Approximation proof of DNN with 1 width and N depth

rajeshk.mishra // rajeshpremlatamishra@gmail.com

August 2024

1 Introduction

Neural Network (NN) has ability to approximate any continuous functions, There exist various universal approximation theorem on what kind of function NN can express,depending on depth of NN,width of NN and class of activation function.This paper [Elb+21] present the approximation properties of NN in most comprehensive way by extending the idea of [Yar17],[Sch20] and [Tel15]. Our method of proof is different in two ways that, 1) we don't assume any form/class on the function, instead we show for error between sample points chosen from function can be reduced to any small value. 2) We use the interpolation property of sample points to get polynomial then that polynomial is approximated using NN.

2 Proof Sketch

Lets define the neural network with n hidden layers and each hidden layer has an activation unit $\sigma.\sigma$ is sub-additive and bounded.

Short introduction: $f : R \to R$

$$f(x+y) \leq f(x) + f(y), \quad \forall x, y \in R \tag{1}$$
$$f(cx) \leq cf(x), \quad x \quad \text{and} \quad c \in R \tag{2}$$

Defining the network as follows,

$$f_0 = x \tag{3}$$
$$f_1 = \sigma(w_1 x + b_1) \tag{4}$$
$$f_{n+1} = \sigma(w_n f_n + b_n) \tag{5}$$
$$\tag{6}$$

Let $F(x)$ be actual function to be approximated using this NN, We know F only through a set of sample points i.e. $(x_i, F(x_i))$ and $i = 1 \dots K$.
To prove that NN f_n uniformly converges to $F(x)$. We need to show n, w_i and

b_i, can be constructed (or they exist) in such a way that norm of the error between $f_n(x)$ and $F(x)$ can be made less than any positive number $\epsilon > 0$, $\epsilon > 0$ does not depend on any x_i

so we need to prove for any $\epsilon > 0$

$$|\sum_{i=1}^{K} F(x_i) - f_n(x_i)| < \epsilon$$

$$\text{there exists } n, w_i, b_i, \text{ for } i = 1, \ldots, K$$

First, We show that for any x neural network can be reduced to an expression less than some polynomial.

$$\begin{aligned} f_1 &= \sigma(w_1 x + b_1) \leq w_1\sigma(x) + \sigma(b_1) && (7)\\ f_2 &= \sigma(w_2 f_1 + b_2) \leq \sigma(w_2[w_1\sigma(x_i) + \sigma(b_1)] + b_2) && (8)\\ &\leq w_2 w_1 \sigma^2(x) + w_2\sigma^2(b_1) + \sigma(b_2) && (9)\\ f_n &= \sigma(w_n f_{n-1} + b_n) && (10)\\ &\leq w_n \ldots w_1 \sigma^n(x) + \sum_{k=1}^{n-1} \sigma^{n-k+1}(b_k) \prod_{j=k+1}^{n} w_j + \sigma(b_n) && (11) \end{aligned}$$

Explanation: in any i bracket we have $\sigma(b_i)$, σ will apply on this function for all $n - i$ outside σ so in total it will be applied by $n - k + 1$ times on b_i,in any i backet, $\sigma(b_i)$ will be multiplied by w_{i+1} to w_n because w_i is in the same bracket as $\sigma(b_i)$ so it won't be multiplied, this expression work up to $n - 1$ and for last b_n, we will have single term $\sigma(b_n)$. $\sigma^i(x)$ means function applied i times. Next, We simplifiy this expression using bounded property of $sigma$, $|\sigma(x)| \leq M$. hence, $|\sigma^{n-k+1)}(b_k)| \leq M^{n-k+1}$ and $|\sigma^n(x)| \leq M^n$.

$$f_n = \sigma(w_n f_{n-1} + b_n) \quad \leq w_n \ldots w_1 M^n + \sum_{k=1}^{n-1} M^{n-k+1} \prod_{j=k+1}^{n} w_j + M \qquad (12)$$

Defining, $w_k = a_k w$. We have,

$$f_n = \sigma(w_n f_{n-1} + b_n) \quad \leq a_n \ldots a_1 M^n w^n + \sum_{k=1}^{n-1} M^{n-k+1} \prod_{j=k+1}^{n} a_j w^{n-k} + M \qquad (13)$$

Lets define $c_n = a_n \ldots a_1 M^n$ and $c_{n-k} = M^{n-k+1} \prod_{j=k+1}^{n} a_j$, we get

$$f_n = \sigma(w_n f_{n-1} + b_n) \quad \leq c_n w^n + \sum_{k=1}^{n-1} c_{n-k} w^{n-k} + M \qquad (14)$$

A polynomial of n degree can pass though exactly n arbitrary possible using Legendre interpolation method . In our case.From this fact it can also be shown

that A polynomial of n degree can pass though exactly n point as close as possible.(Point 2)

$$L_i(w) = \prod_{\substack{j=0 \\ j\neq i}}^{n-1} \frac{w - x_j}{x_i - x_j} \qquad \text{and } P(w) = \sum_{i=0}^{n-1} y_i L_i(w) \tag{15}$$

we need to choose P^{new} such that $|P^{new}(w) - P(w)| < \epsilon$.For any number$|\Delta| \leq \epsilon$ if $P^{new}(w) = P(w) + \Delta$. For Legendre interpolation polynomial, we have

$$|\sum_{i=1}^{K} F(x_i) - P(w)| = 0 \quad \exists \quad n \quad w_i \quad b_i$$

$$|\sum_{i=1}^{K} F(x_i) - P^{new}(w)| < \epsilon$$

$$\begin{aligned} P(w) &= a_n w^n + \ldots + a_0 \\ P^{new}(w) &= a_n w^n + \ldots + a_0 + \Delta \end{aligned} \tag{16}$$

c_i can be chosen so that $|P^{new} - C_n w^n + \ldots + M| <= \Delta$ because they are same degree polynomial ((Point 3).Also note that error does not depend on any x_i it depends only the choice of c_i which by construction implies choice of w_i and b_i. this proves that NN can uniformly approximate any function to any $\epsilon > 0$ and ϵ does not depend of x_i

$$|\sum_{i=1}^{K} F(x_i) - f_n(x_i)| < \epsilon \quad \exists \quad n \quad w_i \quad b_i$$

2.1 Comments

Point 1 : Only eq 1 is actually definition of sub-additive function but eq 2 can be proved based on eq1,which is the key property.
Point 2 : This is intuitively clear to me, by changing the coefficient slightly, visually we can see it cab be done, I did not included the proof for this, but I can prove this for when function is bounded,
Point 3 : In this point, I actually tried to construct c_n but could not, so I used point 2, to say that c_n exists because only existence was needed

References

[Tel15] Matus Telgarsky. "Representation benefits of deep feedforward networks". In: *arXiv preprint arXiv:1509.08101* (2015).

[Yar17] Dmitry Yarotsky. "Error bounds for approximations with deep ReLU networks". In: *Neural networks* 94 (2017), pp. 103–114.

[Sch20] Johannes Schmidt-Hieber. "Nonparametric regression using deep neural networks with ReLU activation function". In: *Project euclid* (2020).

[Elb+21] Dennis Elbrächter et al. "Deep neural network approximation theory". In: *IEEE Transactions on Information Theory* 67.5 (2021), pp. 2581–2623.