Error bounds for deep ReLU networks using the Kolmogorov–Arnold superposition theorem

Hadrien Montanelli[a,*], Haizhao Yang[b]

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

Abstract

We prove a theorem concerning the approximation of multivariate continuous functions by deep ReLU networks, for which the curse of the dimensionality is lessened. Our theorem is based on the Kolmogorov–Arnold superposition theorem, and on the approximation of the inner and outer functions that appear in the superposition by very deep ReLU networks.

Keywords: deep ReLU networks, curse of dimensionality, approximation theory, Kolmogorov–Arnold superposition theorem

1. Introduction

At the second International Congress of Mathematicians in Paris 1900, Hilbert presented ten of his 23 problems, including the 13th problem about equations of degree seven. He considered the following equation,

$$x^7 + ax^3 + bx^2 + cx + 1 = 0,$$

and asked whether its solution $x(a, b, c)$, seen as a function of the three parameters a, b and c, can be written as the composition of functions of only two variables.

Hilbert's 13th problem was solved by Kolmogorov and his 19 years old student Arnold in a series of papers in the 1950s. Kolmogorov first proved in 1956 that any continuous function of several variables can be expressed as the composition of functions of three variables [1]. His student Arnold extended his theorem in 1957; three variables were reduced to two [2]. Kolmogorov finally showed later that year that functions of only one variable were needed [3]. The latter result is known as the *Kolmogorov–Arnold superposition theorem*, and states that any continuous functions $f : [0, 1]^n \to \mathbb{R}$ can be decomposed as

$$f(x_1, \ldots, x_n) = \sum_{j=0}^{2n} \phi_j \left(\sum_{i=1}^{n} \psi_{i,j}(x_i) \right),$$

with $2n + 1$ continuous *outer* functions $\phi_j : \mathbb{R} \to \mathbb{R}$ (dependent of f) and $2n^2 + n$ continuous *inner* functions $\psi_{i,j} : [0, 1] \to \mathbb{R}$ (independent of f).

The Kolmogorov–Arnold superposition theorem was further improved in the 1960s and the 1970s. Lorentz showed in 1962 that the outer functions ϕ_j might be chosen to be the same function ϕ, and replaced the inner functions $\psi_{i,j}$ by $\lambda_i \psi_j$, for some positive rationally independent constants $\lambda_i \leq 1$ [4], while Sprecher replaced the inner functions $\psi_{i,j}$ by Hölder continuous functions $x_i \mapsto \lambda^{ij}\psi(x_i + j\epsilon)$ in 1965 [5]. Two years later, Fridman demonstrated that the inner functions could be chosen to be Lipschitz continuous, but his decomposition used $2n+1$ outer functions and $2n^2+n$ inner functions [6]. Finally, Sprecher provided in 1972 a decomposition with Lipschitz continuous functions $x_i \mapsto \lambda^{i-1}\psi(x_i + j\epsilon)$ [7].

Theoretical connections with neural networks started with the work of Hecht–Nielsen in 1987 [8]. He interpreted the Kolmogorov–Arnold superposition theorem as a neural network, whose activation functions were the inner and outer functions. Girosi and Poggio claimed in 1989 that his interpretation was irrelevant for two reasons; first, the inner and outer functions were highly nonsmooth (*i.e.*, these were at least as difficult to approximate as f); second, the outer functions depended on f (*i.e.*, the network architecture could not be parametrized). Kůrková weakened the statement of Girosi and Poggio, in the early 1990s, by giving a direct proof of the universal approximation theorem of multilayer neural networks using the Kolmogorov–Arnold

*Corresponding author
Email addresses: hadrien.montanelli@gmail.com (Hadrien Montanelli), haizhao@nus.edu.sg (Haizhao Yang)

superposition theorem, and by showing that the weight selection reduced to a linear regression problem [9, 10].

Numerical implementations originated with the work of Sprecher in the mid 1990s [11, 12], which was followed, in 2003, by the Kolmogorov's spline network of Igelnik and Parikh [13]. Braun and Griebel proposed an algorithm for computing the Kolmogorov–Arnold theorem in 2009 [14], using Köppen's Hölder continuous inner function [15].

Approximation theory for neural networks started with shallow networks and the 1989 universal approximation theorems of Cybenko [16] and Hornik [17]. In the last few years, the attention has shifted to the approximation properties of deep ReLU networks [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. In particular, one of the most important theoretical problems is to determine why and when deep networks lessen or break the *curse of dimensionality*, characterized by the $O(\epsilon^{-n})$ growth of the network size W as the error $\epsilon \to 0$, in dimension n.[1]

In this paper, we utilize the Kolmogorov–Arnold superposition theorem to show that any continuous function $f : [0,1]^n \to \mathbb{R}$ can be approximated with error ϵ by a very deep ReLU network of depth and size $O(\epsilon^{-\log n})$; the curse of dimensionality is lessened. However, the resulting network is adaptive in the sense that the constants that appear in the estimates depend on f (via the outer functions). Therefore, the network architecture cannot be parametrized.

Prior to the exposition of our main result in Section 4, we will review a specific version of the Kolmogorov–Arnold superposition theorem in Section 2, and show in Section 3 how to approximate the inner and outer functions by very deep ReLU networks.

2. Constructive version of the Kolmogorov–Arnold superposition theorem

We review in this section a constructive version of the Kolmogorov–Arnold superposition theorem that goes back to Sprecher in 1996 and 1997 [11, 12]. The proof he provided at the time was not fully correct; minor modifications were made by Braun and Griebel in 2009 to complete his proof [14, Thm. 2.1], using the inner function suggested by Köppen [15].

For any integer $n \geq 2$, $m \geq 2n$ and $\gamma \geq m + 2$, let

$$a = \frac{1}{\gamma(\gamma - 1)}, \tag{1}$$

$$\lambda_1 = 1, \quad \lambda_i = \sum_{\ell=1}^{\infty} \gamma^{-(i-1)\beta_n(\ell)}, \quad 2 \leq i \leq n, \tag{2}$$

with

$$\beta_n(\ell) = \frac{1 - n^\ell}{1 - n} = 1 + n + \ldots + n^{\ell-1}, \tag{3}$$

and

$$\nu = 2^{-\alpha}(\gamma + 3), \quad \alpha = \log_\gamma 2. \tag{4}$$

We recall that a function $f : [a,b] \to \mathbb{R}$ is said to be (ν, α)-Hölder continuous if and only if there exist scalars $\nu > 0$ and $0 < \alpha \leq 1$, such that $|f(x) - f(y)| \leq \nu |x - y|^\alpha$, for all $x, y \in [a,b]$. (The value $\alpha = 1$ yields ν-Lipschitz continuous functions.)

Theorem 2.1 (Kolmogorov–Arnold superposition theorem)**.** *Let $n \geq 2$, $m \geq 2n$ and $\gamma \geq m + 2$ be given integers, and let a, λ_i $(1 \leq i \leq n)$, ν and α be defined as in Eqs.* (1)–(4)*. Then, there exists a (ν, α)-Hölder continuous inner function $\psi : [0,2) \to [0,2)$, such that for any continuous function $f : [0,1]^n \to \mathbb{R}$, there exist $m + 1$ continuous outer function $\phi_j : [0, 2\frac{\gamma-1}{\gamma-2}) \to \mathbb{R}$, such that*

$$f(x_1, \ldots, x_n) = \sum_{j=0}^{m} \phi_j \left(\sum_{i=1}^{n} \lambda_i \psi(x_i + ja) \right). \tag{5}$$

Let us comment on the two main steps of the constructive proof of Thm. 2.1; for details, see [11, 12, 14].

The first step is the building of the inner function ψ, which involves uniform grids D_k with step sizes γ^{-k},

$$D_k = \{ i\gamma^{-k},\ 0 \leq i \leq \gamma^k - 1 \} \subset [0,1).$$

There are γ^k different points $0 \leq d \leq 1 - \gamma^{-k} < 1$ on each grid D_k, and each point d on D_k is represented in base γ as follows,

$$d = \sum_{\ell=1}^{k} i_\ell \gamma^{-\ell}, \quad i_\ell \in \{0, 1, \ldots, \gamma - 1\}.$$

Proposition 2.2 (Construction of the inner function)**.** *The inner function ψ is first defined at grid points $d \in D_k$ via $\psi(d) = \psi_k(d)$ for all integers $k \geq 1$, where the functions ψ_k are recursively defined by*

$$\psi_k(d) = \begin{cases} d, & d \in D_1, \\ \psi_{k-1}\left(d - i_k\gamma^{-k}\right) + & \\ \quad i_k\gamma^{-\beta_n(k)}, & d \in D_k,\ k > 1,\ i_k < \gamma - 1, \\ \frac{1}{2}\Big[\psi_k\left(d - \gamma^{-k}\right) + & \\ \quad \psi_{k-1}\left(d + \gamma^{-k}\right)\Big], & d \in D_k,\ k > 1,\ i_k = \gamma - 1. \end{cases}$$

[1] We recall that $W = O(\epsilon^{-n})$ means that there exists $c_1(n) > 0$, such that $W \leq c_1(n)\epsilon^{-n}$, for sufficiently small values of ϵ. Alternatively, we shall write $\epsilon = O(W^{-1/n})$ when there exists $c_2(n) > 0$, such that $\epsilon \leq c_2(n)W^{-1/n}$, for sufficiently large values of W.

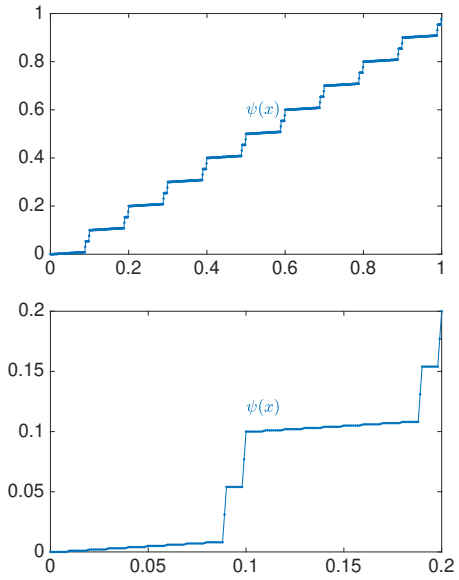

Figure 1: *Plot of the inner function ψ evaluated on the grid D_3 with $n = 2$ and $\gamma = 10$ (top). The second row is a zoomed plot that reveals the fractal structure of the graph of the Hölder continuous inner function ψ as $k \to \infty$.*

The function ψ is then defined at any $x \in [0, 1)$ via

$$\psi(x) = \lim_{k\to\infty} \psi_k\left(\sum_{\ell=1}^{k} i_\ell \gamma^{-\ell}\right),$$

since each $x \in [0, 1)$ has the representation

$$x = \sum_{\ell=1}^{\infty} i_\ell \gamma^{-\ell} = \lim_{k\to\infty} \sum_{\ell=1}^{k} i_\ell \gamma^{-\ell}.$$

Finally, the inner function is extended to $x \in [1, 2)$ by

$$\psi(x) = \psi(x - 1) + 1.$$

The resulting function has domain and range $[0, 2)$.

For points $d = \sum_{\ell=1}^{k} i_\ell \gamma^{-\ell} \in D_k$ whose indices i_ℓ are all strictly smaller than $\gamma - 1$, it is easy to show, by induction, that

$$\psi(d) = \sum_{\ell=1}^{k} i_\ell \gamma^{-\beta_n(\ell)}.$$

For other points, the right-hand side in the equation above is only a lower bound.

The inner function constructed in Prop. 2.2 was introduced by Köppen in 2002 [15]. It is Hölder continuous, a result that can be proved using the techniques introduced by Sprecher in his 1965 paper [5].

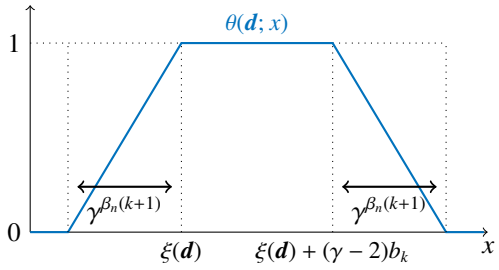

Figure 2: *For each $\boldsymbol{d} \in (D_k^j)^n$, the function $\theta(\boldsymbol{d}; \cdot)$ is compactly supported and piecewise linear with slope $\pm\gamma^{\beta_n(k+1)}$. Therefore, it is ν-Lipschitz continuous with $\nu = \gamma^{\beta_n(k+1)}$.*

Proposition 2.3 (Hölder continuity of the inner function)**.** *The inner function ψ of Prop. 2.2 is (ν, α)-Hölder continuous with $\nu = 2^{-\alpha}(\gamma + 3)$ and $\alpha = \log_\gamma 2$.*

Proof. See [5, Sec. 4]. □

We plot in Fig. 1 the graph of the function ψ evaluated on the grid D_3 for $n = 2$ and $\gamma = 10$. As $k \to \infty$, the graph of ψ approaches a fractal, which is expected since ψ is merely Hölder continuous.

The second step is the iterative construction of the outer functions ϕ_j. For each $0 \le j \le m$, let D_k^j denote the shifted grid defined by

$$D_k^j = D_k + j \sum_{\ell=2}^{k} \gamma^{-\ell}, \quad 0 \le j \le m.$$

Let $(D_k)^n$ and $(D_k^j)^n$ denote the Cartesian products of n copies of D_k and D_k^j, and let

$$\xi(\boldsymbol{d}) = \sum_{i=1}^{n} \lambda_i \psi(d_i), \quad \boldsymbol{d} = (d_1, \ldots, d_n) \in (D_k^j)^n,$$

and

$$b_k = \left(\sum_{\ell=k+1}^{\infty} \gamma^{-\beta_n(\ell)}\right)\left(\sum_{i=1}^{n} \lambda_i\right).$$

For each $\boldsymbol{d} \in (D_k^j)^n$, we define $\theta : x \mapsto \theta(\boldsymbol{d}; x)$ by

$$\theta(\boldsymbol{d}; x) = \sigma\left(\gamma^{\beta_n(k+1)}\left[x - \xi(\boldsymbol{d})\right] + 1\right) - \sigma\left(\gamma^{\beta_n(k+1)}\left[x - \xi(\boldsymbol{d}) - (\gamma - 2)b_k\right]\right),$$

where $\sigma : \mathbb{R} \to [0, 1]$ is the piecewise linear function defined by $\sigma(x) = 0$ for $x \le 0$, $\sigma(x) = x$ for $0 \le x \le 1$, and $\sigma(x) = 1$ for $x \ge 1$. For given $k \ge 1$ and $0 \le j \le m$, the γ^{nk} functions $\theta(\boldsymbol{d}; \cdot)$ have disjoint supports, and are ν-Lipschitz with $\nu = \gamma^{\beta_n(k+1)}$; see Fig. 2.

Proposition 2.4 (Construction of the outer functions)**.** *Let δ and η be two scalars that verify*

$$0 < \delta < 1 - \frac{n}{n-m+1}$$

and

$$0 < \frac{m-n+1}{n+1}\delta + \frac{2n}{m+1} \leq \eta < 1,$$

and $f : [0,1]^n \to \mathbb{R}$ be a continuous function.

Starting with $f_0 = 0$ and $e_0 = f - f_0 = f$, the approximate outer function ϕ_j^r at iteration $r \geq 1$ are defined, for each $0 \leq j \leq m$, as

$$\phi_j^r(x) = \frac{1}{m+1} \sum_{\ell=1}^{r} \sum_{\boldsymbol{d} \in (D_{k_\ell})^n} e_{\ell-1}(\boldsymbol{d}) \theta\left(\boldsymbol{d} + j \sum_{i=2}^{k_\ell} \gamma^{-i}; x\right),$$

for some $k_r = k_r(f)$ chosen such that $\|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \gamma^{-k_r}$ implies $|e_{r-1}(\boldsymbol{x}) - e_{r-1}(\boldsymbol{x}')| \leq \delta \|e_{r-1}\|_{L^\infty([0,1]^n)}$.

This yields an approximate function f_r,

$$f_r(x_1, \ldots, x_n) = \sum_{j=0}^{m} \phi_j^r \left(\sum_{i=1}^{n} \lambda_i \psi(x_i + ja) \right), \tag{6}$$

and its error $e_r = f - f_r$, with

$$\|e_r\|_{L^\infty([0,1]^n)} \leq \eta^r \|f\|_{L^\infty([0,1]^n)}. \tag{7}$$

Taking the limit $r \to \infty$ yields

$$f(x_1, \ldots, x_n) = \sum_{j=0}^{m} \phi_j \left(\sum_{i=1}^{n} \lambda_i \psi(x_i + ja) \right),$$

where $\phi_j = \lim_{r \to \infty} \phi_j^r$.

The approximate outer functions ϕ_j^r of Prop. 2.4 are Lipschitz continuous, as we shall prove next.

Proposition 2.5 (Lipschitz continuity of the outer functions)**.** *For all $r \geq 1$ and $0 \leq j \leq m$, the outer functions ϕ_j^r of Prop. 2.4 have domain $[0, 2\frac{\gamma-1}{\gamma-2})$, and are $\nu_r(f)$-Lipschitz continuous with*

$$\nu_r(f) = \frac{\|f\|_{L^\infty([0,1]^n)}}{m+1} \sum_{\ell=1}^{r} \eta^{\ell-1} \gamma^{\beta_n(k_\ell(f)+1)}. \tag{8}$$

Proof. To prove that the domain is $[0, 2\frac{\gamma-1}{\gamma-2})$, we use the fact that $|\psi(x)| < 2$ for all $x \in [0, 2)$, and

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i &< 1 + \frac{1}{\gamma - 1} + \frac{1}{\gamma^{1+n} - 1} + \frac{1}{\gamma^{1+n+n^2} - 1} + \ldots, \\ &< \frac{\gamma - 1}{\gamma - 2}. \end{aligned}$$

For the Lipschitz constant, we recall that, for given $k_\ell(f)$ and j, the functions $x \mapsto \theta(\boldsymbol{d}; x)$, $\boldsymbol{d} \in (D_{k_\ell}^j)^n$, have disjoint supports, and are $\nu(f)$-Lipschitz continuous with $\nu(f) = \gamma^{\beta_n(k_\ell(f)+1)}$. Using Eq. (7), summing over ℓ and multiplying by $1/(m+1)$ yields the desired result. □

Let us emphasize that the Lipschitz constants $\nu_r(f)$ in Prop. 2.5 depend on f via the integers $k_\ell(f)$, $1 \leq \ell \leq r$.

3. Approximation of the inner and outer functions by very deep ReLU networks

Let $\omega : [0, \infty) \to [0, \infty)$ be a function that is vanishing and continuous at 0, *i.e.*, $\lim_{\delta \to 0+} \omega(\delta) = \omega(0) = 0$, and $B \subset \mathbb{R}^d$ be a compact domain. We say that an uniformly continuous function $f : B \to \mathbb{R}$ has modulus of continuity ω if and only if

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq \omega(\|\boldsymbol{x} - \boldsymbol{x}'\|_2), \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in B.$$

Many classical estimates in approximation theory are based on moduli of continuity. For example, best degree-d polynomial approximation of continuous functions of one variable with modulus of continuity ω yields $O(\omega(d^{-1}))$ errors [29, Thm. 3.9]. The $O(\omega(d^{-1/n}))$ errors in dimension n suffers from the curse of dimensionality, but matches the lower bound obtained by nonlinear widths [30, Thm. 4.2].

In neural network approximation, moduli of continuity appear in the work of Yarotsky. In 2018, he proved that very deep ReLU networks of depth $L = O(W)$ and size W generate $O(\omega(O(W^{-2/n})))$ errors [28, Thm. 2]. This result matches the lower bound based on VC dimension of Anthony and Barlett [31, Thm. 8.7], and improves the $O(W^{-1/n} \log_2^{1/n} W)$ errors he obtained for Lipschitz functions in 2017 [27, Thm. 1].

Let us emphasize that Yarotsky's theorems provide upper bounds for the errors when the same network architecture is used to approximate all functions in a given function space. In other words, the network architecture does not depend on the function being approximated in that space; only the weights do. Moreover, the networks he utilizes are said to be *very deep* because the depth L satisfies $L = O(W)$. We recall his 2018 result below.

Theorem 3.1 (Approximation of continuous functions by very deep ReLU networks in the unit hypercube)**.** *For any continuous function $f : [0,1]^n \to \mathbb{R}$ with modulus of continuity ω_f, there is a deep ReLU network $\tilde{f}$ depth $L \leq c_0(n)W$ and size W, such that*

$$\|f - \tilde{f}\|_{L^\infty([0,1]^n)} \leq c_1(n) \omega_f\left(c_2(n) W^{-2/n}\right),$$

for some $c_0(n), c_1(n), c_2(n) > 0$.

We extend Yarotsky's result to domains $[0, M]^n$.

Corollary 3.2 (Approximation of continuous functions by very deep ReLU networks in scaled hypercubes)**.** *For any continuous function $f : [0, M]^n \to \mathbb{R}$ with modulus of continuity ω_f, there is a deep ReLU network $\tilde{f}$ of depth $L \leq c_0(n)W$ and size W, such that*

$$\|f - \tilde{f}\|_{L^\infty([0,M]^n)} \leq c_1(n)\omega_f\left(c_2(n)MW^{-2/n}\right),$$

with $c_0(n), c_1(n), c_2(n)$ as in Thm. 3.1.

Proof. We use Thm. 3.1 with $g(x) = f(x/M)$ on $[0, 1]^n$. Note that $\omega_g(\delta) = \omega_f(M\delta)$. Therefore, there is a deep ReLU network $\widetilde{g}$ of depth $L \leq c_0(n)W$ and size W, such that

$$\begin{aligned}\|g - \widetilde{g}\|_{L^\infty([0,1]^n)} &\leq c_1(n)\omega_g\left(c_2(n)W^{-2/n}\right),\\ &= c_1(n)\omega_f\left(c_2(n)MW^{-2/n}\right),\end{aligned}$$

with $c_0(n), c_1(n), c_2(n)$ as in Thm. 3.1. Since $g(Mx) - \widetilde{g}(Mx) = f(x) - \widetilde{g}(Mx)$, the network $\tilde{f}(x) = \widetilde{g}(Mx)$ satisfies all requirements in this corollary. □

We shall now apply Cor. 3.2 to the inner and outer functions of Props. 2.2 and 2.4. For simplicity, we shall assume, throughout the rest of the paper, that $m = 2n$ and $\gamma = 2n + 2$.

Proposition 3.3 (Approximation of the inner function by very deep ReLU networks)**.** *Let $n \geq 2$ be an integer and ψ be the inner function defined in Prop. 2.2. Then, for any scalar $0 < \epsilon < 1$, there is a deep ReLU network $\widetilde{\psi}$ that has depth $L \leq c_0(1)W$ and size*

$$W \leq c_3(n)\epsilon^{-[1+\log_2(n+1)]/2},$$

such that $\|\psi - \widetilde{\psi}\|_{L^\infty([0,2])} \leq \epsilon$, with

$$c_3(n) = [(2n + 5)c_1(1)]^{[1+\log_2(n+1)]/2}\, c_2(1)^{1/2}, \quad (9)$$

and $c_0(1), c_1(1), c_2(1)$ as in Thm. 3.1.

Proof. We use Cor. 3.2 with $M = 2$ and the modulus of continuity of Prop. 2.3, *i.e.*,

$$\omega_\psi(\delta) = \nu\delta^\alpha,$$

with $\nu = 2^{-\alpha}(2n + 5)$ and $\alpha = \log_{2n+2} 2$. □

Proposition 3.4 (Approximation of the outer functions by very deep ReLU networks)**.** *Let $n \geq 2$ be an integer, $f : [0, 1]^n \to \mathbb{R}$ be a continuous function that satisfies $\|f\|_{L^\infty([0,1]^n)} \leq 1$, and ϕ_j^r be the $(2n + 1)$ outer functions defined in Prop. 2.4 at iteration r, for some $r \geq 1$. Then, for any scalar $0 < \epsilon < 1$, there are $(2n + 1)$ deep ReLU networks $\widetilde{\phi}_j^r$ that have depth $L \leq c_0(1)W$ and size*

$$W \leq c_4(n, f)\epsilon^{-1/2},$$

such that $\|\phi_j^r - \widetilde{\phi}_j^r\|_{L^\infty([0,2\frac{\gamma-1}{\gamma-2}])} \leq \epsilon$, with

$$c_4(n, f) = \left[\frac{c_1(1)c_2(1)}{n}\sum_{\ell=1}^{r}\eta^{\ell-1}(2n + 2)^{\beta_n(k_\ell(f)+1)}\right]^{1/2}, \quad (10)$$

$c_0(1), c_1(1), c_2(1)$ as in Thm. 3.1, and η and $k_\ell(f)$ as in Prop. 2.4.

Proof. We use Cor. 3.2 with $M = 2\frac{\gamma-1}{\gamma-2}$ and the modulus of continuity corresponding to the Lipschitz continuity described in Prop. 2.5, *i.e.*,

$$\omega_{\phi_j^r}(\delta) = \nu_r(f)\delta,$$

with $\nu_r(f)$ as in Eq. (8). □

4. Main theorem

We present in this section our main theorem about the approximation of multivariate continuous functions by very deep ReLU networks. Our proof is based on the Kolmogorov–Arnold superposition theorem (Thm. 2.1), and on the approximation of the inner and outer functions by very deep ReLU networks (Props. 3.3 and 3.4).

Theorem 4.1 (Approximation of continuous functions using the Kolmogorov–Arnold superposition theorem)**.** *Let $n \geq 2$ be an integer and $f : [0, 1]^n \to \mathbb{R}$ be a continuous function that satisfies $\|f\|_{L^\infty([0,1]^n)} \leq 1$. Then, for any scalar $0 < \epsilon < 1$, there is a deep ReLU network $\tilde{f}_r$ that has depth*

$$L \leq c_0(1)\tilde{c}_3(n, f)\epsilon^{-[1+\log_2(n+1)]/2} + c_0(1)\tilde{c}_4(n, f)\epsilon^{-1/2},$$

and size

$$\begin{aligned}W \leq &(2n^2 + n)\tilde{c}_3(n, f)\epsilon^{-1+\log_2(n+1)]/2}\\ &+ (2n + 1)\tilde{c}_4(n, f)\epsilon^{-1/2},\end{aligned}$$

such that $\|f - \tilde{f}_r\|_{L^\infty([0,1]^n)} \leq \epsilon$, with $c_0(1)$ as in Thm. 3.1,

$$\begin{aligned}\tilde{c}_3(n, f) &= \left[\frac{4n + 2}{n}\sum_{\ell=1}^{r}\eta^{\ell-1}(2n+2)^{\beta_n(k_\ell(f)+1)}\right]^{[1+\log_2(n+1)]/2} c_3(n),\\ \tilde{c}_4(n, f) &= [8n + 4]^{1/2}\, c_4(n, f),\end{aligned}$$

$c_3(n), c_4(n, f)$ as in Eqs. (9) and (10), η and $k_\ell(f)$ as in Prop. 2.4, and $r = \lceil \log 2\epsilon^{-1} / \log \eta^{-1} \rceil$.

Proof. Let $0 < \epsilon < 1$ be a scalar and f be a continuous function that satisfies $\|f\|_{L^\infty([0,1]^n)} \leq 1$. Using Eq. (5) in Thm. 2.1, we write f as

$$f(x_1, \ldots, x_n) = \sum_{j=0}^{2n} \phi_j \left(\sum_{i=1}^{n} \lambda_i \psi(x_i + ja) \right).$$

We first approximate f by f_r using Eq. (6) in Prop. 3.4,

$$f_r(x_1, \ldots, x_n) = \sum_{j=0}^{2n} \phi_j^r \left(\sum_{i=1}^{n} \lambda_i \psi(x_i + ja) \right).$$

If we choose $r = \lceil \log 2\epsilon^{-1} / \log \eta^{-1} \rceil$ then, using Eq. (7), we get $\|f - f_r\|_{L^\infty([0,1]^n)} \leq \epsilon/2$. We denote by $k_\ell(f)$ the integer that defines the step size $\gamma^{-k_\ell(f)}$ at each iteration $1 \leq \ell \leq r$ of Sprecher's algorithm, as in Prop. 2.4.

We now approximate f_r by a deep ReLU network $\tilde{f}_r$ defined by

$$\tilde{f}_r(x_1, \ldots, x_n) = \sum_{j=0}^{2n} \widetilde{\phi}_j^r \left(\sum_{i=1}^{n} \lambda_i \widetilde{\psi}(x_i + ja) \right), \qquad (11)$$

where $\widetilde{\psi}$ and $\widetilde{\phi}_j^r$ approximate ψ and ϕ_j^r to some accuracies $0 < \epsilon_\psi < 1$ and $0 < \epsilon_\phi < 1$ to be determined later. We plot the subnetwork $\widetilde{\phi}_j^r$ in Fig. 3.

Using Props. 3.3 and 3.4, the network $\widetilde{\psi}$ has depth $L_\psi \leq c_0(n) W_\psi$ and size

$$W_\psi \leq c_3(n) \epsilon_\psi^{-[1+\log_2(n+1)]/2},$$

while the networks $\widetilde{\phi}_j^r$ have depth $L_\phi \leq c_0(n) W_\phi$ and size

$$W_\phi \leq c_4(n, f) \epsilon_\phi^{-1/2}.$$

Using the triangle inequality, we compute the accuracy of the network $\tilde{f}_r$ as follows,

$$\begin{aligned}
&|f_r(x_1, \ldots, x_n) - \tilde{f}_r(x_1, \ldots, x_n)|, \\
&\leq \left| \sum_{j=0}^{2n} \phi_j^r \left(\sum_{p=1}^{n} \lambda_i \psi(x_i + ja) \right) - \sum_{j=0}^{2n} \phi_j^r \left(\sum_{i=1}^{n} \lambda_i \widetilde{\psi}(x_i + ja) \right) \right| \\
&+ \left| \sum_{j=0}^{2n} \phi_j^r \left(\sum_{i=1}^{n} \lambda_i \widetilde{\psi}(x_i + ja) \right) - \sum_{j=0}^{2n} \widetilde{\phi}_j^r \left(\sum_{i=1}^{n} \lambda_i \widetilde{\psi}(x_i + ja) \right) \right|, \\
&\leq \frac{(2n+1)^2}{2n} \nu_r(f) \epsilon_\psi + (2n+1) \epsilon_\phi.
\end{aligned}$$

We must choose

$$\epsilon_\psi = \frac{n\epsilon}{2(2n+1)^2 \nu_r(f)}, \quad \epsilon_\phi = \frac{\epsilon}{4(2n+1)},$$

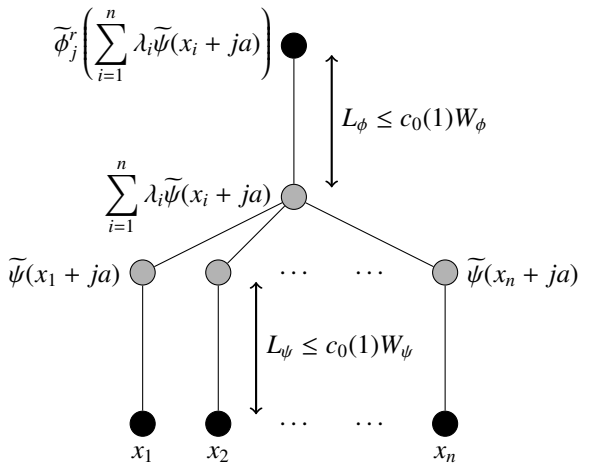

Figure 3: *Subnetwork that approximates the outer function ϕ_j at iteration r. The deep ReLU network in Eq.* (11) *is the sum of* $2n + 1$ *such subnetworks. Each subnetwork has depth $L_\psi + L_\phi$ and size $nW_\psi + W_\phi$, so that the network in Eq.* (11) *has depth $L_\psi + L_\phi$ and size $(2n^2 + n)W_\psi + (2n + 1)W_\phi$.*

to obtain $\|f_r - \tilde{f}_r\|_{L^\infty([0,1]^n)} \leq \epsilon/2$ and $\|f - \tilde{f}_r\|_{L^\infty([0,1]^n)} \leq \epsilon$.

Therefore, the network $\widetilde{\psi}$ has depth $L_\psi \leq c_0(n) W_\psi$ and size

$$W_\psi \leq \tilde{c}_3(n) \epsilon^{-\frac{1}{2}[1+\log_2(n+1)]},$$

with

$$\tilde{c}_3(n, f) = \left\lceil \frac{4n+2}{n} \sum_{j=1}^{r} \eta^{j-1} (2n+2)^{\beta_n(k_j(f)+1)} \right\rceil^{[1+\log_2(n+1)]/2} c_3(n),$$

while the networks $\widetilde{\phi}_j^r$ have depth $L_\phi \leq c_0(n) W_\phi$ and size

$$W_\phi \leq \tilde{c}_4(n, f) \epsilon^{-1/2}, \quad \tilde{c}_4(n, f) = [8n+4]^{1/2} c_4(n, f).$$

Lastly, the network $\tilde{f}_r$ has depth $L \leq c_0(1)(W_\psi + W_\phi)$ and size $W \leq n(2n+1)W_\psi + (2n+1)W_\phi$. □

The upper bounds in Thm. 4.1 show that, for a given dimension n, the depth and the size of the network grow like $O(\epsilon^{-\log n})$; the curse of dimensionality is lessened. Let us emphasize, however, that the constants $\tilde{c}_3(n, f)$ and $\tilde{c}_4(n, f)$ may be considerably large.

Let us also emphasize that the network architecture is adaptive, *i.e.*, the depth and the size of the network depend on f, via the constants $\tilde{c}_3(n, f)$ and $\tilde{c}_4(n, f)$. This explains why our estimates violate the $O(\epsilon^{-n})$ and $O(\epsilon^{-n/2})$ lower bounds obtained for a fixed architecture with continuous [30, Thm. 4.2] or discontinuous weight selection [31, Thm. 8.7].

5. Discussion

We have proven upper bounds for the approximation of multivariate continuous functions $f : [0, 1]^n \to \mathbb{R}$ by deep ReLU networks, for which the curse of dimensionality is lessened. The depth and the size of the networks to approximate such functions f grow like $O(\epsilon^{-\log n})$, as opposed to $O(\epsilon^{-n})$. The proof is based on the ability of very deep adaptive ReLU networks to implement the Kolmogorov–Arnold superposition theorem.

There are many ways in which this work could be profitably continued. If we were able to construct a Lipschitz continuous inner function, we would be able to obtain $O(\epsilon^{-1})$ estimates. Note that Actor and Knepley designed in 2017 an algorithm to compute a Lipschitz continuous inner function [32]. However, they did not provide a method to compute the outer functions. It would also be interesting to derive quantitative estimates for the integers k_ℓ that appear in the construction of the outer functions.

Acknowledgements

The research of the second author is supported by the start-up grant of the Department of Mathematics at the National University of Singapore and by the Ministry of Education in Singapore under the grant MOE2018-T2-2-147.

[1] A. N. Kolmogorov, On the representation of continuous functions of several variables by superposition of continuous functions of a smaller number of variables, Dokl. Akad. Nauk SSSR 108 (1956) 179–182.

[2] V. I. Arnold, On functions of three variables, Dokl. Akad. Nauk SSSR 114 (1957) 679–681.

[3] A. N. Kolmogorov, On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition, Dokl. Akad. Nauk SSSR 114 (1957) 953–956.

[4] G. G. Lorentz, Metric entropy, widths, and superposition of functions, Amer. Math. Montly 69 (1962) 469–485.

[5] D. A. Sprecher, On the structure of continuous functions of several variables, Trans. Amer. Math. Soc. 115 (1965) 340–355.

[6] B. L. Fridman, Improvement in the smoothness of functions in the Kolmogorov superposition theorem, Dokl. Akad. Nauk SSSR 177 (1967) 1019–1022.

[7] D. A. Sprecher, An improvement in the superposition theorem of Kolmogorov, J. Math. Anal. Appl. 38 (1972) 208–213.

[8] R. Hecht-Nielsen, Kolmogorov's mapping neural network existence theorem, in: Proceedings of the International Conference on Neural Networks, IEEE Press, New York, NY, 1987.

[9] V. Kůrková, Kolmogorov's theorem is relevant, Neural Comput. 3 (1991) 617–622.

[10] V. Kůrková, Kolmogorov's theorem and multilayer neural networks, Neural Netw. 5 (1992) 501–506.

[11] D. A. Sprecher, A numerical implementation of Kolmogorov's superpositions, Neural Netw. 9 (1996) 765–772.

[12] D. A. Sprecher, A numerical implementation of Kolmogorov's superpositions II, Neural Netw. 10 (1997) 447–457.

[13] B. Igelnik, N. Parikh, Kolmogorov's spline network, IEEE Trans. Neural Netw. 14 (2003) 725–733.

[14] J. Braun, M. Griebel, On a constructive proof of Kolmogorov's superposition theorem, Constr. Approx. 30 (2009) 653–675.

[15] M. Köppen, On the training of Kolmogorov network, in: J. R. Dorronsoro (Ed.), Artificial Neural Networks—ICANN 2002, Vol. 2415 of Lecture Notes in Computer Science, Springer, Berlin, 2002.

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

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

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

[19] 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.

[20] 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.

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

[22] H. Montanelli, H. Yang, Q. Du, Deep ReLU networks overcome the curse of dimensionality for bandlimited functions, arXiv:1903.00735.

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

[24] 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.

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

[26] 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.

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

[28] D. Yarotsky, Optimal approximation of continuous functions by very deep ReLU networks, in: S. Bubeck, V. Perchet, P. Rigollet (Eds.), 31st Annual Conference on Learning Theory, Proc. Mach. Learn. Res. 75, 2018, pp. 1–11.

[29] A. Gil, J. Segura, N. M. Temme, Numerical methods for special functions, SIAM, Philadelphia, PA, 2007.

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

[31] M. Anthony, P. L. Barlett, Neural network learning: Theoretical foundations, Cambridge University Press, Cambridge, UK, 2009.

[32] J. Actor, M. G. Knepley, An algorithm for computing Lipschitz inner functions in Kolmogorov's superposition theorem, arXiv:1712.08286.