Approximation in $L^p(\mu)$ with deep ReLU neural networks

Felix VOIGTLAENDER
Katholische Universität Eichstätt–Ingolstadt
Ostenstraße 26, 85072 Eichstätt, Germany
felix@voigtlaender.xyz

Philipp PETERSEN
Mathematical Institute, University of Oxford
OX2 6GG, Oxford, UK
pc.petersen.pp@gmail.com

***Abstract*—We discuss the expressive power of neural networks which use the non-smooth ReLU activation function $\varrho(x) = \max\{0, x\}$ by analyzing the approximation theoretic properties of such networks. The existing results mainly fall into two categories: approximation using ReLU networks with a *fixed* depth, or using ReLU networks whose depth *increases* with the approximation accuracy. After reviewing these findings, we show that the results concerning networks with fixed depth—which up to now only consider approximation in $L^p(\lambda)$ for the Lebesgue measure λ—can be generalized to approximation in $L^p(\mu)$, for any finite Borel measure μ. In particular, the generalized results apply in the usual setting of statistical learning theory, where one is interested in approximation in $L^2(\mathbb{P})$, with the probability measure $\mathbb{P}$ describing the distribution of the data.**

I. INTRODUCTION

In recent years, machine learning techniques based on deep neural networks have significantly advanced the state of the art in applications like image classification, speech recognition, and machine translation. The networks used for such applications tend to use the non-smooth ReLU activation function $\varrho(x) = \max\{0, x\}$, since it is empirically observed to improve the training procedure [4].

In this paper, we focus on the *expressive power* of such neural networks. Precisely, given a function class $\mathcal{F}$ and an approximation accuracy $\varepsilon > 0$, we aim to find a *complexity bound* $N = N(\mathcal{F}, \varepsilon)$ such that for any $f \in \mathcal{F}$, one can find a ReLU network Φ_ε^f of complexity at most N satisfying $\|f - \Phi_\varepsilon^f\| \leq \varepsilon$. Here, the *complexity* of the network is measured in terms of its *depth* (the number of layers) and in terms of the number of *neurons* and *weights*. The approximation error will be either measured in the uniform norm or in $L^p(\mu)$ for some measure μ. When we simply write L^p, it is understood that $\mu = \lambda$ is taken to be the Lebesgue measure.

Structure of the paper: We start by reviewing existing results which provide complexity bounds $N(\mathcal{F}, \varepsilon)$ for approximating functions from the class $\mathcal{F} = \mathcal{F}_{d,\beta,B}$ of all C^β functions f on $Q := Q_d := [-\frac{1}{2}, \frac{1}{2}]^d$ that satisfy $\|f\|_{C^\beta} \leq B$. These results fall into two categories: The first considers approximation in L^p using ReLU networks of *fixed* depth, while the second considers *uniform* approximation using networks of *increasing* depth. We also present a novel result, showing that the complexity bounds of the first category also apply for approximation in $L^p(\mu)$; see Theorem II.3.

Note that if $N(\mathcal{F}, \varepsilon)$ is a valid complexity bound, then so is any $N'(\mathcal{F}, \varepsilon) \geq N(\mathcal{F}, \varepsilon)$. Therefore, after reviewing the existing complexity bounds, we also discuss their *optimality*.

In the final section of the paper, we prove Theorem II.3.

II. APPROXIMATION RESULTS USING RELU NETWORKS

In this section, we review the existing findings concerning the approximation properties of ReLU networks. In doing so, we first focus on approximation using ReLU networks with a fixed depth, and then see what changes when the depth of the networks is allowed to grow with the approximation accuracy.

First of all, however, we formally define neural networks and discuss how to measure their complexity. Here and in the remainder of the paper, we write $\underline{m} := \{1, \dots, m\}$.

Definition II.1. A *neural network* Φ with $L = L(\Phi) \in \mathbb{N}$ layers, input dimension $d \in \mathbb{N}$ and output dimension $k \in \mathbb{N}$ is a tuple $\Phi = \big((A_1, b_1), \dots, (A_L, b_L)\big)$, where $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell \in \underline{L}$ and where $N_0 = d$ and $N_L = k$.

Given $\varrho : \mathbb{R} \to \mathbb{R}$ (called the *activation function*), the *ϱ-realization* of Φ is the function $\mathrm{R}_\varrho(\Phi) : \mathbb{R}^d \to \mathbb{R}^k, x \mapsto x_L$, where $x_0 := x \in \mathbb{R}^d$ and $x_L := A_L\, x_{L-1} + b_L \in \mathbb{R}^k$, while

$$x_\ell := \varrho(A_\ell\, x_{\ell-1} + b_\ell) \in \mathbb{R}^{N_\ell} \quad \text{for} \quad \ell \in \underline{L-1},$$

where $\varrho(y) = \big(\varrho(y_1), \dots, \varrho(y_n)\big)$ for $y = (y_1, \dots, y_n) \in \mathbb{R}^n$.

The *number of neurons* of Φ is $N(\Phi) := \sum_{\ell=0}^{L} N_\ell \in \mathbb{N}$, while the *number of (nonzero) weights* of Φ is given by $W(\Phi) := \sum_{i=1}^{L} \big(\|A_i\|_{\ell^0} + \|b_i\|_{\ell^0}\big)$, with $\|A\|_{\ell^0}$ denoting the number of nonzero entries of a matrix or vector A.

Given $\Omega \subset \mathbb{R}$, we say that *all weights of Φ belong to Ω* if all entries of the matrices $A_1, \dots, A_L$ and the vectors $b_1, \dots, b_L$ belong to Ω. Given $s \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2})$, we say that the network Φ is *(s, ε)-quantized*, if all weights of Φ belong to the set $[-\varepsilon^{-s}, \varepsilon^{-s}] \cap 2^{-s\lceil \log_2(1/\varepsilon)\rceil}\mathbb{Z}$.

Weight-quantization is a further notion of complexity, which—when combined with bounds on the number of network weights—restricts the number of bits needed to encode the network.

In the remainder of the paper, we will only consider the ReLU activation function $\varrho : \mathbb{R} \to \mathbb{R}, x \mapsto \max\{0, x\}$.

A. L^p approximation using fixed-depth networks

The following is the main existing result concerning approximation of C^β functions using *fixed-depth* ReLU networks.

Theorem II.2. *([5, Theorem A.9]) Let $\beta, B, p \in (0,\infty)$, $d \in \mathbb{N}$, and $Q := [-\frac{1}{2},\frac{1}{2}]^d$. There are $C > 0$ and $s \in \mathbb{N}$ (depending on d, β, B, p) such that for $\varepsilon \in (0, \frac{1}{2})$ and $f \in C^\beta(Q)$ with $\|f\|_{C^\beta} \leq B$, there is an (s,ε)-quantized network Φ_ε^f with $L(\Phi_\varepsilon^f) \leq 11 + (1 + \lceil \log_2 \beta \rceil)(11 + \frac{\beta}{d})$, and with $\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{L^p} \leq C\varepsilon$ and $N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \leq C\,\varepsilon^{-d/\beta}$.*

In Section IV, we will prove the following generalization:

Theorem II.3. *Let $d \in \mathbb{N}$ and $\beta, B, p \in (0,\infty)$, and let μ be a finite Borel measure on $Q := [-\frac{1}{2},\frac{1}{2}]^d$. There are $C > 0$ and $s \in \mathbb{N}$ (depending on d, p, β, B, μ) such that for $\varepsilon \in (0, \frac{1}{2})$ and $f \in C^\beta(Q)$ with $\|f\|_{C^\beta} \leq B$, there is an (s,ε)-quantized network Φ_ε^f with $L(\Phi_\varepsilon^f) \leq 7 + (1 + \lceil \log_2 \beta \rceil)(11 + \frac{\beta}{d})$, and*

$$\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{L^p(\mu)} \leq C\varepsilon \;\text{ and }\; N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \leq C\,\varepsilon^{-d/\beta}.$$

The optimality of the complexity bound $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/\beta}$ will be discussed in detail in Section III. A related question concerns the optimality of the *depth* of the networks. The next result shows that—up to logarithmic factors—the depth of the networks in Theorems II.2 and II.3 is indeed optimal.

Theorem II.4. *([5, Theorem C.6]; see [7] for the case $p = 2$)*
Let $\varnothing \neq \Omega \subset \mathbb{R}^d$ be open, bounded, and connected. Let $f \in C^3(\Omega)$ be nonlinear and $p \in (0,\infty)$. There is $C_{f,p,d} > 0$ such that for any neural network Φ of depth $L(\Phi)$, we have

$$\|f - \mathrm{R}_\varrho(\Phi)\|_{L^p} \geq C_{f,p,d} \cdot \big(1 + \min\{N(\Phi), W(\Phi)\}\big)^{-2L(\Phi)}.$$

Thus, to attain $\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{L^p} \lesssim \varepsilon$ subject to the complexity bound $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/\beta}$, the networks Φ_ε^f must satisfy $L(\Phi_\varepsilon^f) \geq \beta/2d$, at least for $\varepsilon > 0$ small enough.

In a nutshell, these results show that *ReLU networks achieve better approximation rates for smoother functions. To attain these better rates, however, one has to use deeper networks.*

Further results: One can also derive L^p approximation rates for a certain class of *discontinuous* functions; see [5]. Further, the presented results for fully connected networks are equivalent to approximation results for certain simplified *convolutional* networks [6] that do not employ pooling operations.

We close our tour of approximation results using fixed depth networks with the following result.

Proposition II.5. *(see [9, Proposition 1]) Let $d \in \mathbb{N}$ and $\beta \in (0,1]$, and $Q := [-\frac{1}{2},\frac{1}{2}]^d$. There is $C = C(d,\beta) > 0$ such that for each $f \in C^\beta(Q)$ and $\varepsilon \in (0, \frac{1}{2})$, there is a neural network Φ_ε^f with $L = L(d)$ layers such that*

$$\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{\sup} \leq \varepsilon\,\|f\|_{C^\beta} \;\text{ and }\; N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \leq \tfrac{C}{\varepsilon^{d/\beta}}.$$

Though not explicitly stated in [9], one can show that the same statement holds for certain (s,ε)-quantized networks Φ_ε^f, where $s = s(d,B)$ and $\|f\|_{C^\beta} \leq B$.

Such a uniform approximation is much stronger than approximation in $L^p(\mu)$. It should be noted, however, that the above result only applies for the "low smoothness, slow approximation regime" $\beta \in (0,1]$ where a piecewise affine approximation yields an optimal error. For $\beta > 1$ it is an open problem whether the bounds in Theorems II.2 and II.3 also hold for *uniform* approximation using fixed-depth networks.

B. Uniform approximation using networks of growing depth

While it is open whether fixed-depth networks can satisfy $\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{\sup} \lesssim \varepsilon$ and $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/\beta}$ for $f \in C^\beta$ and $\beta > 1$, this *is* possible with a mild depth-growth as $\varepsilon \to 0$.

Theorem II.6. *(see [8, Theorem 1]) Let $d, k \in \mathbb{N}$ and set $Q := [-\frac{1}{2},\frac{1}{2}]^d$. There is $C = C(d,k) > 0$ such that for any $\varepsilon \in (0, \frac{1}{2})$ and $f \in C^k(Q)$ with $\|f\|_{C^k} \leq 1$, there is a network Φ_ε^f satisfying $N(\Phi_\varepsilon^f) \lesssim W(\Phi_\varepsilon^f) \leq C\big(1 + \ln(1/\varepsilon)\big)\,\varepsilon^{-d/k}$ and $\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{\sup} \leq \varepsilon$, as well as $L(\Phi_\varepsilon^f) \leq C\big(1 + \ln(1/\varepsilon)\big)$.*

Although not stated explicitly in [8, Theorem 1], the proof shows that Φ_ε^f can be chosen to be (s,ε)-quantized for some $s = s(d,k) \in \mathbb{N}$.

We close this section with a surprising result from [9]. In that paper, Yarotsky shows that if one does *not* restrict the growth of the depth as $\varepsilon \to 0$, and if one does *not* insist that the networks be quantized, then one can *significantly* beat the approximation rates stated in Theorems II.6, II.2, and II.3—at least in the "low smoothness" regime $\beta \in (0,1]$:

Theorem II.7. *(see [9, Theorem 2])*
Let $d \in \mathbb{N}$, $\beta \in (0,1]$, and $Q := [-\frac{1}{2},\frac{1}{2}]^d$. There is $C = C(d,\beta) > 0$ such that for any $\varepsilon \in (0, \frac{1}{2})$ and $f \in C^\beta(Q)$ with $\|f\|_{C^\beta} \leq 1$ there is a network Φ_ε^f satisfying $\|f - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{\sup} \leq \varepsilon$ and $W(\Phi_\varepsilon^f) \leq C\,\varepsilon^{-d/(2\beta)}$.

Note that Theorem II.6 only yields $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/\beta}$ instead of $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/(2\beta)}$. Also note that Theorem II.7 does *not* claim that the networks can be chosen to be quantized. In fact, this is *impossible*, as shown in the next section.

III. Optimality of the approximation results

Assuming that the complexity of the individual weights of the network does not grow too quickly as $\varepsilon \downarrow 0$, the complexity bound $W(\Phi_\varepsilon^f) \lesssim \varepsilon^{-d/\beta}$ derived in Theorems II.2 and II.6 is optimal up to a log-factor. In fact, the same arguments as in the proof of [5, Theorem 4.3] show the following:

Proposition III.1. *Let $d, s \in \mathbb{N}$ and $\beta, B, p \in (0,\infty)$ and let $Q := [-\frac{1}{2},\frac{1}{2}]^d$. Then there is a function $f \in C^\beta(Q)$ with $\|f\|_{C^\beta} \leq B$ and a null-sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ such that*

$$\inf\left\{W(\Phi) \;:\; \begin{array}{l}\Phi \text{ an } (s,\varepsilon_k)\text{-quantized neural netw.}\\ \text{and } \|f - \mathrm{R}_\varrho(\Phi)\|_{L^p} \leq \varepsilon_k\end{array}\right\}$$
$$\geq \varepsilon_k^{-d/\beta} \big/ \big(\log(1/\varepsilon_k) \cdot \log(\log(1/\varepsilon_k))\big) \quad \forall\, k \in \mathbb{N}\,.$$

In particular, while Theorem II.3 is not optimal for *every* measure μ, it is optimal for the Lebesgue measure $\mu = \lambda$.

Proposition III.1 also shows that the networks in Theorem II.7 *can not be chosen to be (s,ε)-quantized*.

The proof of Proposition III.1 is *information-theoretic*: On the one hand, [5, Lemma B.4] shows that there are at most $2^{\mathcal{O}(\varepsilon^{-\theta}\log_2(1/\varepsilon))}$ many different realizations of (s,ε)-quantized ReLU networks that have $c\,\varepsilon^{-\theta}$ nonzero weights. On the other hand, [3] yields lower bounds for the cardinality of families that are ε-dense (with respect to the L^p norm) in the set of all C^β-functions f such that $\|f\|_{C^\beta} \leq B$.

Finally, using bounds for the VC-dimension of neural networks (see [1]), Yarotsky showed that the approximation rates derived in Theorem II.7 are optimal; see [9, Theorem 1].

In summary, we note the following:

- For quantized networks, the rates in Theorems II.2 and II.6 are optimal.
- For fixed-depth networks, the rates *for uniform approximation* in Theorem II.6 cannot be improved, even *without* assuming quantized networks; see [8, Part 2 of Theorem 4]. Note, however, that it is open whether these rates *can be attained at all* using fixed-depth networks if $\beta > 1$.
- Since the VC dimension arguments used for proving [8, Theorem 4] do not seem to generalize to L^p-approximation, it is open whether the rates in Theorem II.2 are optimal for bounded-depth networks *without* assuming quantized networks.
- If one neither assumes bounded depth nor quantized networks, then the results in Theorems II.2 and II.5 can be improved, at least for $\beta \in (0,2)$. The optimal rates for this setting and for $\beta \in (0,1]$ are given by Theorem II.7. It is open what the optimal rates for $\beta > 1$ are if one neither assumes bounded depth nor quantized networks.

IV. Proof of Theorem II.3

We begin with the following lemma which shows that ReLU networks can (approximately) localize a function to a cube.

Lemma IV.1. *(modification of [5, Lemma A.6])*
For $a, b \in \mathbb{R}^d$ and $0 < \varepsilon < \min_{i \in \underline{d}} \frac{1}{2}(b_i - a_i)$, set

$$[a,b) := \prod_{i=1}^d [a_i, b_i) \quad \text{and} \quad [a,b]_\varepsilon := \prod_{i=1}^d [a_i + \varepsilon, b_i - \varepsilon].$$

If $B \geq 1$, there is a 4-layer network $\Lambda_{\varepsilon,B}^{(a,b)}$ with 1-dimensional output and $(d+1)$-dimensional input, with $W(\Lambda_{\varepsilon,B}^{(a,b)}) \leq c(d)$ weights, all of which have their absolute values bounded by $d + B + \varepsilon^{-1} \cdot (1 + \|a\|_{\ell^\infty} + \|b\|_{\ell^\infty})$, and such that if $x \in \mathbb{R}^d$ and $y \in [-B, B]$, then

$$\big| \mathrm{R}_\varrho\big(\Lambda_{\varepsilon,B}^{(a,b)}\big)(x,y) - y\, \mathbb{1}_{[a,b)}(x) \big| \leq 2B\, \mathbb{1}_{[a,b) \setminus [a,b]_\varepsilon}(x). \quad \text{(IV.1)}$$

Proof. For $i \in \underline{d}$, define a function $t_i : \mathbb{R} \to \mathbb{R}$ by setting $t_i(x) = \varrho(\frac{x-a_i}{\varepsilon}) - \varrho(\frac{x-a_i-\varepsilon}{\varepsilon}) - \varrho(\frac{x-b_i+\varepsilon}{\varepsilon}) + \varrho(\frac{x-b_i}{\varepsilon})$. Note that t_i is the realization of a two-layer ReLU network with at most 12 nonzero weights, all of which satisfy the required bound on the absolute value. It is easy to see that $0 \leq t_i \leq 1$ and $t_i \equiv 1$ on $[a_i + \varepsilon, b_i - \varepsilon]$, while $t_i \equiv 0$ on $\mathbb{R} \setminus (a_i, b_i)$.

Now, define $T : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ by

$$T(x,y) := \sum_{\ell=0}^{1} (-1)^\ell B\, \varrho\Big(\varrho\big((-1)^\ell \tfrac{y}{B}\big) - d + \textstyle\sum_{i=1}^d t_i(x_i)\Big).$$

By construction, $T = \mathrm{R}_\varrho(\Lambda_{\varepsilon,B}^{(a,b)})$ for a ReLU network $\Lambda_{\varepsilon,B}^{(a,b)}$ as in the statement of the lemma. Furthermore, if $y \in [-B, B]$, then $\varrho\big((-1)^\ell \frac{y}{B}\big) - d + \sum_{i=1}^d t_i(x_i) \leq 1$, which shows $|T(x,y)| \leq B$. Next, if $y \in [-B, B]$ and $x \in [a,b]_\varepsilon$, then $t_i(x_i) = 1$ for all $i \in \underline{d}$, and hence $T(x,y) = y$. Finally, if $y \in [-B, B]$ and $x \in \mathbb{R}^d \setminus [a,b)$, then $t_i(x_i) = 0$ for some $i \in \underline{d}$, which entails $T(x,y) = 0$. This proves Eq. (IV.1). □

In [5], it was used that Estimate (IV.1) shows that $\mathrm{R}_\varrho(\Lambda_{\varepsilon,B}^{(a,b)})(\bullet, f(\bullet))$ and $\mathbb{1}_{[a,b]} \cdot f$ are close in L^p. Our next result shows that *if one properly chooses the endpoints a, b*, one even gets closeness (including an approximation rate) in $L^p(\mu)$. In addition to several results from [5], this observation is the main ingredient for our proof of Theorem II.3.

Proposition IV.2. *Let $d \in \mathbb{N}$, $\gamma > 0$, $p \in [1, \infty)$, and define $k := d + 1 + p\gamma$. For $N \in \mathbb{N}$, let $\Omega_N := \mathbb{Z}^d \cap [0, 2 \cdot 2^N - 1]^d$. For $a \in \mathbb{R}^d$ and $\omega \in \Omega_N$, let $a_N^{\omega,-} := a + \frac{\omega}{2^N} \in \mathbb{R}^d$ and $a_N^{\omega,+} := a + \frac{\omega + (1,\dots,1)}{2^N} \in \mathbb{R}^d$, as well as $I_N^{a,\omega} := [a_N^{\omega,-}, a_N^{\omega,+})$.*

Let μ be a finite Borel measure on $Q := [-\frac{1}{2}, \frac{1}{2}]^d$. For Lebesgue-almost every $a \in [-5,5]^d$, there is a constant $C_a = C_a(a,p,\mu,\gamma) > 0$ such that if $B \geq 1$ and if for each $\omega \in \Omega_N$ a measurable function $f_\omega : Q \to [-B, B]$ is given, then for every $N \in \mathbb{N}$, with $\Lambda_{\varepsilon,B}^{(a,b)}$ as in Lemma IV.1, we have

$$\Big\| \sum_{\omega \in \Omega_N} \Big[\mathrm{R}_\varrho(\Lambda_{2^{-kN},B}^{(a_N^{\omega,-}, a_N^{\omega,+})})(\bullet, f_\omega(\bullet)) - \mathbb{1}_{I_N^{a,\omega}} \cdot f_\omega \Big] \Big\|_{L^p(\mu)} \leq \frac{C_a B}{2^{N\gamma}}.$$

Proof. We consider μ as a Borel measure on $\mathbb{R}^d$, by setting $\mu(A) := \mu(A \cap Q)$. If $x \in [a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}$, then $x_i \notin [(a_N^{\omega,-})_i + 2^{-kN}, (a_N^{\omega,+})_i - 2^{-kN}]$ for some $i \in \underline{d}$; hence, $x_i \in [(a_N^{\omega,-})_i, (a_N^{\omega,-})_i + 2^{-kN}] \cup [(a_N^{\omega,+})_i - 2^{-kN}, (a_N^{\omega,+})_i]$. By definition of $a_N^{\omega,\pm}$, this implies $a_i \in J_N^{(i)}(x,\omega)$, where

$$J_N^{(i)}(x,\omega) := [(\theta_N^{x,\omega})_i - \tfrac{1}{2^{kN}}, (\theta_N^{x,\omega})_i] \cup [(\eta_N^{x,\omega})_i, (\eta_N^{x,\omega})_i + \tfrac{1}{2^{kN}}],$$

with $\theta_N^{x,\omega} := x - 2^{-N}\omega$ and $\eta_N^{x,\omega} := x - 2^{-N}(\omega + (1,\dots,1))$.

Let $R := [-5,5]^d$. Using $\mu(A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x)\, d\mu(x)$, we see

$$\begin{aligned} \circledast &:= \int_R \sum_{N=1}^\infty 2^{Np\gamma} \sum_{\omega \in \Omega_N} \mu\big([a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}\big)\, da \\ &\leq \sum_{N=1}^\infty \sum_{\omega \in \Omega_N} \sum_{i=1}^d 2^{N\gamma p} \underbrace{\int_R \int_{\mathbb{R}^d} \mathbb{1}_{J_N^{(i)}(x,\omega)}(a_i)\, d\mu(x)\, da}_{:= \oplus_N^{(i)}(\omega)}. \end{aligned}$$

Using the Lebesgue measure λ, Fubini's theorem shows

$$\oplus_N^{(i)}(\omega) = \int_{\mathbb{R}^d} \lambda\Big(\big\{a \in [-5,5]^d \colon a_i \in J_N^{(i)}(x,\omega)\big\}\Big)\, d\mu(x),$$

from which we get $\oplus_N^{(i)}(\omega) \leq 2 \cdot 10^{d-1} \mu(\mathbb{R}^d) \cdot 2^{-Nk}$. Thus, $\circledast \lesssim \sum_{N=1}^\infty |\Omega_N| \cdot 2^{N\gamma p} \cdot 2^{-Nk} \lesssim \sum_{N=1}^\infty 2^{N(d+\gamma p - k)} < \infty$, since $k = 1 + d + \gamma p$.

Recalling the definition of $\circledast$, we see that $\sum_{N=1}^\infty \big[2^{Np\gamma} \sum_{\omega \in \Omega_N} \mu\big([a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}\big)\big] < \infty$ for Lebesgue-almost every $a \in [-5,5]^d$. In particular, for Lebesgue-almost every $a \in [-5,5]^d$, there is a constant $C_a = C_a(a,p,\gamma,\mu) > 0$ such that for every $N \in \mathbb{N}$, we have $\sum_{\omega \in \Omega_N} \mu\big([a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}\big) \leq C_a \cdot 2^{-Np\gamma}$.

Let us fix such a point $a \in [-5,5]^d$, and for each $\omega \in \Omega_N$, let $f_\omega : Q \to [-B, B]$ be measurable. Estimate (IV.1) shows

$$\begin{aligned} &\sum_{\omega \in \Omega_N} \big| \mathrm{R}_\varrho\big(\Lambda_{2^{-kN},B}^{(a_N^{\omega,-}, a_N^{\omega,+})}\big)(\bullet, f_\omega(\bullet)) - \mathbb{1}_{[a_N^{\omega,-}, a_N^{\omega,+})} \cdot f_\omega \big| \\ &\leq 2B \sum_{\omega \in \Omega_N} \mathbb{1}_{[a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}} = 2B \cdot \mathbb{1}_P \end{aligned}$$

for $P := \biguplus_{\omega\in\Omega_N} [a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}$, where the union is disjoint. Hence,

$$\begin{aligned}
&\Big\| \sum_{\omega\in\Omega_N} \Big[\mathrm{R}_\varrho(\Lambda^{(a_N^{\omega,-}, a_N^{\omega,+})}_{2^{-kN},B})(\bullet, f_\omega(\bullet)) - \mathbb{1}_{[a_N^{\omega,-}, a_N^{\omega,+})} \cdot f_\omega \Big] \Big\|_{L^p(\mu)} \\
&\leq 2B \cdot \Big(\sum_{\omega\in\Omega_N} \mu\big([a_N^{\omega,-}, a_N^{\omega,+}) \setminus [a_N^{\omega,-}, a_N^{\omega,+}]_{2^{-kN}}\big) \Big)^{1/p} \\
&\leq 2B \cdot C_a^{1/p} \cdot 2^{-N\gamma} \quad \forall\, N \in \mathbb{N}. \qquad \square
\end{aligned}$$

To complete the proof of Theorem II.3, we need two results from [5]. The first result is concerned with an approximate implementation of a family of polynomials.

Lemma IV.3. *(see [5, Lemma A.5]) Let $d, m \in \mathbb{N}$ and $B, \beta > 0$. Set $Q := [-\frac{1}{2}, \frac{1}{2}]^d$, let $(x_\ell)_{\ell\in\underline{m}} \subset Q$, and $(c_{\ell,\alpha})_{\ell\in\underline{m},\alpha\in\mathbb{N}_0^d,|\alpha|<\beta} \subset [-B, B]$.*

Then there are $c = c(d,\beta,B) > 0$, $s = s(d,\beta,B) \in \mathbb{N}$, and $L = L(d,\beta) \in \mathbb{N}$ with $L \leq 1 + (1 + \lceil \log_2 \beta \rceil)(11 + \frac{\beta}{d})$ such that for all $\varepsilon \in (0, \frac{1}{2})$, there is a neural network Φ_ε with d-dimensional input, m-dimensional output, with $L(\Phi_\varepsilon) \leq L$ and $W(\Phi_\varepsilon) \leq c \cdot (m + \varepsilon^{-d/\beta})$, such that all weights of Φ_ε belong to $[-\varepsilon^{-s}, \varepsilon^{-s}]$, and such that

$$\Big| [\mathrm{R}_\varrho(\Phi_\varepsilon)(x)]_\ell - \sum_{|\alpha|<\beta} c_{\ell,\alpha}\, (x - x_\ell)^\alpha \Big| < \varepsilon \quad \forall\, \ell \in \underline{m} \text{ and } x \in Q.$$

Our final ingredient is a consequence of Taylor's theorem.

Lemma IV.4. *(see [5, Lemma A.8]) Let $n \in \mathbb{N}_0$, $\sigma \in (0, 1]$, and $\beta = n + \sigma$. Let $d \in \mathbb{N}$ and $Q := [-\frac{1}{2}, \frac{1}{2}]^d$. There is a constant $C = C(\beta, d) > 0$ such that for each $f \in C^\beta(Q)$ with $\|f\|_{C^\beta} \leq B$ and each $x_0 \in (-\frac{1}{2}, \frac{1}{2})^d$, there is a polynomial $p(x) = \sum_{|\alpha|\leq n} c_\alpha\, (x - x_0)^\alpha$ with $c_\alpha \in [-CB, CB]$ and such that $|f(x) - p(x)| \leq CB \cdot |x - x_0|^\beta$ for all $x \in Q$.*

Proof of Theorem II.3. Let us fix some $a \in \big((-\frac{3}{2}, -\frac{1}{2}] \setminus \mathbb{Q}\big)^d$ and a constant $C_1 = C_1(a, p, \mu, \beta) > 0$ satisfying the conclusion of Proposition IV.2 for the choice $\gamma := \beta$. Such an a exists, since $\big((-\frac{3}{2}, -\frac{1}{2}] \setminus \mathbb{Q}\big)^d$ has positive Lebesgue measure.

Let $N := \lceil \log_2 \varepsilon^{-1/\beta} \rceil \in \mathbb{N}$, whence $\frac{1}{\varepsilon^{1/\beta}} \leq 2^N \leq \frac{2}{\varepsilon^{1/\beta}}$. We observe that $Q \subset a + [0,2)^d \subset \biguplus_{\omega\in\Omega_N} [a_N^{\omega,-}, a_N^{\omega,+})$, since we have $Q - a \subset [0,2)^d$. Next, define $Q^\circ = (-\frac{1}{2}, \frac{1}{2})^d$ and $\Omega_N^* := \{\omega \in \Omega_N : Q \cap [a_N^{\omega,-}, a_N^{\omega,+}) \neq \varnothing\}$. Since $a \in (\mathbb{R} \setminus \mathbb{Q})^d$, we see for each $\omega \in \Omega_N^*$ that there is some $x_\omega \in Q^\circ \cap [a_N^{\omega,-}, a_N^{\omega,+})$. Set $m := |\Omega_N^*|$ and write $\Omega_N^* = \{\omega_1, \dots, \omega_m\}$ for suitable $\omega_1, \dots, \omega_m$. Note $m \leq |\Omega_N| = (2 \cdot 2^N)^d \leq 4^d\, \varepsilon^{-d/\beta}$. For $i \in \underline{m}$, set $x_i := x_{\omega_i}$.

Let $f \in C^\beta(Q)$ with $\|f\|_{C^\beta} \leq B$. Lemma IV.4 yields for each $\ell \in \underline{m}$ a sequence $(c_{\ell,\alpha})_{\alpha\in\mathbb{N}_0^d,|\alpha|<\beta} \subset [-C_2 B, C_2 B]$ such that $|f(x) - p_\ell(x)| \leq C_2 B \cdot |x - x_\ell|^\beta$ for all $x \in Q$, where $p_\ell(x) := \sum_{|\alpha|<\beta} c_{\ell,\alpha}\, (x - x_\ell)^\alpha$. Here, $C_2 = C_2(d,\beta) > 0$.

Next, we apply Lemma IV.3 (with $C_2 B$ instead of B) to obtain a neural network Φ with d-dimensional input and m-dimensional output such that $\big|\big[\mathrm{R}_\varrho(\Phi)(x)\big]_\ell - p_\ell(x)\big| \leq \frac{\varepsilon}{4} < 1$ for all $x \in Q$ and $L(\Phi) \leq 1 + (1 + \lceil \log_2 \beta \rceil)(11 + \beta/d)$, as well as $W(\Phi) \leq C_3 \cdot \big(m + (\varepsilon/4)^{-d/\beta}\big) \leq C_4 \cdot \varepsilon^{-d/\beta}$. Here, $C_i = C_i(d,\beta,B)$ for $i \in \{3,4\}$. Finally, all weights of Φ belong to $[-\varepsilon^{-s_1}, \varepsilon^{-s_1}]$ for some $s_1 = s_1(d,\beta,B) \in \mathbb{N}$.

Next, we use $|p_\ell(x)| \leq |p_\ell(x) - f(x)| + |f(x)|$ to derive

$$|p_\ell(x)| \leq C_2 B\, |x - x_\ell|^\beta + B \leq B(1 + d^\beta C_2),$$

whence $|f_{\omega_\ell}| \leq \frac{\varepsilon}{4} + B(1 + d^\beta C_2) \leq B' = B'(d,\beta,B) \geq 1$ for $f_{\omega_\ell} := \big(\mathrm{R}_\varrho(\Phi)\big)_\ell|_Q$. Let us set $f_\omega \equiv 0$ for $\omega \in \Omega_N \setminus \Omega_N^*$.

Now, set $g := \sum_{\omega\in\Omega_N^*} \mathbb{1}_{[a_N^{\omega,-}, a_N^{\omega,+})}\, f_\omega$ and $k := d + 1 + p\beta$ and furthermore $G := \sum_{\omega\in\Omega_N} \mathrm{R}_\varrho(\Lambda^{(a_N^{\omega,-}, a_N^{\omega,+})}_{2^{-kN},B'})(\bullet, f_\omega(\bullet))$. Then Proposition IV.2 (with B' instead of B) shows $\|G - g\|_{L^p(\mu)} \leq \frac{C_1 B}{2^{N\gamma}} \leq C_5\, \varepsilon$, where $C_5 = C_5(p,\mu,\beta,d,B)$.

For $x \in Q$, we have $x \in [a_N^{\omega_\ell,-}, a_N^{\omega_\ell,+})$ for a unique $\ell \in \underline{m}$, whence $|x - x_\ell| \leq d \cdot \|x - x_\ell\|_{\ell^\infty} \leq d\, 2^{-N} \leq d\, \varepsilon^{1/\beta}$. Therefore, $g(x) = f_{\omega_\ell}(x) = \big[\mathrm{R}_\varrho(\Phi)(x)\big]_\ell$, and hence

$$\begin{aligned}
|f(x) - g(x)| &\leq |f(x) - p_\ell(x)| + |p_\ell(x) - [\mathrm{R}_\varrho(\Phi)(x)]_\ell| \\
&\leq C_2 B \cdot |x - x_\ell|^\beta + \tfrac{\varepsilon}{4} \leq C_6 \cdot \varepsilon,
\end{aligned}$$

where $C_6 = C_6(d,\beta,B)$. Since μ is finite, we thus see $\|f - G\|_{L^p(\mu)} \leq C_7\, \varepsilon$ for $C_7 = C_7(p,\mu,d,\beta,B)$.

It remains to show $\|G - \mathrm{R}_\varrho(\Phi_\varepsilon^f)\|_{L^p(\mu)} \leq \varepsilon$ for a network Φ_ε^f as in the statement of Theorem II.3. First, since the class of neural networks is closed under composition and addition (including control over the complexity of the resulting networks; see the end of the proof of [5, Lemma A.7] for details), we see that $G = \mathrm{R}_\varrho(\Psi_\varepsilon^f)$ for a network Ψ_ε^f with $W(\Psi_\varepsilon^f) \leq C_8 \varepsilon^{-d/\beta}$ and $L(\Psi_\varepsilon^f) \leq 7 + (1 + \lceil \log_2 \beta \rceil)(11 + \frac{\beta}{d})$, where $C_8 = C_8(p,d,\beta,B)$, and such that all weights of Ψ_ε^f lie in $[-\varepsilon^{-s_2}, \varepsilon^{-s_2}]$ for some $s_2 = s_2(p,d,\beta,B,\mu) \in \mathbb{N}$. Here, we used that $|\Omega_N| \lesssim 2^{dN} \lesssim \varepsilon^{-d/\beta}$ and that all weights of the networks $\Lambda^{(a_N^{\omega,-}, a_N^{\omega,+})}_{2^{-kN},B'}$ have absolute value at most $d + B' + 2^{kN}(1 + \|a_N^{\omega,-}\|_{\ell^\infty} + \|a_N^{\omega,+}\|_{\ell^\infty}) \leq d + B' + \frac{1 \cdot 2^k}{\varepsilon^{k/\beta}}$, while all weights of Φ have absolute value at most ε^{-s_1}. Finally, [2, Lemma 3.7] can be used to obtain a quantized network. Precisely, that lemma yields a network Φ_ε^f with the properties stated in Theorem II.3 and such that $\|\mathrm{R}_\varrho(\Phi_\varepsilon^f) - \mathrm{R}_\varrho(\Psi_\varepsilon^f)\|_{\sup} \leq \varepsilon$. $\square$

References

[1] P. L. Bartlett, N. Harvey, C. Liaw, A. Mehrabian *Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks.* arXiv:1703.02930.

[2] H. Boelcskei, P. Grohs, G. Kutyniok, and P. Petersen *Optimal Approximation with Sparsely Connected Deep Neural Networks.* SIAM Journal on Mathematics of Data Science **1**(1) (2019), 8–45.

[3] G.F. Clements *Entropies of several sets of real valued functions.* Pacific J. Math **13** (1963), 1085–1095.

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

[5] P. Petersen and F. Voigtlaender *Optimal approximation of piecewise smooth functions using deep ReLU neural networks.* Neural Netw. **108** (2018), 296–330.

[6] P. Petersen and F. Voigtlaender *Equivalence of approximation by convolutional neural networks and fully-connected networks.* arXiv:1809.00973.

[7] I. Safran and O. Shamir *Depth-width tradeoffs in approximating natural functions with neural networks.* arXiv:1610.09887.

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

[9] D. Yarotsky *Optimal approximation of continuous functions by very deep ReLU networks.* arXiv:1802.03620.