Thalagala B. P.   180631J | EN4553 - Assignment 2.

Question 01

ⓐ Assuming this is a Regression Problem and, that we use only one datum at a time during training.

$$Loss = \frac{1}{2}(y-t)^2 \leftarrow \text{Squared Error Loss.}$$

ⓑ Forward Pass:-

1	8	0.5	6
3	1.1	-1	-8
-0.5	-6	-3	-1.1
-1	-8	-0.5	-6

Maxpooling →

8	6
-0.5	-0.5

X

(i) ↓ Maxpooling

8	6
-0.5	-0.5

: values of the matrix U.

(ii) ↓ LReLU(.) activation function. $= \begin{cases} x & \text{if } x \geq 0 \\ 0.1x & \text{otherwise.} \end{cases}$

8	6
-0.05	-0.05

: Values of the matrix V

(iii) ~~network output~~ y =

(iii) Network output $y = \sum_{i=1}^{4} v_i w_i$

$$W = \begin{bmatrix} 1 & 2 \\ 10 & 4 \end{bmatrix} \qquad y = (8 \times 1) + (6 \times 2) + (-0.05) \times (10) + (-0.05 \times 4)$$

$$= 8 + 12 + (-0.5) + (-0.2)$$

$$\boxed{y = 19.3}$$

(c) Backward Pass :-

Assumption $\Rightarrow$ $\frac{\partial L}{\partial y} = 2.0$ $\left[\text{where} : L = \frac{1}{2}(y-t)^2 \right]$

(i) Consider forward pass data flow

$$X \longrightarrow \text{Maxpool}() \longrightarrow u \longrightarrow \text{LReLU}() \longrightarrow V \longrightarrow \underbrace{\text{Dense}}_{(V.W)} \rightarrow Y \Rightarrow \text{Loss}$$

$\frac{\partial L}{\partial y} \leftarrow$

* Using chain Rule

$$\frac{\partial L}{\partial v} = \frac{\partial L}{\partial y} \times \frac{\partial y}{\partial v}$$

$$= \frac{\partial L}{\partial y} \times \frac{\partial (y = W.V)}{\partial v}$$

$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial y} \times W$$

∴ By Applying elementwise :

$$\frac{\partial L}{\partial v_1} = \frac{\partial L}{\partial y} \times w_1$$

$$= 2 \times 1$$

$$\frac{\partial L}{\partial v_1} = 2$$

(ii) $$\frac{\partial L}{\partial v} = \begin{array}{|c|c|} \hline \frac{\partial L}{\partial v_1} = 2 \times 1 & \frac{\partial L}{\partial v_2} = 2 \times 2 \\ \hline \frac{\partial L}{\partial v_3} = 2 \times 10 & \frac{\partial L}{\partial v_4} = 2 \times 4 \\ \hline \end{array} = \begin{array}{|c|c|} \hline 2 & 4 \\ \hline 20 & 8 \\ \hline \end{array}$$

(iii) Using the chain Rule recursively:

$$\frac{\partial L}{\partial u} = \frac{\partial L}{\partial V} \times \frac{\partial V}{\partial U}$$

$$\downarrow$$

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial V} \times \frac{\partial [LReLU(u)]}{\partial U}$$

Since $V = LReLU(U)$

$u \geq 0$ | $u < 0$

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial V} \times \frac{\partial (V = u\cancel{m})}{\partial u}$$

$$\downarrow$$

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial V} \times \cancel{m}\, 1$$

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial V} \times \frac{\partial (V = 0.1 U)}{\partial U}$$

$$\downarrow$$

$$\frac{\partial L}{\partial U} = \frac{\partial L}{\partial V} \times 0.1$$

$\therefore$
$$\frac{\partial L}{\partial U} = \begin{cases} \frac{\partial L}{\partial V} & ; \text{for } U \geq 0 \\ 0.1 \frac{\partial L}{\partial V} & ; \text{otherwise} \end{cases}$$

$\therefore$
$$\frac{\partial L}{\partial U} = \begin{array}{|c|c|} \hline \frac{\partial L}{\partial V} & \frac{\partial L}{\partial V} \\ \hline 0.1 \frac{\partial L}{\partial V} & 0.1 \frac{\partial L}{\partial V} \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|} \hline 2 & 4 \\ \hline 2 & 0.8 \\ \hline \end{array}$$

(iv) $\dfrac{\partial L}{\partial x} = \dfrac{\partial L}{\partial U} \times \dfrac{\partial U}{\partial x}$ $\quad U = \max(x)$

$$\frac{\partial L}{\partial x} = \begin{cases} \dfrac{\partial L}{\partial U} \times 1 & \text{where } x \text{ is maximum ([illegible])} \\ \dfrac{\partial L}{\partial U} \times 0 & \text{otherwise.} \end{cases}$$

$\therefore \dfrac{\partial L}{\partial x} =$

0	2	0	4
0	0	0	0
2	0	0	0
0	0	0.8	0

Gradient becomes zero as there is no affect from those weights.

(d.)

(i)

$$U = \begin{bmatrix} 14 & -4 \\ -6 & 3 \end{bmatrix} \xrightarrow{\text{PReLU}} \begin{bmatrix} 14 & \alpha_0(-4) \\ \alpha_0(-6) & 3 \end{bmatrix} = \begin{bmatrix} 14 & -4\alpha_0 \\ -6\alpha_0 & 3 \end{bmatrix} = V \leftarrow \boxed{\text{Answer}}$$

$$\boxed{V_i = \begin{cases} u_i & \text{for } u_i \geq 0 \\ \alpha_0 u_i & \text{otherwise} \end{cases}}$$

(ii)

$$\frac{\partial L}{\partial V} = \begin{bmatrix} 2.0 & -2.5 \\ 1.5 & 3.0 \end{bmatrix}$$

* using chain rules $\quad \dfrac{\partial L}{\partial U} = \dfrac{\partial L}{\partial V} \times \dfrac{\partial V}{\partial U} = \begin{cases} \dfrac{\partial L}{\partial V} & \text{if } u \geq 0 \\ \alpha_0 \dfrac{\partial L}{\partial V} & \text{otherwise} \end{cases}$

* using chain rule again : $\quad \cancel{\dfrac{\partial L}{\partial \alpha} = \dfrac{\partial L}{\partial U} \times \dfrac{\partial V}{\partial \alpha}}$

$$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial V} \times \frac{\partial V}{\partial \alpha}$$

if $u_i \geq 0$ otherwise

$$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial V} \times \underbrace{\frac{\partial u}{\partial \alpha}}_{0} \qquad\qquad \frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial V} \times \frac{\partial(\alpha U_i)}{\partial \alpha}$$

$$\therefore \frac{\partial L}{\partial \alpha} = \begin{cases} 0 & \text{for } (u_i \geq 0) \\ \dfrac{\partial L}{\partial V} \times u_i & \text{otherwise} \end{cases} \Rightarrow \frac{\partial L}{\partial \alpha} = \begin{bmatrix} 0 & (-2.5 \times (-4)) \\ 1.5 \times (-6) & 0 \end{bmatrix}$$

$$\boxed{\text{Answer}} \rightarrow \text{So } \left.\frac{\partial L}{\partial \alpha}\right|_{\alpha_0} = \begin{bmatrix} 0 & 10 \\ -9 & 0 \end{bmatrix}$$

Question 02

(a) Definition of softmax = $p_i = \dfrac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)}$

(i)

consider

$$\text{Softmax}(\bar{z}_i) = \frac{\exp(z_i - z_{max})}{\sum_{j=1}^{k} \exp(z_j - z_{max})} \quad ; z_{max} = \max\{z_i\} \quad \forall i \in \{1, 2, \ldots k\}$$

$$= \frac{\exp(z_i)/\exp(z_{max})}{\sum_{j=1}^{k} \exp(z_j)/\exp(z_{max})}$$

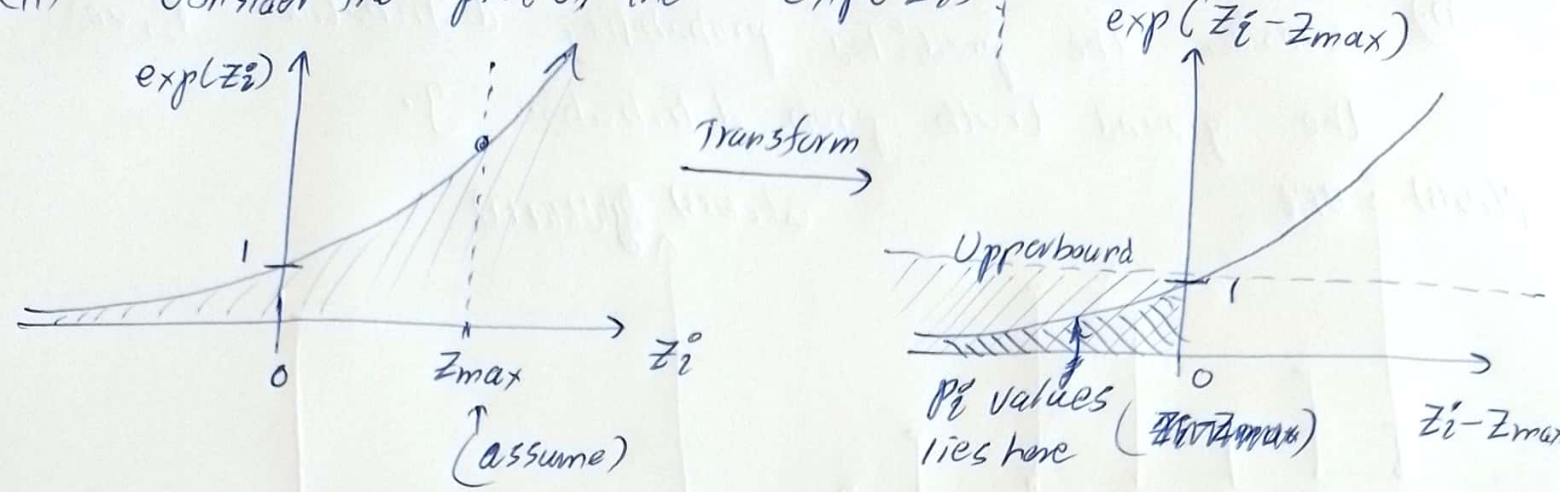
$$= \frac{\cancel{\left\{\frac{1}{\exp(z_{max})}\right\}} \times \exp(z_i)}{\cancel{\left\{\frac{1}{\exp(z_{max})}\right\}} \times \sum_{j=1}^{k} \exp(z_j)}$$

$$= \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)} = \text{softmax}(z_i)$$

$$\Downarrow$$

$\therefore$ $\text{Softmax}(z) = \text{Softmax}(\bar{z})$ //

(ii) Consider the plot of the $\exp(z_i)$



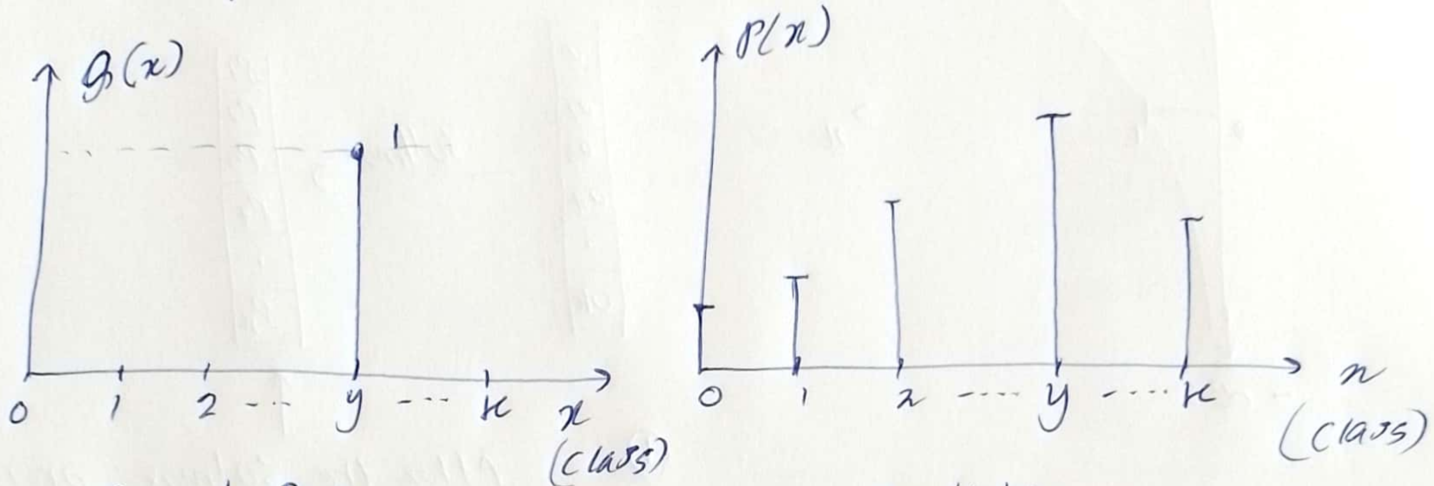
* Subtracting the max of the logit vector is done, to "improve the numerical stability" ~~of~~ when training deep learning models.

* If we don't perform this transformation, the terms in the softmax function become very large leading to numerical overflow.

* This can cause issues when training the model, as the gradient of loss function w.r.t the logit values may become very large or NaN (Not a Number).

(b)

(i) Consider the probability distributions of the Ground Truth and the predicted values.



Ground Truth.   Prediction

* Cross-entropy distance = $H(g,p) = -\sum_{x \in X} g(x) \cdot \log(P(x))$.

* However, due to the nature of the ground truth:

(I) $g(x) = 0 \; : \; \forall \; x \in \{1, 2, 3, \ldots k\} \setminus \{y\}$

(II) $g(y) = 1 \; :$ for the correct class

$\therefore$ Above summation can be further simplified as follows.

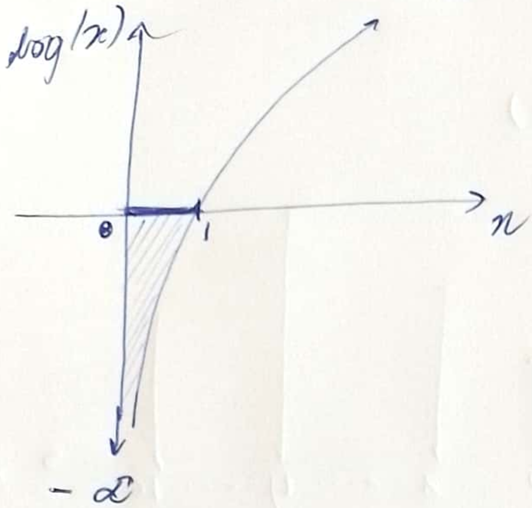
$$H(g,P) = [-\underbrace{g(1)}_{(=0)} \cdot \log(P_1)] + [-\underbrace{g(2)}_{(=0)} \log(P_2)] + \ldots + [-\underbrace{g(y)}_{(=1)} \log(P_y)] + \ldots + [-\underbrace{g(k)}_{=0} \log(P_k)]$$

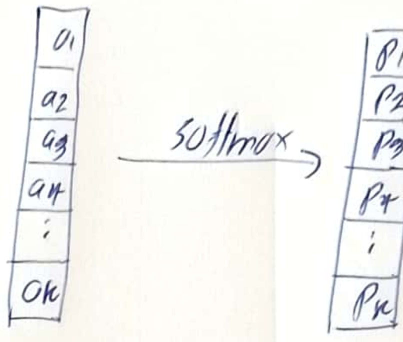
$$\therefore H(g,P) = L(g,P) = -\underbrace{g(y)} \cdot \log(P_y)$$

$$= -1 \times \log(P_y)$$

$$\therefore \underline{\underline{L(g,P) = -\log(P_y)}}$$

(ii.) consider the $\log(x)$ function:





① After the softmax operation all the probability values comes to the range of $[0,1]$.

② The corresponding (log) values for the range $[0,1]$ lies ~~between~~ in the range $(-\infty, 0]$.

③ consider the loss function $\left(\therefore -\log(P_y) \in [0, \infty)\right)$

$-\log(P_y)$

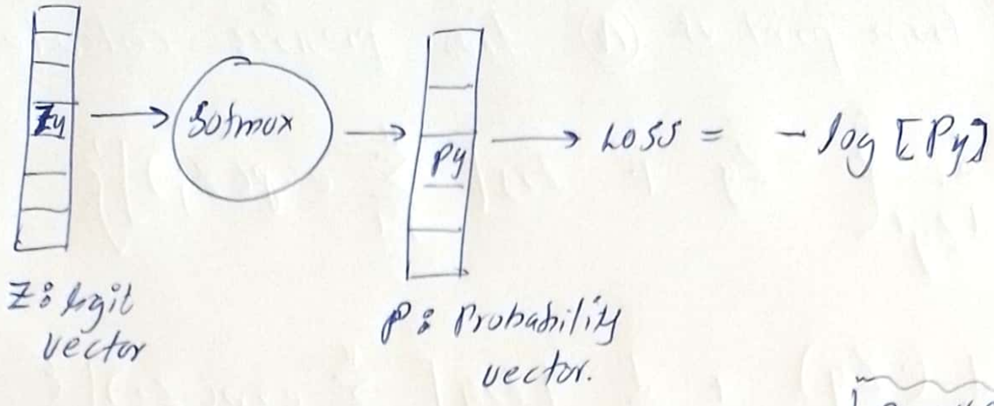
if $(P_y = 1)$ which is the final goal, $(-\log(1) = 0)$. Then the loss is zero.

∴ we try to minimize the $[-\log(P_y).]$

if $0 < P_y < 1$ $[-\log(P_y)]$ becomes a large positive value, making the loss very large.

∴ By making this loss minimum we can increase the Pr(correct class).

(c) (i)



Recall:

$$\frac{d}{dx} \ln(f(x)) = \frac{1}{f(x)} \times f'(x)$$

I. $$\frac{dL}{dz_y} = \frac{d}{dz_y}\left\{-\log[P_y]\right\}$$

$$= \frac{d}{dz_y}\left\{-\log\left[\frac{\exp(z_y)}{\sum_{j=1}^{K}\exp(z_j)}\right]\right\}$$

$$= \frac{d}{dz_y}\left\{-\log[\exp(z_y)] + \log\left[\sum_{j=1}^{K}\exp[z_j]\right]\right\}$$

$$= \frac{d}{dz_y}\left\{-z_y + \log\left[\sum_{j=1}^{K}\exp(z_j)\right]\right\}$$

$$= \frac{d}{dz_y}(-z_y) + \frac{d}{dz_y}\log\left[\sum_{j=1}^{K}\exp(z_j)\right] \longrightarrow A$$

$$= -1 + \frac{1}{\sum_{j=1}^{K}\exp(z_j)} \times \frac{d}{dz_y}\left[\sum_{j=1}^{K}\exp(z_j)\right]$$

$$= -1 + \frac{1}{\sum_{j=1}^{K}\exp(z_j)} \times \frac{d}{dz_y}\underbrace{\left[\exp(z_1) + \exp(z_2) + \dots + \exp(z_y) + \dots + \exp(z_K)\right]}_{\left(\frac{d}{dz_y}(\exp(z_i)) = 0 \ \forall\, i \neq y\right)}$$

$$\therefore \frac{dL}{dz_y} = -1 + \frac{\exp(z_y)}{\sum_{j=1}^{K}\exp(z_j)}$$

$$\boxed{\frac{dL}{dz_y} = -1 + P_y} \leftarrow \text{Answer}$$

(ii) Starting from point of (A) from previous calculations:

$$\frac{dL}{dz_{y'}} = \frac{d}{dz_{y'}}\left\{ -z_y + \log\left[\sum_{j=1}^{k} \exp(z_j)\right]\right\}$$

$$= \underbrace{\frac{d}{dz_{y'}}(-z_y)}_{=0} + \frac{d}{dz_{y'}}\left\{\log\left[\sum_{j=1}^{k} \exp(z_j)\right]\right\}$$

$$= \frac{1}{\sum_{j=1}^{k} \exp(z_j)} \times \frac{d}{dz_{y'}}\left\{\sum_{j=1}^{k} \exp(z_j)\right\}$$

$$= \frac{1}{\sum_{j=1}^{k} \exp(z_j)} \times \frac{d}{dz_{y'}}\{ \exp(z_1) + \exp(z_2) + \dots + \exp(z_{y'}) + \dots \exp(z_y) + \dots + \exp(z_n)\}$$

$$\left(\frac{d}{dz_{y'}} \exp(z_j) = 0 \quad \forall j \neq y' \right)$$

$$\therefore \frac{dL}{dz_{y'}} = \frac{1}{\sum_{j=1}^{k} \exp(z_j)} \times \exp(z_{y'})$$

(softmax for $z_{y'}$)

$$\Downarrow$$

$$\boxed{\therefore \frac{dL}{dz_{y'}} = P_{y'}} \quad \forall\ y' \in \{1, 2, \dots k\} \setminus \{y\}$$