APPROXIMATION OF SMOOTHNESS CLASSES BY DEEP RECTIFIER NETWORKS[*]

MAZEN ALI[†] AND ANTHONY NOUY[†]

Abstract. We consider approximation rates of sparsely connected deep rectified linear unit (ReLU) and rectified power unit (RePU) neural networks for functions in Besov spaces $B_q^\alpha(L^p)$ in arbitrary dimension d, on general domains. We show that deep rectifier networks with a fixed activation function attain optimal or near to optimal approximation rates for functions in the Besov space $B_\tau^\alpha(L^\tau)$ on the critical embedding line $1/\tau = \alpha/d + 1/p$ for *arbitrary* smoothness order $\alpha > 0$. Using interpolation theory, this implies that the entire range of smoothness classes at or above the critical line is (near to) optimally approximated by deep ReLU/RePU networks.

Key words. ReLU Neural Networks, Approximation Spaces, Besov Spaces, Direct Embeddings, Direct (Jackson) Inequalities

AMS subject classifications. 41A65, 41A15 (primary); 68T05, 65D99 (secondary)

1. Introduction. Artificial neural networks (NNs) have become a popular tool in various fields of computational and data science. Due to their popularity and good performance, NNs motivated a lot of research in mathematics – especially in recent years – in an attempt to explain the properties of NNs responsible for their success.

Although many aspects of NNs still lack a satisfactory mathematical explanation, the *expressivity* or approximation theoretic properties of NNs are by now quite well understood. By expressivity we mean the theoretical capacity of NNs to approximate functions from different classes. We do not intend to give a literature overview on this topic and instead refer to the recent survey in [15].

Contribution. In this work, we contribute to the existing body of knowledge on the expressivity of NNs by showing that the very popular rectifier NNs can approximate a wide range of smoothness classes in the Besov scale with (near to) optimal complexity. In the context of this work, "optimality" refers to the notion of *continuous nonlinear widths*[1] introduced in [7]. For the approximation of functions in the Besov space $B_q^\alpha(L^p(\mathbb{R}^d))$, an approximation tool with a continuous parameter selection for the approximand can achieve worst case approximation rates of at most α/d. To make the distinction to existing results clear, we briefly review what is known by now about the approximation of some more standard smoothness classes closely related to our work. In all instances "complexity" is measured by the number of connections, i.e., non-zero weights.

In [19], it was shown that analytic functions on a compact product domain in any dimension can be approximated in the Sobolev norm $W^{k,\infty}$ by ReLU and RePU networks with close to exponential convergence. In [20], it was shown that ReLU networks can approximate any Hölder continuous function with optimal complexity. In [14], it was shown that functions in the Besov space $B_p^\alpha(L^p(\Omega))$ on bounded Lipschitz domains $\Omega \subset \mathbb{R}^d$ in any dimension can be approximated in the L^p-norm by RePU networks with activation function of degree $r \gtrsim \alpha$ with optimal complexity. The spaces $B_p^\alpha(L^p(\Omega))$ correspond to the vertical line in Figure 1, i.e., for $p \geq 1$ these

[†]Centrale Nantes, LMJL UMR CNRS 6629, France ({mazen.ali, anthony.nouy}@ec-nantes.fr).

[1]Closely related to Aleksandrov widths, see [12].

are either the same or slightly larger than the Sobolev spaces $W^{k,p}(\Omega)$.
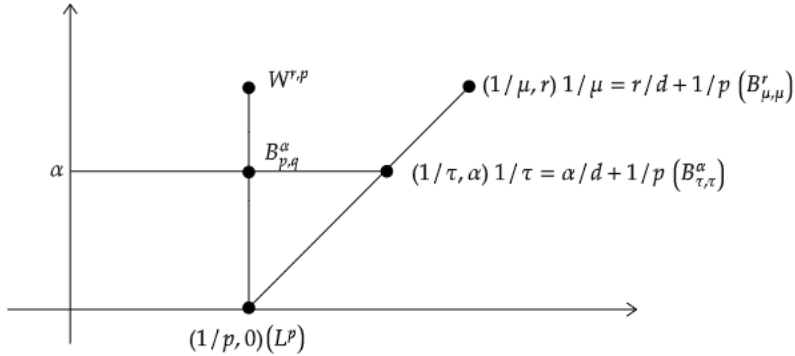

FIG. 1. *DeVore diagram of smoothness spaces [6]. The Sobolev embedding line is the diagonal with the points $(1/\tau, \alpha)$ and $(1/\mu, r)$, i.e., for a fixed p and variable $1/\tau$, the diagonal is the line $\alpha = d(1/\tau - 1/p)$ with slope d and offset $-d/p$. All points above the diagonal line correspond to Besov spaces compactly embedded in L^p, points on the line may or may not be continuously embedded in L^p, and points below the line are never embedded in L^p.*

In [18], it was shown that functions in the Besov space $B_\tau^\alpha(L^\tau(I))$, for $\alpha > 1/\tau - 1/p$ on bounded intervals $I \subset \mathbb{R}$, can be approximated in the (fractional) Sobolev $W^{s,p}(I)$-norm with deep ReLU networks with near to optimal complexity. In [22], the author shows that functions in $B_\tau^\alpha(L^\tau([0,1]^d))$ for $\alpha/d > 1/\tau - 1/p$ and in Besov spaces of dominating mixed smoothness can be approximated in L^p with deep ReLU networks with near to optimal complexity. The space $B_\tau^\alpha(L^\tau(\Omega))$ for $\alpha/d > 1/\tau - 1/p$ and Lipschitz domains Ω is above the *critical embedding line* of functions that barely have enough regularity to be members of L^p, see the diagonal in Figure 1. Spaces above this critical line are embedded in L^p, spaces on this line may or may not be embedded in L^p, and spaces below this line are never embedded in L^p.

It was also shown in [18] that piecewise Gevrey functions can be approximated with close to exponential convergence. Similar results for classical smoothness spaces of univariate functions are contained in [5].

In this work, we show that functions in isotropic Besov spaces $B_q^\alpha(L^\tau(\Omega))$, for $\alpha/d \geq 1/\tau - 1/p$, $q \leq \tau$ and $\Omega \subset \mathbb{R}^d$ an (ε, δ)- or a Lipschitz domain (see Definitions 1.2 and 1.3) in any dimension $d \in \mathbb{N}$, can be approximated by RePU networks with activation function of degree $r \geq 2$ with optimal complexity for any $\alpha > 0$. We show the same for ReLU networks with near to[2] optimal complexity. This completes the picture for rectifier networks expressivity rates for classical isotropic smoothness spaces in the sense that, with regard to L^p approximation, functions from any Besov space on or above the embedding line (see Figure 1), with $q \leq \tau$, can be approximated by ReLU/RePU networks with (near to) optimal complexity, universal in the smoothness order α.

Outline. We begin in Subsection 1.1 and Subsection 1.2 by reviewing the theoretical framework of our work. We then state the main result in Subsection 1.3 that includes a summary of the results on isotropic Besov spaces. To keep the presentation self-contained, we review previous results – that we require for our work – on ReLU approximation in Section 2 and Besov smoothness classes in Section 3. Finally, in Sec-

---

[2]For any approximation rate arbitrarily close to optimal.

tion 4 we derive the main result of this work, stated again in Theorem 4.3. The reader familiar with results on ReLU/RePU approximation and wavelet characterizations of Besov spaces can skip directly to Section 4.

Notation and Terminology. For quantities $A, B \in \mathbb{R}$, we will use the notation $A \lesssim B$ if there exists a constant C that does not depend on A or B such that $A \leq CB$. Similarly for $\gtrsim$ and $\sim$ if both inequalities hold. We use $\mathbb{N}$ for natural numbers and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We use $\mathrm{supp}(f)$ to denote the support of a function $f \in \mathbb{R}^d \to \mathbb{R}$

$$\mathrm{supp}(f) := \overline{\{x \in \mathbb{R}^d : f(x) \neq 0\}},$$

and $|\mathrm{supp}(f)|$ to denote the Lebesgue measure of this set. We use $\#$ to denote the standard counting measure.

We denote by $L^p(\Omega)$ the Lebesgue space of real-valued p-integrable functions for $0 < p \leq \infty$ on open subsets $\Omega \subset \mathbb{R}^d$ and by $H_p(\Omega)$ the real Hardy space (see [13]). Recall that the real Hardy spaces are isomorphic to L^p for $p > 1$. In this work, we will be referring to one of the following three types of domains Ω (see [1, 10, 16, 21]).

DEFINITION 1.1 (Special Lipschitz). *We call an open set $\Omega \subset \mathbb{R}^d$ a special Lipschitz domain if there exists a Lipschitz function $\phi : \mathbb{R}^{d-1} \to \mathbb{R}$ with $|\phi(x_1) - \phi(x_2)| \leq M\|x_1 - x_2\|_2$ for some constant $M > 0$ such that*

$$\Omega = \left\{(x, y) : x : \mathbb{R}^{d-1}, \ y \in \mathbb{R} \ and \ y > \phi(x)\right\}.$$

DEFINITION 1.2 (Strong Local Lipschitz Condition). *An open set $\Omega \subset \mathbb{R}^d$ is said to satisfy the strong local Lipschitz condition – also known as a minimally smooth domain or simply Lipschitz – if there exists $\varepsilon > 0$, $M > 0$, a locally finite open cover $\{U_i : i \in \mathbb{N}\}$ of $\partial\Omega$, and, for each i a real-valued function f_i of $d - 1$ variables, such that*

   (i) for some finite R, every collection of $R + 1$ of the sets U_i has empty intersection;

   (ii) for every pair of points $x, y \in \Omega$ such that $\mathrm{dist}(x, \partial\Omega), \mathrm{dist}(y, \partial\Omega) > \varepsilon$ and $\|x - y\|_2 < \varepsilon$, there exists i such that

$$x, y \in V_i := \left\{z \in U_i : \mathrm{dist}(z, \partial U_i) > \varepsilon\right\};$$

   (iii) each function f_i satisfies a Lipschitz condition with constant M;

   (iv) for some Cartesian coordinate system $(\xi_{i,1}, \ldots, \xi_{i,d})$ in U_i, $\Omega \cap U_i$ is represented by the inequality

$$\xi_{i,d} < f_i(\xi_{i,1}, \ldots, \xi_{i,d-1}).$$

DEFINITION 1.3 ((ε, δ)-Domain). *An open set Ω is called an (ε, δ)-domain if for any $x, y \in \Omega$, satisfying $\|x - y\|_2 \leq \delta$, there exists a rectifiable path $\Gamma \subset \Omega$ of length $\leq C_0\|x - y\|_2$ for some constant $C_0 > 0$, connecting x and y, such that for each $z \in \Gamma$,*

$$\mathrm{dist}(z, \partial\Omega) \geq \varepsilon \min\left\{\|z - x\|_2, \|z - y\|_2\right\}.$$

The inclusions between the different domain types are as follows

$$\text{special Lipschitz} \Rightarrow \text{strong locally Lipschitz} \Rightarrow (\varepsilon, \delta)\text{-domain.}$$

These domain types are not necessarily bounded and the special case $\Omega = \mathbb{R}^d$ trivially satisfies the strong local Lipschitz condition.
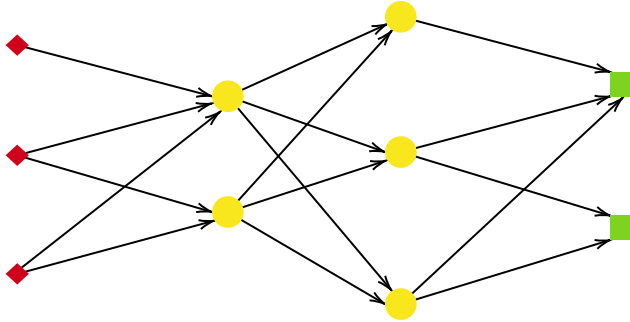
FIG. 2. *Example of a feed-forward neural network. On the left we have the* input *nodes marked in red that represent input data to the network. The yellow nodes are the* neurons *that perform some simple operations on the input. The edges between the nodes represent* connections *that transfer (after possibly applying an affine transformation) the output of one node into the input of another. The final green nodes are the* output *nodes. In this particular example the number of* layers L *is three, with two* hidden layers.

1.1. Neural Networks. We briefly introduce the mathematical description and notation we use for NNs throughout this work. Specifically, we will only consider feed-forward NNs. In Figure 2, we sketch a pictorial representation of a feed-forward NN.

Input values are passed on to the first layer of *neurons* after possibly undergoing an affine transformation. In the neurons, an *activation* function is applied to the transformed input values. The result again undergoes an affine transformation and is passed to the next layer and so on, until the output layer is reached.

The number of inputs and outputs is typically determined by the intended application. Specifying the architecture of such an NN amounts to choosing the number of layers, the number of neurons in each hidden layer, the activation functions and the connections or, equivalently, the position of the non-zero weights in the affine transformations. The process of training then consists of determining said weights.

We formalize our description of the considered mathematical objects. Let $L \in \mathbb{N}$ be the number of layers, N_0 the number of inputs, N_L the number of outputs and $N_1, \ldots, N_{L-1}$ the number of neurons in each hidden layer. A neural network Φ can be described by the tuple

$$\Phi := ((T_1, \boldsymbol{\sigma}_1), \ldots, (T_{L-1}, \boldsymbol{\sigma}_{L-1}), (T_L)),$$

where for each $1 \leq l \leq L$, T_l is an affine transformation

$$(1.1) \qquad T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}, \quad x \mapsto A_l x + b_l, \quad A_l \in \mathbb{R}^{N_l \times N_{l-1}}, \ b_l \in \mathbb{R}^{N_l},$$

and $\boldsymbol{\sigma}_l : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l}$ is a (nonlinear) function, usually applied component-wise as

$$x \mapsto (\sigma_l^{(1)}(x_1), \ldots, \sigma_l^{(N_l)}(x_{N_l})).$$

In this work we will use RePU activation functions, i.e.,

$$(1.2) \qquad \sigma_l^{(i)} \in \{\mathbb{I}_{\mathbb{R}}, \rho_r\}, \quad \rho_r(t) := \max\{0, t\}^r, \quad 1 \leq l \leq L-1, \ r \in \mathbb{N}.$$

where $\mathbb{I}_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$ is the identity map and for $r = 1$, ρ_1 is referred to as the *rectified linear unit (ReLU)*. We allow for the possibility of a non-strict network, i.e., an activation function is either $\mathbb{I}_{\mathbb{R}}$ or ρ_r. Another possibility is a *strict* network where each

activation function is necessarily ρ_r (with the exclusion of the output nodes). But, as was shown in [14], the approximation theoretic properties of both are the same and thus, for our work, it is irrelevant.

Let $\mathsf{Aff}(N_{l-1}, N_l)$ denote the set of affine maps as in (1.1) and $\mathsf{NL}(N_l, r)$ denote the set of activation functions as in (1.2). For fixed N_0, N_L, define

$$\mathsf{RePU}^{r,N_0,N_L} :=$$

$$\bigcup_{L \in \mathbb{N}} \bigcup_{(N_1,\dots,N_{L-1}) \in \mathbb{N}^{L-1}} \mathsf{Aff}(N_0, N_1) \times \mathsf{NL}(N_1, r) \times \cdots \times \mathsf{NL}(N_{L-1}, r) \times \mathsf{Aff}(N_{L-1}, N_L),$$

$$\mathsf{ReLU}^{N_0,N_L} := \mathsf{RePU}^{1,N_0,N_L},$$

and the realization map $\mathcal{R} : \mathsf{RePU}^{r,N_0,N_L} \to (\mathbb{R}^{N_L})^{\mathbb{R}^{N_0}}$ by

$$\mathcal{R}(\Phi) := T_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \boldsymbol{\sigma}_1 \circ T_1.$$

1.2. Approximation Classes. In this work, we will derive results in the approximation theoretic framework introduced in [14]. Before we do so, let us first recall the definition of approximation classes.

Let X be a quasi-normed linear space, $\Sigma_n \subset X$ subsets of X for $n \in \mathbb{N}_0$ and $\Sigma := (\Sigma_n)_{n \in \mathbb{N}_0}$ an approximation tool. Define the best approximation error

$$E(f, \Sigma_n)_X := \inf_{\varphi \in \Sigma_n} \|f - \varphi\|_X .$$

With this we define approximation classes as

DEFINITION 1.4 (Approximation Classes). *For any $f \in X$ and $\alpha > 0$, define the quantity*

$$\|f\|_{A_q^\alpha} := \begin{cases} \left(\sum_{n=1}^\infty [n^\alpha E(f, \Sigma_n)_X]^q \frac{1}{n}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{n \geq 1} [n^\alpha E(f, \Sigma_n)_X], & q = \infty. \end{cases}$$

The approximation classes A_q^α of $\Sigma = (\Sigma_n)_{n \in \mathbb{N}_0}$ are defined by

$$A_q^\alpha(X, \Sigma) := \left\{ f \in X : \|f\|_{A_q^\alpha} < \infty \right\}.$$

The utility of using these classes comes to light only if the sets Σ_n satisfy certain properties. This was discussed in detail in [14] and the relevant properties were shown to hold for RePU networks.

We perform approximation in $X = L^p(\Omega)$ for $1 < p \leq \infty$, we comment on the case $0 < p \leq 1$ in Remark 4.4. For the domain Ω, we require only the existence of an extension operator bounded in the Besov norm. From [9, 10], we know this can be ensured for (ε, δ)-*domains* and for Lipschitz domains, see also Theorem 3.1. We abbreviate

$$E(f, \Sigma_n)_X := E(f, \Sigma_n)_X.$$

As a measure of complexity we will use the number of non-zero weights. For a given $\Phi \in \mathsf{RePU}^{r,d_1,d_2}$ for some $d_1, d_2 \in \mathbb{N}$, the number of non-zero weights is

$$W(\Phi) := \sum_{l=1}^L \|T_l\|_{\ell^0} , \quad \|T_l\|_{\ell^0} := \|A_l\|_{\ell^0} ,$$

$\|A_l\|_{\ell^0}$ being the number of non-zero weights of the matrix A_l. With this we define for any $n \in \mathbb{N}$

$$\mathsf{RePU}_n^{r,d_1,d_2} := \left\{ \Phi \in \mathsf{RePU}^{r,d_1,d_2} : \ \mathcal{R}(\Phi) \in X, \ W(\Phi) \le n \right\},$$
$$\mathsf{ReLU}_n^{d_1,d_2} := \mathsf{RePU}_n^{1,d_1,d_2}.$$

The main result of this work then concerns the approximation classes

$$(1.3) \qquad\qquad A_q^\alpha \left(X, \mathsf{RePU}^{r,d,1} \right).$$

1.3. Main Result. We summarize results on approximation of isotropic Besov spaces – including this work – in Table 1.

X	Domain Ω	Smoothness Class	Approximation Rate $\mathsf{RePU}^{r,d,1}$	$\mathsf{ReLU}^{d,1}$	Reference
$L^p(\Omega)$ $1 < p \le \infty$	(ε, δ) or Lipschitz	$B_\tau^\alpha(L^\tau(\Omega))$ $\alpha \ge 1/\tau - 1/p$	α/d	$\sim \alpha/d$	this work
$L^p(\Omega)$ $0 < p \le \infty$	$[0,1]^d$	$B_\tau^\alpha(L^\tau(\Omega))$ $\alpha > 1/\tau - 1/p$	–	$\sim \alpha/d$	[22]
$L^p(\Omega)$ $0 < p \le \infty$	bounded Lipschitz	$B_p^\alpha(L^p(\Omega))$ $\alpha < r + \min(1, 1/p)$	α/d	α/d	[14]

TABLE 1
Summary of approximation rates for isotropic Besov spaces with deep rectifier networks. Lipschitz refers to the strong locally Lipschitz condition from Definition 1.2. We use $\sim \alpha/d$ to indicate algebraic rates with an additional log term or, in other words, any rate strictly less than α/d can be achieved.

The precise statements for the direct estimates can be found in Lemma 4.2. These estimates imply a range of interpolated smoothness spaces that are continuously embedded in the approximation classes from (1.3)

$$(L^p(\Omega), B_q^\alpha(L^\tau(\Omega)))_{\theta/\alpha, \bar{q}} \hookrightarrow A_{\bar{q}}^{\theta/d}(L^p(\Omega), \mathsf{RePU}^{r,d,1}),$$

see the Theorem 4.3 and definitions in Subsection 3.1.

The required depth to achieve the optimal rates from Table 1 for RePU networks with $r \ge 2$ scales at most logarithmically in the smoothness order α and, in particular, is independent of the approximation error. For ReLU networks, the required depth scales at most logarithmically with the approximation error and at most log-linearly with smoothness order α.

2. Preliminaries on Rectifier Network Approximation. In this section, we review recent results on deep RePU approximation relevant for this work. We use the notation defined in Subsection 1.1. The next theorem states that RePU networks can efficiently reproduce or approximate multiplication.

THEOREM 2.1 (Multiplication [14, 19, 23]). *Let $M_d : \mathbb{R}^d \to \mathbb{R}$ be the multiplication function $x \mapsto \prod_{i=1}^d x_i$. Then, there exists a constant C such that*

(i) *for $r \geq 2$, and $n := Cd$, there exists a RePU network $\Phi_M \in \mathsf{RePU}_n^{r,d,1}$ such that*

$$M_d = \mathcal{R}(\Phi_M),$$

where the depth of Φ_M depends at most logarithmically on d.

(ii) *For $r = 1$, any $K > 0$ and any $0 < \varepsilon < 1$, and $n := Cd\log(dK^d/\varepsilon)$, there exists a ReLU network $\Phi_M^\varepsilon \in \mathsf{ReLU}_n^{d,1}$ with*

$$\|M_d - \mathcal{R}(\Phi_M^\varepsilon)\|_{L^\infty([-K,K]^d)} \leq \varepsilon,$$

where the depth is at most a constant multiple of $(1 + \log(d)\log(dK^d/\varepsilon))$.

This in turn implies RePU networks can efficiently reproduce or approximate piecewise polynomials.

THEOREM 2.2 (Piecewise Polynomials [14, 18, 23]). *Let $v : \mathbb{R} \to \mathbb{R}$ be a piecewise polynomial with N_v pieces, of maximum degree $t \in \mathbb{N}_{\geq 0}$ and with compact support of measure $S := |\operatorname{supp}(v)| < \infty$. Then, there exists a constant $C > 0$ depending on N_v, t and r such that*

(i) *for $r \geq 2$ and any $\varepsilon > 0$, there exists a RePU network $\Phi \in \mathsf{RePU}_C^{r,1,1}$ with*

$$\|v - \mathcal{R}(\Phi)\|_{L^\infty(\mathbb{R})} \leq \varepsilon,$$

where the complexity of the network Φ is independent of ε, the depth is of the order $\mathcal{O}(\log(t))$ and $C = \mathcal{O}(N_v \log(t))$.

(ii) *For $r = 1$, the constant C additionally depends on S and $\|v\|_{L^\infty(\mathbb{R})}$, and for any $0 < \varepsilon < 1$ there exists a ReLU network $\Phi \in \mathsf{ReLU}_n^{1,1}$ with $n := C(1 + \log(\varepsilon^{-1}))$ and the same support as v, such that*

$$\|v - \mathcal{R}(\Phi)\|_{L^\infty(\mathbb{R})} \leq \varepsilon,$$

where the depth of the network is at most of the order $\mathcal{O}(t\log(t)\log(\varepsilon^{-1}))$ and $C = \mathcal{O}(N_v t \log(t))$. For the special case $t = 1$, i.e., if v is piecewise linear, it can be reproduced exactly with $v = \mathcal{R}(\Phi)$ such that $\Phi \in \mathsf{ReLU}_{C(1+N_v)}^{1,1}$ and has depth two.

The previous result states that RePU networks with $r \geq 2$ can reproduce piecewise polynomials of any degree at the same asymptotic cost[3]. This suggests the following *saturation property*.

THEOREM 2.3 (Saturation Property [14]). *For any $r \geq 2$, $\alpha > 0$, any $d_1, d_2 \in \mathbb{N}$, $0 < p \leq \infty$ and $\Omega \subset \mathbb{R}^d$ a Borel-measurable open set with nonzero measure, the approximation spaces defined in* Subsection 1.2 *coincide*

$$A_q^\alpha(L^p(\Omega), \mathsf{RePU}^{2,d_1,d_2}) = A_q^\alpha(L^p(\Omega), \mathsf{RePU}^{r,d_1,d_2}).$$

The saturation property will also be clearly visible in the main result of this work in Theorem 4.3.

We conclude by pointing out that RePU networks can efficiently reproduce *affine systems*, i.e., linear combinations of functions that are generated by dilating and shifting a single *mother* function or, in some cases, a finite number of mother functions.

---

[3]Note that the constants will be, however, affected by the degree.

A prominent example of affine systems are wavelets which will play an important role for the main result in Theorem 4.3.

The reproduction of affine systems by NNs was studied in greater detail in [2]. In the following we only mention the properties relevant for this work.

THEOREM 2.4 (NN calculus [14]). *For any $r \geq 1$, the following properties hold.*
(i) *For any $c \in \mathbb{R}$, $n \in \mathbb{N}$, $d_1, d_2 \in \mathbb{N}$ and any $\Phi_1 \in \mathsf{RePU}_n^{r,d_1,d_2}$, there exists $\Phi_2 \in \mathsf{RePU}_n^{r,d_1,d_2}$ with*

$$c\mathcal{R}(\Phi_1) = \mathcal{R}(\Phi_2),$$

where both Φ_1 and Φ_2 have the same depth.
(ii) *For any $n_1, \ldots, n_N \in \mathbb{N}$, $d_1, d_2 \in \mathbb{N}$ and any $\Phi_1 \in \mathsf{RePU}_{n_1}^{r,d_1,d_2}, \ldots, \Phi_N \in \mathsf{RePU}_{n_N}^{r,d_1,d_2}$, set $C := \min\{d_1, d_2\}(\max_i \operatorname{depth}(\Phi_i) - \min_i \operatorname{depth}(\Phi_i))$. Then, for $n := C + \sum_{i=1}^{N} n_i$, there exists $\Phi_{\sum} \in \mathsf{RePU}_n^{r,d_1,d_2}$ with*

$$\sum_{i=1}^{N} \mathcal{R}(\Phi_i) = \mathcal{R}(\Phi_{\sum}),$$

where the depth of $\Phi_{\sum}$ is bounded by the maximal depth of all Φ_i's.
(iii) *For any $n_1, \ldots, n_N \in \mathbb{N}$, $d, d_1, \ldots, d_N \in \mathbb{N}$ and any*

$$\Phi_1 \in \mathsf{RePU}_{n_1}^{r,d,d_1}, \ldots, \Phi_N \in \mathsf{RePU}_{n_N}^{r,d,d_N},$$

set $K := \sum_{i=1}^{N} d_i$ and $C := \min\{d, K-1\}(\max_i \operatorname{depth}(\Phi_i) - \min_i \operatorname{depth}(\Phi_i))$. Then, for $n := C + \sum_{i=1}^{N} n_i$, there exists $\Phi_{\times} \in \mathsf{RePU}_n^{r,d,K}$ with

$$(\mathcal{R}(\Phi_1), \ldots, \mathcal{R}(\Phi_N)) = \mathcal{R}(\Phi_{\times}),$$

where the depth of $\Phi_{\times}$ is bounded by the maximal depth of all Φ_i's.
(iv) *For any $n_1, n_2 \in \mathbb{N}$, $d_1, d_2, d_3 \in \mathbb{N}$, any $\Phi_1 \in \mathsf{RePU}_{n_1}^{r,d_1,d_2}$ and any $\Phi_2 \in \mathsf{RePU}_{n_2}^{r,d_2,d_3}$, there exists $\Phi \in \mathsf{RePU}_{n_1+n_2}^{r,d_1,d_3}$ such that*

$$\mathcal{R}(\Phi_2) \circ \mathcal{R}(\Phi_1) = \mathcal{R}(\Phi),$$

where the depth of Φ is simply the sum of the depths of Φ_1 and Φ_2.
(v) *Let $D_b^a : \mathbb{R}^d \to \mathbb{R}^d$ denote the affine transformation $x \mapsto ax - b$ for $a \in \mathbb{R}$, $b \in \mathbb{R}^d$. Then, for any $n \in \mathbb{N}$, $d_1, d_2 \in \mathbb{N}$, any $\Phi_1 \in \mathsf{RePU}_n^{r,d,1}$ and any $a \in \mathbb{R}$, $b \in \mathbb{R}^d$, there exists $\Phi_2 \in \mathsf{RePU}_n^{r,d,1}$ with*

$$\mathcal{R}(\Phi_1) \circ D_b^a = \mathcal{R}(\Phi_2),$$

where both Φ_1 and Φ_2 are of the same depth.

3. Besov Spaces and Wavelet Systems. In this section, we recall some classical results on (isotropic) Besov spaces and their characterization with wavelets. As in Section 2, we focus mostly on results relevant to our work. For more details we refer to, e.g., [3].

3.1. Besov Spaces. Let $\Omega \subset \mathbb{R}^d$, $h \in \mathbb{R}^d$ and τ_h a translation operator $(\tau_h f)(x) := f(x+h)$, $\mathbb{I} : \mathbb{R}^d \to \mathbb{R}^d$ the identity operator and define the *m-th difference*

$$\Delta_h^m := (\tau_h - \mathbb{I})^m := \underbrace{(\tau_h - \mathbb{I}) \circ \ldots \circ (\tau_h - \mathbb{I})}_{m \text{ times}}, \quad m \in \mathbb{N}.$$

We use the notation

$$\Delta_h^m(f, x, \Omega) := \begin{cases} (\Delta_h^m f)(x), & \text{if } x, x+h, \ldots, x+rh \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

The *modulus of smoothness* of order m is defined for any $t > 0$ as

$$\omega_m(f, t, \Omega)_X = \sup_{|h| \leq t} \|\Delta_h^m(f, \cdot, \Omega)\|_X,$$

where $|h|$ denotes the standard Euclidean 2-norm. Finally, the Besov quasi-semi-norm is defined for any $0 < q \leq \infty$, any $\alpha > 0$ and $m := \lfloor \alpha \rfloor + 1$ by

$$|f|_{B_q^\alpha(X)} := \begin{cases} \left(\int_0^1 [t^{-\alpha} \omega_m(f, t, \Omega)_X]^q \, \mathrm{d}t/t \right)^{1/q}, & 0 < q < \infty, \\ \sup_{t>0} t^{-\alpha} \omega_m(f, t, \Omega)_X, & q = \infty. \end{cases}$$

Then, the (isotropic) Besov space is defined as

$$B_q^\alpha(X) := \left\{ f \in X : |f|_{B_q^\alpha(X)} < \infty \right\}.$$

and it is a (quasi-)Banach space that we equip with the (quasi-)norm

$$\|f\|_{B_q^\alpha(X)} := \|f\|_X + |f|_{B_q^\alpha(X)}, \quad X = L^p(\Omega),$$
$$\|f\|_{B_q^\alpha(X)} := |f|_{B_q^\alpha(X)}, \qquad X = H_p(\mathbb{R}^d),$$

where $H_p(\mathbb{R}^d)$ is the real Hardy space.

The parameter $\alpha > 0$ is the smoothness order, while the space X reflects the measure of said smoothness. The secondary parameter q is less important and merely provides a finer gradation of smoothness. For $X = L^p(\Omega)$, a few relationships are rather straightforward

$$B_q^{\alpha_1}(L^p(\Omega)) \hookrightarrow B_q^{\alpha_2}(L^p(\Omega)), \quad \alpha_1 \geq \alpha_2,$$
$$B_q^\alpha(L^{p_1}(\Omega)) \hookrightarrow B_q^\alpha(L^{p_2}(\Omega)), \quad p_1 \geq p_2,$$
$$B_{q_1}^\alpha(L^p(\Omega)) \hookrightarrow B_{q_2}^\alpha(L^p(\Omega)), \quad q_1 \leq q_2,$$

where $\hookrightarrow$ denotes a continuous embedding. For non-integer $\alpha > 0$ and $1 \leq p \leq \infty$, $B_p^\alpha(L^p(\Omega))$ is the fractional Sobolev space $W^{\alpha,p}(\Omega)$. For integer $\alpha > 0$, the Besov space $B_\infty^\alpha(L^p(\Omega))$ is slightly larger than $W^{\alpha,p}(\Omega)$. For $p = q = 2$, the Besov space $B_2^\alpha(L^2(\Omega))$ is the same as the Sobolev space $W^{\alpha,2}(\Omega)$.

The Besov spaces $B_\tau^\alpha(L^\tau(\Omega))$, $1/\tau = \alpha/d + 1/p$ are on the *critical embedding line* (see Figure 1). Spaces above this line are embedded in L^p, spaces on this line may or may not be embedded in L^p, and spaces below this line are never embedded in L^p. In this sense, such Besov spaces are quite large as the functions on this line barely have enough regularity to be members of L^p. It is well-known that optimal approximation

– in the sense of continuous nonlinear widths (see also beginning of Section 1) – of functions from such spaces with a continuous parameter selection can only be achieved by nonlinear methods, see [7]. It is the main result of this work that RePU networks achieve optimal approximation for these spaces, while ReLU networks achieve near to optimal approximation.

To transfer results from $\mathbb{R}^d$ to more general domains, we will use the common technique of extension operators.

THEOREM 3.1 (Extension Operator [9, 10]). *Let $\alpha > 0$, $0 < p, q \leq \infty$ and let $\Omega \subset \mathbb{R}^d$ be an (ε, δ)-domain for $0 < p \leq 1$ and a strong locally Lipschitz domain for $1 < p \leq \infty$. Then, there exists an extension operator $\mathcal{E} : B_q^\alpha(L^p(\Omega)) \to B_q^\alpha(L^p(\mathbb{R}^d))$ such that $\mathcal{E}f|_\Omega = f$ and*

$$\|f\|_{B_q^\alpha(L^p(\Omega))} \leq \|\mathcal{E}f\|_{B_q^\alpha(L^p(\mathbb{R}^d))} \leq C \|f\|_{B_q^\alpha(L^p(\Omega))},$$

where C depends only on d, α, p and the domain Ω.

We conclude by noting that Besov spaces combine well with interpolation. To be precise, we briefly define interpolation spaces via the K-functional. Let X be a quasi-normed space and Y be a quasi-semi-normed space with $Y \hookrightarrow X$. The K-functional is defined for any $f \in X$ by

$$K(f, t, X, Y) := \inf_{f = f_0 + f_1} \{\|f_0\|_X + t\, |f_1|_Y\}, \quad t > 0.$$

For $0 < \theta < 1$ and $0 < q \leq \infty$, define the quantity

$$|f|_{(X,Y)_{\theta,q}} := \begin{cases} \left(\int_0^\infty [t^{-\theta} K(f, t, X, Y)]^q \, dt/t\right)^{1/q}, & 0 < q < \infty, \\ \sup_{t>0} t^{-\theta} K(f, t, X, Y), & q = \infty. \end{cases}$$

Then, the spaces

$$(X, Y)_{\theta,q} := \left\{ f \in X : |f|_{(X,Y)_{\theta,q}} < \infty \right\},$$

equipped with the (quasi-)norm

$$\|f\|_{(X,Y)_{\theta,q}} := \|f\|_X + |f|_{(X,Y)_{\theta,q}},$$

are *interpolation spaces*.

Besov spaces provide a relatively complete description of interpolation spaces in the following sense: for $0 < \theta < 1$

$$\begin{aligned} (L^p(\Omega), W^\alpha(L^p(\Omega)))_{\theta,q} &= B_q^{\theta\alpha}(L^p(\Omega)), && 1 \leq p \leq \infty,\ 0 < q \leq \infty, \\ (B_{q_1}^{\alpha_1}(L^p(\Omega)), B_{q_2}^{\alpha_2}(L^p(\Omega)))_{\theta,q} &= B_q^\alpha(L^p(\Omega)), && 0 < \alpha_1 < \alpha_2,\ \alpha := (1-\theta)\alpha_1 + \theta\alpha_2, \\ &&& 0 < p, q, q_1, q_2 \leq \infty, \\ (L^p(\Omega), B_{q_1}^\alpha(L^p(\Omega)))_{\theta,q} &= B_q^{\theta\alpha}(L^p(\Omega)), && 0 < p, q, q_1 \leq \infty. \end{aligned}$$

For Besov spaces on the critical line with $1/\tau = \alpha/d + 1/p$, we obtain

$$(L^p(\Omega), B_\tau^\alpha(L^\tau(\Omega)))_{\theta,q} = B_q^{\theta\alpha}(L^q(\Omega)), \quad \text{if} \quad 1/q = \theta\alpha/d + 1/p.$$

3.2. Wavelets. There are many possible wavelets constructions satisfying different properties depending on the intended application. Said constructions can be rather technical, with the payoff being various favorable analytical and numerical features. We do not intend to cover this topic in-depth and once again only pick out the aspects required for this work. We proceed by briefly reviewing one-dimensional wavelets constructions, after which we turn to wavelets on $\mathbb{R}^d$. Our presentation is somewhat abstract and therefore flexible, but we will also be more specific with some aspects of the construction that we require in Section 4. For more details on the subject we refer to [3].

The starting point of a wavelet construction is typically a multi-resolution analysis (MRA), i.e., a sequence of closed subspaces $V_j \subset V_{j+1}$ of $L^2(\mathbb{R})$ that are nested, dilation- and shift-invariant, dense in $L^2(\mathbb{R})$ and are all generated by a single[4] *scaling function* $\varphi \in V_0$. To be more precise, we assume the system $\{\varphi(\cdot - k) : k \in \mathbb{Z}\}$ is a Riesz basis of V_0 and therefore $\{\varphi(2^j \cdot -k) : k \in \mathbb{Z}\}$ is a Riesz basis of V_j. We use the shorthand notation

$$(3.1) \qquad \varphi_{j,k} := 2^{j/2}\varphi(2^j \cdot -k),$$

where the pre-factor $2^{j/2}$ normalizes φ in L^2. Later we will redefine this to $2^{j/p}$ for normalization in L^p or H_p for any $0 < p \leq \infty$, with the convention $2^{j/\infty} = 1$. Here H_p denotes the real Hardy space which coincides with L^p for $p > 1$, see [13] and Remark 4.4.

Defining a projection $P_j : L^2(\mathbb{R}) \to V_j$ is rather simple if $\{\varphi(\cdot - k) : k \in \mathbb{Z}\}$ forms an orthogonal basis of V_0. Indeed, this property implies that $\{\varphi_{j,k} : k \in \mathbb{Z}\}$ forms an orthogonal basis of V_j, and P_j can be chosen to be the orthogonal projection. However, for numerical reasons, it is sometimes unpractical to construct scaling functions φ such that $\{\varphi(\cdot - k) : k \in \mathbb{Z}\}$ forms an orthogonal basis of V_0 and without this property a constructive definition of P_j is not straightforward.

A way-out are so-called *bi-orthogonal* constructions. A function $\tilde{\varphi} \in L^2(\mathbb{R})$ is dual to φ if it satisfies

$$\langle \varphi(\cdot - k), \tilde{\varphi}(\cdot - l) \rangle_{L^2} = \delta_{k,l}, \quad k, l \in \mathbb{Z},$$

where $\delta_{k,l}$ is the Kronecker delta. We then define the *oblique* projection P_j

$$P_j f := \sum_{k \in \mathbb{Z}} \langle f, \tilde{\varphi}_{j,k} \rangle_{L^2} \varphi_{j,k}.$$

A representation of a function in V_j is typically referred to as a *single-scale* representation. To switch to a *multi-scale* representation, we need to characterize the so-called *detail spaces* defined through the projections

$$Q_j := P_{j+1} - P_j,$$

with the detail spaces defined as $W_j := Q_j(L^2(\mathbb{R}))$. This is achieved by constructing a *wavelet* $\psi \in V_1$

$$\psi := \sum_{k \in \mathbb{Z}} g_k \varphi(2 \cdot -k),$$

---

[4]Multiple scaling functions are possible as well in which case such functions are referred to as *multi-wavelets*, see [11].

for some coefficients $g_k \in \mathbb{R}$ such that

$$N_\psi := \#\{k : g_k \neq 0\} < \infty.$$

Any function $f \in L^2(\mathbb{R})$ can then be decomposed into a sequence of detail coefficients

$$(3.2) \qquad f = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} c_{j,k} \psi_{j,k},$$

where $\psi_{j,k}$ is defined as in (3.1). To simplify notation, one typically introduces the index set $\nabla := \mathbb{Z} \times \mathbb{Z}$. Decomposition (3.2) then simplifies to

$$f = \sum_{\lambda \in \nabla} c_\lambda \psi_\lambda.$$

In order for the wavelets ψ_λ to characterize Besov spaces, they have to satisfy certain assumptions.

ASSUMPTION 3.2 (Characterization). *We assume the scaling function φ and its dual $\tilde{\varphi}$ satisfy the following properties.*
  (W1) *(Integrability) For some $p', p'' \in [1, \infty]$ such that $1/p' + 1/p'' = 1$, we assume $\varphi \in L^{p'}(\mathbb{R})$ and $\tilde{\varphi} \in L^{p''}(\mathbb{R})$.*
  (W2) *(Polynomial Reproduction) We assume φ satisfies* Strang-Fix *conditions of order $L \in \mathbb{N}$ or, equivalently, for any polynomial $P \in \mathbb{P}_{L-1}$ of degree $L - 1$, we have $P \in V_0$.*
  (W3) *(Regularity) For some $s > 0$ and some $0 < \tau, q \leq \infty$, we assume $\varphi \in B_q^s(L^\tau(\mathbb{R}))$.*

These conditions are sufficient to ensure Besov spaces can be characterized by the decay of the wavelet coefficients for the case $p \geq 1$ or given sufficient regularity. For the case $X = H_p(\mathbb{R}^d)$ and $0 < p \leq 1$, we additionally require φ to satisfy[5]

ASSUMPTION 3.3 (Hardy Spaces). *For $0 < p \leq 1$, assume φ satisfies*
  (H1) *$\hat{\varphi}(0) = 1$, $D^\beta \hat{\varphi}(0) = 0$ for every $1 \leq |\beta| < \max\{[d(1/p - 1)] + 1, L\}$, where $\hat{\varphi}$ denotes the Fourier transform of φ.*

For N-term approximation results in the case $p = \infty$, we additionally require

ASSUMPTION 3.4 ($p = \infty$).
  (A1) *We assume φ has compact support and is s times continuously differentiable.*

Finally, for our work we will require two additional conditions that are, however, easy to satisfy for a variety of wavelet families.

ASSUMPTION 3.5 (Piecewise Polynomial). *We additionally assume the scaling function φ satisfies the following properties.*
  (P1) *We assume φ has compact support.*
  (P2) *We assume φ is piecewise polynomial.*

An example of wavelet families that can be constructed to satisfy all of the assumptions (W1)–(W3), (H1), (A1) and (P1)–(P2) are the CDF bi-orthogonal B-spline wavelets from [4]. These constructions allow to choose an arbitrary polynomial reproduction degree $L - 1$, regularity order s and the resulting scaling function φ (and consequently ψ as well) are compactly supported splines of degree $L - 1$.

---

[5]Weaker assumptions are also possible, see [17] for details.

Finally, we briefly describe how to extend the above wavelets to the multivariate case. There are several possible approaches for this, but we describe a specific tensor product construction suitable for isotropic Besov spaces.

For $x \in \mathbb{R}^d$, we define the tensor product scaling function as

$$\phi(x) := \varphi(x_1) \cdots \varphi(x_d),$$

and in the same manner as before, but for a general $0 < p \leq \infty$,

$$(3.3) \qquad \phi_{j,k,p}(x) := 2^{dj/p} \phi(2^j x - k), \quad j \in \mathbb{Z}, \ k \in \mathbb{Z}^d,$$

with the convention $2^{dj/\infty} = 1$. Next, for $e \in \{0,1\}^d \setminus \{0\}$, we define

$$(3.4) \qquad \psi^e(x) := \psi^{e_1}(x_1) \cdots \psi^{e_d}(x_d),$$

with the convention $\psi^1(x_i) := \psi(x_i)$ and $\psi^0(x_i) = \varphi(x_i)$, and $\psi^e_{j,k,p}$ is defined as in (3.3). Simplifying as before with

$$\nabla := \{(e,j,k) : \ e \in \{0,1\}^d \setminus \{0\}, \ j \in \mathbb{Z}, \ k \in \mathbb{Z}^d\},$$

we obtain the d-dimensional wavelet system

$$(3.5) \qquad \Psi := \{\psi_{\lambda,p} : \ \lambda \in \nabla\}.$$

Finally, we define the fixed level sets

$$\nabla_j := \{\lambda = (e,j,k) \in \nabla : \ |\lambda| = j\},$$

where we use the shorthand notation $|\lambda| := |(e,j,k)| := j$.

THEOREM 3.6 (Characterization [3,17]). *Let φ satisfy (W1) for some integrability parameters p', p'', (W2) for order L and (W3) with smoothness order s for primary parameter $0 < p \leq p'$ and any secondary parameter $0 < q \leq \infty$.*

For $\alpha < \min\{s, L\}$, assume additionally either
(i) $X = L^p(\mathbb{R}^d)$, $0 < p \leq \infty$ and $\alpha > d(1/p - 1/p')$,
(ii) or $X = H_p(\mathbb{R}^d)$, $0 < p \leq 1$ and (H1).
(iii) or $X = L^p(\mathbb{R}^d)$, $1 \leq p \leq \infty$, φ bounded and compactly supported.

Then, if $f = \sum_{\lambda \in \nabla} c_{\lambda,p} \psi_{\lambda,p}$ is the wavelet decomposition of f, we have the norm equivalence

$$|f|_{B_q^\alpha(X)} \sim \begin{cases} \left(\sum_{j \in \mathbb{Z}} 2^{j\alpha q} \left(\sum_{\lambda \in \nabla_j} |c_{\lambda,p}|^p \right)^{q/p} \right)^{1/q}, & 0 < q < \infty, \\ \sup_{j \in \mathbb{Z}} 2^{j\alpha} \left(\sum_{\lambda \in \nabla_j} |c_{\lambda,p}|^p \right)^{1/p}, & q = \infty. \end{cases}$$

Note the renormalization relationship

$$(3.6) \qquad c_{\lambda,p} = 2^{-|\lambda|d(1/p - 1/q)} c_{\lambda,q}, \quad 0 < p, q \leq \infty$$

and in particular $c_{\lambda,p} = 2^{-|\lambda|d(1/p-1/2)} c_{\lambda,2}$ with $c_{\lambda,2} = \langle f, \tilde{\psi}_{\lambda,2} \rangle_{L^2}$, the inner product with the L^2-scaled dual wavelet $\tilde{\psi}_{\lambda,2}$.

The above characterization implies the following approximation rates for best N-term wavelet approximations.

THEOREM 3.7 (*N*-term Approximation [3, 8, 17]). *Let $0 < p \le \infty$ and let Ψ be a wavelet system satisfying (W1)–(W3) with regularity s for Besov primary parameter $\tau > 0$, reproduction order L, $0 < \alpha < \min\{s, L\}$ and assume either*

(i) $X = L^p(\mathbb{R}^d)$, $1 < p < \infty$ *and* $d(1/p - 1/p') < \alpha$,

(ii) *or* $X = L^p(\mathbb{R}^d)$, $1 < p < \infty$, φ *is bounded and compactly supported,*

(iii) *or* $X = L^\infty(\mathbb{R}^d)$, *(A1) and* $d \le \alpha \le \min\{s, L\}$,

(iv) *or* $X = H_p(\mathbb{R}^d)$, $0 < p \le 1$ *and (H1).*

Define the set of N-term wavelet expansions as

$$\mathcal{W}_N := \left\{ \sum_{\lambda \in \Lambda} c_{\lambda, p} \psi_{\lambda, p} : \Lambda \subset \nabla, \ \#\Lambda \le N \right\}.$$

Then, for $\alpha/d \ge 1/\tau - 1/p$, $0 < q \le \tau \le p$ and any $f \in B_q^\alpha(Y^\tau)$, it holds

$$E(f, \mathcal{W}_N)_X \lesssim N^{-\alpha/d} |f|_{B_q^\alpha(Y^\tau)},$$

where $Y^\tau := H_\tau(\mathbb{R}^d)$ for $0 < p \le 1$, $Y^\tau := L^\tau(\mathbb{R}^d)$ for $1 < p \le \infty$ and for $p = \infty$ we additionally assume f is continuous.

4. Optimal ReLU Approximation of Smoothness Classes. With the results from Section 2 and Section 3 we have all the tools necessary to derive approximation rates for arbitrary Besov functions. As was reviewed in Section 3, Besov spaces can be characterized by the decay of the wavelet coefficients, and N-term approximations achieve optimal approximation rates for Besov functions.

In this section, we show that a RePU network can reproduce an N-term wavelet expansion with $\mathcal{O}(N)$ complexity. More importantly, we also show that a ReLU network can approximate an N-term wavelet expansion with $\mathcal{O}(N \log(\varepsilon^{-1}))$ complexity, where $\varepsilon > 0$ is the related approximation accuracy. Together with a stability estimate, this will imply RePU and ReLU networks can approximate Besov functions of arbitrary smoothness order with optimal or near to optimal rate, respectively.

LEMMA 4.1 (Wavelet System Complexity). *Let Ψ be a wavelet system as defined in (3.5), with the one-dimensional scaling function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfying (P1) – (P2). Then, for $X = L^p(\mathbb{R}^d)$, $0 < p \le \infty$,*

(i) *for $r \ge 2$, there exists a constant $C > 0$ depending on r, polynomial reproduction order L, dimension d, support of φ, ψ and $\|\varphi\|_{L^\infty(\mathbb{R})}$, $\|\psi\|_{L^\infty(\mathbb{R})}$, such that for any $\psi_\lambda \in \Psi$ and any $\varepsilon > 0$, there exists a RePU network $\Phi_\lambda^\varepsilon \in \mathsf{RePU}_C^{r, d, 1}$ with the same support as ψ_λ such that*

$$\|\psi_\lambda - \Phi_\lambda^\varepsilon\|_{L^p(\mathbb{R}^d)} \le \varepsilon,$$

where the complexity of the network Φ_λ^ε is independent of ε, the depth is at most logarithmic in L and d.

(ii) *For $r = 1$, there exists a constant $C > 0$ with dependencies as above, such that for any $\psi_\lambda \in \Psi$ and any $0 < \varepsilon < 1$, there exists a ReLU network $\Phi_\lambda^\varepsilon \in \mathsf{ReLU}_{C(1+\log(\varepsilon^{-1}))}^{d, 1}$ with the same support as ψ_λ such that*

$$\|\psi_\lambda - \Phi_\lambda^\varepsilon\|_{L^p(\mathbb{R}^d)} \le \varepsilon,$$

where the depth of the network Φ_λ^ε is at most logarithmic in ε^{-1}, log-linear in L and log-linear in d.

Proof. We detail the proof for the case of ReLU networks (ii) as the proof for RePU networks is analogous and more straightforward. In the following we will frequently use the triangle inequality, i.e., assuming $p \geq 1$. For $0 < p < 1$, $\|\cdot\|_{L^p(\mathbb{R}^d)}$ is only a quasi-norm, i.e., the right-hand-side of the triangle inequality is to be multiplied by a constant, and the corresponding complexities are to be adjusted accordingly.

First, due to (P1) – (P2), the mother wavelet $\psi : \mathbb{R} \to \mathbb{R}$ is a compactly supported piecewise polynomial. By Theorem 2.2, for any $0 < \delta < 1$, there exists a ReLU network $\Phi_\psi^\delta \in \mathsf{ReLU}_{C(1+\log(\delta^{-1}))}^{1,1}$ with the same support as ψ such that $\|\psi - \mathcal{R}(\Phi_\psi^\delta)\|_{L^\infty(\mathbb{R})} \leq \delta$. The depth of Φ_ψ^δ is at most logarithmic in δ^{-1} and log-linear in the polynomial degree of ψ. A similar conclusion holds for the scaling function $\varphi : \mathbb{R} \to \mathbb{R}$. For RePU networks, Φ_ψ^δ has complexity independent of δ, with depth at most logarithmic in the polynomial degree of ψ.

Second, recall the tensor product wavelet from (3.4): $\psi^e(x) := \psi^{e_1}(x_1) \cdots \psi^{e_d}(x_d)$, where for each component either $\psi^{e_\nu} = \psi^1 = \psi$ or $\psi^{e_\nu} = \psi^0 = \varphi$. Let Φ_ψ^δ, $\Phi_\varphi^\delta \in \mathsf{ReLU}_{C(1+\log(\delta^{-1}))}^{1,1}$ be the ReLU networks as above with $C = \mathcal{O}(L\log(L))$, approximating ψ and φ, respectively, with accuracy $\delta > 0$ to be specified later. Then, we form the tuple $\Phi_\times^\delta \in \mathsf{ReLU}_{n_\delta}^{d,d}$ as in Theorem 2.4 (iii)

$$\mathcal{R}(\Phi_\times^\delta) = (\mathcal{R}(\Phi_{\psi^{e_1}}^\delta), \ldots, \mathcal{R}(\Phi_{\psi^{e_d}}^\delta)),$$

with $n_\delta := dC(1 + \log(\delta^{-1}))$ and where each $\Phi_{\psi^{e_\nu}}$ has either the depth of Φ_φ or Φ_ψ – both of the same order as discussed above. The depth of $\Phi_\times^\delta$ is then the same as the maximal depth between Φ_φ^δ and Φ_ψ^δ – logarithmic in δ^{-1}.

Next, for some different accuracy $\eta > 0$, we construct an approximate multiplication network $\Phi_M^\eta \in \mathsf{ReLU}_{n_\eta}^{d,1}$ with $n_\eta := Cd\log(dK^d\eta^{-1})$ as in Theorem 2.1 (ii), where $K := \max\{\|\varphi\|_{L^\infty(\mathbb{R})}, \|\psi\|_{L^\infty(\mathbb{R})}\} + \delta$. This choice of K is justified by

$$(4.1) \qquad \left\|\mathcal{R}(\Phi_{\psi^{e_\nu}}^\delta)\right\|_{L^\infty(\mathbb{R}^d)} \leq \delta + \|\psi^{e_\nu}\|_{L^\infty(\mathbb{R})} \leq K.$$

Our final approximation $\Phi_\phi^\varepsilon \in \mathsf{ReLU}_{n_{\delta,\eta}}^{d,1}$ is defined by

$$(4.2) \qquad \mathcal{R}(\Phi_{\psi^e}^\varepsilon) := \mathcal{R}(\Phi_M^\eta) \circ \mathcal{R}(\Phi_\times^\delta),$$

where, according to Theorem 2.4 (iv), $n_{\delta,\eta} = dC(1 + \log(\delta^{-1})) + Cd\log(dK^d\eta^{-1})$, where as above $C = \mathcal{O}(L\log(L))$. The depth of $\Phi_{\psi^e}^\varepsilon$ is the sum of the depths of Φ_M^η and $\Phi_\times^\delta$, i.e., at most a multiple of $(1 + \log(d))\log(dK^d\eta^{-1}) + L\log(L)\log(\delta^{-1})$.

We estimate the resulting error from which it will be clear how to choose $\delta, \eta > 0$ and the resulting cost $n_{\delta,\eta}$. We introduce the auxiliary approximation $\mathcal{R}(\tilde{\Phi}) := M_d \circ \mathcal{R}(\Phi_\times^\delta)$ and the notation $\psi_\delta^{e_\nu} := \mathcal{R}(\Phi_{\psi^{e_\nu}}^\delta)$. Then,

$$(4.3) \qquad \left\|\psi^e - \mathcal{R}(\Phi_{\psi^e}^\varepsilon)\right\|_{L^p(\mathbb{R}^d)} \leq \left\|\psi^e - \mathcal{R}(\tilde{\Phi})\right\|_{L^p(\mathbb{R}^d)} + \left\|\mathcal{R}(\tilde{\Phi}) - \mathcal{R}(\Phi_{\psi^e}^\varepsilon)\right\|_{L^p(\mathbb{R}^d)}.$$

With $S := |\operatorname{supp}(\varphi) \cup \operatorname{supp}(\psi)|$, for the second term we apply Theorem 2.1 (ii) and obtain

$$\left\|\mathcal{R}(\tilde{\Phi}) - \mathcal{R}(\Phi_{\psi^e}^\varepsilon)\right\|_{L^p(\mathbb{R}^d)} \leq \left\|\mathcal{R}(\tilde{\Phi}) - \mathcal{R}(\Phi_{\psi^e}^\varepsilon)\right\|_{L^\infty(\mathbb{R}^d)} \left(\int_{\operatorname{supp}(\mathcal{R}(\tilde{\Phi})) \cup \operatorname{supp}(\mathcal{R}(\Phi_{\psi^e}^\varepsilon))}\right)^{1/p}$$

$$(4.4) \qquad\qquad\qquad \leq \eta S^{d/p}.$$

For the first term, we can write

$$\psi^e - \mathcal{R}(\tilde{\Phi}) = (\psi^{e_\nu} - \psi_\delta^{e_\nu}) \otimes \psi^{e_2} \otimes \cdots \otimes \psi^{e_d} + \psi_\delta^{e_1} \otimes (\psi^{e_2} - \psi_\delta^{e_2}) \otimes \psi_\delta^{e_3} \otimes \cdots \otimes \psi^{e_d}$$
$$+ \psi_\delta^{e_1} \otimes \cdots \otimes \psi_\delta^{e_{d-1}} \otimes (\psi^{e_d} - \psi_\delta^{e_d}).$$

Thus, for estimating (4.3), by a triangle inequality

$$(4.5) \qquad \left\| \psi^e - \mathcal{R}(\tilde{\Phi}) \right\|_{L^p(\mathbb{R}^d)} \leq dK^{d-1}\delta = d(\max\{\|\varphi\|_{L^\infty(\mathbb{R})}, \|\psi\|_{L^\infty(\mathbb{R})}\} + \delta)^{d-1}\delta.$$

From (4.4) and (4.5), we set $\eta := S^{-d/p}\varepsilon/2$ and

$$\delta := \varepsilon(\max\{1, \|\varphi\|_{L^\infty(\mathbb{R})}, \|\psi\|_{L^\infty(\mathbb{R})}\})^{1-d}/(d2^d) < 1.$$

With this δ, we can estimate

$$dK^{d-1}\delta \leq d(\max\{\|\varphi\|_{L^\infty(\mathbb{R})}, \|\psi\|_{L^\infty(\mathbb{R})}\})^{d-1}(\delta^{1/(d-1)} + \delta^{d/(d-1)})^{d-1}$$
$$\leq d(\max\{\|\varphi\|_{L^\infty(\mathbb{R})}, \|\psi\|_{L^\infty(\mathbb{R})}\})^{d-1}(2\delta^{1/(d-1)})^{d-1}$$
$$= d(\max\{\|\varphi\|_{L^\infty(\mathbb{R})}, \|\psi\|_{L^\infty(\mathbb{R})}\})^{d-1}2^{d-1}\delta \leq \varepsilon/2,$$

Thus, overall we obtain for (4.3)

$$\left\| \psi^e - \mathcal{R}(\Phi_{\psi^e}^\varepsilon) \right\|_{L^p(\mathbb{R}^d)} \leq \varepsilon,$$

with number of nonzero weights

$$n_\varepsilon := n_{\delta,\eta} = dC(1 + \log(\delta^{-1})) + Cd\log(dK^d\eta^{-1}) =$$
$$\mathcal{O}\left(dL\log(L)[d\log(1 + \max\{\|\varphi\|_{L^\infty}, \|\psi\|_{L^\infty}\}) + \log(\varepsilon^{-1}) + d\log(S) + \log(d)] \right),$$

and, according to (4.2), depth at most a multiple of

$$1 + \log(d)\log(dK^d\eta^{-1}) + L\log(L)\log(\delta^{-1}) =$$
$$\mathcal{O}\left(\log(d)L\log(L)[d\log(1 + \max\{\|\varphi\|_{L^\infty}, \|\psi\|_{L^\infty}\}) + \log(\varepsilon^{-1}) + d\log(S) + \log(d)] \right).$$

Finally, we define $\Phi_\lambda^\varepsilon \in \mathsf{ReLU}_{C(1+\log(\varepsilon^{-1}))}^{d,1}$ using Theorem 2.4 (i) and (v) such that for $\lambda = (e, j, k)$

$$\mathcal{R}(\Phi_\lambda^\varepsilon) = 2^{dj/p}\mathcal{R}(\Phi_{\psi^e}^\varepsilon) \circ D_k^{2^j}.$$

Note that the error bound for Φ_λ^ε remains unchanged due to the normalization constant $2^{dj/p}$.   □

With this we turn to direct estimates for networks.

LEMMA 4.2 (Direct Estimates RePU/ReLU). *Let $X = L^p(\Omega)$, $1 < p \leq \infty$, $f \in B_q^\alpha(L^\tau(\Omega))$ with $\alpha, \tau, q > 0$ and*

$$\alpha/d \geq 1/\tau - 1/p, \quad 0 < q \leq \tau \leq p,$$

and assume $\Omega \subset \mathbb{R}^d$ is an (ε, δ)-domain for $\tau \leq 1$ and a strong locally Lipschitz domain for $\tau > 1$. For $p = \infty$, assume additionally f is continuous with the convention $1/\infty = 0$. Then,

(i) for $r \geq 2$, there exists a constant $C > 0$ such that

$$E(f, \mathsf{RePU}_n^{r,d,1})_{L^p(\Omega)} \leq Cn^{-\alpha/d}\|f\|_{B_q^\alpha(L^\tau(\Omega))},$$

for all $n \in \mathbb{N}$. The constant C depends on r, α, τ, p, d and Ω. The networks realizing these approximations have depth at most logarithmic in α and d.

(ii) For $r = 1$, there exists a constant $C > 0$ such that

$$E(f, \mathsf{ReLU}_n^{d,1})_{L^p(\Omega)} \leq Cn^{-\bar\alpha/d}\|f\|_{B_q^\alpha(L^\tau(\Omega))},$$

for all $n \in \mathbb{N}$ and for any $0 < \bar\alpha < \alpha$. The constant C depends α, $\bar\alpha$, τ, p, d and Ω. The networks realizing these approximations have depth at most logarithmic in n, log-linear in α and log-linear in d.

Proof. Consider a wavelet system Ψ that satisfies (P1) – (P2), (W1) – (W3) and $(A1)$ for $p = \infty$. Such a wavelet system can be constructed, e.g., as in [4] and we can use Ψ for N-term approximations as in Theorem 3.7. We detail the proof for ReLU networks in part (ii), part (i) follows analogously with fewer technicalities.

Let $\Omega = \mathbb{R}^d$ and $f_N := \sum_{\lambda \in \Lambda_N} c_{\lambda,p}(f)\psi_\lambda$ with $\#\Lambda_N \leq N$ be a best N-term wavelet approximation to f. For $\varepsilon > 0$ to be specified later, let $\Phi_\lambda^\varepsilon \in \mathsf{ReLU}_{C(1+\log(\varepsilon^{-1}))}^{d,1}$ be the ReLU ε-approximation of ψ_λ from Lemma 4.1. Let $\Phi_N^\varepsilon \in \mathsf{ReLU}_{CN(1+\log(\varepsilon^{-1}))}^{d,1}$ be a sum network implemented as in Theorem 2.4 (ii) such that

$$\mathcal{R}(\Phi_N^\varepsilon) = \sum_{\lambda \in \Lambda_N} c_{\lambda,p}(f)\mathcal{R}(\Phi_\lambda^\varepsilon).$$

From Theorem 2.4 (ii), the depth of Φ_N^ε is the same as that of the Φ_λ^ε's. Then,

$$\|f_N - \mathcal{R}(\Phi_N^\varepsilon)\|_{L^p(\Omega)} \leq \varepsilon \sum_{\lambda \in \Lambda_N} |c_{\lambda,p}(f)|.$$

The sum of the coefficients is bounded by a Besov semi-norm of f as we show next.

We use the renormalization relationship from (3.6). First, we renormalize the coefficients in L^τ, multiply by one and split the sum in two

$$\sum_{\lambda \in \Lambda_N} |c_{\lambda,p}(f)| = \sum_{\lambda \in \Lambda_N} |c_{\lambda,\tau}(f)|2^{\alpha|\lambda|}2^{-|\lambda|(\alpha - d[1/\tau - 1/p])}$$

$$= \sum_{\lambda \in \Lambda_N, \, |\lambda| \geq 0} |c_{\lambda,\tau}(f)|2^{\alpha|\lambda|}2^{-|\lambda|(\alpha - d[1/\tau - 1/p])}$$

$$+ \sum_{\lambda \in \Lambda_N, \, |\lambda| < 0} |c_{\lambda,\tau}(f)|2^{\alpha|\lambda|}2^{-|\lambda|(\alpha - d[1/\tau - 1/p])}.$$

We abbreviate $\alpha^* := d[1/\tau - 1/p]$, where by assumption $0 < \alpha^* \leq \alpha$. Next, if $\tau < 1$, we can estimate the first summand as

$$\sum_{\lambda \in \Lambda_N, \, |\lambda| \geq 0} |c_{\lambda,\tau}(f)|2^{\alpha|\lambda|}2^{-|\lambda|(\alpha - d[1/\tau - 1/p])} = \sum_{\lambda \in \Lambda_N, \, |\lambda| \geq 0} |c_{\lambda,\tau}(f)|2^{\alpha^*|\lambda|}$$

$$\leq \left(\sum_{\lambda \in \Lambda_N} |c_{\lambda,\tau}(f)|^\tau 2^{\alpha^*|\lambda|\tau}\right)^{1/\tau},$$

and the second summand as

$$\sum_{\lambda \in \Lambda_N, \, |\lambda|<0} |c_{\lambda,\tau}(f)| 2^{\alpha|\lambda|} 2^{-|\lambda|(\alpha-d[1/\tau-1/p])} \leq \sum_{\lambda \in \Lambda_N, \, |\lambda|<0} |c_{\lambda,\tau}(f)| 2^{-|\lambda|\alpha}$$

$$\leq \left(\sum_{\lambda \in \Lambda_N} |c_{\lambda,\tau}(f)|^\tau 2^{|\lambda|\alpha\tau} \right)^{1/\tau},$$

and thus overall

$$\sum_{\lambda \in \Lambda_N} |c_{\lambda,p}(f)| \leq \left(\sum_{\lambda \in \Lambda_N} |c_{\lambda,\tau}(f)|^\tau 2^{\alpha^*|\lambda|\tau} \right)^{1/\tau} + \left(\sum_{\lambda \in \Lambda_N} |c_{\lambda,\tau}(f)|^\tau 2^{\alpha|\lambda|\tau} \right)^{1/\tau}$$

$$\leq 2 \, |f|_{B_\tau^\alpha(L^\tau(\mathbb{R}^d))},$$

where the last inequality is due to the characterization of the Besov semi-norm from Theorem 3.6, the fact that $\alpha^* \leq \alpha$ and the definition of the Besov semi-norm (see Subsection 3.1).

If $\tau \geq 1$, we apply Hölder with $1 \leq \bar{\tau} \leq \infty$ such that $1/\tau + 1/\bar{\tau} = 1$, the definition and characterization of the Besov semi-norm together with $\alpha^* \leq \alpha$, and obtain

$$\sum_{\lambda \in \Lambda_N} |c_{\lambda,p}(f)| \leq 2N^{1/\bar{\tau}} |f|_{B_\tau^\alpha(L^\tau(\mathbb{R}^d))}.$$

Thus, we set either $\varepsilon := N^{-\alpha/d}/2$ for $\tau < 1$ or $\varepsilon := N^{-\alpha/d-1/\bar{\tau}}/2$ for $\tau \geq 1$ and obtain together with Theorem 3.7

$$\|f - \mathcal{R}(\Phi_N^\varepsilon)\|_{L^p(\mathbb{R}^d)} \lesssim N^{-\alpha/d} |f|_{B_\tau^\alpha(L^\tau(\mathbb{R}^d))}.$$

Due to the definition of the Besov semi-norm, it is straightforward to extend this to Besov semi-norms $|f|_{B_q^\alpha(L^\tau(\mathbb{R}^d))}$ for any $0 < q \leq \tau$ (see also the discussion in Subsection 3.1). The complexity of this network can be bounded by for the case $\tau \geq 1$ as

$$n := CN(1 + \log(\varepsilon^{-1})) = CN(1 + [\alpha/d + 1/\bar{\tau}] \log(N)) \lesssim N^{1+\delta},$$

for any $\delta > 0$, or, equivalently,

$$\|f - \mathcal{R}(\Phi_N)\|_{L^p(\mathbb{R}^d)} \lesssim n^{-\bar{\alpha}/d} |f|_{B_q^\alpha(L^\tau(\mathbb{R}^d))},$$

for any $0 < \bar{\alpha} < \alpha$. The bound for the case $\tau < 1$ is similar, omitting $\bar{\tau}$. This shows the statement for $\Omega = \mathbb{R}^d$.

For $\Omega \subset \mathbb{R}^d$ an (ε, δ)-domain for $0 < \tau \leq 1$ or a locally Lipschitz domain for $1 < \tau \leq \infty$, we use the extension operator from Theorem 3.1 to obtain for any $f \in B_q^\alpha(L^\tau(\Omega))$

$$E(f, \mathsf{ReLU}_n^{d,1})_{L^p(\Omega)} \leq E(\mathcal{E}f, \mathsf{ReLU}_n^{d,1})_{L^p(\mathbb{R}^d)} \lesssim n^{-\bar{\alpha}/d} |\mathcal{E}f|_{B_q^\alpha(L^\tau(\mathbb{R}^d))}$$

$$\lesssim n^{-\bar{\alpha}/d} \|f\|_{B_q^\alpha(L^\tau(\Omega))}. \qquad \square$$

Finally, the direct estimates above immediately imply the main result of this work.

THEOREM 4.3 (Direct Embeddings). *Let $X = L^p(\Omega)$, $1 < p \leq \infty$, $\alpha, \tau, q > 0$ and*

$$\alpha/d \geq 1/\tau - 1/p, \quad 0 < q \leq \tau \leq p,$$

and assume $\Omega \subset \mathbb{R}^d$ is an (ε, δ)-domain for $\tau \leq 1$ and a strong locally Lipschitz domain for $\tau > 1$. Then,

(i) for $r \geq 2$ and $0 < \theta < \alpha$, $0 < \bar{q} \leq \infty$, the following embeddings hold

$$B_q^\alpha(L^\tau(\Omega)) \hookrightarrow A_\infty^{\alpha/d}(L^p(\Omega), \mathsf{RePU}^{r,d,1}),$$

$$(L^p(\Omega), B_q^\alpha(L^\tau(\Omega)))_{\theta/\alpha, \bar{q}} \hookrightarrow A_{\bar{q}}^{\theta/d}(L^p(\Omega), \mathsf{RePU}^{r,d,1}).$$

(ii) For $r = 1$ and $0 < \theta < \bar{\alpha}$, $0 < \bar{q} \leq \infty$, $0 < \bar{\alpha} < \alpha$, the following embeddings hold

$$B_q^\alpha(L^\tau(\Omega)) \hookrightarrow A_\infty^{\bar{\alpha}/d}(L^p(\Omega), \mathsf{ReLU}^{d,1}),$$

$$(L^p(\Omega), B_q^\alpha(L^\tau(\Omega)))_{\theta/\bar{\alpha}, \bar{q}} \hookrightarrow A_{\bar{q}}^{\theta/d}(L^p(\Omega), \mathsf{ReLU}^{d,1}).$$

Remark 4.4 (Hardy Spaces $H_p(\mathbb{R}^d)$ and $0 < p \leq 1$). Theorem 3.6 provides best N-term approximation rates for Besov spaces in the Hardy H_p-norm for $0 < p \leq 1$. ReLU networks can reproduce piecewise linear one-dimensional wavelet systems *exactly* and hence all the embeddings from Theorem 4.3 hold for $r = 1$ and the homogeneous Besov space $B_q^\alpha(H_\tau(\mathbb{R}))$ instead of $B_q^\alpha(L^\tau(\Omega))$ for $0 < p \leq 1$ and $1/\tau - 1/p \leq \alpha < 2$.

For more general results one would require estimates as in Lemma 4.1 in the H_p-norm. However, compactly supported, continuous, piecewise polynomials are not necessarily in H_p. In order to ensure $\mathcal{R}(\Phi_\lambda^\varepsilon)$ is in H_p, we require additionally that the approximands $\mathcal{R}(\Phi_\lambda^\varepsilon)$ have $d(1/p - 1)$ vanishing moments for any ε and λ.

Furthermore, to extend results from $\mathbb{R}^d$ to general domains Ω, one requires constructing extension operators bounded in the homogeneous Besov norm. A detailed investigation of the case of Hardy spaces is thus beyond the scope of this work.

REFERENCES

[1] R. ADAMS AND J. J. F. FOURNIER, *Sobolev Spaces*, Elsevier LTD, Oxford, June 2003.

[2] H. BÖLCSKEI, P. GROHS, G. KUTYNIOK, AND P. PETERSEN, *Optimal Approximation with Sparsely Connected Deep Neural Networks*, SIAM Journal on Mathematics of Data Science, 1 (2019), pp. 8–45.

[3] A. COHEN, *Numerical Analysis of Wavelet Methods*, Elsevier, Amsterdam Boston, 2003.

[4] A. COHEN, I. DAUBECHIES, AND J.-C. FEAUVEAU, *Biorthogonal Bases of Compactly Supported Wavelets*, Communications on Pure and Applied Mathematics, 45 (1992), pp. 485–560.

[5] I. DAUBECHIES, R. DEVORE, S. FOUCART, B. HANIN, AND G. PETROVA, *Nonlinear Approximation and (Deep) ReLU Networks*, arXiv e-prints, (2019), arXiv:1905.02199, p. arXiv:1905.02199.

[6] R. A. DEVORE, *Nonlinear Approximation*, Acta Numerica, 7 (1998), p. 51–150.

[7] R. A. DEVORE, R. HOWARD, AND C. MICCHELLI, *Optimal Nonlinear Approximation*, Manuscripta mathematica, 63 (1989), pp. 469–478.

[8] R. A. DEVORE, P. PETRUSHEV, AND X. M. YU, *Nonlinear Wavelet Approximation in the Space $C(R^d)$*, in Progress in Approximation Theory, A. A. Gonchar and E. B. Saff, eds., New York, NY, 1992, Springer New York, pp. 261–283.

[9] R. A. DEVORE AND R. C. SHARPLEY, *Maximal Functions Measuring Smoothness*, Memoirs of the American Mathematical Society, 47 (1984), pp. 0–0, https://doi.org/10.1090/memo/0293.

[10] R. A. DEVORE AND R. C. SHARPLEY, *Besov Spaces on Domains in R^d*, Transactions of the American Mathematical Society, 335 (1993), pp. 843–864.

[11] G. C. Donovan, J. S. Geronimo, and D. P. Hardin, *Orthogonal Polynomials and the Construction of Piecewise Polynomial Smooth Wavelets*, SIAM Journal on Mathematical Analysis, 30 (1999), pp. 1029–1056.

[12] D. Dung and V. Q. Thanh, *On nonlinear n-widths*, Proceedings of the American Mathematical Society, 124 (1996), pp. 2757–2765, https://doi.org/10.1090/s0002-9939-96-03337-0.

[13] C. Fefferman and E. M. Stein, H^p *Spaces of Several Variables*, Acta Mathematica, 129 (1972), pp. 137 – 193, https://doi.org/10.1007/BF02392215, https://doi.org/10.1007/BF02392215.

[14] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender, *Approximation Spaces of Deep Neural Networks*, arXiv e-prints, (2019), arXiv:1905.01208, p. arXiv:1905.01208.

[15] I. Gühring, M. Raslan, and G. Kutyniok, *Expressivity of Deep Neural Networks*, arXiv e-prints, (2020), arXiv:2007.04759, p. arXiv:2007.04759.

[16] P. W. Jones, *Quasiconformal Mappings and Extendability of Functions in Sobolev Spaces*, Acta Mathematica, 147 (1981), pp. 71–88, https://doi.org/10.1007/bf02392869.

[17] G. C. Kyriazis, *Wavelet Coefficients Measuring Smoothness in $H_p(R^d)$*, Applied and Computational Harmonic Analysis, 3 (1996), pp. 100–119.

[18] J. A. A. Opschoor, P. C. Petersen, and C. Schwab, *Deep ReLU Networks and High-order Finite Element Methods*, Analysis and Applications, (2020).

[19] J. A. A. Opschoor, C. Schwab, and J. Zech, *Exponential ReLU DNN Expression of Holomorphic Maps in High Dimension*, Tech. Report 2019-35, Seminar for Applied Mathematics, ETH Zürich, Switzerland, 2019.

[20] P. Petersen and F. Voigtlaender, *Optimal Approximation of Piecewise Smooth Functions Using Deep ReLU Neural Networks*, Neural Networks, 108 (2018), pp. 296 – 330.

[21] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, 1970, http://www.jstor.org/stable/j.ctt1bpmb07.

[22] T. Suzuki, *Adaptivity of Deep ReLU Network for Learning in Besov and Mixed Smooth Besov Spaces: Optimal Rate and Curse of Dimensionality*, in International Conference on Learning Representations, 2019.

[23] D. Yarotsky, *Error Bounds for Approximations with Deep ReLU Networks*, Neural Networks, 94 (2017), pp. 103 – 114.