Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Babylon
College of Education for Pure Sciences
Department of Mathematics

Function Approximation by General ReLU Neural Network

A Research

Submitted to the Council of the College of Education for Pure Sciences of University of Babylon in Partial Fulfillment of the Requirement for the Degree of Higher Diploma

Education \ Mathematics

By :
Amjed Hamzah Sajat Kaeid

Supervised by :

Dr. Hawraa Abass Fadhil ALMurieb

2021 A.D 1443 A.H

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

﴿ يَرۡفَعِ ٱللَّهُ ٱلَّذِينَ ءَامَنُواْ مِنكُمۡ وَٱلَّذِينَ أُوتُواْ ٱلۡعِلۡمَ دَرَجَٰتٖۚ

وَٱللَّهُ بِمَا تَعۡمَلُونَ خَبِيرٞ ١١ ﴾

صَدَقَ ٱللَّهُ ٱلۡعَظِيمِ

سورة المجادلة: الآيه 11

Dedication

To the one who encouraged me to persevere all my life,

(dear father)

to the giving heart

(My beloved mother)

To those who the made an effort to help me and were best

support

(My brothers and sisters)

To my family, to my friends and colleagues…

To everyone who contributed even a letter to my academic

life…..

To all of them: I dedicate this work, which I ask God

Almighty to accept sincerely….

ACKNOWLEDGEMENTS

Alhamdulillah , Who helped us to accomplish this work. I would like to express my sincere thanks and gratitude to my supervisor Dr. Hawraa Abbas Fadhil for the continuous support. She is always motivating, inspiring, encouraging and her guidance helped me in all the time of the research .

I would like to express my sincere gratitude to my family , for my lovely parents whose hidden prayers bless me throughout my life and My brothers and sisters for encouraging me to finish this work. I would like to thank the staff of College of Education for Pure Sciences, especially those members of Department of Mathematics in their insightful comments and encouragement, I would also like to express my sincere thanks to my teachers who I learned from at all stages of my studiesand .

ABSTRACT

Approximating function by using neural networks is an interesting direction in approximation theory. so it is important to prove that universaly for functions from other wider spaces.

We essential to well choosing the space of the functions that are approximated by neural networks. L_p spaces of functions are fantastic choices to study. It is more interesting to take the value $0 < p < 1$.

On the other hand, the solid formulas of the neural networks as approximates are not less important. We know that we can use the neural networks for the approximation of functions for many types of activation functions. Here, we treat neural networks with simple and effiecient activation function called Rectified Linear Units (ReLU). Generalized ReLU is prove here to define anew formula of neural network that is appropriate for function approximation .

The main objective of this paper is to introduce a type of neural network and we use it to approximate functions and estimate the general approximation error. We will get the optimal approximation if we have a basis independent of the target function .

We prove theorems of existence and uniqueness by estimating the degree of approximation depending on the modulus of smoothness of the functions envolved.

Table Of Contents

LIST OF SYMBOLS

Symbol	Definition	Page
$g_0(\mathbf{t})$	polynomial from the function $x(t)$	**5**
$\|\| \cdot \|\|$	Euclidean Norm	**7**
$\boldsymbol{x}$, $\boldsymbol{y}$	Vectors	**7**
$\boldsymbol{L_p}$	Lebesgue space	**8**
$\|\| \cdot \|\|_p$	L_p- norm	**11**
$\boldsymbol{E_n(.)_p}$	Degree of best approximation	**13**
$\boldsymbol{\varepsilon}$	Epsilon	**13**
$\boldsymbol{\omega_k(.,.)_p}$	k-th modulus of smoothness	**14**
$\Delta_h^{(k)}$	k-th devided difference	**14**
$\mathbf{N(x)}$	Neural Networks	**16**
$\boldsymbol{\sigma}$	Activation function	**16**
$\mathcal{R}^d$	The set of real numbers of dimension d	**18**
$\{f_n\}_{n\geq 1}$	Sequence of functions	**28**
FNN	Feed forward neural network	**29**

INTRODUCTION

There are many studies about the approximation by neural networks with different types of activation functions. The theory of function approximation through neural networks has a long history dating back to the work by McCulloch and Pitts and the seminal paper by Kolmogorov [1], who showed, when interpreted in neural network parlance, that any continuous function of n variables can be represented exactly through a 2-layer neural network of width ($2n + 1$) [2] .

Cybenko [3] was the first who gave a strong theoretical base to researchers by his universal theorem of approximation in 1988. It gives a best neural approximation for any continuous function defined on a compact space X by a single hidden layer. Many papers depended on Cybenko's theorem to make developments in many fields. Multilayer neural networks were the universal approximates of mappings and its derivatives by Hornik and Stinchombe in their joint work in 1990 [4]. Two years later, functions from Hilbert spaces were approximated neutrally by Jones et al in [5] and non Hilbert spaces by[6], while Chen and Chen extended Cybenko's Theorem in [7] and

to functions from several variables in [8]. Three layer neural networks was the type that Suzuki studied and approximate in 1998 [9]. In the beginning of the millennium, researchers became more interested in studying neural approximation.Different neural networks with different activation functions and different function spaces were studied, we cite examples as [10] by Dingankar and Phatak whose neural network was radial basis with simultaneous approximation . In 2006, Ding, Cao and Xu used trigonometric class of hidden layers in [11] and later in [12]. Until the moment of writing this, many researchers are still studying the topic of nural approximation from several directions, see [1]–[13], for more improvements, generalizations, and/or constructions.

In this Research, we investigate the approximation properties of neural networks. In other words, we study how complex networks need to be in order to approximate certain functions well. For this, we focus on networks that use a certain activation function which is possibly the most widely used in applications—the rectified linear unit (ReLU). The research consists of two chapters. In chapter one , we introduce an overview of the basics of function approximation and neural network and activation functions (especially, Rectified Linear

Units) and lists basic elements needed in the neural network constructions considered throughout .

In Chapter two , we define General ReLU Neural Network and introduce function approximation by it. We also prove existence and uniqueness of best approximation.

our main results are listed below ,

2.4 . Existence Theorem

Let $f \in L_p\ [a, b]$, then there exists a simultaneous approximation to any subset $\mathcal{F}$ of L_p by a FNN of the form

$$\mathcal{N}_n(x) = \sum_{i=1}^{n} c_i \mathcal{R}(w_i x + \vartheta_i),$$

2.5. Uniqueness Theorem

The simultaneous best approximation $N^* \in \aleph$ of a subset $\mathcal{F}$ of L_p is unique.

CHAPTER ONE

INTRODUCTION TO FUNCTION APPROXIMATION AND NEURAL NETWORKS

1.1. Function Approximation

The theory of approximation is a very extensive field, which has various applications. The problem of function approximation had been studied throughout polynomial (trigonometric and algebraic), splines, wavelets and neural networks. In this section, we give an introduction to fundamental ideas and aspects of approximation theory in the normed space X. Approximation theory is concerned with how functions can best be approximated with simpler functions, or functions from applicable fields, with quantitively characterizing the errors introduced thereby. More generally, one may want to setup practically useful criterion for the quality of approximations [14] .

Starting from problems concerning certain mechanisms (e.g. the motion of the connecting-rod of a steam engine), P.L . Cebysev was led to state, a century ago [15], the problem of finding, for a real continuous function $X(t)$ on a segment $[a, b\,]$, an algebraic polynomial

$$g_0(\mathrm{t}) = \sum_{i=1}^{n} \alpha_i t^{i+1}$$

of degree $\leq n-1$ such that the "deviation" of the polynomial $g_0(\mathrm{t})$ from the function $X(\,t)$ on the segment $[a, b]$ be the least possible among the deviations of all algebraic polynomials of the from

$$g(\mathrm{t}) = \sum_{i=0}^{n} \alpha_i t^{i+1}$$

of degree$\leq n-1$.

In the mechanical problems considered, for the measuring of the deviation between $g(t)$ and $X(t)$ on the segment $[a,b]$, P. L. Cebysev [16] has found as being the most suitable the number

$$\max_{t\in[a,b]} |x(t) - g(t)|$$

thus, the problem amounts to the minimization of this maximum when $g(t)$ runs over the set of all algebraic polynomials of degree $\leq$ $n-1$. In order to include other important cases as well, the problem has been generalized by other mathematicians, the interval $[a,b]$ being replaced by a compact space Q [16] the real-valued functions by complex-valued functions, as in [17], or by functions with values in more general spaces [18] .

1.2. Normed Spaces

Before going deeper inside our study, we need to give some important concepts that concerns with our space of study.

Definition 1.2.1. [19]

A norm on a vector space V is a function $||\cdot||: V \to R$ satisfying

[i] $||x|| \geq 0$, with equality if and only if x = 0,

[ii] $||x + y|| \leq ||x|| + ||y||$ for all x and y in V .

[iii] $||cx|| = |c|||x||$ for all scalars c and $x \in V$.

Given a norm on a vector space, we get a metric by $\mathrm{d}(\mathrm{x},\mathrm{y}) = ||\mathrm{x} - \mathrm{y}||$

Definition 1.2.2. [13]

A metric space is a set $S \neq \emptyset$ together with a function

$$\rho: S \times S \rightarrow R^{+}$$

(called a metric for S) satisfying the metric laws (axioms):

For any $x, y, and\ z\ in\ S$, we have

(i) $\rho(x,y) \geq 0$ and $\rho(x,y) = 0\ iff\ x = y$;

(ii) $\rho(x,y) = \rho(y,x)$ (symmetry law) ;

(iii) $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$ (triangle law).

Definition 1.2.3.

A quasi-norm on a vector space V is a function $|| \cdot ||: V \rightarrow R$ satisfying

• $||x|| \geq 0$, with equality if and only if $x = 0$.

• $||x + y|| \leq C(||x|| + ||y||)$ for all $x\ and\ y\ in\ V$.

• $||cx|| = |c|||x||$ for all scalars $c\ and\ x \in V$.

Definition 1.2.4. [13]

A Banach space is a vector space V equipped with a norm $||\cdot||$, with respect to the metric defined by $d(\cdot,\cdot)$, V is complete.

1.3. L_p Spaces

Again, as in Example 1.2.2, we go back to the Lebesgue space that we need for our main results. The space L_p are function spaces defined using a natural generalization of the *p*-norm for finite-dimensional vector spaces. They are sometimes called Lebesgue spaces, named after Henri Lebesgue [17], although according to the Bourbaki group they were first introduced by Frigyes Riesz [20]. L_p spaces form an important class of Banach spaces in functional analysis, and of topological vector spaces.

Because of their key role in the mathematical analysis of measure and probability spaces, Lebesgue spaces are used also in the theoretical discussion of problems in physics, statistics, finance, engineering, and other disciplines.

Definition 1.3.1 [19]

The L_p space for $0 < p < \infty$ is given by

$$L_p([a,b]) = \{f: [a,b] \to \mathbb{R}, \quad f \text{ measurable and } \|f\|_p < \infty\}$$

Where

$$\|f\|_p = \left\{ \int_a^b |f|^p \right\}^{1/p}$$

is the L_p-quasi normed space of f.

Set the following space which is a special case of L_p which is called the discrete norm .

$$\mathcal{L}_p([a,b]) = \left\{ f \text{ is measurable} | \left(\sum_{i=1}^{n} \frac{b-a}{n} |f(x_i)|^p \right)^p < \infty \right\}$$

Example 1.3.1. [19]

The space $C[0,1]$ of continuous real-valued functions on [0, 1] has the sup_norm

$$\|f\|_{sup} = sup_{x\in[0,1]} |f(x)|$$

and the L_2-norm

$$\|f\|_2 = \left(\int_0^1 |f(x)|^2 dx \right)^{\frac{1}{2}},$$

While functions that are close in the sup-norm are close in the L_2-norm, the converse is false: a function whose graph is close to the x-axis except for a tall thin spike is near 0 in the L_2-norm but not in the sup-norm.

Example 1.3.2 [19]

Let $(X, \mathcal{M}, \mu)$ be a measure space. For any $p > 0$, set

$$L_p(\mu) \stackrel{\text{def}}{=} \left\{ f: X \to \mathrm{R}: f \text{ measurable and } \|f\|_p = \left(\int_X |f|^p \mathrm{d}\mu \right)^{1/p} < \infty \right\}$$

The spaces L_p for $p \geq 1$ have their metric coming from a norm. For $0 < p < 1$, is L_p is not a normed space, but it is called quasi-normed space .

Example 1.3.3. [19]

The metric used on L_p of Example 1.2.2 is

$$d(f,g) = \begin{cases} \left(\int_X |f-g|^p \mathrm{d}\mu \right)^{1/p}, & \text{if } p \geq 1 \\ \int_X |f-g|^p \mathrm{d}\mu, & \text{if } 0 < p < 1 \end{cases}$$

The reason for these choices, comes from the common properties of L_p for all $p > 0$.

1.3.1. Differences between L_p when $p \geq 1$ and L_p when $p < 1$ [16]

Property	$p \geq 1$	$0 < p < 1$
Triangle Inequality	$\|f+g\|_p \leq \|f\|_p + \|g\|_p$	$\|f+g\|_p \leq 2^{\frac{1}{p}-1}(\|f\|_p + \|g\|_p)$
Holder inequality	$0 < p < \infty$, $0 < q < \infty$, such that $\frac{1}{p}+\frac{1}{q} = 1$, then $\|fg\|_1 \leq \|f\|_p \|g\|_p$ In the case $p = q = 2$ yield the familiar cauchy Schwarz inequality	$0 < p < 1$, $0 < q < 1$, such that $\frac{1}{p}+\frac{1}{q} = 1$, then $\|f\|_p \|g\|_p \leq \int fg$
Minkowiski inequality	$\|f+g\|_p \leq \|f\|_p + \|g\|_p$	$\|f+g\|_p \geq \|f\|_p + \|g\|_p$ Also from triangle inequality $\|f+g\|_p \leq C(\|f\|_p + \|g\|_p)$
Metrability	$d(f,g) = \left(\int_0^1 \|f(t)-g(t)\|^p dt\right)^{\frac{1}{p}}$	$d(f,g) = \int_0^1 \|f(t)-g(t)\|^p dt$
Completeness	L_p is Banach space,	L_p is quasi-Banach space.
Locally Convexity	L_p is locally convex	L_p is not locally convex.

1.4. Existence and Uniqueness of Best Approximation [21]

1.4.1.Existence Theorem

Let Y be a finite-dimensional subspace of a normed linear space X, and let $x \in X$. Then, there exists a (not necessarily unique) $y^* \in Y$ such that $$\|x - y^*\| = \min_{y \in Y} \|x - y\|$$ for all $y \in Y$. That is, there is a best approximation to x by elements of Y .

To state uniqueness theorem, we need the following concept about convexity

Definition 1.4.1 (Strictly Convex Space):

Any norm $\|\cdot\|$ on a vector space X is said to be strictly convex if, for any $x \neq y \in X$ with

$$\|x\| = r = \|y\|$$

we always have

$$\|\lambda x + (1 - \lambda)y\| < r$$

for any $0 < \lambda < 1$.

1.4.2.Uniqueness Theorem [21]

X has a strictly convex norm if and only if the triangle inequality is strict on non-parallel vectors; that is, if and only if

$x \neq \alpha y, y \neq \alpha x,$ for all $\alpha \in R$ Then $\|x + y\| < \|x\| + \|y\|$.

1.4.3 Degree of Approximation [21]

The degree of best approximation of the function $f \in X$ is given by

$$E_n^*(f) = E_n(f) = \inf_{\emptyset \in \mathrm{Y}} \|f - \emptyset\|$$

where $\emptyset \in \mathrm{Y}$ is a function that is a best approximation of f out of Y.

Many theorems studied the existence of best approximation and the degree of approximation beginning with Weierstrass, in 1885, that proved the existence theorm of best approximation out of polynomials as follow

Theorem 1.4.4. (The Weierstrass Approximation Theorem): [21]

Let $P_n[a, b] \in C[a, b]$. Then, for every ε > 0, there is a polynomial(a best approximation from Y) p such that

$$\|f - p\| < \varepsilon$$

Where $P_n[a, b]$ is the space of all polynomials of degree $\leq n$.

1.5. Modulus of Smoothness

Modulus of smoothness is a measure of “Epsilon” in the definition of uniform continuity. The first time to use modulus of continuity to measure the degree of approximation, was Jackson’s Theorem in 1921 in his work [22], he gave an upper bound for the approximation speed. Moreover, in the midst of his work on generalizing Lipschitz condition

functions, Bernstein introduced higher orders of moduli of continuity in his paper [23].

To get a fuller description of function properties, modulus of smoothness is the best measure of the rate of approximation, since it judges the accuracy of the best approximation of a function as the error of approximation is estimated.

In our work, we define the modulus of smoothness in L_p to measure the degree of best approximation, as follow .

Definition 1.5.1 [14]

Let $f \in L_p$, then the kth symmetric difference of f is given by

$$\Delta_h^k(f,x,[a,b]) = \left\{ \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} f\left(x - \frac{kh}{2} + ih\right) , x \pm \frac{kh}{2} \in [a\,,b] \right\}$$

So the kth modulus of smoothness of f is given by

$$\omega_k(f,\delta,[a,b])_p = \sup_{0<h\leq\delta} \|\Delta_h^k(f,.)\|_p$$

for some $\delta \geq 0$.

1 .6 Neural Networks [24]

Let us commence with a provisional definition of what is meant by a neural network" and follow with simple, working explanations of some of the key terms in the definition.

A neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal (or human) neuron.

The processing ability of the network is stored in the interunit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns.

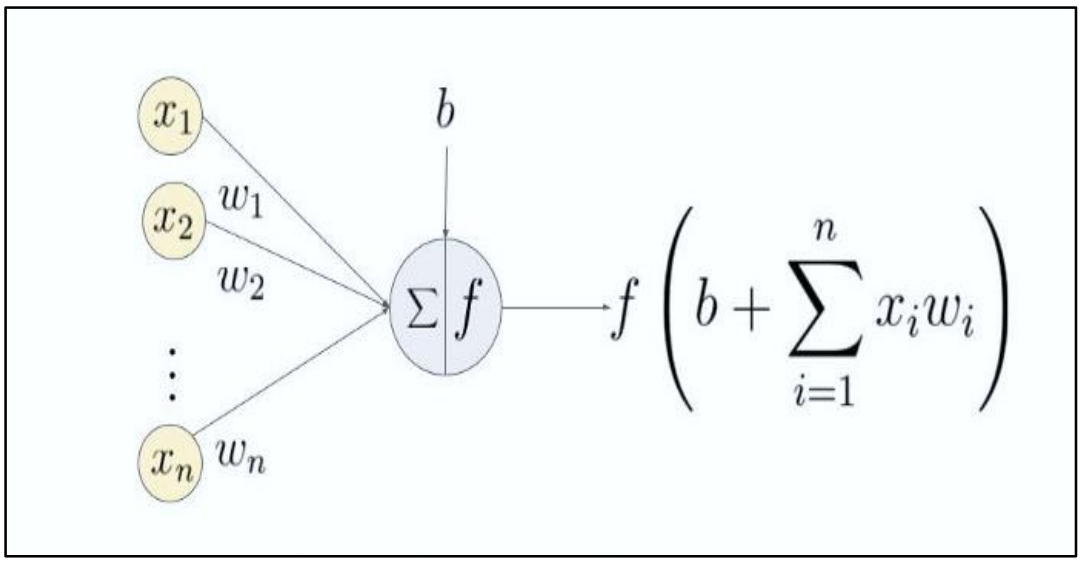
Neural networks are often used for statistical analysis and data modelling, in which their role is perceived as an alternative to standard nonlinear regression or cluster analysis techniques (Cheng & Titterington 1994). Thus, they are typically used in problems that may be couched in terms of classification, or forecasting. Some examples include image and speech recognition, textual character recognition, and domains of human expertise such as medical diagnosis, geological survey for oil, and financial market indicator prediction. This type of problem also falls within the domain of classical artificial intelligence (AI) so that scientists see neural nets as offering a style of parallel distributed computing, thereby providing an alternative to the conventional algorithmic techniques that have dominated in machine intelligence...

1.6.1. Mathematical Expression of Neural Networks [24]

We have to introduce the mathematical formula of neural network, see Figure(1.1), as follow

$$N(x) = \sigma\left(b + \sum_{i=1}^{n} x_i w_i\right)$$

- b = bias
- x = input to neuron
- w = weights
- n = the number of inputs from the incoming layer
- i = a counter from 1 to n
- σ = the activation function



Figure(1.1) Neural Network

Three things are happening here. First, each input is multiplied by a weight:

$$x_1 \rightarrow x_1 * w_1$$

$$x_2 \rightarrow x_2 * w_2$$

$$\vdots$$

$$x_n \rightarrow x_n * w_n$$

Next, all the weighted inputs are added together with a bias *b*:

$$(x_1 * w_1) + (x_2 * w_2) + \cdots + (x_n * w_n) + b$$

Finally, the sum is passed through an activation function:

$$y = \delta(x_1 * w_1 + x_2 * w_2 + \cdots + x_n * w_n + b)$$

1.7 Neural Network and Function Approximation [25]

Hornik (1991) showed that any bounded and regular function $R^d \rightarrow R$ can be approximated at any given precision by a neural network with one hidden layer containing a finite number of neurons, having the same activation function, and one linear output neuron. This result was earlier proved by Cybenko (1989) in the particular case of the sigmoid activation function. More precisely, Hornik's Theorem can be stated as follows

Theorem 1.7.1 [25]

Let ∅ be a bounded, continuous and non decreasing (activation) function. Let K^d be some compact set in R^d and $C(K^d)$ the set of continuous functions on K^d . Let $f \in C(K^d)$. Then for all $\varepsilon > 0$, there exists N, real numbers v_i , b_i and R^d -vectors w_i such that, if we define

$$N(x) = \sum_{i=1}^{N} v_i \, \emptyset((w_i, x) + b_i)$$

then we have

$$\forall x \in K_d \, , |N(x) - f(x)| \leq \varepsilon$$

1.8 Activation Functions [16]

The importance of the activation function comes from its role of converting an input signal to an output signal in each node of each hidden layer. In the next layer, the input signal is the output signal of the previous layer. One asks whether it was possible to dispense the activation function, and what would happen if we omit it from the neural network. If we apply the neural network without any activation function, then the network would be simply a linear function with the following several limitations

a. A neural network without complexity cannot learn complex mappings from data.

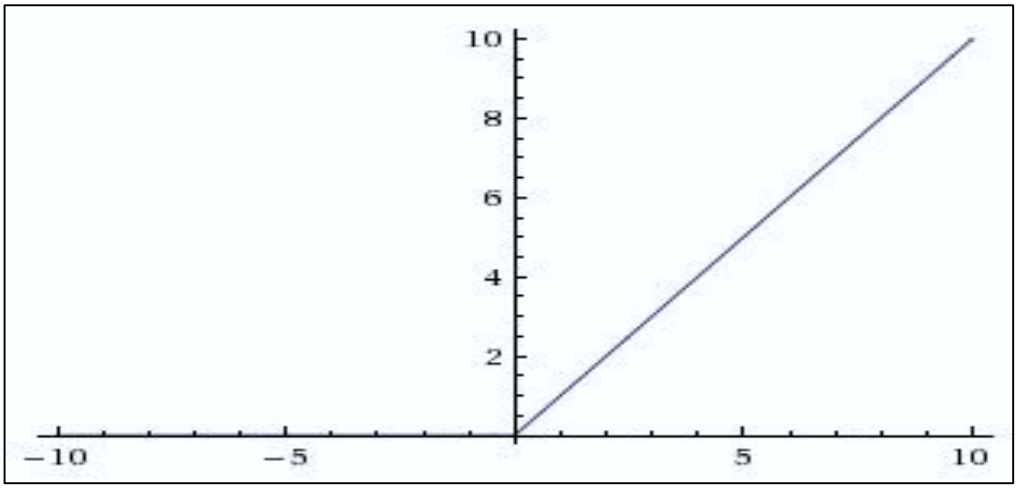
b. A linear regression neural network has a little power that doesn't perform well most of the time.

c. Special data such as images, audios, videos cannot be learned or modeled without activation function.

Those are the reasons for using nonlinear, complicated, multi high dimensional activation functions to model a neural network with a complicated architecture and to obtain knowledge from huge datasets.

1.9. Rectified Linear Units (ReLU) [26]

Rectified linear unit or ReLU is a relatively new activation function which has been shown to enable better training, in particular for networks with many hidden layers, referred to as deep networks [26]. It is nevertheless worthwhile to look into its effects on a network with only a single hidden layer. The function is described by $\mathcal{R}(\mathrm{x}) = \max\{0\,, \mathrm{x}\}$ i.e if $\mathrm{x} < 0$ $\mathcal{R}(\mathrm{x}) = 0\,,$ and if $\mathrm{x} \geq 0$ $\mathcal{R}(\mathrm{x}) = \mathrm{x}$



Figure(1.2) ReLu Activation function

This would look linear function, but ReLU is nonlinear. It is not bound activation function. The output range of ReLU is $[0, \infty)$. This means it can blow up the activation.

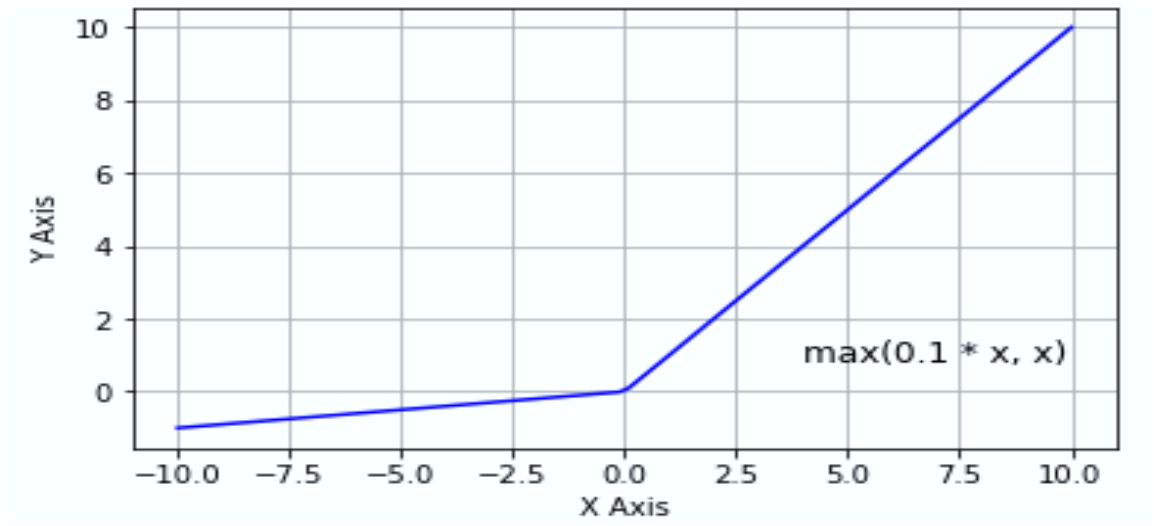
Summary of activation functions is given as follows. ReLU is not purely linear but consists of linear segments. Such functions are called as piecewise linear.

This ReLU function's positive part is an identity function, and negative part is set to zero.

1.9.1. Types of ReLU [26]

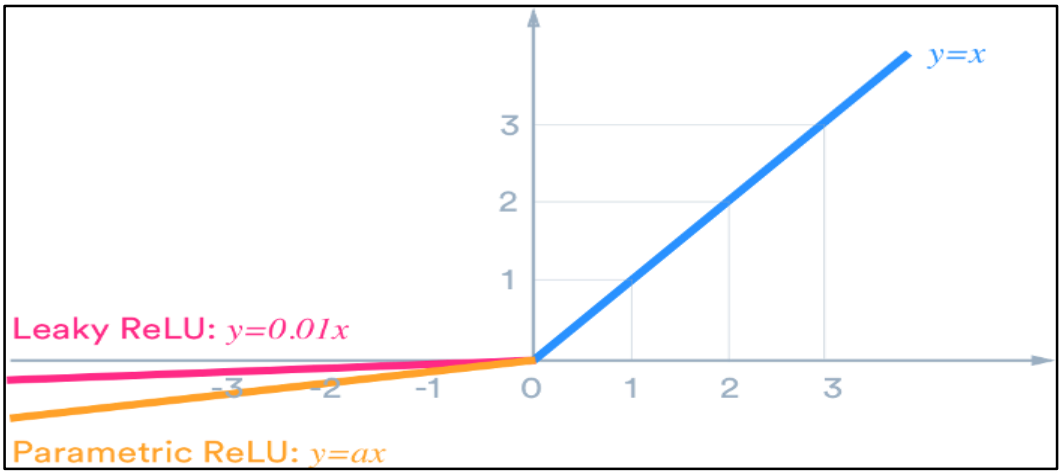
There are other types of ReLU.

• **Leaky ReLU (LReLU)** It assigns a relatively smaller and predefined slop to the negative part. LReLU is given by $R(x) = max(0.1 * x\ , x)$



Figure(1.3) Leaky ReLU Activation function

Parametric ReLU (PReLU) is a type of leaky ReLU that, instead of having a predetermined slope like 0.01, makes it a parameter for the neural network to figure out itself : $y = ax$ when $x < 0$.



Figure(1.4) Parametric ReLU Activation function

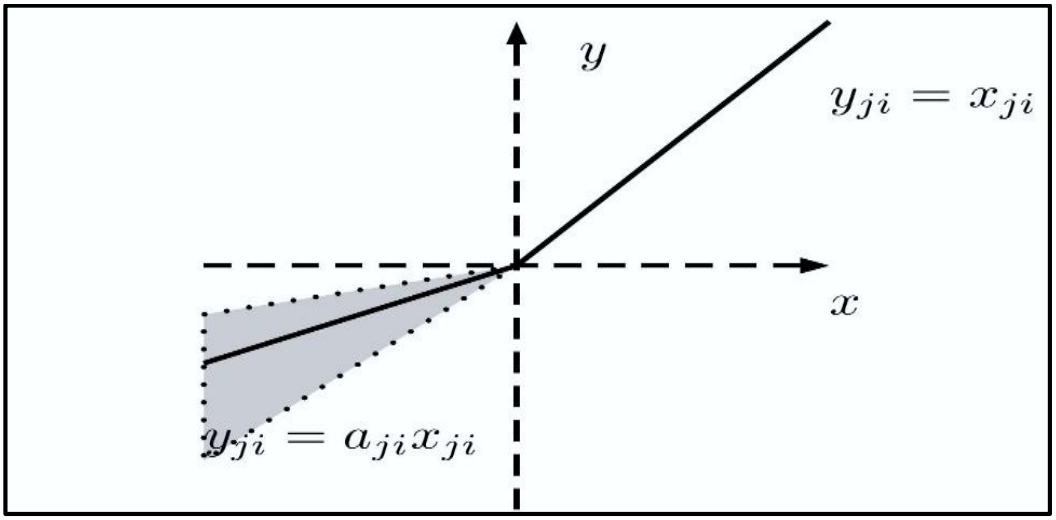
- **Randomized Leaky Rectified Linear** is the randomized version of leaky ReLU. The highlight of RReLU is that in training process,a_{ji} is a random number sampled from a uniform distribution $U(l,u)$. Formally, we have:

$$y_{ji} = x_{ji} \text{ if } x_{ji} \geq 0$$

$$y_{ji} = a_{ji}x_{ji} \text{ if } x_{ji} < 0$$

Where

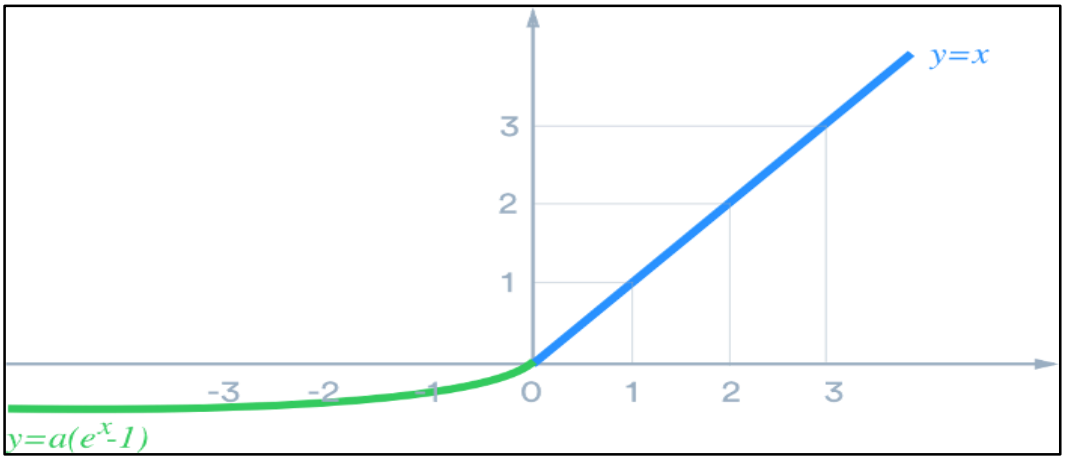
$$a_{ji} \sim U(l,u), l < u \text{ and } l,u \in [0,1)$$



Figure(1.5) Randomized Leaky Rectified

• **Exponential Linear Unit(ELU)** It exponentially reduces the slope from the predefined and fixed threshold to zero and is beneficial to speed up model learning. It has a small slope for negative values. Instead of a straight line, it uses a log curve like the following:

$$y = a(e^x - 1)$$



Figure(1.6) Exponential Linear Unit(ELU)

1.9.2. Advantages and Disadvantages [26]

According to different negative parts of activation functions, there are three categories: 1) Zero-type , 2) linear-type, 3) exponential-type. Among them, ReLu belongs to the first type, LReLU and PReLU belong to the second, and ELU belongs to the third type. ReLU gives us the benefit of the sparsity of the activation. Imagine a network with randomly initialised weights and almost 50% of the network yields 0 activations because of the characteristic of ReLU (output 0 for negative values of x). This means fewer neurons are firing (sparse activation) and the network is lighter. Because of a horizontal line in ReLU(for negative values of X) gradients can be fragile during training and can die. That means weights will not get adjusted and neurons will stop responding to variations in error or input. This is called Dead Neuron or dying ReLu problem. To fix this problem, one modification was introduced by Leaky ReLU. It introduces a small slope to keep the updates alive. One more limitation is that it should only be used within hidden layers of a neural network model.

CHAPTER TWO

Lp Simultanious Approximation by ReLU Neural Networks

2.1. Simultanuous Approximation in L_p

The first notes about simultaneous approximation was done by Dunham in [27]. He generated the classical Chebyshev approximation by approximating two continuous functions f^+ and f^-, with $f^+(x) \leq f^-(x)$, for all $x \in [a,b]$, simultaneously. He also proved that his simultaneous approximation is equivalent to the classical one function Chebyshev approximation when $f^+ = f^-$.

For more specification, Diaz and Mclaughlin [28], [15] proved that the above problem is equivalent to the problem of approximating $\frac{1}{2}|f^+ + f^-|$. Also approximating two appropriate functions simultaneously is equivalent to approximating one function by elements of a certain set $\mathbb{S}$. Moreover, they defined as follow

Definition 2.1.1. [28]

The best simultaneous approximation δ to the set $\mathbb{S}$ is given by

$$\inf_{s\in\mathbb{S}} \sup_{f\in\mathcal{F}} \|f - \mathrm{s}\| = \sup_{f\in\mathcal{F}} \|f - \mathcal{S}\|,$$

where $\mathcal{F}$ is a set of uniformly bounded functions on $[a,b]$ and $\mathbb{S}$ is a set of functions on $[a,b]$. They proved, in the same paper, that $\mathcal{S}$ is equivalent to the best simultaneous approximation of two functions.

Definition 2.1.2 (Linear function)

Let V and W be real vector spaces (their dimensions can be different) ,and let T be a function with domain V and range in W (written $T : V \to W$) . We say T is a linear function if

(a) for all $x, y \in V$, $T(x + y) = T(x) + T(y)$ (T is additive) .

(b) For all $x \in V$, $r \in R$, $T(rx) = rT(x)$ (T is homogeneous)

Example 2.2.1

Let $V = W = R^1$. Define $T(x) = mx$. where m is a fixed real number. let x and y be in R^1 and calculate

$$T(x + y) = m(x + y) = mx + my$$

$$T(x) + T(y) = mx + my$$

Since $T(x + y) = T(x) + T(y)$, that T is additive and homogeneous since $T(rx) = m(rx) = (mr)x = r(mx) = rT(x)$

T is a linear function , but $F(x) = ax + b$, $b \neq 0$ is not a linear functin, where a and b are real number and $b \neq 0$, let x and y be in R^1 and calculate

$$F(x + y) = m(x + y) + b = mx + my + b$$

$$F(x) + F(y) = (mx + b) + (my + b) = mx + my + 2b$$

Since $b \neq 0$,$2b \neq b$ so $F(x + y) \neq F(x) + F(y)$ for all $x , y \in V$, F is not linear .

Definition 2.1.3 [29]

A sequence of functions $\{f_n\}_{n\geq 1}$ from L_p is said to be convergent to some $f \in L_p$ if and only if

$$\lim_{n\to\infty} \|f_n - f\|_p = 0$$

Definition 2.1.4 [29]

A sequence of functions $\{f_n\}_{n\geq 1}$ from L_p is said to be Cauchy Sequence if and only if

$$\lim_{n,m\to\infty} \|f_n - f_m\|_p = 0,$$

Definition 2.1.5 [29]

The space L_p is said to be complete if and only if every Cauchy sequence $\{f_n\}_{n\geq 1}$ from converges to a function that belongs to L_p.

2.2. The ReLU network model

Throughout this chapter, we define a new type of ReLU activation function, that we call, General Rectifier Linear Unit, in short, GReLU. The mathematical formula is given in the following definition , but first ,we need to define some primary definitions .

Definition 2.2.1

The GReLU function is given by

$$\sigma(x) = g^{+}(x) = max\big(0, g(x)\big) = \begin{Bmatrix} 0, & \mathrm{g}(x) \leq 0 \\ g(x), & \mathrm{g}(x) > 0 \end{Bmatrix},$$

Where $g: R \to R$ is any real function that is

(1) Linear

(2) Differentiable

(3) Bounded

(4) Non decreasing

Our generalization for the ReLU activation function serves the linearity of the function, but gives a wide freedom of choice. Moreover, one can still use another activation function in the same network implicitly, but with the great advantages of ReLU function.

GReLu is a general case of ReLU when $g(x) = x$. On the other hand, it is a general case of LReLu when $a = 0$.

2.3. Construction of FNN with ReLU Activation Function [29]

let $\aleph$ be the set of all neural network of the form

$$\mathcal{N} = \sum_{i=1}^{n} c_i \mathcal{R}(w_i x + \vartheta_i)$$

Where$\mathcal{R}(x)$is the GReLU activation function.

Now, we are ready to discuss the essential point in this chapter. Here is the definition of the best simultaneous approximation of the set L_p by elements of $\aleph$ under the norm $\|\cdot\|_p$

Definition 2.3.1

The simultaneous best approximation of a subset $\mathcal{F}$ of L_p is $N^* \in \aleph$ in the expression

$$\inf_{N\in\aleph}\left\{\sup_{f\in\mathcal{F}} \| f - N \|_p\right\} = \sup_{f\in\mathcal{F}}\|f - N^*\|_p$$

2.4 Existence Theorem

Let $f \in L_p\ [a, b]$, then there exists a simultaneous approximation to any subset $\mathcal{F}$ of L_p by a FNN of the form

$$\mathcal{N}_n(x) = \sum_{i=1}^{n} c_i \mathcal{R}(w_i x + \vartheta_i),$$

where $\mathcal{R}$ is the GReLU activation function on $[a, b]$ and the parameters c_i , w_i , and ϑ_i are chosen as follow

$$w_i = -2\frac{hn}{|b-a|}, h = \frac{b}{2n}$$

$$\vartheta_i = \frac{hn}{|b-a|}\left(2a + (2i-1)\frac{b-a}{n}\right),$$

$$c_0 = f(a) - \sum_{i=1}^{n} c_i \mathcal{R}(w_i a + \vartheta_i),$$

$$c_i = \frac{1}{2b}\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i} f\left(x - \frac{kh}{2} + ih\right),$$

We have

$$\|\mathcal{N}_n(x) - f(x)\|_p^p \leq \frac{3}{4} W_k\left(f, \frac{1}{n}\right)_p^p$$

Proof:

Since g is bounded, then $\forall \boldsymbol{x} \in [\boldsymbol{a}, \boldsymbol{b}]$there exists some M > $\mathbf{0}$, s.t .

$$\sup_{x\in[a,b]} |\mathcal{R}(x)| = M$$

Let the partition $a < x_1 < x_2 < \cdots < x_n = b$, such that for all 1 ≤ $i \leq n$

and let $x_i = a + i\frac{b-a}{n}$, choosing

$$c_0 = f(a) - \sum_{i=1}^{n} c_i \mathcal{R}(w_i a + \vartheta_i),$$

implies

$$f(a) = \mathcal{N}_n(a) .$$

For all $x \in [a, b]$, there is $j \in$ N, 0 ≤ $j \leq n$, such that $x \in [x_{j-1}, x_j]$,

and that

$$\mathcal{N}_n(x) = f(a) + \sum_{i=1}^{n}\frac{1}{2b}\sum_{l=0}^{k}\binom{k}{l}(-1)^{k-l} f\left(x - \frac{kh}{2} + lh\right)$$

$[\mathcal{R}(w_i x + \vartheta_i) - \mathcal{R}(w_i a + \vartheta_i)]$

$$= f(a) + \sum_{\mathrm{i}=1}^{\mathrm{j}-1} \frac{1}{2\mathrm{b}} \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} \mathrm{f}\left(x - \frac{\mathrm{kh}}{2} + l\mathrm{h}\right) [\mathcal{R}(w_i x + \vartheta_i)$$

$$- \mathcal{R}(w_i a + \vartheta_i)]$$

$$+ \frac{1}{2b} \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} f\left(x_j - \frac{kh}{2} + lh\right) [\mathcal{R}(w_i x + \vartheta_i)$$

$$- \mathcal{R}(w_i a + \vartheta_i)]$$

$$+ \sum_{i=j+1}^{n} \frac{1}{2b} \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} f\left(x - \frac{kh}{2} + lh\right) [\mathcal{R}(w_i x + \vartheta_i)$$

$$- \mathcal{R}(w_i a + \vartheta_i)]$$

$$= f(a) + S_1 + S_2 + S_3$$

To estimate

$$|\mathcal{R}(w_i x + \vartheta_i) - \mathcal{R}(w_i a + \vartheta_i)|,$$

we have two cases

Case i. For $i > j$, we have, $x \leq x_j \leq x_{i-1}$, so by monotonicity of $\mathcal{R}$ and our choices of the parameters w_i , ϑ_i and x_i, we get

$$0 < \mathcal{R}(w_i x + \vartheta_i) - \mathcal{R}(w_i a + \vartheta_i)$$

$$\leq \mathcal{R}\left(w_i x_j + \vartheta_i\right) - \mathcal{R}(w_i a + \vartheta_i)$$

$$\leq \mathcal{R}(w_i x_{i-1} + \vartheta_i) - \mathcal{R}(w_i a + \vartheta_i)$$

$$= \mathcal{R}(h) - \mathcal{R}(2hi - h)$$

$$\leq g(h) - g(2hi - h)$$

$$\leq g(2h - 2hi)$$

$$\leq 2(1 - \mathrm{i})g(h)$$

$$\leq \frac{2(1 - \mathrm{i})}{2n} g(h)$$

$$\leq \frac{(1 - \mathrm{i})}{n} M$$

Case ii. For $i < j$, we have, $x_i \leq x_{j-1} \leq x$, then

$$\mathcal{R}(w_i a + \vartheta_i) - \mathcal{R}(w_i x + \vartheta_i)$$

$$\leq \mathcal{R}(w_i \mathrm{a} + \vartheta_i) - \mathcal{R}(w_i x_{i-1} + \vartheta_i)$$

$$= \mathcal{R}(-h) - \mathcal{R}(h)$$

$$\leq g(-h) - g(h)$$

$$\leq -2g(h)$$

$$\leq -\frac{2}{2n} g(b)$$

$$\leq -\frac{1}{n} M$$

For the two cases, we conclude that

$$|\mathcal{R}(w_i x + \vartheta_i) - \mathcal{R}(w_i a + \vartheta_i)| \leq h = \frac{b}{2n}$$

Now, we are ready to estimate S_1, S_2 and S_3

$$|S_1| \leq \frac{1}{2b} \sum_{i=1}^{j-1} \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} \left| f\left(x - \frac{kh}{2} + lh\right) \right| |\mathcal{R}(w_i x + \vartheta_i)$$

$$- \mathcal{R}(w_i a + \vartheta_i)|$$

$$\leq \frac{1}{4}\Delta_h^k(f,x,[a,b])$$

$$|S_2| \leq \frac{1}{2b}\sum_{l=0}^{k}\binom{k}{l}(-1)^{k-l}\left|f\left(x_j-\frac{kh}{2}+lh\right)\right||\mathcal{R}(w_i x+\vartheta_i)$$

$$-\mathcal{R}(w_i a+\vartheta_i)|$$

$$\leq \frac{1}{4}\Delta_h^k(f,x,[a,b])$$

$$|S_3| \leq \frac{1}{2b}\sum_{i=j+1}^{n}\sum_{l=0}^{k}\binom{k}{l}(-1)^{k-l}\left|f\left(x-\frac{kh}{2}+lh\right)\right||\mathcal{R}(w_i x+\vartheta_i)$$

$$-\mathcal{R}(w_i a+\vartheta_i)|$$

$$\leq \frac{1}{4}\Delta_h^k(f,x,[a,b])$$

Finally, let ∈ L_p , then we get from above

$$\|\mathcal{N}_n(x)-f(x)\|_p^p \leq \int_a^b |\mathcal{N}_n(x)-f(x)|^p\,dx$$

$$\leq \frac{1}{2b}\sum_{i=1}^{n}[\mathcal{R}(w_i x+\vartheta_i)-\mathcal{R}(w_i a+\vartheta_i)]$$

So that

$$\int_a^b \sum_{l=1}^{k}\binom{k}{l}(-1)^{k-l}\left|f\left(x-\frac{kh}{2}+lh\right)\right|^p dx \leq \frac{3}{4}\omega_k\left(f,\frac{1}{n}\right)_p^p$$

2.5. Uniqueness Theorem

The simultaneous best approximation $N^* \in \aleph$ of a subset $\mathcal{F}$ of L_p is unique.

Proof

To prove that $N^* \in \aleph$ is unique, suppose that $N_1, N_2 \in \aleph$ be two simultaneous approximations to F, then by Definition 2.1.3.

$$\lim_{n\to\infty} \|N_1 - f\|_p = 0$$

and

$$\lim_{n\to\infty} \|N_2 - f\|_p = 0$$

So by condition [ii] Definition 1.2.3, there exists k ≥ 1,

$$\|N_1 - N_2\|_p \leq k\big(\|N_1 - f\|_p + \|N_2 - f\|_p\big)$$

By taking limits to both sides as n tends to infinity, then by Definition 2.1.3, we get that

$$\lim_{n\to\infty} \|N_1 - N_2\|_p = 0$$

So $N_1 = N_2$, and the best simultaneous approximation to F out of $\aleph$ is unique.

CONCLUSION

The neural networks are considered to be universal approximators. We know that we can use the neural networks for the approximation of functions for many types of activation functions. Here, we treat neural networks with simple and effiecient activation function called Rectified Linear Units (ReLU). Generalized ReLU is prove here to define anew formula of neural network that is appropriate for function approximation .

The main objective of this paper is to introduce a type of neural network and we use it to approximate functions and estimate the general approximation error. We will get the optimal approximation if we have a basis independent of the target function .

We prove theorems of existence and uniqueness by estimating the degree of approximation depending on the modulus of smoothness of the functions envolved.

References

[1] 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, vol. 114, no. 5, pp. 953–956, 1957.

[2] G. Cybenko, “Approximation by Superpositions of A Sigmoidal Function,” Math. Control, Signals Systems., vol. 2, no. 4, pp. 303–314, 1989.

[3] G. Cybenko, “Continuous Valued Neural Networks: Approximation Theoretic Results,” in Computer Science and Statistics: proceedings of the 20th Symposium on the Interface, pp. 174–183, 1988.

[4] W. H. Hornik K., Stinchcombe M., “Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforword Networks.,” Neural Networks, vol. 3, no. 5, pp. 551–560, 1990.

[5] L. K, Jones “ A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training,” The annals of Statistics, vol. 30, no. 1, pp. 608–613, 1992.

[6] M. J. Donahue, C. Darken, L. Gurvits, and E. Sontag, “Rates of Convex Approximation in Non-Hilbert Spaces,” Constructive Approximation, vol. 13, no. 2, pp. 187–220, 1997.

[7] T. Chen, H, Chen, R, Liu, “A Constructive Proof and an Extension of Cybenko’s Approximation Theorem,” Computing Science and Statistics, New York, Springer-Verlag, pp. 163–168, 1992.

[8] T. Chen and H. Chen, “Approximation Capability to Functions of Several Variables , Nonlinear Functionals and Operators by Radial

Basis Function Neural Networks," IEEE Transactions on Neural Networks, vol. 6, no. 4, pp. 904–910, 1995.

[9] S. Suzuki, "Constructive Function Approximation by Three Layer Artificial Neural Networks," Neural Networks, vol. 11, no. 6, pp. 1049–1058, 1998.

[10] A. Dingankar, D. Phatak, and B. County, "Simultaneous Approximation with Neural Networks,"IEEE , vol. 4, pp.232-237, 2000.

[11] C. Ding, F. Cao, and Z. Xu, "The Essential Approximation Order for Neural Networks with Trigonometric Hidden Layer Units," Springer-Verlag Berlin Heidelberg, vol. 3971, pp. 72–79, 2006.

[12] E. S. Bhaya, and H. A. Almurieb "Neural Network Trigonometric Approximation," Journal of Babylon University/Pure and Applied Sciences, vol. 24, no. 9, pp. 2395–2399, 2016.

[13] E. Zakon "Mathematical Analysis II'' University of Windsor , 1975.

[14] H. A. Almurieb, "Simultaneous Approximation of Order m by Artificial Neural Network," Journal of Babylon University/Pure and Applied Sciences, vol. 56, no. 4,pp. 38-44, 2017.

[15] J. B. Diaz and H. W. Mclaughlin, "On Simultaneous Chebyshev Approximation and Chebyshev Approximation with an Additive Weight," Journal of Approximation Theory, vol. 6, no. 1, pp. 68–71, 1972.

[16] E. S. Bhaya, "On the Constrained and Unconstrained Approximation", PD Thesis, University of Baghdad, 2003..

[17] R. Walter, "Real and Complex Analysis (3rd ed.)", New York ,(1987) .

[18] I . Singer “Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces” New york ,vol 1,pp. L-3A, 1970.

[19] K. Conrad, “Lp-Spaces for 0 < p < 1,” p.p 1–15, 2003.

[20] Riesz,Frigyes , "Untersuchungen über Systeme Integrierbarer Funktionen" , Mathematische Annalen,vol, 69 ,no (4): pp 449–497,1910.

[21] N. L. Carothers “A Short Course on Approximation Theory” Math 682 Summer ,Citeseer , 1998 .

[22] D. Jacksons, “The General Theory of Approximation by Polynomials and Trigonometric Sums,” Bulletin of the American Mathematical Society, vol. 27, pp. 415–431, 1921.

[23] S. Bernstein, “Sur L’ordre de la Meilleure Approximation des Fonctions Continues par des Polynômes de Degré Donné,” Hayez, imprimeur des academies royales, vol. 4, 1912.

[24] K . Gurney “An Introduction to Neural Networks”, CRC Press, 2018.

[25] K. Ciesielski, J. P. Sacha, K. J. Cios, “Synthesis of Feedforward Networks in Supermom Error Bound”IEEE Transactions on neural networks,vol. 11, no. 6, pp . 1213-1227, 2000 .

[26] K Gayatri, and D Vora. "Activation functions and training algorithms for deep neural network." *UGC approved journal, International Journal of Computer Engineering In Research trends* vol. 5 , no. 4, pp: 98-104, (2018).

[27] C. B. Dunham, “Simultaneous Chebyshev Approximation of Functions on an Interval,” Proceedings of the American Mathematical Society, vol. 18, no. 3, pp 472-477, 1967.

[28] J. B. Diaz and H. W. Mclaughlin, “Simultaneous Approximation of a Set of Bounded Real Functions,” Mathematics of Computation, vol. 23, no. 107, pp. 583–594, 1969.

[29] H. A. Almurieb ,& E. S. Bhaya, ”The Degree of Best Neural Networks Approximation With Application ” PhD Thesis , University of Babylon, 2020 .

المستخلص

يعد تقريب الدوال باستخدام الشبكات العصبية اتجاهًا مثيرًا للاهتمام في نظرية التقريب. كما تم إثباته سابقًا ، الشبكات العصبية عبارة عن تقديرات تقريبية شاملة للدوال المستمرة ، لذلك من المهم إثبات شمولية الدوال من فضاءات أخرى أوسع .

من الضروري اختيار فضاء الدوال التي تقربها الشبكات العصبية بشكل جيد. هذا ويعد فضاء الدوال الليبيكية L_p من الاختيارات الرائعة للدراسة. من المثير للاهتمام أن تأخذ القيمة $0 < p < 1$.

من ناحية أخرى ، فإن الصيغ الأساسية للشبكات العصبية المقربة ليست أقل أهمية. نعلم أنه يمكننا استخدام الشبكات العصبية لتقريب الدوال لأنواع عديدة من دوال التنشيط. هنا ، نتعامل فقط مع الشبكات العصبية بوظيفة تنشيط بسيطة وفعالة تسمى الوحدات الخطية المصححة والتي تختصر بـ ReLU . تم اثباته هنا أن ReLU المعممة تحدد صيغة جديدة للشبكة العصبية مناسبة لتقريب الوظيفة.

الهدف الرئيسي من هذا البحث هو تقديم نوع من الشبكات العصبية ونستخدمها في تقريب الدوال وتقدير خطأ التقريب العام. سنحصل على التقريب الأمثل إذا كان لدينا أساس مستقل عن الدالة المستهدفة. نثبت نظريات الوجود والوحدانية من خلال تقدير درجة التقريب اعتمادًا على معامل نعومة الدوال الليبيكية .

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية التربية للعلوم الصرفة
قسم الرياضيات

تقريب الدوال بأستخدام شبكات ReLU العصبية المعممة

بحث

مقدم الى مجلس التربية للعلوم الصرفة ـ جامعة بابل

كجزء من متطلبات نيل درجة الدبلوم العالي

تربية / الرياضيات

من قبل

أمجد حمزه ساجت كعيد

بإشراف

م.د. حوراء عباس فاضل المرعب

١٤٤٢ هـ