CSE 543: Deep Learning	Spring 2022
Lecture 2	
Prof. Simon Du	*Scribe: Yuhao Wan*

1 1D Approximation

Theorem 1. *Let $g : [0, 1] \to \mathbb{R}$, and ρ-Lipschitz. For any $\epsilon > 0$, 2-layer neural network f with $\lceil \frac{\rho}{\epsilon} \rceil$ nodes, threshold activation $\sigma(z) : z \to \mathbf{1}\{z \geq 0\}$ such that $\sup_{x \in [0,1]} |f(x) - g(x)| \leq \epsilon$.*

Proof. Proof idea: divide the $[0, 1]$ interval into equal length of $\frac{\epsilon}{\rho}$ sub-intervals. Then construct a piece wise constant function f on each interval to approximate our target function g, which can be represented by a 2-layer neural network with a threshold activation function.

Define $m := \lceil \frac{\rho}{\epsilon} \rceil$, and $x_i := \frac{(i-1)\epsilon}{\rho}$ for $i \in \{0, \ldots, m-1\}$, and $a_0 = g(0), \qquad a_i = g(x_i) - g(x_{i-1})$, and lastly define our neural network $f(x) := \sum_{i=0}^{m-1} a_i \mathbf{1}[x - x_i \geq 0]$. This is saying that if $x < x_1$, all except x_0 is 0. So $f(x) = a_0 = g(x_0)$. If $x_1 \leq x < x_2, f(x) = g(x_1)$. Thus, on each sub-interval, this constant function will equal to part of the target function applies to the left of the interval. Then for any $x \in [0, 1]$, letting x_i be the largest index so that $x_i \leq x$,

$$\begin{aligned}
|g(x) - f(x)| &= |g(x) - f(x_i)| \text{ where } x_i \leq x \text{ and closest to } x \text{ on the left} \\
&\leq |g(x) - g(x_i)| + |g(x_i) - f(x_i)| \text{ by triangle inequality} \\
&\leq \rho |x - x_i| \text{ by Lipschitzness of } g \\
&\leq \rho \cdot \frac{\epsilon}{\rho} \\
&= \epsilon.
\end{aligned}$$

Note: the length of the sub-interval depends on Lipschitzness of the target function. If target function is smooth, then we don't need many sub-intervals, and vice versa.

□

2 Multivariate Approximation

Theorem 2. *Let g be a continuous function that satisfies $||x - x'||_\infty \leq \delta \Rightarrow |g(x) - g(x')| \leq \epsilon$ (Lipschitzness). Then there exists a 3-layer ReLU neural network with $\mathcal{O}(\frac{1}{\delta^d})$ nodes that satisfy*

$$\int_{[0,1]^d} |f(x) - g(x)| dx = ||f - g||_1 \leq \epsilon$$

Proof. Proof idea: 1) Use $h(x) = \sum_i \alpha_i \mathbf{1}_{R_i}(x)$ in the partition lemma. 2) Use a 2-layer neural network to approximate the threshold indicator function $x \mapsto \mathbf{1}_{R_i}(x)$, then find some linear combination to represent h.

Our goal is to show

$$||f - g||_1 \leq ||f - h||_1 + ||h - g||_1$$

where the second term on the right is bounded by ϵ by the partition lemma. Thus we want to show $||f - h||_1 \leq \epsilon$.

Formally, define $f(x) = \sum_{i=1}^{N} \alpha_i f_i(x)$ where the α_i's are the same as defined in $h(x)$ above. Then

$$\begin{aligned} ||f - g||_1 &= ||\sum_{i=1}^{N} \alpha_i(\mathbf{1}_{R_i} - f_i)||_1 \\ &\leq \sum_{i=1}^{N} |\alpha_i| ||\mathbf{1}_{R_i} - f_i||_1 \end{aligned}$$

We want $||\mathbf{1}_{R_i} - f_i|| \leq \frac{\epsilon}{\sum_{i=1}^{N} |\alpha_i|}$.

To construct f_i's, we define a region $R_i := [a_1 b_1] \times [a_2 b_2] \times \cdots [a_d b_d]$, which is a Cartesian product of 1-d intervals.

The idea is that we would use a bump function (through our smoother ReLU) to approximate the indicator function, which is non-smooth.

In the 1 dimensional case, given $\gamma > 0$, we define $g_{\gamma,j}(z) = \sigma(\frac{z-(a_j-\gamma)}{\gamma}) - \sigma(\frac{z-a_j}{\gamma}) - \sigma(\frac{z-b_j}{\gamma}) + \sigma(\frac{z-(b_j+\gamma)}{\gamma})$, where σ function is ReLU.

In general, where x is a d-dimensional vector, we define $g_\gamma(x) = \sigma(\sum_{j=1}^{d} g_{\gamma,j}(x^j) - (d-1))$. We can check by definition that

$$g_\gamma(x) = \begin{cases} 1 & \text{if } x_i \\ 0 & \text{if } x \notin [\alpha_1 - \gamma, b_1 + \gamma] \times \cdots [a_d - \gamma, b_d + \gamma] \end{cases}$$

Since $\gamma \to 0, g_\gamma \to \mathbf{1}_{R_i}$ by definition, so , $\exists \gamma$ s.t. $||g_\gamma - \mathbf{1}_{R_i}||_1 \leq \frac{\epsilon}{\sum_i |\alpha_i|}$

Therefore we define $f_i = g_\gamma, f := \sum_{i=1}^{N} \alpha_i f_i$. □