Report for Deep Learning (CS541)_Homework2:

Report By:
Swapneel Dhananjay Wagholikar (WPI ID: 257598983)
Anuj Pradeep Pai Raikar (WPI ID: 181758784)

Solution to Question 2:

We have executed grid search on following values of hyperparameters:

Number of epochs = [100,250,500,1000]
Learning Rates = [0.0001,0.0005,0.0008,0.0006]
Alphas = [0.5,2,5,10]
Mini batch sizes = [25, 50, 75, 100]

We found the best sets of hyperparameters as:

[epochs, learning rate, alpha, mini batch size] = [1000, 0.0006, 10, 25]

Its corresponding half fMSE was 108.33326793577211

The unregularized half fMSE on testing dataset is 130.40361309865452

```
Lowest Fmse for epochs (m) 1000  Learning rate 0.0006 alpha 10  mini_batch_size 25
Min_FMSE 108.33326793577211
Lowest Fmse for epochs (m) 1000  Learning rate 0.0006 alpha 10  mini_batch_size 50
Min_FMSE 111.46059695275072
Lowest Fmse for epochs (m) 1000  Learning rate 0.0006 alpha 10  mini_batch_size 75
Min_FMSE 115.36119579828805
Lowest Fmse for epochs (m) 1000  Learning rate 0.0006 alpha 10  mini_batch_size 100
Min_FMSE 117.77291528770985
fMSE on Testing dataset is:  130.40361309865452
```

PAGE NO
DATE / /

Q.1] XOR problem:

$$J(\theta) = \frac{1}{4} \sum_{x \in X} \left(f^*(\theta) - f(x;\theta)\right)^2$$

$$= \frac{1}{4} \Sigma (x^T w + b - y)^2 = \frac{1}{4} \Sigma (y - (x^T w + b))^2$$

According to XOR truth table,

x_1	x_2	y
0	0	0
1	0	1
0	1	1
1	1	0

$$\therefore x = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$\therefore J(\theta) = \frac{1}{4} \sum_{x \in X} \left(\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \begin{bmatrix} b \\ b \\ b \\ b \end{bmatrix} \right)^2$$

$$= \frac{1}{4} \sum_{x \in X} \left(\begin{bmatrix} -b \\ 1 - w_2 - b \\ 1 - w_1 - b \\ -w_1 - w_2 - b \end{bmatrix} \right)^2$$

$$= \frac{1}{4} \left[(-b)^2 + (1 - w_2 - b)^2 + (1 - w_1 - b)^2 + (-w_1 - w_2 - b)^2 \right]$$

By expanding all individual square brackets

$$J(\theta) = \frac{w_1^2}{2} + \frac{w_2^2}{2} + b^2 + \frac{w_1 w_2}{2} + w_2 b + w_1 b - \frac{w_2}{2} - \frac{w_1}{2} - b + \frac{1}{2}$$

Taking gradient,

$$\Delta_{w_1} J(\theta) = w_1 + \frac{w_2}{2} + b - \frac{1}{2} = 0$$

$$\therefore 2w_1 + w_2 + 2b - 1 = 0 \longrightarrow \text{(i)}$$

$$\Delta_{w_2} J(\theta) = \frac{w_1}{2} + w_2 + b - \frac{1}{2} = 0$$

$$\therefore w_1 + 2w_2 + 2b - 1 = 0 \longrightarrow \text{(ii)}$$

$$\Delta_b J(\theta) = w_1 + w_2 + 2b - 1 = 0 \longrightarrow \text{(iii)}$$

by putting eq^n (iii) in eq^n (ii),

$$w_2 + 0 = 0$$

$$\therefore \boxed{w_2 = 0}$$

by putting eq^n (iii) in eq^n (i),

$$w_1 + 0 = 0$$

$$\therefore \boxed{w_1 = 0}$$

Substituting w_1 & w_2 in (iii),

$$2b - 1 = 0$$

$$\therefore \boxed{b = \frac{1}{2}}$$

$\therefore$ For minimizing J, the values are

$$w_1 = w_2 = 0$$

$$b = \frac{1}{2}$$

Q.3] (a) To prove : $\sigma(-x) = 1 - \sigma(x) \quad \forall x$

when $\sigma(x) = \dfrac{1}{1+e^{-x}}$

Proof :

$$1 - \sigma(x)$$

$$= 1 - \frac{1}{1+e^{-x}}$$

$$= \frac{1+e^{-x}-1}{1+e^{-x}}$$

$$= \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{x}} = \sigma(-x)$$

$$\boxed{\therefore 1 - \sigma(x) = \sigma(-x)}$$

(b) To prove : $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)) \quad \forall x$

Proof :

$$\sigma'(x) = \frac{\partial (1+e^{-x})^{-1}}{\partial x}$$

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

$$= \frac{e^{-x}}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}}$$

$$= \frac{1}{1+e^{x}} \cdot \frac{1}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$

$$\boxed{\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)}$$

Q.4] L_2 penalty term $= \frac{\alpha}{2} w^T w = \frac{\alpha}{2} w^T I w$

For this case, penalty term $= \frac{1}{2} \cdot \alpha \, w^T S w$

where $S = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

By logic,
if weights are symmetric, penalty term $\downarrow$
if weights are asymmetric, penalty term $\uparrow$
if weights are equal $(w_1 = w_2)$, penalty term $= 0$

Considering example of (1×2) image,
$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$

Now, penalty term

$$= \frac{\alpha}{2} w^T S w$$

$$= \frac{\alpha}{2} \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= \frac{\alpha}{2} \begin{bmatrix} w_1 a + w_2 c & w_1 b + w_2 d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= \frac{\alpha}{2} (w_1^2 a + w_1 w_2 c + w_1 w_2 b + w_2^2 d)$$

$$= \frac{\alpha}{2} (w_1 - w_2)^2 \quad \left(\because \text{ if we consider } a = d = 1, b = c = -1 \right)$$

This will be the term satisfying all conditions of symmetric, asymmetric & equal weights.

$$\Rightarrow \boxed{S = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}}$$

Q.5] $P(y|x) = N(\mu = x^T w, \sigma^2)$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y - x^T w)^2}{2\sigma^2}\right)$$

$\rightarrow P(D|w, \sigma^2) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}, w, \sigma^2)$

(∵ applying argmax)

$$\therefore \log P(D|w, \sigma^2) = \log \prod_{i=1} P(y^{(i)}|x^{(i)}, w, \sigma^2)$$

$$= \sum_{i=1}^{n} \log P(y^{(i)}|x^{(i)}, w, \sigma^2)$$

$$= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y^{(i)} - x^{(i)T} w)^2}{2\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sigma} - \frac{(y^{(i)} - x^{(i)T} w)^2}{2\sigma^2}\right]$$

Taking gradient,

$$\nabla_w = \sum_{i=1}^{n} -\frac{(y^{(i)} - x^{(i)T} w)}{\sigma^2} \cdot (-x) = 0$$

$$\therefore XY = XX^T w$$

$$\boxed{\therefore w = (XX^T)^{-1} XY}$$

$$\nabla_\sigma = \sum_{i=1}^{n} -\frac{1}{\sigma} + \sum_{i=1}^{n} \frac{(y - x^T w)^2}{\sigma^3}$$

$$= -\frac{n}{\sigma} + (y - x^T w)^2 \cdot \frac{n}{\sigma^3} = 0$$

$$\sigma^2 n + \Sigma(y - x^T w)^2 = 0$$

$$\therefore \boxed{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x^T w - y)^2}$$