1) a) $g(x) := GELU(x) = x\,\sigma(1.702\,x)$

$$\frac{dg(x)}{dx} = \sigma(1.702x) + x\left(\sigma(1.702x)(1-\sigma(1.702x))\right) \times 1.702$$

For G.D.

$$x_{t+1} = x_t - \eta \nabla g(x_t)$$

t = 0

$x_0 = 0, \quad g(0) = 0$

t = 1

$\Rightarrow x_1 = x_0 - (0.1)\nabla g(0)$

$= 0 - 0.1\left[\sigma(0) + 0\right] = -0.1\left(\frac{1}{2}\right)$

$\Rightarrow \boxed{x_1 = -0.05}$

t = 2

$x_2 = x_1 - (0.1)\nabla g(-0.05)$

~~$= -0.05 - (0.1)[\sigma(-0.05) - 0.05(\sigma(-0.05))(1-\sigma($~~

$= -0.05 - 0.1\left[\sigma(-0.0851) - 0.05\left(\sigma(-0.0851)\cdot(1-\sigma(-0.0851)\right)\times 1.702\right]$

$= -0.05 - 0.1\left[0.4787 - 0.05(0.4787(1-0.4787))\times 1.702\right]$

$\boxed{x_2 = -0.0957}$

$\underline{t=3}$

$$x_3 = x_2 - (0.1)\nabla g(-0.0957)$$

$$\cancel{\Rightarrow x_3 = -0.0957 - 0.1\left[\sigma(-0.1629)(1-\sigma(\right.}$$

$$x_3 = -0.0957 - 0.1\left[\sigma(-0.1629) - 0.0957\left(\sigma(0.1629)(1-\sigma(0.1629))\right)\times 1.702\right]$$

$$\boxed{x_3 = -0.1376}$$

function values,

$$\boxed{\begin{aligned} g(x_1) &= -0.0239 \\ g(x_2) &= -0.0439 \\ g(x_3) &= -0.0608 \end{aligned}}$$

b) with $\underline{\eta = 1.0}$, $x_0 = 0$

$\underline{t=1}$

$$x_1 = 0 - (1.0)\nabla g(0) = \cancel{-0.5}$$

$$\Rightarrow \boxed{x_1 = -0.5} \text{ and } \boxed{g(x_1) = -0.1496}$$

$\underline{t=2}$

$$x_2 = -0.5 - (1.0)\left[\sigma(\overset{-0.851}{\cancel{-0.5}}) - 0.5\left(\sigma(-0.851)(1-\sigma(-0.851))\times 1.702\right)\right]$$

$$\boxed{x_2 = -0.6208} \Rightarrow \boxed{g(x_2) = -0.1601}$$

Scanned by CamScanner

$\underline{t=3}$

$$x_3 = (-0.6208) - (1.0)\left[\sigma(-1.057) - (0.6208)(\sigma(-1.057))(1-\sigma(-1.057)) \times 1.702\right]$$

$$\boxed{x_3 = -0.6765} \Rightarrow \boxed{g(x_3) = -0.1625}$$

with $\eta = 1.0$, the function ~~moved~~ value decreased by almost an order of magnitude faster than with $\eta = 0.1$

c) i) $x_0 = -3$, $\eta = 0.1$

$\underline{t=1}$

$$x_1 = (-3) - (0.1)\,\nabla g(0.1)$$

$$\boxed{x_1 = -2.9975} \Rightarrow \boxed{g(x_1) = -0.0181}$$

$\underline{t=2}$

$$x_2 = (-2.9975) - (0.1)\,\cancel{g}\,\nabla g(-2.9975)$$

$$\Rightarrow \boxed{x_2 = -2.9951} \Rightarrow \boxed{g(x_2) = -0.0182}$$

$\underline{t=3}$

$$x_3 = (-2.9951) - (0.1)\,\nabla g(-2.9951)$$

$$\Rightarrow \boxed{x_3 = -2.9926} \Rightarrow \boxed{g(x_3) = -0.0183}$$

ii) GD with momentum, $\underline{\beta = 0.9}$, $\underline{\eta = 0.1}$, $\underline{x_0 = -3}$

$$\therefore \quad v_0 = \nabla g(x_0) = -0.0245548$$

<u>t=1</u>

$$v_1 = (0.9)\, v_0 + (0.1)\, \nabla g(x_0) =$$

$$\Rightarrow \boxed{v_1 = -0.024548}$$

$$x_1 = x_0 - (0.1)\, v_1$$

$$\Rightarrow \boxed{x_1 = -2.99755}\,, \quad \boxed{g(x_1) = -0.0181}$$

<u>t=2</u>

$$v_2 = (0.9)\, v_1 + (0.1)\, \nabla g(x_1)$$

$$v_2 = -0.024556$$

$$x_2 = x_1 - (0.1)\, v_2$$

$$\Rightarrow \boxed{x_2 = -2.995089}\,, \quad \boxed{g(x_2) = -0.018192}$$

<u>t=3</u>

$$v_3 = (0.9)\, v_2 + (0.1)\, \nabla g(x_2)$$

$$\boxed{v_3 = -0.02457}$$

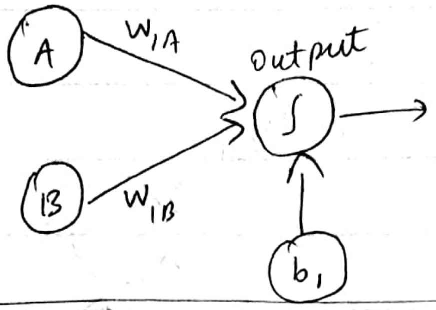
$$x_3 = x_2 - (0.1)\, v_3$$

$$\Rightarrow \boxed{x_3 = -2.992632} \Rightarrow \boxed{g(x_3) = -0.018253}$$

iii) In this case both the methods perform almost the same as could be seen in the x_t and $g(x_t)$ values.

2)

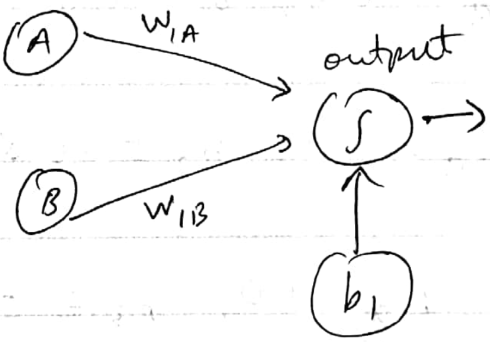
AND gate



{ ∫ : sigmoid }

$$\boxed{W_{1A} = 1.0, \quad W_{1B} = 1.0, \quad b_1 = -1.5}$$

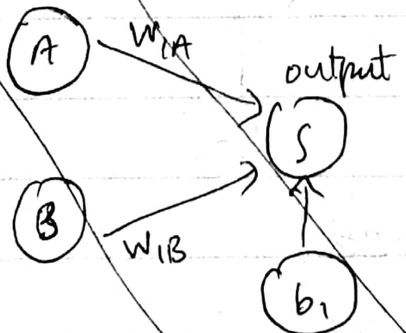
OR gate



{ ∫ : sigmoid }

$$\boxed{W_{1A} = 1.0, \quad W_{1B} = 1.0, \quad b_1 = -0.5}$$

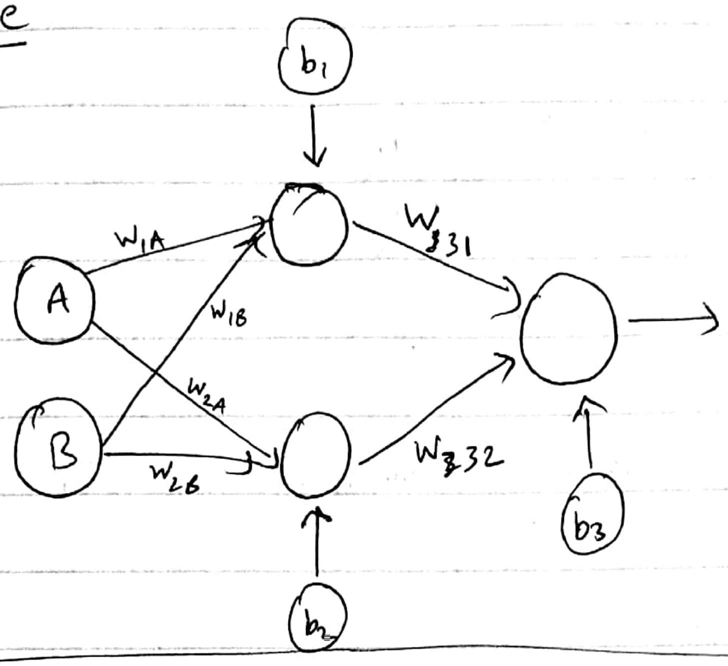
XOR gate



{ ∫ : sigmoid }

$W_{1A} = -0.4$, W_{2A}

XOR gate

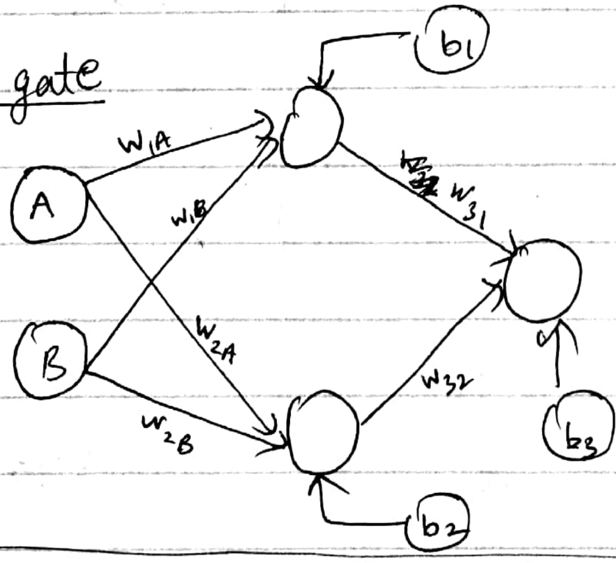


$W_{1A} = -0.4$, $W_{1B} = -0.4$, $b_1 = 0.6$

$W_{2A} = 1.0$, $W_{2B} = 1.0$, $b_2 = -0.5$

$W_{31} = 1.0$, $W_{32} = 1.0$, $b_3 = -1.5$

XNOR gate



$W_{1A} = 1.0$, $W_{1B} = 1.0$, $b_1 = -1.5$

$W_{2A} = -0.7$, $W_{2B} = -0.7$, $b_2 = 0.6$

$W_{31} = 1.0$, $W_{32} = 1.0$, $b_3 = -0.5$

3) a) $E(\mathbf{o}) = -\sum_{i=1}^{K} y_i \log o_i$

$$\nabla_{f_2} E = \left(\nabla_{f_2} O\right)\left(\nabla_{\mathbf{O}} E\right)$$

Now, $\nabla_O E = \begin{bmatrix} \partial E/\partial o_1 \\ \vdots \\ \partial E/\partial o_i \\ \vdots \\ \partial E/\partial o_k \end{bmatrix} = \begin{bmatrix} -y_1/o_1 \\ \vdots \\ \vdots \\ -y_k/o_k \end{bmatrix}$

$$\Rightarrow \nabla_O E_i = -y_i/o_i$$

and $\nabla_{f_2} O = \begin{bmatrix} \frac{\partial S(f_{21})}{\partial f_{21}} & \cdots & \frac{\partial S(f_{2k})}{\partial f_{21}} \\ \vdots & & \\ \frac{\partial S(f_{21})}{\partial f_{2k}} & \cdots & \frac{\partial S(f_{2k})}{\partial (f_{2k})} \end{bmatrix}$

$$\nabla_{f_{2j}} O_i = \frac{\partial S(f_{2i})}{\partial f_{2j}} = \cancel{\frac{e^{f_{2i}}\left(\frac{\partial x_i}{\partial x_j}\right)e^{f}}{\left(\sum_{p=1}^{K} e^{x_p}\right)^2}}$$

$$= \frac{e^{f_{2i}}}{\sum_{p=1}^{k} e^{f_{2p}}}\left(\frac{\partial f_{2i}}{\partial f_{2j}}\right) - \frac{e^{f_{2i}}\, e^{f_{2j}}}{\left(\sum_{p=1}^{k} e^{f_{2p}}\right)^2}$$

if $i=j$, $\Rightarrow \frac{\partial f_{2i}}{\partial f_{2j}} = 1 \Rightarrow \nabla_{f_{2i}} O_i = \frac{e^{f_{2i}}}{\sum_{p=1}^{k} e^{f_{2p}}}\left[1 - \left(\frac{e^{f_{2i}}}{\sum_{p=1}^{k} e^{f_{2p}}}\right)\right]$

$\cancel{\Rightarrow \frac{\partial f_{2i}}{\partial f_{2i}} =}$

Scanned by CamScanner

$$\Rightarrow \quad \nabla_{f_{2i}} O_i = O_i(1-O_i)$$

<u>if $i \neq j$</u>, $\Rightarrow \dfrac{\partial f_{2i}}{\partial f_{2j}} = 0 \Rightarrow \nabla_{f_{2j}} O_i = -O_i O_j$

$$\Rightarrow \quad \nabla_{f_2} O = \begin{bmatrix} O_1(1-O_1) & \cdots & -O_1 O_k \\ \vdots & & \\ -O_1 O_k & \cdots & O_k(1-O_k) \end{bmatrix}$$

$$\text{or} \quad \nabla_{f_{2j}} O_i = \begin{cases} O_i(1-O_i) & ; \text{ if } i = j \\ -O_i O_j & ; \text{ if } i \neq j \end{cases}$$

$$\Rightarrow \quad \nabla_{f_2} E = \begin{bmatrix} O_1(1-O_1) & \cdots & O_1 O_k \\ \vdots & & \vdots \\ -O_1 O_k & & O_k(1-O_k) \end{bmatrix} \begin{bmatrix} -y_1/O_1 \\ \vdots \\ -y_k/O_k \end{bmatrix}$$

~~In classification, Let $y_q = 1$ and $y_{i\neq q} = 0$~~

$\cancel{\Rightarrow \nabla_{f_2} E =}$

$$\cancel{\Rightarrow \nabla_{f_{2j}} E_\xi = \left[\sum_{\substack{p=1,\\ p\neq j}}^{K} O_p (1-O_p)\left(-\frac{y_p}{O_p}\right) \right] O_j O}$$

$$\Rightarrow \nabla_{f_{2i}} E = \left[\sum_{\substack{j=1,\\ j\neq i}}^{K} (-o_i o_j)\left(-\frac{y_j}{o_j}\right)\right] + o_i(1-o_i)\left(-\frac{y_i}{o_i}\right)$$

$$\Rightarrow \nabla_{f_{2i}} E = \sum_{\substack{j=1,\\ j\neq i}}^{K} y_j o_i + y_i o_i - y_i$$

$$\Rightarrow \boxed{\nabla_{f_{2i}} E = \left[\sum_{j=1}^{K} y_j o_i\right] - y_i}$$

b) $\nabla_x E = (\nabla_x f_1)(\nabla_{f_1} a)(\nabla_a f_2)(\nabla_{f_2} E)$

Now $\rightarrow \nabla_x f_1 = \nabla_x (xW_1 + b_1) = W_1 \quad -(1)$

$\rightarrow \nabla_{f_1} a = \nabla_{f_1} \sigma(f_1) = \cancel{\sigma(f_1)(\mathbb{1}_M - \sigma(f_1))}\ \sigma(f_1) I (I - \sigma(f_1) I)$

$\Rightarrow \nabla_{f_1} a = aI(I - aI) \quad -(2)$ $\quad \left\{ \cancel{\mathbb{1} = [1\ 1 \ldots 1]^T}_{M\ \text{times}} \right\}$

$\rightarrow \nabla_a f_2 = \nabla_a (aW_2 + b_2)$

$\Rightarrow \nabla_a f_2 = W_2$

$$\Rightarrow \boxed{\nabla_x E = W_1\, aI(I - aI)\, W_2\, \nabla_{f_2} E}$$

4) ~~a)~~ Given the feature map has same Height and width (H_{in})

$$H_{out} = \frac{H_{in} + \cancel{2 \times p} \; 2p - d(k-1) - 1}{s} + 1 \quad -①$$

a) given $s = 1$, d = dialation = 1, and $H_{out} = H_{in}$, and $k = 3$

$$\cancel{s}(H_{in}) = \frac{H_{in} + 2p - (3-1) - 1}{1} + 1$$

$$2p = \cancel{3}\, 2 \Rightarrow \boxed{p = 1}$$

$$F_{out} = \begin{bmatrix} 13.5 & -18.5 & 9 & -12.0 \\ 4 & 21.5 & 5 & 11.5 \\ 20.5 & 10.5 & 24.5 & 17 \\ 6 & 24 & 17 & 13.5 \end{bmatrix}$$

b) for filter 2, $k = 2$

Using ① with $H_{out} = H_{in} = 4$, $p = 1$, $d = 1$

$$4 \; \cancel{H_{out}} = \frac{\overset{4}{\cancel{H_{in}}} + 2 - 1(2-1) - 1}{s} + 1$$

$$3 = \frac{4}{s} \Rightarrow \boxed{s = \frac{4}{3}} \Rightarrow$$ s is not an integer so it is not possible to get a feature map with $H_{out} = H_{in}$

Scanned by CamScanner

c) $F' = \begin{bmatrix} 13.5 & -18.5 & 9 & -12 \\ 4 & 21.5 & 5 & 11.5 \\ 20.5 & 10.5 & 24.5 & 17 \\ 6 & 24 & 17.0 & 13.5 \end{bmatrix}$

$\text{Avgpool}(F') = \begin{bmatrix} 5.125 & 3.375 \\ 15.25 & 18.0 \end{bmatrix}$

d) convolution filter for average pooling,

$\text{Filter} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$

padding = 0 , stride = 2