CSE 543: Deep Learning Spring 2022, Winter 2023

Lecture 2

Prof. Simon Du *Scribe: Yuhao Wan, Yifang Chen*

1 Classical approximations

1.1 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, there exists a threshold activation function $\sigma(z) : z \mapsto \mathbf{1}[z \geq 1]$ such that*

$$\sup_{x \in [0,1]} |f(x) - g(x)| \leq \epsilon$$

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.

Proof. Define $m := \lceil \frac{\rho}{\epsilon} \rceil$, and $x_i := \frac{(i-1)\epsilon}{\rho}$ for $i \in \{0, \ldots, m-1\}$, and $a_0 = g(0)$, $a_i = g(x_i) - g(x_{i-1})$, 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.

□

1.2 Multivariate Approximation

Theorem 2. *Let g be a continuous function that satisfies $||x - x'||_\infty \leq \delta \Rightarrow |g(x) - g(x')| \leq \epsilon$ for any pairs of input x, x' (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 Idea The proof idea is similar to the 1D special case. **Step 1:** we can partition the whole $[0,1]^d$ continuous space into sub-spaces R_i. So that the overall function $g(x)$ on $[0,1]^d$ can be represented by a linear combination of multiple activations functions $x \mapsto \mathbf{1}_{R_i}(x)$, that is $h(x) = \sum_i \alpha_i \mathbf{1}_{R_i}(x)$. **Step 2:** we show that such h can be approximated by a 2-layer NN, denoted as f.

Proof. For Step 1, we first construct a partition $\{R_i\}_{i=1}^N$ that for any two $x, x' \in R_i$, $\|x - x'\|_\infty \le \delta$. Then by Lemma 3, it is easy to see that there exists some corresponding $(\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$ that,

$$\|g - h\|_1 \le \epsilon, \quad \text{where } h = \sum_{i=1}^N \alpha_i \mathbf{1}_{R_i}$$

And therefore,

$$||f - g||_1 \le ||f - h||_1 + ||h - g||_1 \le ||f - h||_1 + \epsilon$$

Now we start the proof for Step 2 to bound $||f - h||_1$. 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 above. Then

$$||f - h||_1 = ||\sum_{i=1}^N \alpha_i(\mathbf{1}_{R_i} - f_i)||_1 \le \sum_{i=1}^N |\alpha_i| ||\mathbf{1}_{R_i} - f_i||_1$$

Now we want to construct f_i that such $||\mathbf{1}_{R_i} - f_i|| \le \frac{\epsilon}{\sum_{i=1}^N |\alpha_i|}$. Firstly, it is easy to see that when $\sum_{i=1}^N |\alpha_i| = 0$, we have $g(x_i) = 0$ for all i and $|g(x) - 0| \le \epsilon$, so setting $f = 0$ always work. Otherwise, when $\sum_{i=1}^N |\alpha_i| \ne 0$, for each region R_i, we define $R_i := [a_1^i b_1^i] \times [a_2^i b_2^i] \times \cdots [a_d^i b_d^i]$, which is a Cartesian product of 1-d intervals. Then it is easy to see that there exists a bump function $g_\gamma^i(x)$ defined as in Def. 4 that

$$g_\gamma^i(x) = \begin{cases} 1 & \text{if } x \in R_i = x \in [a_1^i, b_1^i] \times \cdots [a_d^i, b_d^i] \\ 0 & \text{if } x \notin [a_1^i - \gamma, b_1^i + \gamma] \times \cdots [a_d^i - \gamma, b_d^i + \gamma] \\ [0,1] & \text{otherwise} \end{cases}$$

Since $\gamma \to 0, g_\gamma^i \to \mathbf{1}_{R_i}$ by definition, so , $\exists \gamma_i$ s.t. $||g_{\gamma_i}^i - \mathbf{1}_{R_i}||_1 \le \frac{\epsilon}{\sum_i |\alpha_i|}$. Therefore, choosing $f_i = g_{\gamma_i}^i$ satisfies the requirement. □

Lemma 3 (Partition Lemma). *let g, δ, ϵ be given. Under the same assumption of Theorem 2, we have for any partition P of $[0,1]^d$, $P = (R_1, \dots, R_N)$ with all side length smaller than δ, there exists $(\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$ such that*

$$\sup_{x \in [0,1]^d} |g(x) - h(x)| \le \epsilon \textit{ with } h(x) := \sum_{i=1}^N \alpha_i \mathbf{1}_{R_i}(x).$$

Proof. For each R_i, pick $x_i \in R_i$ that satisfies $\alpha_i = g(x_i)$. Therefore we have

$$\begin{aligned} \sup_{x \in [0,1]^d} |g(x) - h(x)| &= \sup_{i \in [N]} \sup_{x \in R_i} |g(x) - h(x)| \\ &= \sup_{i \in [N]} \sup_{x \in R_i} (|g(x) - g(x_i)| + |g(x_i) - h(x)|) \\ &\le \epsilon + 0 = \epsilon \end{aligned}$$

Here the first term of the last inequality comes from the construct of partition and the fact that $||x - x'||_\infty \leq \delta \Rightarrow |g(x) - g(x')| \leq \epsilon$. The second term comes from the definition of α_i. □

Definition 4 (Bump function)**.** *For some $\gamma > 0$ and a relu function σ, we define*

$$g_{\gamma,j}(z) = \sigma(\frac{z-(a_j-\gamma)}{\gamma}) - \sigma(\frac{z-a_j}{\gamma}) - \sigma(\frac{z-b_j}{\gamma}) + \sigma(\frac{z-(b_j+\gamma)}{\gamma}) \quad \textit{1-dim case}$$

$$g_\gamma(x) = \sigma(\sum_{j=1}^{d} g_{\gamma,j}(x^j) - (d-1)) \quad \textit{d-dim case}$$

2 Barron's Theorem

Theorem 5. *For any $g : \mathbb{B}_1 \to \mathbb{R}$ where $\mathbb{B}_1 = \{x \in \mathbb{R} : \|x\|_2 \leq 1\}$ is the unit ball, there exists a 3-layer neural network f with $O\left(\frac{C^2}{\epsilon}\right)$ neurons and sigmoid activation function such that*

$$\int_{\mathbb{B}_1} (f(x) - g(x))^2 dx \leq \epsilon$$

Proof Idea The overall proof ideas can be decomposed into three steps:

- Step 1: show any continuous function can be written as an infinite width neural network with cosine-like activation functions. (Tool: Fourier representation)
- Step 2: Show that a function with a small Barron constant can be approximated by a convex combination of a small number (finite width neural network) of cosine-like activation functions. (Tool: subsampling / probabilistic method.)
- Step 3 Show that the cosine function can be approximated by sigmoid functions. (Tool: classical approximation theory.

Proof. For Step 1, we will approximate any continuous function $f(x)$ using infinite width NN by Fourier representation. That is,

$$\hat{f}(w) := \int \exp\left(-2\pi i w^\top x\right) f(x) \mathrm{d}x \qquad \text{(Fourier transform)}$$

$$f(x) := \int \exp\left(2\pi i w^\top x\right) \hat{f}(w) \mathrm{d}w \qquad \text{(inverse Fourier transform)}$$

Now recall the definition of infinite NN as follows.

Definition 6. *An infinite-wide neural network is defined by a signed measure ν over neuron weights (w, b),*

$$f(x) = \int_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sigma\left(w^\top x + b\right) d\nu(w, b).$$

You can see that the inverse Fourier form is very close to this infinite NN, except that the inverse Fourier transform has its activations over complex space. By using polar decomposition as shown in Lemma 7, we can rewrite the representation as

$$f(x) = f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| \underbrace{(\cos(b_w + \langle w, x \rangle) - \cos(b_w))}_{\text{cos-like activation functions}} dw$$

Now in Step 2, we will approximate this $\int_{\mathbb{R}^d} |\hat{f}(w)| (\cos(b_w + \langle w, x\rangle) - \cos(b_w))\, dw$ by a finite neural network by writing the function as the expectation of a random variable.

$$f(x) = f(0) + \int_{\mathbb{R}^d} \frac{|\hat{f}(w)| \|w\|_2}{C} \left(\left(\frac{C}{\|w\|_2} \right) (\cos(b_w + \langle w, x\rangle) - \cos(b_w)) \right) dw$$

where $C = \int_{\mathbb{R}^d} \|\hat{f}(w)|\|w\|_2 dw$. So $\frac{|\hat{f}(w)|\|w\|_2}{C}$ is the probability on w. Denote the overall distribution by D_w. Now by Lemma 8, we have that, when sampling each $w_i \sim D_m$ for $O(\frac{C^2}{\epsilon})$ times, the empirical mean can be ϵ-close to the expectation. That is

$$f(x) - \epsilon \lessapprox f(0) + \frac{1}{r} \sum_{i=1}^{r} \frac{C}{\|w_i\|} (\cos(b_{w_i} + \langle w_i, x\rangle) - \cos(b_{w_i})) \lessapprox f(x) + \epsilon$$

The final Step 3 is to transform the cosine-like activation to sigmoid activation, which is the commonly used activation in the standard neural net. This can be easily done by using a 2-layer NN as shown in Lemma 9.

Therefore, combining this 2-layer approximation for cos-like activation and a 2-layer cosine-like $O(\frac{C^2}{\epsilon})$ nodes NN gives the 3-layer network g. □

Lemma 7 (Fourier representation in forms of infinite NN)**.** *The function* $f(x) = \int_{\mathbb{R}^d} \hat{f}(w) e^{i\langle w, x\rangle} dw$ *can be written as*

$$f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| (\cos(b_w + \langle w, x\rangle) - \cos(b_w))\, dw$$

Proof. This can be done by some direct rearrangements and replacements.

$$\begin{aligned}
\int_{\mathbb{R}^d} \hat{f}(w) e^{i\langle w, x\rangle} dw &= \int_{\mathbb{R}^d} \hat{f}(w) dw + \int_{\mathbb{R}^d} \hat{f}(w)(e^{i\langle w, x\rangle} - 1) dw \\
&= f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| e^{ib_w}(e^{i\langle w, x\rangle} - 1) dw \\
&= f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| (e^{i(b_w + \langle w, x\rangle)} - e^{ib_w}) dw \\
&= f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| (\cos(b_w + \langle w, x\rangle) - \cos(b_w))\, dw
\end{aligned}$$

Here the third equality comes from the polar decomposition of $\hat{f}(w)$ – the $|\hat{f}(w)|$ is the magnitude part and e^{ib_w} the radial part (b_w as the angle), and the last inequality comes from $e^{iz} = \cos(z) + i\sin(z)$. □

Lemma 8 (Approximation error of Sub-sampling)**.** *Sample one $w \in \mathbb{R}^d$ with probability $\frac{|\hat{f}(w)| \|w\|_2}{C}$ for r times, denote them as $\{w_i\}_{i \in [r]}$. Then, as long as $r = O(\frac{C^2}{\epsilon})$, we have with high probability,*

$$f(x) - \epsilon \lessapprox f(0) + \frac{1}{r} \sum_{i=1}^{r} \frac{C}{\|w_i\|} \left(\cos\left(b_{w_i} + \langle w_i, x \rangle\right) - \cos\left(b_{w_i}\right)\right) \lessapprox f(x) + \epsilon$$

Proof. The results come from standard concentration inequality and we omit the details here. □

Lemma 9. *Given $g_w(x) = \frac{C}{\|w\|_2} \left(\cos\left(b_w + \langle w, x \rangle\right) - \cos\left(b_w\right)\right)$, there exists a 2-layer neural network f_0 of size $O(1/\epsilon)$ with sigmoid activations, such that $\sup_{x \in [-1,1]} |f_0(y) - h_w(y)| \leq \epsilon$.*