Neural networks: deep, shallow, or in between?

Guergana Petrova and Przemysław Wojtaszczyk

October 31, 2023

Abstract

We give estimates from below for the error of approximation of a compact subset from a Banach space by the outputs of feed-forward neural networks with width W, depth ℓ and Lipschitz activation functions. We show that, modulo logarithmic factors, rates better that entropy numbers' rates are possibly attainable only for neural networks for which the depth $\ell \to \infty$, and that there is no gain if we fix the depth and let the width $W \to \infty$.

1 Introduction

The fascinating new developments in the area of Artificial Intelligence (AI) and other important applications of neural networks prompt the need for a theoretical mathematical study of their potential to reliably approximate complicated objects. Various network architectures have been used in different applications with substantial success rates without significant theoretical backing of the choices made. Thus, a natural question to ask is whether and how the architecture chosen affects the approximation power of the outputs of the resulting neural network.

In this paper, we attempt to clarify how the width and the depth of a feed-forward neural network affect its worst performance. More precisely, we provide estimates from below for the error of approximation of a compact subset $\mathcal{K} \subset X$ of a Banach space X by the outputs of feed-forward neural networks (NNs) with width W, depth ℓ, bound $w(W, \ell)$ on their parameters, and Lipschitz activation functions. Note that the ReLU function is included in our investigation since it is a Lipschitz function with a Lipschitz constant $L = 1$.

To prove our results, we assume that we know lower bounds on the entropy numbers of the compact sets $\mathcal{K}$ that we approximate by the outputs of feed-forward NNs. Such bounds are known for a wide range of classical and novel classes $\mathcal{K}$ and Banach spaces X, and are usually of the form $n^{-\alpha}[\log n]^{\beta}$, $\alpha > 0$, $\beta \in \mathbb{R}$. We refer the reader to [8, Chapters 3,4], [10, Chapter 15],[5, Section 5], [18, Theorem 9], or [6, 9], where such examples are provided.

It is a well known fact that the number n of parameters of a feed-forward NN with width W and depth ℓ is

$$n \asymp \begin{cases} W^2\ell, & \text{when} \quad \ell > 1, \\ W, & \text{when} \quad \ell = 1. \end{cases} \tag{1}$$

Let us denote by $\Sigma(W, \ell, \sigma; w)$ the set of functions that are outputs of a such a NN with bounds $w = w(W, \ell)$ on its parameters and with Lipschitz activation function. We prove estimates from below for the error $E(\mathcal{K}, \Sigma(W, \ell, \sigma; w))_X$ of approximation of a class $\mathcal{K}$ by the functions from $\Sigma(W, \ell, \sigma; w)$, see Theorem 4.1. Our conclusion is that under a moderate growth of the bound $w \asymp n^{\delta}$, $\delta \geq 0$, one can possibly obtain rates of approximation that are better than the corresponding entropy numbers' rates only when the depth of the NN is let to grow. If the

rate of approximation of $\mathcal{K}$ by outputs of feed-forward NNs is better than the decay rate of its entropy numbers, then we say that we have super convergence. In fact, since we only obtain estimates from below, we claim that super convergence is possibly attainable in such cases. If the depth ℓ is fixed, then the rates of decay of $E(\mathcal{K}, \Sigma(W, \ell, \sigma; w))_X$ cannot be better (modulo logarithmic factors) than the rates of the entropy numbers of $\mathcal{K}$. If both the width W and depth ℓ are allowed to grow, then an improvement of the rates of decay of $E(\mathcal{K}, \Sigma(W, \ell, \sigma; w))_X$ in comparison to the entropy numbers' decay is possible. Of course, the bound w on the NN's parameters also has an effect and a fast growing bound, for example $w \asymp 2^n$, could lead to improved convergence in all cases. However, one needs to be aware of the fact that NNs with such bounds are computationally infeasible.

We show that the mapping assigning to each choice of neural network parameters the function that is an output of a feed-forward NN with these parameters is a Lipschitz mapping, see Theorem 3.1. This allows us to study the approximation properties of such NNs via the recently introduced Lipschitz widths, see [14, 15]. We have utilized this approach in [15] to discuss deep ($W = W_0$ is fixed and $\ell \to \infty$) and shallow ($W \to \infty$ and $\ell = 1$) NNs with bounded Lipschitz or ReLU activation functions and their limitations in approximating compact sets $\mathcal{K}$. Here, we implement the developed technique to treat NNs for which both $W, \ell \to \infty$. Results in this direction are available for shallow and deep NNs, and we refer the reader to the series of works [19, 2, 22, 20, 21, 16, 1, 7, 12, 13], where various estimates from below are given for the error of approximation for particular classes $\mathcal{K}$ and Banach spaces X.

The paper is organized as follows. In §2, we introduce our notation, recall the definitions of NNs, entropy numbers and Lipschitz widths, and state some known results about them. We show in §3 that feed-forward NNs are Lipschitz mappings. Finally, in §4, we use results for Lipschitz widths to derive estimates from below for the error of neural network approximation for a compact class $\mathcal{K}$.

2 Preliminaries

In this section, we introduce our notation and recall some known facts about NNs, Lipschitz widths and entropy numbers. In what follows, we will denote by $A \gtrsim B$ the fact that there is an absolute constant $c > 0$ such that $A \geq cB$, where A, B are some expressions that depend on some variable which tends to infinity. Note that the value of c may change from line to line, but is always independent on that variable. Similarly, we use the notation $A \lesssim B$ (defined in an analogues way) and $A \asymp B$ if $A \gtrsim B$ and $A \lesssim B$.

We also write $A = A(B)$ to stress the fact that the quantity A depends on B. For example, if C is a constant, the expression $C = C(d, \sigma)$ means that C depends on d and σ.

2.1 Entropy numbers

We recall, see e.g. [3, 4, 10], that the *entropy numbers* $\epsilon_n(\mathcal{K})_X$, $n \geq 0$, of a compact set $\mathcal{K} \subset X$ are defined as the infimum of all $\epsilon > 0$ for which 2^n balls with centers from X and radius ϵ cover $\mathcal{K}$. Formally, we write

$$\epsilon_n(\mathcal{K})_X = \inf\{\epsilon > 0 \ : \ \mathcal{K} \subset \bigcup_{j=1}^{2^n} B(g_j, \epsilon), \ g_j \in X, \ j = 1, \dots, 2^n\}.$$

2.2 Lipschitz widths

We denote by $(\mathbb{R}^n, \|.\|_{Y_n})$, $n \in \mathbb{N}$, the n-dimensional Banach space with a fixed norm $\|\cdot\|_{Y_n}$, by

$$B_{Y_n}(r) := \{y \in \mathbb{R}^n : \ \|y\|_{Y_n} \leq r\},$$

its ball with radius r, and by

$$\|y\|_{\ell_\infty^n} := \max_{j=1,\dots,n} |y_j|,$$

the ℓ_∞ norm of $y = (y_1, \dots, y_n) \in \mathbb{R}^n$. The Lipschitz widths $d_n^\gamma(\mathcal{K})_X$ of the compact set $\mathcal{K}$ with respect to the norm $\|\cdot\|_X$ is defined as

$$d_n^\gamma(\mathcal{K})_X := \inf_{\mathcal{L}_n,\, r>0,\, \|\cdot\|_{Y_n}} \sup_{f\in\mathcal{K}} \inf_{y\in B_{Y_n}(r)} \|f - \mathcal{L}_n(y)\|_X, \tag{2}$$

where the infimum is taken over all γ/r-Lipschitz maps $\mathcal{L}_n : (B_{Y_n}(r), \|\cdot\|_{Y_n}) \to X$, all $r > 0$, and all norms $\|\cdot\|_{Y_n}$ in $\mathbb{R}^n$. We have proven, see Theorem 9 in [15], the following result which relates the behavior of the entropy numbers of $\mathcal{K}$ and its Lipschitz widths with a Lipschitz constant $\gamma = 2^{\varphi(n)}$.

Theorem 2.1. *For any compact set $\mathcal{K} \subset X$, we consider the Lipschitz width $d_n^{\gamma_n}(\mathcal{K})_X$ with Lipschitz constant $\gamma_n = 2^{\varphi(n)}$, where $\varphi(n) \geq c \log_2 n$ for some fixed constant $c > 0$. Let $\alpha > 0$ and $\beta \in \mathbb{R}$. Then the following holds:*

$$\text{(i) } \epsilon_n(\mathcal{K})_X \gtrsim \frac{(\log_2 n)^\beta}{n^\alpha}, \quad n \in \mathbb{N} \quad \Rightarrow \quad d_n^{\gamma_n}(\mathcal{K})_X \gtrsim \frac{[\log_2(n\varphi(n))]^\beta}{[n\varphi(n)]^\alpha}, \quad n \in \mathbb{N}; \tag{3}$$

$$\text{(ii) } \epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha}, \quad n \in \mathbb{N} \Rightarrow \quad d_n^{\gamma_n}(\mathcal{K})_X \gtrsim [\log_2(n\varphi(n))]^{-\alpha}, \quad n \in \mathbb{N}. \tag{4}$$

2.3 Neural networks

Let us denote by $C(\Omega)$ the set of continuous functions defined on the compact set $\Omega \subset \mathbb{R}^d$, equipped with the uniform norm.

A feed-forward NN with activation function $\sigma : \mathbb{R} \to \mathbb{R}$, width W, depth ℓ and bound $w = w(W, \ell)$ on its parameters generates a family $\Sigma(W, \ell, \sigma; w)$ of continuous functions

$$\Sigma(W, \ell, \sigma; w) := \{\Phi_\sigma^{W,\ell}(y) : \ y \in \mathbb{R}^n\} \subset C(\Omega), \quad \Omega \subset \mathbb{R}^d,$$

where the number of parameters n satisfies (1). Each $y \in \mathbb{R}^n$, $\|y\|_{\ell_\infty^n} \leq w$ determines a continuous function $\Phi_\sigma^{W,\ell}(y) \in \Sigma(W, \ell, \sigma; w)$, defined on Ω, of the form

$$\Phi_\sigma^{W,\ell}(y) := A^{(\ell)} \circ \bar{\sigma} \circ A^{(\ell-1)} \circ \ldots \circ \bar{\sigma} \circ A^{(0)}, \tag{5}$$

where $\bar{\sigma} : \mathbb{R}^W \to \mathbb{R}^W$ is given by

$$\bar{\sigma}(z_1, \dots, z_W) = (\sigma(z_1), \dots, \sigma(z_W)), \tag{6}$$

and $A^{(0)} : \mathbb{R}^d \to \mathbb{R}^W$, $A^{(j)} : \mathbb{R}^W \to \mathbb{R}^W$, $j = 1, \dots, \ell - 1$, and $A^{(\ell)} : \mathbb{R}^W \to \mathbb{R}$ are affine mappings. The coordinates of $y \in \mathbb{R}^n$ are the entries of the matrices and offset vectors (biases) of the affine mappings $A^{(j)}$, $j = 0, \dots, \ell$, taken in a pre-assigned order. The entries of $A^{(j)}$ appear before those of $A^{(j+1)}$ and the ordering for each $A^{(j)}$ is done in the same way. We refer the reader to [7] and the references therein for detailed study of such NNs with fixed width $W = W_0$ and depth $\ell \to \infty$.

We view a feed-forward NN as a mapping that to each vector of parameters $y \in \mathbb{R}^n$ assigns the output $\Phi_\sigma^{W,\ell}(y) \in \Sigma(W, \ell, \sigma; w)$ of this network,

$$y \to \Phi_\sigma^{W,\ell}(y), \tag{7}$$

where all parameters (entries of the matrices and biases) are bounded by $w(W, \ell)$, namely

$$\Sigma(W, \ell, \sigma; w) = \Phi_\sigma^{W,\ell}(B_{\ell_\infty^n}(w(W, \ell))),$$

with $\Phi_\sigma^{W,\ell}$ being defined in (5).

Lower bounds for the error of approximation of a class $\mathcal{K} \subset X$ by the outputs of DNNs (when $W = W_0$ for a fixed W_0 and $\ell \to \infty$, in which $n \asymp \ell$) and SNNs (when $\ell = 1$ and $W \to \infty$, in which $n \asymp W$) have been discussed in [15] in the case of bounded Lipschitz or ReLU activation functions. In this paper, we state similar results for any feed-forward NN with general Lipschitz activation function. We use the approach from [15] and first show that the mapping (7) is a Lipschitz mapping.

3 Feed-forward NNs are Lipshitz mappings

Let us denote by

$$L := \max\{L', |\sigma(0)|\}, \tag{8}$$

where L' is the Lipschitz constant of σ. Then the following theorem is a generalization of Theorems 3 and 5 from [15] to the case of any feed-forward NN.

Theorem 3.1. *Let X be a Banach space such that $C([0,1]^d) \subset X$ is continuously embedded in X. Then the mapping $\Phi_\sigma^{W,\ell} : (B_{\ell_\infty^n}(w(W,\ell)), \|\cdot\|_{\ell_\infty^n}) \to X$, defined in (5) with a Lipschitz function σ, is an L_n-Lipschitz mapping, that is,*

$$\|\Phi_\sigma^{W,\ell}(y) - \Phi_\sigma^{W,\ell}(y')\|_X \le L_n \|y - y'\|_{\ell_\infty^n}, \quad y, y' \in B_{\ell_\infty^n}(w(W,\ell)).$$

Moreover, there are constants $c_1, c_2 > 0$ such that

$$2^{c_1 \ell \log_2(W(w+1)))} < L_n < 2^{c_2 \ell \log_2(W(w+1)))}, \quad w = w(W,\ell),$$

provided $LW \ge 2$.

Proof: Let us first set up the notation $\|g\| := \max\limits_{1 \le i \le W} \|g_i\|_{C(\Omega)}$, where g is the vector function $g = (g_1, \dots, g_W)^T$ whose coordinates $g_i \in C(\Omega)$. We also will use

$$w := w(W,\ell), \quad \text{and} \quad \tilde{w} := w + 1.$$

Let y, y' be the two parameters from $B_{\ell_\infty^n}(w(W,\ell))$ that determine the continuous functions $\Phi_\sigma^{W,\ell}(y)$, $\Phi_\sigma^{W,\ell}(y') \in \Sigma(W,\ell,\sigma;w)$. We fix $x \in \Omega$ and denote by

$$\eta^{(0)}(x) := \overline{\sigma}(A_0 x + b^{(0)}), \quad \eta'^{(0)}(x) := \overline{\sigma}(A_0' x + b'^{(0)}),$$

$$\eta^{(j)} := \overline{\sigma}(A_j \eta^{(j-1)} + b^{(j)}), \quad \eta'^{(j)} := \overline{\sigma}(A_j' \eta'^{(j-1)} + b'^{(j)}), \quad j = 1, \dots, \ell - 1,$$

$$\eta^{(\ell)} := A_\ell \eta^{(\ell-1)} + b^{(\ell)}, \quad \eta'^{(\ell)} := A_\ell' \eta'^{(\ell-1)} + b'^{(\ell)}.$$

Note that $A_0, A_0' \in \mathbb{R}^{W \times d}$, $A_j, A_j' \in \mathbb{R}^{W \times W}$, $b^{(j)}, b'^{(j)} \in \mathbb{R}^W$, for $j = 0, \dots, \ell - 1$, while $A_\ell, A_\ell' \in \mathbb{R}^{1 \times W}$, and $b^{(\ell)}, b'^{(\ell)} \in \mathbb{R}$. Each of the $\eta^{(j)}, \eta'^{(j)}$, $j = 0, \dots, \ell - 1$, is a continuous vector function with W coordinates, while $\eta^{(\ell)}, \eta'^{(\ell)}$ are the outputs of the NN with activation function σ and parameters y, y', respectively.

Since, see (8),

$$|\sigma(t)| \le |\sigma(t) - \sigma(0)| + |\sigma(0)| \le L(|t| + 1), \quad |\sigma(t_1) - \sigma(t_2)| \le L|t_1 - t_2|, \quad t_1, t_2 \in \mathbb{R},$$

it follows that for any m, vectors $\bar{y}, \hat{y}, \eta \in \mathbb{R}^m$ and numbers $y_0, \hat{y}_0 \in \mathbb{R}$, where $\bar{y}, y_0$ and $\hat{y}, \hat{y}_0$ are subsets of the coordinates of $y, y' \in \mathbb{R}^n$, respectively, we have

$$\begin{aligned} |\sigma(\bar{y} \cdot \eta + y_0)| &\le L(|\bar{y} \cdot \eta + y_0| + 1) \le L(m\|\eta\|_{\ell_\infty^m} + 1)\|y\|_{\ell_\infty^n} + L \\ &\le L(m\|\eta\|_{\ell_\infty^m} + 1)w + L < L\tilde{w}m\|\eta\|_{\ell_\infty^m} + L\tilde{w} \end{aligned} \tag{9}$$

and

$$|\sigma(\bar{y}\cdot\eta+y_0)-\sigma(\hat{y}\cdot\eta+\hat{y}_0)|\le L(m\|\eta\|_{\ell_\infty^m}+1)\|y-y'\|_{\ell_\infty^n}. \tag{10}$$

Then we have $\|\eta'^{(0)}\| < L\tilde{w}d + L\tilde{w}$ (when $m = d$ and $\eta = x$) and

$$\|\eta'^{(j)}\| < LW\tilde{w}\|\eta'^{(j-1)}\| + L\tilde{w}, \quad j = 1,\dots,\ell,$$

(when $m = W$ and $\eta = \eta'^{(j-1)}$). One can show by induction that for $j = 1,\dots,\ell$,

$$\|\eta'^{(j)}\| \le dW^j[L\tilde{w}]^{j+1} + L\tilde{w}\sum_{i=0}^{j}[LW\tilde{w}]^i.$$

Therefore, we have that

$$\|\eta'^{(j)}\| \le dW^j[L\tilde{w}]^{j+1} + 2L\tilde{w}[LW\tilde{w}]^j = (d+2)L\tilde{w}[LW\tilde{w}]^j, \tag{11}$$

since $LW\tilde{w} > LW \ge 2$. The above inequality also holds for $j = 0$.

Clearly, we have

$$\|\eta^{(0)} - \eta'^{(0)}\| \le L(d+1)\|y-y'\|_{\ell_\infty^n} =: C_0\|y-y'\|_{\ell_\infty^n}.$$

Suppose we have proved the inequality

$$\|\eta^{(j-1)} - \eta'^{(j-1)}\| \le C_{j-1}\|y-y'\|_{\ell_\infty^n},$$

for some constant C_{j-1}. Then we derive that

$$\begin{aligned}
\|\eta^{(j)} - \eta'^{(j)}\| &\le L\|A_j\eta^{(j-1)} + b^{(j)} - A'_j\eta'^{(j-1)} - b'^{(j)}\| \\
&\le L\|A_j(\eta^{(j-1)} - \eta'^{(j-1)})\| + L\|(A_j - A'_j)\eta'^{(j-1)}\| + L\|b^{(j)} - b'^{(j)}\| \\
&\le LW\|y\|_{\ell_\infty^n}\|\eta^{(j-1)} - \eta'^{(j-1)}\| + LW\|y-y'\|_{\ell_\infty^n}\|\eta'^{(j-1)}\| + L\|y-y'\|_{\ell_\infty^n} \\
&\le (LW\tilde{w}C_{j-1} + LW(d+2)L\tilde{w}[LW\tilde{w}]^{j-1} + L)\|y-y'\|_{\ell_\infty^n} \\
&= L(W\tilde{w}C_{j-1} + (d+2)[LW\tilde{w}]^j + 1)\|y-y'\|_{\ell_\infty^n} \\
&=: C_j\|y-y'\|_{\ell_\infty^n},
\end{aligned}$$

where we have used that $\|y\|_{\ell_\infty^{\tilde{n}}} \le w$, the bound (11), and the induction hypothesis. The relation between C_j and C_{j-1} can be written as

$$C_0 = L(d+1), \quad C_j = L(W\tilde{w}C_{j-1} + (d+2)[LW\tilde{w}]^j + 1), \quad j = 1,\dots,\ell.$$

Clearly,

$$C_1 = L((d+1)LW\tilde{w} + (d+2)LW\tilde{w} + 1) < (d+2)L(2LW\tilde{w} + 1),$$

and we obtain by induction that

$$C_\ell \;<\; (d+2)L\left(\ell[LW\tilde{w}]^\ell + \sum_{i=0}^{\ell}[LW\tilde{w}]^i\right).$$

If we use the fact $2 \le LW < LW\tilde{w}$, we derive the inequality

$$C_\ell < (d+2)L(\ell+2)[LW\tilde{w}]^\ell.$$

Finally, we have

$$\begin{aligned}\|\Phi_\sigma^{W,\ell}(y)-\Phi_\sigma^{W,\ell}(y')\|_{C(\Omega)} &= \|\eta^{(\ell)}-\eta'^{(\ell)}\|\le C_\ell\|y-y'\|_{\ell_\infty^n}\\ &< (d+2)L(\ell+2)[LW\tilde{w}]^\ell\|y-y'\|_{\ell_\infty^n},\end{aligned}$$

and therefore

$$\|\Phi_\sigma^{W,\ell}(y)-\Phi_\sigma^{W,\ell}(y')\|_X\le c_0\|\Phi_\sigma^{W,\ell}(y)-\Phi_\sigma^{W,\ell}(y')\|_{C(\Omega)}\le\tilde{C}\ell[LW\tilde{w}]^\ell\|y-y'\|_{\ell_\infty^n},$$

where $\tilde{C}=\tilde{C}(d,\sigma)$. Clearly, the Lipschitz constant $L_n:=\tilde{C}\ell[LW\tilde{w}]^\ell$ is such that $2^{c_1\ell\log_2(W(w+1))}<L_n<2^{c_2\ell\log_2(W(w+1))}$ for some $c_1,c_2>0$, and the proof is completed. □

Remark 3.2. *Note that the proof of Theorem* 3.1 *holds also in the case when every coordinate of $\bar{\sigma}$, see (6), is chosen to be a different Lipschitz function σ as long as $LW\ge 2$, where L is defined via (8).*

4 Estimates from below for neural network approximation

In this section, we consider Banach spaces X such that $C([0,1]^d)$ is continuously embedded in X. Let us denote by

$$E(f,\Sigma(W,\ell,\sigma;w))_X:=\inf_{y\in B_{\ell_\infty}^n(w)}\|f-\Phi_\sigma^{W,\ell}(y)\|_X,$$

the error of approximation in the norm $\|\cdot\|_X$ of the element $f\in\mathcal{K}$ by the set of outputs $\Sigma(W,\ell,\sigma;w)$ of a feed-forward NN with width W, depth ℓ, activation function σ, and a bound w on its parameters y, that is $\|y\|_{\ell_\infty^n}\le w$. We also denote by

$$E(\mathcal{K},\Sigma(W,\ell,\sigma;w))_X:=\sup_{f\in\mathcal{K}}\ E(f,\Sigma(W,\ell,\sigma;w))_X,$$

the error for the class $\mathcal{K}\subset X$. It follows from Theorem 3.1 that

$$E(\mathcal{K},\Sigma(W,\ell,\sigma;w))_X\ge d_n^{\gamma_n}(\mathcal{K})_X,\quad\text{with}\quad\gamma_n=2^{c\ell\log_2(W(w+1))}=:2^{\varphi(n)},\tag{12}$$

for some $c>0$. Therefore, see (1),

$$n\varphi(n)=\begin{cases}cn\ell\log_2(W(w+1)), & n\asymp W^2\ell,\quad \ell>1,\\ cn\log_2(n(w+1)), & n\asymp W,\quad \ell=1,\end{cases}$$

and we can state the following corollary of (12) and Theorem 2.1.

Theorem 4.1. *Let $\Sigma(W,\ell,\sigma;w)$ be the set of outputs of an n parameter NN with width W, depth ℓ, Lipschitz activation function σ and weights and biases bounded by w, where $LW\ge 2$. Then, the error of approximation $E(\mathcal{K},\Sigma(W,\ell,\sigma;w))_X$ of a compact subset $\mathcal{K}$ of a Banach space X by $\Sigma(W,\ell,\sigma;w)$ satisfies the following estimates from below, provided we know the following information about the entropy numbers $\epsilon_n(\mathcal{K})_X$ of $\mathcal{K}$:*

- *if for $\alpha>0$ and $\beta\in\mathbb{R}$ we have*

$$\epsilon_n(\mathcal{K})_X\gtrsim\frac{[\log_2 n]^\beta}{n^\alpha},\ n\in\mathbb{N},$$

then

$$E(\mathcal{K},\Sigma(W,\ell,\sigma;w))_X\gtrsim\begin{cases}\frac{1}{n^\alpha\ell^\alpha}\cdot\frac{[\log_2(n\ell\log_2(W(w+1)))]^\beta}{[\log_2(W(w+1))]^\alpha}, & n\asymp W^2\ell,\quad \ell>1,\\ \frac{1}{n^\alpha}\cdot\frac{[\log_2(n\log_2(nw))]^\beta}{[\log_2(n(w+1))]^\alpha}, & n\asymp W,\quad \ell=1.\end{cases}$$

- *if for $\alpha > 0$ we have*

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N},$$

then

$$E(\mathcal{K}, \Sigma(W,\ell,\sigma;w))_X \gtrsim \begin{cases} [\log_2(n\ell\log_2(W(w+1)))]^{-\alpha}, & n \asymp W^2\ell, \quad \ell > 1, \\ [\log_2(n\log_2(n(w+1)))]^{-\alpha}, & n \asymp W, \quad \ell = 1. \end{cases}$$

Proof: The proof follows directly from (12) and Theorem 2.1. □

Remark 4.2. *Theorem 4.1 gives various estimates from below depending on the behavior of the bound $w = w(W,\ell)$ on the absolute values of the parameters of the NN. Here we state only one particular case. Under the conditions of Theorem 4.1 with $w = w(W,\ell) = \text{const}$, we have:*

- *if for $\alpha > 0$ and $\beta \in \mathbb{R}$ we have*

$$\epsilon_n(\mathcal{K})_X \gtrsim \frac{[\log_2 n]^\beta}{n^\alpha},\ n \in \mathbb{N},$$

then

$$E(\mathcal{K}, \Sigma(W,\ell,\sigma;w))_X \gtrsim \begin{cases} \frac{1}{n^\alpha \ell^\alpha} \cdot \frac{[\log_2(n\ell\log_2 W)]^\beta}{[\log_2 W]^\alpha}, & n \asymp W^2\ell, \quad \ell > 1, \\ \frac{1}{n^\alpha} \cdot [\log_2 n]^{\beta-\alpha}, & n \asymp W, \quad \ell = 1. \end{cases}$$

- *if for $\alpha > 0$ we have*

$$\epsilon_n(\mathcal{K})_X \gtrsim [\log_2 n]^{-\alpha},\ n \in \mathbb{N},$$

then

$$E(\mathcal{K}, \Sigma(W,\ell,\sigma;w))_X \gtrsim \begin{cases} [\log_2(n\ell\log_2 W)]^{-\alpha}, & n \asymp W^2\ell, \quad \ell > 1, \\ [\log_2 n]^{-\alpha}, & n \asymp W, \quad \ell = 1. \end{cases}$$

Acknowledgments: G.P. was supported by the NSF Grant DMS 2134077 and ONR Contract N00014-20-1-278.

References

[1] Achour E-M., Foucault A., Gerchinovitz S., Malgouyres F. (2022). A general approximation lower bound in L_p norm, with applications to feed-forward neural networks. arXiv:2206.04360.

[2] Bartlett P., Harvey N., Liaw C., Mehrabian A. (2019). Nearly-tight vc-dimension and pseudo dimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1), 2285–2301.

[3] Carl B. (1981). Entropy numbers, s-numbers, and eigenvalue problems. *J. Funct. Anal.*, 41, 290–306.

[4] Carl B., Stephani I. (1990).*Entropy, compactness and the approximation of operators.* Cambridge University Press.

[5] Cobos F. O. Dominguez and T. Kuhn (2018). Approximation and entropy numbers of embeddings between approximation spaces. *Constructive Approximation*, 47, 453–486.

[6] Cobos F., Kuhn T. (2009). Approximation and entropy numbers in Besov spaces of generalized smoothness. *J. Approx. Theory*, 160, 56–70.

[7] DeVore R., Hanin B., Petrova G. (2021). Neural Network Approximation. *Acta Numerica*, 30, 327–444.

[8] Edmunds D., Triebel H. (1996). *Function spaces, Entropy numbers and differenial operators.* Cambridge Tracts in Mathematics 120.

[9] Gao F. (2008). Entropy estimate for k-monotone functions via small ball probability of integrated Brownian motion. *Elect. Comm. in Probab.*, 13, 121–130.

[10] Lorentz G., Golitschek M., Makovoz Y. (1996). *Constructive Approximation.* Springer Verlag.

[11] Lu J., Shen Z., Yang H., Zhang S. (2020). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5), 5465–5506.

[12] Maiorov V. (1999). On best approximation by ridge functions. *J. Approx. Theory*, 99(1), 68–94.

[13] Maiorov V., Meir R., Ratsaby J. (1999). On the approximation of functional classes equipped with a uniform measure using ridge functions. *J. Approx. Theory*, 99, 95–111.

[14] Petrova G., Wojtaszczyk P. (2023). Lipschitz widths. *Constructive Approximation*, 7, 759–805.

[15] Petrova G., Wojtaszczyk P. (2022). Limitations on approximation by deep and shallow neural networks. arXiv:2212.02223v1.

[16] Shen Z., Yang H., Zhang S. (2022). Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathematiques Pures et Appliquees*,

[17] Siegel J. (2022). Optimal Approximation Rates for Deep ReLU Neural Networks on Sobolev and Besov Spaces. arXiv:2211.14400.

157, 101–135.

[18] Siegel J., Xu J. (2022). Sharp bounds on the approximation rates, metric entropy and n widths of shallow neural networks. *Journal of FOCM*. arXiv:2101.12365v9.

[19] Yang Y., Barron A. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5), 1564–1599.

[20] Yarotsky D. (2017). Error bounds for approximations with deep relu networks. *Neural networks*, 97, 103–114.

[21] Yarotsky D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. *Proceedings of the 31st Conference On Learning Theory*, PMLR, 75, 639–649.

[22] Yarotsky D., and Zhevnerchuk A. (2020). The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33, 13005–13015.

Affiliations:
Guergana Petrova, Department of Mathematics, Texas A&M University, College Station, TX 77843, gpetrova@math.tamu.edu.

Przemysław Wojtaszczyk, Institut of Mathematics, Polish Academy of Sciences, ul. Śniadeckich 8, 00-656 Warszawa, Poland, wojtaszczyk@impan.pl