Vaibhav Nandkumar Kadam
vkadam@wpi.edu

Aadesh Varude
avarude@wpi.edu

Q1 . Solution.

$$J(\theta) = \frac{1}{4}\sum(f^*(x) - f(x,\theta))^2$$

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

$$W = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \text{ and we have, } f^*(x) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

We know $y = W^T X + b$ can be represented as $y = W^T X$ with b augmented in W, Thus we have,

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$W = \begin{bmatrix} w_1 & w_2 & b \end{bmatrix} \text{ and we have, } y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

So we get,

$$W = (X^T X)^{-1} X y$$

$$(X^T X) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 1 & 0 & -1/2 \\ 0 & 1 & -1/2 \\ -1/2 & -1/2 & 3/4 \end{bmatrix}$$

$$Xy = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$W = (X^T X)^{-1} X y = \begin{bmatrix} 1 & 0 & -1/2 \\ 0 & 1 & -1/2 \\ -1/2 & -1/2 & 3/4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1/2 \end{bmatrix}$$

Thus we get $w_1 = 0$, $w_2 = 0$ and $b = 0.5$

For Q 2a and Q 2b please refer last part of the document .

Q2 c. Solution. Training Loss will converge as gradient descent will try minimize log loss as far as the function is differentiable and will satisfies the constraints for optimization.

Q2 d. Solution.

Testing loss convergence significantly depends on whether it is overfitted. Due to overfitting the testing loss would increase. Thus it highly depends on the type and amount of training data its being trained on. Thus the testing loss may or may not converge.

Q3 . Solution.

Deriving $\nabla_{w^{(l)}} z_k$

$$\nabla_{w^{(l)}} z_k = \frac{\partial}{\partial w^{(l)}}[x^T w^k + b_k] \tag{1}$$

$$\nabla_{w^{(l)}} z_k = x^{(i)T} \qquad \text{for } l = k \tag{2}$$

$$\nabla_{w^{(l)}} z_k = 0 \qquad \text{for } l \neq k \tag{3}$$

Deriving $\nabla_{w^{(l)}}[\sum_{k'=1}^{c} e^{z_k}]$

$$\nabla_{w^{(l)}}[\sum_{k'=1}^{c} e^{z_k}] = e^{z_k} \nabla_{w^{(l)}} z_1 \cdot \cdot = e^{z_k} \nabla_{w^{(l)}} z_c \tag{4}$$

$$\begin{aligned} \nabla_{w^{(l)}}[\sum_{k'=1}^{c} e^{z_k}] &= e^{z_k} x^{(i)} \qquad \text{for } k \neq l \\ &= e^{z_l} x^{(i)} \qquad \text{for } k = l \end{aligned} \tag{5}$$

Deriving $\nabla_{w^{(l)}} \hat{y_k}$ for $k = l$

$$\nabla_{w^{(l)}} \hat{y_k} = \frac{e^{z_k}}{\sum_{k'=1}^{c} e^{z_k}} \tag{6}$$

By $\frac{u}{v}$ derivative of division rule.

$$\nabla_{w^{(l)}} \hat{y_k} = \frac{\sum_{k=1}^{c} e^{z_k}[\nabla_{w^{(l)}} e^{z_k}] - e^{z_k}[\nabla_{w^{(l)}}[\sum_{k=1}^{c} e^{z_k}]]}{[\sum_{k=1}^{c} e^{z_k}]^2}$$

From eq. 5

$$\begin{aligned} &= \frac{\sum_{k=1}^{c} e^{z_k}[e^{z_l} x^{(i)}] - e^{z_l}[e^{z_l} x^{(i)}]}{[\sum_{k=1}^{c} e^{z_k}]^2} \\ &= \frac{e^{z_l} x^{(i)}}{\sum_{k=1}^{c} e^{z_k}} - \frac{e^{z_l}}{\sum_{k=1}^{c} e^{z_k}} \frac{e^{z_l}}{\sum_{k=1}^{c} e^{z_k}} x^{(i)} \\ &= \hat{y_l}^{(i)} x^{(i)} - (\hat{y_l}^{(i)})^2 x^{(i)} \end{aligned}$$

$$\boxed{\nabla_{w^{(l)}} \hat{y_k} = \hat{y_l}^{(i)} x^{(i)} (1 - \hat{y_l}^{(i)}) \text{for } l = k} \tag{7}$$

Hence proved.

Deriving $\nabla_{w^{(l)}} \hat{y_k}$ for $k \neq l$

$$\nabla_{w^{(l)}} \hat{y}_k = \frac{\sum_{k=1}^{c} e^{z_k} [\nabla_{w^{(l)}} e^{z_k}] - e^{z_k} [\nabla_{w^{(l)}} [\sum_{k=1}^{c} e^{z_k}]]}{[\sum_{k=1}^{c} e^{z_k}]^2}$$

From eq. 5

$$\begin{aligned}
&= \frac{-e^{z_k}[e^{z_l} x^{(i)}]}{[\sum_{k=1}^{c} e^{z_k}]^2} \\
&= \frac{-e^{z_k}}{\sum_{k=1}^{c} e^{z_k}} \frac{e^{z_l}}{\sum_{k=1}^{c} e^{z_k}} x^{(i)} \\
&= -\hat{y_k}^{(i)} \cdot \hat{y_l}^{(i)} x^{(i)}
\end{aligned}$$

$$\boxed{\nabla_{w^{(l)}} \hat{y}_k = -\hat{y_k}^{(i)} \cdot \hat{y_l}^{(i)} x^{(i)} \text{for } l \neq k} \tag{8}$$

Hence proved.

Deriving $\nabla_{w^{(l)}} f_{CE}(W, b)$

$$\begin{aligned}
\nabla_{w^{(l)}} f_{CE}(W, b) &= \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k{}^{(i)} \nabla_{w^{(l)}} \log \hat{y_k}^{(i)} \\
&= \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} y_k{}^{(i)} \cdot \frac{1}{\hat{y_k}^{(i)}} \cdot \nabla_{w^{(l)}} \hat{y_k}^{(i)} \\
&= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} \cdot \frac{1}{\hat{y_l}^{(i)}} \cdot \nabla_{w^{(l)}} \hat{y_l}^{(i)} + \sum_{k \neq l} y_k^{(i)} \cdot \frac{1}{\hat{y_k}^{(i)}} \nabla_{w^{(l)}} y_k^{(i)} \right] \\
&= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} \cdot \frac{1}{\hat{y_l}^{(i)}} \cdot \left(x^{(i)} y_l^{(i)} \left(1 - \hat{y_l}^{(i)}\right) \right) + \sum_{k \neq l} y_k^{(i)} \cdot \frac{1}{\hat{y_k}^{(i)}} \cdot \left(-\hat{y_l}^{(i)} \hat{y_k}^{(i)} x^{(i)} \right) \right] \\
&= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} \left(x^{(i)} - \left(1 - \hat{y_l}^{(i)}\right) + \sum_{k \neq l} y_k^{(i)} \left(-\hat{y_k}^{(i)} x^{(i)} \right) \right) \right] \\
&= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} x^{(i)} - y_l^{(i)} \hat{y_l}^{(i)} x^{(i)} - \sum_{k \neq l} y_k^{(i)} y_k^{(i)} \hat{y_l}^{(i)} x^{(i)} \right] \\
&= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} x^{(i)} - \sum_{k=1}^{c} y_k^{(i)} \hat{y_l}^{(i)} x^{(i)} \right] \\
&= \frac{-1}{n} \sum_{i=1}^{n} \left[y_l^{(i)} x^{(i)} - x^{(i)} \hat{y_l}^{(i)} \sum_{k=1}^{c} y_k^{(i)} \right]
\end{aligned}$$

We know $\sum_{k=1}^{c} y_k^{(i)} = 1$
Thus

$$\boxed{\nabla_{w^{(l)}} f_{CE}(W, b) = \frac{-1}{n} \sum_{i=1}^{n} x^{(i)} \left[y_l^{(i)} - \hat{y_l}^{(i)} \right]} \tag{9}$$

Deriving $\nabla_b f_{CE}(W, b)$

$$
\begin{aligned}
\boldsymbol{\nabla}_b f_{CE}(W,b) &= \frac{-1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k{}^{(i)} \boldsymbol{\nabla}_b \log \hat{y_k}{}^{(i)} \\
&= \frac{-1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k{}^{(i)} \cdot \frac{1}{\hat{y_l}{}^{(i)}} \boldsymbol{\nabla}_b \hat{y_l}{}^{(i)} \\
&= \frac{-1}{n}\sum_{i=1}^{n}\left[y_l^{(i)} \cdot \frac{1}{\hat{y_l}{}^{(i)}} \cdot \boldsymbol{\nabla}_{b^{(l)}} \hat{y_l}{}^{(i)} + \sum_{k\neq l} y_k^{(i)} \cdot \frac{1}{\hat{y_k}{}^{(i)}} \boldsymbol{\nabla}_{b^{(l)}} y_k^{(i)} \right]
\end{aligned}
$$

Similarly we can derive $\boldsymbol{\nabla}_b[\sum_{k'=1}^{c} e^{z_k}]$ and $\boldsymbol{\nabla}_b y_k^{(i)}$

$$
\begin{aligned}
\boldsymbol{\nabla}_b \hat{y_k} &= \frac{\sum_{k=1}^{c} e^{z_k}[\boldsymbol{\nabla}_b e^{z_k}] - e^{z_k}[\boldsymbol{\nabla}_b[\sum_{k=1}^{c} e^{z_k}]]}{[\sum_{k=1}^{c} e^{z_k}]^2} \\
&= \hat{y_l}{}^{(i)} - (\hat{y_l}{}^{(i)})^2
\end{aligned}
$$

$$
\boldsymbol{\nabla}_{w^{(l)}} \hat{y_k} = \hat{y_l}{}^{(i)}(1 - \hat{y_l}{}^{(i)}) \quad \text{for } l = k \tag{10}
$$

Therefore further deriving same as $\boldsymbol{\nabla}_{w^{(l)}} f_{CE}(W,b)$ for $\boldsymbol{\nabla}_b f_{CE}(W,b)$ we get

$$
\begin{aligned}
\boldsymbol{\nabla}_b f_{CE}(W,b) &= \frac{-1}{n}\sum_{i=1}^{n}\left[y_l^{(i)} - \hat{y_l}{}^{(i)} \sum_{k=1}^{c} y_k^{(i)} \right] \\
\boldsymbol{\nabla}_b f_{CE}(W,b) &= \frac{-1}{n}\sum_{i=1}^{n}\left[y_l^{(i)} - \hat{y_l}{}^{(i)} \right]
\end{aligned}
$$

$$
\boxed{\boldsymbol{\nabla}_b f_{CE}(W,b) = \frac{-1}{n}\sum_{i=1}^{n}\left[y_l^{(i)} - \hat{y_l}{}^{(i)} \right]} \tag{11}
$$

Q4 . Solution.
The unregularized cross-entropy on the test set is 0.47814 loss is and per cent correctly classified examples 83.78% (0.8378).

Q2a and Q2b. Solution.

P.T.O

Monday, February 6, 2023 3:07 PM

Q2

a) For the bias term :-

We can find the gradient of the log loss as $y=1$ as all the training examples are +ve.

$\therefore$ log loss $= -\left[y \log \hat{y} - (1-y) \log(1-\hat{y})\right]$

$\therefore \nabla_b f_{loss} = \nabla_b \left[-(1)\log \hat{y} - 0 \cdot \log(1-\hat{y})\right]$

$\therefore \nabla_b f_{loss} = \nabla_b \left[-\log \hat{y}\right]$

$= -\nabla_b \log \sigma(x^T w + b)$

$= -\dfrac{\sigma(x^T w+b)(1-\sigma(x^T w+b))}{\sigma(x^T w+b)}$

$\therefore \nabla_b f_{loss} = (\sigma(x^T w+b) - 1)$

1) Now, if we assume $w = 0$

$\nabla_b f_{loss} = \sigma(b) - 1$

The bias update is

$b_{new} = b_{old} - [\text{learning rate} \times (\sigma(b)-1)]$

$b_{new} = b_{old} - \alpha(\sigma(b)-1)$

As $0 < \sigma(b) < 1 \quad \therefore \quad \sigma(b)-1 < 0$; let $\alpha(b)-1 = a$

$b_{new} = b_{old} + \alpha \times a$

If we assume b convergence as 10000 using the above equation we won't be able to provide an upper bound for iterations.

2) If $w \neq 0$

$\nabla_b f_{loss} = \sigma(x^T w+b) - 1$

The bias update is:

$b_{new} = b_{old} - \left(\left[\sigma(x^T w+b) - 1\right] \times \alpha\right)$

Now here we can see that the convergence of bias will depend on the values of the training example as well.

$\therefore$ The convergence of the bias will depend and cannot be guaranteed.

Monday, February 6, 2023 5:43 PM

b) For the weight vector :-

As i) we can calculated the gradient of loss function

$\nabla_w f_{loss} = x(\hat{y} - y)$

$\nabla_w f_{loss} = x[\hat{y} - y] = x[\hat{y} - 1]$

$\nabla_w f_{loss} = x[\sigma(x^T w + b) - 1]$

Assume $b = 0$ for simplicity.

$\nabla_w f_{loss} = x[\sigma(x^T w) - 1]$

Let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$; $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

$\nabla_w f_{loss} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} [\sigma([x_1 \; x_2]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}) - 1]$

$\nabla_w f_{loss} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} [\sigma(w_1 x_1 + w_2 x_2) - 1]$

Update for w

$w = w_0 - \alpha \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} [\sigma(x_1 w_1 + w_2 x_2) - 1]$

Now value of $\sigma(x_1 w_1 + w_2 x_2) - 1 < 0$

$w = w_0 - \alpha_{new} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $\quad \alpha_{new} = \alpha \times a$; $a = [\sigma(x_1 w_1 + x_2 w_2) - 1]$

This equation denotes that the convergence of w is dependent on the exact values of X.

If $x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ it will never converge.

If $x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ or some suitable value it might converge.

∴ Convergence of w depends on the value of training examples.

Figure 2: Q 2b solution