SCIENCE CHINA
Mathematics

Progress of Projects Supported by NSFC
• ARTICLES •

October 2018 Vol. 61 No. 10: 1733–1740
https://doi.org/10.1007/s11425-018-9387-x

Exponential convergence of the deep neural network approximation for analytic functions

Dedicated to Professor TaTsien Li on the Occasion of His 80th Birthday

Weinan E[1,2,3] & Qingcan Wang[4]

[1]*Department of Mathematics and PACM, Princeton University, Princeton, NJ* 08544*, USA;*
[2]*Center for Big Data Research, Peking University, Beijing* 100871*, China;*
[3]*Beijing Institute of Big Data Research, Beijing* 100871*, China;*
[4]*Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ* 08544*, USA*

Email: weinan@math.princeton.edu, qingcanw@princeton.edu

Received July 5, 2018; accepted September 1, 2018; published online September 4, 2018

Abstract We prove that for analytic functions in low dimension, the convergence rate of the deep neural network approximation is exponential.

Keywords neural networks, approximation theory, analytic functions

MSC(2010) Primary 65D15; secondary 41-04

Citation: E W, Wang Q. Exponential convergence of the deep neural network approximation for analytic functions. Sci China Math, 2018, 61: 1733–1740, https://doi.org/10.1007/s11425-018-9387-x

1 Introduction

The approximation properties of deep neural network models are among the most tantalizing problems in machine learning. It is widely believed that deep neural network models are more accurate than shallow ones. Yet convincing theoretical support for such a speculation is still lacking. Existing work on the superiority of the deep neural network models are either for very special functions such as the examples given in [7], or special classes of functions such as the ones having a specific compositional structure. For the latter, the most notable are the results proved by Poggio et al. [6] that the approximation error for deep neural network models is exponentially better than the error for the shallow ones for a class of functions with specific compositional structure. However, given a general function f, one cannot calculate the distance from f to such class of functions. In the more general case, Yarotsky [8] considered C^β-differentiable functions, and proved that the number of parameters needed to achieve an error tolerance of ε is $\mathcal{O}(\varepsilon^{-\frac{d}{\beta}}\log\frac{1}{\varepsilon})$. Montanelli and Du [5] considered functions in the Koborov space. Using connection with sparse grids, they proved that the number of parameters needed to achieve an error tolerance of ε is $\mathcal{O}(\varepsilon^{-\frac{1}{2}}(\log\frac{1}{\varepsilon})^d)$.

For shallow networks, there has been a long history of proving the so-called universal approximation theorem, going back to the 1980s (see [2]). For networks with one hidden layer, Barron [1] established a convergence rate of $\mathcal{O}(n^{-\frac{1}{2}})$ where n is the number of hidden nodes. Such universal approximation theorems can also be proved for deep networks. Lu et al. [4] considered networks of width $d+4$ for

functions in d dimension, and proved that these networks can approximate any integrable function with sufficient number of layers. However, they did not give the convergence rate with respect to depth. To fill in this gap, we give a simple proof that the same kind of convergence rate for shallow networks can also be proved for deep networks.

The main purpose of this paper is to prove that for analytic functions, deep neural network approximations converge exponentially fast. The convergence rate deteriorates as a function of the dimensionality of the problem. Therefore the present result is only of significance in low dimension. However, this result does reveal a real superior approximation property of the deep networks for a wide class of functions.

Specifically, this paper contains the following contributions:

(1) We construct neural networks with fixed width $d+4$ to approximate a large class of functions, where the convergence rate can be established.

(2) For analytic functions, we obtain exponential convergence rate, i.e., the depth needed only depends on $\log\frac{1}{\varepsilon}$ instead of ε itself.

2 The setup of the problem

We begin by defining the network structure and the distance used in this paper, and proving the corresponding properties for the addition and composition operations.

We will use the following notation:

(1) Colon notation for subscript: Let $\{x_{m:n}\}=\{x_i : i=m,m+1,\dots,n\}$ and $\{x_{m_1:n_1,m_2:n_2}\}=\{x_{i,j} : i=m_1,m_1+1,\dots,n_1, j=m_2,m_2+1,\dots,n_2\}$.

(2) Linear combination: Denote $y\in\mathcal{L}(x_1,\dots,x_n)$ if there exist $\beta_i\in\mathbb{R}$, $i=1,\dots,n$, such that $y=\beta_0+\beta_1x_1+\cdots+\beta_nx_n$.

(3) Linear combination with ReLU activation: Denote $\tilde{y}\in\tilde{\mathcal{L}}(x_1,\dots,x_n)$ if there exists $y\in\mathcal{L}(x_1,\dots,x_n)$ and $\tilde{y}=\mathrm{ReLU}(y)=\max(y,0)$.

Definition 2.1. Given a function $f(x_1,\dots,x_d)$, if there exist variables $\{y_{1:L,1:M}\}$ such that

$$y_{1,m}\in\tilde{\mathcal{L}}(x_{1:d}),\quad y_{l+1,m}\in\tilde{\mathcal{L}}(x_{1:d},y_{l,1:M}),\quad f\in\mathcal{L}(x_{1:d},y_{1:L,1:M}),\tag{2.1}$$

where $m=1,\dots,M$, $l=1,\dots,L-1$, then f is said to be in the *neural nets class* $\mathcal{F}_{L,M}\left(\mathbb{R}^d\right)$, and $\{y_{1:L,1:M}\}$ is called a set of *hidden variables* of f.

A function $f\in\mathcal{F}_{L,M}$ can be regarded as a neural net with skip connections from the input layer to the hidden layers, and from the hidden layers to the output layer. This representation is slightly different from the one in standard fully-connected neural networks where connections only exist between adjacent layers. However, we can also easily represent such f using a standard network without skip connection.

Proposition 2.2. *A function $f\in\mathcal{F}_{L,M}\left(\mathbb{R}^d\right)$ can be represented by a ReLU network with depth $L+1$ and width $M+d+1$.*

Proof. Let $\{y_{1:L,1:M}\}$ be the hidden variables of f that satisfies (2.1), where

$$f=\alpha_0+\sum_{i=1}^{d}\alpha_ix_i+\sum_{l=1}^{L}\sum_{m=1}^{M}\beta_{l,m}y_{l,m}.$$

Consider the following variables $\{h_{1:L,1:M}\}$:

$$h_{l,1:M}=y_{l,1:M},\quad h_{l,M+1:M+d}=x_{1:d}$$

for $l=1,\dots,L$, and

$$h_{1,M+d+1}=\alpha_0+\sum_{i=1}^{d}\alpha_ix_i,\quad h_{l+1,M+d+1}=h_{l,M+d+1}+\sum_{m=1}^{M}\beta_{l,m}h_{l,m}$$

for $l=1,\dots,,L-1$. One can see that $h_{1,m}\in\tilde{\mathcal{L}}(x_{1:d})$, $h_{l+1,m}\in\tilde{\mathcal{L}}(h_{l,1:M+d+1})$, $m=1,\dots,M+d+1$, $l=1,\dots,L-1$, and $f\in\mathcal{L}(h_{L,1:M+d+1})$, which is a representation of a standard neural net. □

Proposition 2.3 (Addition and composition of neural net class $\mathcal{F}_{L,M}$). (1) *We have*

$$\mathcal{F}_{L_1,M}+\mathcal{F}_{L_2,M}\subseteq\mathcal{F}_{L_1+L_2,M}, \tag{2.2}$$

i.e., if $f_1\in\mathcal{F}_{L_1,M}\left(\mathbb{R}^d\right)$ *and* $f_2\in\mathcal{F}_{L_2,M}\left(\mathbb{R}^d\right)$, *then* $f_1+f_2\in\mathcal{F}_{L_1+L_2,M}$.

(2) *We have*

$$\mathcal{F}_{L_2,M}\circ\mathcal{F}_{L_1,M+1}\subseteq\mathcal{F}_{L_1+L_2,M+1}, \tag{2.3}$$

i.e., if $f_1(x_1,\dots,x_d)\in\mathcal{F}_{L_1,M+1}\left(\mathbb{R}^d\right)$ *and* $f_2(y,x_1,\dots,x_d)\in\mathcal{F}_{L_2,M}\left(\mathbb{R}^{d+1}\right)$, *then*

$$f_2(f_1(x_1,\dots,x_d),x_1,\dots,x_d)\in\mathcal{F}_{L_1+L_2,M+1}\left(\mathbb{R}^d\right).$$

Proof. For the addition property, denote the hidden variables of f_1 and f_2 as $\{y^{(1)}_{1:L_1,1:M}\}$ and $\{y^{(2)}_{1:L_2,1:M}\}$, respectively. Let

$$y_{1:L_1,1:M}=y^{(1)}_{1:L_1,1:M},\quad y_{L_1+1:L_1+L_2,1:M}=y^{(2)}_{1:L_2,1:M}.$$

By definition, $\{y_{1:L_1+L_2,1:M}\}$ is a set of hidden variables of f_1+f_2. Thus $f_1+f_2\in\mathcal{F}_{L_1+L_2,M}$.

For the composition property, denote the hidden variables of f_1 and f_2 as $\{y^{(1)}_{1:L_1,1:M+1}\}$ and $\{y^{(2)}_{1:L_2,1:M}\}$, respectively. Let

$$\begin{aligned}&y_{1:L_1,1:M+1}=y^{(1)}_{1:L_1,1:M+1},\quad y_{L_1+1:L_1+L_2,1:M}=y^{(2)}_{1:L_2,1:M},\\&y_{L_1+1,M+1}=y_{L_1+2,M+1}=\cdots=y_{L_1+L_2,M+1}=f_1(x_1,\dots,x_d).\end{aligned}$$

One can see that $\{y_{1:L_1+L_2,1:M+1}\}$ is a set of hidden variables of $f_2(f_1(\boldsymbol{x}),\boldsymbol{x})$, thus the composition property (2.3) holds. □

Definition 2.4. Given a continuous function $\varphi(\boldsymbol{x})$, $\boldsymbol{x}\in[-1,1]^d$ and a continuous function class $\mathcal{F}([-1,1]^d)$, define the L_∞ *distance*

$$\operatorname{dist}(\varphi,\mathcal{F})=\inf_{f\in\mathcal{F}}\max_{\boldsymbol{x}\in[-1,1]^d}|\varphi(\boldsymbol{x})-f(\boldsymbol{x})|. \tag{2.4}$$

Proposition 2.5 (Addition and composition properties for distance function). (1) *Let* φ_1 *and* φ_2 *be continuous functions. Let* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *be two continuous function classes. Then*

$$\operatorname{dist}(\alpha_1\varphi_1+\alpha_2\varphi_2,\mathcal{F}_1+\mathcal{F}_2)\leqslant|\alpha_1|\operatorname{dist}(\varphi_1,\mathcal{F}_1)+|\alpha_2|\operatorname{dist}(\varphi_2,\mathcal{F}_2), \tag{2.5}$$

where α_1 *and* α_2 *are two real numbers.*

(2) *Assume that* $\varphi_1(\boldsymbol{x})=\varphi_1(x_1,\dots,x_d)$, $\varphi_2(y,\boldsymbol{x})=\varphi_2(y,x_1,\dots,x_d)$ *satisfy* $\varphi_1([-1,1]^d)\subseteq[-1,1]$. *Let* $\mathcal{F}_1([-1,1]^d)$ *and* $\mathcal{F}_2([-1,1]^{d+1})$ *be two continuous function classes. Then*

$$\operatorname{dist}(\varphi_2(\varphi_1(\boldsymbol{x}),\boldsymbol{x}),\mathcal{F}_2\circ\mathcal{F}_1)\leqslant L_{\varphi_2}\operatorname{dist}(\varphi_1,\mathcal{F}_1)+\operatorname{dist}(\varphi_2,\mathcal{F}_2), \tag{2.6}$$

where L_{φ_2} *is the Lipschitz norm of* φ_2 *with respect to* y.

Proof. The additional property obviously holds. Now we prove the composition property. For any $f_1\in\mathcal{F}_1$ and $f_2\in\mathcal{F}_2$, one has

$$\begin{aligned}&|\varphi_2(\varphi_1(\boldsymbol{x}),\boldsymbol{x})-f_2(f_1(\boldsymbol{x}),\boldsymbol{x})|\\&\quad\leqslant|\varphi_2(\varphi_1(\boldsymbol{x}),\boldsymbol{x})-\varphi_2(f_1(\boldsymbol{x}),\boldsymbol{x})|+|\varphi_2(f_1(\boldsymbol{x}),\boldsymbol{x})-f_2(f_1(\boldsymbol{x}),\boldsymbol{x})|\\&\quad\leqslant L_{\varphi_2}\|\varphi_1(\boldsymbol{x})-f_1(\boldsymbol{x})\|_\infty+\|\varphi_2(y,\boldsymbol{x})-f_2(y,\boldsymbol{x})\|_\infty.\end{aligned}$$

Take $f_1^\star=\operatorname{argmin}_f\|\varphi_1(\boldsymbol{x})-f(\boldsymbol{x})\|_\infty$ and $f_2^\star=\operatorname{argmin}_f\|\varphi_2(y,\boldsymbol{x})-f(y,\boldsymbol{x})\|_\infty$. Then

$$|\varphi_2(\varphi_1(\boldsymbol{x}),\boldsymbol{x})-f_2^\star(f_1^\star(\boldsymbol{x}),\boldsymbol{x})|\leqslant L_{\varphi_2}\operatorname{dist}(\varphi_1,\mathcal{F}_1)+\operatorname{dist}(\varphi_2,\mathcal{F}_2).$$

Thus the composition property (2.6) holds. □

Now we are ready to state the main theorem for the approximation of analytic functions.

Theorem 2.6. *Let f be an analytic function over $(-1,1)^d$. Assume that the power series $f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{N}^d} a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}$ is absolutely convergent in $[-1,1]^d$. Then for any $\delta > 0$, there exists a function $\hat{f}$ that can be represented by a deep ReLU network with depth L and width $d+4$, such that*

$$|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| < 2\sum_{\boldsymbol{k}\in\mathbb{N}^d} |a_{\boldsymbol{k}}| \cdot \exp(-d\delta(\mathrm{e}^{-1}L^{\frac{1}{2d}} - 1)) \tag{2.7}$$

for all $\boldsymbol{x} \in [-1+\delta, 1-\delta]^d$.

3 Proof

The construction of $\hat{f}$ is motivated by the approximation of the square function $\varphi(x) = x^2$ and multiplication function $\varphi(x,y) = xy$ proposed in [3, 8]. We use this as the basic building block to construct approximations to monomials, polynomials, and analytic functions.

Lemma 3.1. *The function $\varphi(x) = x^2$, $x \in [-1,1]$ can be approximated by deep neural nets with an exponential convergence rate*:

$$\mathrm{dist}(\varphi = x^2, \mathcal{F}_{L,2}) \leqslant 2^{-2L}. \tag{3.1}$$

Proof. Consider the function

$$g(y) = \begin{cases} 2y, & 0 \leqslant y < 1/2, \\ 2(1-y), & 1/2 \leqslant y \leqslant 1. \end{cases}$$

Then $g(y) = 2y - 4\mathrm{ReLU}(y - 1/2)$ in $[0,1]$. Define the hidden variables $\{y_{1:L,1:2}\}$ as follows:

$$\begin{aligned} &y_{1,1} = \mathrm{ReLU}(x), \quad y_{1,2} = \mathrm{ReLU}(-x), \\ &y_{2,1} = \mathrm{ReLU}(y_{1,1} + y_{1,2}), \quad y_{2,2} = \mathrm{ReLU}(y_{1,1} + y_{1,2} - 1/2), \\ &y_{l+1,1} = \mathrm{ReLU}(2y_{l,1} - 4y_{l,2}), \quad y_{l+1,2} = \mathrm{ReLU}(2y_{l,1} - 4y_{l,2} - 1/2) \end{aligned}$$

for $l = 2, 3, \ldots, L-1$. Using induction, one can see that $|x| = y_{1,1} + y_{1,2}$ and

$$g_l(|x|) = \underbrace{g \circ g \circ \cdots \circ g}_{l}(|x|) = 2y_{l+1,1} - 4y_{l+1,2}, \quad l = 1, \ldots, L-1$$

for $x \in [-1,1]$. Now let

$$f(x) = |x| - \sum_{l=1}^{L-1} \frac{g_l(|x|)}{2^{2l}}.$$

Then $f \in \mathcal{F}_{L,2}$, and $|x^2 - f(x)| \leqslant 2^{-2L}$ for $x \in [-1,1]$. □

Lemma 3.2. *For the multiplication function $\varphi(x,y) = xy$, we have*

$$\mathrm{dist}(\varphi = xy, \mathcal{F}_{3L,2}) \leqslant 3 \cdot 2^{-2L}. \tag{3.2}$$

Proof. Notice that

$$\varphi = xy = 2\left(\frac{x+y}{2}\right)^2 - \frac{1}{2}x^2 - \frac{1}{2}y^2.$$

Applying the addition properties (2.2), (2.5) and Lemma 3.1, we obtain (3.2). □

Now we use the multiplication function as the basic block to construct monomials and polynomials.

Lemma 3.3. *For a monomial $M_p(\boldsymbol{x})$ of d variables with degree p, we have*

$$\mathrm{dist}(M_p, \mathcal{F}_{3(p-1)L,3}) \leqslant 3(p-1) \cdot 2^{-2L}. \tag{3.3}$$

Proof. Let

$$M_p(\boldsymbol{x}) = x_{i_1}x_{i_2}\cdots x_{i_p}, \quad i_1,\ldots,i_p \in \{1,\ldots,d\}.$$

Using induction, we assume that the lemma holds for the degree-p monomial M_p, and consider a degree-$(p+1)$ monomial $M_{p+1}(\boldsymbol{x}) = M_p(\boldsymbol{x})\cdot x_{i_{p+1}}$. Let $\varphi(y,x) = yx$. Then $M_{p+1}(\boldsymbol{x}) = \varphi(M_p(\boldsymbol{x}), x_{i_{p+1}})$. From the composition properties (2.3), (2.6) and Lemma 3.2, we have

$$\begin{aligned}\operatorname{dist}(M_{p+1}, \mathcal{F}_{3pL,3}) &\leqslant \operatorname{dist}(\varphi(M_p(\boldsymbol{x}), x_{i_{p+1}}), \mathcal{F}_{3L,2}\circ\mathcal{F}_{3(p-1)L,3})\\ &\leqslant L_\varphi \operatorname{dist}(M_p, \mathcal{F}_{3(p-1)L,3}) + \operatorname{dist}(\varphi, \mathcal{F}_{3L,2})\\ &\leqslant 3p\cdot 2^{-2L}.\end{aligned}$$

Note that the Lipschitz norm $L_\varphi = 1$ since $x_{i_{p+1}} \in [-1,1]$. □

Lemma 3.4. *For a degree-p polynomial*

$$P_p(\boldsymbol{x}) = \sum_{|\boldsymbol{k}|\leqslant p} a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}, \quad \boldsymbol{x}\in[-1,1]^d, \quad \boldsymbol{k} = (k_1,\ldots,k_d)\in\mathbb{N}^d,$$

we have

$$\operatorname{dist}(P_p, \mathcal{F}_{\binom{p+d}{d}(p-1)L,3}) < 3(p-1)\cdot 2^{-2L}\sum_{|\boldsymbol{k}|\leqslant p}|a_{\boldsymbol{k}}|. \tag{3.4}$$

Proof. The lemma can be proved by applying the addition properties (2.2), (2.5) and Lemma 3.3.

Note that the number of monomials of d variables with degree less than or equal to p is $\binom{p+d}{d}$. □

Now we are ready to prove Theorem 2.6.

Proof of Theorem 2.6. Let

$$\varepsilon = \exp(-d\delta(\mathrm{e}^{-1}L^{\frac{1}{2d}} - 1)).$$

Then $L = [\mathrm{e}(\frac{1}{d\delta}\log\frac{1}{\varepsilon}+1)]^{2d}$. Without loss of generality, assume $\sum_{\boldsymbol{k}}|a_{\boldsymbol{k}}| = 1$. We will show that there exists $\hat{f}\in\mathcal{F}_{L,3}$ such that $\|f-\hat{f}\|_\infty < 2\varepsilon$.

Denote

$$f(\boldsymbol{x}) = P_p(\boldsymbol{x}) + R(\boldsymbol{x}) := \sum_{|\boldsymbol{k}|\leqslant p} a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}} + \sum_{|\boldsymbol{k}|>p} a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}.$$

For $\boldsymbol{x}\in[-1+\delta,1-\delta]^d$, we have $|R(\boldsymbol{x})| < (1-\delta)^p$, thus truncation to $p = \frac{1}{\delta}\log\frac{1}{\varepsilon}$ will ensure $|R(\boldsymbol{x})| < \varepsilon$.

From Lemma 3.4, we have $\operatorname{dist}(P_p, \mathcal{F}_{L,3}) < 3(p-1)\cdot 2^{-2L'}$, where

$$\begin{aligned}L' &= L\binom{p+d}{p}^{-1}(p-1)^{-1} > L\left[\frac{(p+d)^d}{(d/e)^d}\right]^{-1}p^{-1}\\ &= L\left[\mathrm{e}\left(\frac{1}{d\delta}\log\frac{1}{\varepsilon}+1\right)\right]^{-d}\left(\frac{1}{\delta}\log\frac{1}{\varepsilon}\right)^{-1} = \left[\mathrm{e}\left(\frac{1}{d\delta}\log\frac{1}{\varepsilon}+1\right)\right]^{d}\left(\frac{1}{\delta}\log\frac{1}{\varepsilon}\right)^{-1}\\ &\gg \log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\end{aligned}$$

for $d\geqslant 2$ and $\varepsilon \ll 1$, and then

$$\operatorname{dist}(P_p, \mathcal{F}_{L,3}) < 3(p-1)\cdot 2^{-2L'} \ll \varepsilon.$$

In the case of $d = 1$, note that the polynomial

$$P_p(x) = a_0 + a_1x + \cdots + a_px^p = a_0 + x(a_1 + x(\cdots(a_{p-1}+a_px)\cdots)).$$

Following the proof of Lemma 3.3, one can see that

$$\operatorname{dist}(P_p, \mathcal{F}_{3(p-1)L',3}) \leqslant 3(p-1)\cdot 2^{-2L'}$$

for $d=1$ and $\sum_{k=1}^{p}|a_k|\leqslant 1$. Thus

$$L'=\frac{L}{3(p-1)}=\frac{1}{3}\left[\mathrm{e}\left(\frac{1}{\delta}\log\frac{1}{\varepsilon}+1\right)\right]^2\left(\frac{1}{\delta}\log\frac{1}{\varepsilon}-1\right)^{-1}\gg\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}$$

still holds, and $\mathrm{dist}(P_p,\mathcal{F}_{L,3})\ll\varepsilon$.

Therefore, there exists $\hat{f}\in\mathcal{F}_{L,3}$ such that $\|P_p-\hat{f}\|_\infty<\varepsilon$, and

$$\|f-\hat{f}\|_\infty\leqslant\|f-P_p\|_\infty+\|P_p-\hat{f}\|_\infty<2\varepsilon.$$

The proof is completed. □

One can also formulate Theorem 2.6 as follows:

Corollary 3.5. *Assume that the analytic function $f(\boldsymbol{x})=\sum_{\boldsymbol{k}\in\mathbb{N}^d}a_{\boldsymbol{k}}\boldsymbol{x}^{\boldsymbol{k}}$ is absolutely convergent in $[-1,1]^d$. Then for any $\varepsilon,\delta>0$, there exists a function $\hat{f}$ that can be represented by a deep ReLU network with depth $L=[\mathrm{e}(\frac{1}{d\delta}\log\frac{1}{\varepsilon}+1)]^{2d}$ and width $d+4$, such that $|f(\boldsymbol{x})-\hat{f}(\boldsymbol{x})|<2\varepsilon\sum_{\boldsymbol{k}}|a_{\boldsymbol{k}}|$ for all $\boldsymbol{x}\in[-1+\delta,1-\delta]^d$.*

4 The convergence rate for the general case

Here, we prove that for deep neural networks, the approximation error decays like $\mathcal{O}((N/d)^{-\frac{1}{2}})$ where N is the number of parameters in the model. The proof is quite simple but the result does not seem to be available in the existing literature.

Theorem 4.1. *Given a function $f:\mathbb{R}^d\to\mathbb{R}$ with the Fourier representation*

$$f(\boldsymbol{x})=\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{x}}\hat{f}(\boldsymbol{\omega})d\boldsymbol{\omega},$$

and a compact domain $B\subset\mathbb{R}^d$ containing 0, let

$$C_{f,B}=\int_B|\boldsymbol{\omega}|_B|\hat{f}(\boldsymbol{\omega})|d\boldsymbol{\omega},$$

where $|\boldsymbol{\omega}|_B=\sup_{\boldsymbol{x}\in B}|\boldsymbol{\omega}\cdot\boldsymbol{x}|$. Then there exists a ReLU network $f_{L,M}$ with width $M+d+1$ and depth L, such that

$$\int_B|f(\boldsymbol{x})-f_{L,M}(\boldsymbol{x})|^2d\mu(\boldsymbol{x})\leqslant\frac{8C_{f,B}^2}{ML},\tag{4.1}$$

where μ is an arbitrary probability measure.

Here, the number of parameters N satisfies

$$N=(d+1)(M+d+1)+(M+d+2)(M+d+2)(L-1)+(M+d+2)=\mathcal{O}((M+d)^2L).$$

Taking $M=d$, we will have $L=\mathcal{O}(N/d^2)$ and the convergence rate becomes

$$\mathcal{O}((ML)^{-\frac{1}{2}})=\mathcal{O}((N/d)^{-\frac{1}{2}}).$$

Note that in the universal approximation theorem for shallow networks with one hidden layer, one can prove the same convergence rate

$$\mathcal{O}(n^{-\frac{1}{2}})=\mathcal{O}((N/d)^{-\frac{1}{2}}).$$

Here, n is the number of hidden nodes and $N=(d+2)n+1$ is the number of parameters.

Theorem 4.1 is a direct consequence of the following theorem by Barron [1] for networks with one hidden layer and sigmoidal type of the activation function. Here, a function σ is *sigmoidal* if it is bounded measurable on $\mathbb{R}$ with $\sigma(+\infty)=1$ and $\sigma(-\infty)=0$.

Theorem 4.2. *Given a function f and a domain B such that $C_{f,B}$ is finite, given a sigmoidal function σ, there exists a linear combination*

$$f_n(\boldsymbol{x}) = \sum_{j=1}^{n} c_j \sigma(\boldsymbol{a}_j \cdot \boldsymbol{x} + b_j) + c_0, \quad \boldsymbol{a}_j \in \mathbb{R}^d, \quad b_j, c_j \in \mathbb{R},$$

such that

$$\int_B |f(\boldsymbol{x}) - f_n(\boldsymbol{x})|^2 d\mu(\boldsymbol{x}) \leqslant \frac{4C_{f,B}^2}{n}. \tag{4.2}$$

Notice that

$$\sigma(z) = \mathrm{ReLU}(z) - \mathrm{ReLU}(z-1)$$

is sigmoidal, so we have the following corollary.

Corollary 4.3. *Given a function f and a set B with $C_{f,B}$ finite, there exists a linear combination of n ReLU functions*

$$f_n(\boldsymbol{x}) = \sum_{j=1}^{n} c_j \mathrm{ReLU}(\boldsymbol{a}_j \cdot \boldsymbol{x} + b_j) + c_0,$$

such that

$$\int_B |f(\boldsymbol{x}) - f_n(\boldsymbol{x})|^2 d\mu(\boldsymbol{x}) \leqslant \frac{8C_{f,B}^2}{n}.$$

Next, we convert this shallow network to a deep one.

Lemma 4.4. *Let $f_n : \mathbb{R}^d \to \mathbb{R}$ be a ReLU network with one hidden layer (as shown in the previous corollary). For any decomposition $n = m_1 + \cdots + m_L$, $n_k \in \mathbb{N}^\star$, f_n can also be represented by a ReLU network with L hidden layers, where the l-th layer has $m_l + d + 1$ nodes.*

Proof. Denote the input by $\boldsymbol{x} = (x_1, \ldots, x_d)$. We construct a network with L hidden layers in which the l-th layer has $m_l + d + 1$ nodes $\{h_{l,1:m_l+d+1}\}$. Similar to the construction in Proposition 2.2, let

$$h_{L,1:d} = h_{L-1,1:d} = \cdots = h_{1,1:d} = x_{1:d}, \quad h_{l,d+j} = \mathrm{ReLU}(\boldsymbol{a}_{l,j} \cdot \boldsymbol{x} + b_{l,j})$$

for $j = 1, \ldots, m_l$, $l = 1, \ldots, L$, and

$$h_{1,d+m_1+1} = c_0, \quad h_{l+1,d+m_{l+1}+1} = h_{l,d+m_l+1} + \sum_{j=1}^{m_l} c_{l,j} h_{l,d+j}$$

for $l = 1, \ldots, L-1$. Here, we use the notation $\boldsymbol{a}_{l,j} = \boldsymbol{a}_{m_1+\cdots+m_{l-1}+j}$ (the same for $b_{l,j}$ and $c_{l,j}$). One can see that

$$h_{1,m} \in \tilde{\mathcal{L}}(x_{1:d}), \quad h_{l+1,m} \in \tilde{\mathcal{L}}(h_{l,1:d+m_l+1}), \quad m = 1, \ldots, d + m_l + 1, \quad l = 1, \ldots, L-1$$

and

$$h_{l,d+m_l+1} = c_0 + \sum_{j=1}^{m_1+\cdots+m_{l-1}} c_j \mathrm{ReLU}(\boldsymbol{a}_j \cdot \boldsymbol{x} + b_j).$$

Thus

$$f_n = h_{L,d+m_L+1} + \sum_{j=1}^{m_L} c_{L,j} h_{L,d+j} \in \mathcal{L}(h_{L,1:d+m_L+1})$$

can be represented by this deep network. □

Now consider a network with L layers where each layer has the same width $M+d+1$. From Lemma 4.4, this network is equivalent to a one-layer network with ML hidden nodes. Applying Corollary 4.3, we obtain the desired approximation result for deep networks stated in Theorem 4.1.

Acknowledgements This work was supported by Office of Naval Research (ONR) (Grant No. N00014-13-1-0338) and Major Program of National Natural Science Foundation of China (Grant No. 91130005). The authors are grateful to Chao Ma for very helpful discussions during the early stage of this work. The authors are also grateful to Jinchao Xu for his interest, which motivated us to write up this paper.

References

1 Barron A R. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans Inform Theory, 1993, 39: 930–945

2 Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signal Syst, 1989, 2: 303–314

3 Liang S, Srikant R. Why deep neural networks for function approximation? ArXiv:1610.04161, 2016

4 Lu Z, Pu H, Wang F, et al. The expressive power of neural networks: A view from the width. In: Advances in Neural Information Processing Systems, vol. 30. Long Beach: Neural Information Processing Systems Foundation, 2017, 6232–6240

5 Montanelli H, Du Q. Deep ReLU networks lessen the curse of dimensionality. ArXiv:1712.08688, 2017

6 Poggio T, Mhaskar H, Rosasco L, et al. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. Internat J Automat Comput, 2017, 14: 503–519

7 Telgarsky M. Benefits of depth in neural networks. ArXiv:1602.04485, 2016

8 Yarotsky D. Error bounds for approximations with deep relu networks. Neural Networks, 2017, 94: 103–114