Bounds on the Approximation Power of Feedforward Neural Networks

Mohammad Mehrabi [1] **Aslan Tchamkerten** [2] **Mansoor I. Yousefi** [2]

Abstract

The approximation power of general feedforward neural networks with piecewise linear activation functions is investigated. First, lower bounds on the size of a network are established in terms of the approximation error and network depth and width. These bounds improve upon state-of-the-art bounds for certain classes of functions, such as strongly convex functions. Second, an upper bound is established on the difference of two neural networks with identical weights but different activation functions.

1. Introduction

It is well-known that sufficiently large multi-layer feedforward networks can approximate any function with desired accuracy (Hornik et al., 1989). An important problem then is to determine the smallest neural network for a given task and accuracy. The standard guideline is the approximation power (variously known as expressiveness) of the network which quantifies the size of the neural network, typically in terms of depth and width, in order to approximate a class of functions within a given error. In particular, several works provided evidence that deeper networks perform better than shallow ones, given a fixed number of hidden units (Bianchini & Scarselli, 2014; Delalleau & Bengio, 2011; Liang & Srikant, 2017; Mhaskar et al., 2016; Pascanu et al., 2014; Telgarsky, 2015; 2016; Yarotsky, 2017).[1]

A popular activation function is the rectified linear unit (ReLU), partly because of its low complexity when coupled with backpropagation training (Krizhevsky et al., 2012). It has, therefore, become of interest to determine the power of neural networks with ReLU's and, more generally, with piecewise linear activation functions.

Determining the capacity of a neural networks with a piecewise linear activation function typically involves two steps. First, evaluate the number of linear pieces (or break points) that the network can produce and, second, tie this number to the approximation error. The works (Montufar et al., 2014; Pascanu et al., 2014) recently showed that a linear increase in depth results in an exponential growth in the number of linear pieces as opposed to width which results only in a polynomial growth. Accordingly, the approximation capacity exhibits a similar tradeoff between depth and width. For related works with respect to classification error see (Telgarsky, 2015; 2016) and with respect to function approximation error see (Liang & Srikant, 2017; Mhaskar et al., 2016; Yarotsky, 2017).

In this paper we consider general feedforward neural networks with piecewise linear activation functions and establish bounds on the size of the network in terms of the approximation error, the depth d, the width, and the dimension of the input space to approximate a given function. We first establish an improved upper bound on the number of break points that such a network can produce which is a multiplicative factor d^d smaller than the currently best known from (Yarotsky, 2017). This upper bound is obtained by investigating neuron state transitions as introduced in (Raghu et al., 2017). Combining this upper bound with lower bounds in terms of error and dimension, we obtain necessary conditions on the depth, width, error, and dimension for a neural network to approximate a given function. These bounds significantly improve on the corresponding state-of-the-art bounds for certain classes of functions (Theorems 1,2 and Corollaries 1,2,3).

The second contribution of the paper (Theorem 3) is an upper bound on the difference of two neural networks with identical weights but different activation functions. This problem is related to "activation function simulation" investigated in (DasGupta & Schnitger, 1993) which leverages network topology to compensate a change in activation function.

The paper is organized as follows. In Section 2 we briefly introduce the setup. In Section 3 we present the main results which are then compared with the corresponding ones in the recent literature in Section 4. Finally, Section 5 contains the proofs.

[1] Department of Electrical Engineering, Sharif University of Technology, Iran [2] Department of Communications and Electronics, Telecom ParisTech, France. Correspondence to: Mohammad Mehrabi <mohamadmehrabi4@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

[1] For a nice counterexample see (Lu et al., 2017).

2. Preliminaries

Throughout the paper $\mathcal{R}$ denotes a compact convex set in $\mathbb{R}^n$, $n \geq 1$, and $\mathbb{F}_\sigma$ denotes the set of feedforward neural networks with input $\mathcal{R}$, output $\mathbb{R}$, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Feedforward here refers to the fact that the neural network contains no cycles; connections are allowed between non-neighbouring layers. It is assumed that σ is a piecewise linear (not necessarily continuous) function with $t \geq 1$ linear pieces. The set of all such activation functions is denoted by Σ_t.

A neural network $f \in \mathbb{F}_\sigma$ consists of a set of input units $\mathcal{I}_f$, a set of hidden units $\mathcal{H}_f$ that operate according to σ, non-zero weights representing connections, and a single output unit which just weight-sums its inputs. To simplify the notation we use f to represent both a neural network and the function that it represents.

For instance, in the neural network shown in Fig. 1, we have $\mathcal{I}_f = \{x_1, x_2, x_3\}$ and $\mathcal{H}_f = \{u_{ij},\ \forall i, j\}$.

Definition 1 (Depth and width)**.** *Given a neural network $f \in \mathbb{F}_\sigma$, the depth of a hidden unit $h \in \mathcal{H}_f$, denoted as $d_f(h)$, is the length of the longest path from any $i \in \mathcal{I}_f$ to h. The depth of f is*

$$d_f \overset{def}{=} \max \Big\{ d_f(h) \big| h \in \mathcal{H}_f \Big\}.$$

The set of hidden units with depth i is

$$\mathcal{H}_f^i \overset{def}{=} \Big\{ h \in \mathcal{H}_f \big| d_f(h) = i \Big\}.$$

The width of the network is

$$\omega_f \overset{def}{=} \frac{|\mathcal{H}_f|}{d_f} \overset{def}{=} \frac{\sum_{i=1}^{d_f} \omega_i}{d_f} \tag{1}$$

where

$$\omega_i \overset{def}{=} |\mathcal{H}_f^i|.$$

For instance, in Fig. 1, the hidden unit u_{23} can be reached by inputs x_1 and x_3, by following the paths $x_1 \to u_{23}$, $x_3 \to u_{11} \to u_{23}$, or $x_3 \to u_{12} \to u_{23}$. Therefore, $d_f(u_{23}) = 2$. The hidden units of maximum depth are u_{31}, u_{32}, and u_{33} and hence $d_f = 3$, $\mathcal{H}_f^3 = \{u_{31}, u_{32}, u_{33}\}$ and $\omega_f = 8/3$.

The following simple inequality is frequently used in the paper.

Lemma 1. *For any $t \geq 1$, $d_f \geq 1$, and $|\mathcal{H}_f| \geq 1$*

$$((t-1)\omega_f + 1)^{d_f} \leq t^{|\mathcal{H}_f|}.$$

Proof. Set $\omega_f = \frac{|\mathcal{H}_f|}{d_f}$ and observe that

$$\left((t-1)\frac{|\mathcal{H}_f|}{d_f} + 1 \right)^{d_f}$$

is a non-decreasing function of d_f and that $d_f \leq |\mathcal{H}_f|$. □

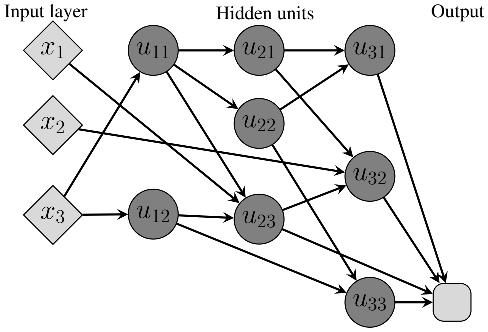

Figure 1. A feedforward network f with $|\mathcal{I}_f| = 3$ inputs, $|\mathcal{H}_f| = 8$ hidden units, depth $d_f = 3$, and width $\omega_f = 8/3$.

Definition 2 (Affine ε-approximation)**.** *Function $f \in \mathbb{F}_\sigma$ is an affine ε-approximation of a function $g : \mathcal{R} \to \mathbb{R}$ if*

$$\sup_{\boldsymbol{x} \in \mathcal{R}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \leq \varepsilon.$$

Definition 3 (Break point)**.** *Given $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{R}^2$, function $f : \mathcal{R} \to \mathbb{R}$ admits a* break point *at $\alpha_0 \in (0,1)$ relative to the segment $[\boldsymbol{x}, \boldsymbol{y}]$ if the first order derivative of $f((1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y})$ does not exist at $\alpha = \alpha_0$. The total number of break points of f on the (open) segment $]\boldsymbol{x}, \boldsymbol{y}[$ is denoted by $B_{\boldsymbol{x}\to\boldsymbol{y}}(f)$. Finally, we let $\bar{B}_{\boldsymbol{x}\to\boldsymbol{y}}(f) \overset{def}{=} B_{\boldsymbol{x}\to\boldsymbol{y}}(f) + 1$.*

Since f is piecewise linear $\bar{B}_{\boldsymbol{x}\to\boldsymbol{y}}(f)$ simply counts the number of linear pieces that f produces as the input ranges from $\boldsymbol{x}$ to $\boldsymbol{y}$.

3. Main Results

Theorems 1,2 and Corollaries 2,3 provide bounds on the size of a neural network to approximate a given function. These bounds are expressed in terms of the approximation error and width and depth of the network, but hold irrespectively of the weights. Recall that connections are allowed between non-neighboring layers.

As a notational convention we use $C^2(\mathcal{R})$ to denote the set of functions $\mathcal{R} \to \mathbb{R}$ whose second order partial derivatives are continuous over $\mathring{\mathcal{R}}$ (the interior of $\mathcal{R}$).

Theorem 1. *Let $f \in \mathbb{F}_\sigma$, $\sigma \in \Sigma_t$, be an ε-approximation of a function $g \in C^2(\mathcal{R})$ and let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{R}$. Then,*

$$\Big((t-1)\omega_f + 1 \Big)^{d_f} \geq \bar{B}_{\boldsymbol{x}\to\boldsymbol{y}}(f) \tag{2}$$

$$\geq \frac{||\boldsymbol{x} - \boldsymbol{y}||_2}{4\sqrt{\varepsilon}} \cdot \Psi(g, \boldsymbol{x}, \boldsymbol{y}), \tag{3}$$

where

$$\Psi(g,\boldsymbol{x},\boldsymbol{y}) \overset{def}{=} \sqrt{\inf_{0\le\alpha\le 1}\Big(\max\big\{0,\gamma(\alpha)\delta(\alpha)\big\}\Big)}, \tag{4}$$
$$\gamma(\alpha) \overset{def}{=} \min\big\{|\alpha_1(\alpha)|,|\alpha_2(\alpha)|\big\},$$
$$\delta(\alpha) \overset{def}{=} \operatorname{sign}\big(\alpha_1(\alpha)\alpha_2(\alpha)\big),$$

and where $\alpha_1(\alpha)$ and $\alpha_2(\alpha)$ are the largest and smallest eigenvalues of the hessian matrix $\nabla^2 g\big((1-\alpha)\boldsymbol{x}+\alpha\boldsymbol{y}\big)$, respectively.

Maximizing the right-hand side of (3) over $\boldsymbol{x},\boldsymbol{y}$ and using Lemma 1 we obtain:

Corollary 1. *Under the assumptions of Theorem 1 we have*

$$|\mathcal{H}_f| \ge \log_t\left(\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{R}^2}\left\{\frac{||\boldsymbol{x}-\boldsymbol{y}||_2}{4\sqrt{\varepsilon}}\cdot\Psi(g,\boldsymbol{x},\boldsymbol{y})\right\}\right).$$

A function $g:\mathcal{R}\to\mathbb{R}$ that is twice differentiable is said to be strongly convex with parameter μ if $\nabla^2 g(\boldsymbol{x})\succeq\mu I$ for all $\boldsymbol{x}\in\mathcal{R}$.

Corollary 2. *Let $f\in\mathbb{F}_\sigma$, $\sigma\in\Sigma_t$, be an ε-approximation of a function $g\in C^2(\mathcal{R})$ that is strongly convex with parameter $\mu>0$. Then,*

$$|\mathcal{H}_f| \ge \frac{1}{2}\log_t\Big(\frac{\mu\cdot(\operatorname{diam}(\mathcal{R}))^2}{16\varepsilon}\Big),$$

where

$$\operatorname{diam}(\mathcal{R}) \overset{def}{=} \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{R}}||\boldsymbol{x}-\boldsymbol{y}||_2.$$

Proof. By strong convexity $\Psi(g,\boldsymbol{x},\boldsymbol{y})\ge\sqrt{\mu}$. The result then follows from Theorem 1 and Lemma 1. □

As an example, consider $g(\boldsymbol{x})=\boldsymbol{x}\cdot\boldsymbol{x}$ over $[0,1]^n$. The Hessian matrix is $2I_{n\times n}$ and from Corollary 2 we get

$$|\mathcal{H}_f| \ge \log_2\Big(\sqrt{\frac{n}{8\varepsilon}}\Big).$$

Corollary 3. *Let $\mathcal{R}=[0,1]^n$. Let $f\in\mathbb{F}_\sigma$, $\sigma\in\Sigma_2$,[2] be an ε-approximation of a function $g\in C^2(\mathcal{R})$ such that $\nabla g(x)\succ 0$ for any $x\in\mathring{\mathcal{R}}$. Then,*

$$|\mathcal{H}_f| \ge q(g)d_f\varepsilon^{-\frac{1}{2d_f}} \tag{5}$$

where $q(g)>0$ is a constant that only depends on g.

Proof of Corollary 3. From Theorem 1 we get

$$\Big(\frac{\mathcal{H}_f}{d_f}+1\Big)^{d_f} \ge \frac{c(g)}{\sqrt{\varepsilon}},$$

where $c(g)>0$ is some strictly positive constant, since the Hessian of g is positive definite everywhere over $\mathring{\mathcal{R}}$. Since $\mathcal{H}_f/d_f\ge 1$ the above inequality implies

$$\Big(2\frac{|\mathcal{H}_f|}{d_f}\Big)^{d_f} \ge \frac{c}{\sqrt{\varepsilon}}.$$

Since $\frac{1}{2}c^{\frac{1}{d_f}}\ge q$ where $q=\frac{1}{2}\min(c,1)$, the above inequality yields the desired result. □

Theorem 2. *Let $\mathcal{R}=[0,1]^n$. Let $f\in\mathbb{F}_\sigma$, $\sigma\in\Sigma_t$, be an ε-approximation of a function $g:\mathcal{R}\to\mathbb{R}$ such that $|D^J(g)(\boldsymbol{x})|\le\delta$ for any $\boldsymbol{x}\in[0,1]^n$ and any multi-index[3] J such that $|J|=3$. Then,*

$$\big((t-1)\omega_f+1\big)^{d_f} \ge \sqrt{\frac{\Big(\max\limits_{\boldsymbol{x}\in[0,1]^n}\big|\Delta(g)(\boldsymbol{x})\big|n^{-1}-\delta n^{\frac{3}{2}}\Big)^+}{16\varepsilon}}, \tag{6}$$

where

$$\Delta(g)(\boldsymbol{x}) = \sum_{k=1}^n\frac{d^2g}{dx_k^2}, \tag{7}$$

is the Laplacian of g and where $a^+=\max(a,0)$.

For instance, approximating

$$g(x_1,x_2)=10x_1^2+x_1^2x_2^2+10x_2^2$$

over $[0,1]^2$ requires $\log_t\Big(\frac{0.82}{\sqrt{\varepsilon}}\Big)$ hidden units—combine Theorem 2 with Lemma 1.

Whether it is Theorem 1 or Theorem 2 which provides a better approximation bound depends on g. For instance, for $g_1(x_1,x_2)=20x_1^2-2x_2^2+x_1^2x_2^2$ Theorem 1 gives a trivial (zero) lower bound since the two eigenvalues of the Hessian matrix $\nabla^2(g_1)$ have always different signs. Theorem 2 instead gives $\frac{0.737}{\sqrt{\varepsilon}}$. On the other hand, for $g_2(x_1,x_2)=10x_1^2+10x_2^2+x_1^2x_2^2$ Theorem 1 gives $\frac{1.37}{\sqrt{\varepsilon}}$ as lower bound while Theorem 2 gives $\frac{0.82}{\sqrt{\varepsilon}}$.

The next theorem quantifies the effect of a change of activation function on the output of the neural network. Here, the activation functions need not be piece-wise affine.

Theorem 3. *Let $f_1\in\mathbb{F}_{\sigma_1}$ and $f_2\in\mathbb{F}_{\sigma_2}$ be two neural networks with identical architectures and weights. Suppose that σ_1 is a δ-Lipschitz continuous function and suppose that the weights belong to some bounded interval $[-A,+A]$, $A>0$. Then,*

$$||f_1-f_2||_\infty \le \frac{||\sigma_1-\sigma_2||_\infty}{\delta}\left(\Big(\delta\cdot A\cdot\omega_f+1\Big)^{d_f}-1\right). \tag{8}$$

[2] Recall that Σ_2 includes ReLU's.

[3] *E.g.*, for $J=(2,1)$ we have $D^J(g(x_1,x_2))=\frac{\partial^3 g}{\partial^2 x_1\partial x_2}$.

A slightly weaker version of (8) is

$$||f_1 - f_2||_\infty \leq \frac{||\sigma_1 - \sigma_2||_\infty}{L} \left(\left(L^2 \cdot \omega_f + 1 \right)^{d_f} - 1 \right),$$

where $L = \max\{A, \delta\}$ denotes the *Lipschitz-bound* defined in (DasGupta & Schnitger, 1993).

As an illustration of Theorem 3 consider a feedforward neural network f_1 with 100 hidden units, a maximum depth of 5, and the *sigmoid* as activation function. Suppose the weights belong to interval $[-1, 1]$. Replacing the sigmoid with a 32-bit quantized function results in an error of at most 0.0001—which can readily be obtained from Theorem 3 with $\delta = \frac{1}{4}, A = 1, ||\sigma_1 - \sigma_2||_\infty = 2^{-32}$.

4. Comparison with Previous Works

Consider first the inequality (2). Restricting attention to neural networks with d hidden layers, at most ω units per layer, and where connections are allowed only between neighbouring layers, this inequality gives

$$\bar{B}_{\boldsymbol{x} \to \boldsymbol{y}}(f) \leq \Big((t-1)\omega + 1 \Big)^d. \tag{9}$$

This is to be compared with the previously best known bound (Lemma 3.2 in (Telgarsky, 2016))

$$2(2(t-1)\omega)^d$$

which is larger by a multiplicative factor that is exponential in d whenever $\omega > 1$, $t \geq 2$. For $n = 1$, Lemma 2.1 in (Telgarsky, 2015) gives $(t\omega)^d$ which still differs from (9) by a multiplicative factor that is exponential is d for $\omega > 1$, $t \geq 2$.

For general feedforward neural networks the previously best known bound (see Lemma 4 of (Yarotsky, 2017)) was

$$\bar{B}_{\boldsymbol{x} \to \boldsymbol{y}}(f) \leq \Big(t \cdot \omega \cdot d_f \Big)^{d_f}$$

which is a multiplicative factor ${d_f}^{d_f}$ larger than (2).

Now consider the approximation power of neural networks in terms of number of hidden units required to approximate a given function within a given error. Theorem 11 in (Liang & Srikant, 2017) states that to approximate a function $[0,1]^n \to \mathbb{R}$, assumed to be differentiable and strongly convex with parameter μ, with a neural network f requires

$$|\mathcal{H}_f| \geq \frac{1}{2} \log_2 \big(\frac{\mu}{16\varepsilon} \big),$$

regardless of the dimension n. Corollary 2 improves this bound to

$$\frac{1}{2} \log_2 \big(\frac{\mu \cdot n}{16\varepsilon} \big)$$

which incorporates dimension as well—albeit the dependency on dimension is arguably small.

Table 1. Bounds comparisons

	Previous	This paper
Regular:	(Telgarsky, 2016)	(Theorem 1)
$\bar{B}_{\boldsymbol{x} \to \boldsymbol{y}}(f) \leq$	$2(2(t-1)\omega)^d$	$\Big((t-1)\omega + 1\Big)^d$
General:	(Yarotsky, 2017)	(Theorem 1)
$\bar{B}_{\boldsymbol{x} \to \boldsymbol{y}}(f) \leq$	$\Big(t \cdot \omega \cdot d_f\Big)^{d_f}$	$\Big((t-1)\omega_f + 1\Big)^{d_f}$
$g \in C^2([0,1]^n)$		
over μ-convex	(Liang & Srikant, 2017)	(Corollary 2)
$\lvert\mathcal{H}_f\rvert \geq$	$\frac{1}{2}\log_2\big(\frac{\mu}{16\varepsilon}\big)$	$\frac{1}{2}\log_2\big(\frac{\mu \cdot n}{16\varepsilon}\big)$
$g \in C^2([0,1]^n)$		
Hess$(g) \succ 0$, Σ_2	(Yarotsky, 2017)	(Corollary 3)
$\lvert\mathcal{H}_f\rvert \geq$	$q_1 \varepsilon^{\frac{-1}{2d_f}}$	$d_f q_2 \varepsilon^{\frac{-1}{2d_f}}$

Corollary 3 provides a lower bound for ReLU types of networks in terms of the error, the depth, and a constant term which only depends on g. This bound can be compared with the bound of Theorem 6 in (Yarotsky, 2017) which is of order $\epsilon^{-\frac{1}{2d_f}}$.[4] Hence, Corollary 3 provides a linear (in d_f) improvement which is particularly relevant in the deep regime where $d_f = \Omega(\log(1/\varepsilon))$. Table 1 summarizes the above discussion.

To the best of our knowledge Theorem 3 is the first result to bound the effect of a change in the activation function for given network topology and weights. Noteworthy perhaps, this bound is essentially universal in the weights since it only depends on their range.

Finally, compared to the cited papers it should perhaps be stressed that the proofs here (see next section) are relatively elementary—*e.g.*, they do not hinge on VC dimension analysis—and hold true for general feedforward networks.

5. Analysis

We first establish a few lemmas to prove Proposition 1 which will provide an upper bound on the number of break points. Then we establish Propositions 2 and 3 which will give lower bounds on the number of break points in terms of the approximation error. Combining these propositions will give Theorems 1 and 2. Finally, we prove Theorem 3.

Definition 4 (Intermediate set of units)**.** *Given $f \in \mathbb{F}_\sigma$ and*

[4] Theorem 6 of (Yarotsky, 2017) provides a bound of the form $q\epsilon^{-\frac{1}{2d_f}}$ where q is a constant that depends on both g and d_f. However, a close inspection of the proof of this theorem reveals that q depends only on g.

$\mathcal{U} \subseteq \mathcal{H}_f$ we define the set of hidden units that lie on a path between the input and $\mathcal{U}$ as

$$\mathrm{in}(\mathcal{U}) \overset{def}{=} \Big\{ v \in \mathcal{H}_f \backslash \mathcal{U} \,|\, \exists i \in \mathcal{I}_f, u \in \mathcal{U} \text{ s.t. } v \in (i \rightarrow u) \Big\}$$

where $(i \rightarrow u)$ denotes the set of intermediate hidden nodes on the path from i to u.

For instance, in Fig. 1 we have

$$\mathrm{in}(\{u_{32}\}) = \{u_{11}, u_{12}, u_{21}, u_{23}\}.$$

The following lemma follows from the above definition.

Lemma 2. *Given $\mathcal{U} \subseteq \mathcal{H}_f$ we have*

$$\mathrm{in}(\mathrm{in}(\mathcal{U}) = \emptyset$$

and

$$\mathrm{in}(u) \subseteq (\mathcal{U} \cup \mathrm{in}(\mathcal{U}))$$

for any $u \in \mathcal{U}$.

Definition 5 (State)**.** *Any $\sigma \in \Sigma_t$ partitions the real line (its input) into t intervals $I_1, I_2, ..., I_t$ such that on each of these intervals σ is affine. The state of a unit with activation function σ is defined to be $s \in \{1, 2, \ldots, t\}$ if its input belongs to I_s. By extension, the state of $\mathcal{U} \subseteq \mathcal{H}_f$ is defined to be the vector of length $|\mathcal{U}|$ whose components are the state of each unit in $\mathcal{U}$.*

The following definition is inspired by the notion of pattern transition introduced in (Raghu et al., 2017):

Definition 6 (Transition)**.** *Let $f \in \mathbb{F}_\sigma$, $\mathcal{U} \subseteq \mathcal{H}_f$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{R}$. Let $\boldsymbol{z}_\alpha = (1-\alpha)\boldsymbol{x} + \alpha\boldsymbol{y}$ be a parametrization of the line segment $[\boldsymbol{x}, \boldsymbol{y}]$ as α goes from 0 to 1. We say that the state of $\mathcal{U}$ experiences a transition at point $\boldsymbol{z}_{\alpha^*}$ for some $\alpha^* \in (0, 1]$ if the state vector of $\mathcal{U}$ changes at $\boldsymbol{z}_{\alpha^*}$ while the state vector of in$(\mathcal{U})$ does not change at $\boldsymbol{z}_{\alpha^*}$. The number of state transitions of $\mathcal{U}$ on the segment $[\boldsymbol{x}, \boldsymbol{y}]$, denoted by $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U})$, is defined to be the number of state transitions of $\mathcal{U}$ as the input changes from $\boldsymbol{x}$ to $\boldsymbol{y}$ on $\boldsymbol{z}_\alpha$. If $\mathrm{in}(\mathcal{U}) = \emptyset$, then $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U})$ is defined to be the number of state transitions of $\mathcal{U}$ as the input changes from $\boldsymbol{x}$ to $\boldsymbol{y}$.*

Note that if the state vectors of both $\mathcal{U}$ and $\mathrm{in}(\mathcal{U})$ change at α, $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U})$ does not change at that α. For example, consider the neural network f in Fig. 1. Suppose that $\mathcal{U} = \{u_{11}, u_{12}\}$ and suppose that the state of u_{11} and u_{12} changes exactly once along segment z_α for some $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively at α_1 and α_2. Then $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\{u_{11}\}) = 1$ and $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\{u_{12}\}) = 1$. If $\alpha_1 = \alpha_2$, $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U}) = 1$, otherwise $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U}) = 2$. If $\mathcal{U}' = \{u_{21}, u_{22}, u_{23}\}$, and the state of each of u_{21}, u_{22} and u_{23} changes exactly once at either α_1 or α_2, then $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U}') = 0$ since the state vector of $\mathrm{in}(\mathcal{U}') = \mathcal{U}$ has also changed at both α_1 and α_2.

Lemma 3. *Given $f \in \mathbb{F}_\sigma$ and $\mathcal{U}_1, \mathcal{U}_2 \subseteq \mathcal{H}_f$ such that $\mathrm{in}(\mathcal{U}_2) = \emptyset$ and $\mathrm{in}(\mathcal{U}_1) \subseteq \mathcal{U}_2$, we have*

$$N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_1 \cup \mathcal{U}_2\Big) \leq N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_1\Big) + N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_2\Big).$$

Proof. Suppose $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_1 \cup \mathcal{U}_2\Big)$ increases by one at $\alpha = \alpha^*$. If $\mathcal{U}_2$ undergoes a state transition at α^* then, because $\mathrm{in}(\mathcal{U}_2) = \emptyset$, we have that $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_2\Big)$ also increases by one at α^*. Instead, if no state change happens in $\mathcal{U}_2$ at α^* then, due to the state change of $\mathcal{U}_1 \cup \mathcal{U}_2$ at α^*, the state of $\mathcal{U}_1$ must change as well at α^*. Since $\mathrm{in}(\mathcal{U}_1) \subseteq \mathcal{U}_2$ and no change in the state of $\mathcal{U}_2$ is observed at α^* we have that $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_1\Big)$ necessarily increases by one at α^*. □

Lemma 4. *Given $f \in \mathbb{F}_\sigma$ and $\mathcal{U}_1, \mathcal{U}_2 \subseteq \mathcal{H}_f$ such that $\mathcal{U}_1 \subseteq \mathcal{U}_2$ and $\mathrm{in}(\mathcal{U}_2) = \emptyset$ we have*

$$N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_1\Big) \leq N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_2\Big).$$

Proof. Suppose $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_1\Big)$ increases by one at α^*. Since $\mathcal{U}_1 \subseteq \mathcal{U}_2$ the state of $\mathcal{U}_2$ changes as well at α^*. Since $\mathrm{in}(\mathcal{U}_2) = \emptyset$ we deduce that $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}\Big(\mathcal{U}_2\Big)$ increases at α^* by one, thereby concluding the proof. □

Lemma 5. *Given $f \in \mathbb{F}_\sigma$, for any $\mathcal{U} \subseteq \mathcal{H}_f$ we have*

$$N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U}) \leq \sum_{u \in \mathcal{U}} N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(u).$$

Proof. Suppose that $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathcal{U})$ increases by one at α^*. Let $\mathcal{V} \subseteq \mathcal{U}$ be the set of units that experience a transition at α^*. Since we have a transition in the state of $\mathcal{U}$ at α^* we have $\mathcal{V} \neq \emptyset$. Now, because the neural network is cycle-free,[5] there exists some $v \in \mathcal{V}$ such that $\mathrm{in}(v) \cap \mathcal{V} = \emptyset$. We claim that the state of $\mathrm{in}(v)$ has not changed at α^*. To prove this note that by Lemma 2 we have $\mathrm{in}(v) \subseteq \mathrm{in}(\mathcal{U}) \cup \mathcal{U}$ and since $\mathrm{in}(v) \cap \mathcal{V} = \emptyset$ we deduce that $\mathrm{in}(v) \subseteq (\mathrm{in}(\mathcal{U}) \cup \mathcal{U} \backslash \mathcal{V})$. On the other hand neither $\mathcal{U} \backslash \mathcal{V}$ nor $\mathrm{in}(\mathcal{U})$ has a transition at α^*. This implies that $\mathrm{in}(v)$ has no transition at α^* and therefore $N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(v)$ increases by one at α^*. This concludes the proof since $v \in \mathcal{U}$. □

Lemma 6. *Given $f \in \mathbb{F}_\sigma$, for any $u \in \mathcal{H}_f$ we have*

$$N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(u) \leq (t-1)\Big(N_{\boldsymbol{x}\rightarrow\boldsymbol{y}}(\mathrm{in}(u)) + 1\Big).$$

Proof. To establish the lemma we show that between transitions of $\mathrm{in}(u)$ there are at most $t-1$ transitions of u.

[5] Recall that throughout the paper neural networks are feedforward.

Suppose, by way of contradiction, that at least t transitions in the state of u happen while $\mathrm{in}(u)$ experiences no change. Then there exists an increasing sequence of real numbers $\alpha_1, ..., \alpha_{t+1}$ from interval $[0, 1]$ and an increasing set of integers $k_1, k_2, ..., k_{t+1}$ from $S = \{1, 2, ..., t\}$, with $k_i \neq k_{i+1}$, such that for particular $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ we have

$$\begin{aligned} \boldsymbol{x_i} &\overset{\text{def}}{=} (1-\alpha_i)\boldsymbol{x} + \alpha_i \boldsymbol{y} \\ \boldsymbol{w} \cdot \boldsymbol{x_i} &+ b \in I_{k_i} \end{aligned}$$

where I_i is defined in Definition 5. Since $|S| = t$ there exists $i < j$ such that $k_i = k_j$. Now since $k_i \neq k_{i+1}$ we deduce that $j \neq i+1$ and therefore $j > i+1$. But $\boldsymbol{w} \cdot \boldsymbol{x_{i+1}} + b$ lies between $\boldsymbol{w} \cdot \boldsymbol{x_i} + b$ and $\boldsymbol{w} \cdot \boldsymbol{x_j} + b$ since the sequence $\alpha_1, \alpha_2, ..., \alpha_{t+1}$ is increasing. Since $\boldsymbol{w} \cdot \boldsymbol{x_j} + b$ and $\boldsymbol{w} \cdot \boldsymbol{x_i} + b$ belong to I_{k_i}, by the connectedness property of the set I_i we deduce that that $\boldsymbol{w} \cdot \boldsymbol{x_{i+1}} + b \in I_i$. Therefore, we get $k_i = k_{i+1} = k_j$, a contradiction. □

Since a break point of $f \in \mathbb{F}_\sigma$ necessarily implies a change in the state of the units we get:

Lemma 7. *Given* $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{R}^2$ *and* $f \in \mathbb{F}_\sigma$ *we have*

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f) \leq N_{\boldsymbol{x}\to\boldsymbol{y}}(\mathcal{H}_f).$$

Propositions 1 and 2 establish inequalities (2) and (3) of Theorem 1.

Proposition 1. *Given* $f \in \mathbb{F}_\sigma$, $\sigma \in \Sigma_t$, *we have*

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f) \leq \Big((t-1)\omega_f + 1\Big)^{d_f} - 1. \tag{10}$$

Proof of Proposition 1. Fix $f \in \mathbb{F}_\sigma$ where $\sigma \in \Sigma_t$. Referring to Definition 1, consider the partition

$$\cup_{i=1}^{d} \mathcal{H}_f^i$$

of $\mathcal{H}_f$ according to unit depth where $d = d_f$.

Fix $u \in \mathcal{H}_f^{i+1}$, $0 \leq i < d$. From the definitions of $\mathrm{in}(u)$ and $\mathcal{H}_f^i$ we get

$$\mathrm{in}(u) \subseteq \bigcup_{j=1}^{i} \mathcal{H}_f^j \tag{11}$$

$$\mathrm{in}\Big(\mathcal{H}_f^{i+1}\Big) \subseteq \bigcup_{j=1}^{i} \mathcal{H}_f^j$$

$$\mathrm{in}\Big(\bigcup_{j=1}^{i} \mathcal{H}_f^j\Big) = \emptyset.$$

Applying Lemma 3 with $\mathcal{U}_1 = \mathcal{H}_f^{i+1}$ and $\mathcal{U}_2 = \bigcup_{j=1}^{i} \mathcal{H}_f^j$ we get

$$N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i+1} \mathcal{H}_f^j) \leq N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i} \mathcal{H}_f^j) + N_{\boldsymbol{x}\to\boldsymbol{y}}(\mathcal{H}_f^{i+1}).$$

From Lemma 5

$$N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i+1} \mathcal{H}_f^j) \leq N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i} \mathcal{H}_f^j) + \sum_{u \in \mathcal{H}_f^{i+1}} N_{\boldsymbol{x}\to\boldsymbol{y}}(u)$$

and applying Lemma 6 to the previous inequality

$$\begin{aligned} N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i+1} \mathcal{H}_f^j) &\leq N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i} \mathcal{H}_f^j) \\ &+ \sum_{u \in \mathcal{H}_f^{i+1}} (t-1)\Big(N_{\boldsymbol{x}\to\boldsymbol{y}}\big(\mathrm{in}(u)\big) + 1\Big). \end{aligned}$$

Then, using (11) and Lemma 4 we get

$$\begin{aligned} &N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i+1} \mathcal{H}_f^j) \\ &\leq N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i} \mathcal{H}_f^j) + \sum_{u \in \mathcal{H}_f^{i+1}} (t-1)\Big(N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i} \mathcal{H}_f^j) + 1\Big) \\ &= \big(\omega_{i+1}(t-1) + 1\big) N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{i} \mathcal{H}_f^j) + \omega_{i+1}(t-1). \end{aligned} \tag{12}$$

For $u \in \mathcal{H}_f^1$ we have $\mathrm{in}(u) = \emptyset$ and according to Lemma 6 we deduce that $N_{\boldsymbol{x}\to\boldsymbol{y}}(\mathcal{H}_f^1) \leq (t-1)\omega_1$. With this initial condition and the recursive relation in (12) we get

$$\begin{aligned} &N_{\boldsymbol{x}\to\boldsymbol{y}}(\bigcup_{j=1}^{d} \mathcal{H}_f^j) \\ &\leq \sum_{j=1}^{d} \Bigg(\prod_{1 \leq \alpha_1 < \alpha_2 < \ldots < \alpha_j \leq d} \omega_{\alpha_1}\omega_{\alpha_2}\cdots\omega_{\alpha_j} \big(t-1\big)^j \Bigg) \\ &\leq \sum_{j=1}^{d} \Bigg(\binom{d}{j} \big(\omega_f(t-1)\big)^j \Bigg) = \Big(\omega_f(t-1) + 1\Big)^d - 1 \end{aligned}$$

with ω_f as width of f. Finally, apply Lemma 7 to obtain

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f) \leq \Big((t-1)\omega_f + 1\Big)^{d_f} - 1.$$

□

Proposition 2. *Let* $\mathcal{R}$ *be a convex region in* $\mathbb{R}^n$. *For any affine* ε*-approximation* $f : \mathcal{R} \to \mathbb{R}$ *of a function* $g \in C^2(\mathcal{R})$ *we have*

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f) \geq \frac{||\boldsymbol{x}-\boldsymbol{y}||_2}{4\sqrt{\varepsilon}} \cdot \Psi(g, \boldsymbol{x}, \boldsymbol{y}) - 1 \tag{13}$$

where $\Psi(g, \boldsymbol{x}, \boldsymbol{y})$ *is defined in* (4).

Proof of Proposition 2. We partition $\mathcal{R}$ into *convex* subregions $\mathcal{R}_i$, such that in each subregion $f(\boldsymbol{x})$ is an affine function. These convex subregions partition a segment $[\boldsymbol{x}, \boldsymbol{y}]$ into sub-segments with end points $\Big\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_s\Big\}$, where $\boldsymbol{x}_0 = \boldsymbol{x}, \boldsymbol{x}_s = \boldsymbol{y}$ and $s = B_{\boldsymbol{x}\to\boldsymbol{y}}(f)+1$. In the sub-segment $i \in \{0, 1, ..., s-1\}$,

$$f(\boldsymbol{x}) = \boldsymbol{p}_i.\boldsymbol{x} + q_i, \quad \boldsymbol{x} \in [\boldsymbol{x}_i, \boldsymbol{x}_{i+1}], \tag{14}$$

for some $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$. Let $\boldsymbol{x}_i(r) = (1-r)\boldsymbol{x}_i + r\boldsymbol{x}_{i+1}$, $r \in [0,1]$, and define

$$\begin{aligned} f_i(r) &= (1-r)g(\boldsymbol{x}_i) + rg(\boldsymbol{x}_{i+1}), \\ h_i(r) &= g\big(\boldsymbol{x}_i(r)\big), \\ l_i(r) &= f(\boldsymbol{x}(r)). \end{aligned}$$

From the definition of ε-approximation, $||h_i(r) - l_i(r)||_\infty \le \varepsilon$. Thus

$$\begin{aligned} ||f_i(r) - h_i(r)||_\infty &\le ||f_i(r) - l_i(r)||_\infty + ||l_i(r) - h_i(r)||_\infty \\ &\overset{(a)}{\le} \max\big\{|f_i(0) - l_i(0)|, |f_i(1) - l_i(1)|\big\} + \varepsilon \\ &\le 2\varepsilon, \end{aligned} \tag{15}$$

where $||k(r)||_\infty = \sup\limits_{0\le r\le 1} k(r)$ and step (a) follows because $f_i(r)$ and $l_i(r)$ are both line segments and the maximum distance between them is achieved at end points.

As $h(r)$ on $(0,1)$ is differentiable so there exists $r_i^* \in (0,1)$ such that $h_i'(r_i^*) = h_i(1) - h_i(0)$. Consider $\boldsymbol{x}_i^* = (1 - r_i^*)\boldsymbol{x}_i + r_i^*\boldsymbol{x}_{i+1}$. From (15) we obtain

$$\begin{aligned} &|(1 - r_i^*)\big(g(\boldsymbol{x}_i) - g(\boldsymbol{x}_{i+1})\big) - g(\boldsymbol{x}_i^*) + g(\boldsymbol{x}_{i+1})| \le 2\varepsilon, \\ &|r_i^*\big(g(\boldsymbol{x}_{i+1}) - g(\boldsymbol{x}_i)\big) + g(\boldsymbol{x}_i) - g(\boldsymbol{x}_i^*)| \le 2\varepsilon. \end{aligned}$$

Then, from the definition of r_i^* we have

$$|(r_i^* - 1)\nabla g(\boldsymbol{x}_i^*).(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) - g(\boldsymbol{x}_i^*) + g(\boldsymbol{x}_{i+1})| \le 2\varepsilon \tag{16}$$

$$|r_i^*\nabla g(\boldsymbol{x}_i^*).(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) - g(\boldsymbol{x}_i^*) + g(\boldsymbol{x}_i)| \le 2\varepsilon. \tag{17}$$

Since $g \in C^2(\mathcal{R})$ a Taylor expansion of $g(\boldsymbol{x}_i)$ and $g(\boldsymbol{x}_{i+1})$ around x_i^* gives

$$\begin{aligned} g(\boldsymbol{x}_i) &= g(\boldsymbol{x}_i^*) - r_i^*\nabla g\big(\boldsymbol{x}_i^*\big).(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) \\ &+ \frac{r_i^{*2}}{2}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i), \\ g(\boldsymbol{x}_{i+1}) &= g(\boldsymbol{x}_i^*) + (1 - r_i^*)\nabla g\big(\boldsymbol{x}_i^*\big).(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) \\ &+ \frac{(1 - r_i^*)^2}{2}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{x}_i(\beta_i)\big)(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i), \end{aligned}$$

where $0 \le \alpha_i \le r_i^* \le \beta_i \le 1$.

Substituting the above relations in inequalities (16) and (17) we get

$$|(1 - r_i^*)^2(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{x}_i(\beta_i)\big)(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)| \le 4\varepsilon, \tag{18}$$

$$|r_i^{*2}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)| \le 4\varepsilon. \tag{19}$$

Use the *Rayleigh quotient* and the definitions of $\theta(\alpha), \gamma(\alpha)$ to obtain

$$\begin{aligned} &\Big|\frac{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)}{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)^T(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)}\Big| \\ &\ge \inf_{0\le\alpha\le 1}\Big(\max\big\{0, \theta(\alpha)\gamma(\alpha)\big\}\Big). \end{aligned}$$

Combining the above inequality with (18) and (19) and the fact that $r_i^{*2} + (1 - r_i^*)^2 \ge \frac{1}{2}$ we get

$$||\boldsymbol{x}_{i+1} - \boldsymbol{x}_i||_2{}^2. \inf_{0\le\alpha\le 1}\Big(\max\big\{0, \theta(\alpha)\gamma(\alpha)\big\}\Big) \le 16\varepsilon.$$

Accordingly,

$$\sum_{i=0}^{s-1}\left(\frac{||\boldsymbol{x}_{i+1} - \boldsymbol{x}_i||_2}{4\sqrt{\varepsilon}}.\sqrt{\inf_{0\le\alpha\le 1}\Big(\max\big\{0, \theta(\alpha)\gamma(\alpha)\big\}\Big)}\right) \le s,$$

which gives

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f) \ge \frac{||\boldsymbol{x} - \boldsymbol{y}||_2}{4\sqrt{\varepsilon}}\Psi(g, \boldsymbol{x}, \boldsymbol{y}) - 1.$$

□

Proposition 3. *Let $g : [0,1]^n \to \mathbb{R}$ be such that $D^J(g)(\boldsymbol{x}) \le \delta$ for any $\boldsymbol{x} \in [0,1]^n$ and any multi-index J such that $|J| = 3$. Then, for any affine ε-approximation f*

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f) \ge \sqrt{\frac{\Big(\max\limits_{\boldsymbol{x}\in[0,1]^n}\big|\Delta(g)(\boldsymbol{x})\big| \cdot n^{-1} - \delta \cdot n^{\frac{3}{2}}\Big)^+}{16\varepsilon}} - 1$$

for any $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^n$, where Δ denotes the Laplace operator (7).

Proof of Proposition 3. Define

$$\boldsymbol{z} \overset{\text{def}}{=} \arg\max_{\boldsymbol{x}\in\mathcal{R}} \rho\big(\nabla^2 g(\boldsymbol{x})\big)$$

where $\rho(\cdot)$ denotes the spectral radius. Let $\boldsymbol{u}$ be a normalized eigenvector corresponding to an eigenvalue λ where $|\lambda| = \rho\big(\nabla^2 g(\boldsymbol{z})\big)$, *i.e.*,

$$\nabla^2 g(\boldsymbol{z})\boldsymbol{u} = \lambda\boldsymbol{u}, \quad ||\boldsymbol{u}|| = 1. \tag{20}$$

Consider any segment $[\boldsymbol{x}, \boldsymbol{y}]$ in $\mathcal{R}$ in the direction of $\boldsymbol{u}$, *i.e.*, such that $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{u}$. The convex subregions of f, defined in the proof of Proposition 2, divide this segment into sub-segments with end points $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ..., \boldsymbol{x}_s\}$ where $\boldsymbol{x}_0 = \boldsymbol{x}, \boldsymbol{x}_s = \boldsymbol{y}$ and $s = B_{\boldsymbol{x}\to\boldsymbol{y}}(f) + 1$. Using the same analysis as in the proof of Proposition 2, from (14)–(19) we obtain (18) and (19). On the other hand, note that

$$
\begin{aligned}
&|(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)| \\
&\geq |(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)^T\nabla^2 g\big(\boldsymbol{z}\big)(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)| \\
&-|(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)^T\Big(\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)-\nabla^2 g\big(\boldsymbol{z}\big)\Big)(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)| \\
&= |\lambda|\cdot||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2 \\
&-\Big|\mathrm{tr}\big\{\big(\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)-\nabla^2 g\big(\boldsymbol{z}\big)\big)(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)^T\big\}\Big| \\
&\overset{(a)}{\geq} |\lambda|\cdot||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2 \\
&-\big|\big|\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)-\nabla^2 g\big(\boldsymbol{z}\big)\big|\big|_{\mathrm{F}}\big|\big|(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)(\boldsymbol{x}_{i+1}-\boldsymbol{x}_i)^T\big|\big|_{\mathrm{F}} \\
&= |\lambda|\cdot||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2 \\
&-\big|\big|\nabla^2 g\big(\boldsymbol{x}_i(\alpha_i)\big)-\nabla^2 g\big(\boldsymbol{z}\big)\big|\big|_{\mathrm{F}}||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2 \\
&= ||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2\cdot\Big(|\lambda|-n\delta\cdot||\boldsymbol{z}-\boldsymbol{x}_i(\alpha_i)||\Big) \\
&\geq ||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2\cdot\Big(|\lambda|-\delta\cdot n^{\frac{3}{2}}\Big),
\end{aligned}
$$

where in step (a) we used the inequality

$$\Big|\mathrm{tr}\big(AB\big)\Big| \leq ||A||_F||B||_F,$$

$||\cdot||_F$ stands for Frobenius norm.

Combining the above relation with (18), (19) and the fact that ${r_i^*}^2+(1-r_i^*)^2\geq\frac{1}{2}$ we get

$$16\varepsilon \geq ||\boldsymbol{x}_{i+1}-\boldsymbol{x}_i||^2\cdot\Big(|\lambda|-\delta\cdot n^{\frac{3}{2}}\Big),$$

which gives

$$4\sqrt{\varepsilon}\cdot\big(B_{\boldsymbol{x}\to\boldsymbol{y}}(f)+1\big)\geq||\boldsymbol{x}-\boldsymbol{y}||\cdot\sqrt{\Big(|\lambda|-\delta\cdot n^{\frac{3}{2}}\Big)^+}.$$

Finally, rewriting the above inequality we get

$$B_{\boldsymbol{x}\to\boldsymbol{y}}(f)\geq\frac{1}{4\sqrt{\varepsilon}}\cdot\sqrt{\Big(|\lambda|-\delta\cdot n^{\frac{3}{2}}\Big)^+}-1.$$

Since $|\lambda|=\rho\big(\nabla^2 g(\boldsymbol{z})\big)=\max_{\boldsymbol{x}\in[0,1]^n}\rho\big(\nabla^2 g(\boldsymbol{x})\big)$ and

$$|\Delta(g)(\boldsymbol{x})|=|\mathrm{tr}(\nabla^2 g(\boldsymbol{x}))|\leq\rho(\nabla^2 g(\boldsymbol{x}))\cdot n,$$

we obtain the desired result. □

Proofs of Theorems 1 and 2

Propositions 1 and 2 give Theorem 1 and Propositions 1 and 3 give Theorem 2. □

Proof of Theorem 3

Given a neural network f we use o to denote the output unit, $\mathrm{w}(u,v)$ to denote the weight of two connected units u and v, and $b(u)$ to denote the bias of unit u. Furthermore, given $u\in\mathcal{H}_f$ and $\boldsymbol{x}\in\mathcal{R}$ let $f_1^u(\boldsymbol{x})$ denote the output of unit u when the input to f_1 is $\boldsymbol{x}$, and similarly for $f_2(\boldsymbol{x})$. Finally, define the maximum change in hidden layer i as

$$\varepsilon_i(\boldsymbol{x})\overset{\mathrm{def}}{=}\max_{u\in\mathcal{H}_f^i}\Big\{\big|f_1^u(\boldsymbol{x})-f_2^u(\boldsymbol{x})\big|\Big\}.$$

Fix $1\leq i\leq d_f-1$ and $v\in\mathcal{H}_f^{i+1}$. Then,

$$
\begin{aligned}
&\big|f_1^v(\boldsymbol{x})-f_2^v(\boldsymbol{x})\big| \\
&=\Bigg|\sigma_1\Big(\sum_{u\in\bigcup_{j=1}^{i}\mathcal{H}_f^j} w(u,v)\cdot f_1^u(\boldsymbol{x})+b(v)\Big) \\
&-\sigma_2\Big(\sum_{u\in\bigcup_{j=1}^{i}\mathcal{H}_f^j} w(u,v)\cdot f_2^u(\boldsymbol{x})+b(v)\Big)\Bigg| \\
&\leq\varepsilon+\delta\cdot\Big(\sum_{u\in\bigcup_{j=1}^{i}\mathcal{H}_f^j}|w(u,v)|\cdot\big|f_1^u(\boldsymbol{x})-f_2^u(\boldsymbol{x})\big|\Big) \\
&\leq\varepsilon+\delta A\cdot\Big(\sum_{j=1}^{i}\sum_{u\in\mathcal{H}_f^j}\big|f_1^u(\boldsymbol{x})-f_2^u(\boldsymbol{x})\big|\Big) \\
&\leq\varepsilon+\delta A\cdot\Big(\sum_{j=1}^{i}\omega_j\varepsilon_j(\boldsymbol{x})\Big)
\end{aligned}
$$

where the first inequality holds since σ_1 is δ-Lipschitz and assuming that $||\sigma_1-\sigma_2||_\infty\leq\varepsilon$. Hence we get the recursion between ε_i's

$$\varepsilon_{i+1}(\boldsymbol{x})\leq\varepsilon+\delta A\cdot\Big(\sum_{j=1}^{i}\omega_j\varepsilon_j(\boldsymbol{x})\Big)\tag{21}$$

for $1\leq i\leq d_f-1$. Now, since $\varepsilon_1(\boldsymbol{x})\leq\big|\sigma_1(\boldsymbol{x})-\sigma_2(\boldsymbol{x})\big|$ we get $\varepsilon_1(\boldsymbol{x})\leq\varepsilon$. From this initial condition and (21)

$$\varepsilon_{i+1}(\boldsymbol{x})\leq\varepsilon(1+\delta A\omega_1)(1+\delta A\omega_2)\cdots(1+\delta A\omega_i).\tag{22}$$

On the other hand we have

$$
\begin{aligned}
&|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|=\Big|\sum_{u\in\bigcup_{j=1}^{d_f}\mathcal{H}_f^j}\mathrm{w}(u,o)\cdot\big(f_1^u(\boldsymbol{x})-f_2^u(\boldsymbol{x})\big)\Big| \\
&\leq A\big(\varepsilon_1(\boldsymbol{x})\omega_1+\varepsilon_2(\boldsymbol{x})\omega_2+...+\varepsilon_d(\boldsymbol{x})\omega_{d_f}\big)
\end{aligned}
$$

and from (22) we finally get

$$
\begin{aligned}
&|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})| \\
&\leq\frac{\varepsilon}{\delta}\Big((1+\delta A\omega_1)(1+\delta A\omega_2)...(1+\delta A\omega_{d_f})-1\Big) \\
&\leq\frac{||\sigma_1-\sigma_2||_\infty}{\delta}\Bigg(\Big(\delta\cdot A\cdot\omega_f+1\Big)^{d_f}-1\Bigg)
\end{aligned}
$$

which gives the desired result. □

References

Bianchini, Monica and Scarselli, Franco. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565, 2014.

DasGupta, Bhaskar and Schnitger, Georg. The power of approximating: A comparison of activation functions. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 615–622, 1993.

Delalleau, Olivier and Bengio, Yoshua. Shallow vs. deep sum-product networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 666–674, 2011.

Hornik, Kurt, Stinchcombe, Maxwell, and White, Halbert. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

Liang, Shiyu and Srikant, R. Why deep neural networks for function approximation? In *5th International Conference on Learning Representations (ICLR)*, pp. 1–13, 2017.

Lu, Zhou, Pu, Hongming, Wang, Feicheng, Hu, Zhiqiang, and Wang, Liwei. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, pp. 6232–6240, 2017.

Mhaskar, Hrushikesh, Liao, Qianli, and Poggio, Tomaso. Learning functions: When is deep better than shallow. *arXiv preprint, arXiv:1603.00988*, 2016.

Montufar, Guido F, Pascanu, Razvan, Cho, Kyunghyun, and Bengio, Yoshua. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2924–2932, 2014.

Pascanu, Razvan, Montufar, Guido, and Bengio, Yoshua. On the number of inference regions of deep feed forward networks with piece-wise linear activations. In *International Conference on Learning Representations*, 2014.

Raghu, Maithra, Poole, Ben, Kleinberg, Jon, Ganguli, Surya, and Sohl-Dickstein, Jascha. On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICML)*, pp. 2847–2854, 2017.

Telgarsky, Matus. Representation benefits of deep feedforward networks. *arXiv preprint, arXiv:1509.08101*, 2015.

Telgarsky, Matus. Benefits of depth in neural networks. *Journal of Machine Learning Research (JMLR)*, 49:1–23, 2016.

Yarotsky, Dmitry. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.