AMA 564 Assignment 1 24133101g Wu Yifan

1. (a) (1) Sigmoid activation function $y = f(x) = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1}$

$$\frac{dy}{dx} = -(1+e^{-x})^{-2} \cdot -e^{-x} = e^{-x}(1+e^{-x})^{-2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\lim_{x\to+\infty} \frac{dy}{dx} = \lim_{x\to+\infty} \frac{e^{-x}}{(1+e^{-x})^2} = 0 \qquad \lim_{x\to-\infty} \frac{dy}{dx} = \lim_{x\to-\infty} \frac{e^{-x}}{(1+e^{-x})^2} = 0$$

(2) Tanh activation function $y = f(x) = \frac{e^{2x}-1}{e^{2x}+1}$

$$\frac{dy}{dx} = \frac{2e^{2x}(e^{2x}+1) - (e^{2x}-1)2e^{2x}}{(e^{2x}+1)^2} = \frac{4e^{2x}}{(e^{2x}+1)^2}$$

$$\lim_{x\to+\infty} \frac{dy}{dx} = \lim_{x\to+\infty} \frac{4e^{2x}}{(e^{2x}+1)^2} = 0 \qquad \lim_{x\to-\infty} \frac{dy}{dx} = \lim_{x\to-\infty} \frac{4e^{2x}}{(e^{2x}+1)^2} = 0$$

(3) Leaky ReLU activation function

$$y = f(x) = \max\{ax, x\} = \begin{cases} x, & x > 0 \\ ax, & x \le 0 \end{cases} \quad \text{for some } \alpha \in (0,1)$$

$$\frac{dy}{dx} = \begin{cases} 1, & x > 0 \\ a, & x \le 0 \end{cases}$$

$$\lim_{x\to+\infty} \frac{dy}{dx} = 1, \qquad \lim_{x\to-\infty} \frac{dy}{dx} = a \quad \text{for some } a \in (0,1)$$

(b)

For hidden layer:

$$h = \begin{bmatrix} -0.8 & 0.5 & -1 \\ 1.2 & -0.7 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \begin{bmatrix} -0.4 \\ 0.9 \end{bmatrix} = \begin{bmatrix} -3.2 \\ 1.3 \end{bmatrix}$$

ReLU: $\mathrm{ReLU}(h) = \begin{bmatrix} 0 \\ 1.3 \end{bmatrix}$

For output layer:

$$O = \mathrm{ReLU}\left\{ \begin{bmatrix} 0.6 & 1.1 \end{bmatrix} \begin{bmatrix} 0 \\ 1.3 \end{bmatrix} + \begin{bmatrix} -0.1 \end{bmatrix} \right\} = 1.33$$

Back-propagation

loss fuction: $L = \frac{1}{2}(f(x;\theta) - y)^2$, $L' = f(x;\theta) - y$

ReLU function: $z(x) = \max(0, x)$, $z'(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases}$

For output layer:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial O} \cdot \frac{\partial O}{\partial z} \cdot \frac{\partial z}{\partial W} = 1 \cdot 1 \cdot \begin{bmatrix} 0 \\ 1.3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1.3 \end{bmatrix}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial O} \cdot \frac{\partial O}{\partial z} \cdot \frac{\partial z}{\partial b} = 1 \cdot 1 \cdot 1 = 1$$

For hidden layer:

for the output of hidden layer h:

$$\frac{\partial L}{\partial h} = \frac{\partial L}{\partial o} \cdot \frac{\partial o}{\partial z} \cdot \frac{\partial z}{\partial x} = 1 \cdot W^{(2)} = \begin{bmatrix} 0.6 \\ 1.1 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_h} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial W_h} = \begin{bmatrix} 0.6 \\ 1.1 \end{bmatrix} \odot \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 1.1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1.1 & 2.2 & 3.3 \end{bmatrix}$$

$$\frac{\partial L}{\partial b_h} = \frac{\partial L}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial b_h} = \begin{bmatrix} 0.6 \\ 1.1 \end{bmatrix} \odot \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot 1 = \begin{bmatrix} 0 \\ 1.1 \end{bmatrix}$$

2.

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

$$R(f) = \int_{-\infty}^{+\infty} \rho_\tau(Y - f(x))\, dF_{Y|X}(y)$$

$$= \int_{-\infty}^{f(x)} (\tau - 1)(Y - f(x))\, dF_{Y|X}(y)$$

$$+ \int_{f(x)}^{+\infty} \tau (Y - f(x))\, dF_{Y|X}(y)$$

$$\frac{dR(f)}{df} = \int_{-\infty}^{f(x)} (\tau - 1)(-1)\, dF_{Y|X}(y)$$

$$+ \int_{f(x)}^{+\infty} \tau(-1)\, dF_{Y|X}(y)$$

$$= (\tau - 1) \cdot -F_{Y|X}(f(x)) - \tau(1 - F_{Y|X}(f(x)))$$

$$= -\tau F_{Y|X}(f(x)) + \tau F_{Y|X}(f(x)) + F_{Y|X}(f(x)) - \tau = 0$$

$\Rightarrow F_{Y|X}(f(x)) = \tau$, which means $P(Y \leq f(x)) = \tau$.

Thus, $f^*(x)$ is the conditional τ-th quantile.

3. At step k, according to lemma 3.1, we have

$$f(\theta^{k+1}) - f(\theta^k) \le -\frac{1}{2L}\|\nabla f(\theta^k)\|^2$$

Consider $k=1,\dots,T$, then

$$f(\theta^1) - f(\theta^0) \le -\frac{1}{2L}\|\nabla f(\theta^0)\|^2$$

$$f(\theta^2) - f(\theta^1) \le -\frac{1}{2L}\|\nabla f(\theta^1)\|^2$$

$$\vdots \qquad\qquad \vdots$$

$$f(\theta^{T+1}) - f(\theta^T) \le -\frac{1}{2L}\|\nabla f(\theta^T)\|^2$$

Sum them up

$$f(\theta^{T+1}) - f(\theta^0) \le -\frac{1}{2L}\left[\|\nabla f(\theta^0)\|^2 + \cdots + \|\nabla f(\theta^T)\|^2\right]$$

$$\le -\frac{1}{2L}\left(T \cdot \min_{0\le k\le T}\|\nabla f(\theta^k)\|^2\right)$$

Consider $\bar{f}\in\mathbb{R}$ satisfy $f(\theta)\ge\bar{f}>-\infty$, then we have

$$\bar{f} - f(\theta^0) \le -\frac{1}{2L}\left(T \cdot \min_{0\le k\le T}\|\nabla f(\theta^k)\|^2\right)$$

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