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

$$\text{GELU}(x) \approx x\,\sigma(1.702x)$$

$$\boxed{\text{GELU}(x_0) = 0} \; ; \; x_0 = 0$$

$$\frac{\partial\, \text{GELU}(x)}{\partial x} = \sigma(1.702x) + x\,\sigma'(1.702x)$$

$$= \sigma(1.702x) + 1.702x\,\sigma(1.702x)(1-\sigma(1.702x))$$

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

Using GD:

$$x_1 = x_0 - \eta \left.\frac{\partial\, \text{GELU}(x)}{\partial x}\right|_{x_0}$$

$$\boxed{x_1 = 0 - 0.1\left[\frac{1}{2}\right] = -0.05}$$

$$\text{GELU}(x_1) = -0.05\,\sigma(1.702(-0.05))$$

$$\boxed{= -0.024}$$

$$x_2 = x_1 - \eta \frac{\partial GELU(x)}{\partial x}\bigg|_{x_1}$$

$$= -0.05 - 0.1\left[0.48\left[1 + 1.702(-0.05) - 1.702(-0.05)\,0.48\right]\right]$$

$$\boxed{= -0.096}$$

$$Gelu(x_2) = -0.096\,\sigma(1.702(-0.096))$$

$$\boxed{= -0.044}$$

$$x_3 = x_2 - \eta \frac{\partial GELU(x)}{\partial x}\bigg|_{x_2}$$

$$= -0.096 - 0.1\left[0.46\left[1 + 1.702(-0.096) - 1.702(-0.096)\,0.46\right]\right]$$

$$= -0.138$$

$$Gelu(x_3) = -0.138\,\sigma(1.702(-0.138))$$

$$\boxed{= -0.061}$$

b) Using learning rate 1 instead of 0.1

$$x_1 = x_0 - \eta \left.\frac{\partial GELU(x)}{\partial x}\right|_{x_0}$$

$$x_1 = 0 - 1\left[\frac{1}{2}\right] \boxed{= -0.5}$$

$$GELU(x_1) = -0.5\ \sigma(1.702(-0.5))$$
$$= -0.15$$

$$x_2 = x_1 - \eta \left.\frac{\partial GELU(x)}{\partial x}\right|_{x_1}$$

$$= -0.5 - 1 \times 0.12$$
$$\boxed{= -0.62}$$

$$Gelu(x_2) = -0.62\ \sigma(1.702(-0.62))$$
$$= -0.16$$

$$x_3 = x_2 - \eta \left.\frac{\partial GELU(x)}{\partial x}\right|_{x_2}$$

$$= -0.62 - 1 \times 0.056$$

$$\boxed{= -0.68}$$

$$\text{Gelu}(x_3) = -0.68\,\sigma(1.702(-0.68))$$
$$= -0.162$$

with a higher learning rate the function decreases faster and achieves a lower value of -0.162 compared to -0.061

1c) Using learning rate 0.1 & $x_0 = -3$

Normal GD

$$x_1 = x_0 - \eta \left.\frac{\partial \text{GELU}(x)}{\partial x}\right|_{x_0}$$

$$\boxed{x_1 = -3 - .1(-0.0245) = -2.997}$$

$$\text{GELU}(x_1) = -2.997\,\sigma(1.702(-2.997))$$
$$\boxed{= -0.0181}$$

$$x_2 = x_1 - \eta \left.\frac{\partial \text{GELU}(x)}{\partial x}\right|_{x_1}$$

$$= -2.997 - .1\times(-0.0246)$$
$$\boxed{= -2.9951}$$

$$\text{Gelu}(x_2) = -2.9951\,\sigma(1.702(-2.9951))$$
$$\boxed{= -0.0182}$$

$$x_3 = x_2 - \eta \left.\frac{\partial GELU(x)}{\partial x}\right|_{x_2}$$

$$= -2.9951 - 1 \times (-0.0247)$$

$$\boxed{= -2.993}$$

$$Gelu(x_3) = -2.993\,\sigma(1.702(-2.993))$$

$$\boxed{= -0.0183}$$

– GD with momentum $(\beta = 0.9)$

$$v_0 = \left.\frac{\partial\, Gelu(x)}{\partial x}\right|_{x_0} = -0.0245$$

$$v_1 = 0.9 v_0 + 0.1 \left.\frac{\partial\, Gelu(x)}{\partial x}\right|_{x_0}$$

$$= -0.0245$$

$$x_4 = x_0 - 0.1 v_1$$

$$= 3 - 0.1(-0.0245)$$

$$\boxed{= -2.997}$$

$$Gelu(x_1) = -2.997\,\sigma(1.702(-2.997))$$
$$\boxed{= -0.0181}$$

$$v_2 = 0.9 v_1 + 0.1 \left.\frac{\partial Gelu(x)}{\partial x}\right|_{x_1}$$
$$= 0.9(-0.0245) + 0.1 \times (-0.0246)$$
$$= -0.0246$$

$$x_2 = x_1 - 0.1 v_2$$
$$= -2.997 - 0.1 \times (-0.0246)$$
$$\boxed{= -2.995}$$

$$\boxed{Gelu(x_2) = -0.0182}$$

$$v_3 = 0.9 v_2 + 0.1 \times \left.\frac{\partial Gelu(x)}{\partial x}\right|_{x_2}$$
$$= 0.9(-0.0246) + 0.1 \times (-0.0247)$$
$$= -0.0246$$

$$x_3 = x_2 - 0.1 V_3$$
$$= -2.995 - 0.1(-0.0246)$$
$$\boxed{= -2.9926}$$

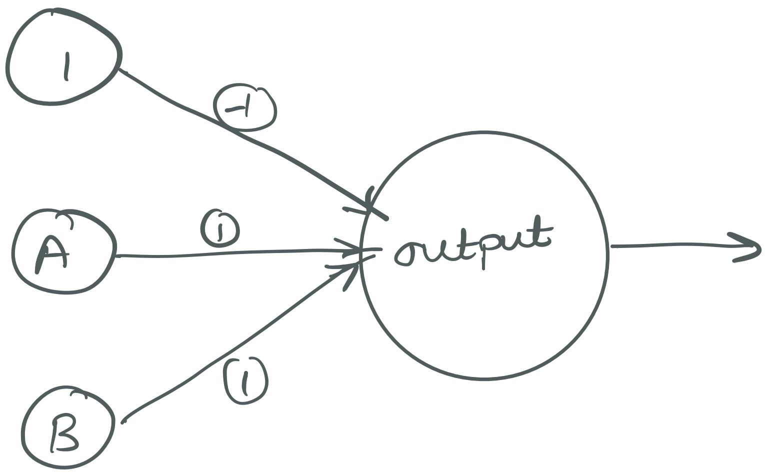
$$\boxed{Gelu(x_3) = -0.0183}$$

From the values we can see that GD with momentum is slightly faster however since we are in a flat region of the Gelu function & we are using a relatively small step size; the difference between the two methods is minor

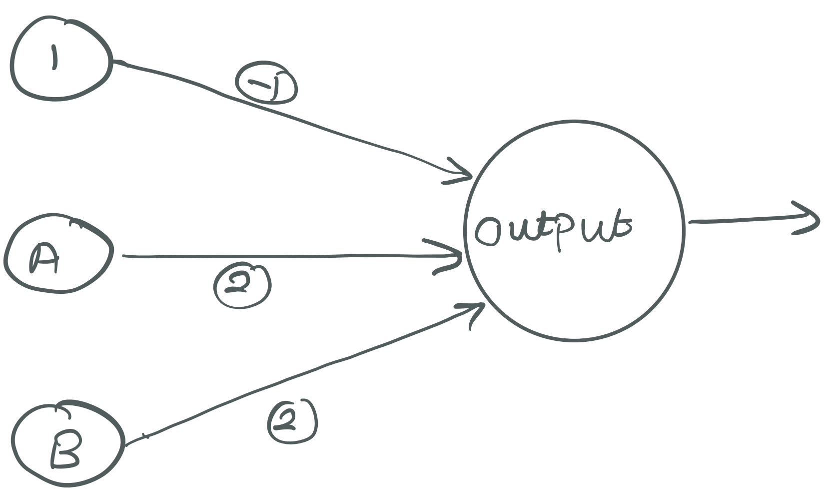
Further the momentum smooths out the update & so GD is less affected by sharp changes in the derivative

2) I have considered gate will activate for greater than 0.5

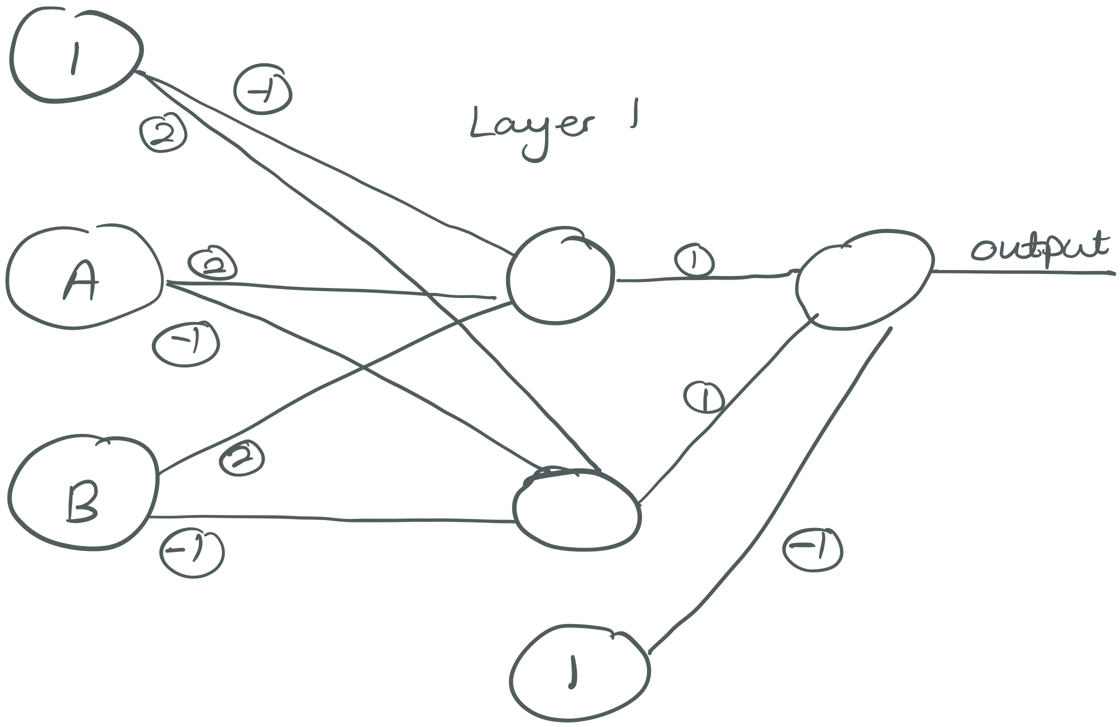
For AND Gate



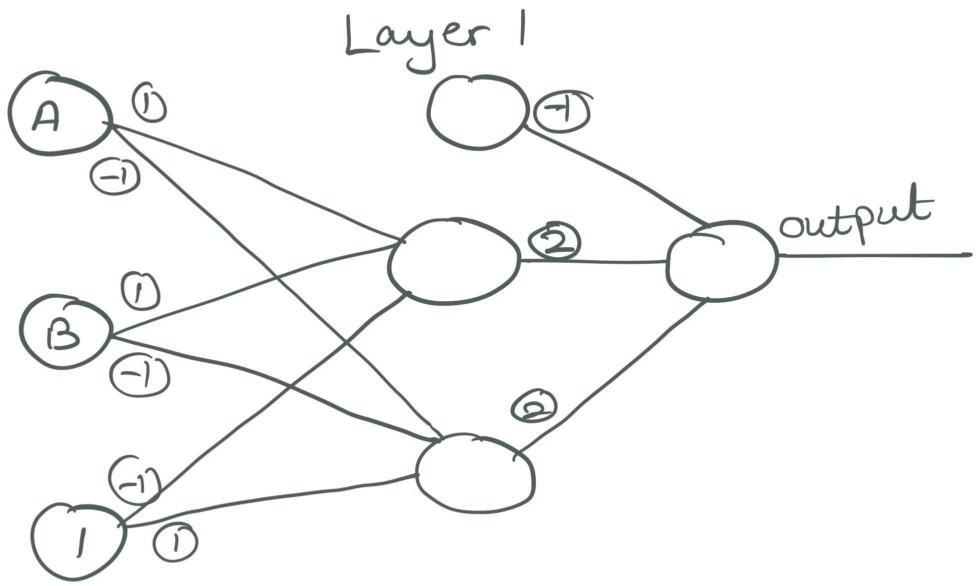
For OR Gate



For XOR Gate



For XNOR Gate



3)

$$f_1 = xW_1 + b_1$$

$$a = \sigma(f_1)$$

$$= \frac{1}{1+e^{-(xW_1+b_1)}}$$

$$f_2 = aW_2 + b_2$$

$$O = S(f_2)$$

$$E(O) = -\sum_{1}^{K} y_i \log o_i$$

$$\frac{\partial E(x)}{\partial f_2} = \frac{\partial E(O)}{\partial O} \cdot \frac{\partial S(f_2)}{\partial f_2}$$

$$= E'(O) \cdot S'(f_2)$$

$$= -\sum_{K} y_K \cdot \frac{1}{O_K} \cdot S'(f_2)$$

$$= -\sum_{K} y_K \cdot \frac{\sum_j \exp(f_2)}{\exp(f_{2-K})} \cdot S'(f_2)$$

$$= -\sum_{K} y_K \cdot \frac{\sum_j \exp(f_{2-j})}{\exp(f_{2-K})} \cdot S(f_2)\left(\delta_{ij} - S(f_2)\right)$$

where δ is Kronecker delta.

If we substitute f_2 as aW_2+b & simplify:

$$\frac{\partial E}{\partial a_i} = -y_i(1-O_i) - \sum_{k \neq i} y_k \frac{1}{O_k}(-o_k o_i)$$

$$= -y_i + y_i o_i + \sum_{k \neq 1} y_k o_i$$

$$\boxed{= o_i - y_i}$$ for one-hot encoded vector

Using (1) we calculate the $\frac{\partial E(x)}{\partial x}$

$$\frac{\partial E(x)}{\partial x} = \frac{\partial E}{\partial a} \times \frac{1}{w_2} \times \left(\frac{w_1 w_2 \exp(-b_1 - w_1 x)}{(\exp(-b_1 - w_1 x) + 1)^2} \right)$$

$$= \begin{bmatrix} \frac{\partial E}{\partial a_1} \\ \frac{\partial E}{\partial a_2} \\ \vdots \\ \frac{\partial E}{\partial a_k} \end{bmatrix} \times \left(\frac{w_1 \exp(-b_1 - w_1 x)}{(\exp(-b_1 - w_1 x) + 1)^2} \right)$$

$$= \begin{bmatrix} o_1 - y_1 \\ o_2 - y_2 \\ \vdots \\ o_k - y_k \end{bmatrix} \times \left(\frac{w_1 \exp(-b_1 - w_1 x)}{(\exp(-b_1 - w_1 x) + 1)^2} \right)$$

4) $F = \begin{bmatrix} 3 & 5 & 2 & 3 \\ 9 & 1 & 8 & 4 \\ 6 & 4 & 3 & 7 \\ 7 & 0 & 2 & 4 \end{bmatrix}$

Filter 1 = $\begin{bmatrix} -1 & 0.5 & -2 \\ 2 & 0 & 1 \\ 0 & 1 & 1.5 \end{bmatrix}$

Filter 2 = $\begin{bmatrix} -1 & 0.5 \\ 2 & 0 \end{bmatrix}$

a) $s = 1$

$$\text{output} = \frac{(W - K + 2P)}{S} + 1$$

since we want same output size as input

we have: $W = 4$, $k = 3$

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

$$3 = 1 + 2P$$

$$2P = 2$$

$$P = 1$$

$\therefore$ we need a zero padding of size 1 to get output of same size as input

The output F' after convolution is :
↳ (cross-correlation)

$$F' = \begin{bmatrix} 15.5 & 21 & 27 & 8 \\ 4.5 & 30 & 9.5 & 22.5 \\ 13.5 & -6.5 & 18 & 4 \\ -5 & 6 & -12.5 & 4.5 \end{bmatrix}$$

b) $p = 1$ (given)

If we want output of same size :

$$4 = \frac{4 - 2 + 2\times 1}{s} + 1 \qquad (\because W = 4,\ K = 2,\ P = 1)$$

$$3 = \frac{4}{s}$$

$$\Rightarrow s = 4/3$$

Thus it is not possible to get an output of the same size as the input because s is not an integer.

i.e. we would need a fractional stride size which does not make physical sense.

c) Carrying out average pooling of size 2 & stride 2 we get :

$$\text{output} = \begin{bmatrix} 5.125 & 3.375 \\ 15.25 & 18 \end{bmatrix}$$

d) If we use a filter

$$f_3 = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}$$

with our calculated F' with no padding and a stride of 2