2025 Spring AMA564 Assignment 1

Department of Applied Mathematics, The Hong Kong Polytechnic Univeristy

Due 23:59, Sunday, March 02, 2025

1. (a) **Example:** For ReLU activation function $\sigma(x) = \max\{0, x\}$, its derivative is $\frac{d}{dx}\sigma(x) = I(x > 0)$ where $I(x > 0) = 1$ if $x > 0$ and $I(x > 0) = 0$ if $x \leq 0$. $\lim_{x \to +\infty} \frac{d}{dx}\sigma(x) = 1$ and $\lim_{x \to -\infty} \frac{d}{dx}\sigma(x) = 0$.
Question: Please calculate the derivative of the following activation functions (ignore the places where the activation function is not differentiable). Please also calculate the function value of the derivatives at infinity.
(1) (**1 mark**) Sigmoid activation function $\frac{1}{1+e^{-x}}$.
(2) (**1 mark**) Tanh activation function $\frac{e^{2x}-1}{e^{2x}+1}$.
(3) (**1 mark**) Leaky ReLU activation function $\max\{ax, x\}$ for some $a \in (0, 1)$.

(b) Let $f(x; \theta) = W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}$ be a ReLU activated multi-layer perceptron with one hidden layer where $\theta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ denotes the parameters of multi-layer perceptron f and $\sigma(a) = \max\{0, a\}$ be the ReLU activation function ($\sigma(a) = (\sigma(a_1), \sigma(a_2), \cdots, \sigma(a_d))^\top$ if $a = (a_1, a_2, \cdots, a_d)^\top$ is a d-dimensional vector). Figure 1 illustrates the architecture of the multi-layer perceptron f.

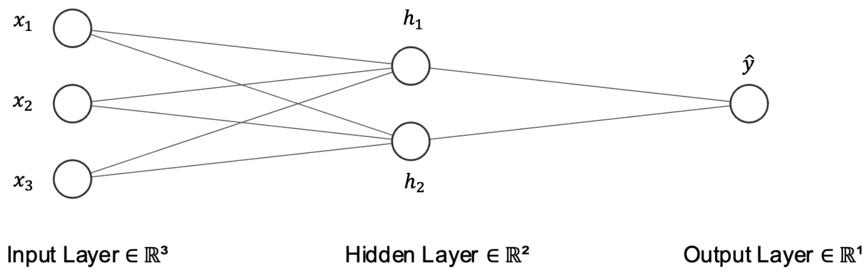

Figure 1: Architecture of the multi-layer perceptron f.

Question: (3 marks) Suppose the value of $\theta = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ are

$$W^{(1)} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} & w_{13}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} & w_{23}^{(1)} \end{bmatrix} = \begin{bmatrix} -0.8 & 0.5 & -1 \\ 1.2 & -0.7 & 0.2 \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix} = \begin{bmatrix} -0.4 \\ 0.9 \end{bmatrix},$$

and

$$W^{(2)} = \begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.6 & 1.1 \end{bmatrix}, \quad b^{(2)} = \begin{bmatrix} b_1^{(2)} \end{bmatrix} = \begin{bmatrix} -0.1 \end{bmatrix}.$$

Now we have a sample (x, y) with $x = (x_1, x_2, x_3)^\top = (1, 2, 3)^\top$ and $y = 0.33$. Define the loss $L = (f(x;\theta) - y)^2/2$. Please calculate the value of the forward pass, i.e. the value of $f(x;\theta)$. Please use back-propagation algorithm to calculate the derivatives of the loss L with respect to each weight and bias, i.e. $dL/dw_{11}^{(1)}, \cdots, dL/dw_{23}^{(1)}, dL/db_1^{(1)}$, $dL/db_2^{(1)}$, $dL/dw_{11}^{(2)}, dL/dw_{12}^{(2)}$ and $dL/db_1^{(2)}$.

2. **Background:** Let $\{(X_i, Y_i)\}_{i=1}^n$ be an independently and identically distributed (i.i.d) sample drawn from the joint distribution of (X, Y). The objective of deep quantile regression is to minimize the empirical risk

$$\mathcal{R}_n(f(\cdot;\theta)) := \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - f(X_i;\theta)),$$

over a class of neural networks $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \mathbb{R}^{\mathcal{S}}\}$ where $f(\cdot;\theta)$ is a neural network parameterized by $\theta \in \mathbb{R}^{\mathcal{S}}$ with size $\mathcal{S}$ and $\rho_\tau(a) = (\tau - I(a < 0)) \cdot a$ is the quantile check loss function with some $\tau \in (0, 1)$.

Now we consider the risk minimization problem at the population level and wonder what is the target of the risk minimization problem with quantile loss function. Given a function f, we define the risk of f by

$$\mathcal{R}(f) := \mathbb{E}\{\rho_\tau(Y - f(X))\},$$

where the expectation $\mathbb{E}$ is taken with respect to (X, Y). And we define the minimizer of the risk (target) by

$$f^* = \arg\min_f \mathcal{R}(f) = \arg\min_f \mathbb{E}\{\rho_\tau(Y - f(X))\}.$$

Assumption: Suppose $X \in \mathbb{R}^d$ is a random vector, $Y \in \mathbb{R}$ is a continuous random variable satisfying $\mathbb{E}\{|Y| \mid X = x\} < \infty$ for each $x \in \mathbb{R}^d$.

Question: (**2 marks**) Please prove that for each $x \in \mathbb{R}^d$, the $f^*(x)$ is the conditional τ-th quantile of the random variable Y given $X = x$.

Hints: Follow and modify the proof for Least Absolute Deviation loss in Lecture Note 2.

3. **Background:** Given data $\{(X_i, Y_i)\}_{i=1}^n$, we are interested in minimizing an empirical loss

$$f(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta : X_i, Y_i)$$

over $\theta \in \mathbb{R}^{\mathcal{S}}$ where $\ell(\cdot)$ is some loss function. Now we use gradient decent algorithm to optimize the problem. We start from some random initialization $\theta^0 \in \mathbb{R}^{\mathcal{S}}$ and for $k = 0, 1, \ldots, T-1$, we update as

$$\theta^{k+1} = \theta^k - \frac{1}{L}\nabla f(\theta^k),$$

where in each update we choose the a fixed stepsize $1/L$. Then we obtain a sequence $\{\theta^k\}_{k=0}^T$ generated by the gradient descent algorithm.

Assumption: Suppose f has a finite lower bound, i.e., there exists $\bar{f} \in \mathbb{R}$ such that $f(\theta) \geq \bar{f} > -\infty$ for any θ in the domain of f. Also, suppose f is a L-smooth function for some $L > 0$, i.e., f is continuously differentiable and its gradient ∇f is Lipschitz continuous ($\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$).

Question: (**2 marks**) Please prove that

$$\min_{1 \leq k \leq T} \|\nabla f(\theta^k)\|^2 \leq \frac{2L\{f(\theta^0) - \bar{f}\}}{T}.$$

Hints: (1) Apply Lemma 3.1 at step k. (2) Sum them up for $k = 0, 1, \ldots, T$. (3) Note that f is bounded from below by $\bar{f}$.