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:

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

□

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:

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

□

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

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

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. □

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:

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

   Symmetrically, we have $\mathbb{E}\left[(x^T w^*)^2\right] = \frac{1}{2\pi}(\pi - \theta^* + \sin(\theta^*)\cos(\theta^*))\|w^*\|_2^2$. For the cross term:

$$\begin{aligned}
\mathbb{E}\left[x^T w x^T w^*\right] &= \mathbb{E}\left[\|x\|_2^2 \|w\|_2 \|w^*\|_2 \cos(\theta_x - \theta_w)\cos(\theta_x - \theta_{w^*})\right] = \\
&= \|w\|_2 \|w^*\|_2 \frac{1}{2\pi} \int_0^\infty r^3 e^{-r^2/2} \mathrm{d}r \int_{-\pi/2+\theta^*}^{\pi/2} \cos(\theta)\cos(\theta - \theta^*) \mathrm{d}\theta = \\
&= \|w\|_2 \|w^*\|_2 \frac{1}{\pi}\frac{1}{2}(\sin(\theta^*) + (\pi - \theta^*)\cos\theta^*)
\end{aligned} \tag{8}$$

2. Let $x^T w \geq 0$ and $x^T w^* < 0$, then $\theta_x - \theta_w \in [-\pi/2; -\pi/2 + \theta^*$. Hence, the conditional expectation

$$\mathbb{E}\left[(\sigma(x^T w) - \sigma(x^T w^*))^2\right] = \mathbb{E}\left[(x^T w)^2\right] = \frac{\|w\|^2}{\pi} \int_{-\pi/2}^{-\pi/2+\theta^*} 1 \mathrm{d}\theta = \frac{\|w\|^2}{2\pi}(\theta^* - \sin\theta\cos\theta^*) \tag{9}$$

   Symmetrically, the case of $x^T w < 0$ and $x^T w^* \geq 0$ yields

$$\mathbb{E}\left[(\sigma(x^T w) - \sigma(x^T w^*))^2\right] = \frac{\|w^*\|^2}{2\pi}(\theta^* - \sin\theta\cos\theta^*) \tag{10}$$

Now we open the expectation up using the full probability formula. Combining the pieces above together and cancelling out matching terms gives us

$$\begin{aligned}
&\mathbb{E}\left[(\sigma(x^T w) - \sigma(x^T w^*))^2\right] = \frac{1}{2\pi}(\pi - \theta^* + \sin(\theta^*)\cos(\theta^*))\|w\|_2^2 - \\
&- 2\|w\|_2\|w^*\|_2 \frac{1}{\pi}\frac{1}{2}(\sin(\theta^*) + (\pi - \theta^*)\cos\theta^*) + \frac{1}{2\pi}(\pi - \theta^* + \sin(\theta^*)\cos(\theta^*))\|w^*\|_2^2 + \\
&+ \frac{\|w\|^2}{2\pi}(\theta^* - \sin\theta\cos\theta^*) + \frac{\|w^*\|^2}{2\pi}(\theta^* - \sin\theta\cos\theta^*) = \\
&= \frac{1}{2}\|w\|^2 - \|w\|_2\|w^*\|_2 \frac{1}{\pi}(\sin(\theta^*) + (\pi - \theta^*)\cos\theta^*) + \frac{1}{2}\|w^*\|^2
\end{aligned} \tag{11}$$

which is what we want to show.

Problem 2 continued on next page. . .

Now we take the derivative of the formula above with respect to w:

$$\begin{aligned}\nabla_w f(w) &= w + \cancel{\nabla_w(\|w^*\|_2^2)}^{0} - \nabla_w \left(\|w\|_2 \|w^*\|_2 \frac{1}{\pi}(\sin(\theta^*) + (\pi - \theta^*)\cos\theta^*) \right) = \\ &= w - \frac{1}{\pi}\|w^*\|_2 \frac{w}{\|w\|_2}(\sin\theta^* + (\pi-\theta^*)\cos\theta^*) + \\ &+ \frac{1}{\pi}\|w\|_2\|w^*\|_2(\nabla(\sin\theta^*) + \pi\nabla(\cos\theta^*) - \nabla(\sin\theta^*) - \theta^*\nabla(\cos\theta^*)) =\end{aligned} \tag{12}$$

We can evaluate $\nabla_w(\cos\theta^*)$ by opening it up as a scalar product:

$$\nabla_w(\cos\theta^*) = \nabla_w\left(\frac{w^T w^*}{\|w\|_2\|w^*\|_2}\right) = \frac{w^*}{\|w\|_2\|w^*\|_2} - \underbrace{\frac{w^T w^*}{\|w\|_2\|w^*\|_2}}_{\cos\theta^*}\frac{w}{\|w\|_2^2} \tag{13}$$

Substituting this result back we get

$$\begin{aligned}&= w - \frac{1}{\pi}\|w^*\|_2\frac{w}{\|w\|_2}(\sin\theta^* + \cancel{(\pi-\theta^*)\cos\theta}) + \frac{1}{\pi}\|w\|_2\|w^*\|_2(\pi-\theta^*)\left[\frac{w^*}{\|w\|_2\|w^*\|_2} - \cancel{\cos\theta^*\frac{w}{\|w\|_2^2}}\right] = \\ &= w - \frac{w}{\pi}\frac{\|w^*\|_2}{\|w\|_2}\sin\theta^* - \frac{w^*}{\pi}(\pi-\theta^*)\end{aligned} \tag{14}$$

which is what we want. □

Solution for 2.2 The set of critical points is a solution set for $\nabla f(w) = 0$. As asked, let's assume $w \neq 0$. Notice that the same equation for the gradient can be written as

$$\alpha w = \beta w^* \tag{15}$$

It implies that w should be collinear to w^* for each critical point. In other words $\theta^* = 0$ for these points. Evaluating α and β with this condition makes it clear that there is only one such point: $w = w^*$.

$$\begin{aligned}\alpha &= 1 - \frac{1}{\pi}\frac{\|w^*\|_2}{\|w\|_2}\sin 0 = 1 \\ \beta &= \frac{1}{\pi}(\pi - 0) = 1\end{aligned} \tag{16}$$

□

Solution for 2.3 Let's notice that the equation for w_{t+1} can be written as:

$$w_{t+1} = w_t\alpha(w_t) + \beta(w_t)w^* = w_t(1 - \eta g(w_t)) + \beta(w_t)w^*, \quad \alpha(w_t),\ \beta(w_t) \in \mathbb{R} \tag{17}$$

where

$$g(w_t) = 1 - \frac{\sin\theta(w_t, w^*)}{\pi}\frac{\|w^*\|_2}{\|w_t\|_2} \in [0, 1] \tag{18}$$

It means that for any $\eta < 1$ $\alpha = (1 - \eta g(w_t)) < 1$. Then we have that $w_{t+1} = \alpha_t w_t + \beta_t w^* = \alpha_t(\alpha_{t-1}w_{t-1} + \beta_{t-1}w^*) + \beta_t w^* = w\prod_{j=1}^t \alpha_j + w^*(\beta_t \sum_{j=t-1}^1 \beta_j\alpha_{j+1}\beta_j)$. The sequence of angles $\theta_t = \theta(w_t, w^*)$ is non-increasing because the product of α_j is non-increasing for any $\alpha_j < 1$. Hence, $\theta(w_{t+1}, w^*) \le \theta(w_t, w^*)$. □

Problem 3

First, we notice that since we're given that the matrix M has an eigendecomposition $M = \sum_{i=1}^{d} \lambda_i u_i u_i^T$ it implies that the matrix M is symmetric: $M^T = \sum_{i=1}^{d} \lambda_i (u_i u_i^T)^T = \sum_{i=1}^{d} \lambda_i u_i u_i^T = M$. Next, we use this fact ($M_{ij} = M_{ji}$) to simplify the gradient:

$$\begin{aligned}\frac{\partial f}{\partial x_i} &= \frac{1}{2}\sum_{j\neq i}^{d} x_j(x_i x_j - M_{ij}) + \frac{1}{2}\sum_{j\neq i}^{d} x_j(x_j x_i - M_{ji}) + x_i(x_i^2 - M_{ii}) = \\ &= \sum_{j=1}^{d} x_j(x_i x_j - M_{ij})\end{aligned} \tag{19}$$

The gradient in the vector form is:

$$\nabla f(x) = x\|x\|_2^2 - Mx \tag{20}$$

The stationary equation for such system gives us that the set of stationary points match to the set of eigenvectors multiplied by the square root of respective eigenvalues (plus zero):

$$Mx = \|x\|x \implies x_0^* = 0,\ x_j^* = \sqrt{\lambda_j} u_j \tag{21}$$

where we used the fact that the matrix U in the eigendecomposition of M is orthonormal, hence all eigenvectors are unit and orthogonal.
To establish qualities of these critical points (minima, maxima, saddles) we use Hessian:

$$[\nabla^2 f(x)]_{ik} = \left[\sum_{j=1}^{d} x_j(x_i x_j - M_{ij})\right]_k' = \{k \neq i\}(2x_k x_i - M_{ik}) + \{k = i\}(\sum_{j\neq k} x_j x_j + 3x_k^2 - M_{kk}) \tag{22}$$

Hence, Hessian in the matrix from looks like

$$\nabla^2 f(x) = 2xx^T - M + \|x\|_2^2 I \tag{23}$$

Now let's consider all possible stationary points:

- $x = x_0^* = 0$. In this case, $\nabla^2 f(x) = -M$ where M is a PSD matrix. It implies that $x^* = 0$ is a local maximum, and so any direction that corresponds to, say, non-zero eigenvector of M will be a descent direction.

- $x = x_1^* = \sqrt{(\lambda_1)} u_i$. It's easy to show that this point is a global minimum: out of all critical points $\{x_i^*\}_{i=1}^d$ the function $f(x)$ achieves minimal value at $x = x_1^*$.

$$f(x_1^*) = \frac{1}{4}\|\lambda_1 u_1 u_1^T - \sum_{i=1}^{d} \lambda_i u_i u_i^T\| = \frac{1}{4}\|\sum_{i=2}^{d} \lambda_i u_i u_i^T\|_F^2 = \sum_{i=2}^{d} \lambda_i^2 \leq \sum_{i\neq j} \lambda_i^2 = f(x_j^*), \quad j \neq 1 \tag{24}$$

- Now we show that the rest of the critical points are saddle points and that they're strict. Let's consider $x = x_i^*$, $i \neq 1$. Hessian at this point looks like:

$$\nabla^2 f(x_i^*) = \lambda_i I + \lambda_i u_i u_i^T - \sum_{j\neq i} \lambda_j u_j u_j^T \tag{25}$$

  It's clear that the direction $w = \sqrt{\lambda_1} u_1$ is a descent direction from any of the points x_i^*, $i \neq 1$

$$\lambda_1 u_1 \nabla^2 f(x_i^*) u_1 = \lambda_i \lambda_1 + 0 - \lambda_1^2 < 0 \tag{26}$$

  which is what we need.

□