Some university exam questions

Question 1

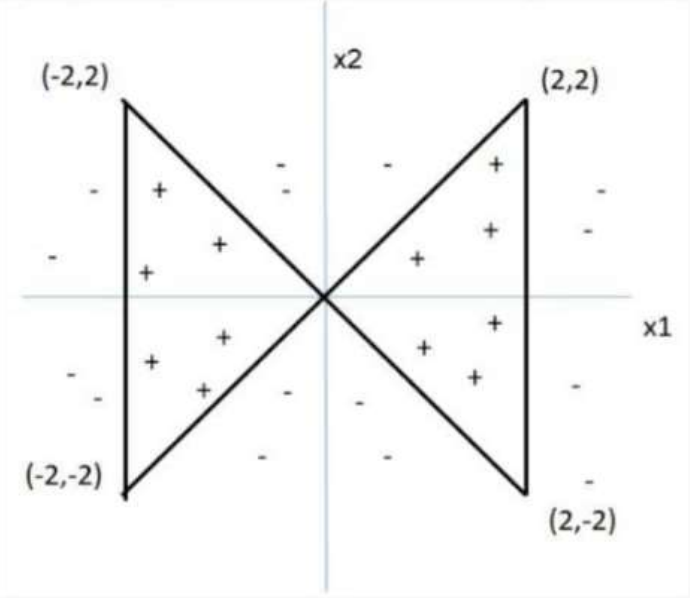
Qtext :

(x1, x2) are input features and target classes are either +1 or -1 as shown in the figure. **(Marks: 1+1+1+4=7)**

(A) What is the minimum number of hidden layers and hidden nodes required to classify the following dataset with 100% accuracy using a fully connected multilayer perceptron network? Step activation functions are used at all nodes, i.e., output=+1 if total weighted input >= bias b at a node, else output = -1.

(B) Show the minimal network architecture by organizing the nodes in each layer horizontally. Show the node representing x1 at the left on the input layer. Organize the hidden nodes in ascending order of bias at that node.

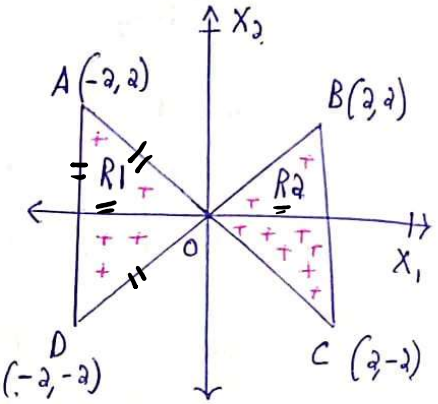
(C) Specify all weights and bias values at all nodes. Weights can be only -2.5, 2.5 or 0, and bias +ve/-ve multiples of 2.5.



We will consider R1 first.

R1 is made up of AD, AO & OD

equation of line passing through A(-2, 2) & D(-2, -2) will be $\boxed{x_1 = -2}$



equation of line passing through A(-2, 2) & O(0, 0) will be.

$$\frac{x_2 - 0}{x_1 - 0} = \frac{2 - 0}{-2 - 0}$$

$$\Rightarrow \quad -2x_2 = 2x_1$$

$$\Rightarrow \quad \boxed{x_1 + x_2 = 0}$$

equation of line passing through O(0, 0) & D(-2, -2) will be.

$$\frac{x_2 - 0}{x_1 - 0} = \frac{-2 - 0}{-2 - 0} \Rightarrow \boxed{x_1 - x_2 = 0}$$

We will now consider region R2. R2 is made of OB, BC & OC.

equation of CB is $\boxed{x_1 = 2}$

We will now consider O(0, 0) & C(2, -2). The equation of line through OC will be. $\frac{x_2 - 0}{x_1 - 0} = \frac{-2 - 0}{2 - 0} \Rightarrow -x_1 - x_2 = 0 \Rightarrow \boxed{x_1 + x_2 = 0}$

The equation of line through O(0, 0) & B(2, 2) is exactly the equation of line through OD. We get $\boxed{x_1 - x_2 = 0}$

The equation of lines is of form. $w_1x_1 + w_2x_2 + b = 0$.

Further, we are told weights can be only $-2.5, 2.5$ & 0.

& bias can be a +ve/-ve multiple of 2.5.

We reformulate the equations as below.

Line	region	equation	New equation.
OA	R1	$x_1 + x_2 = 0$.	$2.5x_1 + 2.5x_2 = 0$
OD	R1	$x_1 - x_2 = 0$	$2.5x_1 - 2.5x_2 = 0$
AD	R1	$x_1 = -2$.	$2.5x_1 + 0x_2 + 5 = 0$
OB	R2	$x_1 - x_2 = 0$.	$2.5x_1 - 2.5x_2 = 0$
OC	R2	$x_1 + x_2 = 0$.	$2.5x_1 + 2.5x_2 = 0$
BC	R2.	$x_1 = 2$	$2.5x_1 + 0x_2 - 5 = 0$

Decision rule = $\begin{cases} +1 & \text{if } w_1x_1 + w_2x_2 - b \geq 0 \\ -1 & \text{if } w_1x_1 + w_2x_2 - b < 0 \end{cases}$

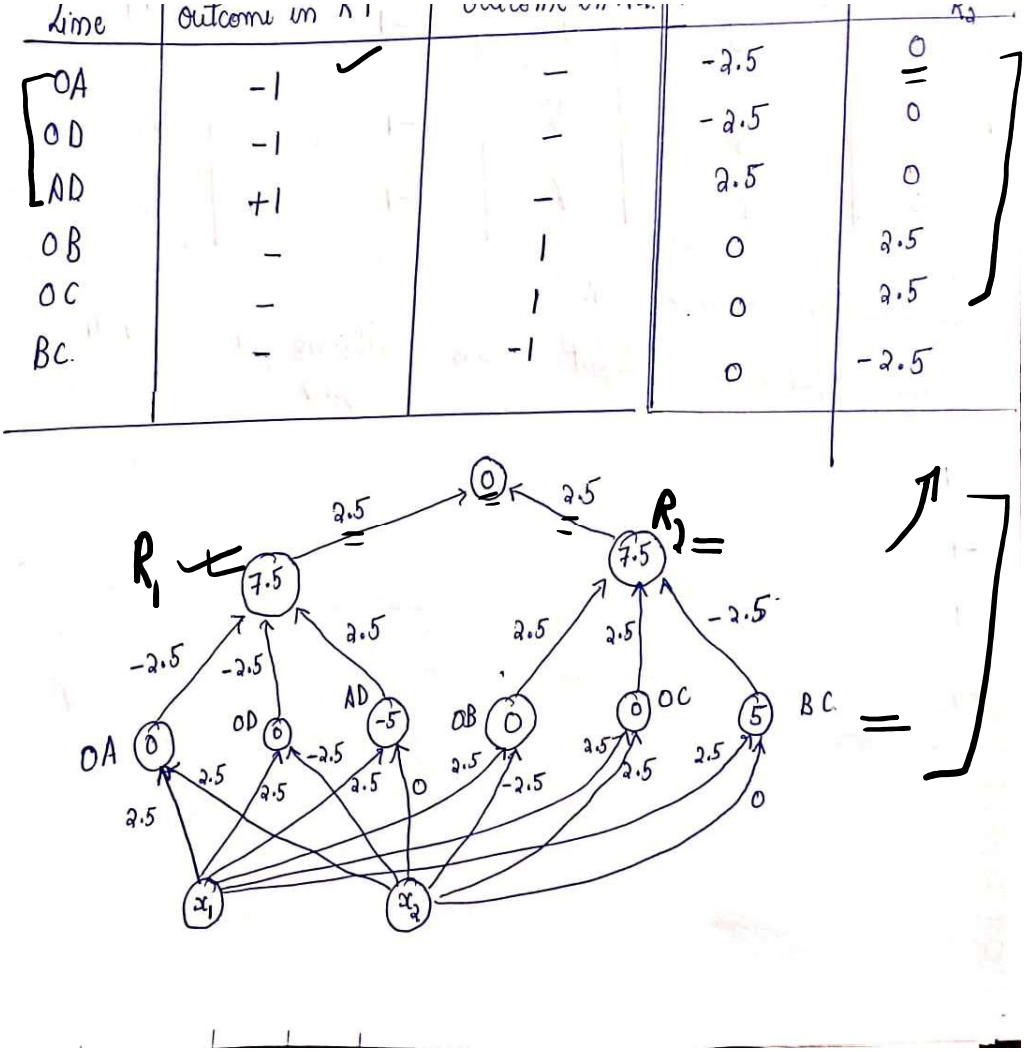
Line	equation	w_1	w_2	$-b$	Decision rule = $\begin{cases} +1 \text{ if } w_1x_1 + w_2x_2 - b \geq 0 \\ -1 \text{ else.} \end{cases}$
OA	$2.5x_1 + 2.5x_2 = 0$	2.5	2.5	0.	$\begin{cases} +1 \text{ if } 2.5x_1 + 2.5x_2 \geq 0 \\ -1 \text{ else.} \end{cases}$
OD	$2.5x_1 - 2.5x_2 = 0$	2.5	-2.5	0	$\begin{cases} +1 \text{ if } 2.5x_1 - 2.5x_2 \geq 0 \\ -1 \text{ else.} \end{cases}$
AD	$2.5x_1 + 5 = 0$	2.5	0	-5	$\begin{cases} +1 \text{ if } 2.5x_1 + 0x_2 - (-5) \geq 0 \\ -1 \text{ else.} \end{cases}$

} R_1

OB	$2.5x_1 - 2.5x_2 = 0$	2.5	-2.5	0	$\begin{cases} +1 \text{ if } 2.5x_1 - 2.5x_2 \geq 0 \\ -1 \text{ else.} \end{cases}$
OC.	$2.5x_1 + 2.5x_2 = 0$	2.5	2.5	0	$\begin{cases} +1 \text{ if } 2.5x_1 + 2.5x_2 \geq 0 \\ -1 \text{ else.} \end{cases}$
BC	$2.5x_1 - 5 = 0$	2.5	0	5	$\begin{cases} +1 \text{ if } 2.5x_1 + 0x_2 \geq 5 \\ -1 \text{ else.} \end{cases}$

} R_2

Line	outcome in R1	outcome in R2.	Recalculated R_1	Recalculated R_2
OA	-1 ✓	—	-2.5	0



Question2

2. Computational Graph (20 points)

$h = 2x + 1$

$z = x^2 + h$

$y = \frac{1}{1 + e^{-h}}$

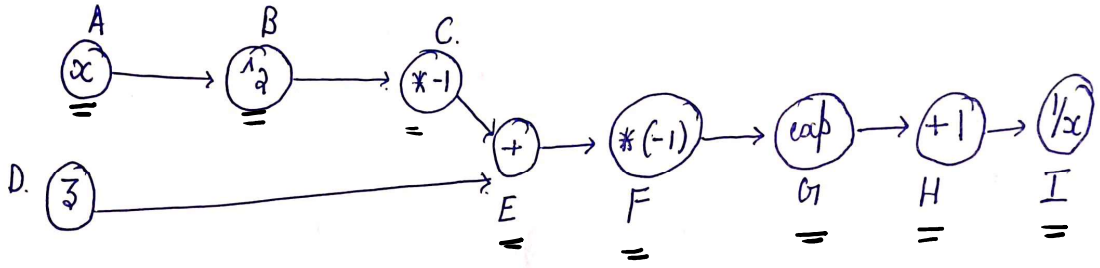
(1) Draw the computational graph based on the above three equations (1 point)

(2) What is $\frac{\partial y}{\partial z}$ from the graph? (19 points)

Let's have. $h = 2x + 1$ ✓

$z = x^2 + h$. ✓

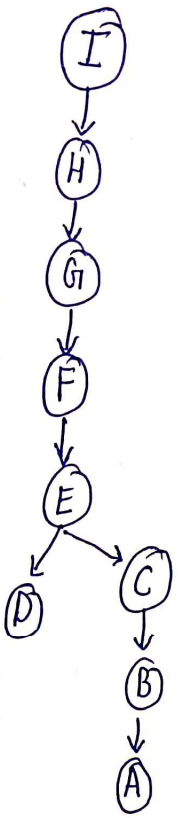
$y = \frac{1}{1+e^{-h}} = \frac{1}{1+\exp[-1 \times (z - x^2)]}$



Node	Value	Actual Value
A	x.	x. ✓
B	A^2.	x^2. ✓
C	$-B$.	$-x^2$. ✓
D	z	z.
E	$C + D$.	$z - x^2$.
F	$-E$	$-(z - x^2)$
G	$\exp(F)$	$\exp(-(z - x^2))$
H	$1 + G$.	$1 + \exp(-(z - x^2))$
I	$1/H$	$\frac{1}{1+\exp[-(z-x^2)]}$

We have to compute $\partial y/\partial z$ i.e. $\partial I/\partial D$ from our graph.

We get dependency graph showing dependency between node & its successors as below.



The path possible from I to D is

I H G F E D.

$$\Rightarrow \frac{\partial I}{\partial D} = \frac{\partial I}{\partial H}\frac{\partial H}{\partial G}\frac{\partial G}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial z}$$

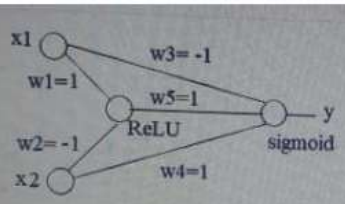
$$= \frac{-1}{H^2}(1)\,e^F(-1)(1)(1)$$

$$= \frac{e^F}{H^2}$$

$$= \frac{e^{-(z-x^2)}}{\left[1+e^{-(z-x^2)}\right]^2}$$

$$= \frac{e^{-h}}{\left[1+e^{-h}\right]^2} = \frac{\partial y}{\partial z}$$

Question 3



Hidden node and output node use, respectively, ReLU and sigmoid activation functions.

Bias values at hidden and output nodes are zero. Weights for the current iteration are given in the above figure. Target output d is specified as 0. Learning rate is 0.3.

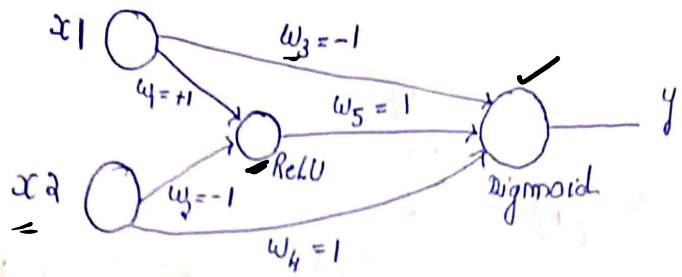
A. Calculate the actual output y for the current iteration with input (x1, x2) = (1,1).

B. Calculate the binary cross-entropy error for the current iteration.

C. Assuming L1 regularization constant = 0.2, calculate the new w_3 for the next iteration.

D. Assuming L2 regularization constant = 0.2, calculate the w_1 for the next iteration.

E. Assuming both L1 (with regularization constant=0.2) and L2 (regularization constant=0.2) are applied, calculate the value of w5 in next iteration.



Let us first write the forward propagation steps

$$z_1 = w_1 x_1 + w_2 x_2$$

$$a_1 = ReLU(z_1) = \begin{cases} 0 & \text{if } z_1 \le 0 \\ z_1 & \text{if } z_1 > 0 \end{cases}$$

$$z_2 = w_3 x_1 + w_4 x_2 + w_5 a_1$$

$$a_2 = \frac{1}{1 + \exp(-z_2)}$$

We have $x_1 = x_2 = 1$ & actual output $= 0$ [i.e. $d = 0$]

Loss function at terminal node $= -d \log a_2 - (1-d) \log(1-a_2)$

$$\boxed{L = -\log(1-a_2)}$$

$- y \log \hat{y} - (1-y) \log(1-\hat{y})$

Answer to part A)

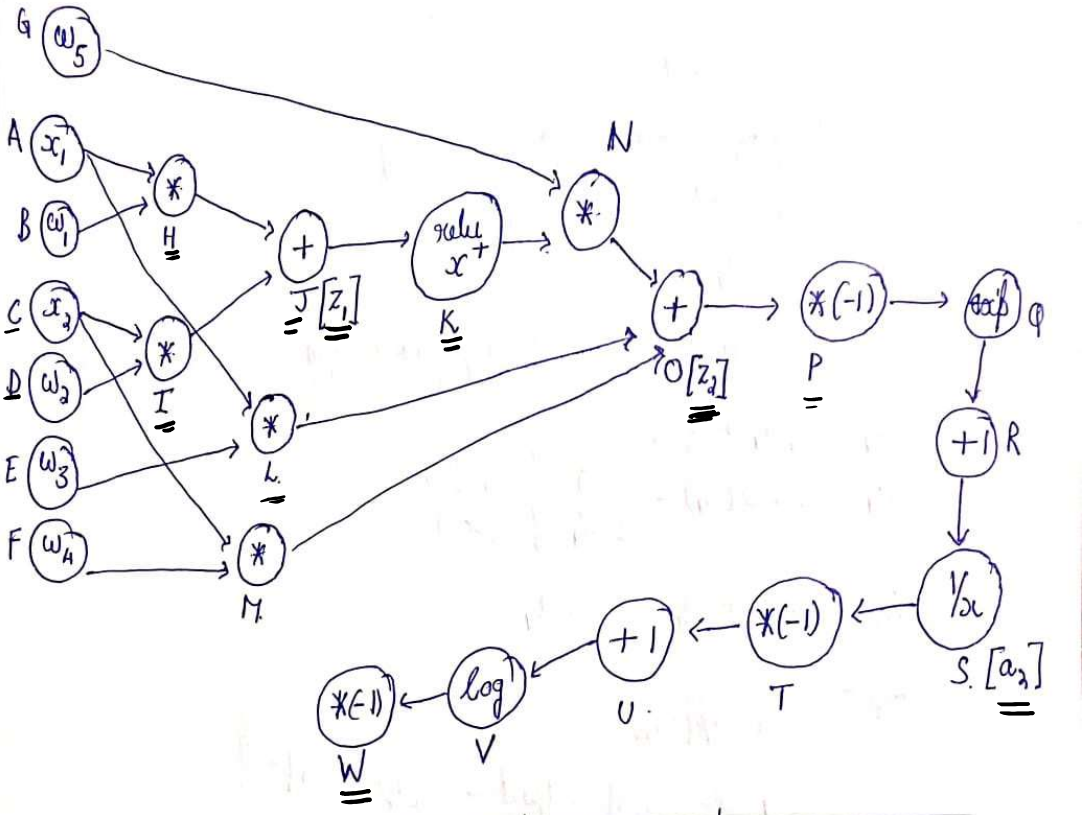
$z_1 = 1(1) - 1(1) = 0$

$a_1 = \text{relu}(0) = 0$

$z_2 = -1(1) + 1(1) + 0(1) = 0$

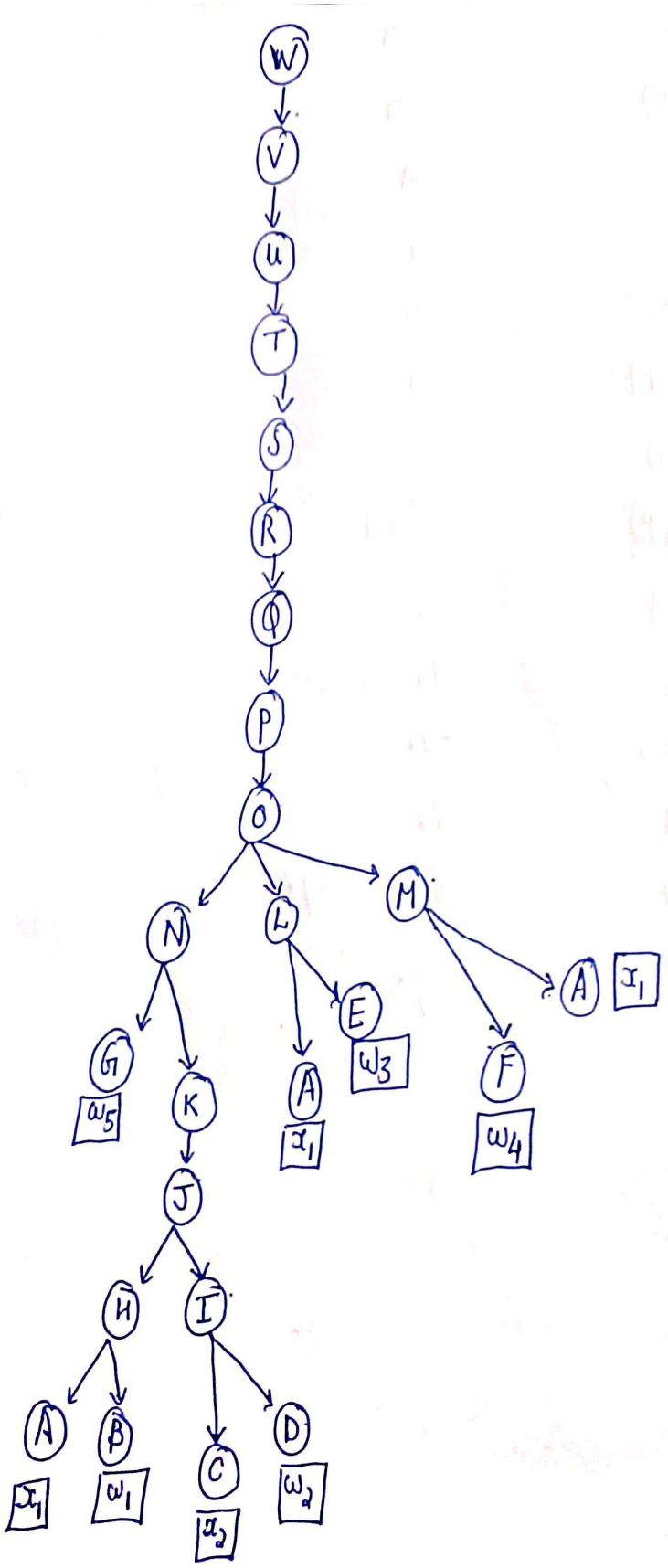
$a_2 = \frac{1}{1+e^{-0}} = \frac{1}{2}$

Answer to part B) $L = -\log\left(1 - \frac{1}{2}\right) = -\log \frac{1}{2} = \log 2$



Node	Value substituted (predecessor)	Value saved
A	x_1	1
B	w_1	1
C	x_2	1
D	w_2	-1
E	w_3	-1
F	w_4	1
G	w_5	1
H	A * B	1
I	C * D	-1

J	H + I	0
K	Rdu(J)	0.
L	A * E	-1
M	C * F	1
N	K * G	0
O	N + L + M	0
P	O * (-1)	0.
Q	exp(P)	$e^0 = 1$
R	1 + Q	2
S	1/R.	1/2
T	S * (-1)	-1/2.
U	1 + T	1/2.
V	log U.	log 1/2 = -log 2.
W	-1 * V.	log 2.


W
V
U
T
S
R
Q
P
O
N
L
M
G
ω_5
K
A
x_1
E
ω_3
A
x_1
F
ω_4
J
H
I
A
x_1
B
ω_1
C
ω_2
D
ω_2

Answer to Q C.

We have $\underline{w} = (w_1, w_2, \cdots w_5)$ ✓

$$l(w) = -\log 2.$$

$$l_R(w) = l(w) + \alpha |w|$$

Now $\frac{\partial}{\partial w_3} l_R(w) = \frac{\partial}{\partial w_3} l(w) + \alpha \, \text{sign}(w_3)$ $\left[\because \frac{\partial}{\partial w}|w| = \text{sign}(w)\right]$

We need to compute first $\frac{\partial}{\partial w_3} l(w) = \frac{\partial W}{\partial E}$

Path possible from W to E

= N V U T S R Q P O L E

$$\frac{\partial W}{\partial E} = \frac{\partial W}{\partial V}\frac{\partial V}{\partial U}\frac{\partial U}{\partial T}\frac{\partial T}{\partial S}\frac{\partial S}{\partial R}\frac{\partial R}{\partial Q}\frac{\partial Q}{\partial P}\frac{\partial P}{\partial O}\frac{\partial O}{\partial L}\frac{\partial L}{\partial E}\frac{\partial E}{\partial w_3}$$

$$= (-1)\frac{1}{U}(1)(-1)\left(\frac{-1}{R^2}\right) 1 \; e^P (-1)(1) A(1)$$

$$= \frac{e^0 (1)}{1/2 \; 2^2} = \frac{1}{2}$$

Updation rule is

$$w_3(t+1) = w_3(t) - \eta \frac{\partial}{\partial w_3} l_R(w)$$

$$= -1 - 0.3\left[\frac{1}{2} + \alpha(-1)\right]$$

$$= -1 - 0.3\left[\frac{1}{2} + (-0.2)\right]$$

$$= -1.09$$

Answer to Q1

We have $l_R(w) = l(w) + \frac{\alpha}{2}\|w\|_2^2$

$$\nabla l_R(w) = \nabla l(w) + \alpha' w$$

Now $\boxed{\frac{\partial l_R(w)}{\partial w_1} = \frac{\partial l(w)}{\partial w_1} + \alpha w_1}$ ✓

Now $\frac{\partial l(w)}{\partial w_1} = \frac{\partial W}{\partial B}$ ✓

Path possible from W to B

= W V U T S R Q P O N K J H B

$$\frac{\partial W}{\partial B} = \frac{\partial W}{\partial V}\frac{\partial V}{\partial U}\frac{\partial U}{\partial T}\frac{\partial T}{\partial S}\frac{\partial S}{\partial R}\frac{\partial R}{\partial Q}\frac{\partial Q}{\partial P}\frac{\partial P}{\partial O}\frac{\partial O}{\partial N}\frac{\partial N}{\partial K}\frac{\partial K}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial B}\frac{\partial B}{\partial w_1}$$

$$= (-1)\frac{1}{a}(1)(-1)\left(\frac{-1}{R^2}\right)(1)e^P(-1)(1)(1)(0)(1)(1)1$$

$$= 0 \qquad \left[\because \frac{\partial K}{\partial J} = \frac{\partial}{\partial J}\max(0, J) = 0 \text{ as } J = 0\right]$$

Now $\frac{\partial l_R(w)}{\partial w_1} = 0.2(1) = 0.2$

Now $w_1(t+1) = w_1(t) - \eta \frac{\partial l_R(w)}{\partial w_1}$

$$= +1 - 0.3(0.2)$$

$$= 0.94$$ ✓

Answer to Q 6

a) Let us consider l_2 regularization first

$$l_R(w) = l(w) + \frac{\alpha}{2}\|w\|_2^2$$

Now $$\boxed{\frac{\partial l_R(w)}{\partial w_5} = \frac{\partial l(w)}{\partial w_5} + \alpha(w_5)}$$

We need to compute $\frac{\partial l(w)}{\partial w_5} = \frac{\partial W}{\partial G_1}$

Path from W to G_1 = W V U T S R Q P O N G_1

$$\frac{\partial W}{\partial G_1} = \frac{\partial W}{\partial V}\frac{\partial V}{\partial u}\frac{\partial U}{\partial T}\frac{\partial T}{\partial S}\frac{\partial S}{\partial R}\frac{\partial R}{\partial \Phi}\frac{\partial \Phi}{\partial P}\frac{\partial P}{\partial O}\frac{\partial O}{\partial N}\frac{\partial N}{\partial G_1}\frac{\partial G_1}{\partial w_5}$$

$$= (-1)\frac{1}{u}\cdot(1)(-1)\left(\frac{-1}{R^2}\right)(1)\,e^P(-1)(1)\,K\,(1)$$

$$= \frac{e^0}{(1/2)}\frac{1}{2^2}(0) = 0 \qquad [\because K = 0]$$

b) If we consider l_1 regularization.

$$l_R(w) = l(w) + \alpha|w|$$

$$\Rightarrow \boxed{\frac{\partial l_R(w)}{\partial w_5} = \frac{\partial l(w)}{\partial w_5} + \alpha.}$$

We need to consider both regularizations

i.e. $l_R(w) = l(w) + \frac{\alpha}{2}\|w\|_2^2 + \alpha|w|$

Here $\frac{\partial l_R(w)}{\partial w_5} = \frac{\partial l(w)}{\partial w_5} + \alpha w_5 + \alpha.$

$$\frac{\partial l_R(w)}{\partial w_5} = 0 + \alpha(w_5 + 1)$$

Now $w_5(t+1) = w_5(t) - \eta \frac{\partial l_R(w)}{\partial w_5}$

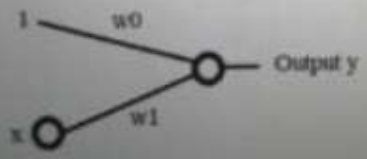
$= 1 - 0.3[0.2(1+1)]$

$= 0.88.$

Question 4:

Question The training data and network architecture are given below. Sigmoid activation function $y(z) = \frac{1}{1+e^{-z}}$ is used at the output node and binary cross-entropy is the loss function. Learning rate is 0.25 and momentum rate is 0.75. (w0, w1) at iteration t=4 is (0, 0) and at iteration t=5 is (-1, +1). Batch training is used.

Input x	Target Output d
0	0.0
1	1.0

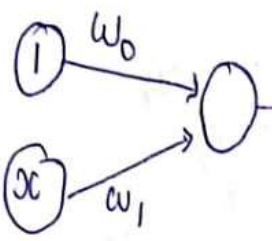


a Calculate the value of training loss at t=5

b Calculate the gradient of the training loss function at t=5

c Calculate the new values of w0 and w1 at iteration t=6 when momentum based gradient descent is used.

d Calculate the new values of w0 and w1 at iteration t=6 when Nestorov accelerated gradient descent is used. [1.5+1.5+1+2=6



[Actual value = d]

x_i	d_i

We have $z_i = w_0 + w_1 x_i$

x_i	d_i
0	0
1	1

We have y_i

$$y_i = \frac{1}{1+\exp(-w_0 - w_1 x_i)}$$

a) Cross entropy loss at any time t

$$L_t = \sum_i \{ d_i \log y_i - (1-d_i) \log (1-y_i) \}$$

$$= -d_0 \log y_0 - (1-d_0) \log (1-y_0) - d_1 \log y_1 - (1-d_1) \log (1-y_1)$$

Here $w_0 = -1$ & $w_1 = 1$ at $t = 5$.

$$\Rightarrow y_i(t) = \frac{1}{1+e^{-(-1+w_1 x_i)}} = \frac{1}{1+\exp(1-x_i)}$$

Now $y_0(t) = \frac{1}{1+e}$ & $y_1(t) = \frac{1}{1+e^0} = \frac{1}{2}$

$$\Rightarrow L_t = -(1-0) \log \left(1 - \frac{1}{1+e}\right) - 1 \log \frac{1}{2}$$

$$= -\log \left(\frac{e}{1+e}\right) + \log 2$$

x_i	y_i	d_i
0	$1/[1+\exp(-\omega_0-\omega_1(0))] = \frac{1}{1+e^{-\omega_0}}$	0
1	$\frac{1}{1+e^{-\omega_0-\omega_1(1)}} = \frac{1}{1+e^{-\omega_0-\omega_1}}$	1

$$Loss = -\log\left(1-\frac{1}{1+e^{-\omega_0}}\right) - \log\left(\frac{1}{1+e^{-\omega_0-\omega_1}}\right)$$

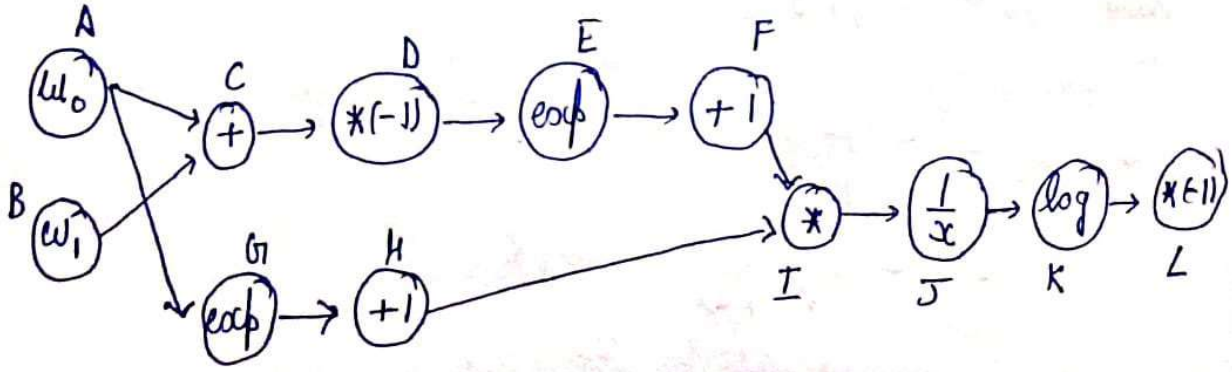
$$= -\log\left(\frac{e^{-\omega_0}}{1+e^{-\omega_0}}\right) - \log\left(\frac{1}{1+e^{-\omega_0-\omega_1}}\right)$$

$$= -\log\left[\frac{e^{-\omega_0}}{(1+e^{-\omega_0})(1+e^{-\omega_0-\omega_1})}\right]$$

$$L = -\log\left[\frac{1}{(e^{\omega_0}+1)(1+e^{-\omega_0-\omega_1})}\right]$$

[Multiplying denominator & Numerator by e^{ω_0}

We draw computational graph for the same.



A	w_0	w_0
B	w_1	w_1
C	A + B	$(w_0 + w_1)$
D	− C	$-(w_0+w_1)$
E	exp(D)	$\exp(-(w_0+w_1))$
F	1 + E	$1 + \exp(-(w_0+w_1))$
G	exp(A)	$\exp(w_0)$
H	1 + G	$1 + \exp(w_0)$
I	F * H	$[1+\exp(w_0)] * [1+\exp(-w_0-w_1)]$
J	1/I	$1/[[1+\exp(w_0)] * [1+\exp(-w_0-w_1)]]$
K	log J.	$\log[1/[(1+e^{w_0})(1+e^{-w_0-w_1})]]$
L.	− K.	$-\log\left[\frac{1}{(1+e^{w_0})(1+e^{-w_0-w_1})}\right]$ = loss.

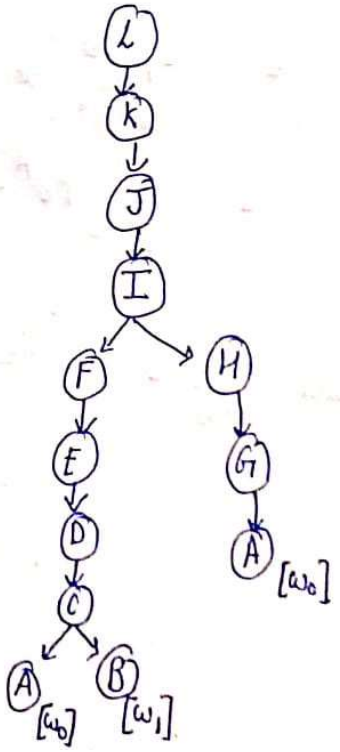
Now $\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial B}$

$= \frac{\partial L}{\partial k}\frac{\partial k}{\partial J}\frac{\partial J}{\partial I}\frac{\partial I}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial C}\frac{\partial C}{\partial B}\frac{\partial B}{\partial w_1}$

$-(-1)\frac{1}{J}\left(\frac{-1}{I^2}\right) H.(1)\, e^{D}(-1)(1)$

$= -\frac{He^{D}}{JI^2}$

$= -\frac{He^{D}}{I}$ $[\because J = 1/I]$



Also $\frac{\partial L}{\partial \omega_0} = \frac{\partial L}{\partial A}$

Paths from L to A -

Path 1:- L K J I F E D C A

Path 2:- L K J I H G A

$$\frac{\partial L}{\partial A} = \frac{\partial L}{\partial K}\frac{\partial K}{\partial J}\frac{\partial J}{\partial I}\frac{\partial I}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial C}\frac{\partial C}{\partial A}\frac{\partial A}{\partial \omega_0} +$$

$$\frac{\partial L}{\partial K}\frac{\partial K}{\partial J}\frac{\partial J}{\partial I}\frac{\partial I}{\partial H}\frac{\partial H}{\partial G}\frac{\partial G}{\partial A}\frac{\partial A}{\partial \omega_0}.$$

$$= (-1)\frac{1}{J}\left(\frac{-1}{I^2}\right) F\, H\, (1)\, e^D (-1)(1)(1) +$$

$$(-1)\frac{1}{J}\left(\frac{-1}{I^2}\right) F\, (1)\, e^A (1)$$

$$= \frac{-He^D}{JI^2} + \frac{Fe^A}{JI^2}$$

$$= \frac{1}{I}\left[Fe^A - He^D\right]$$

Now $\frac{\partial L}{\partial \omega} = \begin{pmatrix} \frac{Fe^A - He^D}{I} = \frac{\partial L}{\partial \omega_0} \\ \frac{-He^D}{I} = \frac{\partial L}{\partial \omega_1} \end{pmatrix}$

Answer to Q b)

At $t = 5$, $w_0 = -1$ & $w_1 = 1$

$$\Rightarrow \frac{\partial L}{\partial w_0} = \frac{Fe^A - He^D}{I} = \frac{(1+e^{-w_0-w_1})e^{w_0} - (1+e^{w_0})e^{-w_0-w_1}}{(1+e^{w_0})(1+e^{-w_0-w_1})}$$

$$= \frac{(1+e^0)e^{-1} - (1+e^{-1})e^0}{(1+e^{-1})(1+e^{-(0)})}$$

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

Now $\frac{\partial L}{\partial w_1} = \frac{-He^D}{I} = \frac{-(1+e^{w_0})e^{-(w_0+w_1)}}{(1+e^{w_0})(1+e^{-(w_0+w_1)})}$

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

Answer to Q e) for momentum based updation, we have following rule.

$$w_{t+1} = w_t + \beta \Delta w_t - \eta g_t \quad \text{where } g_t = \nabla L$$

Here $\Delta w_5 = w_5 - w_4 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$

Also, $\frac{\partial L}{\partial w_0}(\text{at } t=5) = \frac{-1}{2} + \frac{1}{1+e}$ & $g_5 = \begin{pmatrix} \frac{1}{1+e} - \frac{1}{2} \\ -1/2 \end{pmatrix}$

$\frac{\partial L}{\partial w_1}(\text{at } t=5) = \frac{-1}{2}$

Hence;

$$w_6 = w_5 + \beta \Delta w_5 - \eta g_5 \quad \left[\text{Here } \begin{matrix}\beta = 0.75 \\ \eta = 0.25\end{matrix}\right]$$

$$= \begin{pmatrix}-1\\1\end{pmatrix} + 0.75\begin{pmatrix}-1\\1\end{pmatrix} - 0.25\begin{pmatrix}\frac{1}{1+e^{-1/2}} - \frac{1}{2}\end{pmatrix}$$

$$= \begin{pmatrix}-1 + 0.75(-1) - 0.25\left(\frac{1}{1+e} - \frac{1}{2}\right)\\ +1 + 0.75(1) - 0.25\left(\frac{-1}{2}\right)\end{pmatrix}$$

$$= \begin{pmatrix}-1.692\\1.875\end{pmatrix}$$

Answer to Qd The updation rule for Nesterov's Accelerating Gradient

$$w_{t+1} = w_t + \beta \Delta w_t - \eta \nabla l(w_t + \beta \Delta w_t)$$

$$\Rightarrow w_6 = w_5 + \beta \Delta w_5 - \eta \nabla l(w_5 + \beta \Delta w_5)$$

Now $w_5 + \beta \Delta w_5 = \begin{pmatrix}-1\\1\end{pmatrix} + 0.75\begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}-1.75\\1.75\end{pmatrix}$

We have $\frac{\partial L}{\partial w_0} = \frac{(1+e^{-w_0-w_1})e^{w_0} - (1+e^{w_0})e^{-w_0-w_1}}{(1+e^{w_0})(1+e^{-w_0-w_1})}$

$$= \frac{2e^{-1.75} - (1+e^{-1.75})}{(1+e^{-1.75})(1+e^{0})} = \frac{2e^{-1.75} - 1 - e^{-1.75}}{2(1+e^{-1.75})}$$

$$= \frac{e^{1.75}}{1+e^{1.75}} - \frac{1}{2}$$

$$= \frac{1}{1+e^{1.75}} - \frac{1}{2}$$

Also $\frac{\partial L}{\partial w_1} = \frac{-(1+e^{w_0})\left(e^{-(w_0+w_1)}\right)}{(1+e^{w_0})\left(1+e^{-(w_0+w_1)}\right)} = \frac{-1}{1+1} = \frac{-1}{2}$

Now $w_1 = \begin{pmatrix} -1.75 \\ 1.75 \end{pmatrix} - 0.25 \begin{pmatrix} \frac{1}{1+e^{1.75}} - \frac{1}{2} \\ -1/2 \end{pmatrix}$

$$= \begin{pmatrix} -1.75 + \frac{0.25}{2} - \frac{0.25}{1+e^{1.75}} \\ 1.75 + \frac{0.25}{2} \end{pmatrix}$$

$$= \begin{pmatrix} -1.66 \\ 1.875 \end{pmatrix}$$

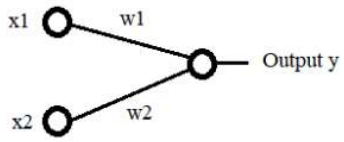
Question 5

Question 4. [2+3+0.5+0.5 = **6 marks**]

Consider the following training dataset with input X=(x1, x2) and target (desired) output d neuron with two inputs and one output is used for this training dataset. Activation function a linear function with zero bias. Sum of square error is used as the loss function.

Input x1	Input x2	Target Output d
0	0	0.0

1	0	1.0
0	1	1.0
1	1	0.0



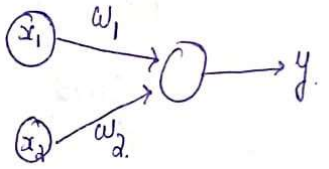
A. If back propagation is used, what will be the weights (w1, w2) after convergence?

B. What will be the nature of the loss function? What is the value of learning rate which leads to convergence in least number of iterations? Show all calculation steps.

C. To achieve convergence in least number of iterations, will you use batch gradient descent, stochastic gradient descent or mini batch gradient and why?

Answer

x_1	x_2	d
0	0	0
1	0	1
0	1	1
1	1	0



a) Loss function = $l(\omega) = (0 - 0\omega_1 - 0\omega_2)^2 + (1 - 1\omega_1 - 0\omega_2)^2 + (1 - 0\omega_1 - 1\omega_2)^2 + (0 - \omega_1 - \omega_2)^2$

$$\boxed{l(\omega) = (1-\omega_1)^2 + (1-\omega_2)^2 + (\omega_1+\omega_2)^2}$$

After applying back propagation, weights are supposed to converge to local minima.

$$\nabla l(\omega) = \begin{pmatrix} \frac{\partial l(\omega)}{\partial \omega_1} \\ \frac{\partial l(\omega)}{\partial \omega_2} \end{pmatrix} = \begin{pmatrix} -2(1-\omega_1) + 2(\omega_1+\omega_2) \\ -2(1-\omega_2) + 2(\omega_1+\omega_2) \end{pmatrix}$$

for extremum points: $\frac{\partial l(\omega)}{\partial \omega_1} = 0$

$$\Rightarrow 2(\omega_1+\omega_2) - 2(1-\omega_1) = 0$$

$$\Rightarrow 2\omega_1 + \omega_2 = 1$$

$$\Rightarrow \boxed{\omega_2 = 1 - 2\omega_1}$$

Also, $\frac{\partial l(w)}{\partial w_2} = 0 \Rightarrow w_1 + w_2 = 1 - w_2$

$\Rightarrow w_1 + w_2 = 1 - 1 + 2w_1$

$\Rightarrow w_1 + 1 - 2w_1 = 2w_1$

$\Rightarrow w_1 - 2w_1 - 2w_1 = -1$

$\Rightarrow \boxed{w_1 = 1/3}$

We get $w_2 = 1 - 2w_1 = 1 - \frac{2}{3} = 1/3 \Rightarrow \boxed{w_1 = 1/3}$

Now $H = \begin{pmatrix} \frac{\partial^2 l}{\partial w_1^2} & \frac{\partial^2 l}{\partial w_2 \partial w_1} \\ \frac{\partial^2 l}{\partial w_1 \partial w_2} & \frac{\partial^2 l}{\partial w_2^2} \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$

eigen values of H will be roots of $\lambda^2 - 8\lambda + 12 = 0$.

$(\lambda - 6)(\lambda - 2) = 0$

The eigenvalues of H are 6 & 2. $\Rightarrow$ H is +ve definite

$\Rightarrow w_1$ & w_2 attained are local minima.

B) Loss function is Quadratic in nature. We can approximate using Taylor's approximation.

$$l(w_{K+1}) = l(w_K) + (w_{K+1} - w_K)\nabla l(w_K) + \frac{1}{2}(w_{K+1} - w_K)^T H (w_{K+1} - w_K)$$

Here w_K is weights at kth iteration.

w_{K+1} is unknown weights at (k+1)th iteration

We have gradient updation rule.

$$\omega_{k+1} = \omega_k - \eta \nabla l(\omega_k).$$

We substitute ω_{k+1} in Taylor's theorem with equation above to get

$$l(\omega_{k+1}) = l(\omega_k - \eta \nabla l(\omega_k))$$

$$= l(\omega_k) + (\omega_k - \eta \nabla l(\omega_k) - \omega_k)^T \nabla l(\omega_k) +$$

$$(\omega_k - \eta \nabla l(\omega_k) - \omega_k)^T H (\omega_k - \eta \nabla l(\omega_k) - \omega_k)$$

$$\boxed{l(\omega_{k+1}) = l(\omega_k) - \eta g^T g + \eta^2 g^T H g} \quad [\text{where } g = \nabla l(\omega_k)]$$

for extremum points; $\frac{\partial}{\partial \eta} l(\omega_{k+1}) = 0$

$$\Rightarrow -g^T g + 2\eta g^T H g = 0$$

$$\Rightarrow \boxed{\eta = \frac{g^T g}{g^T H g}}$$

We have $H = \Phi D \Phi^T$ where Φ is eigenvector of diagonal matrix D

$$g^T H g = g^T \Phi D \Phi^T g.$$

$$= X^T D X \quad \text{where } X = \Phi^T g.$$

$$= \sum_i \lambda_i x_i^2$$

Here, $\lambda_{max} = \max_i \lambda_i$

$\lambda_i \leq \lambda_{max}$.

$\Rightarrow \sum_i^n \lambda_i x_i^2 \leq \lambda_{max} \sum_{i=1}^n x_i$

$\Rightarrow X^T D X \leq \lambda_{max} X^T X$

$\Rightarrow g^T \phi D \phi^T g \leq \lambda_{max} \cdot g^T \phi \phi^T g$

$\Rightarrow g^T H g \leq \lambda_{max} g^T g$. [$\because \phi$ is orthogonal]

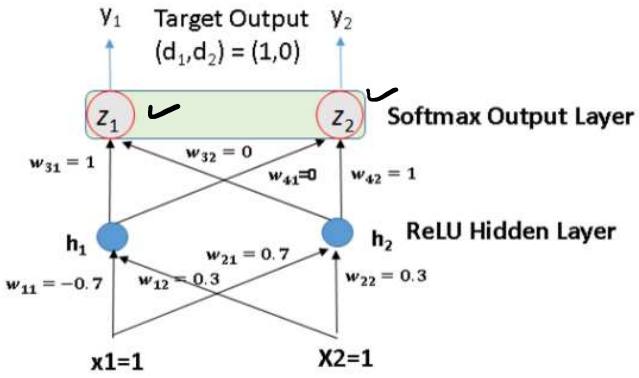
$\Rightarrow \boxed{\eta \geq \frac{1}{\lambda_{max}}}$

In our case, $\lambda_{max} = 6$

$\Rightarrow \boxed{\eta \geq 1/6}$

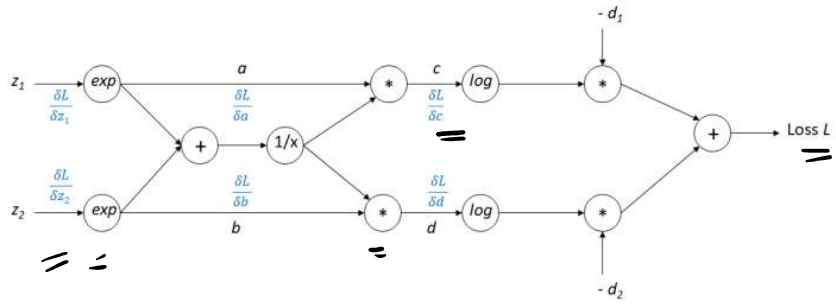
Optimal learning rate is $1/6$

Question 6

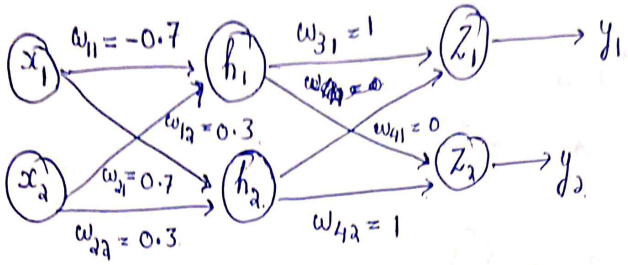


A. Calculate the actual output (y_1, y_2) for the current iteration with input (x1, x2) = (1, 1).
$y_1 = 1 / (1+e)$ $y_2 = e / (1+e)$

B. With the (z_1, z_2) calculated in A., use the following computation graph to calculate the loss L, $\frac{\delta L}{\delta c}, \frac{\delta L}{\delta d}, \frac{\delta L}{\delta a}, \frac{\delta L}{\delta b}, \frac{\delta L}{\delta z_1}, \frac{\delta L}{\delta z_2}$.



Relu hidden layer



We have $h_1 = w_{11} x_1 + w_{12} x_2 = -0.7(1) + 0.3(1) = -0.4$

$h_2 = w_{21} x_1 + w_{22} x_2 = 0.7(1) + 0.3(1) = 1$

We have $a_1 = \text{relu}(h_1) = 0$

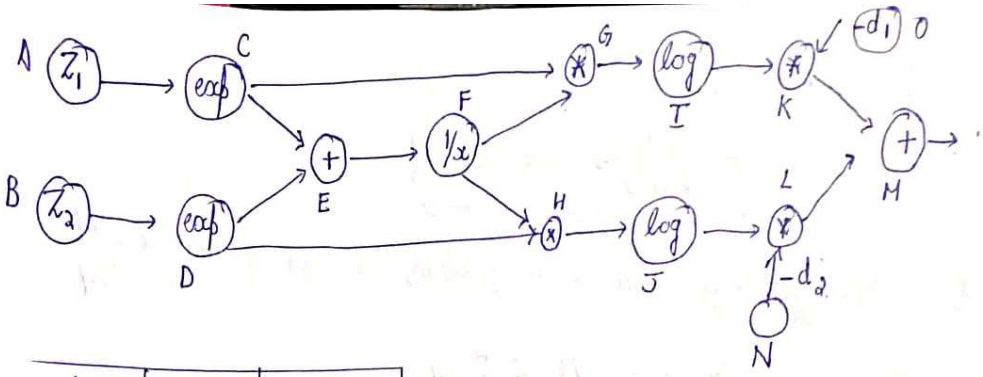
$a_2 = \text{relu}(h_2) = 1$

Here $z_1 = w_{31} a_1 + w_{32} a_2 = 1(0) + 0(1) = 0$

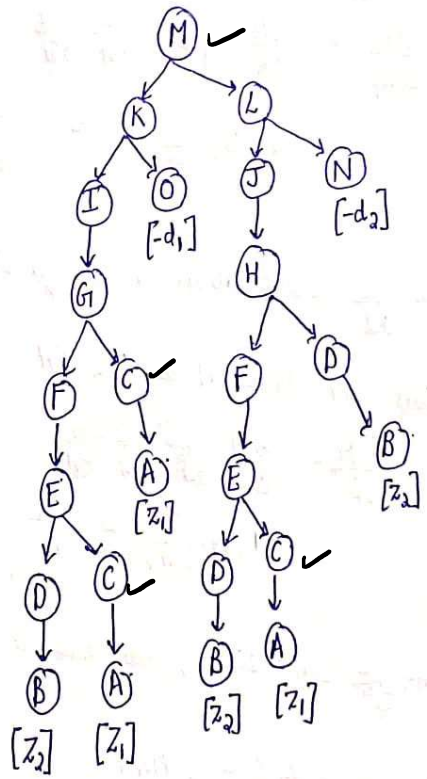
$z_2 = w_{41} a_1 + w_{42} a_2 = 0(0) + 1(1) = 1$

$$y_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} = \frac{e^0}{e^0 + e^1} = \frac{1}{e+1}$$

$$y_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2}} = \frac{e}{1+e}$$



A	Z_1	0
B	Z_2	1
C	e^A	1
D	e^B	e
E	C + D	1 + e
F	1/E	1/(1+e)
G	C * F	1/(1+e)
H	D * F	e/(1+e)
I	log G	$\log\left(\frac{1}{1+e}\right)$
J	log H	$\log\left(\frac{e}{1+e}\right)$
O	$-d_1$	-1
N	$-d_2$	0
K	I * O	$-\log\left(\frac{1}{1+e}\right)$



L	J*N	0
M	K+L	$-\log\left(\frac{1}{1+e}\right)$ = Loss calculated

a) For $\frac{\partial L}{\partial c}$ computation as per question, we have to compute $\frac{\partial M}{\partial G_1}$

Path from M to G_1 is M K I G_1

$$\frac{\partial M}{\partial G_1} = \frac{\partial M}{\partial K}\cdot\frac{\partial K}{\partial I}\cdot\frac{\partial I}{\partial G_1}\cdot\frac{\partial G_1}{\partial G_1}$$

$$= 1.(0)\,\frac{1}{G_1}.(1) = \frac{-1}{1/(1+e)} = -(1+e)$$

b) For $\frac{\partial L}{\partial d}$ computation as per question, we have to compute $\frac{\partial M}{\partial H}$

Path from M to H is M L J H

$$\frac{\partial M}{\partial H} = \frac{\partial M}{\partial L}\cdot\frac{\partial L}{\partial J}\cdot\frac{\partial J}{\partial H}\cdot\frac{\partial H}{\partial H}$$

$$= 1(N)\left(\frac{1}{H}\right) = \frac{N}{H} = 0$$

c) For $\frac{\partial L}{\partial a}$ computation, we have to compute $\frac{\partial M}{\partial C}$

Path from M to C – 1) M K I G_1 C
2) M K I G F E C
3) M L J H F E C

Now, $\frac{\partial H}{\partial C} = \frac{\partial M}{\partial K}\frac{\partial K}{\partial I}\frac{\partial I}{\partial G}\frac{\partial G}{\partial C}\frac{\partial C}{\partial C} +$

$\frac{\partial M}{\partial K}\frac{\partial K}{\partial I}\frac{\partial I}{\partial G}\frac{\partial G}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial C}\frac{\partial C}{\partial C} +$

$\frac{\partial M}{\partial L}\frac{\partial L}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial C}\frac{\partial C}{\partial C}$

$= 1(0)\frac{1}{G}F + 1(0)\frac{1}{G}C.\left(\frac{-1}{E^2}\right) +$

$1(N)\left(\frac{1}{H}\right)D.\left(\frac{-1}{E^2}\right)(1)$

$= -\frac{F}{G} + \frac{C}{GE^2} - \frac{ND}{HE^2}$

$= \frac{-1}{(1+e)\frac{1}{1+e}} + \frac{1}{\frac{1}{1+e}(1+e)^2} - 0$

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

d) For $\frac{\partial L}{\partial b}$ computation, we compute $\frac{\partial M}{\partial D}$

We have following paths from M to D.

1) M K I G F E D
2) M L J H F E D
3) M L J H D

$$\frac{\partial M}{\partial D} = \frac{\partial M}{\partial K}\frac{\partial K}{\partial I}\frac{\partial I}{\partial G}\frac{\partial G}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial D} +$$

$$\frac{\partial M}{\partial L}\frac{\partial L}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial D} +$$

$$\frac{\partial M}{\partial L}\frac{\partial L}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial D}\frac{\partial D}{\partial D}$$

$$= 1(0)\frac{1}{G}\,C\left(\frac{-1}{E^2}\right)(1)(1) + 1(N)\frac{1}{H}\cdot D\left(\frac{-1}{E^2}\right)1 +$$

$$1(N)\frac{1}{H}F(1)$$

$$= -\frac{CO}{GE^2} = \frac{-1(-1)}{\frac{1}{(1+e)}(1+e)^2} = \frac{1}{1+e}$$

e) for $\partial L/\partial z_1$ we have to compute $\partial M/\partial A$

We have following paths from M to A

1) MKIGCA
2) MKIGFECA
3) MLJHFECA

$$\frac{\partial M}{\partial A} = \frac{\partial M}{\partial K}\frac{\partial K}{\partial I}\frac{\partial I}{\partial G}\frac{\partial G}{\partial C}\frac{\partial C}{\partial A}\frac{\partial A}{\partial A} +$$

$$\frac{\partial M}{\partial K}\frac{\partial K}{\partial I}\frac{\partial I}{\partial G}\frac{\partial G}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial C}\frac{\partial C}{\partial A}\frac{\partial A}{\partial A} +$$

$$\frac{\partial M}{\partial L}\frac{\partial L}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial C}\frac{\partial C}{\partial A}\frac{\partial A}{\partial A}$$

$$= +1(0)\frac{1}{G_1} F e^A + 1(0)\frac{1}{G_1} C\left(\frac{-1}{E^2}\right)(1) e^A +$$

$$1(N)\frac{1}{H} D\left(-1/E^2\right)(1) e^A$$

$$= +\frac{0 F e^A}{G_1} + \frac{-0 C e^A}{G_1 E^2} + 0$$

$$= +(-1)\frac{1}{1+e}\frac{e^{z_1}}{\frac{1}{1+e}} - \frac{(-1)1 e^{z_1}}{\frac{1}{1+e}(1+e)^2}$$

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

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

f) Similarly; $\partial L/\partial z_2 = \partial M/\partial B$.

$$\frac{\partial M}{\partial B} = \frac{\partial M}{\partial K}\frac{\partial K}{\partial I}\frac{\partial I}{\partial G_1}\frac{\partial G_1}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial B}\frac{\partial B}{\partial B} +$$

$$\frac{\partial M}{\partial L}\frac{\partial L}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial F}\frac{\partial F}{\partial E}\frac{\partial E}{\partial D}\frac{\partial D}{\partial B}\frac{\partial B}{\partial B} +$$

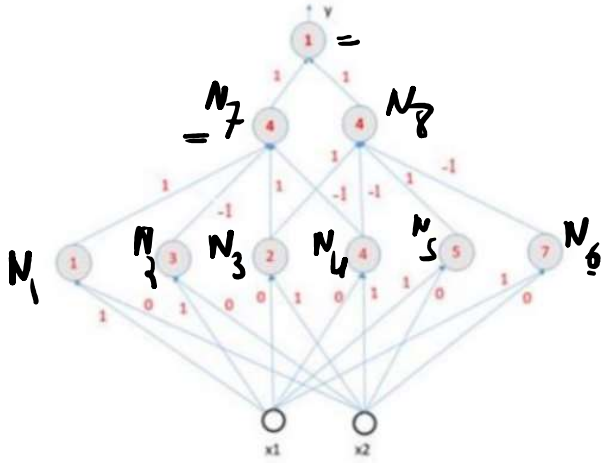
$$\frac{\partial M}{\partial L}\frac{\partial L}{\partial J}\frac{\partial J}{\partial H}\frac{\partial H}{\partial D}\frac{\partial D}{\partial B}\frac{\partial B}{\partial B}.$$

$$= 1(0)\frac{1}{G_1} C\left(\frac{-1}{E^2}\right)(1) e^B + N(N)\,0 + 0 \quad \left[\because \frac{\partial L}{\partial J} = N = 0\right]$$

$$= -\frac{0 C e^B}{G_1 E^2} = -\frac{(-1)1 e^1}{\frac{1}{1+e}(1+e)^2} = \frac{e}{1+e}$$

Question 7

Refer to the following multilayer perceptron network with two hidden layers. All nodes use a step activation function, i.e., output = +1 if total input >= bias output = -1 if total input < bias Bias at each node is indicated inside the node.



A. What will be the output y, if (x1,x2)=(2,3), (x1,x2)=(6,3), and (x1,x2)=(4,3)

B. Express output y as a decision rule of input (x1, x2). Recall, decision rule is expressed as an if-then-else statement, e.g., y = (x1> 2) OR (x2 < 7) AND (x1 + x2 = 7)

C. Can this decision rule be realized using a multilayer perceptron with one hidden layer? If no, why? If yes, how many hidden node will be needed in that one hidden layer?

Layer 1			Layer 2			Layer 3		
Node	Input	output	Node	Input	output	Node	Input	output
N_1	2	1	N_7	4	1	N_9	0	-1
N_2	2	-1	N_8	2	-1			
N_3	3	1						
N_4	3	-1						
N_5	2	-1						
N_6	2	-1						