathbb{R}, \ x \mapsto a \cdot (A_2^{nk})_{i,j} \cdot (a_{k,j}, x) + (b_{1,k})_j \quad \text{for all } k \in \mathbb{N}.$$

Next, for arbitrary $0 < \delta < B$, we define $K_\delta := \left[ -(B-\delta), B-\delta \right]^d$. Note that since $S_{\alpha(i),\beta(i)} \cap K^o \neq \emptyset$ for all $i \in R$, there is some $\delta_0 > 0$ such that $S_{\alpha(i),\beta(i)} \cap \left( (B-\delta), B+\delta \right)^d \neq \emptyset$ for all $i \in R$ and all $0 < \delta \leq \delta_0$. For the remainder of this step, we will consider a fixed $\delta \in (0,\delta_0]$, and we claim that there is some $N_1 = N_1(\delta) > 0$ such that

$$\text{sign} \left( (a_{k,j}, x) + (b_{1,k})_j \right) = \text{sign} \left( (b_{1,k})_j \right) \neq 0 \quad \forall \ j \in J, k \geq N_1, \text{ and } x \in K_\delta,$$

where $\text{sign} x = 1$ if $x > 0$, $\text{sign} x = -1$ if $x < 0$, and $\text{sign} 0 = 0$. Note that once this is shown, it is not hard to see

$$(A_2^{nk})_{i,j} \cdot g_a( (a_{k,j}, x) + (b_{1,k})_j ) \neq 0 \quad \forall \ j \in J, k \geq N_1, \text{ and } x \in K_\delta,$$

simply because $g_a(x) = x$ if $x \geq 0$, and $g_a(x) = ax$ if $x < 0$. Therefore, the affine-linear function

$$g_{r+1}^{(k)} := b_{2,k}^{nk} + \sum_{j \in J} \phi_j^{(k)}(x) \text{ satisfies } g_{r+1}^{(k)}(x) = \sum_{j \in J} \left( A_2^{nk}_2 \right)_{i,j} g_a \left( (a_{k,j}, x) + (b_{1,k})_j \right) \quad \forall k \geq N_1(\delta) \text{ and } x \in K_\delta.$$

To prove Equation (C.15), we distinguish two cases for each $j \in J$:

**Case 1:** We have $(b_{1,j})_j \in [\pm \infty]$. Because of $(b_{1,k})_j \to (b_{1,k})_j$, there is thus some $k_j \geq 0$ with $|b_{1,k})_j| \geq 2d \cdot B$ for all $k \geq k_j$. Since we have $|a_{k,j}| = 1$ and $|x| \leq \sqrt{d}B \leq dB$ for $x \in K$, this implies

$$|a_{k,j}, x) + (b_{1,k})_j| \geq |b_{1,k})_j| - |a_{k,j}, x) | \geq 2d \cdot B - |x| \geq dB > 0 \quad \forall x \in K = [-B,B]^d \text{ and } k \geq k_j.$$
Now, since the function $x \mapsto \langle a_{k,j}, x \rangle + (b_{1}^{n})_{j}$ is continuous, since $K$ is connected (in fact convex), and since $0 \in K$, this implies $\langle a_{k,j}, x \rangle + (b_{1}^{n})_{j} = \text{sign}(b_{1}^{n})_{j}$ for all $x \in K$ and $k \geq k_{j}$.

**Case 2:** We have $(b_{1})_{j} \in \mathbb{R}$, but $S_{a_{j},(b_{1})_{j}} \cap K^{\circ} = \emptyset$, and hence $S_{a_{j},(b_{1})_{j}} \cap K_{\delta} = \emptyset$. In view of Lemma [C.4], there is thus some $\varepsilon_{j,\delta} > 0$ satisfying $K_{\delta} \subset U_{\alpha(j),\delta}^{(\varepsilon)}$, that is, $\left| \langle a_{j}, x \rangle + (b_{1})_{j} \right| \geq \varepsilon_{j,\delta} > 0$ for all $x \in K_{\delta}$.

Since $a_{k,j} \rightarrow a_{j}$ and $(b_{1}^{n})_{j} \rightarrow (b_{1})_{j}$ as $k \rightarrow \infty$, there is some $k_{j} = k_{j}(\varepsilon_{j,\delta}) = k_{j}(\delta) \in \mathbb{N}$ such that $|a_{k,j} - a_{j}| \leq \varepsilon_{j,\delta}/(4dB)$ and $|(b_{1}^{n})_{j} - (b_{1})_{j}| \leq \varepsilon_{j,\delta}/4$ for all $k \geq k_{j}$. Therefore,

$$\left| \langle a_{k,j}, x \rangle + (b_{1}^{n})_{j} \right| \geq \left| \langle a_{j}, x \rangle + (b_{1})_{j} \right| - \left| \langle a_{j} - a_{k,j}, x \rangle + (b_{1})_{j} - (b_{1}^{n})_{j} \right| \geq \varepsilon_{j,\delta} - \frac{\varepsilon_{j,\delta}}{4dB} \cdot dB = \frac{\varepsilon_{j,\delta}}{2} > 0 \quad \forall x \in K_{\delta} \text{ and } k \geq k_{j}.$$

With the same argument as at the end of Case 1, we thus see $\text{sign}(a_{k,j}, x) + (b_{1}^{n})_{j}) = \text{sign}(b_{1}^{n})_{j}$ for all $x \in K_{\delta}$ and $k \geq k_{j}(\delta)$.

Together, the two cases prove that Equation [C.15] holds if we set $N_{1}(\delta) = \max_{j \in J} k_{j}(\delta)$.

**Step 5 (Showing that the $j$-th neuron is affine-linear on $U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$ for $j \in J_{i}$):** We claim that for each $\varepsilon > 0$, there is some $N_{2}(\varepsilon) \in \mathbb{N}$ such that:

If $i \in \mathbb{R}$, $j \in J_{i}$ and $k \geq N_{2}(\varepsilon)$, then $\nu_{j}^{(k)}(\varepsilon) := \alpha_{i}((a_{k,j}, \bullet) + (b_{1}^{n})_{j})$ is affine-linear on $K \cap U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$.

To see this, let $\varepsilon > 0$ be arbitrary, and recall $J^{c} = \bigcup_{i=1}^{r} J_{i}$. By definition of $J_{i}$, there is for each $i \in \mathbb{R}$ and $j \in J_{i}$ some $\sigma_{j} \in \{\pm 1\}$ satisfying

$$(a_{k,j}, (b_{1}^{n})_{j}) \xrightarrow{k \rightarrow \infty} (a_{j}, (b_{1})_{j}) = \sigma_{j} \cdot (\alpha(i), \beta(i)).$$

Thus, there is some $k(j)(\varepsilon) \in \mathbb{N}$ such that $|a_{k,j} - \sigma_{j} \alpha(i)| \leq \varepsilon/4dB$ and $|(b_{1}^{n})_{j} - \sigma_{j} \beta(i)| \leq \varepsilon/4$ for all $k \geq k(j)(\varepsilon)$.

Define $N_{2}(\varepsilon) := \max_{j \in J^{c}} k(j)(\varepsilon)$. Then, for $k \geq N_{2}(\varepsilon)$ and arbitrary $x \in K \cap U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$, we have on the one hand $|\sigma_{j} \cdot (\alpha(i), x) + \beta(i)| \geq \varepsilon$, and on the other hand

$$\left| \left( \langle a_{k,j}, x \rangle + (b_{1}^{n})_{j} \right) - \sigma_{j} \cdot (\alpha(i), x) + \beta(i) \right| \leq dB \cdot |a_{k,j} - \sigma_{j} \alpha(i)| + |(b_{1}^{n})_{j} - \sigma_{j} \beta(i)| \leq \varepsilon/2,$$

since $|x| \leq \sqrt{\alpha} \cdot B \leq dB$. In combination, this shows $\left| \langle a_{k,j}, x \rangle + (b_{1}^{n})_{j} \right| \geq \varepsilon/2 > 0$ for all $x \in K \cap U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$. But since $K \cap U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$ is connected (in fact convex), and since the function $x \mapsto \langle a_{k,j}, x \rangle + (b_{1})_{j}$ is continuous, it must have a constant sign on $K \cap U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$. This easily implies that $\nu_{j}^{(k)} = \alpha_{i}((a_{k,j}, \bullet) + (b_{1}^{n})_{j})$ is indeed affine-linear on $K \cap U_{\alpha(i),\beta(i)}^{(\varepsilon,\pm)}$ for $k \geq N_{2}(\varepsilon)$.

**Step 6 (Proving the “almost convergence” of the sum of all $j$-th neurons for $j \in J_{i}$):** For $i \in \mathbb{R}$ define

$$g_{t}^{(k)} : \mathbb{R}^{d} \rightarrow \mathbb{R}, x \mapsto \sum_{j \in J_{i}} (A_{2}^{k})_{1,j} \alpha_{i}((a_{k,j}, x) + (b_{1}^{n})_{j}) = \sum_{j \in J_{i}} (A_{2}^{k})_{1,j} \nu_{j}^{(k)}(x).$$

In combination with Equation (C.10), we see

$$f_{n_{k}}(x) = R_{\alpha_{i}}(\Phi_{n_{k}})(x) = \sum_{t=1}^{r+1} g_{t}^{(k)}(x) \quad \forall x \in K_{\delta} \text{ and } k \geq N_{1}(\delta),$$

with $g_{r+1}$ being affine-linear.
Recall from Step 4 that $K_{\delta_0}^n \cap S_{\alpha(i),\beta(i)} \neq \emptyset$ for all $i \in \mathbb{N}$ by choice of $\delta_0$. Therefore, Lemma 3.4 shows (because of $U_{\alpha,\beta}^\sigma \subset (U_{\alpha,\beta}^\varepsilon)^\circ$ for $\varepsilon < \sigma$) for each $i \in \mathbb{N}$ that

$$K_i := K_{\delta_0}^n \cap S_{\alpha(i),\beta(i)} \cap \bigcap_{\ell \in \mathbb{N} \setminus \{i\}} \left(U_{\alpha(i),\beta(i)}^{\varepsilon(i)}\right)^\circ \neq \emptyset \quad \text{for a suitable } \varepsilon_i > 0.$$  

Let us fix some $x_i \in K_i$ and some $\varepsilon_i > 0$ such that $\overline{B}_{\varepsilon_i}(x_i) \subset K_{\delta_0}^n \cap \bigcap_{\ell \in \mathbb{N} \setminus \{i\}} \left(U_{\alpha(i),\beta(i)}^{\varepsilon(i)}\right)^\circ$; this is possible, since the set on the right-hand side is open. Now, since $\overline{B}_{\varepsilon_i}(x_i)$ is connected, we see for each $\ell \in \mathbb{N} \setminus \{i\}$ that either $\overline{B}_{\varepsilon_i}(x_i) \subset U_{\alpha(i),\beta(i)}^{\varepsilon(i)}$ or $\overline{B}_{\varepsilon_i}(x_i) \subset U_{\alpha(i),\beta(i)}^{\varepsilon(i)-\ell}$. Therefore, as a consequence of the preceding step, we see that there is some $N_3(i) \in \mathbb{N}$ such that $g_i^{(k)}(x)$ is affine-linear on $\overline{B}_{\varepsilon_i}(x_i)$ for all $i \in \mathbb{N} \setminus \{i\}$ and all $k \geq N_3(i)$.

Thus, setting $N_3 := \max\{N_1(\delta_0), \max_{i=1,\ldots,N_3(i)}\}$, we see as a consequence of Equation (C.17), and because of $\overline{B}_{\varepsilon_i}(x_i) \subset K_{\delta_0}^n$ for each $i \in \mathbb{N}$ that for any $k \geq N_3$, there is an affine-linear map $g_i^{(k)} : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$f_{\eta_k}(x) = \sum_{i=k}^{k+1} g_i^{(k)}(x) = g_i^{(k)}(x) + q_i^{(k)}(x) \quad \forall x \in \overline{B}_{\varepsilon_i}(x_i) \text{ and } k \geq N_3. \quad (C.18)$$

Next, note that the preceding step implies for arbitrary $\varepsilon > 0$ that for all $k$ large enough (depending on $\varepsilon$), $g_i^{(k)}$ is affine-linear on $B_{\varepsilon_i}(x_i) \cap U_{\alpha(i),\beta(i)}^{\varepsilon(i)}$. Since $f(x) = \lim_k f_{\eta_k}(x) = \lim_k g_i^{(k)}(x) + q_i^{(k)}(x)$, we thus see that $f$ is affine-linear on $B_{\varepsilon_i}(x_i) \cap U_{\alpha(i),\beta(i)}^{\varepsilon(i)}$ and continuous on $K \supset B_{\varepsilon_i}(x_i)$, and we have $x_i \in B_{\varepsilon_i}(x_i) \cap S_{\alpha(i),\beta(i)} \neq \emptyset$. Thus, Lemma 3.8 shows that there are $c_i \in \mathbb{R}$, $\zeta_i \in \mathbb{R}^d$, and $\kappa_i \in \mathbb{R}$ such that

$$f(x) = G_i(x) \quad \forall x \in B_{\varepsilon_i}(x_i), \quad \text{ with } G_i : \mathbb{R}^d \to \mathbb{R}, x \mapsto c_i \cdot g_i\left(\alpha(i), x + \beta(i)\right) + \zeta_i, x + \kappa_i. \quad (C.19)$$

We now intend to make use of the following elementary fact: If $(\psi_k)_{k \in \mathbb{N}}$ is a sequence of maps $\psi_k : \mathbb{R}^d \to \mathbb{R}$, if $\Omega \subset \mathbb{R}^d$ is such that each $\psi_k$ is affine-linear on $\Omega$, and if $U \subset \Omega$ is a nonempty open subset such that $\psi(x) := \lim_{k \to \infty} \psi_k(x) \in \mathbb{R}^d$ for all $x \in U$, then $\psi$ can be extended to an affine-linear map $\psi : \mathbb{R}^d \to \mathbb{R}$, and we have $\psi(x) \to \psi(x)$ for all $x \in \Omega$, even with locally uniform convergence. Essentially, what is used here is that the vector space of affine-linear maps $\mathbb{R}^d \to \mathbb{R}$ is finite-dimensional, so that the (Hausdorff) topology of pointwise convergence on $U$ coincides with that of locally uniform convergence on $\Omega$; see [41, Theorem 1.21].

To use this observation, note that Equations (C.18) and (C.19) show that $g_i^{(k)} + q_i^{(k)}$ converges pointwise to $G_i$ on $B_{\varepsilon_i}(x_i)$. Furthermore, since $x_i \in S_{\alpha(i),\beta(i)}$, it is not hard to see that there is some $\varepsilon_0 > 0$ with $U_{\alpha(i),\beta(i)}^{\varepsilon_0} \cap B_{\varepsilon_i}(x_i) \neq \emptyset$ for all $\varepsilon \in (0, \varepsilon_0)$; for the details, we refer to Step 1 in the proof of Lemma 3.8.

Finally, as a consequence of Step 5, we see for arbitrary $\varepsilon \in (0, \varepsilon_0)$ that $g_i^{(k)} + q_i^{(k)}$ and $G_i$ are both affine-linear on $U_{\alpha(i),\beta(i)}^{\varepsilon(i)}$, at least for $k$ large enough (depending on $\varepsilon$). Thus, the observation from above implies that $g_i^{(k)} + q_i^{(k)} \to G_i$ pointwise on $U_{\alpha(i),\beta(i)}^{\varepsilon(i)}$, for arbitrary $\varepsilon \in (0, \varepsilon_0)$.

Because of $\bigcup_{\varepsilon \in \{\pm\}} U_{\alpha(i),\beta(i)}^{\varepsilon(i)} = \mathbb{R}^d \setminus S_{\alpha(i),\beta(i)}$, this implies

$$g_i^{(k)} + q_i^{(k)} \xrightarrow{k \to \infty} G_i \quad \text{pointwise on } \mathbb{R}^d \setminus S_{\alpha(i),\beta(i)}.$$  

Step 7 (Finishing the proof): For arbitrary $\delta \in (0, \delta_0)$, let us set

$$\Omega_\delta := K_\delta \setminus \bigcup_{i=1}^{r} S_{\alpha(i),\beta(i)}.$$
Then, Equation (C.17) implies
\[ g_{r+1}^{(k)} - \sum_{i=1}^{r} q_{i}^{(k)} = \sum_{i=1}^{r+1} g_{i}^{(k)} - \left( \sum_{i=1}^{r} g_{i}^{(k)} + q_{i}^{(k)} \right) = f_{nk} - \left( \sum_{i=1}^{r} g_{i}^{(k)} + q_{i}^{(k)} \right) \xrightarrow{\text{pointwise on } \Omega_{\delta}} f - \sum_{i=1}^{r} G_{i}. \]

But since \( g_{r+1}^{(k)} \) and all \( q_{i}^{(k)} \) are affine-linear, and since \( \Omega_{\delta} \) is an open set of positive measure, this implies that there is an affine-linear map \( \psi : \mathbb{R}^{d} \to \mathbb{R} \) satisfying \( f - \sum_{i=1}^{r} G_{i} = \psi \) on \( \Omega_{\delta} \), for arbitrary \( \delta \in (0, \delta_0) \). Note that \( \psi \) is independent of the choice of \( \delta \), and thus
\[ f = \psi + \sum_{i=1}^{r} G_{i} \quad \text{on} \quad \bigcup_{0<\delta<\delta_0} \Omega_{\delta} = K^{\circ} \bigcup_{i=1}^{r} S_{\alpha(i),\beta(i)}. \]

But the latter set is dense in \( K \) (since its complement is a null-set), and \( f \) and \( \psi + \sum_{i=1}^{r} G_{i} \) are continuous. Hence,
\[ f(x) = \psi(x) + \sum_{i=1}^{r} G_{i}(x) = \left( \kappa + \sum_{i=1}^{r} \kappa_i \right) + \left( \zeta + \sum_{i=1}^{r} \zeta_i, x \right) + \sum_{i=1}^{r} c_i \cdot g_{a}(\langle \alpha(i), x \rangle + \beta(i)) \quad \forall x \in K. \]

Recalling from Steps 3 and 4 that \( r < N_0 = N - d - 1 \), this implies \( f \in \mathcal{R} \mathcal{N} \mathcal{N}^{K}_{d,2,d+(r+1)+1,\varrho_a} \subset \mathcal{R} \mathcal{N} \mathcal{N}^{K}_{d,2,N,\varrho_a} \), as claimed. Here, we implicitly used that
\[ \langle \alpha, x \rangle + \beta = g_{a}(\langle \alpha, x \rangle + dB |\alpha|) + \beta - dB |\alpha| \quad \forall x \in K \text{ and arbitrary } \alpha \in \mathbb{R}^{d}, \beta \in \mathbb{R}, \]

since \( \langle \alpha, x \rangle + dB |\alpha| \geq 0 \) for \( x \in K = [-B, B]^{d} \), so that \( g_{a}(\langle \alpha, x \rangle + dB |\alpha|) = \langle \alpha, x \rangle + dB |\alpha| \).

## D Proofs of the results in Section 5

### D.1 Proof of Proposition 5.1

**Step 1:** We first show that if \((f_n)_{n \in \mathbb{N}} \) and \((g_n)_{n \in \mathbb{N}} \) are sequences of continuous functions \( f_n : \mathbb{R}^{d} \to \mathbb{R}^{N} \) and \( g_n : \mathbb{R}^{N} \to \mathbb{R}^{d} \) that satisfy \( f_n \to f \) and \( g_n \to g \) with locally uniform convergence, then also \( g_n \circ f_n \to g \circ f \) locally uniformly.

To see this, let \( R, \varepsilon > 0 \) be arbitrary. On \( \overline{B_{R}(0)} \subset \mathbb{R}^{d} \), we then have \( f_n \to f \) uniformly. In particular, \( C := \sup_{n \in \mathbb{N}} \sup_{x \in \mathbb{R}^{d}} |f_n(x)| < \infty \). But on \( \overline{B_{C}(0)} \subset \mathbb{R}^{N} \), we have \( g_n \to g \) uniformly, so that there is some \( n_1 \in \mathbb{N} \) with \( |g_n(y) - g(y)| < \varepsilon \) for all \( n \geq n_1 \) and all \( y \in \mathbb{R}^{N} \) with \( |y| \leq C \). Furthermore, \( g \) is uniformly continuous on \( \overline{B_{C}(0)} \), so that there is some \( \delta > 0 \) with \( |g(y) - g(z)| < \varepsilon \) for all \( y, z \in \overline{B_{C}(0)} \) with \( |y - z| \leq \delta \). Finally, by the uniform convergence of \( f_n \) to \( f \) on \( \overline{B_{R}(0)} \), we get some \( n_2 \in \mathbb{N} \) with \( |f_n(x) - f(x)| \leq \delta \) for all \( n \geq n_2 \) and all \( x \in \mathbb{R}^{d} \) with \( |x| \leq R \).

Overall, these considerations show for \( n \geq \max\{n_1, n_2\} \) and \( x \in \mathbb{R}^{d} \) with \( |x| \leq R \) that
\[ |g_n(f_n(x)) - g(f(x))| \leq |g_n(f_n(x)) - g(f_n(x))| + |g(f_n(x)) - g(f(x))| \leq \varepsilon + \varepsilon. \]

**Step 2:** We prove the continuity of \( \mathcal{R}^{K}_{\varrho} \). Assume that \( \Phi_n = ((A^{(n)}_1, b^{(n)}_1),\ldots,(A^{(n)}_L, b^{(n)}_L)) \) satisfies \( \Phi_n \to \Phi = ((A_1, b_1),\ldots,(A_L, b_L)) \). For \( \ell \in \{1,\ldots,L-1\} \) set
\[
\alpha^{(n)}_{\ell} : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, x \mapsto \varrho_{\ell}(A_{\ell}^{(n)} x + b_{\ell}^{(n)}),
\]
\[
\alpha_{\ell} : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, x \mapsto \varrho(A_{\ell} x + b_{\ell}),
\]
where \( \varrho := \varrho \otimes \cdots \otimes \varrho \) denotes the \( N_{\ell} \)-fold tensor product of \( \varrho \). Likewise, set
\[
\alpha^{(n)}_{L} : \mathbb{R}^{N_{L-1}} \to \mathbb{R}^{N_L}, x \mapsto A_{L}^{(n)} x + b_{L}^{(n)} \quad \text{and} \quad \alpha_{L} : \mathbb{R}^{N_{L-1}} \to \mathbb{R}^{N_L}, x \mapsto A_{L} x + b_{L}.
\]
By what was shown in Step 1, it is not hard to see for every \( \ell \in \{1, \ldots, L\} \) that \( \alpha^{(n)}_\ell \to \alpha_\ell \) locally uniformly as \( n \to \infty \). By another (inductive) application of Step 1, this shows

\[
R^K_\varphi (\Phi_n) = \alpha^{(n)}_L \circ \cdots \circ \alpha^{(n)}_1 \to \alpha_L \circ \cdots \circ \alpha_1 = R^K_\varphi (\Phi)
\]

with locally uniform convergence. Since \( K \) is compact, this implies uniform convergence on \( K \), and thus completes the proof of the first claim.

**Step 3:** Let \( \varrho_\ell := \varrho \otimes \cdots \otimes \varrho \) be the \( N_\ell \)-fold tensor product of \( \varrho \) in case of \( \ell \in \{1, \ldots, L-1\} \), and set \( \varrho_L := \text{id}_{\mathbb{R}^{N_L}} \). For arbitrary \( x \in K \) and \( \Phi = ((A_1, b_1), \ldots, (A_L, b_L)) \in \mathcal{N}(N_0, \ldots, N_L) \) define inductively \( \alpha_x^{(0)} (\Phi) := x \in \mathbb{R}^d = \mathbb{R}^{N_0} \), and

\[
\alpha_x^{(\ell + 1)} (\Phi) := \varrho_{\ell + 1} (A_{\ell + 1} \alpha_x^{(\ell)} (\Phi) + b_{\ell + 1}) \in \mathbb{R}^{N_{\ell + 1}} \quad \text{for} \quad \ell \in \{0, \ldots, L - 1\}.
\]

Let \( R > 0 \) be fixed, but arbitrary. We will prove by induction on \( \ell \) that

\[
\|\alpha_x^{(\ell)} (\Phi)\|_{R^\infty} \leq C_{\ell, R} \quad \text{and} \quad \|\alpha_x^{(\ell)} (\Phi) - \alpha_x^{(\ell)} (\Psi)\|_{R^\infty} \leq M_{\ell, R} \|\Phi - \Psi\|_{\text{total}}
\]

for suitable \( C_{\ell, R}, M_{\ell, R} > 0 \) and arbitrary \( x \in K \) and \( \Phi, \Psi \in \mathcal{N}(N_0, \ldots, N_L) \) with \( \|\Phi\|_{\text{total}}, \|\Psi\|_{\text{total}} \leq R \).

This will imply that \( R^K_\varphi \) is locally Lipschitz, since

\[
\|R^K_\varphi (\Phi) - R^K_\varphi (\Psi)\|_{\sup} = \sup_{x \in K} |\alpha^{(L)}_x (\Phi) - \alpha^{(L)}_x (\Psi)| \leq M_{L, R} \|\Phi - \Psi\|_{\text{total}}.
\]

The case \( \ell = 0 \) is trivial: On the one hand, \( |\alpha_x^{(0)} (\Phi) - \alpha_x^{(0)} (\Psi)| = 0 \leq \|\Phi - \Psi\|_{\text{total}} \). On the other hand, since \( K \) is bounded, we have \( |\alpha_x^{(0)} (\Phi)| = |x| \leq C_0 \).

For the induction step, let us write \( \Psi = ((B_1, c_1), \ldots, (B_L, c_L)) \), and note

\[
\|A_{\ell + 1} \alpha_x^{(\ell)} (\Phi) + b_{\ell + 1}\|_{R^\infty} \leq N_\ell \|A_{\ell + 1}\|_{\text{max}} \cdot \|\alpha_x^{(\ell)} (\Phi)\|_{R^\infty} + \|b_{\ell + 1}\|_{R^\infty} \leq (1 + N_\ell C_{\ell, R}) \cdot \|\Phi\|_{\text{total}} := K_{\ell + 1, R}.
\]

Clearly, the same estimate holds with \( A_{\ell + 1}, b_{\ell + 1} \) and \( \Phi \) replaced by \( B_{\ell + 1}, c_{\ell + 1} \) and \( \Psi \), respectively. Next, observe that with \( \varrho \) also \( \varrho_{\ell + 1} \) is locally Lipschitz. Thus, there is \( \Gamma_{\ell + 1, R} > 0 \) with

\[
|\varrho_{\ell + 1} (x) - \varrho_{\ell + 1} (y)| \leq \Gamma_{\ell + 1, R} \cdot |x - y| \quad \text{for all} \quad x, y \in \mathbb{R}^{N_{\ell + 1}} \quad \text{with} \quad |x|, |y| \leq K_{\ell + 1, R}.
\]

Therefore,

\[
\|\alpha_x^{(\ell + 1)} (\Phi) - \alpha_x^{(\ell + 1)} (\Psi)\|_{R^\infty}
\]

\[
= \|\varrho_{\ell + 1} (A_{\ell + 1} \alpha_x^{(\ell)} (\Phi) + b_{\ell + 1}) - \varrho_{\ell + 1} (B_{\ell + 1} \alpha_x^{(\ell)} (\Psi) + c_{\ell + 1})\|_{R^\infty}
\]

\[
\leq \Gamma_{\ell + 1, R} \cdot \|\varrho_{\ell + 1} (A_{\ell + 1} \alpha_x^{(\ell)} (\Phi) + b_{\ell + 1}) - (B_{\ell + 1} \alpha_x^{(\ell)} (\Psi) + c_{\ell + 1})\|_{R^\infty}
\]

\[
\leq \Gamma_{\ell + 1, R} \cdot \left( \|A_{\ell + 1} - B_{\ell + 1}\| \cdot \|\alpha_x^{(\ell)} (\Phi)\|_{R^\infty} + \|B_{\ell + 1} \alpha_x^{(\ell)} (\Phi) - \alpha_x^{(\ell)} (\Psi)\|_{R^\infty} + \|b_{\ell + 1} - c_{\ell + 1}\|_{R^\infty} \right)
\]

\[
\leq \Gamma_{\ell + 1, R} \cdot \left( N_\ell \cdot \|\Phi - \Psi\|_{\text{total}} \cdot \|\alpha_x^{(\ell)} (\Phi)\|_{R^\infty} + N_\ell \cdot \|\Psi\|_{\text{total}} \cdot \|\alpha_x^{(\ell)} (\Phi) - \alpha_x^{(\ell)} (\Psi)\|_{R^\infty} + \|\Phi - \Psi\|_{\text{total}} \right)
\]

\[
\leq \Gamma_{\ell + 1, R} \cdot (N_\ell C_{\ell, R} + R N_\ell M_{\ell, R} + 1) \cdot \|\Phi - \Psi\|_{\text{total}}.
\]

**Step 4:** Let \( \varrho \) be Lipschitz with Lipschitz constant \( M \), where we assume without loss of generality that \( M \geq 1 \). With the functions \( \varrho_\ell \) from the preceding step, it is not hard to see that each \( \varrho_\ell \) is \( M \)-Lipschitz, where we use the \( \|\cdot\|_{R^\infty} \) norm on \( \mathbb{R}^{N_\ell} \).

Let \( \Phi = ((A_1, b_1), \ldots, (A_L, b_L)) \in \mathcal{N}(N_0, \ldots, N_L) \), and set \( \alpha_\ell : \mathbb{R}^{N_{\ell - 1}} \to \mathbb{R}^{N_\ell}, x \mapsto \varrho_\ell (A_\ell x + b_\ell) \). Then \( \alpha_\ell \) is Lipschitz with Lipschitz constant \( M \cdot \|A_\ell\|_{R^\infty} \leq M \cdot N_{\ell - 1} \cdot \|A\|_{\text{max}} \leq M N_{\ell - 1} \cdot \|\Phi\|_{\text{scaling}} \). Thus, we finally see that \( R^K_\varphi (\Phi) = \alpha_L \circ \cdots \circ \alpha_1 \) is Lipschitz with Lipschitz constant \( M^L \cdot N_0 \cdots N_{L - 1} \cdot \|\Phi\|^L_{\text{scaling}} \). This proves the final claim of the lemma when choosing the \( \ell^\infty \)-norm on \( \mathbb{R}^d \) and \( \mathbb{R}^{N_\ell} \). Of course, choosing another norm than the \( \ell^\infty \)-norm can be done, at the cost of possibly enlarging the constant \( C \) in the statement of the lemma. □
D.2 Proof of Theorem 5.2

Step 1: For \( a > 0 \), define

\[
 f_a : \mathbb{R} \to \mathbb{R}, \ x \mapsto g(x + a) - 2g(x) + g(x - a).
\]

Our claim in this step is that there is some \( a > 0 \) with \( f_a \not\equiv \text{const} \).

Let us assume towards a contradiction that this fails, i.e., \( f_a \equiv c_a \) for all \( a > 0 \). Since \( g \) is Lipschitz continuous, it is at most of linear growth, so that \( g \) is a tempered distribution. Elementary properties of the Fourier transform (for tempered distributions) show

\[
c_a \cdot \delta_0 = \hat{f}_a = \hat{g} \cdot g_a \quad \text{with} \quad g_a : \mathbb{R} \to \mathbb{R}, \ \xi \mapsto e^{2\pi i a \xi} - 2 + e^{-2\pi i a \xi}.
\]

Next, note for \( z(\xi) := e^{2\pi i a \xi} \neq 0 \) that

\[
g_a(\xi) = z(\xi) - 2 + z^{-1}(\xi) = z^{-1}(\xi) \cdot (z^2(\xi) - 2z(\xi) + 1) = z^{-1}(\xi) \cdot (z(\xi) - 1)^2 \neq 0,
\]

as long as \( z(\xi) \neq 1 \), that is, as long as \( \xi \not\in a^{-1}\mathbb{Z} \).

Let \( \varphi \in C_c^\infty(\mathbb{R} \setminus \{0\}) \) be fixed, but arbitrary. This implies supp \( \varphi \subset \mathbb{R} \setminus a^{-1}\mathbb{Z} \) for some sufficiently small \( a > 0 \). Since \( g_a \neq 0 \) on the compact set supp \( \varphi \), it is not hard to see that there is some smooth, compactly supported function \( h \) with \( h \cdot g_a \equiv 1 \) on the support of \( \varphi \). All in all, we thus get

\[
(\hat{g} \cdot \varphi)_{S',S} = (\hat{g} \cdot g_a)_{S',S} = (\hat{f}_a, h \cdot \varphi)_{S',S} = c_a \cdot h(0) \cdot \varphi(0) = 0.
\]

Since \( \varphi \in C_c^\infty(\mathbb{R} \setminus \{0\}) \), we have shown supp \( \varphi \subset \{0\} \). But by [19, Corollary 2.4.2], this implies that \( g \) is a polynomial. Since only affine-linear polynomials are Lipschitz continuous (on the whole real line), \( g \) must be affine-linear, contradicting the prerequisites of the proposition.

Step 2: In this step we construct certain continuous functions \( F_n : \mathbb{R}^d \to \mathbb{R} \) which satisfy \( \text{Lip}(F_n|_K) \to \infty \) and \( \|F_n\|_{L^\infty(y)} \to 0 \). We will then use these functions in the next step to construct the desired networks \( \Phi_n \).

We first note that each function \( f_a \) from Step 1 is bounded. In fact, if \( g \) is \( M \)-Lipschitz, then

\[
|f_a(x)| \leq |g(x + a) - g(x)| + |g(x - a) - g(x)| \leq 2M|a|.
\]

Next, since \( g \) is Lipschitz continuous, it is (locally) absolutely continuous. Thus, \( g \) is differentiable almost everywhere and satisfies \( g(y) - g(x) = \int_x^y g'(t) \, dt \) for \( x < y \). By assumption, \( g \) is not constant. Thus, there is some \( t_0 \in \mathbb{R} \) with \( g'(t_0) \neq 0 \). Thus, Proposition 2.4.3 shows that there is a neural network \( \Phi \in \mathcal{N}'_1.L^{-1,L} \) such that \( \psi := R^\infty_a(\Phi) \) is differentiable at the origin with \( \psi(0) = 0 \) and \( \psi'(0) = 1 \). By definition, this means that there is a function \( \delta : \mathbb{R} \to \mathbb{R} \) with \( \psi(x) = x + x \cdot \delta(x) \) with \( \delta(x) \to 0 = \delta(0) \) as \( x \to 0 \).

Next, since \( K \) has nonempty interior, there is some \( x_0 \in \mathbb{R}^d \) and some \( r > 0 \) with \( x_0 + [-r,r]^d \subset K \). Let us now choose \( a > 0 \) with \( f_a \neq \text{const} \) (the existence of such an \( a > 0 \) is implied by the previous step), and define

\[
F_n : \mathbb{R}^d \to \mathbb{R}, \ x \mapsto \psi\left(n^{-1} \cdot f_a(n^2 \cdot (x - x_0)_{1})\right).
\]

Since \( f_a \) is not constant, there are \( b, c \in \mathbb{R} \) with \( b < c \) and \( f_a(b) \neq f_a(c) \). Since \( \delta(x) \to 0 \) as \( x \to 0 \) and by the boundedness of \( f_a \) (see Equation (D.1)), we see that there is some \( \kappa > 0 \) and some \( n_1 \in \mathbb{N} \) with

\[
|f_a(b) - f_a(c)| - |f_a(b)| \cdot |\delta(f_a(b)/n)| - |f_a(c)| \cdot |\delta(f_a(c)/n)| \geq \kappa > 0 \quad \text{for all} \ n \geq n_1.
\]

Let us set \( x_n := x_0 + n^{-2} \cdot (b,0,\ldots,0) \in \mathbb{R}^d \) and \( y_n := x_0 + n^{-2} \cdot (c,0,\ldots,0) \in \mathbb{R}^d \), and observe \( x_n, y_n \in K \) for \( n \in \mathbb{N} \) large enough. We have \( |x_n - y_n| = n^{-2} \cdot |b - c| \). Furthermore, using the expansion \( \psi(x) = x + x \cdot \delta(x) \), and noting \( f_a(n^2(x_n - x_0)_{1}) = f_a(b) \) as well as \( f_a(n^2(y_n - x_0)_{1}) = f_a(c) \), we get

\[
|F_n(x_n) - F_n(y_n)| = |\psi(f_a(b)/n) - \psi(f_a(c)/n)|
\]

\[
= \left| f_a(b) \cdot \frac{b}{n} + f_a(c) \cdot \frac{c}{n} - f_a(b) \cdot \frac{b}{n} - f_a(c) \cdot \frac{c}{n} \right|
\]

\[
\geq 1 \cdot \left( |f_a(b) - f_a(c)| - |f_a(b)| \cdot |\delta(f_a(b)/n)| - |f_a(c)| \cdot |\delta(f_a(c)/n)| \right) \geq \kappa/n,
\]
as long as \( n \geq n_1 \) is so large that \( x_n, y_n \in K \). But this implies

\[
\text{Lip}(F_n | K) \geq \frac{|F_n(x_n) - F_n(y_n)|}{|x_n - y_n|} \geq \frac{\kappa/n}{n^{-2} \cdot |b - c|} = n \cdot \frac{\kappa}{|b - c|} \to \infty.
\]

It remains to show \( F_n | K \to 0 \) uniformly. To this end, let \( \varepsilon > 0 \) be arbitrary. By continuity of \( \psi \) at 0, there is some \( \delta > 0 \) with \( |\psi(x)| \leq \varepsilon \) for \( |x| \leq \delta \). But equation (D.1) shows \( |n^{-1} \cdot f(n^{-2} \cdot (x-x_0)_1)| \leq n^{-1} \cdot 2M |a| \leq \delta \) for all \( x \in \mathbb{R}^d \) and all \( n \geq n_0 \), with \( n_0 = n_0(M, a, \delta) \in \mathbb{N} \) suitable. Hence, \( |F_n(x)| \leq \varepsilon \) for all \( n \geq n_0 \) and \( x \in \mathbb{R}^d \). We have thus shown \( F_n \to 0 \) uniformly on \( \mathbb{R}^d \) and not only on \( K \).

**Step 3:** In this step, we give the construction of the networks \( \Phi_n \). For \( n \in \mathbb{N} \) define

\[
A_1^{(n)} := n^2 \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{3 \times d} \quad \text{and} \quad b_1^{(n)} := \begin{pmatrix} -n^2 \cdot (x_0)_1 + a \\ -n^2 \cdot (x_0)_1 \\ -n^2 \cdot (x_0)_1 - a \end{pmatrix} \in \mathbb{R}^3,
\]

as well as \( A_2^{(n)} := n^{-1} \cdot (1, -2, 1) \in \mathbb{R}^{1 \times 3} \) and \( b_2^{(n)} := 0 \in \mathbb{R}^1 \). A direct calculation shows

\[
R_{\tilde{\theta}}^\varepsilon (\Phi_n^{(0)})(x) = n^{-1} \cdot f_a(n^2 \cdot (x - x_0)_1) \quad \text{for all } x \in \mathbb{R}^d \text{ with } \Phi_n^{(0)} := ((A_1^{(n)}, b_1^{(n)}), (A_2^{(n)}, b_2^{(n)})).
\]

Thus, with the concatenation operation introduced in Definition 2.8, the network \( \Phi_1^{(1)} := \Phi \bullet \Phi_n^{(0)} \) satisfies \( R_{\tilde{\theta}}^\varepsilon (\Phi_1^{(1)}) = F_n | K \). Furthermore, it is not hard to see that \( \Phi_1^{(1)} \) has \( L \) layers and has the architecture \((d, 3, 1, \ldots, 1)\). From this and because of \( N_1 \geq 3 \), by Lemma 2.9 there is a network \( \Phi_n \) with architecture \( S = (d, N_1, \ldots, N_{L-1}, 1) \) and \( R_{\tilde{\theta}}^\varepsilon (\Phi_n) = F_n | K \). By Step 2, this implies \( R_{\tilde{\theta}}^\varepsilon (\Phi_n) = F_n | K \to 0 \) uniformly on \( K \), as well as \( \text{Lip}(R_{\tilde{\theta}}^\varepsilon (\Phi_n)) \to \infty \) as \( n \to \infty \).

**Step 4:** In this step, we establish the final property which is stated in the proposition. For this, let us assume towards a contradiction that there is a family of networks \((\Psi_n)_{n \in \mathbb{N}}\) with architecture \( S \) and \( R_{\tilde{\theta}}^\varepsilon (\Psi_n) = R_{\tilde{\theta}}^\varepsilon (\Phi_n) \), some \( C > 0 \), and a subsequence \((\Psi_n)_r \subseteq \mathbb{N} \) with \( \|\Psi_n\|_{\text{scaling}} \leq C \) for all \( r \in \mathbb{N} \). In view of the last part of Proposition 5.1 there is a constant \( C' = C'(\tilde{\theta}, S) > 0 \) with

\[
\text{Lip}(R_{\tilde{\theta}}^\varepsilon (\Psi_n)) = \text{Lip}(R_{\tilde{\theta}}^\varepsilon (\Psi_n)) \leq C' \cdot \|\Psi_n\|_{\text{scaling}} \leq C' \cdot C^L,
\]

which is in contradiction to \( \text{Lip}(R_{\tilde{\theta}}^\varepsilon (\Phi_n)) \to \infty \). \(\square\)

**D.3 Proof of Corollary 5.3**

Let us denote the range of the realization map by \( R \). By definition, \( R_{\tilde{\theta}}^\varepsilon \) is a quotient map if and only if

\[
\forall M \subseteq R : \quad M \text{ open } \iff (R_{\tilde{\theta}}^\varepsilon)^{-1}(M) \text{ open}.
\]

Clearly, by switching to complements, we can equivalently replace replace “open” by “closed” everywhere.

Now, let us choose neural net \((\Phi_n)_{n \in \mathbb{N}}\) as in Proposition 5.2 and write \( F_n := R_{\tilde{\theta}}^\varepsilon (\Phi_n) \). Because of \( \text{Lip}(F_n | K) \to \infty \), we have \( F_n | K \neq 0 \) for all \( n \geq n_0 \) with \( n_0 \in \mathbb{N} \) suitable. Define \( M := \{F_n | K : n \geq n_0\} \subset R \).

Note that \( M \subseteq C(K) \) is not closed, since \( F_n | K \to 0 \) uniformly, but \( 0 \notin M \). Once we show that \( (R_{\tilde{\theta}}^\varepsilon)^{-1}(M) \) is closed, we will have shown that \( R_{\tilde{\theta}}^\varepsilon \) is not a quotient map.

Thus, let \((\Psi_n)_{n \in \mathbb{N}}\) be a sequence in \((R_{\tilde{\theta}}^\varepsilon)^{-1}(M)\) and assume \( \Psi_n \to \Psi \) as \( n \to \infty \). In particular, \( \|\Psi_n\|_{\text{scaling}} \leq C \) for some \( C > 0 \) and all \( n \in \mathbb{N} \). We want to show \( \Psi \in (R_{\tilde{\theta}}^\varepsilon)^{-1}(M) \) as well. Since \( \Psi_n \in (R_{\tilde{\theta}}^\varepsilon)^{-1}(M) \), there is for each \( n \in \mathbb{N} \) some \( r_n \in \mathbb{N} \) with \( R_{\tilde{\theta}}^\varepsilon (\Psi_n) = F_{r_n} | K \). Now there are two cases:

**Case 1:** The family \((r_n)_{n \in \mathbb{N}}\) is infinite. But in view of Proposition 5.1 we have

\[
\text{Lip}(F_{r_n} | K) \leq \text{Lip}(R_{\tilde{\theta}}^\varepsilon (\Psi_n)) \leq C' \cdot \|\Psi_n\|_{\text{scaling}} \leq C' \cdot C^L.
\]
for a suitable constant $C' = C'(\varrho, N_0, \ldots, N_L)$, in contradiction to the fact that $\text{Lip}(F_n|_K) \to \infty$ as $r_n \to \infty$. Thus, this case cannot occur.

**Case 2:** The family $(r_n)_{n \in \mathbb{N}}$ is finite. Thus, there is some $N \in \mathbb{N}$ with $r_n = N$ for infinitely many $n \in \mathbb{N}$, that is, $R^K_{\varrho}(\Psi_n) = F_{r_n}|_K = F_N|_K$ for infinitely many $n \in \mathbb{N}$. But since $R^K_{\varrho}(\Psi_n) \to R^K_{\varrho}(\Psi)$ as $n \to \infty$ (by the continuity of the realization map), this implies $R^K_{\varrho}(\Psi) = F_N|_K \in M$, as desired. □

**References**

[1] H. Amann and J. Escher. *Analysis. III*. Birkhäuser Verlag, Basel, 2009.

[2] P. M. Anselone and J. Korevaar. Translation invariant subspaces of finite dimension. *Proc. Amer. Math. Soc.*, 15:747–752, 1964.

[3] M. Anthony and P. L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.

[4] A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.

[5] J. Bergstra, G. Desjardins, P. Lamblin, and Y. Bengio. Quadratic polynomials learn better image features. Technical Report 1337, Département d’Informatique et de Recherche Opérationnelle, Université de Montréal, Apr. 2009.

[6] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *preprint arXiv:1705.01714*, 2018.

[7] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[8] B. Carlile, G. Delamarter, P. Kinney, A. Marti, and B. Whitney. Improving deep learning by inverse square root linear units (ISRLUs). *arXiv preprint arXiv:1710.09967*, 2017.

[9] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015.

[10] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on Learning Theory*, pages 698–728, 2016.

[11] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Am. Math. Soc.*, 39:1–49, 2002.

[12] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal*, 2(4):303–314, 1989.

[13] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Audio, Speech, Language Process.*, 20(1):30–42, 2012.

[14] I. Daubechies. *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[15] W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat.*, 5(4):349–380, 2017.

[16] G. Folland. *Real Analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999.
[17] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011.

[18] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. [http://www.deeplearningbook.org](http://www.deeplearningbook.org)

[19] L. Grafakos. *Classical Fourier analysis*, volume 249 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2008.

[20] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[22] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97, 2012.

[23] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, 1989.

[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[25] K. Kawaguchi. Deep learning without poor local minima. In *Adv. Neural Inf. Process. Syst.*, pages 586–594, 2016.

[26] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inf. Process. Syst. 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[27] I. E. Lagaris, A. Likas, and D. I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Netw.*, 9(5):987–1000, 1998.

[28] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[29] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.*, 6(6):861 – 867, 1993.

[30] B. Liao, C. Ma, L. Xiao, R. Lu, and L. Ding. An arctan-activated WASD neural network approach to the prediction of Dow Jones Industrial Average. In *Advances in Neural Networks - ISNN 2017 - 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21-26, 2017, Proceedings, Part I*, pages 120–126, 2017.

[31] A. Maas, Y. Hannun, and A. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[32] V. Maiorov and A. Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1-3):81–91, 1999.

[33] S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[34] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943.
[35] H. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.*, 8(1):164–177, 1996.

[36] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80, Feb 1993.

[37] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS’14, pages 2924–2932, Cambridge, MA, USA, 2014. MIT Press.

[38] V. Nair and G. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, ICML’10, pages 807–814, USA, 2010. Omnipress.

[39] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *arXiv:1709.05289*, 2018.

[40] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[41] W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991.

[42] I. Safran and O. Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2979–2987, 2017.

[43] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural netw.*, 61:85–117, 2015.

[44] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] N. Usunier, G. Synnaeve, Z. Lin, and S. Chintala. Episodic exploration for deep deterministic policies: An application to StarCraft micromanagement tasks. *arXiv preprint arXiv:1609.02993*, 2016.

[47] R. Vidal, J. Bruna, R. Giryes, and S. Soatto. Mathematics of deep learning. *arXiv preprint arXiv:1712.04741*, 2017.

[48] T. Wiatowski and H. Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Trans. Inf. Theory*, 64(3):1845–1866, 2018.

[49] C. Wu, P. Karanasou, M. J. Gales, and K. C. Sim. Stimulated deep neural network for speech recognition. Technical report, University of Cambridge, 2016.

[50] G. N. Yannakakis and J. Togelius. *Artificial Intelligence and Games*. Springer, 2017.

[51] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.