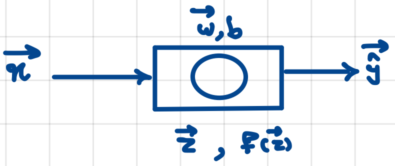
Mahdi Tabatabaei - 400101515

Deep Learning - Assignment 2

Dr. Fatemizadeh

Q1



$$\vec{z} = \vec{w}\vec{x} + b \quad , \quad \sigma(\vec{z}) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\vec{w}\vec{x}+b)}}$$

$$E_{w,b} = -\sum y_n \ln \hat{y}(x_n) + (1-y_n)\ln(1-\hat{y}_n(x_n))$$

To show that $E_{w,b}$ has minimum, we have to show it is convex.
As we know sum of convex functions is convex $\rightarrow$ we have to show convexity of $-(y_n \ln \hat{y}_n + (1-y_n)\ln(1-\hat{y}))$

$$E = -(y\ln\hat{y} + (1-y)\ln(1-\hat{y})) \quad , \quad \hat{y} = \sigma(z) \rightarrow E = -(y \ln(\sigma(z)) + (1-y)\ln(1-\sigma(z)))$$

$$\frac{\partial E}{\partial z} = \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \rightarrow \begin{cases} \dfrac{\partial E}{\partial \hat{y}} = -\dfrac{y}{\hat{y}} + \dfrac{1-y}{1-\hat{y}} \\ \dfrac{\partial \hat{y}}{\partial z} = \sigma(z)(1-\sigma(z)) = \hat{y}(1-\hat{y}) \end{cases} \rightarrow \frac{\partial E}{\partial z} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)\hat{y}(1-\hat{y}) = \hat{y} - y$$

$$\frac{\partial^2 E}{\partial z^2} = \frac{d}{dz}(\hat{y}-y) = \hat{y}(1-\hat{y}) \geqslant 0 \rightarrow E \text{ is convex} \rightarrow E_{w,b} \text{ is convex} \rightarrow E_{w,b} \text{ has minimum.}$$

We use gradient descent as a recursive rule for update.

$$\frac{\partial E_{w,b}}{\partial w} = \frac{\partial E_{w,b}}{\partial \hat{y}_n} \cdot \frac{\partial \hat{y}_n}{\partial z} \cdot \frac{\partial z}{\partial w} \qquad \frac{\partial E_{w,b}}{\partial b} = \frac{\partial E_{w,b}}{\partial \hat{y}_n} \cdot \frac{\partial \hat{y}_n}{\partial z} \cdot \frac{\partial z}{\partial b} \rightarrow \begin{cases} \dfrac{\partial E_{w,b}}{\partial \hat{y}} = -\sum \dfrac{y_n}{\hat{y}_n} - \dfrac{1-y_n}{1-\hat{y}_n} \\ \dfrac{\partial \hat{y}_n}{\partial z} = \hat{y}_n(1-\hat{y}_n) \\ \dfrac{\partial \hat{y}_n}{\partial w} = x_n \ , \ \dfrac{\partial \hat{y}_n}{\partial b} = 1 \end{cases}$$

$$\Rightarrow \frac{\partial E_{w,b}}{\partial w} = \sum_n (\hat{y}_n - y_n)x_n \quad , \quad \frac{\partial E_{w,b}}{\partial b} = \sum_n (\hat{y}_n - y_n)$$

* η is learning rate.

$$\begin{cases} w = w - \eta \dfrac{\partial E_{w,b}}{\partial w} \rightarrow w := w - \eta \sum_n (\hat{y}_n - y_n)x_n \\ b = b - \eta \dfrac{\partial E_{w,b}}{\partial b} \rightarrow b := b - \eta \sum_n (\hat{y}_n - y_n) \end{cases}$$

Q2

A) Covariate shift happens when the input distribution changes as data flows through different layers in neural network. This makes it harder for the network to learn because each layer has to adapt to the changing distribution.
Batch Normalization $\rightarrow$ normalize the input of each layer to have a consistent distribution. by keeping the distribution stable, it makes learning easier and faster.

B) BN can act as a regularization term, indirectly (mean and var of each mini-batch is different from the whole dataset $\rightarrow$ cause a small amount of noise) $\rightarrow$ prevent overfitting.
It can cause smoother learning and faster convergence $\rightarrow$ good for generalization.

C)
$$\hat{x}_i = x_i - \mu \qquad \mu = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$y_i = \gamma \hat{x}_i + \beta = \gamma\left(x_i - \frac{1}{n}\sum x_j\right) + \beta \rightarrow \frac{\partial y_i}{\partial x_j} = \begin{cases} \gamma(1-1/n) & i=j \\ -\gamma/n & i \neq j \end{cases}$$

$$\frac{\partial L}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i} = \frac{\partial L}{\partial y_i}(\gamma(1-1/n)) + \sum_{j \neq i} \frac{\partial L}{\partial y_j}(-\gamma/n) = \gamma\left(\frac{\partial L}{\partial y_i}(1-1/n) - 1/n \sum_{j \neq i}^{n} \frac{\partial L}{\partial j}\right)$$

D) $n=1$: $\dfrac{\partial L}{\partial x_1} = -\gamma \dfrac{\partial L}{\partial y_1}$ $\rightarrow$ makes the gradient of Loss simpler, and less stable due to lack of averaging.

$n \rightarrow \infty$: $\dfrac{\partial L}{\partial x_i} = \gamma \dfrac{\partial L}{\partial y_i}$ $\rightarrow$ The influence of each individual input on the batch mean become smaller effect of centering the batch is less significant for each x_i.

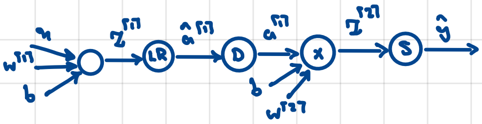
Q3

First layer $\rightarrow \begin{cases} z^{(1)} = w^{(1)}x + b^{(1)} \\ \hat{a}^{(1)} = \text{Leaky Relu}(z^{(1)}, \alpha = 0.1) \\ a^{(1)} = \text{Dropout}(\hat{a}^{(1)}, p = 0.2) \end{cases}$

Output layer $\rightarrow \begin{cases} z^{(2)} = w^{(2)}a^{(1)} + b^{(2)} \\ \hat{y} = \text{Softmax}(z^{(2)}) \end{cases}$

Cross-entropy loss $\rightarrow L = \sum_{i=1}^{K} -y_i \log(\hat{y}_i)$

A) $\dfrac{\partial y_k}{\partial z_i^{(2)}} = ?$ $\qquad \hat{y}_k = \dfrac{e^{z_k^{(2)}}}{\sum_{j=1}^{K} e^{z_j^{(2)}}}$



if $k = i \rightarrow \dfrac{\partial \hat{y}_k}{\partial z_k^{(2)}} = \dfrac{e^{z_k^{(2)}}(1 - \sum e^{z_k^{(2)}})}{(\sum e^{z_k^{(2)}})^2} = \hat{y}_k(1 - \hat{y}_k)$

if $k \neq i \rightarrow \dfrac{\partial \hat{y}_k}{\partial z_i^{(2)}} = \dfrac{-e^{z_i^{(2)} + z_k^{(2)}}}{(\sum e^{z_k^{(2)}})^2} = -\hat{y}_k \hat{y}_i$

$\rightarrow \dfrac{\partial \hat{y}_k}{\partial z_i^{(2)}} = \begin{cases} \hat{y}_k(1 - \hat{y}_k) & k = i \\ -\hat{y}_k \hat{y}_i & k \neq i \end{cases} = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & -\hat{y}_1\hat{y}_2 & \cdots & -\hat{y}_1\hat{y}_k \\ -\hat{y}_2\hat{y}_1 & & & \\ \vdots & & \ddots & \\ -\hat{y}_k\hat{y}_1 & & & \hat{y}_k(1-\hat{y}_k) \end{bmatrix}$

B) $y = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \rightarrow$ k-th element $\rightarrow L = -\log(\hat{y}_k)$

$i = k \quad \dfrac{\partial L}{\partial z_k^{(2)}} = \dfrac{\partial L}{\partial \hat{y}_k} \cdot \dfrac{\partial \hat{y}_k}{\partial z_k^{(2)}} = -\dfrac{1}{\hat{y}_k} \cdot \hat{y}_k(1 - \hat{y}_k) = \hat{y}_k - 1$

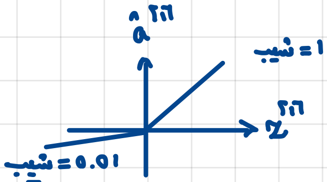
$i \neq k \quad \dfrac{\partial L}{\partial z_i^{(2)}} = \dfrac{\partial L}{\partial \hat{y}_k} \cdot \dfrac{\partial \hat{y}_k}{\partial z_i^{(2)}} = -\dfrac{1}{\hat{y}_k} \cdot (-\hat{y}_k \hat{y}_i) = \hat{y}_i$

$\rightarrow \dfrac{\partial L}{\partial z_i^{(2)}} = \begin{cases} \hat{y}_i - 1 & i = k \\ \hat{y}_i & i \neq k \end{cases} = \hat{y} - \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \hat{y} - y$

C) $\dfrac{\partial L}{\partial w^{[1]}} = \dfrac{\partial z^{[1]}}{\partial w^{[1]}} \times \dfrac{\partial \hat{a}^{[1]}}{\partial z^{[1]}} \times \dfrac{\partial a^{[1]}}{\partial \hat{a}^{[1]}} \times \dfrac{\partial z^{[2]}}{\partial a^{[1]}} \times \underbrace{\dfrac{\partial \hat{y}}{\partial z^{[2]}} \times \dfrac{\partial L}{\partial \hat{y}}}_{\hat{y} - y}$

$z^{[2]} = w^{[2]} a^{[1]} + b \rightarrow \dfrac{\partial z^{[2]}}{\partial a^{[1]}} = w^{[2]T}$

$\dfrac{\partial a^{[1]}}{\partial \hat{a}^{[1]}} = pI = 0.2I \qquad \hat{a}^{[1]} = \text{Leaky Relu}(z^{[1]}, \alpha = 0.01)$



$\dfrac{\partial \hat{a}^{[1]}}{\partial z^{[1]}} = \begin{bmatrix} I(z_1^{[1]} > 0) + \alpha I(z_1^{[1]} < 0) & & \\ & \ddots & \\ & & I(z_m^{[1]} > 0) + \alpha I(z_m^{[1]} < 0) \end{bmatrix} = A$

$z^{[1]} = w^{[1]} x + b \rightarrow \dfrac{\partial z^{[1]}}{\partial w^{[1]}} = x^T$

$\dfrac{\partial L}{\partial w^{[1]}} = pA\, w^{[2]T} (\hat{y} - y) x^T$

due to dimension.

Q4

$\nabla y(u,v,z) = \begin{bmatrix} \partial y / \partial u \\ \partial y / \partial v \\ \partial y / \partial z \end{bmatrix} \xrightarrow{J} \begin{bmatrix} \dfrac{\partial^2 y}{\partial u^2} & \dfrac{\partial^2 y}{\partial v \partial u} & \dfrac{\partial^2 y}{\partial z \partial u} \\ \dfrac{\partial^2 y}{\partial u \partial v} & \dfrac{\partial^2 y}{\partial v^2} & \dfrac{\partial^2 y}{\partial z \partial v} \\ \dfrac{\partial^2 y}{\partial z \partial u} & \dfrac{\partial^2 y}{\partial v \partial z} & \dfrac{\partial^2}{\partial z^2} \end{bmatrix}$

$H(y(u,v,z)) = \begin{bmatrix} \dfrac{\partial^2 y}{\partial u^2} & \dfrac{\partial^2 y}{\partial v \partial u} & \dfrac{\partial^2 y}{\partial z \partial u} \\ \dfrac{\partial^2 y}{\partial u \partial v} & \dfrac{\partial^2 y}{\partial v^2} & \dfrac{\partial^2 y}{\partial z \partial v} \\ \dfrac{\partial^2 y}{\partial z \partial u} & \dfrac{\partial^2 y}{\partial v \partial z} & \dfrac{\partial^2}{\partial z^2} \end{bmatrix}$ which are the same.

Q5

$$J_1 = 0.5\left(y_d - \sum_{k=1}^{n} \delta_k W_k x_k\right)^2 =$$

$\delta_k \sim \text{Normal}(1, \omega^2)$ $\quad \mathbb{E}(\nabla J_1) = ?$ $\quad \omega^2 = \mathbb{E}\{\delta_k^2\} - \mathbb{E}\{\delta_k\}^2 \rightarrow \mathbb{E}\{\delta_k^2\} = \omega^2 + 1$

$$\frac{\partial J_1}{\partial w_i} = \frac{1}{2} \times 2 \times \delta_i x_i \left(y_d - \sum_{k=1}^{n} \delta_k W_k x_k\right) \longrightarrow \mathbb{E}\left\{\frac{\partial J_1}{\partial w_i}\right\} = -x_i y_d\, \mathbb{E}\{\delta_i\} - \mathbb{E}\{-\delta_i x_i \textstyle\sum \delta_k w_k x_k\}$$

$$-x_i y_d\, \mathbb{E}\{\delta_i\} = -x_i y_d$$

$$\mathbb{E}\{-\delta_i x_i \textstyle\sum \delta_k w_k x_k\} = \{\delta_i^2 x_i^2 w_i + \displaystyle\sum_{\substack{k \neq i \\ k=1}}^{n} \delta_i \delta_k w_k x_i x_k\} = (\omega^2+1) x_i^2 w_i + \sum_{\substack{k \neq i \\ i=1}}^{n} w_k x_i x_k$$

$$\rightarrow \mathbb{E}\left\{\frac{\partial J_1}{\partial w_i}\right\} = -x_i y_d + \omega^2 x_i^2 w_i + \sum_{\substack{k=1 \\ k \neq i}}^{n} w_k x_i x_k$$

without Drop-out $\rightarrow J = \frac{1}{2}\left(y_d - \sum_{k=1}^{n} w_k x_k\right)^2 \rightarrow \frac{\partial J}{\partial w_i} = -y_d x_i + \sum_{k=1}^{n} w_k x_k x_i$

$$\frac{\partial J_1}{\partial w_i} = -x_i y_d + \sum_{k=1}^{n} w_k x_i x_k + \omega^2 x_i^2 w_i = \frac{\partial J}{\partial w_i} + \omega^2 x_i^2 w_i$$

$$\rightarrow J_1 = J + \underbrace{\tfrac{1}{2} \omega^2 \textstyle\sum x_k^2 w_k^2}_{\text{Regularization Term}}$$

Q6

$f(x) = g'(x)$

$$x_{k+1} = x_k - \frac{g'(x)}{g''(x)} = x_k - \frac{f(x)}{f'(x)}$$

$$e_k = x_k - x^* \Rightarrow x^* = x_k - e_k$$

$$f(x^*) = f(x_k - e_k) \simeq f(x_k) - e_k f'(x_k) + \frac{e_k^2}{2} f''(\xi_k) \quad \text{which} \quad x_k < \xi_k < x^*$$

$$f(x^*) = 0 \rightarrow f(x_k) - e_k f'(x_k) + \frac{e_k^2}{2} f(\xi_k) = 0 \xrightarrow{1/f'} \frac{f(x_k)}{f'(x_k)} - e_k + \frac{e_k^2}{2f'(x_k)} f''(\xi_k) = 0 \rightarrow x_k - x_{k-1} - x_k + x^* + \frac{e_k^2}{2} \frac{f''(\xi_k)}{f'(x_k)} = 0$$

$$\rightarrow x_{k-1} - x^* = \frac{1}{2}(x_k - x^*)^2 \frac{f''(\xi_k)}{f'(x_k)} \rightarrow |x_{k-1} - x^*| \leq \frac{|f''(\xi_k)|}{2|f'(x_k)|} |x_k - x^*|^2$$

$k \rightarrow \infty$: $f'(x_k) \rightarrow f'(x^*)$

$$\xi_k \rightarrow x^* \Rightarrow f''(\xi_k) \rightarrow f''(x^*) \Rightarrow |x_{k-1} - x^*| \leq M |x_k - x^*|^2$$

$\hookrightarrow M > \frac{|f''(x_k)|}{2|f'(x_k)|}$

Q7

A) $\hat{y}_i = \dfrac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$

$$\mathcal{L}(z,y) = -\sum_{k=1}^{K} y_k \log\left(\frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}\right) = -\sum_{k=1}^{K} y_k \left(z_k - \ln \sum_{j=1}^{K} e^{z_j}\right)$$

$$\frac{\partial L}{\partial z_i} = -y_i + \sum_{k=1}^{K} y_k \frac{e^{z_i}}{\sum e^{z_j}} = -y_i + \frac{e^{z_i} \overbrace{\sum y_k}^{1}}{\sum_{j=1}^{k} e^{z_j}} = -y_i + \hat{y}_i \longrightarrow \nabla_z L = \hat{y} - y$$

$$\nabla_z L = \nabla_z (\hat{y} - y) = \frac{\partial \hat{y}}{\partial z} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial z_1} & \frac{\partial \hat{y}_2}{\partial z_1} & \cdots \\ \vdots & \ddots & \end{bmatrix}$$

$$\frac{\partial \hat{y}_i}{\partial z_i} = \frac{e^{z_i}(\sum e^{z_i}) - e^{z_i} e^{z_i}}{(\sum e^{z_i})^2} = \hat{y}_i - \hat{y}_i^2 = \hat{y}_i(1-\hat{y}_i)$$

$$H = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & -\hat{y}_1\hat{y}_2 & \cdots & \hat{y}_1\hat{y}_k \\ \vdots & & & \\ \hat{y}_k\hat{y}_1 & \cdots & & \hat{y}_k(1-\hat{y}_k) \end{bmatrix} = \text{diag}(y) - \hat{y}\hat{y}^T$$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{e^{-z_j} e^{-z_i}}{(\sum e^{z_j})^2} = \hat{y}_i y_i$$

B) PSD $\longrightarrow u^T H u \geq 0 \longrightarrow v^T(\text{diag}(\hat{y}) - \hat{y}\hat{y}^T)u \geq 0 \longrightarrow v^T \text{diag}(\hat{y}) v = \sum_{i=1}^{k} \hat{y}_i v_i^2$, $\hat{y}_i \geq 0 \to v^T \text{diag}(\hat{y}) v \geq 0$

$\longrightarrow v^T \hat{y} \hat{y}^T v = (v^T \hat{y})^2$ cauchy-shwartz $\longrightarrow (a^T b)^2 \leq (a^T a)(b^T b)$

$\longrightarrow v^T \hat{y} = v^T \hat{y}^{1/2} \hat{y}^{1/2} = a^T b \longrightarrow a = [v_1\sqrt{\hat{y}_1}, v_2\sqrt{\hat{y}_2}, \dots, v_k\sqrt{\hat{y}_k}]$
, $b = [\sqrt{y_1}, \sqrt{y_2}, \dots, \sqrt{y_k}]$

$$\longrightarrow (v^T \hat{y})^2 \leq \left(\sum_{i=1}^{k} \hat{y}_i v_i^2\right)\underbrace{\left(\sum_{i=1}^{k} \hat{y}_i\right)}_{1} = \sum_{i=1}^{k} \hat{y}_i v_i^2$$

$\longrightarrow (v^T \hat{y})^2 \leq \sum_{i=1}^{k} \hat{y}_i v_i^2 \longrightarrow v^T H v \geq 0 \longrightarrow H$ is PSD.

C) H is PSD $\longrightarrow$ convex