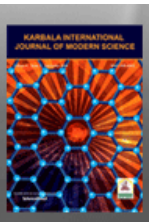

Karbala International Journal of Modern Science

Volume 6 | Issue 4 Article 9

L_p_ Approximation by ReLU Neural Networks

Eman Samir Bhaya
University of Babylon, emanbhaya@itnet.uobabylon.edu.iq

Zainab Abdulmunim Sharba
University of Babylon, zainab.abd@uobabylon.edu.iq

L_p_ Approximation by ReLU Neural Networks

Abstract

We know that we can use the neural networks for the approximation of functions for many types of activation functions. Here, we treat only neural networks with simple and particular activation function called rectified linear units (ReLU). The main aim of this paper is to introduce a type of constructive universal approximation theorem and estimate the error of the universal approximation. We will obtain optimal approximation if we have a basis independent of the target function. We prove a type of Debao Chen's theorem for approximation.

Keywords

Activation function; B spline; Modulus of smoothness; Neural networks

Cover Page Footnote

The authors would like to thank Dr. Dunya Muhammed Miqdad for assisting in the linguistic review.

1. Introduction

There are many studies about the approximation by neural networks with different types of activation functions, for example, in Ref. [1] the authors investigate the error arising from the method of approximation operators with logarithmic sigmoidal function. By the constructive method, the authors in Ref. [2] introduced a direct theorem for simultaneous pointwise approximation using neural networks with one hidden layer. In Ref. [3], a new sigmoidal function is introduced with parameter and considering constructive feedforward approximation on a closed interval. The Proof for the universal approximation ability of recurrent neural networks in the state space model form is introduced in Ref. [4]. Also, versions of direct and inverse inequalities using neural networks are proved in Ref. [5,6]. The questions about a quadratic network approximation are asked in Ref. [7] with demonstrating the merits of a quadratic network.

In the last decades, the ability of approximation by single hidden layer feedforward neural networks (SLFNNs) was studied in numerous works.

Since 1960, many papers proved that for any continuous function, there exists a multilayer neural network with one hidden layer as an approximation of continuous real function on $\mathbb{R}^n$. As the proofs given by Cybenko [8], White [9], and Hornik [10]. These articles proved that the above results with few conditions on activation function are not constructive and are not elementary. They did not describe the number of new neurons that should be used in the hidden layer and they did not estimate the degree of the best approximation depending on the number of neurons. After many years, many improvements of universal approximation were introduced and their applications to lay many types of research (see for example ([11−16]), with a restriction on the set of weights of the neural network. In these articles, the authors proved that any restriction on the weight of SLFNNs does not affect the universal approximation property. In Ref. [17], Stinchcombe and White proved that the SLFNNs with a polygonal, spline polynomial, or analytic activation function and a bounded set of weights still have the property of universal approximation. In Ref. [18], Ito introduced the results of universal approximation using SLFNNs with monotone sigmoidal functions, which converge to 0 at $-\infty$ and converge to 1 at ∞; he used only weights on the unit sphere. In Ref. [12−14], the authors used SLFNNs with various weights on a set of directions and gave many conditions to get a good approximation by such neural networks.

Such results hold for a wide variety of activations, among these activation functions, a type called rectified linear unit (ReLU) activation function [19]. In Ref. [20], the authors present an approximation of ReLU by relating wavelet. Some local approximation constructions are proposed to represent general functions including piecewise linear trapezoid [21,22], and piecewise linear spike-shaped unit [23]. There are some other constructions that first approximate polynomials [24−27] and then used them as media to approximate more general functions. It has also been shown that compared with these deep ReLU network constructions, shallow networks have to be exponentially wider to achieve identical approximation accuracy. This comes from the fact that for sufficiently smooth functions, there exist lower bounds of approximation errors that are determined by the number approximating linear pieces, which in turn are dominated by the depth. Recently [28], establish L^{∞}, L^{2} direct theorems of multivariate functions by ReLU combination [29]. used the approximation by deep convolution neural networks for functions in the Sobolev space Hr (Ω) with $r > 2 + d/2$ [30]. studied the approximation by ReLU neural networks depending on depth and weights of functions in L^2 spaces.

The rectified linear unit is an interesting choice for the activation functions of neural networks. For their efficiency and simplicity practically and/or theoretically, we consider this important type of activation functions which is given by $\mathrm{ReLU}(x) = \max(x, 0)$ and we prove our results on $L_p(I)$, space for $0 < p \leq \infty$. The main aim of this paper is to introduce an approximation theoretic structure for single hidden layer neural networks.

Now, let us introduce some preliminaries that we need in this work.

2. Preliminaries

Let $L_p(\mathrm{I})$, $0 < p \leq \infty$ denotes the space of all measurable functions f on I, where $\mathrm{I} = [a, b]$ *or* $\mathbb{R}$ such that:

https://doi.org/10.33640/2405-609X.2362

$$\|f\|_{L_p(I)} := \begin{cases} \left(\int_I |f(x)|^p dx \right)^{1/p}, & 0<p<\infty, \\ ess \sup_{x\in I} |f(x)|, & p=\infty. \end{cases}$$

Is finite, we also denote $\|f\|_p = \|f\|_{L_p(I)}$, and let $L_p^k(I)$ be the space of functions, which are $k-$ fold integrable of $L_p(I)$ *or* $L_p(\mathbb{R})$ functions where $\mathbb{R}$ is the set of real numbers. Also, we need to use the modulus of smoothness given by:

$$\omega_{r,m}^{\varphi}(f^{(m)},t)_p := \sup_{0\le h\le t} \left\| W_{rh}^m(\cdot)\Delta_{h\varphi(\cdot)}^r(f^{(m)},.) \right\|_p, \text{ [31]}$$

where $f\in\mathbb{B}_p^m$,

$\mathbb{B}_p^m := \{f : \|f^{(m)}\varphi^{(m)}\|_p < +\infty\}$,

$W_\delta(x) := ((1-x-\delta\varphi(x)/2)(1+x-\delta\varphi(x)/2))^{1/2}$,

$\varphi(x) := \sqrt{1-x^2}$.

and

$$\Delta_h^r(g,y,I) := \begin{cases} \sum_{\ell=0}^{r} \binom{r}{\ell} (-1)^{r-\ell} g(y+(l-r/2)h) & if\ y\mp rh/2\in I, \\ 0 & otherwise. \end{cases} \text{ [31]}$$

We can construct the best approximation by choosing the parameters of the approximation $(0\le \ell\le n)$:

Let $y_\ell = \frac{\ell-1}{n}$, and $\mathscr{W}_\ell = Z(f,\ell) - \sum_{j=1}^{\ell-1} \mathscr{W}_j$, where

$$Z(f,\ell) = \sup_{y\in[y_\ell,y_{\ell+1})} \left\{ 2f(y-r/2) + \sum_{\ell=1}^{r} \binom{r}{\ell} (-1)^{r-\ell} f(y+(\ell-r/2)h) \right\}$$
$$+ \inf_{y\in[y_\ell,y_{\ell+1})} \left\{ 2f(y-r/2) + \sum_{\ell=1}^{r} \binom{r}{\ell} (-1)^{r-\ell} f(y+(\ell-r/2)h) \right\},$$

and

$$(A_n f)(y) = \sum_{\ell=1}^{n} \mathscr{W}_\ell \sigma(y-y_\ell).$$

Besides, let us use the B-splines that are piecewise functions coming from extending the following functions:

$$B_1(x) = \begin{cases} 1 & if\ \frac{-1}{2} \le x < \frac{1}{2} \\ 0 & otherwise \end{cases}$$

Similarly, for any $n\in\mathbb{N}$, we write

$B_n = B_1 * B_1 * \ldots B_1$ (n − times).

We point out that m is equal to zero, and. $\lim_{n\to\infty} a_n = 0$ then $\lim_{n\to\infty} \omega_{r,m}^{\varphi}(f^{(m)},a_n)_p = 0$ *also* $\forall \epsilon > 0$ $\omega_{r,m}^{\varphi}(I,\epsilon)_p = \epsilon$, where I is the identity function $(I(x) = x)$. Recall that the best approximation of $f\in L_p(I)$ by a polynomial $h\in\Pi_n$ given by

$E_n(g)_p = \inf_{h\in\Pi_n} g - h_p$. [32]

where Π_n denotes the set of all algebraic polynomials of degree $\le n$.

3. Main results

Here, let us introduce the following results:

Theorem 3.1*(approximation with ReLU)*

Let $\sigma(.)$ be the ReLU function, then for every $g\in L_p^k(R), \epsilon > 0$,

there exists $n\in N$, $\mathscr{W}_\ell$, $b_{\ell,}\in\mathbb{R}$, $\ell\in\{0,\ldots,n\}$, and

$$(A_n g)(y) = \sum_{\ell=1}^{n} \mathscr{W}_\ell \sigma(y + b_\ell),$$

as an approximation of $f(.)$, *which* satisfies

$$\left\|(A_n f)(y) - f(y)_p\right\| < \epsilon.$$

Proof

Since $g \in L_p^k(\mathbb{R})$, there exists $K \in \mathbb{Z}^+$ such that $|g(y)| < \Delta_{h\varphi(.)}^r g(y)$. When $|y| > K$, we divide the interval $[-K, K]$ into $2K^2$ equal points with length $1/K$.

Let $-K = y_0 < y_1 \cdots < y_{2K^2} = K$, and let $b_\ell = \frac{y_{\ell-1}+y_\ell}{2}$ $(1 \le \ell \le 2K^2)$. Since σ is a bounded activation function, so we assume that there exists a positive integer N s.t. $|\sigma(x)| < 1/2\, K$, for $|x| \ge N$, choose a positive integer M such that $\frac{M}{2Z} > N$.

Now, we construct the neural network as:

$$(A_n g)(y) = \sum_{\ell=1}^{2r^2} \Delta_{h\varphi(\cdot)}^r g(y) \sigma(M(y - b_\ell)),$$

Then, $|M(y - b_\ell)| \ge N$ and hence $|\sigma(M(y - b_\ell)| < 1/2K$, for $\ell = 1, 2, \ldots, 2K^2$

Choose b_ℓ such that $\sum_{\ell=1}^{2r^2} \sigma(M(y - b_\ell)) = c$, where c is constant. So

$$\|(A_n g)(y) - g(y)\|_p$$

$$\le \left\| \sum_{\ell=1}^{2r^2} \Delta_{h\varphi()}^r g(y) \sigma(M(y - b_\ell))_p \right\|$$

$$\le c(p,k) \omega_{r,m}^{\varphi}(g,t)_p,$$

c(p,k) is a constant that depends on *p and k* and it is different from a step to others.

since $\Delta_{h\varphi(.)}^r g$ converges to zero for $g \in L_p^k(\mathbb{R})$ so that:

$$\|(A_n g)(y) - g(y)\|_p < \in.$$

We can prove the above result by another method but we need the following Remark.

Remark 3.2: It is well known that if f is a continuous function, then f can be approximated by a linear combination of its translation and dilation.

Proof of Theorem 3.1

We can build the 1^{st} order B-spline mother function

$$\psi\left(x, 1, \frac{1}{2}\right) = \sigma\left(x + \frac{1}{2}\right) - \sigma\left(x - \frac{1}{2}\right) = B_1(x),$$

by using the previous remark, the proof is complete.

Theorem 3.3 *(estimating the error)*

for any $n \in N$, *and* $f \in L_p(I)$, we have:

$$\left\|f - A_n f\right\|_p \le \omega_r^{\varphi}(f,t)_p.$$

Proof:

First, we must prove for all $\ell \in \{1, \ldots, n\}$ and for every $x \in [x_\ell, x_{\ell+1})$,

$$(A_n f)(x) = Z(f, \ell).$$

Fix $\ell \in \{1, \ldots, n\}$ and let $x \in [x_\ell, x_{\ell+1})$, from the construction of the activation function

$$\sigma(x - x_\ell) = \begin{cases} x - x_\ell & j \le l \\ 0 & j > l. \end{cases}$$

we have:

$$(A_n f)(x) = \sum_{j=1}^{n} \mathscr{W}_j \sigma(x - x_j)$$

$$= \sum_{j=1}^{\ell} \mathscr{W}_j (x - x_j)$$

$$= \mathscr{W}_\ell (x - x_\ell) + \sum_{j=1}^{\ell-1} \mathscr{W}_j (x - x_j)$$

$$= Z(f, \ell) - \sum_{j=1}^{\ell-1} \mathscr{W}_j (x - x_j) + \sum_{j=1}^{\ell-1} \mathscr{W}_j (x - x_j)$$

$$= Z(f, \ell).$$

Consequently:

$$\|f - A_n f\|_p \le c(p) \|f - Z(f, \ell)\|_p$$

$$\le \omega_r^{\varphi}(f,t)_p.$$

Theorem 3.4 *(optimality)*

for every $n \in N$, $A_n f$ is optimal, in the sense that if the basis is independent of f, then there is $\ell \in \{1, \ldots n\} : x_\ell \ne \frac{\ell-1}{n}$ *or there is* $\ell \in \{1, \ldots, n\}$:

$$\mathscr{W}_\ell \ne Z(f, \ell) - \sum_{i=1}^{\ell-1} \mathscr{W}_i,$$

then, there exists $f \in L_p(I)$ such that:

$$\|f - A_n f\|_p > \omega_r^{\varphi}\left(f, \frac{1}{n}\right)_p.$$

Proof:

Since $\omega_1 > \omega_r$, for $r \ge 2$;

therefore, we prove our result for case $r = 1$ and hence it is true for any. r.

Case 1: suppose. $\ell \in \{1, \ldots, n\} : x_\ell \ne \frac{\ell-1}{n}$.

In that case, there exists $i \in \{0, \ldots, n\}$, such that:

$$x_{i+1} - x_i > \frac{1}{n}, \ (x_0 = 0, x_{n+1} = 1).$$

Let $\delta = x_{i+1} - x_i$. It is clear that $A_n f$ is constant on. $[x_i, x_{i+1})$.

Let $f = I$, then. $f - A_n f_p \geq \delta > \frac{1}{n} = \omega_r^{\varphi}\left(I, \frac{1}{n}\right)_p$.

Case 2: suppose that *for all* $\ell \in \{1, \ldots, n\} : x_\ell = \frac{\ell - 1}{n}$ otherwise, we already proved in case 1.

Assume that there exists $\ell \in \{1, \ldots, n\}$: $\mathscr{W}_\ell \neq Z(f, \ell) - \sum_{i=1}^{\ell-1} \mathscr{W}_i$, *let* $f = I$,

in this case, *there exists* $i \in \{0, \ldots, n\}$, *for all* $x \in [x_i, x_{i+1})$, such that:

$$(A_n f)(x) = c \neq Z(I, \ell) = x_i + \frac{1}{n}.$$

If $c > Z(f, \ell)$, then. $|f(x_i) - (A_n f)(x_i)| = c - x_i > x_i + \frac{1}{n} - x_i = \frac{1}{n}$.

If $c < Z(f, \ell)$, then $\lim |f(x) - (A_n f)(x)| = x_{i+1} - c > x_{i+1} + \frac{1}{n} - x_{i+1} = \frac{1}{n}$.

4. Debao Chen's theorem

In this section, we prove a type of Debao Chen's theorem for estimating the degree of approximation by multilayer feedforward artificial neural network with some hidden units.

Definition 4.1

If we fix the activation function σ and the integer number n, we will get the class of function to have the following form:

$$h(x) = \sum_{i=1}^{n} a_i \sigma(mx + r_i), \quad (x \in R).$$

For various types of parameters $a_i \in \mathbb{R}$, $m \in \mathbb{N}$ *and* $r_i \in \mathbb{Z}$, we get a type of class which will be denoted by $\Phi(\sigma, n)$. Now, what about the degree of approximation of $f \in L_p[0, 1]$ by elements of $\Phi(\sigma, n)$? To answer this question, we must define Dist$(f, \Phi(\sigma, n) = inf \{\|f - h_p\| : h \in \Phi(\sigma, n)\}$, and prove the following theorem. First, we have to introduce the following remark.

Remark 4.2: φ is called an activation function if and only if it satisfies $\lim_{x \to \infty} \varphi(x) = a$, $\lim_{x \to -\infty} \varphi(x) = b$, $a \neq b$, addition to its boundedness.

Theorem 4.3(Debao Chen)

For each f in $L_p[0, 1]$

$$\text{Dist}\left(f, \Phi(\sigma, n) \leq \|\sigma\|_p \omega_{r,m}^{\varphi}\left(f, \frac{1}{n}\right)_p.\right.$$

$$Here, \|\sigma\|_p = \left(\int_0^1 |\sigma(x)|^p dx\right)^{1/p}, \ x \in R$$

And Dist$(f, \Phi(\sigma, n)$ is the distance between f and $\Phi(\sigma, n)$.

Proof:

Define:

$$i = i/n, \quad f_i = f(x_i), \ (0 \leq i \leq n),$$

and

$$(L_m f)(x) = f_1 \sigma(mx + m) \sum_{i=1}^{n-1} \sum_{\ell=1}^{r} \binom{r}{\ell}$$

$$(-1)^{r-\ell} f(x_\ell + (\ell - r/2)h) \sigma(mx - mx_i). (4.3.1)$$

Assume that if m is a multiple of, then $L_m f \in \Phi(\sigma, n)$; so the operators $L_m f$ are linear.

By Remark (4.2), an activation function σ is a bounded function. The following function values converge to 0 when t goes to $+\infty$:

$$\vartheta(t) = \max\{\max_{x \geq t} |\sigma(x) - 1|, \max_{x \leq -t} |\sigma(x)|\}.$$

Now, to prove our theorem, let us prove the following estimate, in which $\omega_{r,m}^{\varphi}$ is the modulus of smoothness of f, and $0 < \delta < 1/2n$.

$$\|L_m f - f\|_p \leq \|\sigma\|_p \omega_{r,m}^{\varphi}\left(f, \frac{1}{n}\right)_p + \omega_{r,m}^{\varphi}(f, \delta)_p$$

$$+ \vartheta(m\delta)\left[\|f\|_p + n\omega_{r,m}^{\varphi}\left(f, \frac{1}{n}\right)_p\right], \ (4.3.2)$$

set $= 1/\sqrt{m}$, then $m \to \infty$ through the multiples of n. Since Dist$(f, \Phi(\sigma, n)) \leq \|L_m f - f\|_p$, and the bound in (4.3.2), this distance converges to. $\|\sigma\|_p \omega_{r,m}^{\varphi}\left(f, \frac{1}{n}\right)_p$.

Now to prove (4.3.2), we write:

$$(L_m f-f)(x)=f_k-f(x)+f_1[\sigma(mx+m)-1]+\sum_{i=1}^{k-1}\sum_{\ell=1}^{r}\left(\binom{r}{\ell}-1\right)(-1)^{r-\ell}fx_\ell+(\ell-r/2)h\Big)[\sigma mx-mx_i)-1]$$
$$+\sum_{i=k}^{n-1}\sum_{\ell=1}^{r}\binom{r}{\ell}(-1)^{r-\ell}f\left(x_\ell+(\ell-r/2)h\right)\sigma\left(mx-mx_i\right) \quad 4.3.3$$

Fix m and x. Let $0<\delta<1/2n$, we have two cases

Case 1 $|x-x_k|<\delta$ *for some k in* $\{1,2,\ldots,n\}$.

We will write equation (4.3.3), as follows:

Case 2

Suppose $|x-x_i|\geq\delta$ for all *i in* $\{1,2,\ldots,n\}$.

Choose k, such that $x_{k-1}\leq x\leq x_k$.

$$(L_m f-f)(x)=f_k-f(x)+f_1[\sigma(mx+m)-1]+\sum_{i=1}^{k-1}\sum_{\ell=1}^{r}\binom{r}{\ell}(-1)^{r-\ell}f(x_\ell+(\ell-r/2)h)[\sigma(mx-mx_i)-1]$$
$$+(f_{k+1}-f_k)\sigma(mx-mx_k)+\sum_{i=k+1}^{n-1}\sum_{\ell=1}^{r}\binom{r}{\ell}(-1)^{r-\ell}f(x_\ell+(\ell-r/2)h)\sigma(mx-mx_i). \quad 4.3.4$$

For $1\leq i\leq k-1$, we have

$$x-x_i\geq x-x_{k-1}\geq x_k-\delta-x_{k-1}=\frac{1}{n}-\delta\geq\frac{1}{n}-\frac{1}{2n}$$
$$=\frac{1}{2n}>\delta.$$

Hence.$(x-x_i)\geq m\delta$ *and* $|\sigma(mx-mx_i)-1|\leq\vartheta(m\delta)$.

Similarly, if $k+1\leq i\leq n-1$, then

$$x-x_i\leq x-x_{k+1}\leq x_k+\delta-x_{k+1}=-\frac{1}{n}+\delta\leq-\delta.$$

In this case $m(x-x_i)\leq-m\delta$ *and* $|\sigma(mx-mx_i)|\leq\vartheta(m\delta)$.

Also, $+m\geq m$, so that $|\sigma(mx+m)-1|\leq\vartheta(m)$. From (4.3.4), it follows that:

As in the previous steps and by using (4.3.3), we obtain:

$$|(L_m f-f)(x)|\leq\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p+\|f\|_p\vartheta(m\delta)$$
$$+n\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\vartheta(m\delta)\leq\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\|\sigma\|_p$$
$$+\vartheta(m\delta)\left[\|f\|_p+n\omega^{\varphi}_{m,r}\left(f,\frac{1}{n}\right)_p\right]+\omega^{\varphi}_{r,m}(f,\delta)_p.$$

Thus, from cases 1 and 2, the same upper bound has been obtained and the estimate (4.3.2) is proved. This leads to the proof of Theorem 4.3.

$$|(L_m f-f)(x)|\leq\omega^{\varphi}_{r,m}(f,\delta)_p+\|f\|_p\vartheta(m)+(k-1)\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\vartheta(m\delta)+\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\|\sigma\|_p$$
$$+(n-k-1)\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\vartheta(m\delta)\leq\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\|\sigma\|_p+\vartheta(m\delta)\left[\|f\|_p+n\omega^{\varphi}_{r,m}\left(f,\frac{1}{n}\right)_p\right]+\omega^{\varphi}_{r,m}(f,\delta)_p.$$

5. Conclusions

We know that we can use neural networks with an appropriate activation function for approximating continuous functions. In our work, we present function approximation on L_p space by using the ReLU activation function. We conclude that L_p universal approximation using neural networks with the ReLU activation function can be estimated. This approximation is optimal in terms of basis independent of the original function, then Debao Chen's theorem type for this approximation can be proved.

References

[1] Z. Chen, F. Cao, The approximation operators with sigmoidal functions, Math. Appl. 58 (2009) 758−765.

[2] F. Cao, Z. Xu, Y. Li, Pointwise Approximation for Neural Networks, International Symposium on Neural Networks, Springer, Berlin, Heidelberg, 2005, pp. 39−44.

[3] B. Yun, A neural network approximation based on a parametric sigmoidal function, Mathematics 7 (2019) 262.

[4] A. Schfer, H. Zimmermann, Recurrent Neural Networks Are Universal Approximators, International Conference on Artificial Neural Networks, Springer, Berlin, Heidelberg, 2006, pp. 632−640.

[5] E.S. Bhaya, Z.H. Al-sadaa, Stechkin- Marchaud inequality in terms of neural networks approximation in Lp-space for 0< p< 1, Mater. Sci. Eng. 571 (2020) 1−5.

[6] H.A. Almurieb, E.S. Bhaya, SoftMax neural best approximation, IOP Conf. Ser. Mater. Sci. Eng. 871 (2020) 1−10.

[7] F. Fan, J. Xiong, G. Wang, Universal approximation with quadratic deep neural networks, Neural Network. 124 (2020) 383−392.

[8] G. Cybenko, Approximation by superposition of sigmoidal functions, Math. Control, Signals, Syst. 2 (1989) 303−314.

[9] H. White, Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings, Neural Network. 3 (1990) 535−549.

[10] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Network. 4 (1991) 251−257.

[11] S. Draghici, On the capabilities of neural networks using limited precision weights, Neural Network. 15 (2002) 395−414.

[12] V.E. Ismailov, Approximation by neural networks with weights varying on a finite set of directions, J. Math. Anal. Appl. 389 (2012) 72−83.

[13] V.E. Ismailov, Approximation by ridge functions and neural networks with a bounded number of neurons, Appl. Anal. 94 (2015) 2245−2260.

[14] V.E. Ismailov, E. Savas, Measure theoretic results for approximation by neural networks with limited weights, Numer. Funct. Anal. Optim. 38 (2017) 819−830.

[15] B. Jian, C. Yu, Y. Jinshou, Neural networks with limited precision weights and its application in embedded systems, Proceedings of the second international workshop on education technology and computer science, Wuhan 1 (2010) 86−91.

[16] Y. Liao, S.C. Fang, H.L. Nuttle, A neural network model with bounded-weights for pattern classification, Comput. Oper. Res. 31 (2004) 1411−1426.

[17] M. Stinchcombe, H. White, Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights, in: Proceedings of the 1990 IEEE international joint conference on neural networks 3, 1990, pp. 7−16.

[18] Y. Ito, Approximation of continuous functions on Rd by linear combinations of shifted rotations of a sigmoid function with and without scaling, Neural Network. 5 (1992) 105−115.

[19] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proc. 27th Int. Conference Mach. Learn, 2010, pp. 807−814.

[20] U. Shaham, A. Cloninger, R. Colifman, Provable approximation properties for deep neural networks, Appl. Comput. Harmon. Anal. 44 (2018) 537−557.

[21] Z. Lu, H. Pu, F. Wang, Z. Hu, L. Wang, The expressive power of Neural network: avew from the width, in: Proceeding of the 31st International Conference on Neural Information Proceeding Systems, 2017, pp. 6231−6239.

[22] H. Lin, S. Jeglka, Resnet with one- neuron hidden layers is a universal approximation, Adv. Neural Inf. Process. Syst. 31 (2018) 6169−6178.

[23] D. Yarostky, Optimal approximation of continuous functions by very deep ReLU networks, in: Conference on Learning Theory (COLT) 75, 2018, pp. 1−11.

[24] S. Liang, R. Srikant, Why deep neural networks for function approximation?, in: International Conference on Machine Learning 70, 2017, pp. 2979−2987.

[25] D. Yarostky, Error bounds for approximations with deep ReLU networks, Neural Network. 94 (2017) 103−114.

[26] I. Safran, O. Shamir, Depth-width tradeoffs in approximating natural functions with neural networks, in: International Conference on Machine Learning (ICML) 70, 2017, pp. 2979−2987.

[27] H. Lin, M. Tegmark, D. Rolnick, Why closed deep and cheap learning work so well? J. Stat. Phys. 168 (2017) 1223−1247.

[28] J.M. Klusowski, A.R. Barron, Approximation by combinations of ReLU and squared ReLU ridge functions with l^1 and l^0 controls, in: IEEE Trans. Inf. Theor. 64, 2018, pp. 7649−7656.

[29] D.X. Zhou, Universality of deep convolutional neural networks, Appl. Comput. Harmon. Anal. 48 (2020) 787−794.

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

[31] K.A. Kopotun, D. Leviatan, I.A. Shevchuk, New moduli of smoothness: weighted DT moduli revisited and applied, Constr. Approx. 42 (2015) 129−159.

[32] K.A. Kopotun, Monotone polynomial and spline approximation in Lp, 0< p <∞(quasi) norm, Approx. Theor. VIII (1995) 295−302.