Deep ReLU networks overcome the curse of dimensionality for generalized bandlimited functions

Hadrien Montanelli[a], Haizhao Yang[b], Qiang Du[a]

[a]Department of Applied Physics and Applied Mathematics, Columbia University, New York, United States
[b]Department of Mathematics, Purdue University, United States[1], and National University of Singapore, Singapore[2]

Abstract

We prove a theorem concerning the approximation of generalized bandlimited multivariate functions by deep ReLU networks for which the curse of the dimensionality is overcome. Our theorem is based on a result by Maurey and on the ability of deep ReLU networks to approximate Chebyshev polynomials and analytic functions efficiently.

Keywords: machine learning, deep ReLU networks, curse of dimensionality, approximation theory, bandlimited functions, Chebyshev polynomials

1. Introduction

The curse of dimensionality is a vital bottleneck in scientific computing. Standard numerical algorithms whose cost is exponential in the dimension d are prohibitive when d is large. As a mesh-free function parametrization tool, neural networks are believed to be a suitable approach to conquer the curse of dimensionality if the number of parameters required in the network to maintain an ϵ approximation accuracy is bounded by $\mathcal{O}\left(p\left(\frac{1}{\epsilon}\right)\right)$, where p is a polynomial with a degree independent of d. In this paper, we show that ReLU networks overcome the curse of dimensionality for *generalized bandlimited functions*, which we shall define at the end of the introduction.

Shallow networks are approximations $\widetilde{f}_W$ of multivariate functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$\widetilde{f}_W(\boldsymbol{x}) = \sum_{i=1}^{W} \alpha_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} + \theta_i), \tag{1}$$

for a certain *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$, weights $\alpha_i, \theta_i \in \mathbb{R}$, $\boldsymbol{w}_i \in \mathbb{R}^d$ and integer $W \geq 1$. Each operation $\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x} + \theta_i)$ is called a *unit* and the W units in (1) form a *hidden layer*; this is a special form of nonlinear approximation [1, 2]. *Deep networks* are compositions of shallow networks and have several hidden layers, and each unit of each layer performs an operation of the form $\sigma(\boldsymbol{w} \cdot \boldsymbol{x} + \theta)$. Following Yarotsky [3], we allow connections between units in non-neighboring layers. We define the *depth* L of a network as the number of hidden layers and the *size* W as the total number of units. In this paper, we shall consider networks with depth $L = \mathcal{O}(1)$ as shallow, and networks with $L \gg 1$ as deep.

Before the revolution of deep learning [4], most research concerned the approximation power of shallow networks with $L = 1$ and various sigmoid activation functions. Recently, in deep learning, networks using the *REctifier Linear Unit* (*ReLU*) activation function $\sigma(x) = \max(0, x)$ have become the most popular tool, partly because sigmoid activation functions lead to severe gradient degeneracy in the optimization process. It was shown in [5] that deep ReLU networks can produce sparsity that helps a wide range of machine learning applications, while smooth activation functions including smoothed ReLU functions cannot. This is why we focus on ReLU networks in this paper.

[1] Current institute.

[2] The work was done at this institute.

Email addresses: hadrien.montanelli@gmail.com (Hadrien Montanelli), haizhao@purdue.edu (Haizhao Yang), qd2125@columbia.edu (Qiang Du)

The theory of approximating functions using shallow networks goes back to 1989 when Cybenko showed that any continuous functions can be approximated by shallow networks [6], while Hornik, Stinchcombe, and White proved a similar result for Borel measurable functions [7]. In the 1990s, the attention shifted to *convergence rates*[3] of approximations by shallow networks [8, 9, 10, 11]. Of particular interest was the discussion of the absence of the curse of dimensionality when one-hiddel-layer sigmoid neural networks are applied to approximate functions with fast decaying Fourier coefficients [12].

Fast forward to the 2010s and the success of deep networks, one of the most important theoretical problems is to determine why and when deep networks can lessen or break the curse of dimensionality, especially for ReLU networks. One may focus on a particular set of functions which have a very special structure (such as compositional or polynomial), and show that for this particular set deep networks overcome the curse of dimensionality [13, 14, 15, 16, 17, 18, 19, 20]. Alternatively, one may consider a function space that is more generic for multivariate approximation in high dimensions, such as Korobov spaces [21], and prove convergence results for which the curse of dimensionality is lessened [22].

In this paper, we may consider *generalized bandlimited* functions $f : B = [0,1]^d \to \mathbb{R}$ of the form

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} F(\boldsymbol{w}) K(\boldsymbol{w} \cdot \boldsymbol{x}) d\boldsymbol{w}, \tag{2}$$

$$\operatorname{supp} F \subset [-M, M]^d, \quad M \geq 1, \tag{3}$$

for some integrable function $F : [-M, M]^d \to \mathbb{C}$ and analytic kernel $K : \mathbb{R} \to \mathbb{C}$. In Section 3, we shall show that for any measure μ such functions can be approximated to accuracy ϵ in the $L^2(B, \mu)$-norm by deep ReLU networks of depth $L = \mathcal{O}\left(\log_2^2 \frac{1}{\epsilon}\right)$ and size $W = \mathcal{O}\left(\frac{1}{\epsilon^2} \log_2^2 \frac{1}{\epsilon}\right)$, up to some constants that depend on F, K, μ and B.

We review properties of deep ReLU networks in Section 2, providing new proofs of existing results (Prop. 2.2 and Prop. 2.3), as well as new results (Prop. 2.4, Prop. 2.5 and Thm. 2.6). We recall an existing theorem (Thm. 3.1) and prove our main theorem (Thm. 3.2) in Section 3.

2. Approximation properties of deep ReLU networks

The ability of deep ReLU networks to implement the multiplication of two real numbers with an amplitude at most M was proved by Yarotsky in [3, Prop. 1]. Liang and Srikant proved a similar result for $M = 1$ using networks with rectifier linear as well as binary step units in [16, Thm. 1]. In the rest of the paper, "with accuracy ϵ" or "bounded" should be understood in the L^∞-norm, unless stated otherwise.

Proposition 2.1 (Multiplication in two dimensions). *For any scalar $M \geq 1$, $N \geq 1$ and $0 < \epsilon < 1$, there is a deep ReLU network $\widetilde{\pi}(x_1, x_2)$ with inputs $(x_1, x_2) \in [-M, M] \times [-N, N]$, that has depth*

$$L = \mathcal{O}\left(\log_2 \frac{MN}{\epsilon}\right) \tag{4}$$

and size

$$W = \mathcal{O}\left(\log_2 \frac{MN}{\epsilon}\right) \tag{5}$$

such that

$$\|\widetilde{\pi}(x_1, x_2) - x_1 x_2\|_{L^\infty([-M,M]\times[-N,N])} \leq \epsilon. \tag{6}$$

Equivalently, if the network has depth $L = \mathcal{O}\left(\log_2 \frac{1}{\epsilon}\right)$ and size $W = \mathcal{O}\left(\log_2 \frac{1}{\epsilon}\right)$, the approximation error satisfies $\|\widetilde{\pi}(x_1, x_2) - x_1 x_2\|_{L^\infty([-M,M]\times[-N,N])} \leq MN\epsilon$.

[3] For a real-valued function f in $\mathbb{R}^d$ whose smoothness is characterized by some integer $m \geq 1$, and for some prescribed accuracy $\epsilon > 0$, one shows that there exists a shallow network $\widetilde{f_W}$ of size $W = W(d, m)$ that satisfies $\|f - \widetilde{f_W}\| \leq \epsilon$ for some norm $\|\cdot\|$.

We generalize the proposition of Yarotsky to the d-dimensional case.

Proposition 2.2 (Multiplication in $d \geq 2$ dimensions)**.** *For any scalar $M \geq 1$ and $0 < \epsilon < 1$, and any integer $d \geq 2$, there is a deep ReLU network $\widetilde{\Pi}(x_1, \ldots, x_d)$ with inputs $(x_1, \ldots, x_d) \in [-M, M]^d$, that has depth*

$$L = O\left(d \log_2 \frac{d}{\epsilon} + d^2 \log_2 M\right) \tag{7}$$

and size

$$W = O\left(d \log_2 \frac{d}{\epsilon} + d^2 \log_2 M\right) \tag{8}$$

such that

$$\left\|\widetilde{\Pi}(x_1, \ldots, x_d) - x_1 \ldots x_d\right\|_{L^\infty([-M,M]^d)} \leq \epsilon. \tag{9}$$

Proof. Let $M \geq 1$ and $0 < \epsilon < 1$ be two scalars and $d \geq 2$ an integer. For any scalar $A \geq 1$ and $B \geq 1$, let us call $\widetilde{\pi}$ the network of Prop. (2.1) that implements the multiplication xy, $x \in [-A, A]$, $y \in [-B, B]$, with accuracy $AB\epsilon_0$ for some scalar $0 < \epsilon_0 < 1$ to be determined later. This network has depth and size $O\left(\log_2 \frac{1}{\epsilon_0}\right)$.

We construct the network $\widetilde{\pi}(x_1, \ldots, x_d)$ that implements the multiplication $x_1 x_2 \ldots x_d$ as follows,

$$\begin{aligned}
y_1 &= \widetilde{\pi}(x_1, x_2), & |y_1| &\leq M^2(1+\epsilon_0),\\
y_2 &= \widetilde{\pi}(y_1, x_3), & |y_2| &\leq M^3(1+\epsilon_0)^2,\\
y_3 &= \widetilde{\pi}(y_2, x_4), & |y_3| &\leq M^4(1+\epsilon_0)^3,\\
&\vdots & &\vdots\\
y_{d-1} &= \widetilde{\pi}(y_{d-2}, x_d), & |y_{d-1}| &\leq M^d(1+\epsilon_0)^{d-1},
\end{aligned}$$

and set $\widetilde{\Pi}(x_1, \ldots, x_d) = y_{d-1}$.

The network $\widetilde{\Pi}(x_1, \ldots, x_d)$ has accuracy

$$\begin{aligned}
|y_{d-1} - x_1 \ldots x_d| \leq{} & |y_{d-1} - y_{d-2}x_d| + |x_d||y_{d-2} - y_{d-3}x_{d-1}|\\
& + \ldots + |x_d x_{d-1} \ldots x_5||y_3 - y_2 x_4|\\
& + |x_d x_{d-1} \ldots x_4||y_2 - y_1 x_3|\\
& + |x_d x_{d-1} \ldots x_3||y_1 - x_1 x_2|,\\
<{} & M^d(1+\epsilon_0)^{d-2}\epsilon_0 + M^d(1+\epsilon_0)^{d-3}\epsilon_0\\
& + \ldots + M^d(1+\epsilon_0)^2\\
& + M^d(1+\epsilon_0) + M^d\epsilon_0,\\
<{} & dM^d(1+\epsilon_0)^d\epsilon_0 \quad \text{(crude estimate)}.
\end{aligned}$$

We choose $\epsilon_0 = \epsilon/(dM^d e)$ to obtain accuracy ϵ.

The depth and the size of the resulting network are equal to $(d-1)$ times the depth and size of the network defined at the beginning of the proof. With accuracy ϵ_0 defined above, this gives depth and size

$$O\left(d \log_2 \frac{dM^d e}{\epsilon}\right) = O\left(d \log_2 \frac{d}{\epsilon} + d^2 \log_2 M\right). \tag{10}$$

The proof is complete. □

The network of Prop. 2.2 computes $x_1 \ldots x_d$ as well as all the intermediate products $x_1 \ldots x_k$, $2 \leq k \leq d-1$, to the same accuracy ϵ. This allows us to prove the following result about polynomials[4] (similar to [16, Thm. 2]).

[4] In the rest of the paper, we shall exclude the trivial cases $n = 0$ and $n = 1$.

Proposition 2.3 (Polynomials). *For any scalar $M \geq 1$, $C \geq 0$ and $0 < \epsilon < 1$, any integer $n \geq 2$, and any polynomial $p_n(x)$ of degree n with input $x \in [-M, M]$ and of the form*

$$p_n(x) = \sum_{k=0}^{n} c_k x^k, \quad \max_{0 \leq k \leq n} |c_k| \leq C, \tag{11}$$

there is a deep ReLU network $\widetilde{p}(x_1, \ldots, x_n)$ with inputs $(x_1, \ldots, x_n) \in [-M, M]^n$, that has depth

$$L = O\left(n \log_2 \frac{Cn}{\epsilon} + n^2 \log_2 M\right) \tag{12}$$

and size

$$W = O\left(n \log_2 \frac{Cn}{\epsilon} + n^2 \log_2 M\right) \tag{13}$$

such that

$$\|\widetilde{p}_n(x, \ldots, x) - p_n(x)\|_{L^\infty([-M,M])} \leq \epsilon. \tag{14}$$

Proof. Let $M \geq 1$, $C \geq 0$ and $0 < \epsilon < 1$ be three scalars, $n \geq 2$ an integer and consider a polynomial

$$p_n(x) = \sum_{k=0}^{n} c_k x^k, \quad \max_{0 \leq k \leq n} |c_k| \leq C. \tag{15}$$

We construct $\widetilde{p}(x_1, \ldots, x_n)$ as follows,

$$\widetilde{p}_n(x_1, \ldots, x_n) = c_0 + c_1 x_1 + \sum_{k=2}^{n} c_k y_{k-1}(x_1, \ldots, x_k), \tag{16}$$

where $y_{k-1}(x_1, \ldots, x_k)$ approximates $x_1 \ldots x_k$ with the network of Prop. 2.2 to accuracy $0 < \epsilon_0 < 1$ to be determined later. (Note that when the inputs are the same $y_{k-1}(x, \ldots, x)$ approximates x^k.)

The network $\widetilde{p}(x, \ldots, x)$ has accuracy

$$\begin{aligned} |\widetilde{p}_n(x, \ldots, x) - p_n(x)| &\leq C \sum_{k=2}^{n} |y_{k-1}(x, \ldots, x) - x^k|, \\ &< nC\epsilon_0. \end{aligned}$$

We choose $\epsilon_0 = \epsilon/(Cn)$ to obtain accuracy ϵ.

The resulting network has depth and size

$$O\left(n \log_2 \frac{Cn^2 M^n}{\epsilon}\right) = O\left(n \log_2 \frac{Cn}{\epsilon} + n^2 \log_2 M\right). \tag{17}$$

The proof is complete. □

The Chebyshev polynomials of the first kind play a central role in approximation theory [23]. They are defined on $[-1, 1]$ via the three-term recurrence relation

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n \geq 2, \tag{18}$$

with $T_0 = 1$ and $T_1(x) = x$. We show next how deep ReLU networks can efficiently implement Chebyshev polynomials, using the three-term recurrence (18).

Proposition 2.4 (Chebyshev polynomials). *For any scalar $0 < \epsilon < 1$, any integer $n \geq 2$ and any Chebyshev polynomial $T_n(x)$ of degree n with input $x \in [-1,1]$, there is a deep ReLU network $\widetilde{T}_n(x_1,\ldots,x_n)$ with inputs $(x_1,\ldots,x_n) \in [-1,1]^n$, that has depth*

$$L = O\left(n\log_2\frac{n}{\epsilon} + n^2\right) \tag{19}$$

and size

$$W = O\left(n\log_2\frac{n}{\epsilon} + n^2\right) \tag{20}$$

such that

$$\left\|\widetilde{T}_n(x,\ldots,x) - T_n(x)\right\|_{L^\infty([-1,1])} \leq \epsilon. \tag{21}$$

Proof. Let $0 < \epsilon < 1$ be a scalar and $n \geq 2$ be an integer. For any scalar $A \geq 1$ and $B \geq 1$, let us call $\widetilde{\pi}$ the network of Prop. (2.1) that implements the multiplication xy, $x \in [-A,A]$, $y \in [-B,B]$, with accuracy $AB\epsilon_0$ for some scalar $0 < \epsilon_0 < 1$ to be determined later. This network has depth and size $O\left(\log_2\frac{1}{\epsilon_0}\right)$.

We construct the network $\widetilde{T}_n(x,\ldots,x)$ that approximates $T_n(x)$ as follows,

$$\begin{aligned}
\widetilde{T}_0 &= 1, & |\widetilde{T}_0| &\leq 1,\\
\widetilde{T}_1(x) &= x, & |\widetilde{T}_1| &\leq 1,\\
\widetilde{T}_2(x,x) &= 2\widetilde{\pi}(x,\widetilde{T}_1) - \widetilde{T}_0, & |\widetilde{T}_2| &< (1+\epsilon_0)^2,\\
\widetilde{T}_3(x,x,x) &= 2\widetilde{\pi}(x,\widetilde{T}_2) - \widetilde{T}_1, & |\widetilde{T}_3| &< 3(1+\epsilon_0)^3,\\
&\vdots & &\vdots\\
\widetilde{T}_n(x,\ldots,x) &= 2\widetilde{\pi}(x,\widetilde{T}_{n-1}) - \widetilde{T}_{n-2}, & |\widetilde{T}_n| &< 3^{n-2}(1+\epsilon_0)^n.
\end{aligned}$$

Let us now estimate the accuracy e_n of the network $\widetilde{T}_n(x,\ldots,x)$, where $e_n = |\widetilde{T}_n(x,\ldots,x) - T_n(x)|$. We have

$$\begin{aligned}
e_n &= |2\widetilde{\pi}(x,\widetilde{T}_{n-1}) - \widetilde{T}_{n-2} - 2xT_{n-1} + T_{n-2}|,\\
&\leq 2|\widetilde{\pi}(x,\widetilde{T}_{n-1}) - x\widetilde{T}_{n-1}| + 2|x||\widetilde{T}_{n-1} - T_{n-1}| + e_{n-2},\\
&\leq 2\epsilon_0|\widetilde{T}_{n-1}| + 2e_{n-1} + e_{n-2},\\
&< 2\epsilon_0 3^{n-3}(1+\epsilon_0)^{n-1} + 2e_{n-1} + e_{n-2},\\
&< n4^n(1+\epsilon_0)^n\epsilon_0 \quad \text{(crude estimate)}.
\end{aligned}$$

We choose $\epsilon_0 = \epsilon/(n4^n e)$ to obtain accuracy ϵ.

The depth and the size of the resulting network are equal to $(n+1)$ times the depth and size of the network defined at the beginning of the proof. With accuracy ϵ_0 defined above, this gives depth and size

$$O\left(n\log_2\frac{n4^n e}{\epsilon}\right) = O\left(n\log_2\frac{n}{\epsilon} + n^2\right). \tag{22}$$

The proof is complete. □

Note that we could have proven Prop. 2.4 using Prop. 2.3 and an estimate for the size C of the coefficients of the expansion of T_n in the monomial basis (the leading term grows like 2^{n-1} while the other terms grow at most like c^n for some $c < 4$).

Since Prop. 2.4 implements T_n as well as the intermediate T_k's, $0 \leq k \leq n-1$, to the same accuracy ϵ, we have the following result about truncated Chebyshev series.

Proposition 2.5 (Truncated Chebyshev series)**.** *For any scalar* $0 < \epsilon < 1$*, any integer* $n \geq 2$*, and any truncated Chebyshev series* $f_n(x)$ *of degree n with input* $x \in [-1, 1]$ *and of the form*

$$f_n(x) = \sum_{k=0}^{n} c_k T_k(x), \quad \max_{0 \leq k \leq n} |c_k| \leq C, \tag{23}$$

there is a deep ReLU network $\widetilde{f_n}(x_1, \ldots, x_n)$ *with inputs* $(x_1, \ldots, x_n) \in [-1, 1]^n$*, that has depth*

$$L = O\left(n \log_2 \frac{Cn}{\epsilon} + n^2\right) \tag{24}$$

and size

$$W = O\left(n \log_2 \frac{Cn}{\epsilon} + n^2\right) \tag{25}$$

such that

$$\left\|\widetilde{f_n}(x, \ldots, x) - f_n(x)\right\|_{L^\infty([-1,1])} \leq \epsilon. \tag{26}$$

Proof. Let $C \geq 0$ be a scalar, $n \geq 2$ an integer and consider a truncated Chebyshev series

$$f_n(x) = \sum_{k=0}^{n} c_k T_k(x), \quad \max_{0 \leq k \leq n} |c_k| \leq C. \tag{27}$$

We construct $\widetilde{f}(x_1, \ldots, x_n)$ as follows,

$$\widetilde{f_n}(x_1, \ldots, x_n) = c_0 + c_1 x_1 + \sum_{k=2}^{n} c_k \widetilde{T}_k(x_1, \ldots, x_k), \tag{28}$$

where $\widetilde{T}_k$ approximates T_k with the network of Prop. 2.4 to accuracy $0 < \epsilon_0 < 1$ to be determined later.

The network $\widetilde{f}(x, \ldots, x)$ has accuracy

$$\begin{aligned} |\widetilde{f_n}(x, \ldots, x) - f_n(x)| &\leq C \sum_{k=2}^{n} |\widetilde{T}_k - T_k|, \\ &< nC\epsilon_0. \end{aligned}$$

We choose $\epsilon_0 = \epsilon/(Cn)$ to obtain accuracy ϵ.

The resulting network has depth and size

$$O\left(n \log_2 \frac{Cn^2}{\epsilon} + n^2\right) = O\left(n \log_2 \frac{Cn}{\epsilon} + n^2\right). \tag{29}$$

The proof is complete. □

Chebyshev series lies at the heart of polynomial approximation. Lipschitz continuous functions $f(x)$ with input $x \in [-M, M]$ have a unique absolutely and uniformly convergent (scaled) Chebyshev series and we write $f(x) = \sum_{k=0}^{\infty} c_k T_k(x/M)$ [23, Thm. 3.1]. For analytic functions, the truncated (scaled) Chebyshev series $f_n(x) = \sum_{k=0}^{n} c_k T_k(x/M)$ are *exponentially accurate* approximations [23, Thm. 8.2].

More precisely, let us define

$$a_s^M = M \frac{s + s^{-1}}{2}, \quad b_s^M = M \frac{s - s^{-1}}{2}, \tag{30}$$

and the *Bernstein s-ellipse scaled to* $[-M, M]$,

$$E_s^M = \left\{ x + iy \in \mathbb{C} \,:\, \frac{x^2}{(a_s^M)^2} + \frac{y^2}{(b_s^M)^2} = 1 \right\}. \tag{31}$$

(It has foci $\sqrt{(a_r^M)^2 - (b_r^M)^2} = \pm M$, semi-major axis a_s^M and semi-minor axis b_s^M.) If a function $f(x)$ is analytic in $[-M, M]$ and analytically continuable to the open Bernstein s-ellipse E_s^M for some $s \geq 1$ where it satisfies $|f(x)| < C_f$ for some $C_f > 0$, then for each $n \geq 0$ the truncated Chebyshev series f_n satisfy

$$\|f_n(x) - f(x)\|_{L^\infty([-M,M])} \leq \frac{2C_f s^{-n}}{s-1}. \tag{32}$$

Using Prop. 2.5 and Eq. (32) we prove a result about the approximation of analytic functions by deep ReLU networks. Our result below could be generalized to multiple dimensions, which would be interesting future work. In [24], it was shown that deep ReLU networks can approximate multivariate analytic functions with exponential convergence, a result similar to our theorem below. However, we would like to emphasize that it is not possible to apply the result in [24] to prove our main theorem in Section 3 because the result in [24] is only valid on an open interval of $[-1, 1]$, instead of an arbitrary closed interval $[-M, M]$.

Theorem 2.6 (Deep networks for analytic functions)**.** *For any scalar* $0 < \epsilon < 1$ *and* $M \geq 1$*, and any analytic function* $f(x)$ *with input* $x \in [-M, M]$ *that is analytically continuable to the open Bernstein s-ellipse* E_s^M *for some* $s > 1$ *where it satisfies* $|f(x)| \leq C_f$ *for some* $C_f > 0$*, there is a deep ReLU network* $\widetilde{f}(x_1, \ldots, x_n)$ *with inputs* $(x_1, \ldots, x_n) \in [-M, M]^n$*, that has depth*

$$L = O\left(\frac{1}{\log_2^2 s} \log_2^2 \frac{C_f}{\epsilon} \right) \tag{33}$$

and size

$$W = O\left(\frac{1}{\log_2^2 s} \log_2^2 \frac{C_f}{\epsilon} \right) \tag{34}$$

such that

$$\left\| \widetilde{f_n}(x, \ldots, x) - f(x) \right\|_{L^\infty([-M,M])} \leq \epsilon. \tag{35}$$

Proof. Let $0 < \epsilon < 1$ and $M \geq 1$ be two scalars, and f be an analytic function defined on $[-M, M]$ that is analytically continuable to the open Bernstein s-ellipse E_s^M for some $s > 1$ where it satisfies $|f(x)| \leq C_f$ for some $C_f > 0$. We first approximate f by a truncated Chebyshev series f_n and then approximate f_n by a deep ReLU network $\widetilde{f_n}$ using Prop. 2.5.

Since f is analytic in the open Bernstein s-ellipse E_s^M then for any integer $n \geq 2$

$$\|f_n(x) - f(x)\|_{L^\infty([-M,M])} \leq \frac{2C_f s^{-n}}{s-1} = O\left(C_f s^{-n}\right). \tag{36}$$

Therefore if we take $n = O\left(\frac{1}{\log_2 s} \log_2 \frac{2C_f}{\epsilon}\right)$ then the term above is bounded by $\epsilon/2$.

Let us now approximate $f_n(x)$ by a deep ReLU network $\widetilde{f_n}(x, \ldots, x)$. We first write

$$f_n(x) = \sum_{k=0}^{n} c_k T_k\left(\frac{x}{M}\right), \tag{37}$$

with

$$\max_{0 \leq k \leq n} |c_k| = O\left(C_f s\right) \text{ via [23, Thm. 8.1].} \tag{38}$$

We then define our network $\widetilde{f_n}(x, \ldots, x)$ as in Prop. 2.5 with extra scaling x/M,

$$\widetilde{f_n}(x, \ldots, x) = \sum_{k=0}^{n} c_k \widetilde{T}_k\left(\frac{x}{M}, \ldots, \frac{x}{M}\right), \tag{39}$$

where the $\widetilde{T}_k$'s are computed to accuracy $\epsilon/2$ so that

$$|\widetilde{f_n}(x, \ldots, x) - f_n(x)| \leq \frac{\epsilon}{2}. \tag{40}$$

This yields

$$\begin{aligned} |\widetilde{f_n}(x, \ldots, x) - f(x)| &\leq |\widetilde{f_n}(x, \ldots, x) - f_n(x)| \\ &\quad + |f_n(x) - f(x)|, \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

With $n = O\left(\frac{1}{\log_2 s} \log_2 \frac{C_f}{\epsilon}\right)$, the resulting network has depth and size

$$O\left(n \log_2 \frac{C_f n}{\epsilon} + n^2\right) = O\left(n^2\right) = O\left(\frac{1}{\log_2^2 s} \log_2^2 \frac{C_f}{\epsilon}\right). \tag{41}$$

The proof is complete. □

Let us highlight that in general the constants s and C_f depend on M. Let us look at two examples, a function with a singularity on the imaginary axis and an *entire* function (*i.e.*, a function that is analytic over the whole complex plane).

A typical example of an analytic function with singularities on the imaginary axis is the Runge-like function $f(x) = 1/(1 + \frac{x^2}{\beta^2}), \beta > 1$, whose singularities are located at $x = \pm i\beta$. The function f is analytic on the interval $[-M, M]$ and analytically continuable to the open Bernstein s-ellipse E_s^M with

$$s(M) = \frac{\sqrt{(4M^2 - 2)r^2 + r^4 + 1} + r^2 - 1}{2Mr} \tag{42}$$

and $r = \beta + \sqrt{\beta^2 + 1}$. Since f increases along the imaginary axis we may take

$$C_f(M) = f\left(M \frac{s(M) - s(M)^{-1}}{2}\right). \tag{43}$$

The complex exponential $f(x) = e^{ix}$ is an entire function. Hence, any $s > 1$ works but $C_f(s, M)$ must grow with s and M. As f increases along the imaginary axis we may choose

$$C_f(s, M) = f\left(M \frac{s - s^{-1}}{2}\right) = e^{M \frac{s - s^{-1}}{2}}. \tag{44}$$

In this case the network of Thm. 2.6 has depth and size

$$O\left(\frac{1}{\log_2^2 s}\left(M \frac{s - s^{-1}}{2} + \log_2 \frac{1}{\epsilon}\right)^2\right). \tag{45}$$

We also would like to mention that the ReLU activation function is not an optimal choice for constructing neural networks to approximate smooth functions. For example, Thm. 2.3 of [9] shows that one-hidden-layer shallow networks with $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ parameters can approximate analytic functions with ϵ accuracy when a smooth activation function is used. The disadvantage of the ReLU activation function in this scenario is not unexpected since it is not a natural choice to use a function that is not differentiable at the origin to approximate a smooth function. However, from the point of view of deep learning and optimization, ReLU is a much better choice, as discussed in [24]. The study in this paper should be regarded as a good complement to existing approximation theory, using a more modern approach.

3. Approximation of generalized bandlimited functions by deep ReLU networks

A famous theorem of Carathéodory states that if a point $x \in \mathbb{R}^d$ lies in the the convex hull of a set P then x can be written as the convex combination of at most $d+1$ points in P. Maurey's theorem [26] is an extension of Carathéodory's result to the infinite-dimensional case. It was used in the context of shallow network approximations by Barron in 1993 [12]. We quote the theorem below without a proof. The reader is referred to [12] for its proof.

Theorem 3.1. *Let H be a Hilbert space with norm $\|\cdot\|$. Suppose there exists $G \subset H$ such that for every $g \in G$, $\|g\| \leq b$ for some $b > 0$. Then for every f in the convex hull of G and every integer $n \geq 1$, there is a f_n in the convex hull of n points in G and a constant $c > b^2 - \|f\|^2$ such that $\|f - f_n\|^2 \leq \frac{c}{n}$.*

We are now ready to prove our main theorem about the approximation of generalized bandlimited functions of the form (2)–(3) by deep ReLU networks.

Theorem 3.2 (Deep networks for generalized bandlimited functions)**.** *Let $B = [0,1]^d$ and $f : B \to \mathbb{R}$ be a generalized bandlimited function of the form*

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} F(\boldsymbol{w}) K(\boldsymbol{w} \cdot \boldsymbol{x}) d\boldsymbol{w}, \tag{46}$$

$$\operatorname{supp} F \subset [-M, M]^d, \quad M \geq 1, \tag{47}$$

for some functions $F : [-M,M]^d \to \mathbb{C}$ and $K : \mathbb{R} \to \mathbb{C}$. Suppose that K is analytic in $t = \boldsymbol{w} \cdot \boldsymbol{x} \in [-dM, dM]$ and satisfies the assumption of Thm. 2.6 *for some $s > 1$ and $C_K > 0$. Suppose also that K is bounded by some constant $0 < D_K \leq 1$ on the real axis, and that*

$$\int_{\mathbb{R}^d} |F(\boldsymbol{w})| d\boldsymbol{w} = \int_{[-M,M]^d} |F(\boldsymbol{w})| d\boldsymbol{w} = C_F < \infty. \tag{48}$$

Then, for any measure μ and any scalar $0 < \epsilon < 1$, there exists a deep ReLU network $\widetilde{f}(\boldsymbol{x})$ with inputs $\boldsymbol{x} \in B$, that has depth

$$L = O\left(\frac{1}{\log_2^2 s} \log_2^2 \frac{C_F C_K \sqrt{\mu(B)}}{\epsilon}\right) \tag{49}$$

and size

$$W = O\left(\frac{C_F^2 \mu(B)}{\epsilon^2 \log_2^2 s} \log_2^2 \frac{C_F C_K \sqrt{\mu(B)}}{\epsilon}\right) \tag{50}$$

such that

$$\left\|\widetilde{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right\|_{L^2(\mu,B)} = \sqrt{\int_B |\widetilde{f}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 d\mu(\boldsymbol{x})} \leq \epsilon. \tag{51}$$

Proof. Let $F(\boldsymbol{w}) = |F(\boldsymbol{w})| e^{i\theta(\boldsymbol{w})}$. We may write

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^d} F(\boldsymbol{w}) K(\boldsymbol{w} \cdot \boldsymbol{x}) d\boldsymbol{w}, \tag{52}$$

$$= \int_{\mathbb{R}^d} C_F e^{i\theta(\boldsymbol{w})} K(\boldsymbol{w} \cdot \boldsymbol{x}) \frac{|F(\boldsymbol{w})|}{C_F} d\boldsymbol{w}. \tag{53}$$

The integral in (2) represents $f(\boldsymbol{x})$ as an infinite convex combination of functions in the set

$$G(\boldsymbol{w}) = \{\gamma e^{i\beta} K(\boldsymbol{w} \cdot \boldsymbol{x}),\ |\gamma| \leq C_F,\ \beta \in \mathbb{R}\}. \tag{54}$$

In other words $f(\boldsymbol{x})$ is in the closure of the convex hull of $G(\boldsymbol{w})$. Since functions in $G(\boldsymbol{w})$ are bounded in the $L^2(\mu, B)$-norm by $C_F\sqrt{\mu(B)}$ (since $D_K \leq 1$), Thm. 3.1 tells us that there exists[5]

$$f_{\epsilon_0}(\boldsymbol{x}) = \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} b_j K(\boldsymbol{w}_j \cdot \boldsymbol{x}), \quad \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} |b_j| \leq C_F, \tag{55}$$

for some $0 < \epsilon_0 < 1$ to be determined later, such that

$$\left\| f_{\epsilon_0}(\boldsymbol{x}) - f(\boldsymbol{x}) \right\|_{L^2(\mu,B)} \leq C_F\sqrt{\mu(B)}\epsilon_0. \tag{56}$$

We now approximate $f_{\epsilon_0}(\boldsymbol{x})$ by a deep ReLU network $\widetilde{f}(\boldsymbol{x})$. Using Thm. 2.6, each $K(\boldsymbol{w}_j \cdot \boldsymbol{x})$ can be approximated to accuracy ϵ_0 using a network $\widetilde{K}(\boldsymbol{w}_j \cdot \boldsymbol{x})$ of depth and size

$$\mathcal{O}\left(\frac{1}{\log_2^2 s}\log_2^2\frac{C_K}{\epsilon_0}\right). \tag{57}$$

We define the deep ReLU network $\widetilde{f}(\boldsymbol{x})$ by

$$\widetilde{f}(\boldsymbol{x}) = \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} b_j \widetilde{K}(\boldsymbol{w}_j \cdot \boldsymbol{x}). \tag{58}$$

This network has depth $L = \mathcal{O}\left(\frac{1}{\log_2^2 s}\log_2^2\frac{C_K}{\epsilon_0}\right)$ and size $W = \mathcal{O}\left(\frac{1}{\epsilon_0^2\log_2^2 s}\log_2^2\frac{C_K}{\epsilon_0}\right)$, and

$$\begin{aligned}|\widetilde{f}(\boldsymbol{x}) - f_{\epsilon_0}(\boldsymbol{x})| &\leq \sum_{j=1}^{\lceil 1/\epsilon_0^2 \rceil} |b_j||\widetilde{K}(\boldsymbol{w}_j \cdot \boldsymbol{x}) - K(\boldsymbol{w}_j \cdot \boldsymbol{x})|, \\ &\leq C_F\epsilon_0,\end{aligned}$$

which yields

$$\left\| \widetilde{f}(\boldsymbol{x}) - f_{\epsilon_0}(\boldsymbol{x}) \right\|_{L^2(\mu,B)} \leq C_F\sqrt{\mu(B)}\epsilon_0. \tag{59}$$

The total approximation error satisfies

$$\left\| \widetilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right\|_{L^2(\mu,B)} \leq 2C_F\sqrt{\mu(B)}\epsilon_0. \tag{60}$$

We take

$$\epsilon_0 = \frac{\epsilon}{2C_F\sqrt{\mu(B)}} \tag{61}$$

to complete the proof. □

Let us end this section with comments on the constants C_F, C_K and $\mu(B)$; we start with C_F. If F is a mollifier then $C_F = 1$, whereas if F is a normal distribution truncated to $[-M, M]^d$ then $C_F < 1$. In general, however, C_F might grow algebraically or exponentially with the dimension d.

We continue with C_K. Consider for example the complex exponential kernel $K(t) = e^{it}$, $t \in [-dM, dM]$. Eq. 44 yields

$$C_K(s, dM) = e^{dM\frac{s-s^{-1}}{2}}, \quad \text{for any } s > 1. \tag{62}$$

[5] We use Thm. 3.1 with $c = b^2 > b^2 - \|f\|^2$, $b = C_F\sqrt{\mu(B)}$ and $\|\cdot\| = \|\cdot\|_{L^2(\mu,B)}$.

The resulting network to approximate a function to accuracy ϵ in the $L^2(\mu, B)$-norm with such a kernel has depth

$$L = O\left(\frac{1}{\log_2^2 s}\left(dM\frac{s - s^{-1}}{2} + \log_2\frac{C_F\sqrt{\mu(B)}}{\epsilon}\right)^2\right) \tag{63}$$

and size

$$W = O\left(\frac{C_F^2\mu(B)}{\epsilon^2\log_2^2 s}\left(dM\frac{s - s^{-1}}{2} + \log_2\frac{C_F\sqrt{\mu(B)}}{\epsilon}\right)^2\right). \tag{64}$$

We conclude with $\mu(B)$. If μ is a probability measure, then $\mu(B) \leq 1$ for any compact domain B. If μ is Lebesgue measure, then $\mu(B) = 1$ for the domain $B = [0, 1]^d$ we considered, but grows exponentially with the dimension d if $B = [0, L]^d$, $L > 1$. This is a common drawback in the approximation theory of neural networks for conquering the curse of dimensionality, e.g., [12].

4. Discussion

We have proven new upper bounds for the approximation of bandlimited functions of the form (2)–(3), for which the curse of dimensionality is overcome. Our proof is based on Maurey's theorem and on the ability of deep ReLU networks to approximate Chebyshev polynomials and analytic functions efficiently.

There are many ways in which this work could be profitably continued. The space of bandlimited functions is a type of Reproducing kernel Hilbert space (RKHS) and therefore a possible extension would be to look at different types of RKHS. One could also relax the bandlimited assumption (3), e.g., to functions F whose derivatives are rapidly decreasing. In this case, the kernel K could be approximated on the real line by Chebyshev polynomials on truncated intervals or Hermite polynomials. The latter is another example of classical orthogonal polynomials, which can be represented by a three-term recurrence relation similar to (18) and efficiently implemented by deep ReLU networks.

Let us conclude this paper with a comment on deep versus shallow networks in the context of parallel computing efficiency. Since the depth L grows like $O\left(\log_2^2\frac{1}{\epsilon}\right)$ in Thm. 3.2, the approximation accuracy for deep networks can be root-exponentially improved if L increases. Hence, very deep networks are more efficient than shallow networks when both parallel computing efficiency and approximation efficiency are considered. This is in contrast with the more general case of continuous functions, the approximation of which via very deep networks might be less attractive in terms of parallel computing [27].

Acknowledgements

We thank the members of the CM3 group at Columbia University and our colleague Mikael Slevinsky for fruitful discussions. The first author is much indebted to former PhD supervisor Nick Trefethen for his inspirational contributions to numerical analysis and in particular to the field of approximation theory.

The research of the second author was partially supported by Grant R-146-000-251-133 in the Department of Mathematics at the National University of Singapore, by the Ministry of Education in Singapore under the grant MOE2018-T2-2-147, and the start-up grant of the Department of Mathematics at Purdue University.

The research of the third author is supported in part by US NSF DMS-1719699 and the NSF TRIPODS program CCF-1704833.

[1] R. A. DeVore, R. Howard, C. Micchelli, Optimal nonlinear approximation, Manuscripta Math. 63 (1989) 469–478.

[2] R. A. DeVore, Nonlinear approximation, Acta Numer. 7 (1998) 51–150.

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

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

[5] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: G. Gordon, D. Dunson, M. Dudík (Eds.), Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 15, Fort Lauderdale, FL, 2011, pp. 315–323.

[6] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Syst. 2 (1989) 303–314.

[7] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366.

[8] H. N. Mhaskar, Approximation properties of a multilayered feedforward articial neural network, Adv. Comput. Math. 1 (1993) 61–80.

[9] H. N. Mhaskar, Neural networks for optimal approximation of smooth and analytic functions, Neural Comput. 8 (1996) 164–177.

[10] H. N. Mhaskar, C. A. Micchelli, Approximation by superposition of sigmoidal and radial basis functions, Adv. Appl. Math. 13 (1992) 350–373.

[11] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.

[12] A. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inf. Theory 39 (1993) 930–945.

[13] F. Bach, Breaking the curse of dimensionality with convex neural networks, J. Mach. Learn. Res. 18 (2017) 1–53.

[14] N. Cohen, O. Sharir, A. Shashua, On the expressive power of deep learning: A tensor analysis, in: V. Feldman, A. Rakhlin, O. Shamir (Eds.), 29th Annual Conference on Learning Theory, Proc. Mach. Learn. Res. 49, Columbia University, New York, 2016, pp. 698–728.

[15] R. Eldan, O. Shamir, The power of depth for feedfoward neural networks, in: V. Feldman, A. Rakhlin, O. Shamir (Eds.), 29th Annual Conference on Learning Theory, Proc. Mach. Learn. Res. 49, Columbia University, New York, 2016, pp. 907–940.

[16] S. Liang, R. Srikant, Why deep neural networks for function approximation?, arXiv:1610.0416.

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

[18] T. Poggio, H. N. Mhaskar, L. Rosasco, B. Miranda, Q. Liao, Why and when can deep—but not shallow—networks avoid the curse of dimensionality: A review, International Journal of Automation and Computing 14 (2017) 503–519.

[19] U. Shaham, A. Cloninger, R. R. Coifman, Provable approximation properties for deep neural networks, Appl. Comput. Harm. Anal. 44 (2018) 537–557.

[20] M. Telgarsky, Benefits of depth in neural networks, in: V. Feldman, A. Rakhlin, O. Shamir (Eds.), 29th Annual Conference on Learning Theory, Proc. Mach. Learn. Res. 49, Columbia University, New York, 2016, pp. 1517–1539.

[21] N. M. Korobov, On the approximate solution of integral equations, Dokl. Akad. Nauk SSSR 128 (1959) 233–238.

[22] H. Montanelli, Q. Du, New error bounds for deep ReLU networks using sparse grids, SIAM J. Math. Data Sci. 1 (2019) 78–92.

[23] L. N. Trefethen, Approximation Theory and Approximation Practice, SIAM, Philadelphia, PA, 2013.

[24] W. E, Q. Wang, Exponential convergence of the deep neural network approximation for analytic functions, Sci. China Math. 61 (2018) 1733–1740.

[25] P. P. Petrushev, V. A. Popov, Rational Approximation of Real Functions, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1988. doi:10.1017/CBO9781107340756.

[26] G. Pisier, Remaques sur un résultat non publié de B. Maurey, Tech. Rep. 5, Séminaire d'analyse fonctionelle, École Polytechnique (1981).

[27] Z. Shen, H. Yang, S. Zhang, Nonlinear approximation via compositions, arXiv:1902.10170.