OPTIMAL APPROXIMATION COMPLEXITY OF HIGH-DIMENSIONAL FUNCTIONS WITH NEURAL NETWORKS

VINCENT P.H. GOVERSE*, JAD HAMDAN†, AND JARED TANNER†

* *Imperial College London*

† *University of Oxford*

Abstract. We investigate properties of neural networks that use both ReLU and x^2 as activation functions and build upon previous results to show that both analytic functions and functions in Sobolev spaces can be approximated by such networks of constant depth to arbitrary accuracy, demonstrating optimal order approximation rates across all nonlinear approximators, including standard ReLU networks. We then show how to leverage low local dimensionality in some contexts to overcome the curse of dimensionality, obtaining approximation rates that are optimal for unknown lower-dimensional subspaces.

1. Introduction

The number of parameters needed to approximate smooth high-dimensional functions, $W^{n,\infty}([0,1]^d)$, within a prescribed ϵ accuracy in the ℓ_∞ norm was lower bounded by [6] to have a dependence on ϵ that is proportional to $\epsilon^{-d/n}$. [16] has subsequently shown that a simple feedforward neural network with $\mathrm{ReLU}(x) := \max\{0, x\}$ nonlinear activation is nearly optimal in terms of the number of parameters needed, requiring only $c(n,d) = \epsilon^{-d/n}\log(1/\epsilon)$ parameters[1], see [16][Theorem 1]. Subsequently, [4] reduced the number of parameters needed by a feedforward neural network to achieve ϵ accuracy to being proportional to $c(n,d) = \epsilon^{-d/n}\log(\log(1/\epsilon))$ by using trainable rational function as nonlinear activations, see [4][Theorem 4].

Here we further adapt the proof by Yarotsky to achieve the optimal dependence of $\epsilon^{-d/n}$ proven by [6], using a feedforward network that makes use of *two* nonlinear activations (henceforth referred to as *bi-activation* networks). Specifically, we allow some layers to use the ReLU nonlinear activation to localize $f(x)$ through a partition of unity, and the quadratic activation x^2 to allow for efficient computation of localized high degree polynomial approximations.

Specifically, following the notation of [6] and [16], we consider nonlinear approximation methods $M_p(a)$ that have a continuous dependence[2] on the p parameters $a \in \mathbb{R}^p$ and which approximate high dimensional functions $f(\cdot)$ within the unit ball of the Sobolev space $W^{n,\infty}([0,1]^d)$,

$$||f||_{W^{n,\infty}([0,1]^d)} = \max_{\mathbf{n}\in|\mathbf{n}|\leq n} \mathrm{esssup}_{\mathbf{x}\in[0,1]^d} |D^{\mathbf{n}} f(\mathbf{x})| \tag{1.1}$$

2010 *Mathematics Subject Classification.* 37H05, 47B65, 60J05.

Key words and phrases. Machine Learning, Universal Approximation, bi-activation, Neural Networks .

[1]The function $c(n,d)$ depends on the smoothness, n, and the dimension of $f(x)$, but not on the desired accuracy ϵ.

[2]The continuous dependence of $M_n(a)$ on a is introduced in [6] to avoid space filling curves and can be viewed as ensuring the parameters a can be learned from a sufficiently near estimate; for details see [6].

where $\mathbf{n} = (n_1, ..., n_d) \in \{0, 1, ...\}^d$, $|\mathbf{n}| = \sum_i^d n_i$ and $D^{\mathbf{n}} f$ the respective weak derivative. The foundational lower bound on the number of elements in *any* nonlinear approximation method $M_p(a)$ that depends smoothly on $a \in \mathbb{R}^p$ is given in Theorem 1.1.

Theorem 1.1 (Optimal non-linear approximation lower bound, [6]). *For function $f(x)$ with $||f||_{W^{n,\infty}([0,1]^d)} \leq 1$, and $M_p(a)$ depending continuously on $a \in \mathbb{R}^p$, approximating $f(x)$ with bound*

$$\inf_{M_p(a), a} \max_{x \in [0,1]^d} |f(x) - M_p(a)(x)| \leq \epsilon$$

then necessarily $M_p(\cdot)$ has $p \geq C_1(d,n)\epsilon^{-d/n}$ where $C_1(d,n)$ may depend on d and n, but not on ϵ.

As a method to explain the value of depth in deep learning, [16] constructed a feed forward networks with is near optimal order number of parameters as a function of approximation accuracy ϵ. In particular,

Theorem 1.2 (Near optimal non-linear approximation with ReLU-networks, [16]). *For function $f(x)$ with $||f||_{W^{n,\infty}([0,1]^d)} \leq 1$, there exists $M_{\mathcal{C}_Y,\mathrm{ReLU}}(a)$ formed as a feed-forward network with at most $\mathcal{C}_Y = C_2(d,n)\epsilon^{-d/n}(1 + \log(1/\epsilon))$ elements $a \in \mathbb{R}^{\mathcal{C}_Y}$ for which*

$$\min_{a \in \mathbb{R}^{\mathcal{C}_Y}} \max_{x \in [0,1]^d} |f(x) - M_{\mathcal{C}_Y, ReLU}(a)(x)| \leq \epsilon$$

where $C_2(d,n)$ may depend on d and n, but not on ϵ.

The feed-forward network $M_{\mathcal{C}_Y,ReLU}$ constructed in [16] has hidden layers $h_{i+1} = \mathrm{ReLU}(W_i h_i + b_i)$ for $i = 0, \dots L$ with input $h_0 := x$, W_i being matrices of width bounded independent of ϵ, and depth $L \leq c(d,n)(\log(1/\epsilon) + 1)$. The feed-forward network is constructed analogously to the proof in [6] where there the input $x \in \mathbb{R}^d$ is partitioned into exponentially many localized portions, each of which then has a local polynomial constructed to approximate $f(\cdot)$. The ReLU nonlinear activation allows for partitions of the input space $[0,1]^d$ and the logarithmic depth is needed to construct high-degree local polynomial approximations using the saw-tooth functions developed by Telgarsky [15]; for details, see [16].

Our main contribution here is a feed-forward network $M_{\mathcal{C}_F,bi-\sigma}$ where the layers $h_{i+1} = \sigma_i(W_i h_i + b_i)$ have non-linear activations $\sigma_i(x)$ which are either $\mathrm{ReLU}(x)$ or x^2 depending on the layer. This choice of nonlinear activations is made to simplify the proof in [16] by retaining the ability to localize $\mathbb{R}^d$ while more efficiently computing higher-order polynomial functions with bounded depth L. Other choices of localizing and approximation activations are possible, see the details of the proof of Theorem 1.3.

Theorem 1.3 (Optimal approximation order bi-activation networks). *For function $f(x)$ with $||f||_{W^{n,\infty}([0,1]^d)} \leq 1$, there exists $M_{\mathcal{C}_F,bi-\sigma}(a)$ formed as a feed-forward network with $\mathcal{C}_F = C_3(d,n)\epsilon^{-d/n}$ elements $a \in \mathbb{R}^{\mathcal{C}_F}$ for which*

$$\min_{a \in \mathbb{R}^{\mathcal{C}_F}} \max_{x \in [0,1]^d} |f(x) - M_{\mathcal{C}_F, bi-\sigma}(a)(x)| \leq \epsilon$$

where $C_3(d,n)$ may depend on d and n, but not on ϵ.

The Proof of Theorem 1.3 is given in Section 2.1, making use of a key lemma from the proof of Theorem 1.2 by Yarotsky.

We further extend Theorem 1.3 in two separate directions, by considering $f(x)$ to be analytic or $f(x)$ to be contained on the union of $d_{\mathrm{eff}} < d$ dimensional canonical subspaces of $\mathbb{R}^d$.

Theorem 1.4 (Optimal approximation order bi-activation networks: Analytic functions)**.** *Let $f(x)$ be an analytic function on $[0,1]^d$, characterised* [1] *by*

$$\sup_{x\in[0,1]^d}\left|\frac{\partial^{\mathbf{n}} f}{\partial x^{\mathbf{n}}}(x)\right| \leq C_f^{|\mathbf{n}|+1}\mathbf{n}! \quad \textit{for all } n \tag{1.2}$$

where C_f depends on the particular choice of $f(x)$. Then for any d, and $\epsilon \in (0,1)$, there exists $M_{\mathcal{C}_A,bi-\sigma}(a)$ formed as a feed-forward network with $\mathcal{C}_A = C_4(d,C_f)\left((2\epsilon)^{\log^{-\frac{1}{2}}\left(\frac{2^d}{\epsilon}\right)}\log^{\frac{d}{2}}\left(\frac{1}{\epsilon}\right)\right)$ elements $a \in \mathbb{R}^{\mathcal{C}_A}$ for which

$$\min_{a\in\mathbb{R}^{\mathcal{C}_\mathcal{A}}}\max_{x\in[0,1]^d}|f(x) - M_{\mathcal{C}_A,bi-\sigma}(a)(x)| \leq \epsilon$$

where $C_4(d,C_f)$ does not depend on ϵ.

Theorem 1.4 differs from Theorem 1.3 primarily in the lack of dependence on smoothness n as the number of parameters $\mathcal{C}_\mathcal{A}$ needed in the network has been minimized over all admissible n. The consequence of choosing the optimal smoothness n is that the ϵ and d dependence of the number of parameters $\mathcal{C}_\mathcal{A}$ decreases from $(\epsilon^{-1/n})^d$ to predominantly $\log(1/\epsilon)^{d/2}$.

Next, for $d_{\text{eff}} < d$ we define the canonical subspace of $[0,1]^d$ of dimension d_{eff}; that is

$$x \in \chi^d_{d_{\text{eff}},e} := \{x \in [0,1]^d : \text{ with if } i \notin e, x_i = 0\}.$$

Where e is a subset of $\{1,\ldots,d\}$, with d_{eff} elements. $I^d_{d_{\text{eff}}}$ is the collections of all e. Then if $f(x)$ is nonzero on only one known subspace $\chi^d_{d_{\text{eff}},e}$ Lemma 2.2 holds. In the case that $f(x)$ is nonzero on the union of all $\binom{d}{d_{\text{eff}}}$ such subspaces

$$\bar{\chi}^d_{d_{\text{eff}}} := \bigcup_{e\in I}\chi^d_{d_{\text{eff}},e},$$

the number of parameters $\mathcal{C}_M$ needed to compute an ϵ approximation of $f(x)$ over one or all canonical subspaces is given by $\mathcal{C}_M = C_5(d,d_{\text{eff}},n)\epsilon^{-d_{\text{eff}}/n})$ (see Lemma 2.2 and Theorem 1.5).

Theorem 1.5 (Optimal approximation order bi-activation networks: low-dimensional subspaces)**.** *For function $f(x)$ with $||f||_{W^{n,\infty}([0,1]^d)} \leq 1$ where x is restricted to $\bar{\chi}^d_{d_{eff}}$, there exists $M_{\mathcal{C}_M,bi-\sigma}(a)$ formed as a feed-forward network with $\mathcal{C}_M = C_5(d,n)\epsilon^{-d_{eff}/n}$ elements $a \in \mathbb{R}^{\mathcal{C}_M}$ for which the error restricted on $\chi^d_{d_{eff}}$ is*

$$\min_{a\in\mathbb{R}^{\mathcal{C}_M}}\max_{x\in\bar{\chi}^d_{d_{eff}}}|f(x) - M_{\mathcal{C}_M,bi-\sigma}(a)(x)| \leq \epsilon$$

where $C_5(d,n)$ may depend on d and n, but not on ϵ.

This restricted subspace model is motivated by natural image inputs with prescribed compression on a known orthogonal basis, such as JPEG compression. This union of subspace model $\bar{\chi}^d_{d_eff}$ is also widely used in the theory of compressed sensing, see [7] and references therein, and has also been used to increase robustness against adversarial attacks on image classification by [9].

2. Approximation power of bi-activation networks

The proof of Theorem 1.3 being adapted from that 1.2 in [16], an understanding of the former is essential in order to explain the latter.

As mentioned previously, [16] first partitions the input $x \in [0,1]^d$ into exponentially many localized portions using a partition of unity $\{\phi_{\mathbf{m}}\}$, where each $\phi_{\mathbf{m}}$ is piecewise linear and expressible by a ReLU network with a constant number of parameters (see Proposition 1 in [16]). The aim is then to approximate the function f by Taylor polynomials locally, giving the following representation for an approximation of f.

Lemma 2.1 ([16])**.** *Let $\epsilon > 0$ be arbitrary and $f \in W^{n,\infty}([0,1]^d)$. Then there exists a function $\tilde{f}$ expressible as*

$$\tilde{f}(x) = \sum_{\boldsymbol{m}\in\{0,\dots,N\}^d} \sum_{\boldsymbol{n}:|\boldsymbol{n}|<n} a_{\boldsymbol{m},\boldsymbol{n}}\phi_{\boldsymbol{m}}(x)\left(x - \frac{\boldsymbol{m}}{N}\right)^{\boldsymbol{n}},$$

where $a_{m,n} \in \mathbb{R}$, $|a_{m,n}| \leq 1$, $\{\phi_{\boldsymbol{m}}\}_{\boldsymbol{m}\in\{0,1,\dots,N\}^d}$ is a partition of unity such that each $\phi_{\boldsymbol{m}}$ is given by a product of d piecewise linear univariate factors. Furthermore, $\tilde{f}$ is such that

$$|f(x) - \tilde{f}(x)| \leq \frac{2^d d^n}{n!}\left(\frac{1}{N}\right)^n \max_{\boldsymbol{n}:|\boldsymbol{n}|=n} \textit{ess sup}_{x\in[0,1]^d}|D^{\boldsymbol{n}}f(x)|. \tag{2.1}$$

The proof of this lemma is included in the appendix for completeness.

Showing that ReLU networks can approximate monomials (and, in turn, polynomials) would then complete the proof. Indeed, in Section 3.1 of [16], the author does so by first showing that $f(x) = x^2$ can be approximated by a ReLU network of complexity $O(\ln(1/\epsilon))$. Using the following identity to recover multiplication from squaring:

$$xy = \frac{1}{2}\big((x+y)^2 - x^2 - y^2\big) \tag{2.2}$$

the author then shows how a ReLU network of complexity $O(\ln(1/\epsilon))$ can in fact approximate terms of the form $\phi_{\mathbf{m}}(x)\left(x - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}}$.

Lastly, note that in lemma 2.1, $\tilde{f}$ is a linear combination of at most $d^n(N+1)^d$ such terms. N is a smoothness parameter that can be chosen so that the upper bound in (2.1) becomes $|f(x) - \tilde{f}(x)| < \epsilon$. In Yarotsky's case, this corresponds to choosing

$$N = N(\epsilon, d, n) = \left\lceil \left(\frac{n!}{2^d d^n}\epsilon\right)^{-1/n} \right\rceil, \tag{2.3}$$

which also yields

$$d^n(N+1)^d = d^n\left(\frac{n!}{2^d d^n}\epsilon\right)^{-d/n} = O(\epsilon^{-d/n}),$$

and the final ReLU network used approximate f therefore consists of $\mathcal{C}_Y = O(\epsilon^{-d/n}\ln(1/\epsilon))$ parameters due to the $\log(1/\epsilon)$ depth needed to approximate x^2 within ϵ using a ReLU network.

2.1. **Proof of Theorem 1.3, Optimal approximation order bi-activation networks.**

Proof of Theorem 1.3. Let $\tilde{f}$ be the approximation to f given by Lemma 2.1. Since f is in the unit-ball in $W^{n,\infty}$, $\max_{\mathbf{n}:|\mathbf{n}|=n} \text{ess sup}_{x\in[0,1]^d}|D^{\mathbf{n}}f(x)| \leq 1$. Choosing the same N as in 2.3, we find that $||f - \tilde{f}||_\infty \leq \epsilon$.

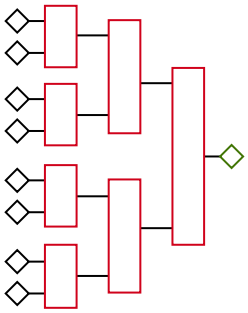

FIGURE 1. Multiplication of k elements (depicted in black) using $O(k)$ subnetworks (depicted in red) each of constant size (independent of d and n), giving a network of depth $O(\ln_2(k))$. Here, $k = 8$.

In contrast to ReLU networks, we claim that bi-activation networks can represent terms of the form $\phi_{\mathbf{m}}(x)(x - \mathbf{m}/N)^{\mathbf{n}}$ *exactly* using a constant number of trainable parameters. Indeed, each of these terms is itself a product of at most $d+n-1$ piecewise linear univariate factors: a product of d functions defining each $\phi_{\mathbf{m}}$ and at most $n-1$ functions $x_k - m_k/N$. These products can be implemented by a bi-activation network with a complexity of the order of $(n + d)$ and depth of the order of $\log_2(n + d)$ (in both cases, $O(1)$ with respect to ϵ), by repeatedly pairing up the terms and multiplying them in tournament fashion (see figure 1). The multiplication of two terms can be achieved by a bi-activation network of constant size using (2.2)[3].

Therefore, $\tilde{f}$ can be written by a bi-activation network $M_{\mathcal{C}_F, bi-\sigma}(a)$ with $\mathcal{C}_F = O(d^n(N+1)^d)$ parameters as follows. The network uses parallel subnetworks that each compute a term in the series defining $\tilde{f}$, and computes the final output by summing the outputs of these subnetworks, weighted with the appropriate $a_{\mathbf{m},\mathbf{n}}$. Since there are not more than $d^n(N + 1)^d$ subnetworks, $\mathcal{C}_F = C_3(d, n)d^n(N + 1)^d$ weights and computation units, for some constant $C_3(d, n)$. For our choice of N in (2.3) to achieve an ϵ accurate approximation, $\mathcal{C}_F = O(\epsilon^{-d/n})$. □

2.2. **Proof of Theorem 1.4, Optimal approximation order bi-activation networks: Analytic functions.**

Proof of Theorem 1.4. Once again, let $\tilde{f}$ be the approximation to f given by Lemma 2.1, noting that $f \in W^{n,\infty}([0, 1]^d)$ for all n as it is analytic. Then applying the bound on $|f(x) - \tilde{f}(x)|$ given by the same Lemma and the bound on smoothness for analytic functions (1.2), we find that

$$\begin{aligned}
|f(x) - \tilde{f}(x)| &\leq \frac{2^d d^n}{n!} \left(\frac{1}{N}\right)^n \max_{\mathbf{n}:|\mathbf{n}|=n} \operatorname{ess\,sup}_{x\in[0,1]^d} |D^{\mathbf{n}} f(x)| \\
&\leq \frac{2^d d^n}{n!} \left(\frac{1}{N}\right)^n C_f^{n+1} n! \\
&\leq 2^d d^n \left(\frac{C_f}{N}\right)^{n+1},
\end{aligned}$$

[3]More specifically, we can use a network with activation function x^2 which has one hidden layer. The inputs x and y connect fully to the hidden layer with three nodes, and weights $[0, 1], [1, 0]$ and $[1, 1]$. The three nodes are connected to the output with weight $[-1/2, -1/2, 1/2]$.

where C_f is a constant depending on f.

Notice that in this case the result holds for all n. This means that, when picking N, we can optimize over n to minimize the number of trainable parameters needed by our network. To begin with, choosing

$$N_1 = N(C_f, \epsilon, d, n) = \frac{1}{C}\left\lceil \left(\frac{\epsilon}{2^d d^n}\right)^{-1/(n+1)} \right\rceil \tag{2.4}$$

we get that $||f - \tilde{f}||_\infty \leq \epsilon$.

Arguing in the exact same manner as in the proof of Theorem 1.3, we know that $\tilde{f}$ can be written as a bi-activation neural network $M_{\mathcal{C}_A, bi-\sigma}(a)$. The total number of parameters $\mathcal{C}_A$ then needed by the network to represent $\tilde{f}$ is equal to

$$\mathcal{C}_A = C_4(C_f, n, d) d^n (N+1)^d \tag{2.5}$$

for some constant $C_4 = C_4(C_f, n, d)$ that does not depend on ϵ. Substituting the choice of N in (2.4) in (2.5) and minimizing over n, we find that $\mathcal{C}_\mathcal{A}$ is minimal for

$$n_{\min} = \sqrt{\frac{d\left(d\log(2) + \log\left(\frac{1}{\epsilon}\right)\right)}{\log(d)}}. \tag{2.6}$$

Substituting (2.6) and (2.4) into (2.5), gives us

$$\mathcal{C}_A = C_4 \cdot 2^{\frac{d^{3/2}\sqrt{\log(d)}}{\sqrt{d\log(2)+\log\left(\frac{1}{\epsilon}\right)}}} \epsilon^{\frac{\sqrt{d}\sqrt{\log(d)}}{\sqrt{d\log(2)+\log\left(\frac{1}{\epsilon}\right)}}} \log^{\frac{d}{2}}\left(\frac{1}{\epsilon}\right),$$

which grows as $\epsilon \to 0$ in the order of

$$(2\epsilon)^{\log^{-\frac{1}{2}}\left(\frac{2^d}{\epsilon}\right)} \log^{\frac{d}{2}}\left(\frac{1}{\epsilon}\right),$$

concluding the proof.

□

2.3. **Proof of Theorem 1.5, Optimal approximation order bi-activation networks: low-dimensional subspaces.** For clarity, first consider the simplest case of $x \in \chi^d_{d_{\text{eff}},e}$, for a known $e \in I$. Without loss of generality this can be the first d_{eff} dimensions of $\mathbb{R}^d$ being nonzero, that is $f \circ A(x) := f(Ax)$ for $A \in \mathbb{R}^{d \times d_{\text{eff}}}$ given by

$$A = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \\ 0 & 0 & & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}. \tag{2.7}$$

In this case we have $f|_{\chi^d_{d_{\text{eff}},e}} = f \circ A$. When we consider that the function we try to approximate is of the form $f \circ A$, we get the following lemma.

Lemma 2.2 (Optimal approximation order bi-activation networks: low-dimensional single subspace)**.** *For function $f(x)$ with $||f||_{W^{n,\infty}([0,1]^d)} \leq 1$ where x is restricted to a single*

canonical subspace $\chi^d_{d_{eff},e}$*, there exists* $M_{\mathcal{C}_S,bi-\sigma}(a)$ *formed as a feed-forward network with* $\mathcal{C}_S = C_6(d,n)\epsilon^{-d_{eff}/n}$ *elements* $a \in \mathbb{R}^{\mathcal{C}_S}$ *for which*

$$\min_{a\in\mathbb{R}^{\mathcal{C}_S}} \max_{x\in\chi^d_{d_{eff},e}} |f(x) - M_{\mathcal{C}_S,bi-\sigma}(a)(x)| \leq \epsilon$$

where $C_6(d,n)$ *may depend on* d *and* n*, but not on* ϵ*.*

We prove Lemma 2.2, by showing that $\|f\circ A\|_{W^{n,\infty}[0,1]^{d_{\text{eff}})}} \leq 1$ and then applying Theorem 1.3.

Proof. For a fixed $d, n \in \mathbb{N}$, $d_{\text{eff}} \in \mathbb{N}$ such that $d_{\text{eff}} < d$ and $\epsilon \in (0,1)$. We consider without loss of generality a f and A as prescribed, then by upper bounding $\|f\circ A\|_{W^{n,\infty}([0,1]^{d_{\text{eff}})}}$ by 1, we can apply Theorem 1.3. We have for a $\mathbf{n}$, with $|\mathbf{n}| = n$ that

$$\begin{aligned}
&\operatorname{esssup}_{x\in[0,1]^{d_{\text{eff}}}} |D^{\mathbf{n}}(f\circ A)(x)| \\
&= \operatorname{esssup}_{x\in[0,1]^{d_{\text{eff}}}} \left|\partial^{n_1}_{x_1}\partial^{n_2}_{x_2}\dots\partial^{n_{d_{\text{eff}}}}_{x_{d_{\text{eff}}}}(f\circ A)(x)\right| \\
&= \operatorname{esssup}_{x\in[0,1]^{d_{\text{eff}}}} \left|\partial^{n_1}_{x_1}\partial^{n_2}_{x_2}\dots\partial^{n_{d_{\text{eff}}}-1}_{x_{d_{\text{eff}}}}\sum_{i=1}^{d}\partial_{x_i}(f)(Ax)\cdot A_{id_{\text{eff}}}\right| \\
&= \operatorname{esssup}_{x\in[0,1]^{d_{\text{eff}}}} \left|\partial^{n_1}_{x_1}\partial^{n_2}_{x_2}\dots\partial^{n_{d_{\text{eff}}}-1}_{x_{d_{\text{eff}}}}\partial_{x_{d_{\text{eff}}}}(f)(Ax)\cdot A_{d_{\text{eff}}d_{\text{eff}}}\right| \\
&= \operatorname{esssup}_{x\in[0,1]^{d_{\text{eff}}}} \left|\partial^{n_1}_{x_1}\partial^{n_2}_{x_2}\dots\partial^{n_{d_{\text{eff}}}}_{x_{d_{\text{eff}}}}(f)(Ax)\right| \qquad (2.8)\\
&= \operatorname{esssup}_{x\in[0,1]^{d_{\text{eff}}}} |D^{\mathbf{n}}(f)(Ax)| \\
&\leq \operatorname{esssup}_{x\in[0,1]^{d}} |D^{\mathbf{n}}(f)(x)| \leq 1.
\end{aligned}$$

Here in (2.8) we use the argument above $|\mathbf{n}|$ times. Taking the maximum over $\mathbf{n}$ gives us that

$$\|f\circ A\|_{W^{n,\infty}([0,1]^{d_{\text{eff}}})} \leq 1.$$

To finish the proof we apply Theorem 1.3. □

The reason we introduce the previous lemma is that for all canonical subspaces of dimension d_{eff}, we can assume without loss of generality that there exists a matrix A of the form of (2.7).

Proof of Theorem 1.5. For any $d, n, d_{\text{eff}} \in \mathbb{N}$ such that $d_{\text{eff}} < d$ and $\epsilon \in (0,1)$, we define $\hat{f} : [0,1]^d \to \mathbb{R}$, as $\hat{f}|_{\bar{\chi}^d_{d_{\text{eff}},e}} = \tilde{f}$, for all $e \in I$ where $\tilde{f}$ is as in Lemma 2.2, and zero elsewhere. Then for $\hat{f}$ we have:

$$\sup_{x\in\bar{\chi}^d_{d_{\text{eff}}}} |\hat{f}(x) - f(x)| \leq \sum_{e\in I} \sup_{x\in\chi^d_{d_{\text{eff}},e}} |\hat{f}(x) - f(x)| \leq \frac{2^{d_{\text{eff}}} d_{\text{eff}}^n}{n!}\left(\frac{1}{N}\right)^n \binom{d}{d_{\text{eff}}}. \qquad (2.9)$$

Setting

$$N = N(\epsilon, d, d_{\text{eff}}, n) = \left\lceil \left(\frac{n!\binom{d}{d_{\text{eff}}}}{2^{d_{\text{eff}}} {d_{\text{eff}}}^n}\epsilon\right)^{-1/n} \right\rceil$$

and plugging N in (2.9), we get $\sup_{x\in\chi^d_{d_{\text{eff}}}} |\hat{f}(x) - f(x)| \leq \epsilon$. Furthermore, by Lemma 2.2 $\tilde{f}$ can be implemented as a feed-forward network $M_{\mathcal{C}_S,bi-\sigma}(a)(x)$. Then $\hat{f}$ can be formed as the product

of these networks, which results in a total feed-forward network $M_{\mathcal{C}_M,bi-\sigma}(a)(x)$, where

$$\mathcal{C}_M = \binom{d}{d_{\text{eff}}} {d_{\text{eff}}}^n (N+1)^{d_{\text{eff}}} = C_5(d, d_{\text{eff}}, n)\epsilon^{-d_{\text{eff}}/n},$$

which finishes our proof. □

Remark 2.3. Although the networks in the case of Lemma 2.2 and Theorem 1.5 have the same ϵ functional dependence in their number of parameters, the total size of the network $\mathcal{C}_{\mathcal{S}}$ and $\mathcal{C}_{\mathcal{M}}$ will be different, as they also depend in a different way on d, d_{eff} and n.

3. Conclusions

We have shown that bi-activation networks, which use both the ReLU and x^2 as activation functions, have greater approximation power than ReLU networks. By repurposing a proof of [16] for ReLU networks, we have derived upper bounds for the number of parameters needed by bi-activation networks to approximate functions in the unit ball of the Sobolev space $W^{n,\infty}([0,1]^d)$ achieving the optimal order $O(\epsilon^{-d/n})$ number of parameters as lower bounded by [6]. We also extended our result to analytic functions on $[0,1]^d$ for yet superior ϵ dependence and to low-dimensional subspaces to overcome the curse of dimensionality.

Natural extensions of these results are 1) to determine if a feedforward, or another network, with a single nonlinear activation can achieve the optimal order $O(\epsilon^{-d/n})$ number of parameters, and 2) to consider further low-complexity models of $f(x)$ beyond the union of subspaces, see for instance the nested structure considered in [14].

Acknowledgments

VG and JH are also supported by the EPSRC Centre for Doctoral Training in Mathematics of Random Systems: Analysis, Modelling and Simulation (EP/S023925/1) JT is supported by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA) and thanks UCLA Department of Mathematics for kindly hosting him during the completion of this manuscript.

References

[1] Lars Valerian Ahlfors. *Complex Analysis*. McGraw-Hill Book Company, 2 edition, 1966.
[2] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge university press, 2009.
[3] Yoshua Bengio, Patrice Simard, , and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5:157–166, 1994.
[4] Nicolas Boulle, Yuji Nakatsukasa, and Alex Townsend. Rational neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14243–14253. Curran Associates, Inc., 2020.
[5] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals and Systems*, 2:303–314, 1989.
[6] Ronald A DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63:469–478, 1989.
[7] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. 2013.
[8] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.
[9] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

[10] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990.
[11] Hikosaburo Komatsu. A characterization of real analytic functions. *Proceedings of the Japan Academy*, 36(3):90–93, January 1960.
[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.
[13] G. Leoni. *A First Course in Sobolev Spaces*. Graduate studies in mathematics. American Mathematical Soc., 2009.
[14] Tomaso A. Poggio, Hrushikesh Narhar Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14:503–519, 2016.
[15] Matus Telgarsky. Representation benefits of deep feedforward networks. *CoRR*, abs/1509.08101, 2015.
[16] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Appendix

Proof of Lemma 2.1. Begin by defining a partition of unity $\phi_{\mathbf{m}}$ on the domain $[0,1]^d$:

$$\sum_{\mathbf{m}} \phi_{\mathbf{m}}(\mathbf{x}) \equiv 1, \quad \mathbf{x} \in [0,1]^d$$

Here $\mathbf{m} = (m_1, \ldots, m_d) \in \{0, 1, \ldots, N\}^d$, and $\phi_{\mathbf{m}}$ is defined as

$$\phi_{\mathbf{m}}(\mathbf{x}) = \prod_{k=1}^{d} \psi\left(3N\left(x_k - \frac{m_k}{N}\right)\right),$$

where

$$\psi(x) = \begin{cases} 1, & |x| < 1 \\ 0, & 2 < |x| \\ 2 - |x|, & 1 \leq |x| \leq 2. \end{cases}$$

Furthermore, note that $||\psi||_\infty = 1$ and $||\phi_{\mathbf{m}}||_\infty = 1$ for all $\mathbf{m}$, and that

$$\text{supp } \phi_{\mathbf{m}} \subseteq \left\{x : \left|x_k - \frac{m_k}{N}\right| < \frac{1}{N} \forall k\right\}.$$

For any $\mathbf{m} \in \{0, ..., N\}^d$, consider the degree$-(n-1)$ Taylor polynomial for the function f at $\mathbf{x} = \mathbf{m}/N$:

$$P_{\mathbf{m}}(x) = \sum_{\mathbf{n}:|\mathbf{n}|<n} \left.\frac{D^{\mathbf{n}} f}{\mathbf{n}!}\right|_{x=\mathbf{m}/N} \left(x - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}},$$

with the usual conventions $\mathbf{n}! = \prod_{k=1}^{d} n_k!$ and $(x - \frac{\mathbf{m}}{N})^{\mathbf{n}} = \prod_{k=1}^{d} \left(x_k - \frac{m_k}{N}\right)^{n_k}$. Now define an approximation to f by

$$f_1 = \sum_{\mathbf{m} \in \{0,\ldots,N\}^d} \phi_{\mathbf{m}} P_{\mathbf{m}}.$$

We bound the approximation error using the Taylor expansion of f:

$$\begin{aligned} |f(x) - f_1(x)| &= \left|\sum_{\mathbf{m}} \phi_{\mathbf{m}}(x)\big(f(x) - P_{\mathbf{m}}(x)\big)\right| \\ &\leq \sum_{\mathbf{m}:|x_k - m_k/N| < 1/N \, \forall k} |f(x) - P_{\mathbf{m}}(x)| \end{aligned}$$

$$\leq 2^d \max_{\mathbf{m}:|x_k - m_k/N| < \frac{1}{N} \forall k} |f(x) - P_{\mathbf{m}}(x)|$$
$$\leq \frac{2^d d^n}{n!} \left(\frac{1}{N}\right)^n \max_{\mathbf{n}:|\mathbf{n}|=n} \operatorname{esssup}_{x \in [0,1]^d} |D^{\mathbf{n}} f(x)|$$

In the second step, we used the support property for $\phi_{\mathbf{m}}$ and the uniform bound on its supremum norm. In the third step, we used the observation that any $x \in [0,1]^d$ belongs to the support of at most 2^d functions $\phi_{\mathbf{m}}$, in the fourth a standard bound for the Taylor remainder.

Note that, the coefficients of the polynomials $P_{\mathbf{m}}$ are uniformly bounded for all f:

$$P_{\mathbf{m}}(x) = \sum_{\mathbf{n}:|\mathbf{n}|<n} a_{\mathbf{m},\mathbf{n}} \left(x - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}}, \quad |a_{\mathbf{m},\mathbf{n}}| \leq 1.$$

Expanding f_1 as follows

$$f_1(x) = \sum_{\mathbf{m} \in \{0,\ldots,N\}^d} \sum_{\mathbf{n}:|\mathbf{n}|<n} a_{\mathbf{m},\mathbf{n}} \phi_{\mathbf{m}}(x) \left(x - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}}.$$

completes the proof.

□

Department of Mathematics, Imperial College London.
Email address: vincent.goverse21@imperial.ac.uk, hamdan@maths.ox.ac.uk, tanner@maths.ox.ac.uk