This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this article is published in
Sampling Theory, Signal Processing, and Data Analysis,
and is available online at
https://doi.org/10.1007/s43670-022-00020-y

On Sharpness of an Error Bound for Deep ReLU Network Approximation

Steffen Goebbels

Abstract Zuowei Shen, Haizhao Yang, and Shijun Zhang recently proved a sharp estimate in terms of a modulus of continuity for the error of approximating continuous functions with ReLU-activated deep neural networks in their paper "Optimal approximation rate of ReLU networks in terms of width and depth", Journal de Mathématiques Pures et Appliqués (2021). The sharpness was established based on a VC dimension estimate by showing that for each choice of certain fixed width and depth parameters, no general smaller error bound applies. In principle, the obtained counterexamples can be different for different parameters. However, for a given convergence rate (e.g., determined by a Lipschitz class), the counterexamples can be condensed to a single counterexample using a quantitative variant of the uniform boundedness principle. Such theorems were developed at the institute of Paul Butzer at RWTH Aachen. When network width and depth simultaneously tend to infinity in certain ratios, the condensed counterexamples show that the approximation order is not in the little-o class of the modulus of continuity used in the error estimate, i.e., the convergence cannot be faster than stated.

Keywords Neural Networks · Sharpness of Error Bounds · Counterexamples · Rates of Convergence · Uniform Boundedness Principle

Mathematics Subject Classification (2010) MSC 41A25 · 41A50 · 62M45

Steffen Goebbels
Niederrhein University of Applied Sciences, Faculty of Electrical Engineering and Computer Science, Institute for Pattern Recognition, D-47805 Krefeld
Tel.: +49-2151-8224648
E-mail: Steffen.Goebbels@hsnr.de

1 Introduction

A trained neural network represents a function that maps input values to an output value. The function depends on parameters called weights and biases, which are determined in a learning phase. In supervised learning, the target outputs belonging to sample input values are given. Training then consists of minimizing the differences between the network output and the target output, often by applying gradient descent (back-propagation). Neural networks are thus in principle able to reconstruct functions from sample values within certain error bounds. The error is composed from approximation, optimization and generalization errors. Instead of finding optimal network weights and biases in the learning phase, gradient descent can lead to local optima and hence optimization errors. Since the network is trained only on sample data, it may compute undesirable outputs when applied to other data. Then it does not generalize. This error is also known as overfitting. However, the paper focuses on the approximation error and not on aspects of network learning, i.e., it focuses on the ability to approximate a function by the space of all functions that can be realized with the network by varying weights and biases. Already function spaces belonging to single layer neural networks can be used to approximate continuous functions arbitrary well. This is known as the universal approximation property, see [3,6,10,11]. Associated rates of convergence have also been studied by many authors. Although the strength of approximation with neural network lies in their non-linearity, moduli of continuity or smoothness have nevertheless been proven to be suitable to express these convergence rates, see the literature overviews in [7,12].

For $\mathbf{x} \in \mathbb{R}^d$ let $\|\mathbf{x}\|_2 := \sqrt{\sum_{k=1}^d x_k^2}$ be the 2-norm. Let a real-valued continuous function f on $[0,1]^d$ be given, i.e., $f \in C([0,1]^d)$. The modulus of continuity is defined via first order differences:

$$\omega(f,\delta) := \max\{|f(\mathbf{x}) - f(\mathbf{y})| : \mathbf{x},\mathbf{y} \in [0,1]^d, \|\mathbf{x}-\mathbf{y}\|_2 \le \delta\}.$$

For the properties of this modulus see, e.g., [9].

In this paper, deep ReLU-activated neural networks with d real input values, (at most) $\tilde{L}$ hidden layers and $\tilde{N}_k$ neurons in the hidden layer $k \in \{1,\dots,\tilde{L}\}$ are discussed. The well-known ReLU (Rectified Linear Unit) activation function σ is defined as

$$\sigma(x) := \begin{cases} 0, & x < 0 \\ x, & x \ge 0. \end{cases}$$

Although non-differentiable in $x = 0$, the computational simplicity makes the ReLU function σ the standard activation function in most current deep neural network applications.

The activation function σ is applied component-wise to a column vector, i.e.,

$$\sigma\left((x_1, x_2, \dots)^\top\right) := (\sigma(x_1), \sigma(x_2), \dots)^\top.$$

Let $\tilde{N}_0 := d$ and $\tilde{N}_{\tilde{L}+1} := 1$. Each layer k has a real-valued weight matrix $\mathbf{W}_k \in \mathbb{R}^{\tilde{N}_k, \tilde{N}_{k-1}}$, with $\tilde{N}_k$ rows and $\tilde{N}_{k-1}$ columns, and a bias vector $\mathbf{b}_k \in \mathbb{R}^{\tilde{N}_k}$ with $\tilde{N}_k$ components in one column. Additionally, $\mathbf{W}_{\tilde{L}+1} \in \mathbb{R}^{1,\tilde{N}_{\tilde{L}}}$ represents weights of the output layer and the only component of $\mathbf{b}_{\tilde{L}+1}$ is the bias to be added to the output value. Thus, the paper is restricted to networks that realize a real-valued and not a vector-valued function. Let $\mathbf{x} \in \mathbb{R}^d$ be an input vector, then the network computes the function value in $\mathbf{y}_{\tilde{L}+1} := \mathbf{W}_{\tilde{L}+1}\mathbf{y}_{\tilde{L}} + \mathbf{b}_{\tilde{L}+1}$ from input $\mathbf{x} \in \mathbb{R}^d$ via $(1 \le k < \tilde{L})$

$$\mathbf{y}_1 := \sigma\left(\mathbf{W}_1 \cdot \mathbf{x} + \mathbf{b}_1\right), \quad \mathbf{y}_{k+1} := \sigma\left(\mathbf{W}_{k+1} \cdot \mathbf{y}_k + \mathbf{b}_{k+1}\right). \tag{1}$$

The following error bound is proved in [12]: For each function $f \in C([0,1]^d)$, and each choice of $N, L \in \mathbb{N} := \{1, 2, \dots\}$, there exists a function Φ given by a ReLU-activated neural network with d input nodes, at most $\hat{N}(N) := 3^{d+3}\max\{d\lfloor\sqrt[d]{N}\rfloor, N+2\}$ neurons per hidden layer and at most $\hat{L}(L) := 11L + 18 + 2d$ hidden layers such that

$$\begin{aligned}\|f - \Phi\|_\infty &:= \sup\{|f(\mathbf{x}) - \Phi(\mathbf{x})| : \mathbf{x} \in [0,1]^d\} \\ &\le 131\sqrt{d} \cdot \omega\left(f, \frac{1}{\sqrt[d]{N^2L^2\log_3(N+2)}}\right).\end{aligned} \tag{2}$$

Thus, parameters N and L are used to specify the maximum number $\hat{N}$ of hidden neurons per layer and the maximum number $\hat{L}$ of hidden layers, respectively. Based on the VC dimension, the bound is shown to be best possible in [12] for certain fixed choices of L and N. This is possible because of the logarithm in the denominator of the modulus parameter $\delta = (\sqrt[d]{N^2L^2\log_3(N+2)})^{-1}$. This logarithm also occurs in VC dimension estimates but is difficult to obtain in direct estimates.

Shen et al. also provide bounds for L^p norms of the error in [12], but these errors are estimated against a sup-norm modulus of continuity rather than an L^p modulus. Therefore, only sup-norms are considered in what follows.

In the next sections, parameters L and N are coupled and the existence of counterexample functions is proved for which the convergence order determined by the modulus of continuity in (2) cannot be improved in terms of a little-o estimate. This is done on the basis of an existing VC dimension estimate combined with a quantitative uniform boundedness principle that condenses a counterexample from a series of test functions. Quantitative extensions of the uniform boundedness principle (known from Functional Analysis) and their applications to Approximation Theory and Fourier Analysis were the subject of a research group at the institute of Paul Butzer at RWTH in Aachen. The group consisted over time of Rolf Nessel and Werner Dickmeis, as well as various graduate students, including Erich van Wickeren, Herbert Mevissen, Bernhard Büttgenbach, Gerald Lüttgens, Ralf Zeler, and the author. The Banach-Steinhaus theorem with rates by Paul Butzer, Karl Scherer and Ursula Schmidt-Westphal [2] inspired the work of this group, see [4]. While most of the quantitative uniform boundedness theorems have been

formulated for sub-linear bounded functionals on Banach spaces, sub-linearity is violated when it comes to non-linear approximation with neural networks. In the following sections, a non-linear version of the theorem in [5, p. 108] is used. This non-linear variant was proved in [8]. It was specialized in [7] to be directly applicable to VC dimension estimates, and was then applied to analyze the error of best multivariate approximation with single hidden layer neural networks. The given paper demonstrates that it can also be used to analyze multi-layer networks that are applied as deep neural networks in many current machine learning applications.

2 Sharpness Result

To couple width parameter N and depth parameter L, these values are written as functions $N(n)$ and $L(n)$ of a common parameter $n \in \mathbb{N}$. There may be a gap between the convergence rate of the direct estimate and the VC dimension estimate from which sharpness is derived. This gap does not occur when the depth parameter $L(n)$ is upper bounded or when the width parameter $N(n)$ is either upper bounded or when the maximum width is lower bounded by $L(n)^\gamma$ for a real power $\gamma > 0$, i.e., $N(n) \geq L(n)^\gamma$, see [12]. While the simpler case of bounded depth is briefly discussed in the conclusions section, the rest of the paper is mainly concerned with the effect of increasing depth by choosing $L(n) := n$ and $N(n) := M\lfloor 1 + n^\gamma \rfloor$ for a fixed $\gamma \geq 0$ (the case of bounded width is included with $\gamma = 0$) and a constant $M \in \mathbb{N}$, $M \geq 3$, that is chosen large enough to get $d\sqrt[d]{M} \leq M$, i.e., $M \geq d^{\frac{d}{d-1}}$ if $d > 1$, and thus

$$\tilde{N}(n) := \hat{N}(N(n)) = 3^{d+3}(N(n) + 2) = 3^{d+3}(M\lfloor 1 + n^\gamma \rfloor + 2).$$

For the later proof of condition (9), a slightly larger depth than required by the direct estimate is chosen:

$$\tilde{L}(n) := (29 + 2d)n \geq \hat{L}(L(n)) = \hat{L}(n) = 11n + 18 + 2d.$$

Then the direct bound (2) with $N = N(n) = M\lfloor 1 + n^\gamma \rfloor$ and $L = L(n) = n$ also holds for larger networks with width $\tilde{N}(n)$ and depth $\tilde{L}(n)$.

Let V_n be the space of all (continuous) functions on $[0, 1]^d$ which can be represented by a ReLU neural network (1) with maximum width $\tilde{N}(n)$ and maximum number of hidden layers $\tilde{L}(n)$. Further let E_n be the error of best approximation measured in the sup-norm, i.e.,

$$E_n(f) := \inf\{\|f - \Phi\|_\infty : \Phi \in V_n\}.$$

Then, from (2), by applying the properties of the modulus of continuity (see [9]), one obtains the following estimate for $\gamma > 0$, each $f \in C([0, 1]^d)$ and

$n \in \mathbb{N}$:

$$\begin{aligned}
E_n(f) &\leq 131\sqrt{d} \cdot \omega\left(f, \frac{1}{\sqrt[d]{M^2\lfloor 1+n^\gamma\rfloor^2 n^2 \log_3(M\lfloor 1+n^\gamma\rfloor+2)}}\right) \\
&\leq 131\sqrt{d} \cdot \omega\left(f, \frac{\sqrt[d]{\ln(3)}}{\sqrt[d]{M^2 n^{2\gamma+2}\ln(Mn^\gamma)}}\right) \\
&= 131\sqrt{d} \cdot \omega\left(f, \frac{\sqrt[d]{\ln(3)}}{\sqrt[d]{M^2 n^{2\gamma+2}[\ln(M)+\gamma\ln(n)]}}\right) \\
&\leq 131\sqrt{d}\left[\sqrt[d]{\frac{\ln(3)}{M^2\min\{\ln(M),\gamma\}}}+1\right] \cdot \omega\left(f, \frac{1}{\sqrt[d]{n^{2\gamma+2}(1+\ln(n))}}\right) \\
&= C_{d,\gamma} \cdot \omega\left(f, \frac{1}{\sqrt[d]{n^{2\gamma+2}(1+\ln(n))}}\right). \qquad (3)
\end{aligned}$$

In the case of bounded width, i.e., $\gamma = 0$, one obtains

$$E_n(f) \leq C_{d,0} \cdot \omega\left(f, \frac{1}{\sqrt[d]{n^2}}\right). \qquad (4)$$

Based on VC dimension estimation, a corollary of the optimality result in [12] is that the error estimate cannot be improved for functions belonging to Lipschitz classes. Let $0 < \alpha \leq 1$ then there exists a constant $c_\alpha > 0$, independent of $n \geq n_0$, such that

$$\begin{aligned}
&\sup\left\{E_n(f) : f \in C([0,1]^d) \wedge \omega(f,\delta) \in O\left(\delta^\alpha\right)\right\} \\
&\geq c_\alpha\left(\sqrt[d]{N(n)^2 L(n)^2 \log_3(N(n)+2)}\right)^{-\alpha}. \qquad (5)
\end{aligned}$$

This estimate does not exclude that for all $f \in C([0,1]^d)$ with $\omega(f,\delta) \in O\left(\delta^\alpha\right)$

$$\lim_{n\to\infty} \frac{E_n(f)}{\left(\sqrt[d]{N(n)^2 L(n)^2 \log_3(N(n)+2)}\right)^{-\alpha}} = 0,$$

i.e, $E_n(f) = o((\sqrt[d]{N(n)^2 L(n)^2 \log_3(N(n)+2)})^{-\alpha})$, could hold true. Despite of (5), that would imply that convergence could be faster for each single function than suggested by the error bound.

To additionally show the sharpness of this estimate with regard to little-o rates, abstract moduli of smoothness ω are applied, see [13, p. 96ff]: An abstract modulus of smoothness is a continuous, increasing function $\omega : [0,\infty) \to [0,\infty)$ such that for $\delta_1, \delta_2 > 0$

$$0 = \omega(0) < \omega(\delta_1) \leq \omega(\delta_1+\delta_2) \leq \omega(\delta_1)+\omega(\delta_2). \qquad (6)$$

When dealing with Lipschitz classes, one chooses $\omega(\delta) := \delta^\alpha$, $0 < \alpha \leq 1$.

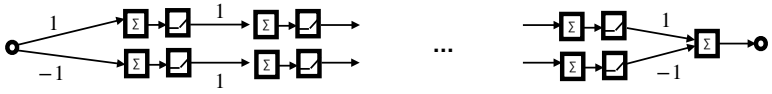

Fig. 1 A value can be passed unchanged along a deep ReLU network with two neurons per layer.

Theorem 1 *Let $\gamma > 0$. For each abstract modulus of continuity ω satisfying*

$$\lim_{\delta \to 0+} \frac{\omega(\delta)}{\delta} = \infty, \tag{7}$$

i.e., $\omega(\delta) = \delta$ is excluded, there exists a function $f_\omega \in C([0,1]^d)$ such that for all $n \in \mathbb{N}$

$$E_n(f_\omega) \le C_1 \omega\left(f_\omega, \frac{1}{\sqrt[d]{n^{2\gamma+2}(1+\ln(n))}}\right) \le C_2 \omega\left(\frac{1}{\sqrt[d]{n^{2\gamma+2}(1+\ln(n))}}\right)$$

(with constants C_1 and C_2 independent of n, cf. (3)) but ($n \to \infty$)

$$E_n(f_\omega) \neq o\left(\omega\left(\frac{1}{\sqrt[d]{n^{2\gamma+2}(1+\ln(n))}}\right)\right).$$

In the case $\gamma = 0$ of bounded width, the same result holds with the $\ln(n)$ expressions replaced by zero.

3 Proof of Theorem 1

Instead of directly dealing with the error functionals E_n, extended networks with additional $2d + 4$ utility neurons per hidden layer are used as shown in Figure 2, i.e., with width between $2d + 4$ and $\tilde{L}(n) + 2d + 4$ and depth bounded by $\tilde{N}(n)$. Some connections from and to these utility neurons have fixed weights -1 or 1 and most biases are set to zero, as shown in the figure. Other weights of the additional connections can be set to zero such that the approximation capability is equal or better than the approximation capability of core network of type (1) with width at most $\tilde{L}(n)$ and depth at most $\tilde{N}(n)$. The connections within the core network are only indicated by the grayed out area in Figure 2.

Let W_n be the corresponding spaces of functions that can be realized by these extended networks and let R_n be the corresponding error of best approximation. Then $V_n \subseteq W_n$ and $R_n(f) \le E_n(f)$ such that error bounds (3) and (4) directly apply for $R_n(f)$.

The ReLU activation function allows a value to be passed unchanged through the network by utilizing two neurons per layer, see Figure 1:

$$x = \sigma(\dots\sigma(\sigma(x))) - \sigma(\dots\sigma(\sigma(-x))). \tag{8}$$

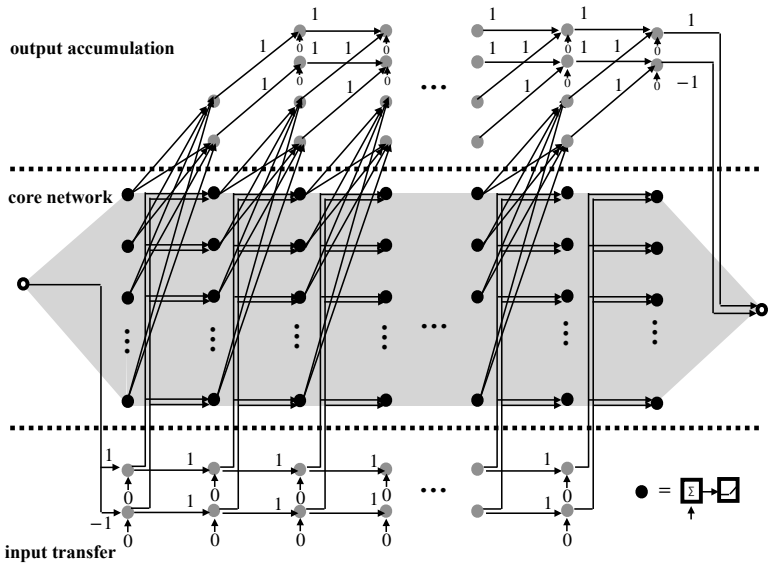

Fig. 2 Network topology for defining function spaces W_n: Each node represents a neuron that adds all input values plus a bias value (if any) and applies the ReLU function to compute its output. One input (i.e., $d = 1$) is passed through the network unchanged with $2d = 2$ neurons per layer, so that the original input is available in all network layers. A weighted output sum of each network layer can be accumulated by four neurons. Before the values are added, they can be split into a difference $a - b$ of two non-negative values a and b which are added separately with the first and the second row of neurons.

This is the idea behind the construction of of spaces W_n. The extended network in Figure 2 consist of $2d$ rows of neurons in the input transfer part, allowing the input values to be forwarded to all hidden layers. The input values are weighted by either 1 or -1 to pass the amount through ReLU activation functions unchanged, cf. (8). Thus, a network can be concatenated from several subnetworks, all of which have access to the original input. Their outputs can be accumulated into a sum with the help of output accumulation neurons that are organized in four rows in Figure 2. Only non-negative values are passed along the first two rows, as they are connected with neurons of the core network via the neurons of the third and fourth rows, which apply ReLU activation.

This allows us to prove the following lemma, which establishes prerequisites for the resonance theorem.

Lemma 1 *The error functionals* $(R_n)_{n=1}^{\infty}$, $R_n : C([0,1]^d) \to [0,\infty)$, *fulfill following conditions for* $m \in \mathbb{N}$, $f, f_1, f_2, \ldots, f_m \in C([0,1]^d)$, *and constants*

$c \in \mathbb{R}$:

$$R_{m \cdot n}\left(\sum_{k=1}^{m} f_k\right) \leq \sum_{k=1}^{m} R_n(f_k), \tag{9}$$

$$R_n(cf) = |c| R_n(f), \tag{10}$$

$$R_n(f) \leq \|f\|_\infty, \tag{11}$$

$$R_n(f) \geq R_{n+1}(f). \tag{12}$$

Proof Condition (10) immediately follows from multiplying or dividing weights and the bias of the output layer by c if $c \neq 0$. When $c = 0$, all weights can be chosen to be zero. By using zero weights and biases, (11) is also established. Since $W_n \subseteq W_{n+1}$ (only upper bounds for depth and width are specified), (12) is obvious. To show (9) for $m > 1$, let $\varepsilon > 0$ and $\Phi_1, \ldots, \Phi_m \in W_n$ be functions realized by extended networks with width at most $\tilde{L}(n) + 2d + 4$ and depth at most $\tilde{N}(n)$ such that $\|f_k - \Phi_k\|_\infty < R_n(f_k) + \varepsilon/m$. Then $\sum_{k=1}^{m} \Phi_k \in W_{m \cdot n}$. To see this, the n extended networks are interpreted as subnetworks that can be concatenated as shown in Figure 3 for two subnetworks. The resulting network has a width of at most $\tilde{N}(n) + 2d + 4 \leq \tilde{N}(n \cdot m) + 2d + 4$ and a depth that is bounded by $m \cdot \tilde{L}(n) = m \cdot (29 + 2d)n = \tilde{L}(n \cdot m)$ so that these parameters fit with $W_{n \cdot m}$.

To provide the first layer of the kth sub-network with weighted input data, as in the stand-alone network-realization of f_k, input transfer neurons are connected to the first layer neurons of this sub-network. The connections are weighted in the same way as the original direct connections from the input node, but because of (8), each weight is used both with a factor of 1 and -1, see Figure 4.

The output of the first $n - 1$ subnetworks is accumulated by means of the neurons of the output accumulation part, again using (8). Instead of connecting the last hidden layers of these subnetworks to an output node, they are connected to the neurons of the third and fourth rows in the output accumulation part, see Figure 5. The connections to the third row are weighted with the original weights of the connections to the output node, and the bias of the output node is applied to the corresponding neuron of the third row. The connections to the fourth row are weighted by the original weights of the connections to the output node, multiplied by -1, and the output node bias times -1 is applied to the corresponding neuron in the fourth row. Thus, the output is represented as a sum $a - b$ of two non-negative values a and b. The value a is added to the accumulator that is represented by the first row, and the value b is added to the sum realized by the second row. Note that the non-negative values passed in the first two rows of neurons of a subnetwork are a superposition of results from other subnetworks and intermediate results of the current subnetwork that become the final result of this subnetwork when the weighted outputs of its last hidden layer and the bias of its output node are added.

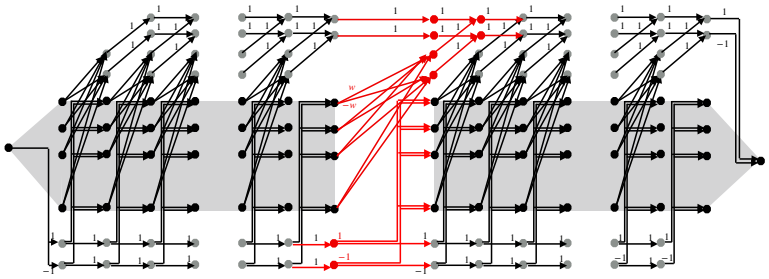

Fig. 3 Subnetworks can be concatenated into larger networks without increasing the width. The image shows a network that is able to add the outputs of two subnetworks computed on the same input value. This can be extended to the concatenation of m subnetworks operating on d input nodes in a straightforward manner.

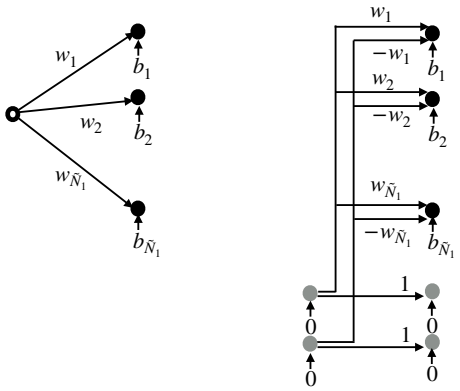

Fig. 4 Each subnetwork can be used with the original input, cf. Figure 3; left: $d = 1$ input neuron is directly connected with first layer neurons; right: instead of a direct connection with the input node, input transfer neurons are connected with neurons of the first layer of a subnetwork.

This construction results in a network that computes $\sum_{k=1}^{m} \Phi_k \in W_{m\cdot n}$. Therefore,

$$R_{m\cdot n}\left(\sum_{k=1}^{m} f_k\right) \leq \left\|\sum_{k=1}^{m} f_k - \sum_{k=1}^{m} \Phi_k\right\|_\infty \leq \sum_{k=1}^{m} \|f_k - \Phi_k\|_\infty$$
$$< \sum_{k=1}^{m} \left(R_n(f_k) + \frac{\varepsilon}{m}\right) = \varepsilon + \sum_{k=1}^{m} R_n(f_k).$$

Since $\varepsilon > 0$ can be chosen arbitrarily, (9) follows. □

An estimate for the VC dimension is applied in order to condense counterexamples. Let V be a set of functions $g : X \to \mathbb{R}$ on a set $X \subseteq \mathbb{R}^d$ and

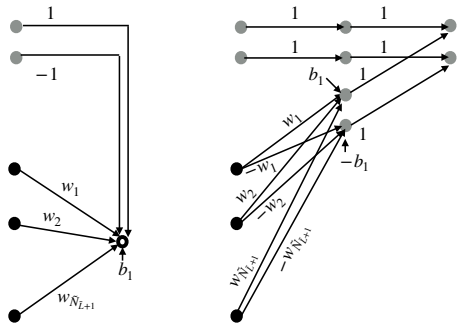

Fig. 5 The output of a subnetwork (left) can be added to the sum of outputs of the preceding subnetworks (right) giving $a - b$ where a is the non-negative output value of the upper right neuron and b is the non-negative output value of the lower right neuron, cf. Figure 3.

$H : \mathbb{R} \to \{0, 1\}$ be the Heaviside-function

$$H(x) := \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Then the VC dimension VC-dim(V) is the largest cardinality of a subset

$$S = \{x_1, \ldots, x_k\} \subseteq X$$

such that for each sign sequence $s_1, \ldots, s_k \in \{-1, 1\}$ a function $g \in V$ can be found that fulfills (cf. [1])

$$H(g(x_i)) = H(s_i), \quad 1 \leq i \leq k.$$

With $X = [0, 1]^d$, a well-known VC dimension estimate for $n \in \mathbb{N}$ is applied, see [12]. For $\gamma > 0$ one gets

$$\begin{aligned}
&\text{VC-dim}(W_n) \\
&\leq C_1 \cdot \min\{(\tilde{N}(n) + 2d + 4)^2 \tilde{L}(n)^2 \ln([\tilde{N}(n) + 2d + 4]\tilde{L}(n)), \\
&\qquad (\tilde{N}(n) + 2d + 4)^3 \tilde{L}(n)^2\} \qquad (13) \\
&\leq C_1 \cdot \left(3^{d+3}(M\lfloor 1 + n^\gamma \rfloor + 2) + 2d + 4\right)^2 (29 + 2d)^2 n^2 \cdot \\
&\qquad \cdot \ln\left((3^{d+3}(M\lfloor 1 + n^\gamma \rfloor + 2) + 2d + 4)(29 + 2d)n\right) \\
&\leq C_2 \cdot n^{2\gamma+2} \cdot \ln(C_3 \cdot n^{\gamma+1}) = C_2 \cdot n^{2\gamma+2} \cdot [\ln(C_3) + (1 + \gamma)\ln(n)] \\
&\leq C_{\text{VC}} \cdot n^{2\gamma+2} \cdot (1 + \ln(n)) \qquad (14)
\end{aligned}$$

with constants independent of n. In the case of bounded width, i.e., $\gamma = 0$, there is

$$\begin{aligned}
&\text{VC-dim}(W_n) \\
&\leq C_1 \cdot (\tilde{N}(n) + 2d + 4)^3 \tilde{L}(n)^2 \leq C_2 \tilde{L}(n)^2 = C_2(29 + 2d)^2 n^2 \leq C_{\text{VC}} n^2. \qquad (15)
\end{aligned}$$

The constant $C_{\mathrm{VC}} \geq 1$ can be chosen such that both (14) and (15) hold.

This upper bound especially holds true if one replaces W_n by a set of functions from W_n that are additionally restricted to a subset of $[0,1]^d$. As subsets, grids are used in the following theorem that is taken from [7, Theorem 5]. It combines a quantitative extension of the uniform boundedness principle with VC dimension estimates.

If one compares inequality (13) with the error estimate (2), then the logarithm $\ln(\tilde{N}(n)\tilde{L}(n))$ differs from $\ln(\tilde{N}(n))$. This is the reason for coupling $\tilde{N}(n)$ and $\tilde{L}(n)$ such that the depth $\tilde{L}(n)$ is bounded or $\tilde{N}(n)$ is, in principle, a power of $\tilde{L}(n)$. In the case of bounded width, one has $(\tilde{N}(n)+2d+4)^3\tilde{L}(n)^2 \in O(\tilde{L}(n)^2)$ such that this problem does not occur.

Theorem 2 (Sharpness due to VC Dimension, [7]) *Let $(F_n)_{n=1}^{\infty}$ be a sequence of (non-linear) function spaces B_n of bounded real-valued functions on $[0,1]^d$ such that (error-)functionals*

$$F_n(f) := \inf\{\|f-g\|_{C([0,1]^d)} : g \in B_n\}$$

fulfill conditions (9)–(12) on the Banach space $C([0,1]^d)$. An equidistant grid $X_n \subseteq [0,1]^d$ with a step size $\frac{1}{\tau(n)}$, $\tau : \mathbb{N} \to \mathbb{N}$, is given via

$$X_n := \left\{\frac{j}{\tau(n)} : j \in \{0,1,\dots,\tau(n)\}\right\} \times \cdots \times \left\{\frac{j}{\tau(n)} : j \in \{0,1,\dots,\tau(n)\}\right\}.$$

Let

$$B_{n,\tau(n)} := \{h : X_n \to \mathbb{R} : \text{ a function } g \in B_n \text{ exists with } h(\mathbf{x}) = g(\mathbf{x}) \text{ for all } \mathbf{x} \in X_n\}$$

be the set of functions that are generated by restricting functions of B_n to this grid. Convergence rates are expressed via a strictly decreasing function $\varphi(x) : [1,\infty) \to (0,\infty)$ with $\lim_{x\to\infty}\varphi(x) = 0$ such that for each $0 < \lambda < 1$ there has to exist a real number $X_0 = X_0(\lambda) \geq \lambda^{-1}$ and a constant $C_\lambda > 0$ such that for all $x > X_0$ there holds

$$\varphi(\lambda x) \leq C_\lambda \varphi(x). \tag{16}$$

Let the VC dimension of $B_{n,\tau(n)}$ and function values of τ and φ be coupled via inequalities

$$\text{VC-dim}(B_{n,\tau(n)}) < \tau(n)^d, \tag{17}$$

$$\tau(4n) \leq \frac{C}{\varphi(n)}, \tag{18}$$

for all $n \geq n_0 \in \mathbb{N}$ with a constant $C > 0$ that is independent of n.

Then, for each abstract modulus of smoothness ω satisfying (6) and (7), there exists a counterexample $f_\omega \in C([0,1]^d)$ such that for $\delta \to 0+$ and $n \to \infty$

$$\omega(f_\omega,\delta) = O\left(\omega(\delta)\right) \text{ and } F_n(f_\omega) \neq o\left(\omega(\varphi(n))\right).$$

In the proof of the theorem, the counterexample f_ω is obtained with a resonance sequence of smooth functions that realize signs at the grid points such that no function of B_n can have the same signs on the grid due to the VC dimension estimate. Then the uniform boundedness principle of [8] condenses the sequence to a single function by utilizing properties (9)–(12). This is proved in the cited paper with the gliding hump method: The counterexample is constructed as an infinite series of scaled functions from the resonance sequence. Then, for some sub-sequence $(n_j)_{j=1}^\infty$ of indices, the error F_{n_j} applied to the jth summand becomes so large (hump) that it dominates the value of F_{n_j} when applied to the remainder of the sum. This results in the desired lower estimate of the error.

Proof (of Theorem 1) Theorem 2 is applied to prove Theorem 1 by choosing function spaces $B_n := W_n$ and functionals $F_n := R_n$ as defined before based on extended networks. Then Lemma 1 shows that conditions (9)–(12) are fulfilled. For $n \in \mathbb{N}$, let

$$\varphi(x) := \begin{cases} \frac{1}{\sqrt[d]{x^{2\gamma+2} \cdot (1+\ln(x))}} & \text{for } \gamma > 0 \\ \frac{1}{\sqrt[d]{x^2}} & \text{for } \gamma = 0, \end{cases}$$

$$\tau(n) := \begin{cases} 2 \cdot \lfloor 2 \sqrt[d]{C_{\mathrm{VC}} n^{2\gamma+2} \cdot (1+\ln(n))} \rfloor & \text{for } \gamma > 0 \\ 2 \cdot \lfloor 2 \sqrt[d]{C_{\mathrm{VC}} n^2} \rfloor & \text{for } \gamma = 0, \end{cases}$$

with the constant C_{VC} from (14) and (15). The function φ ist strictly decreasing with $\lim_{x\to\infty} \varphi(x) = 0$. To prove (16) for $\gamma > 0$, let $0 < \lambda < 1$. Then $\ln(\lambda) < 0$ and for $x > X_0 := \frac{1}{\lambda^2}$ (which is greater than $\frac{1}{\lambda}$), i.e., $\frac{1}{2}\ln(x) > -\ln(\lambda)$, there is $\ln(x) > \frac{1}{2}\ln(x) - \ln(\lambda)$ such that (16) is fulfilled:

$$\begin{aligned} \varphi(\lambda x) &= \frac{1}{\lambda^{\frac{2\gamma+2}{d}} \sqrt[d]{x^{2\gamma+2}(1+\ln(x)+\ln(\lambda))}} < \frac{1}{\lambda^{\frac{2\gamma+2}{d}} \sqrt[d]{x^{2\gamma+2}\left(1+\frac{\ln(x)}{2}\right)}} \\ &\leq \frac{\sqrt[d]{2}}{\lambda^{\frac{2\gamma+2}{d}}} \varphi(x). \end{aligned}$$

In case of bounded width, i.e., $\gamma = 0$, condition (16) also holds true: $\varphi(\lambda x) = \lambda^{-\frac{2}{d}} (\sqrt[d]{x^2})^{-1} = \lambda^{-\frac{2}{d}} \varphi(x)$.

VC dimension bound (17) directly follows from (14) for $\gamma > 0$,

$$\begin{aligned} \text{VC-dim}(W_n) &\leq \left(\sqrt[d]{C_{\mathrm{VC}} \cdot n^{2\gamma+2} \cdot (1+\ln(n))} \right)^d \\ &\leq \left\lfloor 1 + \sqrt[d]{C_{\mathrm{VC}} \cdot n^{2\gamma+2} \cdot (1+\ln(n))} \right\rfloor^d \\ &\leq \left\lfloor 2 \cdot \sqrt[d]{C_{\mathrm{VC}} \cdot n^{2\gamma+2} \cdot (1+\ln(n))} \right\rfloor^d < \tau(n)^d, \end{aligned}$$

and from (15) for $\gamma = 0$:

$$\begin{aligned}\text{VC-dim}(W_n) &\leq \left(\sqrt[d]{C_{\text{VC}} \cdot n^2}\right)^d \leq \left\lfloor 1 + \sqrt[d]{C_{\text{VC}} \cdot n^2} \right\rfloor^d \\ &\leq \left\lfloor 2 \cdot \sqrt[d]{C_{\text{VC}} \cdot n^2} \right\rfloor^d < \tau(n)^d.\end{aligned}$$

It remains to show (18). For $\gamma > 0$ there holds

$$\begin{aligned}\tau(4n) &\leq C_\tau \cdot \sqrt[d]{(4n)^{2\gamma+2} \cdot (1 + \ln(4n))} \\ &\leq C_\tau 4^{\frac{2\gamma+2}{d}} \left(\sqrt[d]{n^{2\gamma+2} \cdot (1 + \ln(4) + \ln(n))} \leq \frac{C_\tau 4^{\frac{2\gamma+2}{d}} \sqrt[d]{1 + \ln(4)}}{\varphi(n)},\right.\end{aligned}$$

whereas the estimate follows for bounded width, i.e., $\gamma = 0$, via

$$\tau(4n) \leq C_\tau \cdot \sqrt[d]{(4n)^2} \leq C_\tau 4^{\frac{2}{d}} \sqrt[d]{n^2} = C_\tau 4^{\frac{2}{d}} \frac{1}{\varphi(n)}.$$

Theorem 2 now provides a counterexample f_ω, for which in case $\gamma > 0$

$$\begin{aligned}&\omega\left(f_\omega, \frac{1}{\sqrt[d]{n^{2\gamma+2}(1 + \ln(n))}}\right) = O\left(\omega\left(\frac{1}{\sqrt[d]{n^{2\gamma+2}(1 + \ln(n))}}\right)\right), \\ &E_n(f_\omega) \geq R_n(f_\omega) \neq o\left(\omega\left(\frac{1}{\sqrt[d]{n^{2\gamma+2}(1 + \ln(n))}}\right)\right),\end{aligned}$$

and for $\gamma = 0$:

$$\omega\left(f_\omega, \frac{1}{\sqrt[d]{n^2}}\right) = O\left(\omega\left(\frac{1}{\sqrt[d]{n^2}}\right)\right), \quad E_n(f_\omega) \geq R_n(f_\omega) \neq o\left(\omega\left(\frac{1}{\sqrt[d]{n^2}}\right)\right).$$

Thus, Theorem 1 follows by considering (3) and (4), respectively. □

4 Conclusions

The given paper has discussed the sharpness of an error bound for the approximation with deep ReLU networks, where the depth grows with a parameter n. If the width grows linearly with n, the depth can even be chosen to be a constant. In this simpler situation (which corresponds to the discussion of single layer networks in [7]), the direct error bound (2) becomes

$$E_n(f) \leq C\omega\left(f, \frac{1}{\sqrt[d]{n^2(1 + \ln(n))}}\right),$$

and the errors of best approximation E_n immediately satisfy the properties in Lemma 1, since for property (9) the m subnetworks can be arranged in parallel to form a network corresponding to parameter $m \cdot n$ such that all subnetworks are connected to the input nodes and the output node. Thus, no additional

utility neurons are required. Using the VC dimension estimate VC-dim$(V_n) \leq Cn^2(1 + \ln(n))$, Theorem 2 proves the existence of a counterexample $f_\omega \in C([0,1]^d)$ for each abstract modulus of continuity ω satisfying (7) such that

$$E_n(f_\omega) \leq C_1\omega\left(f_\omega, \frac{1}{\sqrt[d]{n^2(1+\ln(n))}}\right) \leq C_2\omega\left(\frac{1}{\sqrt[d]{n^2(1+\ln(n))}}\right)$$

but $(n \to \infty)$

$$E_n(f_\omega) \neq o\left(\omega\left(\frac{1}{\sqrt[d]{n^2(1+\ln(n))}}\right)\right).$$

Due to the $\ln(n)$ expression, the error in this case of bounded depth is asymptotically smaller than in the case of bounded width, see (4). This seems to contradict the success of deep neural networks. But the estimates are obtained for networks consisting of fully connected layers. Instead of width and depth, the number of different weights and biases in connection with the role of sparsely connected layers (e.g., convolution and max-pooling layers) could be discussed in future work.

Conflict of Interest Statement

The author states that there is no conflict of interest.

References

1. Bartlett, P.L., Williamson, R.C.: The VC dimension and pseudodimension of two-layer neural networks with discrete inputs. Neural Computation **8**(3), 625–628 (1996)
2. Butzer, P., Scherer, K., Westphal, U.: On the Banach-Steinhaus-theorem and approximation in locally convex spaces. Acta Sci. Math. (Szeged) **34**, 25–34 (1973)
3. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989)
4. Dickmeis, W., Nessel, R.J.: On uniform boundedness principles and Banach-Steinhaus theorems with rates. Numerical Functional Analysis and Optimization **3**(1), 19–52 (1981)
5. Dickmeis, W., Nessel, R.J., van Wickeren, E.: Quantitative extensions of the uniform boundedness principle. Jahresber. Deutsch. Math.-Verein. **89**, 105–134 (1987)
6. Funahashi, K.I.: On the approximate realization of continuous mappings by neural networks. Neural Networks **2**, 183–192 (1989)
7. Goebbels, S.: On sharpness of error bounds for multivariate neural network approximation. Ricerche di Matematica **1827-3491**, 1768–1811 (2020). DOI 10.1007/s11587-020-00549-x
8. Goebbels, S.: On sharpness of error bounds for univariate approximation by single hidden layer feedforward neural networks. Results Math **75**(3) (2020)
9. Johnen, H., Scherer, K.: On the equivalence of the K-functional and moduli of continuity and some applications. In: W. Schempp, K. Zeller (eds.) Constructive Theory of Functions of Several Variables. Proc. Conf. Oberwolfach 1976, pp. 119–140 (1976)
10. Jones, L.K.: Constructive approximations for neural networks by sigmoidal functions. Proceedings of the IEEE **78**(10), 1586–1589, Correction and addition in Proc. IEEE 79 (1991), 243 (1990)

11. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks **6**(6), 861 – 867 (1993)
12. Shen, Z., Yang, H., Zhang, S.: Optimal approximation rate of ReLU networks in terms of width and depth. Journal de Mathématiques Pures et Appliquées (2021). DOI https://doi.org/10.1016/j.matpur.2021.07.009
13. Timan, A.: Theory of Approximation of Functions of a Real Variable. Pergamon Press, New York, NY (1963)