# PROBLEM-1

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

$GELU(x) = x\,\sigma(1.702x)$.

$\Rightarrow \frac{\sigma(x)}{dx} = \sigma(x)(1-\sigma(x)) = \sigma(x) - \sigma^2(x)$.

$\therefore$ For gradient descent :

$$x_{n+1} = x_n - \eta \frac{d\,GELU(x_n)}{dx_n}$$

$$= x_n - \eta\left[\sigma(1.702x) + 1.702x\,\sigma(1.702x)\left[1-\sigma(1.702x)\right]\right]$$

$$= x_n - \eta\left[e^{1.702x}(1.702x + e^{1.702x}+1)/(e^{1.702x}+1)^2\right]$$

$\therefore$ At $x_0 = 0$ $\quad GELU(x_0) = 0$

$\Rightarrow$

$\boxed{x_1 = -0.05 \quad x_2 = -0.0957501 \quad x_3 = -0.13763}$

$\boxed{Gelu(x_1) = -0.02393 \quad Gelu(x_2) = -0.04398 \quad Gelu(x_3) = -0.06079}$

(b) Similar to part (a) we can calculate for $\eta = 1$ & $x_0 = 0$

$\boxed{x_1 = -0.5 \quad x_2 = -0.62077 \quad x_3 = -0.6765}$

$\boxed{Gelu(x_1) = -0.149611 \quad Gelu(x_2) = -0.160139 \quad Gelu(x_3) = -0.1625175}$

The update is faster for $\eta = 1$ compared to $\eta = 0.1$ we get faster convergence to minima

Scanned with CamScanner

(c)

$x_0 = -3 \qquad \eta = 0.1$

$$V_0 = x_0 - \eta\left[e^{1.702x}(1.702x + e^{1.702x} + 1) / (e^{1.702x} + 1)^2\right].$$

$\downarrow$

$x_1 = x_0 - \eta V_0$

$\downarrow$

$V_1 = \beta V_0 + (1-\beta)\nabla f x_0$

Using the above algorithm we get.

| | |
|---|---|
| $x_1 = -2.99$ | $G(x_1) = -0.0181$ |
| $x_2 = -2.9950$ | $G(x_2) = -0.0182$ |
| $x_3 = -2.99249$ | $G(x_3) = -0.0183$ |

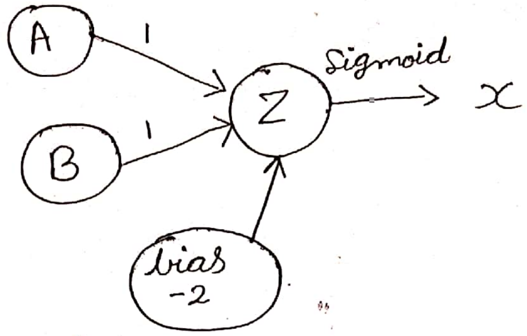
Using momentum update as $\beta = 0.9$.

| | |
|---|---|
| $x_1 = -2.9975$ | $Gelu(x_1) = -0.0181$ |
| $x_2 = -2.9950$ | $Gelu(x_2) = -0.0182$ |
| $x_3 = -2.9925$ | $Gelu(x_3) = -0.0183$ |

The performance is almost same for 2 methods but that can be due to the reason that GELU as per graph has only 1 minima and momentum is generally used to bypass local minima to reach global minima.

# PROBLEM-2

## AND




$\Rightarrow$ $A=0$, $B=0$ $\Rightarrow$ $Z = A(1)+B(1)+(-2)$ $\Rightarrow$ $\sigma(z)=\sigma(-2)=0.119 \Rightarrow \underline{x=0}$

$Z = 0+0-2=-2$

$\therefore$ $\boxed{\begin{matrix} A=0 \\ B=0 \\ x=0 \end{matrix}}$

$\Rightarrow$ $A=1$, $B=0$ $\Rightarrow$ $Z=A(1)+B(1)+(-2)$ $\Rightarrow$ $\sigma(-1)=0.2 \Rightarrow \underline{x=0}$

$Z = 1+0-2$

$\therefore$ $\boxed{\begin{matrix} A=1 \\ B=0 \\ x=0 \end{matrix}}$

$\Rightarrow$ $A=0$, $B=1$ $\Rightarrow$ $Z=A(1)+B(1)+(-2)$ $\Rightarrow$ $\sigma(-1)=0.2 \Rightarrow \underline{x=0}$

$Z = 0+1-2=-1$

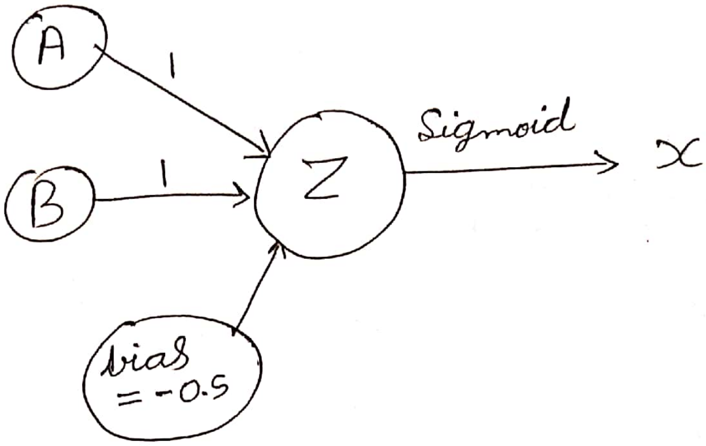
$\therefore$ $\boxed{\begin{matrix} A=0 \\ B=1 \\ x=0 \end{matrix}}$

$\Rightarrow$ $A=1$, $B=1$ $\Rightarrow$ $Z=A(1)+B(1)+(-2)$ $\Rightarrow$ $\sigma(0)=0.5 \Rightarrow \underline{x=1}$

$Z = 1+1-2=0$

$\therefore$ $\boxed{\begin{matrix} A=1 \\ B=1 \\ x=1 \end{matrix}}$

# OR




A = 0, B = 0 $\Rightarrow Z = A(1) + B(1) + (-0.5) \Rightarrow \sigma(-0.5) = 0.3775 \Rightarrow \underline{x = 0}$

$Z = 0 + 0 + (-0.5)$

$$\boxed{\begin{matrix} A = 0 \\ B = 0 \\ x = 0 \end{matrix}}$$

A = 1, B = 0 $\Rightarrow Z = A(1) + B(1) + (-0.5) \Rightarrow \sigma(0.5) = 0.622 \Rightarrow \underline{x = 1}$

$Z = 1 + 0 - 0.5$

$$\boxed{\begin{matrix} A = 1 \\ B = 0 \\ X = 1 \end{matrix}}$$

A = 0, B = 1 $\Rightarrow Z = A(1) + B(1) + (-0.5) \Rightarrow \sigma(0.5) = 0.622 \Rightarrow \underline{x = 1}$

$Z = 0 + 1 - 0.5$

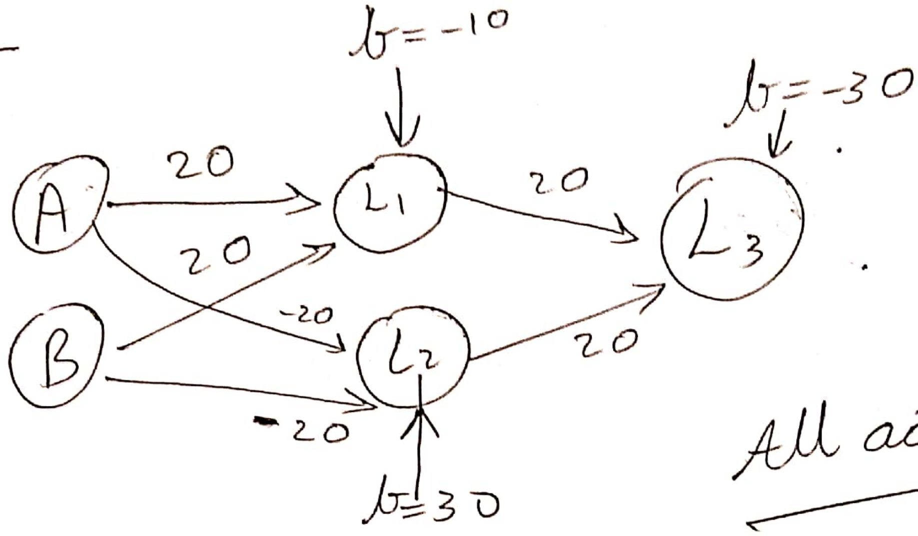
$$\boxed{\begin{matrix} A = 0 \\ B = 1 \\ x = 1 \end{matrix}}$$

A = 0, B = 0 $\Rightarrow Z = A(1) + B(1) + (-0.5) \Rightarrow \sigma(-0.5) = 0.3775 \Rightarrow \underline{x = 0}$

$Z = 0 + 0 - 0.5$

$$\boxed{\begin{matrix} A = 0 \\ B = 0 \\ X = 0 \end{matrix}}$$

## XOR




All activation sigmoid

$\left.\begin{array}{l}A=1\\B=1\end{array}\right\}$ At $L_1$

$= 20+20-10$

$= 30 \xrightarrow{Sig(30)} (>0.5)$ ie 1.

At $L_2$.

$= -20-20+30$

$= -10 \xrightarrow{sig(-10)} < 0.5$ ie 0.

At $L_3$.

$20-30 = -10$

$sig(-10) < 0.5$

$\rightarrow \boxed{out = 0}$.

Similarly for ~~0-0~~ $\boxed{A=0\ ,\ B=0}$ we can say output $\boxed{x=0}$

At $\left.\begin{array}{l}A=1\\B=0\end{array}\right\}$ At $L_1$

$20-10=10$

$\xrightarrow{Sig(10)>0.5} 1$.

At $L_2$

$-20+30=10$.

$\xrightarrow{Sig(10)>0.5} 1$

At $L_3$

$20+20$

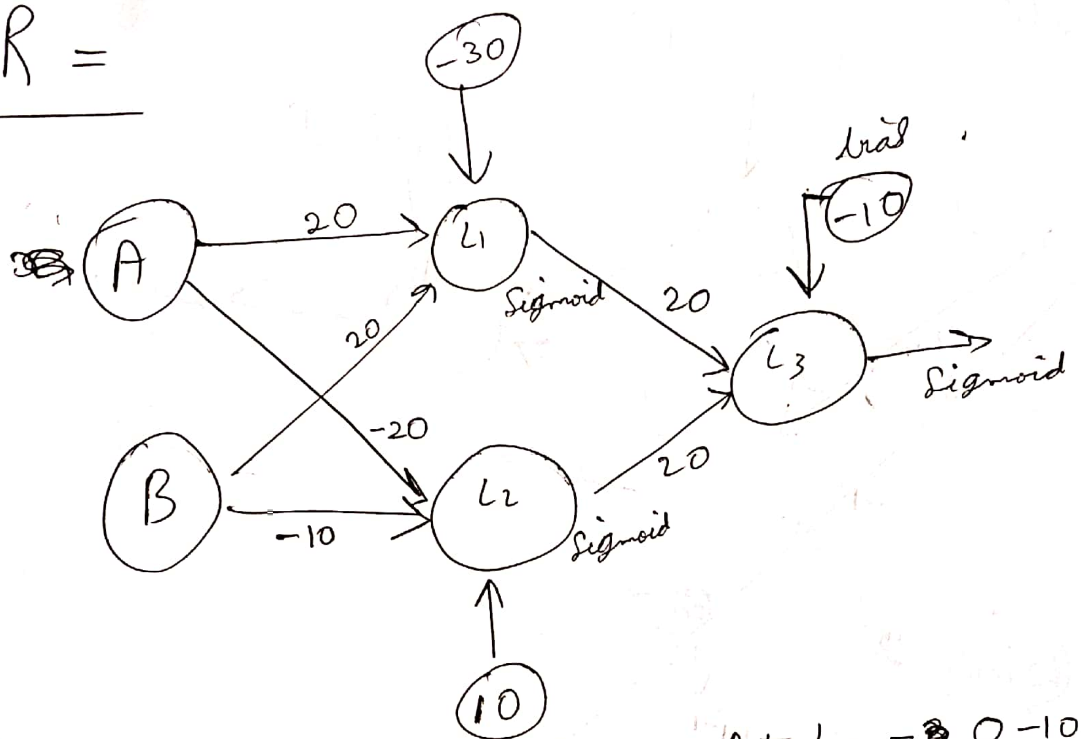
$-30$

$=10 \rightarrow Sig(10) > 0.5$

$\rightarrow \boxed{output = 1}$

Similarly we can prove for A=0 & B=1

# XNOR =




$A = 0$
$B = 0$

At $L_1 = sig(-30) \rightarrow 0$

At $L_2 = sig(10) \rightarrow 1$

At $L_3 = 0 - 10 + 20$
$= 10 \quad sig(10) \rightarrow > 0.5$
$\rightarrow \boxed{out = 1}$

Similarly we can prove for $A=1$, $B=1$

If $A=1$, $B=0$

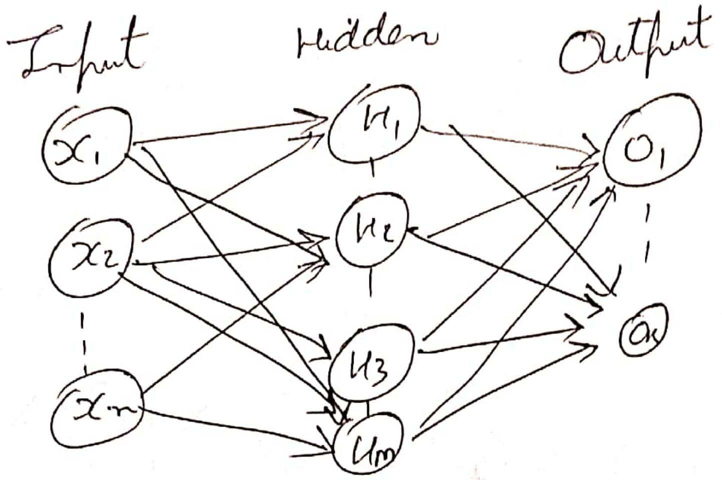
At $L_1 = 20 - 30 = -10 \rightarrow sig(-10) \rightarrow 0$

At $L_3 = -10 \xrightarrow{sig} \boxed{0}$

out = ?

$AL_2 = 10 - 20 = -10 \xrightarrow{sig(-10)} 0.$

Q3)




$\sigma(x) = 1/_{1+e^{-x}}$ $\qquad S(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ $\qquad E(o) = -\sum_1^K y_i \log o_i$

$f_1 = x\omega_1 + b_1$

$a = \sigma f(1)$

$f(2) = a\omega_2 + b2$

$O = S(f_2)$.

a) $E(o) = -\sum_1^K y_i \log(o_i) = -\sum_1^K y_i \log(S(f_2)) = -\sum_1^K y_i \log\left(\frac{e^{x_i}}{\sum_j e^{x_j}}\right)$

$= -\sum_1^x y_i x_i - \log\left(\sum_j e^{x_i}\right)$ --- ①

Now for $\frac{\partial E}{\partial f_2} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial f_2}$ -- ③

we can say

$\frac{\partial o}{\partial f_2} = S(f_2)(1-S(f_2))$ -- ②

On simplyfing ① & ② and putting it in ③ we can write.

$\frac{\partial E}{\partial o} = o - y$.

b) $$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial o} \times \frac{\partial o}{\partial f_2} \times \frac{\partial f_2}{\partial a} \times \frac{\partial a}{\partial f_1} \times \frac{\partial f_1}{\partial x}.$$

$$\boxed{\frac{\partial E}{\partial x} = \underbrace{(O-y)}_{\frac{\partial E}{\partial o}} \times \underbrace{\left[\sigma(f_2) - \sigma^2(f_2)\right]}_{\frac{\partial o}{\partial f_2}} \times \underbrace{w_2}_{\frac{\partial f_2}{\partial a}} \times \underbrace{w_1}_{\frac{\partial f_1}{\partial x}} \times \underbrace{\left[\sigma(f_1) - \sigma^2(f_1)\right]}_{\frac{\partial a}{\partial f_1}}}$$

# PROBLEM-4

a)

| | | | |
|---|---|---|---|
| 3 | 5 | 2 | 3 |
| 9 | 1 | 8 | 4 |
| 6 | 4 | 3 | 7 |
| 7 | 0 | 2 | 4 |

Feature Map (F)

| | | |
|---|---|---|
| -1 | 0.5 | -2 |
| 2 | 0 | 1 |
| 0 | 1 | 1.5 |

Filter 1

Size of feature Map (F) = 4×4

Size of filter 1 = 3×3

Size of new feature map (F') = 4×4

$$\text{Output size} = \frac{W-K+2P}{S}+1$$

where W = input size
K = Kernel size
S = Stride
P = padding.

$$\Rightarrow 4 = \frac{4-3+2P}{1}+1$$

$$\Rightarrow P = \frac{(4-1)-(4-3)}{2} = 1 \Rightarrow \boxed{\text{Padding size} = 1}$$

⇒

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 3 | 5 | 2 | 3 | 0 |
| 0 | 9 | 1 | 8 | 4 | 0 |
| 0 | 6 | 4 | 3 | 7 | 0 |
| 0 | 7 | 0 | 2 | 4 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

Padded feature map

| | | |
|---|---|---|
| -1 | 0.5 | -2 |
| 2 | 0 | 1 |
| 0 | 1 | 1.5 |

⇒

| | | | |
|---|---|---|---|
| 9+1.5+5 = 15.5 | 6+2+1+12 = 21 | 10+3+8+6 = 27 | 4+4 = 8 |
| 1.5-10+1+6+6 = 4.5 | -3+2.5-4+18+8+4+4.5 = 30 | -5+1-6+2+4+3+10.5 = 9.5 | -2+1.5+16+7 = 22.5 |
| 4.5-2+4+7 = 13.5 | -9+.5+12-16+3+3 = -6.5 | -1+4-8+8+7+2+6 = 18 | -8+2+6+4 = 4 |
| 3-8 = -5 | -6+2-6+14+2 = 6 | -4+1.5-14+4 = -12.5 | -3+3.5+4 = 4.5 |

Scanned with CamScanner

(b)

| | |
|---|---|
| -1 | 0.5 |
| 2 | 0 |

Filter 2

Padding size = 1

Given feature map (F) size = 4×4

Final feature map size = 4×4.

$$\Rightarrow \text{Final feature map size} = \frac{W-K+2P}{S}+1$$

$$\Rightarrow 4 = \frac{4-2+2}{S}+1$$

$\underline{S = 4/3} \rightarrow$ Not possible to have fractional stride.

It is not possible to get feature map of same size as input with given parameters.

(c) From (a) the new feature map $F' =$

| | | | |
|---|---|---|---|
| 15.5 | 21 | 27 | 8 |
| 4.5 | 30 | 9.5 | 22.5 |
| 13.5 | -6.5 | 18 | 4 |
| -5 | 6 | -12.5 | 4.5 |

∴ Average Pooling:

| | |
|---|---|
| $\frac{15.5+21+4.5+30}{4} = 17.75$ | $\frac{27+8+9.5+22.5}{4} = 16.75$ |
| $\frac{13.5-6.5-5+6}{4} = 2$ | $\frac{18+4+4.5-12.5}{4} = 4.5$ |

d) New filter = $\begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$

Stride = 2
Padding = 0

Proof:

| | |
|---|---|
| 3.875 + 5.25 + 1.125 + 7.5 = 17.75 | 6.75 + 2 + 2.375 + 5.625 = 16.75 |
| 3.375 + 1.75 + 1.5 + (−.25) = 2 | 4.5 |