Neural Networks 23 (2010) 618–624

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

New study on neural networks: The essential order of approximation☆

Jianjun Wang[a,b], Zongben Xu[b,*]

[a] School of Mathematics & Statistics, Southwest University, Chongqing, 400715, PR China

[b] Institute for Information and System Science, Xi'an Jiaotong University, Xi'an, Shaan'xi, 710049, PR China

ARTICLE INFO

Article history:
Received 1 December 2008
Received in revised form 7 September 2009
Accepted 15 January 2010

Keywords:
The essential order of approximation
Nearly exponential type neural networks
Modulus of smoothness

ABSTRACT

For the nearly exponential type of feedforward neural networks (neFNNs), the essential order of their approximation is revealed. It is proven that for any continuous function defined on a compact set of R^d, there exist three layers of neFNNs with the fixed number of hidden neurons that attain the essential order. Under certain assumption on the neFNNs, the ideal upper bound and lower bound estimations on approximation precision of the neFNNs are provided. The obtained results not only characterize the intrinsic property of approximation of the neFNNs, but also proclaim the implicit relationship between the precision (speed) and the number of hidden neurons of the neFNNs.

1. Introduction

Artificial neural networks have been extensively applied in various fields of science and engineering. It is mainly because feedforward neural networks (FNNs) have universal approximation capability (Attali & Pages, 1997; Cardaliaguet & Euvrard, 1992; Chen, 1994; Chen & Chen, 1995; Chui & Li, 1992, 1993; Cybenko, 1989; Funahashi, 1989; Hornik, Stinchombe, & White, 1989, 1990; Leshno, Lin, Pinks, & Schocken, 1993). A typical example of such universal approximation assertions states that for any given continuous function defined on a compact set $\mathbf{K}$ of $\mathcal{R}^d$, there exists a three-layer of FNN such that it can approximate the function arbitrarily well. A three-layer of FNN with one hidden layer, d inputs and one output can be mathematically expressed as

$$\mathcal{N}(x) = \sum_{i=1}^{m} c_i \sigma\left(\sum_{j=1}^{d} w_{ij} x_j + \theta_i\right), \quad x \in \mathcal{R}^d,\ d \geq 1, \tag{1.1}$$

where $1 \leq i \leq m, \theta_i \in \mathcal{R}$ are the thresholds, $w_i = (w_{i1}, w_{i2}, \ldots, w_{id})^T \in \mathcal{R}^d$ are connection weights of neuron i in the hidden layer with the input neurons, $c_i \in \mathcal{R}$ are the connection strength of neuron i with the output neuron, and σ is the activation function used in the network. The activation function is normally taken as sigmoid type, that is, it satisfies $\sigma(t) \to 1$ as $t \to +\infty$ and $\sigma(t) \to 0$ as $t \to -\infty$. Eq. (1.1) can be further expressed in vector form as

$$\mathcal{N}(x) = \sum_{i=1}^{m} c_i \sigma\left(\mathbf{w}_i \cdot \mathbf{x} + \theta_i\right), \quad x \in R^d.$$

The approximation of multivariate functions by the FNNs (1.1) has been widely studied in past years, with various significant results, concerning density or complexity. For instance, it was proven in Cybenko (1989) that under very mild conditions on the sigmoidal activation function σ, any continuous function defined on a compact set $\mathbf{K}$ of $\mathcal{R}^d$ can be approximated arbitrarily well by the FNNs (1.1). Later, various density and complexity results on approximation of the functions by the FNNs (1.1) were established by many authors and by using different approaches for more or less general situations (Chen, 1994; Funahashi, 1989; Hornik et al., 1990; Yoshifusa, 1991). All these researches are qualitative in feature. From the perspective of application, however, the quantitative research on approximation of the neural networks is more helpful. Particularly, one would like to know what is the degree of the neural networks to approximate a certain type of functions, and how fast they approximate. Also one would like to know how the approximation capability of a neural network is related to the topology of the network (say, how many hidden neurons are needed in order for the network to reach a predetermined approximation precision?). There have been many authors who studied those problems (Barron, 1993; Cao & Xu, 2001; Kůrkova, Kainen, & Kreinovich, 1997; Maiorov & Meir, 1998; Mhaskar, 1996; Mhaskar & Khachikyan, 1995; Mhaskar & Micchelli, 1992, 1994, 1995; Ritter, 1994). The results obtained have basically dealt with the neural networks with logistic (including sigmoid) and no Heaviside step

☆ Supported by National Basic Research Program of China (Grant Nos. 2007CB311000), Natural Science Foundation of China (Grant Nos. 10726040 and 10701062), the Key Project of Chinese Ministry of Education. (No. 108176), China Postdoctoral Science Foundation (No. 20080431237), Natural Science Foundation Project of CQ CSTC (No.CSTC,2009BB2306), Southwest University (China) Development Foundation of China (No. SWUF2007014) and Southwest University Doctoral Foundation of China (No. SWUB2007006).

* Corresponding author.
E-mail address: zbxu@mail.xjtu.edu.cn (Z. Xu).

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.01.004

activation functions. They have offered, however, only certain kind of upper bound estimations on approximation of the neural networks. In those researches, Suzuki (1998) obtained, by using a constructive approach, an upper bound estimation on approximation of the neural networks and explicitly calculated the number of hidden neurons needed for guaranteeing the predetermined approximation precision.

The upper bound estimation results can imply convergence of the neural networks to the function to be approximated and also provide quantitative measurement on how accurate the neural networks approximate the functions. The estimations cannot, however, precisely characterize the essential order of the networks, that is, they cannot decipher the highest approximation accuracy of the neural networks to achieve (Xu & Cao, 2004). In order to get such an essential order of approximation of a neural network, besides upper bound estimation, a lower bound estimation that characterizes the worst approximation precision of the network can also be needed. The essential order of approximation can be obtained when and only when the upper and the lower bound estimations are of the same order. Clearly, obtaining the essential order of a neural network is not easy, but very important, and is significant. In Xu and Cao (2004), such problem was tackled for the neural networks (1.1) when the activation function is sigmoidal and satisfies some other conditions. In the present paper, our aim is to tackle the same problem for more broader types of neural networks when the activation function is nearly exponential. The class of the nearly exponential functions was introduced by Ritter (1994), which are those functions that approximate the exponential function arbitrarily well on the negative half line through appropriate affine transformation at the origin and in the target space (for the more precise definition, see the next section). It includes the normal sigmoid and exponential functions as special cases. A neural network with the form (1.1) henceforth will be called a nearly exponential FNN (denoted by neFNN in brief) whenever the activation function is nearly exponential.

In Ritter (1999), an upper bound estimation on approximation of the neFNNs was developed, but it did not offer any estimation on lower bound of approximation. Consequently, uncovering the essential order of approximation of the neFNNs is still open. In Xu and Wang (2006), we resolved the problem through developing a more precise upper bound estimation on approximation of the FNNs first, and then provided a lower bound estimation of the approximation. Finally, we characterized the conditions under which the upper and the lower bounds have the same order, from which the essential order of the neFNNs will be revealed. In this paper, we characterize accurately the topology selection and the essential order of neFNNs under some conditions; the results obtained not only sharpen the results developed in Ritter (1999), but also clarify the relationship between the approximation speed (precision) and the number of hidden neurons needed for the neural networks.

The remainder of this paper is organized as follows. In Section 2, we present some notations, some basic definitions, the main results and briefly review their significations. Section 3 summarizes some of our previous work and offers two fundamental results for proving our results. In Section 4, the proof of the main results is given by some techniques of approximation theory. Section 5 provides some concluding remarks.

2. Notation and main results

We begin with some comments concerning notation. The symbols $\boldsymbol{N}$, $\mathbf{R}$ and $\mathbf{R}_+$ stand respectively for the sets of nonnegative integers, real numbers, and nonnegative real numbers. We use R^d which denotes d-dimensional Euclidean space ($d \geq 1$). All operations on vectors are taken component wisely. For instance, when

$$x = (x_1, x_2, \ldots, x_d) \in R^d, \qquad y = (y_1, y_2, \ldots, y_d) \in R^d,$$

we define

$$e^x = (e^{x_1}, e^{x_2}, \ldots, e^{x_d}), \qquad xy = (x_1y_1, x_2y_2, \ldots, x_dy_d)$$

and whenever $x_i \geq 0$, $i = 1, 2, \ldots, d$, we define

$$x^y = (x_1^{y_1}, x_2^{y_2}, \ldots, x_d^{y_d}).$$

We use $P_n(d)$ and $T_n(d)$ which respectively denote the spaces of all d-variate algebric and triangular polynomials of order not larger than n. The symbol $P_n^E(d)$ is used to standing for the set of all real, d-variate exponential polynomials of the form $\sum_{\lambda \in l(0,1,\ldots,n)^d} a_\lambda e^{-\lambda \cdot x}$ for some $l > 0$. For any given activation function $\sigma : R \to R$, $R_n^\sigma(d)$ denotes the set of all sums of the form $\sum_{\lambda \in l(0,1,\ldots,n)^d} a_\lambda \sigma(-\lambda \cdot x + b_\lambda)$ ($l > 0$). The infinity norm of R^d is denoted by $\|\cdot\|_\infty$, which is defined by

$$\|f\|_\infty = \sup_{x \in R^d} |f(x)|;$$

For any given continuous function f and a bounded real or complex function set S, the distance from f to S is defined as

$$d_\infty(f, S) = \sup_{g \in S} \|f - g\|_\infty,$$

where g is a bounded function. Given a smooth function f in $\mathcal{R}^d$, the $|\mathbf{m}|$th order partial derivatives of f are expressed as

$$D^{|\mathbf{m}|} f(\mathbf{x}) := \frac{\partial^{|\mathbf{m}|} f}{\partial \mathbf{x}^{|\mathbf{m}|}}(\mathbf{x}) = \frac{\partial^{|\mathbf{m}|} f}{\partial x_1^{m_1} \partial x_2^{m_1} \cdots \partial x_s^{m_1}}(\mathbf{x}).$$

The modulus of smoothness of a function is a measure of continuity and smoothness of the function, which has played a very important role in distinguishing different classes of functions and quantitative studies in approximation theory. Suppose that Q is a metric space with metric d, and $C(Q)$ is the set of continuous functions on Q.

Definition 1. If $f \in C(Q)$, the modulus of smoothness of f, denoted by $\omega(f, t)$, is defined by

$$\omega(f, t) = \sup_{\|x_1 - x_2\| \leq t} |f(x_1) - f(x_2)|.$$

Given a direction $\mathbf{e} \in R^d$, the rth order symmetric difference of f along the direction $\mathbf{e}$ and with the step-length h defined by

$$\Delta_h^r f(\mathbf{x}) = \sum_{i=0}^{r} (-1)^{r-i} \binom{r}{i} f\left(x + \left(\frac{r}{2} - i\right) h\mathbf{e}\right).$$

Moreover, the rth order modulus of smoothness of f is defined by

$$\omega_r(f, t) = \sup_{\mathbf{x} \pm \frac{h\mathbf{e}}{2} \in Q, \|h\| \leq t} \left|\Delta_h^r f(\mathbf{x})\right|.$$

A function f is said to belong to the α-Lipschitz class with order up to 2, denoted by $f \in \mathrm{Lip}(\alpha)_2$, if the second order modulus of smoothness $\omega_2(f, t) = O(t^\alpha)$, where $\alpha \in (0, 2]$ is a real parameter.

Definition 2 (*Ritter, 1999*). A function $\sigma : R \longrightarrow R$ is said to be nearly exponential whenever, for all $\varepsilon > 0$, there exist real numbers $\beta, \gamma, \rho, \tau$ such that

$$|\gamma \sigma(\beta t + \tau) + \rho - e^t| < \varepsilon$$

for all $t \leq 0$.

This means that after suitable rescaling and shifting at the origin and in the target spaces R, the function σ can approximate the exponential function arbitrarily well in the half line of negative reals. It is easy to see that the normal exponential function e^t is nearly exponential (get this through setting $\beta = 1$, $\rho = 0$ and

$\gamma = 1$). The sigmoid function $\sigma(t) = \frac{1}{1+e^{-t}}$ can also be shown to be nearly exponential, provided we observe that through putting $\beta = 1$, $\rho = 0$ and $\gamma = \frac{1}{\sigma(\tau)}$, we can get

$$\left|\frac{\sigma(t+\tau)}{\sigma(\tau)} - e^t\right| = \frac{e^{t+\tau}}{e^{t+\tau}+1}|1-e^t| \le e^{\tau}|e^t - e^{2t}|,$$

which converges to 0 uniformly for $t \le 0$ and $\tau \longrightarrow -\infty$. Furthermore, we can show that most sigmoid-type functions are nearly exponential functions.

In Ritter (1999), using the modulus of smoothness, Ritter obtained the following results of the upper bound estimation for the nearly exponential neural networks.

Proposition 1. *For any $f \in C_{[0,1]^d}$ and $n \in N$, there exists a nearly exponential type of FNN, $R_n^\sigma(d)$, with the form (1.1), whose number of hidden neurons is $m(n) \ge \min_{C_d(f,n)<\varepsilon}(n+1)^d$ (here $C_d(f,n) = \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega\left(f, \frac{1}{n+2}\right)$, n is any integer not less than the reciprocal of the preset approximation precision) such that*

$$d_\infty(f, R_n^\sigma(d)) \le \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega\left(f, \frac{1}{n+2}\right). \tag{2.0}$$

By using higher order modulus of smoothness of a function, we will generalize and sharpen the above result in several different ways. First, we will take the second-order modulus of smoothness instead of the modulus of first order to deduce a more accurate upper bound estimation on approximation of the neFNNs. Second, we develop a lower bound estimation of approximation accuracy of the neFNNs, and then obtain the essential order of the neFNNs. The main results obtained in this paper can be summarized as the following theorem.

Theorem 1. *Suppose V is a compact set of R^d and f is any given continuous function defined on V. Then, there exists a nearly exponential type of FNN with the hidden neuron number $m(n) \ge \min_{B_d(f,n)<\varepsilon}(n+1)^d$ (here $B_d(f,n) = \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2}+1\right)^2\omega_2\left(f, \frac{1}{n+2}\right)$, n is any integer not less than the reciprocal of the preset approximation precision ε) such that*

(i) *the following upper bound estimation holds:*

$$d_\infty(f, R_n^\sigma(d)) \le \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2}+1\right)^2\omega_2\left(f, \frac{1}{n+2}\right); \tag{2.1}$$

(ii) *the following lower bound estimation holds:*

$$\omega_2\left(f, \frac{1}{n+2}\right) \le \frac{C}{n^2}\left\{\sum_{k=1}^{n} k \cdot d_\infty(f, R_k^\sigma(d)) + \|f\|_\infty\right\}; \tag{2.2}$$

(iii) *the following essential order estimation holds:*

$$d_\infty(f, R_n^\sigma(d)) = \bigcirc(n^{-\alpha}) \quad \text{iff } f \in \mathrm{Lip}(\alpha)_2; \tag{2.3}$$

(iv) *if the relation $d_\infty(f, R_n^\sigma(d)) \le (1+1/n)^2 d_\infty(f, R_{n+1}^\sigma(d))$ is correct, we have*

$$\omega_2\left(f, \frac{1}{n+2}\right) \le C\{d_\infty(f, R_n^\sigma(d)) + n^{-2}\|f\|_\infty\}, \tag{2.4}$$

and

$$C\omega_2\left(f, \frac{1}{n+2}\right) - C\frac{1}{n^2}\|f\|_\infty \le d_\infty(f, R_n^\sigma(d)) \le \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega_2\left(f, \frac{1}{n+2}\right); \tag{2.5}$$

(v) *if $d_\infty(f, R_n^\sigma(d))$ is monotonically decreasing, then for $0 \le \delta \le 1$, we have*

$$\omega_2\left(f, \frac{1}{n+2}\right) \le C\left\{\frac{1}{n^{2(1-\delta)}}d_\infty(f, R_1^\sigma(d)) + d_\infty(f, R_{[n^\delta]}^\sigma(d)) + n^{-2}\|f\|_\infty\right\}. \tag{2.6}$$

Here and hereafter, C, C_1, C_2, C_3 are positive constants independent of n, f and x(its value, however, may be different in different contexts). *And $[n^\delta]$ is a maximum integer not less than the number n^δ.*

Remark 1. The assertion (i) in the above theorem offers an upper bound estimation on approximation order of the neFNN, which obviously sharpens the upper bound in Ritter (1999) (i.e. (2.0) of Proposition 1. For example, let $g(x) = x^m$ ($m \in N$) is a polynomial, then $\omega(g,t) = \bigcirc(t)$, and $\omega_2(g,t) = \bigcirc(t^2)$. In general, we have the relationship: $\omega(g,t) \le C\omega_2(g, t^{1/2})$ and $\omega_2(g,t) \le 2\omega(g,t)$). And Eq. (2.1) deciphers explicitly the relationship between the upper bound of approximation speed of the neFNN and the number of hidden neurons m. In particular, it shows that the approximation speed of the neFNN is proportional to the number of hidden neurons and controlled by the second-order modulus of smoothness of f, and it reveals how many hidden neurons are needed in order for the neFNN to achieve a preset approximation precision. Obviously, $d_\infty(f, R_n^\sigma(d)) \to 0$ as $n \to \infty$. It therefore also shows that any continuous function f on V can be approximated arbitrarily well by the neFNNs.

The assertion (ii) of the Theorem 1 provides us a lower bound estimation on approximation accuracy of the neFNN, which implies that the average of the neFNN over parameters n or, equivalently, over different number of neurons, is lower controlled by the second order modulus of smoothness of f and $\frac{1}{n^2}$. Actually, we can get this through rewriting (2.2) as

$$\omega_2\left(f, \frac{1}{n+2}\right) - \frac{C}{n^2}\|f\|_\infty \le C\left(\frac{1}{2}+\frac{1}{n}\right)\left\{\frac{2}{n(n+1)}\sum_{k=1}^{n} k \cdot d_\infty(f, R_k^\sigma(d))\right\}.$$

It is noted that the following identity always holds

$$\lim_{n\to\infty} d_\infty(f, R_n^\sigma(d)) = \lim_{n\to\infty}\frac{1}{n^2}\sum_{k=1}^{n} k \cdot d_\infty(f, R_k^\sigma(d)).$$

So, whenever n (or, equivalently, the number of hidden neurons) is sufficiently large, $d_\infty(f, R_n^\sigma(d))$ and $\frac{1}{n^2}\sum_{k=1}^{n} k \cdot d_\infty(f, R_k^\sigma(d))$ can be viewed approximately as the same. In such cases, statements (2.1) and (2.2) then imply

$$C\omega_2\left(f, \frac{1}{n+2}\right) - C\frac{1}{n^2}\|f\|_\infty \le d_\infty(f, R_n^\sigma(d)) \le \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega_2\left(f, \frac{1}{n+2}\right).$$

This shows that the neFNN can achieve the highest approximation accuracy and the accuracy is controlled by $\omega_2\left(f, \frac{1}{n+2}\right)$.

The assertion (iii) of the Theorem 1 gives an essential order estimation of the neFNN. It shows that whenever f belongs to the α-Lipschitz class, the essential order of approximation of the neFNN is $\bigcirc(n^{-\alpha})$. It shows also that the higher the smoothness of the function to be approximated, the faster the neFNN can approximate and vice versa.

The assertion (iv) of the Theorem 1 offers us a lower bound estimation on approximation accuracy of the neFNN under the assumption of $d_\infty(f, R_n^\sigma(d)) \le (1+1/n)^2 d_\infty(f, R_{n+1}^\sigma(d))$, which implies that the neFNN over parameters n or, equivalently, over

different number of neurons, is lower controlled by the second-order modulus of smoothness of f and $\frac{1}{n^2}$. Actually, we can get this through rewriting (2.4) as

$$\omega_2\left(f, \frac{1}{n+2}\right) - \frac{1}{n^2}\|f\|_\infty \le Cd_\infty(f, R_n^\sigma(d)).$$

According to (2.1), we have

$$C\omega_2\left(f, \frac{1}{n+2}\right) - C\frac{1}{n^2}\|f\|_\infty \le d_\infty(f, R_n^\sigma(d)) \le \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega_2\left(f, \frac{1}{n+2}\right).$$

This shows that the order of approximation by neFNNs can be characterized nearly completely through $\omega_2\left(f, \frac{1}{n+2}\right)$.

The assertion (v) of the Theorem 1 affords us a lower bound estimation on approximation accuracy of the neFNNs when $d_\infty(f, R_n^\sigma(d))$ is monotonically decreasing of n. (v) shows that the neFNNs over $[n^\delta]$ is lower controlled by the second-order modulus of smoothness of f and $\frac{1}{n^{2(1-\delta)}}$ $(0 < \delta < 1)$. According to (2.1), we have

$$C\omega_2\left(f, \frac{1}{n+2}\right) - C\frac{1}{n^{2(1-\delta)}}d_\infty(f, R_1^\sigma(d)) - Cn^{-2}\|f\|_\infty \le d_\infty(f, R_{[n^\delta]}^\sigma(d)) \le \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega_2\left(f, \frac{1}{[n^\delta]+2}\right).$$

Since $d_\infty(f, R_1^\sigma(d))$ is a constant, the above inequalities show that the lower bound and the upper bound of approximation by neFNNs $R_{[n^\delta]}^\sigma(d)$ are determined by $\omega_2\left(f, \frac{1}{n+2}\right)$ and $\omega_2\left(f, \frac{1}{[n^\delta]+2}\right)$, respectively.

Remark 2. Barron (1993) proved the approximation bounds related to the Theorem 1. He used an integral representation by means of indictor functions of half spaces in order to show that the order of uniform approximation of any function $f : [-1, 1]^d \mapsto R$ with Fourier representation $\check{f}$:

$$f(\mathbf{x}) = \int_{R^d} e^{\mathbf{x}\cdot\mathbf{y}}\check{f}(\mathbf{y})d\mathbf{y}$$

such that $\|\mathbf{y}\|\check{f}(\mathbf{y})$ is integrable is at least $O(N^{-1/2})$ (N is the number of hidden neuron). This describes a class of functions for which a dimension-independent approximation order holds (the constants involved in the estimates depend on dimension). Barron's class is described in terms of a global property of its Fourier transform. Barron's bound is controlled by the number of hidden neuron and the dimension, and our bounds depend on the second-order modulus of smoothness of f and the number of hidden neuron. Barron's results also showed that the order of mean-square approximation of a member of the said class is at least $O(n^{-1/2})$. This results are very useful in the statistical classification problems. Our paper addresses an important issue for neural networks: the essential order of approximation (helping to explain why so many neurons are required and perhaps why the speed of convergence behaves the way it does).

The proof of Theorem 1 will be presented in Section 4. Some preliminary results will be given in the next section.

3. Uniform approximation of continuous functions by polynomials

In approximation theory (Jackson, 1912), the well-known Jackson's type theorems describe approximation precision and speed of a continuous function defined on a real interval by polynomials in terms of modulus of smoothness of the function. Multivariate extensions of this type of theorems are due to the references (Feinerman & Newman, 1974; Nikol'skii, 1975; Soardi, 1984) Ditzian and Totik (1987), who introduced a new modulus of smoothness of a function, nowadays known as Ditzian–Totik modulus. Feinerman and Newman (1974) and Soardi (1984) have given the following important proposition, which underlies the research of Ritter (1999).

Proposition 2. *For any continuous function $f : [0, 1]^d \to R$ and all $n \in N$, the following estimation holds:*

$$d_\infty(f, P_n(d)) \le \left(\frac{1}{2} + \frac{\pi^2}{4}\sqrt{d}\right)\omega\left(f, \frac{1}{n+2}\right). \tag{3.1}$$

Using the second-order modulus of smoothness, we sharpened the above result into the following theorem (Xu & Wang, 2006).

Theorem 2. *For any continuous function $f : [a, b]^d \to R$ and all $n \in N$, there holds*

$$d_\infty(f, P_n(d)) \le \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2} + 1\right)^2\omega_2\left(f, \frac{1}{n+2}\right). \tag{3.2}$$

To prove our main theorem, we also need the following result (Xu & Wang, 2006).

Theorem 3. *For any continuous function $f : [a, b]^d \to R$ and all $n \in N$, we have*

$$d_\infty(f, P_n^E(d)) \le \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2} + 1\right)^2\omega_2\left(f, \frac{1}{n+2}\right). \tag{3.3}$$

4. Proof of Theorem 1

Firstly, we prove the estimation (2.1). Let τ denote the Euclidean projection $[0, 1]^d \longrightarrow V \subseteq R^d$, and f be a continuous function defined on the compact set V. Then $f(\tau)$ is a continuous extension of f on V with the same modulus of smoothness as that of f. Using Theorem 3, we know that there is an exponential polynomial

$$p(x) = \sum_{\lambda\in l(0,1,2,\ldots,n)^d} a_\lambda e^{-\lambda\cdot x}$$

such that

$$\|f - p|_C\|_\infty \le \|f(\tau) - p|_{[0,1]^d}\|_\infty \le \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2} + 1\right)^2\omega_2\left(f, \frac{1}{n+2}\right) + \varepsilon.$$

Since σ is nearly exponential, we know that the exponential function e^x is uniformly approximated by an expression of the form $\gamma\sigma(\beta x + \tau) + \rho$ on the half line, with the error $\varepsilon/\sum|a_\lambda|$. So the sum

$$\sum_{\lambda\in l(0,1,2,\ldots,n)^d\setminus\{0\}} \gamma a_\lambda\sigma(\beta\lambda\cdot x + \tau) + \left(\gamma a_0\sigma(\tau) + \rho\sum_{\lambda\in l(0,1,2,\ldots,n)^d} a_\lambda\right)$$

belongs to $R_n^\sigma(d)$ and its distance to f does not exceed the value $\frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2} + 1\right)^2\omega_2\left(f, \frac{1}{n+2}\right) + 2\varepsilon$. Since ε is arbitrary, we obtain (2.1).

In order to prove (2.2), we establish the following lemma first.

Lemma 1. *For the nonnegative sequences* $\{a_n\}$, $\{b_n\}$, *if the inequality* $(p > 0)$

$$a_n \le \left(\frac{k}{n}\right)^p a_k + b_k \quad (1 \le k \le n) \tag{4.1}$$

holds for $n \in \boldsymbol{N}$, *then one has*

$$a_n \le C_p n^{-p} \left\{ \sum_{k=1}^{n} k^{p-1} b_k + a_1 \right\}. \tag{4.2}$$

Proof. For $n \ge 2$, we can choose $N \in \boldsymbol{N}$ such that $2^N \le n < 2^{N+1}$. Then, there is $m_k \in \boldsymbol{N}$, $1 \le k \le n$, such that $\frac{n}{2^{k+1}} \le m_k \le \frac{n}{2^k}$ and $a_{m_k} \le a_j$ $\left(\frac{n}{2^{k+1}} \le j \le \frac{n}{2^k}\right)$. Taking $m_{N+1} = 1$, we then have by (4.1) that

$$\begin{aligned}
a_n &\le \left(\frac{m_0}{n}\right)^p a_{m_0} + b_{m_0} \\
&= n^{-p} \sum_{k=0}^{N} m_k^p \left(a_{m_k} - \left(\frac{m_{k+1}}{m_k}\right)^p a_{m_{k+1}} \right) + b_{m_0} + n^{-p} a_1 \\
&\le \sum_{k=0}^{N} 2^{-pk} b_{m_{k+1}} + b_{m_0} + n^{-p} a_1 \\
&= \sum_{k=1}^{N+1} 2^{-p(k-1)} b_{m_k} + b_{m_0} + n^{-p} a_1 \\
&\le 2^p \sum_{k=0}^{N+1} 2^{-pk} b_{m_k} + n^{-p} a_1 \\
&\le 2^p \sum_{k=0}^{N+1} 2^{-pk} \frac{2^{k+1}}{n} \sum_{\frac{n}{2^{k+1}} \le j \le \frac{n}{2^k}} b_j + n^{-p} a_1 \\
&\le 2^{p+1} \sum_{k=0}^{N+1} n^{-p} \sum_{\frac{n}{2^{k+1}} \le j \le \frac{n}{2^k}} j^{p-1} b_j + n^{-p} a_1 \\
&\le C_p n^{-p} \left\{ \sum_{k=1}^{n} k^{p-1} b_k + a_1 \right\},
\end{aligned}$$

which verifies the equation.

Now we prove the estimation (2.2). For $n \ge 1$, $n \in N$, let

$$u_n(x) = \frac{\left(1 - \cos^2 \frac{\pi}{n+2}\right)\left(t_{n+2}(\cos x) - t_{n+2}(x)\right)}{(n+2)\left(\cos x - \cos \frac{\pi}{n+2}\right)^2}$$

be the Korovkin's kernel, which is seen as a triangular trigonometric polynomial of degree not greater than n. In the kernel, $t_n(x)$ is the Chebyshev polynomial, namely, $t_n(x) = \arccos(nx)$, $u_n(x) \in T_n(1)$, $u_n \ge 0$ and there holds $\frac{1}{2\pi}\int_{-\pi}^{\pi} u_n(x)\mathrm{d}x = 1$. As we know, $u_n(x)$ can also be represented as $u_n(x) = \frac{2}{n+2}\left\{\sum_{-n/2}^{n/2} \cos\frac{k\pi}{n+2} \cos kx\right\}^2$. Hence, if we define the d-fold tensor product $v_n(x_1, x_2, \ldots, x_d) = \overbrace{u_n(x) \times u_n(x) \times \cdots \times u_n(x)}^{d} \in T_n(d)$, then we find $v_n \ge 0$ and

$$(2\pi)^{-d} \int_{[-\pi,\pi]^d} v_n(x)\mathrm{d}x = 1.$$

Recall that the convolution $L(v_n, f)$ of two continuous 2π-periodic functions v_n and f is defined by

$$L(v_n, f) = (2\pi)^{-d} \int_{[-\pi,\pi]^d} f(x-t) v_n(t)\mathrm{d}t, \quad (x \in R^d).$$

Obviously,

$$D^{\nu} L(v_n, f) = (2\pi)^{-d} \int_{[-\pi,\pi]^d} D^{\nu} f(x-t) v_n(t)\mathrm{d}t.$$

For all $1 \le q \le \infty$, since $(2\pi)^{-d}\int_{[-\pi,\pi]^d} v_n(x)\mathrm{d}x = 1$ and $\|L(v_n, f)\|_q \le \|f\|_q \|v_n\|_1$, we have

$$\|D^{|\nu|} L(v_n, f)\|_q \le \|D^{\nu} f\|_q. \tag{4.3}$$

By using the Bernstein inequality, we then obtain

$$\|D^{|\nu|} L(v_n, f)\|_q \le C n^{|\nu|} \|D^{\nu} f\|_q. \tag{4.4}$$

Let $a_n = \frac{1}{n^2}\|D^{|\nu|} L(v_n, f)\|_q$, $|\nu| = 2$, and $b_n = \|L(v_n, f) - f\|_q$. From Eqs. (4.3) and (4.4), we obtain

$$\begin{aligned}
a_n &\le \frac{1}{n^2}\|D^{|\nu|} L(v_n, L(v_k, f))\|_q + \frac{1}{n^2}\|D^{|\nu|} L(v_n, f - L(v_k, f))\|_q \\
&\le \frac{1}{n^2}\|D^{|\nu|} L(v_k, f)\|_q + C\|f - L(v_k, f)\|_q \\
&= \left(\frac{k}{n}\right)^2 a_k + C b_k.
\end{aligned} \tag{4.5}$$

This establishes that the condition of (4.1) is met, so we can now apply the results of Lemma 1. Substituting $p = \nu = 2$ in to (4.1):

$$a_n \le C_2 n^{-2} \left\{ \sum_{k=1}^{n} k b_k + a_1 \right\}.$$

Therefore, substituting for b_k and a_1:

$$\sup_{|\nu|=2} \|D^{|\nu|} L(v_n, f)\|_q \le C_2 \left\{ \sum_{k=1}^{n} k \|L(v_k, f) - f\|_q + \|f\|_q \right\}. \tag{4.6}$$

On the other hand, for $n \ge 2$, there exists $m \in N$ such that $\frac{n}{2} \le m \le n$ and

$$\|f - L(v_m, f)\|_q \le \|f - L(v_k, f)\|_q, \quad \frac{n}{2} \le k \le n. \tag{4.7}$$

Based on Eqs. (4.5)–(4.7), we now can define a K-function as follows:

$$K_2(f, t^2) = \inf_{D^{|m|} g \in \text{A.C.loc}} \left\{ \|f - g\| + t^2 \sup_{|m|=2} \|D^{|m|} g\| \right\},$$

where $g \in$ A.C.loc means that g is $|m|$ times differentiable and $D^{|m|} g$ is absolutely continuous in the finite set.

From Johnen and Scherer (1977), there exists a positive constant C' such that

$$C'^{-1} K_2(f, t^2) \le \omega_2(f, t) \le C' K_2(f, t^2). \tag{4.8}$$

Hence,

$$\begin{aligned}
&K_2\left(f, \frac{1}{(n+2)^2}\right) \\
&\le \|f - L(v_m, f)\|_q + \frac{1}{(n+2)^2} \sup_{|\nu|=2} \|D^{|\nu|} L(v_n, f)\|_q \\
&\le \frac{4}{n^2} \sum_{\frac{n}{2} \le k \le n} k \|f - L(v_k, f)\|_q \\
&\quad + \frac{C_2}{(n+2)^2} \left\{ \sum_{k=1}^{n} k \|L(v_k, f) - f\|_q + \|f\|_q \right\} \\
&\le C_3 \frac{1}{n^2} \left\{ \sum_{k=1}^{n} k \|L(v_k, f) - f\|_q + \|f\|_q \right\} \\
&\le \frac{C_3}{n^2} \left\{ \sum_{k=1}^{n} k \cdot d_{\infty}(f, P_k(d)) + \|f\|_{\infty} \right\}.
\end{aligned}$$

Using Eq. (4.8) and a technique used in Theorem 3, we deduce that

$$\omega_2\left(f, \frac{1}{n+2}\right) \le \frac{C}{n^2}\left\{\sum_{k=1}^{n} k\cdot(d_\infty(f, R_k^\sigma(d)) + \varepsilon) + \|f\|_\infty\right\} \le \frac{C}{n^2}\left\{\sum_{k=1}^{n} k\cdot d_\infty(f, R_k^\sigma(d)) + \|f\|_\infty\right\} + C\varepsilon.$$

Since ε is arbitrary, it gives (2.2). □

The estimation (2.3) can be verified directly by combining the conclusions (2.1) with (2.2).

Now we prove the estimation (2.4). Firstly, we prove the following proposition.

Proposition 3. *Assume that for the nonnegative sequences* $\{a_n\}$, $\{b_n\}$, $\{b_k\}_{k=1}^n$ *satisfied* $b_k \le \left(1+\frac{1}{k}\right)^p b_{k+1}$, *and the inequality*

$$a_n \le Cn^{-2}\left\{\sum_{k=1}^{n} kb_k + M\right\} \tag{4.9}$$

holds for $n \in N$, *then one has*

$$a_n \le C(b_n + n^{-2}M). \tag{4.10}$$

Here $C \ge 1$ *is a constant and* M *is a constant independent of* n, k.

Proof. In fact, we only need to prove the general inequality

$$\sum_{k=1}^{n} k^{p-1}b_k \le Cn^p b_n \tag{4.11}$$

holds for $p \in N$. We prove (4.11) by induction. When $n = 1$, we have

$$b_1 \le Cb_1.$$

Assume (4.11) is correct for $n \in N$, that is,

$$\sum_{k=1}^{n} k^{p-1}b_k \le Cn^p b_n.$$

We now show that (4.11) also holds for $n+1$. It is clear that

$$\begin{aligned}\sum_{k=1}^{n+1} k^{p-1}b_k &= \sum_{k=1}^{n} k^{p-1}b_k + (n+1)^{p-1}b_{n+1}\\ &\le Cn^p b_n + (n+1)^{p-1}b_{n+1}\\ &\le C(n+1)^p b_{n+1}\left(\left(\frac{n}{n+1}\right)^p \frac{b_n}{b_{n+1}} + \frac{1}{C(n+1)}\right)\\ &\le C(n+1)^p b_{n+1}.\end{aligned}$$

With this, inequality (4.11) is completed. Let $p = 2$, using Eqs. (4.9) and (4.11), we have

$$\begin{aligned}a_n &\le C_2 n^{-2}\left\{\sum_{k=1}^{n} k^1 b_k + M\right\}\\ &\le C\cdot C_2 b_n + C\cdot C_2 n^{-2}M\\ &\le C(b_n + n^{-2}M).\end{aligned}$$

Eq. (4.10) is completed.

Now for Eq. (2.2), let $a_n = \omega_2\left(f, \frac{1}{n+2}\right)$, $b_k = d_\infty(f, R_k^\sigma(d))$, and $M = \|f\|_\infty$. Applying Proposition 3, we have

$$a_n \le C(b_n + n^{-2}M).$$

Therefore

$$\omega_2\left(f, \frac{1}{n+2}\right) \le C\{d_\infty(f, R_n^\sigma(d)) + n^{-2}\|f\|_\infty\},$$

which gives (2.4). Combining Eqs. (2.1) and (2.4), the estimation (2.5) is established.

Now we prove (2.6). From (2.2) and the fact that $d_\infty(f, R_n^\sigma(d))$ is monotonically decreasing, we have

$$\begin{aligned}\omega_2\left(f, \frac{1}{n+2}\right) &\le \frac{C}{n^2}\left\{\sum_{k=1}^{n} k\cdot d_\infty(f, R_k^\sigma(d)) + \|f\|_\infty\right\}\\ &= \frac{C}{n^2}\left\{\sum_{k=1}^{[n^\delta]-1} k\cdot d_\infty(f, R_k^\sigma(d)) + \sum_{k=[n^\delta]}^{n} k\cdot d_\infty(f, R_k^\sigma(d)) + \|f\|_\infty\right\}\\ &\le \frac{C}{n^2}\left\{d_\infty(f, R_1^\sigma(d))\sum_{k=1}^{[n^\delta]-1} k + d_\infty(f, R_{[n^\delta]}^\sigma(d))\sum_{k=[n^\delta]}^{n} k + \|f\|_\infty\right\}\\ &\le C\left\{\frac{1}{n^{2(1-\delta)}} d_\infty(f, R_1^\sigma(d)) + d_\infty(f, R_{[n^\delta]}^\sigma(d)) + n^{-2}\|f\|_\infty\right\}.\end{aligned}$$

In order to confirm the number of hidden neurons used, we take n to be the smallest integer larger than the reciprocal of ε (the preset approximation precision) in (2.1). Though computation, we obtain $m(n) = \min_{B_d(f,n)<\varepsilon}(n+1)^d$, where $B_d(f, n) = \frac{1}{2}\left(\frac{\sqrt{d}\pi^2}{2}+1\right)^2 \omega_2\left(f, \frac{1}{n+2}\right)$. With this, the proof of Theorem 1 is completed. □

5. Conclusions

In this work, the essential approximation order of the nearly exponential-type neural networks has been studied. In terms of second-order modulus of smoothness of a function, an upper bound and lower bound estimations on approximation precision and speed of the neural networks are simultaneously developed. Under certain assumption on the neFNNs, we present ideally the upper bound and the lower bound on the degree of approximation. Our research reveals that the approximation precision and speed of the neural networks depend not only on the number of hidden neurons used, but also on the smoothness of the functions to be approximated. We have explicitly given a lower bound estimation on the number of hidden neurons of the network in order to attain a predetermined approximation precision. The results obtained are helpful in understanding the approximation capability and topology construction of the neural networks.

References

Attali, J. G., & Pages, G. (1997). Approximation of functions by a multilayer perceptron: A new approach. *Neural Networks*, *10*, 1069–1081.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*, 930–945.

Cao, F. L., & Xu, Z. B. (2001). Neural network approximation for multivariate periodic functions: Estimates on approximation order. *Chinese Journal of Computers*, *24*(9), 903–908.

Cardaliaguet, P., & Euvrard, G. (1992). Approximation of a function and its derivatives with a neural networks. *Neural Networks*, *5*, 207–220.

Chen, T. P. (1994). Approximation problems in system identification with neural networks. *Science in China, Series A*, *24*(1), 1–7.

Chen, T. P., & Chen, H. (1995). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to a dynamic system. *IEEE Transactions on Neural Networks*, *6*, 911–917.

Chui, C. K., & Li, X. (1992). Approximation by ridge functions and neural networks with one hidden layer. *Journal of Approximation Theory*, *70*, 131–141.

Chui, C. K., & Li, X. (1993). Neural networks with one hidden layer. In K. Jetter, & F. I. Utreras (Eds.), *Multivariate approximation: From CAGD to wavelets* (pp. 7–89). Singapore: World Scientific Press.

Cybenko, G. (1989). Approximation by superpositions of sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314.
Ditzian, Z., & Totik, V. (1987). *Moduli of smoothness*. NewYork: Springer-Verlag.
Feinerman, R. P., & Newman, D. J. (1974). *Polynomial approximation*. Baltimore, MD: Williams & Wilkins.
Funahashi, K. I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.
Hornik, K., Stinchombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximation. *Neural Networks*, *2*, 359–366.
Hornik, K., Stinchombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, *3*, 551–560.
Jackson, D. (1912). On the approximation by trigonometric sums and polynomials. *Transactions of the American Mathematical Society*, *13*, 491–515.
Johnen, H., & Scherer, K. (1977). On the equivalence of the *K*-functional and modulus of continuity and some applications. In W. Schempp, & K. Zeller (Eds.), *Constructive theory of functions of several variable* (pp. 119–140). Berlin: Springer-Verlag.
Kůrkova, V., Kainen, P. C., & Kreinovich, V. (1997). Estimates for the number of hidden units and variation with respect to half-space. *Neural Networks*, *10*, 1068–1078.
Leshno, M., Lin, V. Y., Pinks, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*, 861–867.
Maiorov, V., & Meir, R. S. (1998). Approximation bounds for smooth functions in $C(\mathcal{R}^d)$ by neural and mixture networks. *IEEE Transactions on Neural Networks*, *9*, 969–978.
Mhaskar, H. N. (1996). Neural networks for optimal approximation for smooth and analytic functions. *Neural Computation*, *8*, 164–177.
Mhaskar, H. N., & Khachikyan, L. (1995). Neural networks for functions approximation. In J. Makhoul, E. Manolakos, & E. Wilson (Eds.), *Neural networks for signal processing, Proc. 1995 IEEE workshop, vol. V, F. Girosi (Cambridge, MA)* (pp. 21–29). New York: IEEE Press.
Mhaskar, H. N., & Micchelli, C. A. (1992). Approximation by superposition of a sigmoidal function. *Advances in Applied Mathematics*, *13*, 350–373.
Mhaskar, H. N., & Micchelli, C. A. (1994). Dimension-independent bounds on the degree of approximation by neural networks. *IBM Journal of Research and Development*, *38*, 277–284.
Mhaskar, H. N., & Micchelli, C. A. (1995). Degree of approximation by neural networks with a single hidden layer. *Advances in Applied Mathematics*, *16*, 151–183.
Nikol'skii, S. M. (1975). *Approximation of functions of several variables and imbedding theorems*. Berlin, Heidelberg, New York: Springer.
.Ritter, G. (1994). Jackson's theorems and the number of hidden units in Neural networks for uniform approximation. *Technical Report, MIP9415*. Univ. Passau, Fak. Math. Inform.
Ritter, G. (1999). Efficient estimation of neural weights by polynomial approximation. *IEEE Transactions on Information Theory*, *45*, 1541–1550.
Soardi, P. M. (1984). *Quad.dell'Unione Mat.Italiana*: *Vol. 26*. *Serie di fouroer in piú variabili*. Bologna, Italy: Pitagora Editrice.
Suzuki, S. (1998). Constructive function approximation by three-layer artificial neural networks. *Neural Networks*, *11*, 1049–1058.
Xu, Z. B., & Cao, F. L. (2004). The essential order of approximation for neural networks. *Science in China, Series F*, *47*, 97–112.
Xu, Z. B., & Wang, J. J. (2006). The essential order of approximation for neural networks. *Science in China, Series F. Information Sciences*, *49*(4), 446–460.
Yoshifusa, I. (1991). Approximation of functions on a compact set by finite sums of sigmoid function without scaling. *Neural Networks*, *4*, 817–826.