Neural Networks 140 (2021) 121–129

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

On decision regions of narrow deep neural networks

Hans-Peter Beise [a,*], Steve Dias Da Cruz [b,c], Udo Schröder [b]

[a] Department of Computer Science, Trier University of Applied Sciences, Germany
[b] Department of Basics & Mathematical Models, IEE S.A., Luxembourg
[c] Department of Computer Science, University of Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 21 July 2020
Received in revised form 12 January 2021
Accepted 22 February 2021
Available online 10 March 2021

Keywords:
Expressive power
Approximation by network functions
Neural networks
Decision regions
Width of neural networks

ABSTRACT

We show that for neural network functions that have width less or equal to the input dimension all connected components of decision regions are unbounded. The result holds for continuous and strictly monotonic activation functions as well as for the ReLU activation function. This complements recent results on approximation capabilities by Hanin and Sellke (2017) and connectivity of decision regions by Nguyen et al. (2018) for such narrow neural networks. Our results are illustrated by means of numerical experiments.

1. Introduction

In recent years machine learning experienced a remarkable evolution mainly due to the progress achieved with deep neural networks, c.f. Goodfellow, Bengio, Courville, and Bengio (2016), He, Zhang, Ren, and Sun (2016), Hinton et al. (2012), Krizhevsky, Sutskever, and Hinton (2012), Nguyen, Mukkamala, and Hein (2018) and Schmidhuber (2015) for an overview and theoretical background. In the course of this astonishing success in applications, there has been huge progress in the research towards understanding the mathematical properties of neural network functions. As part of this, the approximation properties, or expressiveness, of neural network functions have attracted intense interest in recent research. The central result in this field is the classic *universal approximation theorem*, which states that any continuous function can be approximated with arbitrary accuracy (in terms of uniform approximation or L_p norms) by neural network functions that have only one hidden layer for nearly every activation function c.f. Cybenko (1989), Hornik (1991), see also Scarselli and Tsoi (1998) for an overview of classical results and Gnecco, Kurková, and Sanguineti (2011) for a comparison with other common approximation techniques. In last years, this kind of results where further improved in several respects (Bölcskei, Grohs, Kutyniok, & Petersen, 2019; Costarelli & Spigler, 2013; Elfwing, Uchibe, & Doya, 2018; Gühring, Kutyniok, & Petersen, 2020; Guliyev & Ismailov, 2018; Petersen & Voigtlaender, 2018, 2020; Yarotsky, 2017, 2018) for some recent results in this direction.

On the other hand, from empirical observations it turned out that depth has a significant impact on the performance of neural networks which is why a lot of research has been dedicated to analyze the effect of depth on the expressive power of neural networks, c.f. Cohen, Sharir, and Shashua (2016), Lin, Tegmark, and Rolnick (2017), Mhaskar and Poggio (2016), Montufar, Pascanu, Cho, and Bengio (2014), Raghu, Poole, Kleinberg, Ganguli, and Sohl-Dickstein (2017), Rolnick and Tegmark (2018) and Telgarsky (2016).

It is however clear that besides depth a neural network needs a certain width in order to be a universal approximator, c.f. Remark 2. The role of width with regard to approximation properties is investigated in a line of works (Hanin, 2019; Hanin & Sellke, 2017; Johnson, 2018; Kidger & Lyons, 2020; Lu, Pu, Wang, Hu, & Wang, 2017; Park, Yun, Lee, & Shin, 2021), and the importance of width from the perspective of decision regions is pointed out in Nguyen and Hein (2017). In this work, we follow the latter work and investigate the expressiveness of networks of bounded width mainly in the latter terms.

Let us introduce the following notation. By $\|\cdot\|$ we denote the Euclidean norm in $\mathbb{R}^d$. For a set $D \subset \mathbb{R}^d$ we denote by D° the set of interior points, by $\overline{D}$ its closure, by $\partial D = \overline{D} \setminus D^\circ$ the boundary of D and for $f : \mathbb{R}^d \to \mathbb{R}^m$ we set $\|f\|_D := \sup\{\|f(x)\| : x \in D\}$. For $W \in \mathbb{R}^{m\times n}$ we denote by $\|W\|_{\mathrm{op}} := \max_{\|x\|\le 1} \|Wx\|$ the operator norm. We consider neural network functions $F : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ where d_{in} is called the input dimension and d_{out} the output dimension. Our network functions have the following form $F := W_L \circ A_{L-1} \circ \ldots \circ A_1$ where $A_j(x) = \sigma(W_j x + b_j)$ with $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$ (weights), $b_j \in \mathbb{R}^{d_j}$ (bias) and $\sigma : \mathbb{R} \to \mathbb{R}$ the activation function. We emphasize that σ and the preimage σ^{-1} are understood to be applied elementwise when applied to vectors or subsets of $\mathbb{R}^d$.

* Corresponding author.
E-mail address: H.Beise@hochschule-trier.de (H.-P. Beise).

https://doi.org/10.1016/j.neunet.2021.02.024
0893-6080/

The widely applied activation function, called *rectified linear unit*, or shortly ReLU, is defined by $t \mapsto \max\{t, 0\}$. We set $d_0 := d_{in}$ and $d_L = d_{out}$ and call d_j the *width* of layer $j = 1, \ldots, L$. The width of the network is defined as $\omega := \omega_F = \max\{d_j : j = 1, \ldots, L\}$ and L is called the depth of the network. Adapting the notation from Hanin and Sellke (2017) to our needs, we define $\omega_{\min}(\sigma, d_{in}, d_{out})$ to be the minimum width such that for every continuous $f : [a, b]^{d_{in}} \to \mathbb{R}^{d_{out}}$ and $\varepsilon > 0$ there exists a network function $F : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ with activation function σ and $\omega_F \leq \omega_{\min}(\sigma, d_{in}, d_{out})$ such that $\|f - F\|_{[a,b]^{d_{in}}} < \varepsilon$, where $a < b$ are some real numbers.

2. Related work

It is a fundamental observation that the expressiveness of a function that implements a classifier model is closely related to the notion of decision regions.

Definition 1. For a network function $F = (F_1, \ldots, F_{d_{out}}) : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ and $j \in \{1, \ldots, d_{out}\}$, the set $C_j := \{x \in \mathbb{R}^{d_{in}} : F_j(x) > F_k(x), \text{for all } k \neq j\}$ is called a decision region (for class j). If $K \subset \mathbb{R}^{d_{in}}$, then $C_j \cap K$ is called the decision region (of class j) in K.

Definition 2. A set $C \subset \mathbb{R}^d$ is said to be connected if there exist no disjoint open sets $U, V \subset \mathbb{R}^d$ such that $C \subset U \cup V$ and $C \cap U$ and $C \cap V$ are non-empty. Let $K \subset \mathbb{R}^d$ be some compact set, then C is said to be connected in K, if such sets U, V do not exist for $C \cap K$.

It is known that open sets can be decomposed into a disjoint family of connected open sets. In what follows, an element of this family is called connected *component*. These components are maximal in the sense that they are no proper subset of another connected subset of the original set.

It should be noted that the notion of connectivity is not equivalent to the more intuitive but more restrictive term of path-connectivity (c.f. Definition 3).

Definition 3. A set $C \subset \mathbb{R}^d$ is said to be path-connected if for all $x_1, x_2 \in C$ there exists a continuous path $\gamma : [0, 1] \to C$ such that $\gamma(0) = x_1$ and $\gamma(1) = x_2$. Let $K \subset \mathbb{R}^d$ be some compact set, then C is said to be path-connected in K, if for all $x_1, x_2 \in C \cap K$ there exists a continuous path $\gamma : [0, 1] \to C \cap K$ such that $\gamma(0) = x_1$ and $\gamma(1) = x_2$.

In Fawzi, Moosavi-Dezfooli, Frossard, and Soatto (2017) the decision regions of deep neural networks are investigated empirically. By experiments with ImageNet networks, the authors observe that two samples that are predicted to belong to the same class can be connected by a continuous path, where the path is found by a dedicated algorithm that is also provided in Fawzi et al. (2017). The latter article further gives some interesting insight regarding the local curvature of decision boundaries.

In contrast to the experimentally driven work, our work is more related to Nguyen et al. (2018) which treats this problem from a theoretical perspective. A central result therein is the following.

Theorem 1 (*Theorem 3.10, Nguyen et al., 2018*). *Let F be a neural network function such that $d_{in} = d_1 \geq d_2 \geq \cdots d_L = d_{out}$ and each weight matrix has full rank. If the activation function σ is continuous, strictly monotonically increasing and satisfies $\sigma(\mathbb{R}) = \mathbb{R}$, then every decision region is connected.*

In parallel to this, the approximation properties of width bounded neural networks have been studied in several works, Hanin (2019), Hanin and Sellke (2017), Johnson (2018), Kidger and Lyons (2020), Lu et al. (2017) and Park et al. (2021). A common goal in these lines of research is to provide upper and lower bounds on the minimum width required to ensure universal approximation in $C(K, \mathbb{R}^d_{out})$, i.e the space of continuous function from a compact set $K \subset \mathbb{R}^{d_{in}}$ to $\mathbb{R}^{d_{out}}$, and in $L_p(\mathbb{R}^{d_{in}}, \mathbb{R}^{d_{out}})$ or $L_p(K, \mathbb{R}^{d_{out}})$ (K again a compact set in $\mathbb{R}^{d_{in}}$). For the case of $C(K, \mathbb{R}^{d_{out}})$, it is proven in Hanin and Sellke (2017) that

$$d_{in} + 1 \leq \omega_{\min}(\text{ReLU}, d_{in}, d_{out}) \leq d_{in} + d_{out}. \tag{1}$$

Further, in Johnson (2018) it is shown that

$$\omega_{\min}(\sigma, d_{in}, d_{out}) \geq d_{in} + 1, \tag{2}$$

for activation functions σ that admit arbitrary accurate uniform approximation by injective continuous functions on arbitrary compact subset of $\mathbb{R}$.

Recently, results as stated in (1) and (2), have been extended and partially improved in Kidger and Lyons (2020) and Park et al. (2021). We refer to the latter work for a good overview.

In this work, we take a similar focus as in Nguyen et al. (2018), namely limitations regarding the decision regions of neural networks of bounded width. Our main result complements the results on the lower bound found in Hanin and Sellke (2017), c.f. (1), and a result from Nguyen et al. (2018), c.f. 1, in terms of decision regions. We show that for network functions with maximum width d_{in} and both types of activation functions, strictly monotonic or ReLU, the decision regions do not need to be connected, as opposed to Nguyen et al. (2018), c.f. (1), but are unbounded and therefore do not admit arbitrary accurate approximation of all continuous functions on compact sets.

The impossibility of universal approximation in the latter case is also obtained in Johnson (2018), c.f. (2). To this end, it is shown in Johnson (2018, Lemma 4) that the contour lines of network functions $F : \mathbb{R}^{d_{in}} \to \mathbb{R}$ are unbounded. Our main result extends this by giving a related statement for decision regions of vector valued network functions.

3. Results

In our main result we show that in case of $\omega_F \leq d_{in}$ and continuous and strictly monotonic activation functions or ReLU activation the components of the decision regions are unbounded. This implies that they intersect the boundary of the natural bounding box of input data.

To this end, we exploit the basic observation of the following lemma and show that for certain narrow neural networks this can be continued to the input domain. The content of the lemma is well-known. We give a short proof for interested readers.

Lemma 1. *Let $A \in \mathbb{R}^{n \times d}$ with $n \leq d$, $b \in \mathbb{R}^n$ and $x \in \mathbb{R}^d$ with $Ax < b$. Then there exists a non-zero $v \in \mathbb{R}^d$ such that $A(x+\lambda v) < b$ for all $\lambda \geq 0$*

The geometrical interpretation of the previous lemma is that a convex set in $\mathbb{R}^d$, that is described by less than $d+1$ hypersurfaces, cannot enclose a point. It directly follows that for every compact set K such that if $C = \{x \in K : Ax < b\} \neq \emptyset$ then $C \cap \partial K \neq \emptyset$.

Proof. In case where A is invertible let w be the unique solution of $Aw = b$. One directly verifies that $v = x - w$ has the desired property. Otherwise, we set v equal to one of the (non-zero) vectors that are orthogonal to the rows of A. $\square$

In what follows, it will be convenient to argue with invertible weight matrices. The following remark clarifies that this is justified in our setting.

Lemma 2. *Let $W \in \mathbb{R}^{d\times d}$, $K \subset \mathbb{R}^d$ be a compact set and $\phi, \tilde{\phi} : K \to \mathbb{R}^d$ be continuous mappings such that for some $\varepsilon_1 > 0$*

$$\|\tilde{\phi} - \phi\|_K < \varepsilon_1.$$

Then for every $\varepsilon_2 > \|W\|_{\mathrm{op}}\,\varepsilon_1$, there exists an invertible $\tilde{W} \in \mathbb{R}^{d\times d}$ such that

$$\|\tilde{W}\tilde{\phi} - W\phi\|_K < \varepsilon_2.$$

Proof. For $\delta = \varepsilon_2 - \|W\|_{\mathrm{op}}\,\varepsilon_1$ we can find an invertible $\tilde{W} \in \mathbb{R}^{d\times d}$ such that

$$\|\tilde{W}\tilde{\phi} - W\tilde{\phi}\|_K < \delta.$$

With the triangle inequality and elementary properties of the operator norm it follows that for every $x \in K$ we have

$$\begin{aligned}&\|\tilde{W}\tilde{\phi}(x) - W\phi(x)\|\\ &\le \|\tilde{W}\tilde{\phi}(x) - W\tilde{\phi}(x)\| + \|W\|_{\mathrm{op}}\,\|\tilde{\phi}(x) - \phi(x)\| < \varepsilon_2.\end{aligned}$$

This concludes the proof. □

Lemma 3. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a neural network function given by $F := A_L \circ \ldots \circ A_1$ where $A_j(x) = \sigma(W_j x + b_j)$ with $W_j \in \mathbb{R}^{d\times d}$ for $j = 1, \ldots, L$, and with a continuous activation function σ. For a compact $K \subset \mathbb{R}^d$ and $\varepsilon > 0$, there exist invertible $\tilde{W}_j \in \mathbb{R}^{d\times d}$, $j = 1, \ldots, L$ such that the network function $\tilde{F} : \mathbb{R}^d \to \mathbb{R}^d$ defined by $\tilde{F} := \tilde{A}_L \circ \ldots \circ \tilde{A}_1$ with $\tilde{A}_j(x) = \sigma(\tilde{W}_j x + b_j)$, $j = 1, \ldots, L$, approximates F in a way that*

$$\|\tilde{F} - F\|_K < \varepsilon.$$

Proof. We set $M := \max\{1, \|W_1\|_{\mathrm{op}}, \|W_2\|_{\mathrm{op}}, \ldots, \|W_L\|_{\mathrm{op}}\}$. We want to use the uniform continuity of σ on compact sets. To this end let $Q_1, Q_2, \ldots, Q_L \subset \mathbb{R}^d$, $Q_0 := K$ be compact sets iteratively defined by

$$\begin{aligned}Q_j := \bigcup_{x\in Q_{j-1}} &\big(\{y : \|W_j x + b_j - y\| \le 1\} \cup \{y : \|\sigma(W_j x + b_j)\\ &-y\| \le 1\}\big),\ j = 1, \ldots, L-1\end{aligned}$$

and $Q := \bigcup_{j=0}^{L-1} Q_j$. This yields a sufficiently large compact set that will contain the output of every intermediate layer, with and without activation, of the network function that will turn out from our below construction for every input $x \in K$. As a component wise applied function from $\mathbb{R}^d$ to $\mathbb{R}^d$, σ is uniformly continuous on Q and we hence find $1 > \delta_{L-1}, \ldots, \delta_1, \delta_0 > 0$ corresponding to $\varepsilon =: \delta_L$ such that $\delta_j > M\delta_{j-1}$, $j = 1, \ldots, L$, and for all $x, y \in Q$ with $\|x - y\| < M\delta_j$ the following estimate holds

$$\|\sigma(x) - \sigma(y)\| < \delta_{j+1}, \quad j = 1, \ldots, L-1. \tag{3}$$

Now, one can iteratively apply Lemma 2 to yield the desired $\tilde{W}_j$. In fact, assume that we have already determined invertible $\tilde{W}_1, \ldots, \tilde{W}_{j-1} \in \mathbb{R}^{d\times d}$ such that the corresponding $\tilde{A}_k(x) = \sigma(\tilde{W}_k x + b_k)$, $k = 1, \ldots, j-1$ satisfy

$$\|\tilde{A}_{j-1} \circ \ldots \tilde{A}_2 \circ \tilde{A}_1 - A_{j-1} \circ \ldots A_2 \circ A_1\|_K < \delta_{j-1}$$

and $\tilde{A}_{j-1} \circ \ldots \tilde{A}_2 \circ \tilde{A}_1(x) \in Q_{j-1}$. Then Lemma 2 delivers an invertible $\tilde{W}_j \in \mathbb{R}^{d\times d}$ such that

$$\|\tilde{W}_j - W_j\|_{Q_{j-1}} < M\delta_{j-1},$$

where the matrices in the norms of the latter inequality are interpreted as the corresponding linear mappings. The application of (3) to $\tilde{A}_j := \sigma(\tilde{W}_j x + b_j)$ yields

$$\|\tilde{A}_j \circ \ldots \tilde{A}_2 \circ \tilde{A}_1 - A_j \circ \ldots A_2 \circ A_1\|_K < \delta_j.$$

From the construction of Q_j and $\delta_j < 1$ it further follows that $\tilde{A}_j(x) \in Q_j$ for all $x \in Q_{j-1}$. This concludes the general step and the assertion follows inductively. □

Proposition 1. *Let $F : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ be a neural network function $F(x) = W_L(A_{L-1} \circ \ldots \circ A_1(x)) + b_L$ with $\omega_F \le d_{in}$ and L layers. Then for a given $\varepsilon > 0$, there exist invertible $\tilde{W}_j \in \mathbb{R}^{d_{in}\times d_{in}}$, $j = 1, \ldots, L-1$ such that $\tilde{F}(x) = W_L(\tilde{A}_{L-1} \circ \ldots \circ \tilde{A}_1(x)) + b_L$ with $\tilde{A}_j(x) = \sigma(\tilde{W}_j x + b_j)$, $j = 1, \ldots, L-1$ satisfies*

$$\|\tilde{F} - F\|_K < \varepsilon.$$

Proof. By padding with zeros rows and zero components, we can consider $x \mapsto W_j x + b_j$, $j = 1, \ldots, L-1$ as mapping from $\mathbb{R}^{d_{in}}$ to $\mathbb{R}^{d_{in}}$ so that $W_j \in \mathbb{R}^{d_{in}\times d_{in}}$ and $b_j \in \mathbb{R}^{d_{in}}$, $j = 1, \ldots, L-1$. Hence, since the final layer is the same for both F and $\tilde{F}$, the result follows immediately from Lemma 3 applied to the first $L-1$ layers. □

Remark 1. Let F be a neural network function as in Proposition 1 and C a fixed connected component of a decision region of F, say C a component C_1, and K a compact subset of $\mathbb{R}^{d_{in}}$ that has non-empty intersection with C. Let further $\varepsilon > 0$. Then by means of Proposition 1 we find a network function $\tilde{F}$ with invertible square weight matrices in the first $L-1$ layers such that

$$\|\tilde{F} - F\|_K < \varepsilon$$

and further the decision region of $\widetilde{F}$ corresponding to the first class has a connected component $\widetilde{C}$ such that $\widetilde{C} \cap K \subset C \cap K$.

In fact, first one selects $\varepsilon = \varepsilon(C) > 0$ so small that a network function that approximates F by ε with respect to norm $\|\cdot\|_K$, automatically has a decision region corresponding to the first class that intersects $C \cap K$. Then one approximates F with an accuracy of $\varepsilon/2$ by a network H by means of Proposition 1. In the network function H one decreases the first entry in the bias vector of the final layer L by $\varepsilon/2$ to give the desired approximating network function $\tilde{F}$. Then by the triangle inequality

$$\|\tilde{F}(x) - F(x)\| < \varepsilon.$$

Further, the first components F_1, H_1, $\tilde{F}_1$ of F, H, $\tilde{F}$, respectively, satisfy

$$\tilde{F}_1(x) < \tilde{F}_1(x) + \varepsilon/2 - (H_1(x) - F_1(x)) = F_1(x)$$

for every $x \in K$.

Our main result now states as follows.

Theorem 2. *Let $F : \mathbb{R}^{d_{in}} :\to \mathbb{R}^{d_{out}}$ be a neural network function with continuous and strictly monotonic activation function σ or σ =ReLU and $\omega_F \le d_{in}$. Then for every decision region C_j, $j = 1, \ldots, d_{out}$, the connected components of C_j are unbounded.*

Proof. For a general d_{in}-dimensional box $K = [a, b]^{d_{in}}$, $a < b$, we show that each connected component of the decision regions intersect ∂K. We consider C_1, the decision region for the first class (c.f. Definition 1), for which we fix a connected component that intersects K. We denote the intersection of this component with K by C. That is, C is non-empty and open in K and $C \cap K^\circ \neq \emptyset$. We assume that each weight matrix W_j, $j = 1, \ldots, L-1$ is in $\mathbb{R}^{d_{in}\times d_{in}}$ and invertible. By Proposition 1 and Remark 1, this covers the remaining cases. By the definition of C_1 and by continuity we have that $\widetilde{C} := A_{L-1} \circ \ldots \circ A_1(C)$ is a connected subset of $\Omega = \{y \in \mathbb{R}^{d_{in}} : b_{1,L} - b_{k,L} > (w_{k,L} - w_{1,L})y,\ k = 2, \ldots, d_{out}\}$ where $w_{j,L}$ denotes the jth row of W_L and $b_{j,L}$ is the jth component of b_L.

Lemma 1 yields that Ω is unbounded, since otherwise the hypersurfaces defined by $b_{1,L} - b_{k,L} = (w_{k,L} - w_{1,L})y$ for $k = 2, \ldots, d_{out}$ would enclose the points in Ω.

First, consider the case when σ is continuous and strictly monotonic. In this case, our assumptions give that $A_{L-1} \circ \ldots \circ A_1$

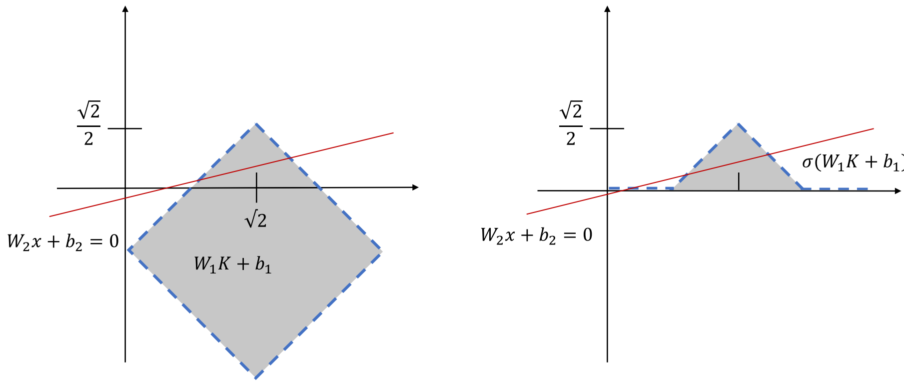


Fig. 1. Example 1: drawing of W_1K+b_1 (left) and $\sigma(W_1K+b_1)$ right (x_1 horizontal, x_2 vertical). The red line depicts the decision hypersurface defined by $W_2\,x+b_2 = 0$.

is injective. As it is the image of a non-empty bounded set under a continuous mapping defined on the whole domain $\mathbb{R}^{d_{in}}$, $\widetilde{C}$ is a non-empty bounded set. Since Ω is unbounded, as it is pointed out above, there exists an $y_0 \in \partial\widetilde{C} \cap \Omega$. The compactness of K and the fact that $A_{L-1} \circ \ldots \circ A_1(C) \subset A_{L-1} \circ \ldots \circ A_1(K)$ ensures the existence of an $x_0 \in K$ with $F(x_0) = y_0$. Further, the Invariance Domain Theorem (also known as Brouwer Invariance Domain Theorem), which applies to $A_{L-1} \circ \ldots \circ A_1$ in this case, implies that inner points of C are mapped to inner points and hence x_0 must be an element of ∂C.

Then, if $x_0 \in \partial K$ the proof for this case is finished. Otherwise x_0 is an interior point in K. Since y_0 is an interior point of Ω, the continuity of $A_{L-1} \circ \ldots \circ A_1$ implies the existence of a small $\varepsilon > 0$ such that $B = \{x : \|x - x_0\|_2 < \varepsilon\}$ is a subset of K and such that $A_{L-1} \circ \ldots \circ A_1(B) \subset \Omega$. Hence, $C \cup B \subset C_1$, since by definition $C_1 = (A_{L-1} \circ \ldots \circ A_1)^{-1}(\Omega)$. But B is not a subset of C which contradicts the fact that C is a connected component of C_1.

Now the case σ = ReLu is considered. If for all $x \in C$, we have $W_jA_{j-1} \circ \ldots \circ A_1(x) + b_j > 0$ for $j = 2, \ldots, L-1$, then $A_{L-1} \circ \ldots \circ A_1$ constitutes a linear affine and invertible map from C to $\widetilde{C}$. Thus, following the same line of arguments as in the previous case, we obtain that $\partial K \cap C \neq \emptyset$. Otherwise, there exist an $x_0 \in C$ and a smallest $l \in \{1, \ldots, L-1\}$ with corresponding $k \in \{1, \ldots, d_{in}\}$ such that $w_{k,l}y_0 + b_{k,l} \leq 0$ where $y_0 := A_{l-1} \circ \ldots \circ A_1(x_0)$ if $l > 1$ and $y_0 = x_0$ otherwise, and where $w_{k,l}$ denotes the l–th row of W_l and $b_{k,l}$ is the l–th component of b_l. Without loss of generality say $k = 1$. Then by the definition of ReLU, the classification does not change on $y_t = y_0 - te_1$, $t > 0$ where $e_1 = (1, 0, \ldots, 0)^T$. More precisely, $A_{L-1} \circ \ldots \circ A_l(\sigma(y_0)) = A_{L-1} \circ \ldots \circ A_l(\sigma(y_t)) \in \Omega$ for all $t > 0$. Since we have chosen l to be minimal, the mapping $A_{l-1} \circ \ldots \circ A_1$, for $l > 1$ or identity otherwise, is linear affine and invertible on C. Its preimage of the half line $\{y_0 - te_1 : t \geq 0\}$ thus intersects ∂C in some point w. Since by the preceding consideration $A_{L-1} \circ \ldots \circ A_1(w) \in \Omega$, we can conclude that $w \in \partial K$ by the same arguments as in the cases above. □

With Theorem 2 we can now extend the lower bound of (1) from Hanin and Sellke (2017) to a wider class of activation functions.

Corollary 1. *For network functions with continuous and strictly monotonic activation function σ or σ =ReLU the lower estimate $d_{in} < \omega_{\min}(\sigma, d_{in}, d_{out})$ holds.*

Proof. For the purpose to show the result by contradiction, let $f : K := [0, 1]^{d_{in}} \to \mathbb{R}$ be continuous with $f(x_h) = -1$ where $x_h = (1/2, \ldots, 1/2)^T$ and $f = 1$ on the boundary of K and assume that

$$\|f - F\|_K < 1/2. \tag{4}$$

Then necessarily $F(x_h) < 0$ and $F(x_b) > 0$ for all x_b in ∂K. By Theorem 2 the preimage of $(-\infty, 0)$ under F (the decision region $F < 0$) is either empty or intersects the boundary of K and hence gives a contradiction, since $F > 1/2$ holds on ∂K. □

Remark 2. It is easily seen that in general $\omega_{\min}(\sigma, d_{in}, d_{out}) < d_{in}$ is impossible for all σ. Indeed, in this case W_1 would have non trivial kernel and therefore every function that is non-constant on all subspaces of $\mathbb{R}^{d_{in}}$, such as $x \mapsto \prod_{j=1}^{d_{in}} x_j^2$, cannot be approximated with arbitrary accuracy.

We now formulate an example that shows that, despite the restrictions given by Theorems 2 and 1, decision regions can be disconnected as subset of a compact input domain (c.f. Definition 2).

Example 1. Let $K = [-1, 1] \times [-1, 1]$ and σ be the ReLU activation function. The weights and bias in the first layer are set as follows: W_1 is the rotation matrix with angle $\alpha = -\pi/4$, i.e.:

$$W_1 = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}, \tag{5}$$

and $b_1 = \sqrt{2}(1, -1/2)^T$. The parameters for the (output) second layer are as follows: $W_2 = 1/\sqrt{2}(1, -4)^T$, $b_2 = -1/4$. The complete model now is written as

$$F : K \to \mathbb{R},\ x \mapsto W_2\,\sigma(W_1x + b_1) + b_2. \tag{6}$$

We further set

$$A_1 : K \to \mathbb{R}^2,\ x \mapsto \sigma(W_1x + b_1). \tag{7}$$

One easily verifies that $A_1(K)$ and the decision hypersurface defined by $W_2x + b_2 = 0$ are as depicted in Fig. 1. Then the region in K that is mapped to $(-\infty, 0)$ is not connected in K as depicted in Fig. 2.

It is straightforward to adapt the above example to the leaky ReLU activation $\sigma_\beta(t) = \max\{t, \beta\,t\}$, $0 < \beta < 1$. The important point is that a convex excerpt of the activation function can be used to create an internal corner in the image of K under the mapping that corresponds to the first layer, c.f. Fig. 1 on the right.

We further extend Example 1 to the case that the mapping starts at $\mathbb{R}^2$ rather than K. The example shows that the surjectivity condition Theorem 1 cannot be dropped in general and the result does not hold for ReLU. It should be mentioned that the authors of Nguyen et al. (2018) are aware of this limitation as they also formulate a counter example for ReLU networks.

Example 2. Let σ be the ReLU activation function and $W_1 = I_2$ the identity in $\mathbb{R}^2$ and $b_1 = (0, 0)^T$. Then $A_1 : \mathbb{R}^2 \to \mathbb{R}^2$, $x \mapsto \sigma(W_1x + b_1)$ maps $\mathbb{R}^2$ to $Q_1 := \{(x_1, x_2)^T \in \mathbb{R}^2 : x_1, x_2 \geq 0\}$.

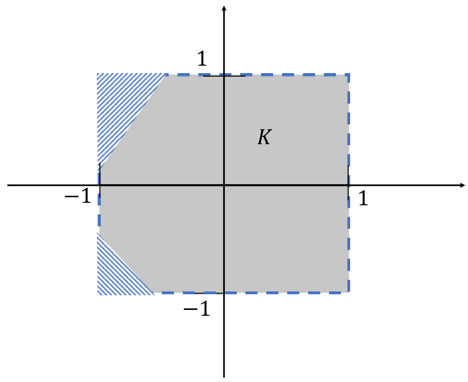


Fig. 2. Drawing of the set K in Example 1 with crosshatched blue area for the preimage of $(-\infty, 0)$ under F (x_1 horizontal, x_2 vertical).

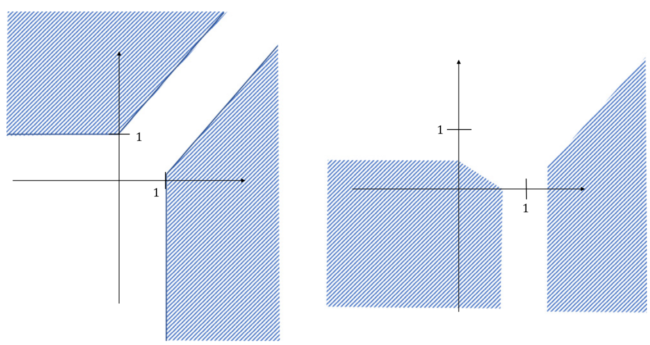


Fig. 3. Example 2 (x_1 horizontal, x_2 vertical): Left, crosshatched blue area depicts the preimage of $[0, 1/\sqrt{2}] \times \{0\} \cup [3/\sqrt{2}, \infty) \times \{0\}$ under G_2. Right: crosshatched blue area depicts the preimage of $(-\infty, 0)$ under F.

Let W_2 be the rotation matrix for angle $\alpha = -3/4\pi$ (c.f. (5)), and $b_2 = \sqrt{2}(1, 1/2)^T$. The resulting image of $\mathbb{R}^2$ under $G_2 : \mathbb{R}^2 \to \mathbb{R}^2$, $x \mapsto \sigma(W_2 A_1(x) + b_2)$ is similar to Fig. 1 right, except that the right bar is continued to $+\infty$. One verifies that the preimage of these bars is as depicted in Fig. 3 left. Following Example 1, we set $W_3 = 1/\sqrt{2}(1, -4)^T$, $b_3 = 1/4$ and define $F : \mathbb{R}^2 \to \mathbb{R}$ as $x \mapsto W_3 G_2(x) + b_3$. Now the preimage of $(-\infty, 0)$ under F is not connected in $\mathbb{R}^2$ as sketched in Fig. 3 right.

The conclusion from the previous examples can be summarized as follows.

Corollary 2. *From Examples 1 and 2 it immediately follows that in general*

1. *for a network function with ReLU or leaky ReLU activation function and width not exceeding the input dimension, the decision areas are not necessarily connected with respect to a bounded input domain (c.f. Definition 2).*
2. *In Theorem 1 the condition that the activation function maps $\mathbb{R}$ surjectively to $\mathbb{R}$ can in general not be dropped. In particular, Theorem 1 does not hold for ReLU activation.*

4. Experiments

In this section, we present the results of some numerical experiments. In the first instance, those experiments where carried out for illustrations purposes. By means of simple artificial data, we show that the limitations that we derived in our main result can be observed in low dimensional setting. Secondly, we briefly investigated if our theoretically derived limitation affect the capabilities of neural networks to approximate common (simple) training data sets.

We defined a dataset consisting of two n-spheres both being centered at the origin. The inner n-sphere (first class) has a radius of $\frac{n}{2}$ and the outer n-sphere (second class) a radius of n. We increase the radius of the n-sphere with the dimension in order to avoid numerical problems when we go to higher dimensions. The two and three dimensional dataset is shown in Fig. 4. The neural network we wanted to train should separate the inner sphere from the outer sphere and we considered a trained network to be successful only when we reached a test accuracy of 100%. According to Theorem 2, this should only be possible if $\omega_F > d_{in}$. If $\omega_F \leq d_{in}$, we should always get an unbounded decision region and consequently not obtain a test accuracy of 100%. We generated 10^7 uniformly distributed points on each sphere for the training data and $25 \cdot 10^5$ uniformly distributed points on each sphere for the test data. Each network was trained for 10 epochs using a batch size of 10^4 and the Adam optimizer with a learning rate of 0.001. For the cost function we used the cross entropy. We trained neural networks of different sizes with fully connected layers and ReLU activation functions only. We combined the following different possible settings:

1. Input dimension: 2, 3, 8 and 20,
2. Number of fully connected layers: 1, 2, 10,
3. All layers of width d_{in} and layers of width $d_{in} + 1$.

Each one of the 24 possible combinations of settings was repeated for 100 trainings of 10 epochs in order to have a meaningful representation of the setting. For each training, we stored the maximal obtained accuracy on the test dataset as the performance of the training. For input data of dimension two and three, Figs. 5, and 6, respectively, illustrate the successful training of a network with $\omega_F > d_{in}$ and an example of an unsuccessful training of a network with $\omega_F \leq d_{in}$. As expected, the networks with $\omega_F \leq d_{in}$ learn an unbounded decision region. Moreover, the results show that unbounded decision regions cannot be avoided throughout the training, even at the network weight initialization. This highlights that the results obtained in Theorem 2 are algorithm-independent. The summary of the complete parameter study is illustrated in Fig. 7. Successful training was only possible when $\omega_F > d_{in}$, although some of the settings which could have been successful did not achieve 100% test accuracy.

In our second experiment we wanted to test the applicability of our results on widely used benchmark datasets. The goal of this experiment was to achieve 100% accuracy on the training datasets for a varying width ω_F of the neural network. When the expressiveness of the neural network is high enough, then we should be able to classify all samples on the training distribution correctly. On the other side, once the width ω_F of the neural network is too small, we should no longer be able to classify all samples correctly. We did not consider the test dataset in this setting, because our work is about approximation capabilities and hence does not consider generalization. This is not a problem for the first experiment, as in the former the manifold of the data is well understood and train and test samples behave similarly. Our investigations were performed as follows: We trained many different neural networks until we reached 100% accuracy on the training dataset or we stopped the training after 500 epochs. We tested different optimizers (Adam, SGD with momentum and Nesterov, and RMSProp), different learning rates, different batch sizes and we alternatively included a scheduler which decreased the learning rate once a plateau was reached. We did not use weight decay or any other regularization techniques. We used neural networks consisting of fully connected layers only with 2 up to 5 hidden layers. The width of all layers was chosen to be

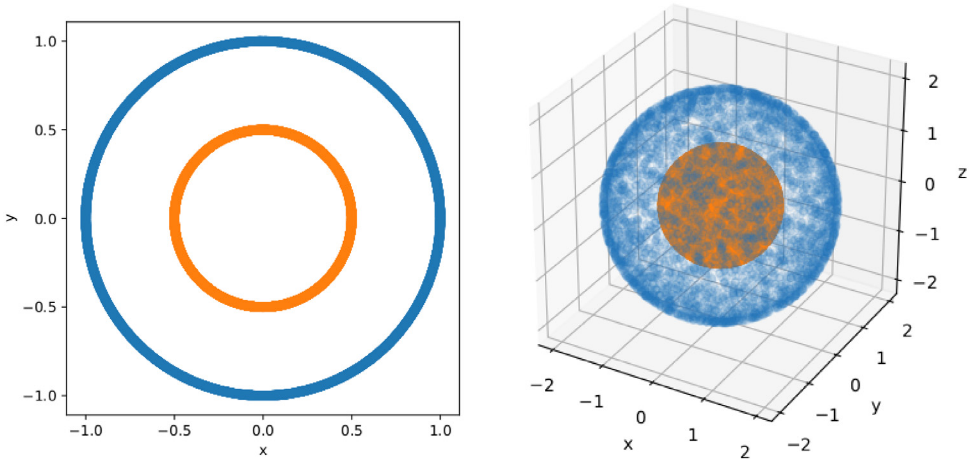
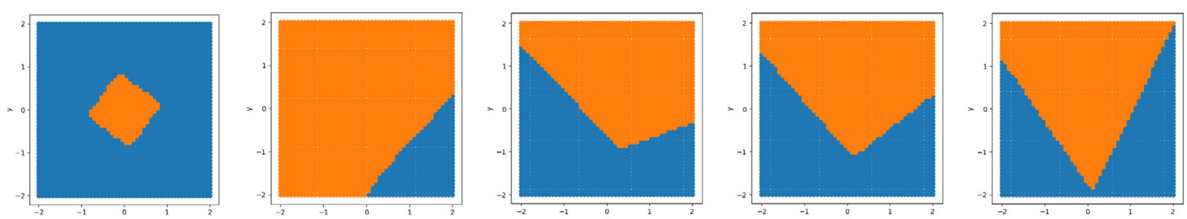
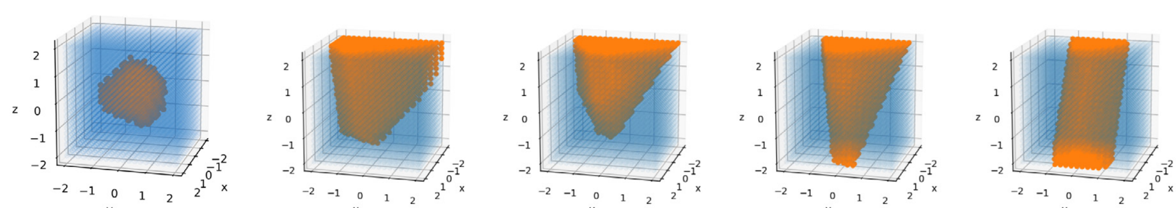


Fig. 4. Input datasets in the two and three dimensional case. The dataset consists of two classes: the inner sphere (orange) labeled as 0 and the outer sphere (blue) labeled as 1. Left: 1-sphere with radius 0.5 for the inner sphere and radius 1.0 for the outer sphere. Right: 2-sphere with radius 1.0 for the inner sphere and radius 2.0 for the outer sphere. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Success - Epoch 10 (b) Failure - Epoch 0 (c) Failure - Epoch 1 (d) Failure - Epoch 2 (e) Failure - Epoch 10

Fig. 5. Decision regions by trained networks for input dimension two. (a): Decision regions for a network with one hidden layer of size $d_{in}+1$ obtaining 100% test accuracy. (b)–(e): Decision regions for a network with one hidden layer of size d_{in} over several epochs. The results show that the decision regions need to be unbounded throughout training even before the first epoch, i.e. at initialization.



(a) Success - Epoch 10 (b) Failure - Epoch 0 (c) Failure - Epoch 1 (d) Failure - Epoch 2 (e) Failure - Epoch 10

Fig. 6. Decision regions by trained networks for input dimension three. (a): Decision regions for a network with one hidden layer of size $d_{in}+1$ obtaining 100% test accuracy. (b)–(e): Decision region for a network with one hidden layer of size d_{in} over several epochs. The results show that the decision regions need to be unbounded throughout the training even before the first epoch, i.e. at initialization.

the same. However, we did not consider neural networks with a width smaller than the number of classes of the dataset used. We tested neural networks mostly using ReLU as an activation function, but did also experiments with tanh, however, both behaved similarly. We tested our approach on MNIST and Fashion MNIST (10 classes) as well as on the EMNIST letters (26 classes) using the cross-entropy as the cost function. From all our experiments, we report in Fig. 8 (MNIST), Fig. 9 (Fashion MNIST) and Fig. 10 (EMNIST letters) in blue (success) all the widths for which we achieved 100% accuracy on the training distribution and mark by red (failure) all the widths for which this was not achieved in our experiments. Achieving 100% accuracy consistently was more challenging then we anticipated and for some widths we did not achieve 100% accuracy although it was achieved for smaller widths. We report only the transition phase between successful and failed widths as the 100% training accuracy was already achievable for widths much smaller than the input dimension of the dataset. However, our results show that the trend in general is the same for all datasets: on the one side the expressiveness decreases drastically when a certain threshold is reached and on the other side, in contrast to our theoretically derived limitations, the input dimension limitation does not provide any restriction for widely used benchmark datasets.

It is a common belief that the true dimension of a dataset does not correspond to the dimension of the input space (e.g. pixel space for images), but that the dataset can be represented by a lower dimensional (non-linear) manifold. The dimension of the latter is commonly called the intrinsic dimension, which is for most datasets unknown, but its determination for example is investigated in Levina and Bickel (2005). Considering our results on the benchmark datasets, we would like to raise the question, and leave this as an open question for future work, whether the theoretical results as derived in the beginning of our work could show limitations that would become relevant for real data in their

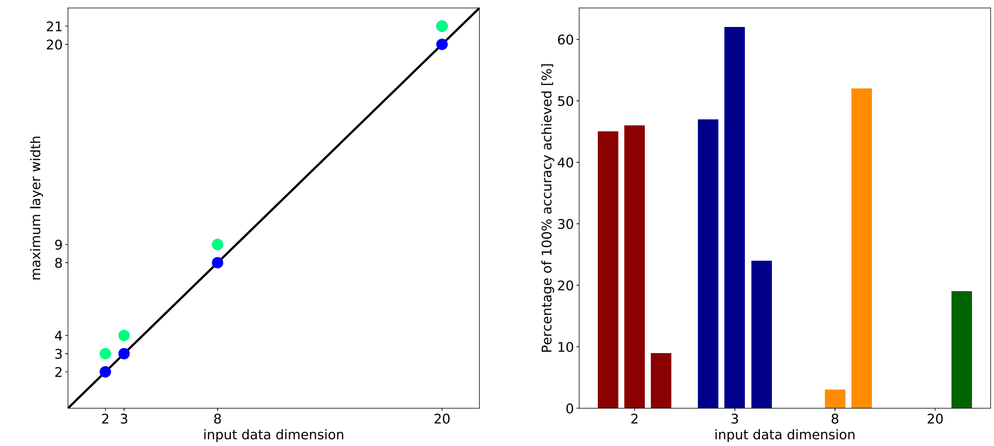


Fig. 7. Results of the complete test run (24 combination of settings trained 100 times for 10 epochs). Left: Input data dimension d_{in} plotted against maximal network width ω_F. Green points are instances for which the network obtained at least once 100% accuracy on the test set. Blue points are instances for which the network never obtained 100% test accuracy. We clearly see the trend that the network is only capable of learning correct decision regions when $\omega_F > d_{in}$. Right: For each input data dimension (2, 3, 8 and 20), we tested three different network sizes (colored in the same color from left to right: 1,2 and 10). The height of the bar plot represents the percentage of networks which obtained 100% accuracy on the test set out of 100 repeated trainings. For $d_{in} = 8$ and $d_{in} = 20$ some of the network settings did never achieve 100% although success should be possible. That is why some bars vanish completely. We only plot networks for which $\omega_F > d_{in}$, since otherwise the percentage is zero. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

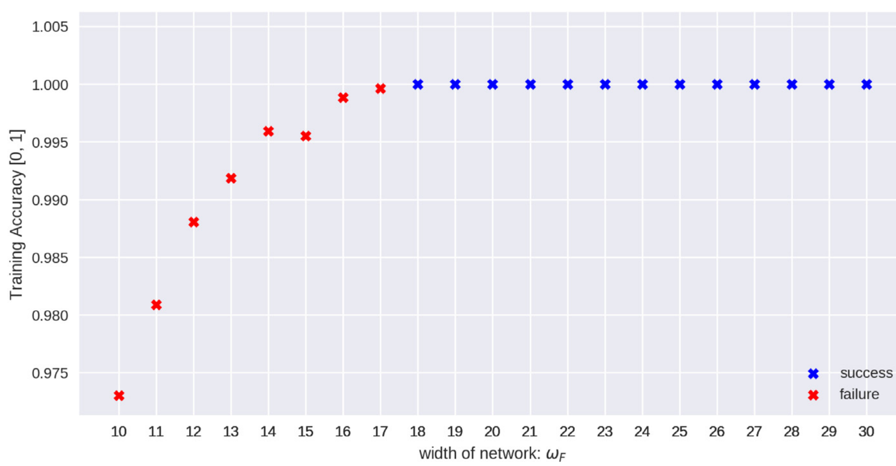


Fig. 8. Results of our investigations on achieving 100% accuracy on the training set for MNIST as a function of the width ω_F of the neural networks: In blue we mark the widths for which our experiments yielded 100% accuracy whereas in red we mark the widths for which we could not train a model to classify all training samples correctly. The dimension of the input data is $28 * 28 = 784$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

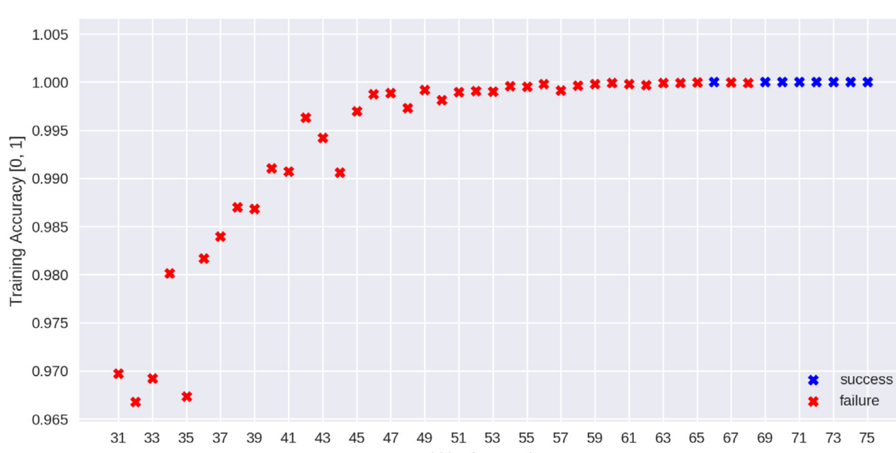


Fig. 9. Results of our investigations on achieving 100% accuracy on the training set for Fashion MNIST as a function of the width ω_F of the neural networks: In blue we mark the widths for which our experiments yielded 100% accuracy whereas in red we mark the widths for which we could not train a model to classify all training samples correctly. The dimension of the input data is $28 * 28 = 784$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

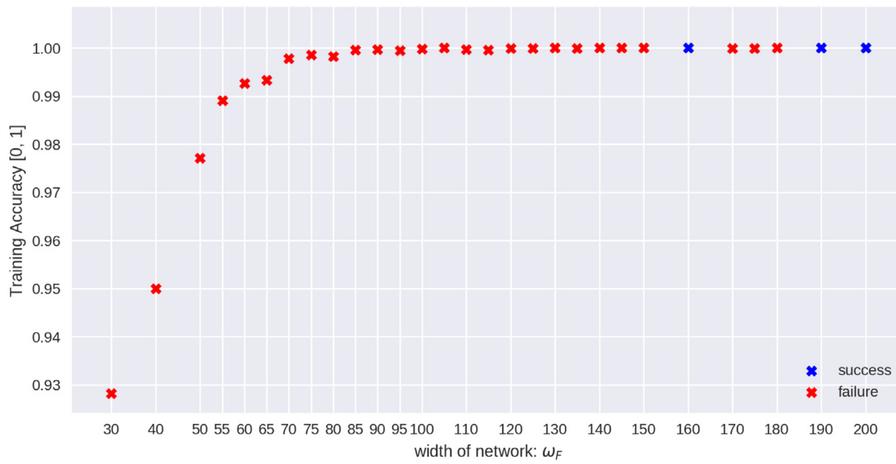


Fig. 10. Results of our investigations on achieving 100% accuracy on the training set for EMNIST as a function of the width ω_F of the neural networks: In blue we mark the widths for which our experiments yielded 100% accuracy whereas in red we mark the widths for which we could not train a model to classify all training samples correctly. The dimension of the input data is $28 * 28 = 784$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

lower dimensional representation corresponding to their intrinsic dimension.

5. Conclusion

For a wide class of activation functions, we have shown that for neural network functions that have a maximum width less than or equal to the input dimension the connected components of the decision regions are unbounded. Hence, for such networks the decision regions intersect the boundary of a natural input domain. This links some recent results from Hanin and Sellke (2017) and Nguyen et al. (2018), where for such narrow neural networks limitations regarding their expressive power for the case of ReLU activation are achieved in the first work, and the connectivity of decision regions is restricted for the case of continuous, monotonically increasing and surjective $\sigma : \mathbb{R} \to \mathbb{R}$ in the second work. We illustrated our findings by numerical experiments with spherical data where it was observed that the input dimension is the critical threshold for network width in order to achieve 100% accuracy. However, in experiments on MNIST, Fashion MNIST and EMNIST letters we could not detect a limitation on the performance for such narrow networks. This raises the question to what extent limitations in terms of connectivity imply a crucial restriction in practical applications. From theoretical perspective, it would be interesting to have a more complete describtion of possible shapes and topological properties of decision regions and to check whether Theorem 2 can be generalised to the case where the input data are located on a lower dimensional manifold.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially supported by the MECO, Luxembourg project "Artificial Intelligence for Safety Critical Complex Systems". The second author is supported by the Luxembourg National Research Fund (FNR) under the grant number 13043281.

References

Bölcskei, Helmut, Grohs, Philipp, Kutyniok, Gitta, & Petersen, Philipp (2019). Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, *1*(1), 8–45.
Cohen, Nadav, Sharir, Or, & Shashua, Amnon (2016). On the expressive power of deep learning: A tensor analysis. In *Conference on Learning Theory* (pp. 698–728).
Costarelli, Danilo, & Spigler, Renato (2013). Approximation results for neural network operators activated by sigmoidal functions. *Neural Networks*, *44*, 101–106.
Cybenko, George (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.
Elfwing, Stefan, Uchibe, Eiji, & Doya, Kenji (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, *107*, 3–11.
Fawzi, Alhussein, Moosavi-Dezfooli, Seyed-Mohsen, Frossard, Pascal, & Soatto, Stefano (2017). Classification regions of deep neural networks. arXiv preprint arXiv:1705.09552.
Gnecco, Giorgio, Kurková, Věra, & Sanguineti, Marcello (2011). Some comparisons of complexity in dictionary-based and linear computational models. *Neural Networks*, *24*(2), 171–182.
Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron, & Bengio, Yoshua (2016). *Deep learning, Vol. 1*. MIT press Cambridge.
Gühring, Ingo, Kutyniok, Gitta, & Petersen, Philipp (2020). Error bounds for approximations with deep ReLU neural networks in W s, p norms. *Analysis and Applications*, *18*(05), 803–859.
Guliyev, Namig J., & Ismailov, Vugar E. (2018). On the approximation by single hidden layer feedforward neural networks with fixed weights. *Neural Networks*, *98*, 296–304.
Hanin, Boris (2019). Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, *7*(10), 992.
Hanin, Boris, & Sellke, Mark (2017). Approximating continuous functions by ReLU nets of minimal width. arXiv preprint arXiv:1710.11278.
He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.
Hornik, Kurt (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.
Johnson, Jesse (2018). Deep, skinny neural networks are not universal approximators. In *International Conference on Learning Representations (ICLR)*.
Kidger, Patrick, & Lyons, Terry (2020). Universal approximation with deep narrow networks. In *Conference on Learning Theory* (pp. 2306–2327). PMLR.
Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).
Levina, Elizaveta, & Bickel, Peter J. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems* (pp. 777–784).

Lin, Henry W., Tegmark, Max, & Rolnick, David (2017). Why does deep and cheap learning work so well?. *Journal of Statistical Physics*, *168*(6), 1223–1247.
Lu, Zhou, Pu, Hongming, Wang, Feicheng, Hu, Zhiqiang, & Wang, Liwei (2017). The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems* (pp. 6232–6240).
Mhaskar, Hrushikesh N., & Poggio, Tomaso (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, *14*(06), 829–848.
Montufar, Guido F, Pascanu, Razvan, Cho, Kyunghyun, & Bengio, Yoshua (2014). On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 2924–2932).
Nguyen, Quynh, & Hein, Matthias (2017). The loss surface and expressivity of deep convolutional neural networks. arXiv preprint arXiv:1710.10928.
Nguyen, Quynh, Mukkamala, Mahesh Chandra, & Hein, Matthias (2018). Neural networks should be wide enough to learn disconnected decision regions. In *Proceedings of Machine Learning Research*: *vol. 80*, *Proceedings of the 35th International Conference on Machine Learning* (pp. 3740–3749). PMLR.
Park, Sejun, Yun, Chulhee, Lee, Jaeho, & Shin, Jinwoo (2021). Minimum width for universal approximation. In *International Conference on Learning Representations (ICLR)*.
Petersen, Philipp, & Voigtlaender, Felix (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, *108*, 296–330.
Petersen, Philipp, & Voigtlaender, Felix (2020). Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the Americal Mathematical Society*, *148*(4), 1567–1581.
Raghu, Maithra, Poole, Ben, Kleinberg, Jon, Ganguli, Surya, & Sohl-Dickstein, Jascha (2017). On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICLR)* (pp. 2847–2854). PMLR.
Rolnick, David, & Tegmark, Max (2018). The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations (ICLR)*.
Scarselli, Franco, & Tsoi, Ah Chung (1998). Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, *11*(1), 15–37.
Schmidhuber, Jürgen (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.
Telgarsky, Matus (2016). Benefits of depth in neural networks. In *Proceedings of Machine Learning Research*: *vol. 49*, *29th Annual Conference on Learning Theory* (pp. 1517–1539). PMLR.
Yarotsky, Dmitry (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, *94*, 103–114.
Yarotsky, Dmitry (2018). Optimal approximation of continuous functions by very deep relu networks. In *Proceedings of Machine Learning Research*: *vol. 75*, *Proceedings of the 31st Conference on Learning Theory* (pp. 639–649). PMLR.