CSE 599 (Neural Networks Theory): Homework #1

Due on February, 15th at 23:59

Prof. Simon S. Du

Alexey Sholokhov

Problem 1

Solution for 1.1 Let $g \in C^2$, $g(0) = g'(0) = 0$, $\sigma(a) \triangleq a\mathbb{I}_{[a>0]}$. Taking the integral from the assignment by parts we get that:

$$\begin{aligned} \int_0^1 \sigma(x-b)g''(b)\mathrm{d}b = \int_0^1 (x-b) \int x - b > 0g \\ (b)\mathrm{d}b = \int_0^x (x-b)g''(b)\mathrm{d}b = (x-b)g'(x)\Big|_0^x - \int_0^x g'(b)\mathrm{d}b = g(x) \end{aligned} \tag{1}$$

$\square$

Solution for 1.2 First, we notice that $|g''| \leq \beta$ implies that g is a β-smooth function, i.e. its gradient g' is β-Lipschitz continuous. Since this is the case, we can use the theorem from Feb. 11th lecture to show that there is a threshold network that approximates the derivative $g'(x)$ ε-well by infinity-norm:

$$\|g(x)' - f(x)\|_\infty = \left\| g(x) - \sum_{i=1}^m a_i \mathbb{I}_{[x-x_i]} \right\|_\infty \leq \varepsilon \tag{2}$$

where $x_i \triangleq (i-1)\varepsilon\beta^{-1}$, $m \triangleq \lceil \beta\varepsilon^{-1} \rceil$, and $a_i = g'(x_i) - g'(x_{i-1})$. We also know that if $\|g - f\|_\infty = \max_{x\in[a,b]} |g(x) - f(x)| \leq \varepsilon$ then

$$\left\| \int (f-g) \right\|_\infty = \max_{x\in[a,b]} \left| \int_a^x (f(b) - g(b))\mathrm{d}b \right| \leq \max_{x\in[a,b]} \int_a^x |f(b) - g(b)|\,\mathrm{d}b \leq \varepsilon * (b-a) \tag{3}$$

In particular, we can apply this lemma to the equation (2) to get an approximation of $g(x)$ with a shallow neural network:

$$\begin{aligned} \max_{x\in[0,1]} \left| \int_0^x (g'(x) - f(x)) \right| &= \max_{x\in[0,1]} \left| g(x) - \cancelto{0}{g(0)} - \int_0^x \sum_{i=1}^m a_i \mathbb{I}_{[b-x_i]}\mathrm{d}b \right| = \\ &= \max_{x\in[0,1]} \left| g(x) - \sum_{i=1}^m a_i \int_0^x \mathbb{I}_{[b-x_i]}\mathrm{d}b \right| = \\ &= \max_{x\in[0,1]} \left| g(x) - \sum_{i=1}^m (g'(x_i) - g'(x_{i-1}))\sigma(x - x_i) \right| \leq \varepsilon(1-0) \leq \varepsilon \end{aligned} \tag{4}$$

$\square$

Solution for 1.3 First, we notice that the equation that we proved in the problem 1.1 is a representation of $g(x)$ with an infinite-wide shallow neural network with ReLU activation function. Now we use Pister's lemma: according to it we can sample coefficients $\{a_i,\ b_i\}_{i=1}^m$ from the signed density function $\mu(b) = g''(b)\mathrm{d}b$ such that

$$\begin{aligned} \left\| g(x) - \frac{1}{m}\sum_{i=1}^m a_i\sigma(x-b_i) \right\|_{L_2}^2 \leq \mathbb{E}\left[\left\| g(x) - \frac{1}{m}\sum_{i=1}^m a_i\sigma(x-b_i) \right\| \right] \leq \|\mu\|_1^2 \sup_b \|\sigma(x-b)\|_{L_2(P_X)} = \\ \frac{1}{m}\left(\int_0^1 |g''(x)|\mathrm{d}x \right)^2 \sup_{b\leq 1} \cancelto{= b^3/3 \leq 1}{\int_0^1 \sigma^2(\xi - b)\mathrm{d}\xi} \leq \varepsilon \end{aligned} \tag{5}$$

According to Pister's lemma, for $\varepsilon > 0$ the above holds for some $m \leq \lceil \varepsilon^{-1} \left(\int_0^1 |g''(x)|\mathrm{d}x \right)^2 \rceil$, which is what we want. $\square$

Problem 2

Solution for 2.1 First, we split the expectation from the problem assignment into a full system of four cases:

1. Let $x^T w \geq 0$ and $x^T w^* \geq 0$. Treating all expectations below as conditionals on the event, we get:

$$\mathbb{E}\left[(\sigma(x^T w) - \sigma(x^T w^*))^2\right] = \mathbb{E}\left[(x^T w)^2\right] - \mathbb{E}\left[2x^T w x^T w^*\right] + \mathbb{E}\left[(x^T w^*)^2\right] \tag{6}$$

We'll evaluate all three expectations using polar coordinates. Let θ_x be the angle of x, θ_w be the angle of w^*, θ_{w^*} be the angle of w^*, and θ^* be the angle between w and w^*. Since $x^T w \geq 0$ we know that $\theta_x - \theta_w \in [-\pi/2; \pi/2]$. Similarly, $x^T w^* \geq 0$ gives us $\theta_x - \theta_{w^*} = \theta_x - \theta_w - \theta^* \in [-\pi/2; \pi/2]$. The intersection of these bounds is $\theta_x - \theta_w \in [-\pi/2 + \theta^*; \pi/2]$, which provides us with bounds for the polar part in the integrals below:

$$\begin{aligned}\mathbb{E}\left[(x^T w)^2\right] &= \frac{1}{2\pi}\int_{x^T w \geq 0, x^T w^* \geq 0} \|x\|_2^2 \|w\|_2^2 e^{-x^T x/2} dx = \\ &= \frac{1}{2\pi}\cancelto{2}{\int_0^\infty r^3 e^{-r^2/2} dr} * \int_{\theta_x - \theta_w = -\pi/2 + \theta^*}^{\pi/2} 1 \mathrm{d}(\theta_x - \theta_w) = \frac{1}{\pi}(\pi - \theta^*)\|w\|_2^2\end{aligned} \tag{7}$$

Symmetrically, we have $\mathbb{E}\left[(x^T w^*)^2\right] = \frac{1}{\pi}(\pi - \theta^*)\|w^*\|_2^2$. It gets more involved with the cross term:

$$\mathbb{E}\left[x^T w x^T w^*\right] = \tag{8}$$