HW7 Gradients questions 1a & 2a

NetID: sh2432

SDS 555 HW7 sh2432 Gradient

1a) $h_1 = ReLU(W_1 X + b_1)$, $h_2 = ReLU(W_2 h_1 + b_2)$, $p = Softmax(W_3 h_2 + b_3)$

$$p(Y=k|x_i) = p(Y=k|X) = \frac{\exp(f(y_i))}{\sum_{i=1}^{K} \exp(f(y_i))} \quad \text{for } k = 1, 2 \dots K$$

$$L = \frac{1}{n}\sum_{i=1}^{n} -\log p(Y=y_i|x_i) = \frac{1}{n}\sum_{i=1}^{n} -\log \frac{\exp(f(y_i))}{\sum_{i=1}^{K}\exp(f(y_i))} = \frac{1}{n}\sum_{i=1}^{n} -f(y_i) + \log\left[\sum_{i=1}^{K}\exp(f(y_i))\right]$$

$$\text{therefore, } \frac{\partial L}{\partial f(y_i)} = \frac{\exp(f(y_i))}{\sum_{i=1}^{K}\exp(f(y_i))} - \frac{1}{n}\mathbb{1}(Y=y_i) = p(Y=y_i|x_i) - \frac{1}{n}\mathbb{1}(Y=y_i)$$

Apply the chain rule:

- $\frac{\partial L}{\partial f} = \begin{pmatrix} \frac{\partial L}{\partial f_1} \\ \vdots \\ \frac{\partial L}{\partial f_K} \end{pmatrix}$ $K \times 1$
- $\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial f} \cdot h_2^T \in \mathbb{R}^{K \times H_2}$
- $\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial f} \in \mathbb{R}^K$
- $\frac{\partial L}{\partial W_2} = \mathbb{1}(h_2>0) \cdot \frac{\partial L}{\partial h_2} \cdot h_3^T = \mathbb{1}(h_2>0) \cdot W_3^T \cdot \frac{\partial L}{\partial f} \cdot h_1^T$
- $\frac{\partial L}{\partial b_2} = \mathbb{1}(h_2>0)\frac{\partial L}{\partial h_2} = \mathbb{1}(h_2>0) \cdot W_3^T \cdot \frac{\partial L}{\partial f}$
- $\frac{\partial L}{\partial W_1} = \mathbb{1}(h_1>0)\frac{\partial L}{\partial h_3} \cdot X^T = \mathbb{1}(h_1>0)\mathbb{1}(h_2>0) W_2^T \cdot \frac{\partial L}{\partial h_2} \cdot X^T$
- $= \mathbb{1}(h_1>0)\mathbb{1}(h_2>0) W_2^T W_3^T \frac{\partial L}{\partial f} X^T$
- $\frac{\partial L}{\partial b_1} = \mathbb{1}(h_1>0)\frac{\partial L}{\partial h_1} = \mathbb{1}(h_1>0)\mathbb{1}(h_2>0) W_2^T \cdot W_3^T \cdot \frac{\partial L}{\partial f}$

2 a) $h = \text{ReLU}(W_1 x + b_1)$, $\hat{x} = \text{ReLU}(W_2 h + b_2)$

$$L = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \text{ReLU}(W_2\,\text{ReLU}(W_1 x_i + b_1) + b_2)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x}\|^2$$

$$\frac{\partial L}{\partial \hat{x}} = \frac{1}{n}\sum_{i=1}^{n} 2(\hat{x}_i - x_i)$$

Apply the chain Rule:

- $\frac{\partial L}{\partial W_2} = \mathbb{1}(\hat{x} > 0) \cdot \frac{\partial L}{\partial \hat{x}} \cdot h^T$
- $\frac{\partial L}{\partial b_2} = \mathbb{1}(\hat{x} > 0) \cdot \frac{\partial L}{\partial \hat{x}}$
- $\frac{\partial L}{\partial W_1} = \mathbb{1}(h > 0) \frac{\partial L}{\partial h} \cdot X^T = \mathbb{1}(h > 0)\mathbb{1}(\hat{x} > 0) W_2^T \frac{\partial L}{\partial \hat{x}} \cdot X^T$
- $\frac{\partial L}{\partial b_1} = \mathbb{1}(h > 0) \frac{\partial L}{\partial h} = \mathbb{1}(h > 0)\mathbb{1}(\hat{x} > 0) W_2^T \cdot \frac{\partial L}{\partial \hat{x}}$