Neural Networks Can Approximate Continuous Functions

Abstract

A clean Tex version of the proof seen during the lectures that neural networks with ReLU activation can approximate continuous functions.

1 Statement

Definition. *$\mathcal{C}([a,b])$ is the set of real continuous functions on $[a,b]$.*

Definition. *$\mathcal{P}(n,l)$ is the set of rectangle 1D-1D perceptrons with l hidden layers and n neurons in each hidden layer.*

Theorem (Particular case of the universal approximation theorem[1])**.**

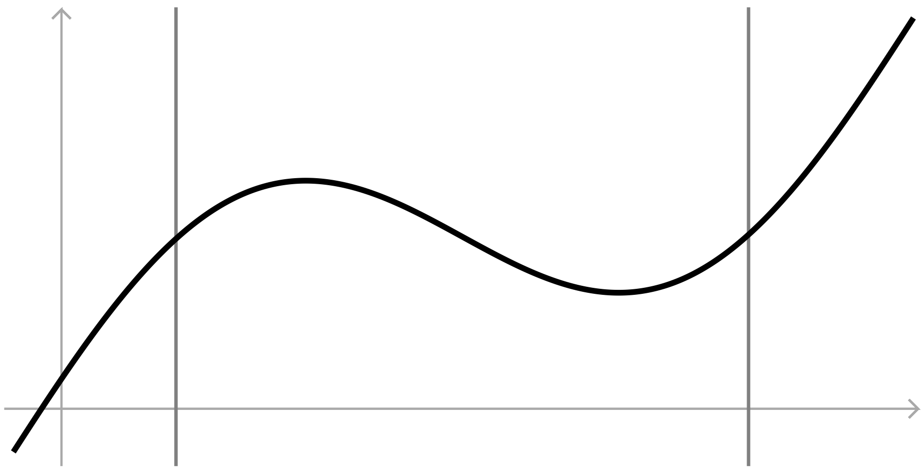
$$\forall f \in \mathcal{C}([a,b]),\ \forall \varepsilon > 0,\ \exists N \in \mathbb{N} \text{ and } p \in \mathcal{P}(N,1) \text{ s.t. } \|p - f\|_\infty < \varepsilon$$

Where $\|g\|_\infty$ is infinity norm of g on $[a,b]$; i.e. $\|g\|_\infty = \max_{x\in[a,b]}(|g(x)|)$.

2 Idea

The idea of the proof is to split the interval $[a,b]$ in sub intervals such that f has variation less than ε in the sub-intervals. Then, choose the weights of the neural network to interpolate linearly f et the beginning and end of the sub-intervals.

[1] See `https://en.wikipedia.org/wiki/Universal_approximation_theorem` for details.



3 Proof

Let $\varepsilon > 0$ and $f \in \mathcal{C}([a,b])$ with $a, b \in \mathbb{R}$, we want $N \in \mathbb{N}$ and $p \in \mathcal{P}(N,1)$ such that $\|p - f\|_\infty < \varepsilon$.

ReLU Network If $p \in \mathcal{P}(N,1)$, then

$$p(x) = \xi + \sum_{k=0}^{N-1} \gamma_k(\alpha_k x + \beta_k)_+$$

where $(x)_+$ is ReLU(x).

All we need to do is to find the right coefficients $\xi, \alpha_k, \beta_k, \gamma_k (0 \leq k < N)$ such that $|f(x) - p(x)| < \varepsilon \quad (\forall x \in [a,b])$.

Uniform Continuity Since f is continuous on a compact set, so f is uniformly continuous. Thus, $\forall x_1, x_2 \in [a,b]$, $\exists \delta > 0$ such that

$$|x_1 - x_2| < \delta \implies |f(x_1) - f(x_2)| < \varepsilon.$$

Sub-intervals Let $c_0 = a$ and $c_{i+1} = c_i + \delta$. Let $N \in \mathbb{N}$ be such that $c_N \geq b$ and redefine $c_N = b$.

Coefficients Take:

- $\alpha_k = 1 \quad (0 \leq k < N)$
- $\beta_k = -c_k \quad (0 \leq k < N)$
- $\tilde{\gamma}_k = \frac{f(c_{k+1}) - f(c_k)}{c_{k+1} - c_k} \quad (0 \leq k < N)$

- $\gamma_0 = \tilde{\gamma 0}$
- $\gamma_k = \tilde{\gamma_{k+1}} - \tilde{\gamma_k} \quad (0 < k < N)$
- $\xi = f(a)$

Thus,

$$p(x) = \xi + \sum_{k=0}^{N-1} \gamma_k (\alpha_k x + \beta_k)_+$$

becomes

$$p(x) = f(a) + \sum_{k=0}^{N-1} \gamma_k (x - c_k)_+.$$

Intermediate Result

Claim ('The network interpolates linearly from c_n to c_{n+1}')**.** *If $x \in [c_n, c_{n+1}]$, then $p(x) = f(c_n) + \tilde{\gamma_n}(x - c_n)$.*

Proof. **Case $k = 0$:**
Let $x \in [c_0, c_1]$, then:

$$p(x) = f(a) + \sum_{k=0}^{N-1} \gamma_k (x - c_k)_+$$

since $x \leq c_1$, $(x - c_k)_+ = 0 \quad \forall k > 0$, and $(x - c_k)_+ = x - c_k$:

$$p(x) = f(a) + \gamma_0 (x - c_0)$$

as $\tilde{\gamma_0} = \gamma_0$ and $a = c_0$, we finally have $p(x) = f(a) + \tilde{\gamma_0}(x - c_0)$, as expected.
Recursion:

- Suppose that if $x \in [c_n, c_{n+1}]$, then $p(x) = f(c_n) + \tilde{\gamma_n}(x - c_n)$.
- We want to show that if $x \in [c_{n+1}, c_{n+2}]$, then $p(x) = f(c_{n+1}) + \tilde{\gamma_{n+1}}(x - c_{n+1})$.

Let $x \in [c_{n+1}, c_{n+2}]$, let's calculate $p(x)$:

$$p(x) = f(a) + \sum_{k=0}^{N-1} \gamma_k (x - c_k)_+$$

if $k > n+1$, then $c_k \geq x$, so $(x - c_k)_+ = 0$; similarly, if $k \leq n+1$, then $c_k < x$, so $(x - c_k)_+ = x - c_k$ thus:

$$p(x) = f(a) + \sum_{k=0}^{n+1} \gamma_k (x - c_k)$$

Now, let's split $(x - c_k)$ to $(x - c_{n+1}) + (c_{n+1} - c_k)$:

$$p(x) = f(a) + \sum_{k=0}^{n+1} \gamma_k (x - c_{n+1}) + \sum_{k=0}^{n+1} \gamma_k (c_{n+1} - c_k)$$

We can add again $(c_{n+1} - c_k)_+$ for $n + 1 < k < N$ (this is just adding zeros) to the second sum to make $p(c_{n+1})$ appear:

$$p(x) = f(a) + \sum_{k=0}^{N-1} \gamma_k (c_{n+1} - c_k) + \sum_{k=0}^{n+1} \gamma_k (x - c_{n+1})$$

$$= f(c_{n+1}) + \sum_{k=0}^{n+1} \gamma_k (x - c_{n+1})$$

Now, $\gamma_k = \tilde{\gamma_k} - \tilde{\gamma_{k-1}} \quad 0 < k < N$ and $\tilde{\gamma_0} = \gamma_0$, so:

$$p(x) = f(c_{n+1}) + \gamma_0 (x - c_{n+1}) + \sum_{k=1}^{n+1} (\tilde{\gamma_k} - \tilde{\gamma_{k-1}})(x - c_{n+1})$$

This is a telescoping series, after simplification, we have:

$$p(x) = f(c_{n+1}) + \tilde{\gamma_{n+1}} (x - c_{n+1})$$

As expected. $\square$

Therefore, we have that $\forall\, 0 \leq n < N$, if $x \in [c_n, c_{n+1}]$, then $p(x) = f(c_n) + \tilde{\gamma_n}(x - c_n)$.

Bounds for $f(x)$ and $p(x)$ on $[c_n, c_{n+1})$ Take $x \in [c_n, c_{n+1})$, then we have: $|c_n - x| < \delta$, so $|f(x) - f(c_n)| < \varepsilon$.

WLOG[2], take $f(c_n) \leq f(c_{n+1})$: - $f(c_n) \leq p(x) \leq f(c_{n+1})$ and $|f(c_{n+1}) - f(c_n)| \leq \varepsilon$ so $|p(x) - f(c_n)| \leq \varepsilon$.

Bound for $f(x) - p(x)$ on $[a, b]$

$$|f(x) - p(x)| = |f(x) - f(c_n) + f(c_n) - p(x)|$$
$$\leq |f(x) - f(c_n)| + |p(x) - f(c_n)|$$
$$< \varepsilon + \varepsilon = 2\varepsilon$$

This is true $\forall\, 0 \leq n < N,\ \forall\, x \in [c_n, c_{n+1})$, it is true $\forall\, x \in [a, b)$.

Moreover, $p(b) = f(b)$ (from the property above), so $|f(x) - p(x)| < 2\varepsilon$ on all of $[a, b]$.

Thus, we finally have $\|f - p\|_\infty < 2\varepsilon$.

Conclusion Now, taking $\tilde{\varepsilon} = \frac{1}{2}\varepsilon$, we get $\|f - p\|_\infty < \tilde{\varepsilon}$ with the same reasoning.

Hence, f can be ϵ approximated by a single hidden layer perceptron.

[2]Without Loss Of Generalities