Journal of Interdisciplinary Mathematics
ISSN: 0972-0502 (Print), ISSN: 2169-012X (Online)
Vol. 27 (2024), No. 4, pp. 635–641
DOI : 10.47974/JIM-1734

Enhancing L_p approximation via dual hidden layers in neural networks

Sara Saleh Mahdi *
Eman S. Bhaya
Department of Mathematics
College of Education for Pure Sciences
University of Babylon
Babylon
Iraq

Abstract

Many researchers work on the approximation by the one hidden layer neural network. However, little work is done on multilayer neural network approximation. Here, we approximate multilayer neural networks by 2-hidden layers with sigmoid function output. That is what we shall introduce in our work.

Subject Classification: *92B20, 90B18.*

Keywords: *Best approximation, Universal approximation, Neural approximation.*

1. Introduction

Firstly, let us introduce the backpropagation algorithm, which is modified by the gradient descent method and modified weights and thresholds. So, the error is minimized between the signal of the network and the desired output. Let us define B(x) as the output function. The sigmoid function f(x) is

$$f(x) = \frac{1}{(1+e^{-x})}$$

In [1], we can find many mathematical models of neural networks. The researchers in [2] and [6] can find many applications, such as Pattern recognition, and show the mechanism of human information processing using these models. In [2]the authors introduced the back propagational algorithm (generalized delta rule)and learning rule for multilayer networks.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

* *E-mail:* **sara.salih638@gmail.com**

The authors in [2] introduced little theoretical research on multilayer networks. Lippmann in [7] showed that the arbitrary complex decision regions that contain concave regions, can use 4-hidden layers networks. The researchers in [4] introduced an example of 3-layer networks with thresholding units in which Partition space into concave subsets. In [3], the researchers showed that Pattern recognition applications can form several complex decision regions that 3-layer networks by simulations. In [5], the author proposed a based as generalized delta rule on the same principle. In [6], Sejnowski and Rosenberg apply multilayer networks for forming mappings, such as Net.

The authors in [10,11,14], introduced solutions to Hilbert's thirteenth Problem and proved an estimate for the approximation of continuous function using 4-layer neural networks.

In [8,13], the authors pointed to a problem: the unit of the neural network approximation is not a sigmoid function. In [14], the authors used a 3-layer neural network approximation. The condition of absolute integrability must satisfy the output function $B(x)$, so it is not a sigmoid function.

In [9], The researchers showed by a multilayer network, it has been known that any piecewise - linear decision region can be realized.

There is a different Point from the [1] model, which uses the output function of units, which is why it is a learning algorithm for multilayer networks. In this work, we started from [9], the integral formula, and proved the approximate realization of mappings in L_p spaces 3- layers networks are sigmoid functions whose output function is for hidden layers and output function for input. We use Kolmogorov and Sprecher theorem to prove our results. Here, we approximate multilayer neural networks in L_p by two hidden layers with sigmoid function output.

For a real-valued function $f: K \subset \mathbb{R}^n \rightarrow \mathbb{R}$

We Can use the following norm

$$\|f\|_{L_p(k)} = \left(\int_k |f(x_1, x_2, \ldots, x_n)|^p d_{x1}, \ldots, dx_n\right)^{\frac{1}{p}}.$$

2. Multilayer Neural Networks

Generally, the networks have input, hidden, and output layers. Each layer consists of computation units. The relationship between the input-output is through by connection, w_i weights, threshold b, inputs x_i, output y and differentiable function B as follows

$$Y = B\left(\sum_{i=1}^{k} w_i x_i - b\right).$$

The algorithm of back Propagation is the learning rule of this network in [2]. The relationship between the input units r and output units s for multilayer networks defines a continuous mapping from r-dimensional Euclidean space to s-dimensional Euclidean space. This mapping is called the input-output mapping of the network.

In the study of mappings of networks, output functions for hidden layer is $B(x)$, and output functions for input-output layers defined by multilayer networks to consider are linear.

Let Points for r-dimensional Euclidean space $\mathbb{R}^r$ be denoted by $X = (x_1, x_2, \ldots, x_n)$ and the norm of x defined by $\|x\| = \left(\sum_{i=0}^{r} x_i^2\right)^{\frac{1}{2}}$

In this work, we prove the following theorems and corollaries.

Theorem 2.1 (Kolmogorov): *Any* $f \in L_p(\mathbb{R}^n)$, $f(x_1, x_2, x_3, \ldots, x_n) \in \mathbb{R}^n$ *defined on* $(n \geq 2)$ *of several variables, we can be represented by [11]*

$$f(x) = \sum_{j=1}^{2n+1} xi\left(\sum_{i=1}^{n} \psi_{ij}(x_i)\right).$$

Where

$x_i\psi_{ij} \in L_p(\mathbb{R}^n)$ with one Variable and ψ_{ij} is sigmoid function, which not dependent on f.

Theorem 2.2 (Sprecher): *For each integer* $(n \geq 2)$, *there exists areal function* $B(x), B([0,1]) = [0,1]$ *dependent on n and having the following property*

If $\delta > 0$, there exists an arational number $w_k\, 0 < w_k < \delta$, every function f(x) of n variable on I^n, we can be represented by[14]

$$\boldsymbol{f(x) = \sum_{j=1}^{2n+1} x\left[\sum_{i=1}^{n} \lambda\, B\big(x_i + w_k\,(j-1)\big) + j - 1\right]}.$$

Where

λ and the function $X \in L_p(\mathbb{R}^n)$ is an independent constant of f.

In [10], the mapping $f: x \in I^n \to (f_1\,(x), \ldots, f_n(x)) \in \mathbb{R}^m$ is a 4-layers neural network with hidden units whose output function are Ψ, $x_i(i = 1, \ldots .\, m)$,

Where x_i is given in the theorem [14] for $f_i(x)$, $x_i(i = 1, \ldots, m)$ are used for the 2- hidden layers and Ψ for the one hidden layer.

Lemma 2.3: *If* $f(x) \in L_p(\mathbb{R}^n)$ *then* $F(w) \in L_p(\mathbb{R}^n)$

Proof: We have

$$\left\|\widehat{T}_f(w)\right\|_p = \|\, T_f(\widehat{w})\|_p = \left\|\int_{I.R.^n} f(x)\widehat{w}\,(x)\,dx\right\|_p,$$

$$\leq \left|\int_{I.R.^n} f(x)\widehat{w}\,(x)\,dx\right|,$$

$$\leq \|f\|_2\,\|\widehat{w}\|_2$$

$$= \|f\|_2\,\|w\|_2 \leq \|f\|_1\,\|w\|_1.$$

Since $f, w \in L_p(\mathbb{R}^n)$ and since $L_2(\mathbb{R}^n) \subset L_p(\mathbb{R}^n)$,

So, that $\|f\|_2, \|w\|_2 < \infty$.

This implies to

$$F(w) \in L_p(\mathbb{R}^{\boldsymbol{n}}).$$

Theorem 2.4: *Let k be a compact subset of R^n , Let $B(x)$ be anon Constant function and $f(x_1, x_2, x_3, \dots, x_n)$ be a real valued function in $L_p(k)$. Then there exist $M \in \mathbb{Z}^+$ and real Constants b_i , $c_i (i = 1, \dots, M), w_{ij} (i = 1, \dots, D, j = 1, \dots, n)$ such that*

$$f^*(x_1, \dots, x_n) = \sum_{i=1}^{M} C_i \ B\left(\sum_{i=1}^{n} w_{ij} x_j - b_i\right)$$

Satisfies

$$\|f(x_1, \dots, x_n) - f^*(x_1, \dots, x_n)\|_p < c(p) w_v (f, \delta)_p,$$

where

For $f: K \subseteq \mathbb{R}^n \to \mathbb{R}$, there exists a two - layers network whose a linear function with output functions for input-output layers which has an input-output function $f^*(x_1, \dots, x_n)$ and the function $B(x)$, whose output functions for the hidden layer

Such that

$$\|f(x_1, \dots, x_n) - f^*(x_1, \dots x_n)\|_p < c(p) w_v (f, \delta)_p.$$

Proof: We can find $H > 0$, and we define $I_H \ (x_1, \dots, x_n)$, $I_{\infty,H} \ (x_1, \dots, x_n)$ and $J_H \ (x_1, \dots, x_n)$ as

$$I_H \ (\boldsymbol{x_1}, \dots, \boldsymbol{x_n}) = \int_{-H}^{H} \cdots \int_{-H}^{H} \boldsymbol{\psi} \left(\sum_{i=1}^{n} \boldsymbol{x_i} \boldsymbol{w_i} - \boldsymbol{w_0}\right)$$

where

$\psi(x) \in L^1$ −topology is defined by $\psi(x) = B\left(\frac{x}{\delta + \propto}\right) - B\left(\frac{x}{\delta - \propto}\right)$ for some δ and $\propto$ by the Paley - wiener theorem in [12], the real analytic is the Fourier transform$F(w)\big(w = (w_1, \dots, w_n)\big)$of $f(x)$, such that

$$\left\| I_{\infty,H} \ (x_1, \dots, x_n) - f \ (x_1, \dots, x_n) \right\|_p < w_v (f, \frac{k}{n})_p. \tag{1}$$

By finite integrals on k, we will approximate $I_{\infty,H}$.

$I_{H*,H} \quad (x_1, \dots, x_n) =$

$$\int_{-H}^{H} \cdots \int_{-H}^{H} \left[\int_{-H^*}^{H*} \psi \left(\sum_{i=1}^{n} x_i w_i - w_0\right) \times \frac{1}{(2\Pi)^n \, \psi(1)} F(w_1, \dots, w_n) \times \exp(i w_0) d w_0 \right] d w_1, \dots, d w_n.$$

We can take $H^* > 0$ so that

$$\left\| I_{H*,H} \ (x_1, \dots, x_n) - I_{\infty,H} (x_1, \dots, x_n) \right\|_p < w_v (f, \frac{k}{n})_p. \tag{2}$$

By using the equation [9]

$$\int_{-H^*}^{H*} \psi \left(\sum_{i=1}^{n} x_i w_i - w_0\right) \exp(i w_0) d w_0 =$$

$$\int_{\sum_{i=1}^{n} x_i w_i - H^*}^{\sum_{i=1}^{n} x_i w_i + H^*} \psi(t) \exp(it) dt . \exp(i \sum_{i=1}^{n} x_i w_i),$$

From $F(x) \in L_p$ topology and compactness of $[-H, H]^{\,n} * k$, We can take H^*, so

$\int_{-H^*}^{H*} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0\right) \exp(iw_0)\, dw_0 - \int_{-\infty^*}^{\infty} \psi\left(\sum_{i=1}^{n} x_i w_i - w_0\right) \exp(iw_0)\, dw_0$

$\leq \|\psi(\sum_{i=1}^{n} x_i w_i - w_0) \exp(iw_0)\, dw_0 - \psi(\sum_{i=1}^{n} x_i w_i - w_0)\exp(iw_0)\, dw_0\|_1$

$\leq \frac{w_v\left(f,\frac{k}{n}\right)(2\Pi)^n \|\psi(1)\|_1}{\left(2\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\|F(x)\|_1 dx+1\right)} \times \int_{-H}^{H}\cdots\int_{-H}^{H}\|F(x)\|_1\, dx$ **on** k.

Since ψ and F are belong to finite dimensional space then

$$\left\|\frac{w_v\left(f,\frac{k}{n}\right)(2\Pi)^n \|\psi(1)\|_1}{\left(2\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\|F(x)\|_1 dx+1\right)}\right\|_p \leq \frac{\|F\|_p}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\|F(x)\|_p} \times \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\|F(x)\|_p \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\|F(x)\|_1$$

$$\left\|I_{-H*,H}\ \ (x_1,\ldots,x_n) - I_{\infty,H}\ \ (x_1,\ldots,x_n)\right\|_p \leq \frac{w_v\left(f,\frac{k}{n}\right)}{\left(2\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\|F(x)\|_p dx+1\right)} \times \int_{-H}^{H}\cdots\int_{-H}^{H}\|F(x)\|_p\, dx < c(p) w_v\left(f,\frac{k}{n}\right)$$

from (1) and (2) we get for any $H, H^* > 0$,

$$\left\|f\ \ (x_1,\ldots,x_n) - I_{H^*,H}\ \ (x_1,\ldots,x_n)\right\|_p < c(p)\, w_v(f,\delta)_p.$$

So, we can approximate $f(x)$ by the finite integrals $I_{H^*,H}$, (x) on K.

It can be represented by the Riemann sum and by two-layer networks.

Then $f(x)$ can be represented approximately by two-layer networks.

Proposition 2.5: Let $f(x) \in Lp\ (\mathbb{R})$ and $B(x)$ is a sigmoid function for an arbitrary compact subset K of I.R. and an arbitrary $w_k(f,\delta)_p > 0$, there exist$M \in\ \mathbb{Z}^+$ and real constant

$a_i, b_i, c_i (i = 1,\ldots, M)$, such that

$\|f(x) - \sum_{i=1}^{M} c_i\, B\,(a_i X\ \ + b_i)\|_p < c(p) w_k(f,\delta\,)_p$, hold on compact subset K.

Proof: We suppose that $K\ =\ [0,1]^n$, and we apply Sprecher's theorem to $g_p(x)$ $(p = 1,\ldots N)$ and represent $g_p(x)$ by the form

$$g_p(x) = \sum_{i=1}^{2m+1} x_p \left[\sum_{i=1}^{m} h^i \psi\ \left(x_i + t\ \ (j-1)\right) + j - 1\right], (p = 1,\ldots N)$$

Where

h,t are constants. We can approximate the functions by using a sigmoid function$B$and apply our Proposition to the function Y_p, ψ.

Let $K_j(j\ =\ 1,\ldots 2m\ +1)$ be the images of $[\ 0,1\ ]^n$ by mappings

$$\lambda_j : x \to \sum_{i=1}^{m} h^i\, \psi\,(x_i + t(j-1)) +\ j - 1, (j\ = 1,2,\ldots,2m+1)$$

we can take k_δ of δ neighbourhood of $K \in L_p(\mathbb{R})(P = 1, \dots, N)$ are approximated by

$$Y_{p,M}(\, X\,) = \sum_{i=1}^{M} C_{i,M}\, B(\, a_{i,M}\mathrm{x} \, + b_{i,M}) \tag{3}$$

So,

$$\left\| Y_p(x) - Y_{p,M}(x) \right\|_p < c(p)\, w_k\, (f, \delta)_{p.} \tag{4}$$

We can approximate $\lambda_j$on $[0,1]^n$ by λ_j, S so that

$$\left\| \lambda_j(x) - \lambda_j\, S(x) \right\|_p < \min(r, \delta). \tag{5}$$

Where r sufficiently small, and we can approximate $\psi(x)$ by

$$\psi S(x) = \sum_{i=1}^{s} c_i^*\, B(a_i^* X + b_i^*) \tag{6}$$

On ℓ neighborhood of [0,1] and set

$$\lambda_{j,s}(x) = \sum_{i=1}^{m} h^i\, \psi_s\big(x_i + t(j-1)\big) + j - 1 \tag{7}$$

So, by (5) is satisfied.

By using transformation

$$\sum_{i=1}^{2m+1} y_p\left[\lambda_j(x)\right] - \sum_{j=1}^{2m+1} y_{p,M}\left[\lambda_{j,s}(x)\right]$$

$$= \sum_{i=1}^{2m+1} y_p\left[\lambda_j(x)\right] - \sum_{j=1}^{2m+1} y_{p.M}\left[\lambda_j(x)\right] + \sum_{j=1}^{2m+1} y_{p.M}\left[\lambda_j(x)\right] - \sum_{j=1}^{2m+1} y_{p.M}\left[\lambda_{j,s}(x)\right]$$

Then, we get $g_p(x)(p = 1, \dots \dots .. N)$, are approximated by

$$\sum_{j=1}^{2m+1} Y_{p,M}\left[\lambda_{j,s}(x)\right](p = 1, \dots N) \text{ on } [0,1]^n$$

3. Conclusions

If $f \in L_p(\mathbb{R}), 0 < p < 1$, then there exists a neural network of 2 hidden layers with a sigmoid activation function as the best approximation function of f.

4. Future Works

As a future work, we can use neural networks with more than 2 hidden layers of another activation function kind as the best approximation functions in quasi-normed spaces.

References

[1] W. S. McCulloch and W. Pitts. A logical calculus of the idea immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5; pp. 115-133 (1943).

[2] D.E.Rumelhart, G . E.Hinton, and R.J. Williams, Learning representations by error Propagation. In D. E. Rumelhart, J.L. McClelland and

the PDP Research Group (Eds.), Parallel distributed Processing Cambridge, MA: MIT Press, Vol . 1, P.P. 318-362 (1986).

[3] W.Y.Huang, and R.P. Lippmann, Neural Net and traditional classifiers. In D.Z. Anderson (Ed.), Neural information Processing System, Denver, Colorado, New York: American Institute of Physics, P.P. 387-396 (1987).

[4] A. Wieland, and R. Leighton, Geometric analysis of neural network capabilities, *IEEE First International Conference on Neural Networks,* 3, pp . 385-392 (1987).

[5] S. Amari, A theory of adaptive Pattern classifiers, *IEEE Transactions on Electronic Computers*, E.C. - 16, pp. 299-307 (1967).

[6] T.J. Sejnowski, and C.R. Rosenberg, Parallel networks that learn to Pronounce English text. Complex System, 1, pp.145-168 (1987).

[7] R.P. Lippmann, An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4, pp. 4-22 (1987 April).

[8] Y. Uesaka, Analog Perceptrons: On the additive representation of functions. *Information and Control,* 19, pp. 41-65 (1971).

[9] R.O.Duda, and H. Fossum , Pattern classification by iteratively determined linear and Piecewise linear discriminant functions. *IEEE Transactions on Electronic Computers*, E.C. - 15, pp. 220-232 (1966).

[10] R. Hecht - Nielsen, Kolmogorov mapping neural network existence theorem IEEE First International Conference on Neural Networks, 3, pp. 11-13 (1987).

[11] A.N. Kolmogorov, On the representation of continuous functions of many variables by superposition of continuous functions of one Variable and addition. Doklady Akademii Nauk SSSR, 144, pp.679-681 (1957). *American Mathematical Society Translation*, 28, pp. 55-59 (1963).

[12] K.Yosida, Functional analysis. New York: Springer- Verlag, (1968).

[13] T. Poggio, Visual algorithms In O.J. Braddick and A.C. Sleigh (Eds.) Physical and Biological Processing of Images. New York: Springer-Verlag, pp. 128-135 (1983).

[14] Shubham Sharma & Ahmed J. Obaid. Mathematical modelling, analysis and design of fuzzy logic controller for the control of ventilation systems using MATLAB fuzzy logic toolbox, *Journal of Interdisciplinary Mathematics*, 23 : 4, 843-849 (2020), DOI: 10.1080/09720502.2020.1727611.

Received October, 2023