Problem ①

Given,

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

$$GELU \approx x\sigma(1.702x) \approx \frac{x}{1+e^{-1.702x}}$$

$$\frac{\partial(GELU)}{\partial x} = \frac{\left[\frac{\partial x}{\partial x} \times (1+e^{-1.702x})\right] - \left[x \frac{\partial}{\partial x}\left[1+e^{-1.702x}\right]\right]}{(1+e^{-1.702x})^2}$$

.... gradient with respect to x of GELU

$$\therefore \frac{\partial(GELU)}{\partial x} = \frac{1+e^{-1.702x}(1+1.702x)}{(1+e^{-1.702x})^2}$$

We know that,

$$x_{t+1} = x_t - \eta \frac{\partial(GELU)}{\partial x} \quad \text{--- ①} \quad \eta = \text{learning rate}$$

(a) Initial guess $x_0 = 0$

$$GELU\big|_{x_0} = x\sigma(1.702x) = 0$$

$$\eta = 0.1$$

To calculate x_1 we use equation (1)

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_0} = \frac{1+e^{0}(1+0)}{(1+e^{0})^2} = \frac{2}{4} = 0.5$$

$$\therefore x_1 = x_0 - \eta \left.\frac{\partial GELU}{\partial x}\right|_{x_0} = 0 - 0.1 \times 0.5$$

$$x_1 = -0.05$$

$$GELU(x_1) = -0.02394$$

To calculate x_2 we have,

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_1} = \frac{1+e^{-1.702\times(-0.05)}\left(1+1.702(-0.05)\right)}{\left(1+e^{-1.702\times(-0.05)}\right)^2}$$

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_1} = 0.4575$$

$$\therefore x_2 = x_1 - \eta \left.\frac{\partial GELU}{\partial x}\right|_{x_1} = -0.05 - 0.1 \times 0.4575$$

$$x_2 = -0.09575$$

$$GELU(x_2) = -0.04398$$

To calculate x_3

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_2} = \frac{1+e^{-1.702\times(-0.09575)}\left(1+1.702(-0.09575)\right)}{\left(1+e^{-1.702\times(-0.09575)}\right)^2}$$

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_2} = 0.418875$$

$$x_3 = x_2 - \eta \times \left.\frac{\partial GELU}{\partial x}\right|_{x_2} = -0.13764$$

$$GELU(x_3) = -0.06079$$

(b) Initial guess $x_0 = 0$, $GELU(x_0) = 0$

$\eta = 1$,

Gradient of GELU w.r.t x when $x = x_0$ is

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_0} = +0.5$$

$$\therefore \quad x_1 = x_0 - \eta \left.\frac{\partial GELU}{\partial x}\right|_{x=x_0} = 0 - 1 \times 0.5 = -0.5$$

$$GELU(x_1) = -0.14961$$

Calculation of x_2 :

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_1} = 0.120778$$

$$x_2 = x_1 - \eta \left.\frac{\partial GELU}{\partial x}\right|_{x_1} = -0.62078$$

$$GELU(x_2) = -0.16014$$

Calculation of x_3 :

$$\left.\frac{\partial GELU}{\partial x}\right|_{x_2} = 0.055719$$

$$x_3 = x_2 - \eta \left.\frac{\partial GELU}{\partial x}\right|_{x_2} = -0.6765$$

$GELU(x_3) =$ -0.16252

When learning rate = 1, the model quickly reaches minima as compared to learning rate = 0.1, the gap between the value of x is greater with learning rate = 1. But larger learning rate can also end up bypassing the global minima.

c) Gradient Descent with momentum

(i) Initial Guess, $x_0 = -3$, $\eta = 0.1$

$GELU(x_0) = -0.01807$

$$\frac{\partial(GELU)}{\partial x}\Big|_{x_0} = -0.02455$$

$$x_1 = x_0 - \eta \frac{\partial(GELU)}{\partial x}\Big|_{x_0} = -2.99755$$

$GELU(x_1) = -0.01813$

Calculation of x_2:

$$\frac{\partial(GELU)}{\partial x}\Big|_{x_1} = -0.02462$$

$$x_2 = x_1 - \eta \frac{\partial(GELU)}{\partial x}\Big|_{x_1} = -2.99508$$

$GELU(x_2) = -0.01819$

Calculation of x_3:

$$\frac{\partial(GELU)}{\partial x}\Big|_{x_2} = -0.0247$$

$$x_3 = x_2 - \eta \frac{\partial(GELU)}{\partial x}\Big|_{x_2} = -2.99261$$

$GELU(x_2) = -0.01825$

(ii) Initial guess: $x_0 = -3$, $\eta = 0.1$, $\beta = 0.9$

$$v_0 = \nabla f_{x_0} = \frac{\partial (GELU)}{\partial x}\Big|_{x_0} = -0.02455, \quad GELU(x_0) = -0.01807$$

$$v_1 = \beta v_0 + (1-\beta)\nabla f_{x_0} = -0.02455$$

$$x_1 = x_0 - \eta v_0 = -3 - 0.1 \times (-0.02455)$$

$$x_1 = -2.99755$$

$$GELU(x_1) = -0.01813$$

$$\frac{\partial (GELU)}{\partial x}\Big|_{x_1} = -0.02462$$

$$x_2 = x_1 - \eta v_1 = -2.99755 - 0.1 \times (-0.02455)$$

$$x_2 = -2.99509$$

$$GELU(x_2) = -0.01819$$

$$\frac{\partial (GELU)}{\partial x}\Big|_{x_2} = -0.0247$$

$$v_2 = \beta v_1 + (1-\beta)\frac{\partial (GELU)}{\partial x}\Big|_{x_1} = -0.02456$$

$$x_3 = x_2 - \eta v_2 = -2.99263$$

$$GELU(x_3) = -0.01825$$

(iii) From the above results it can be seen that the performance of both the methods are same.

Problem 2

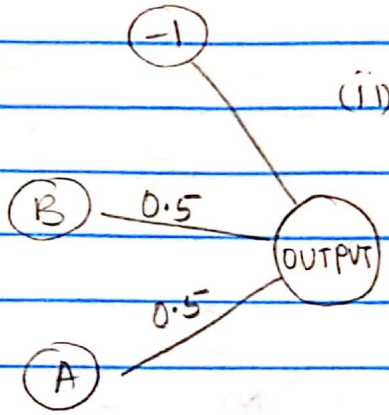
① AND Gate

(i) No. of layers → 1 input → 2 neurons
1 output → 1 neuron

(ii) Activation function → Sigmoid

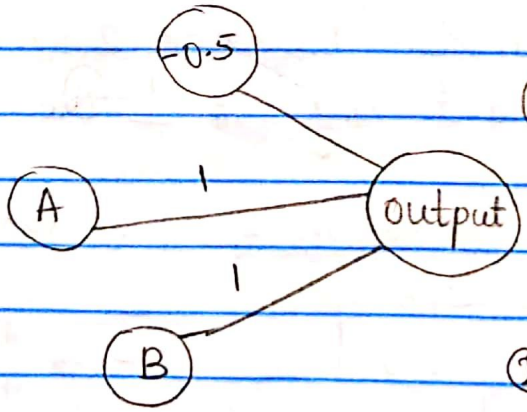
$\sigma(x) \geq 0.5 = 1$

$\sigma(x) < 0.5 = 0$



B	A	$\Sigma Wx+b$	$\sigma(\Sigma Wx+b)$	X
0	0	-1	0.269	0
0	1	-0.5	0.376	0
1	0	-0.5	0.376	0
1	1	0	0.5	1

② OR Gate



① No. of layers :-

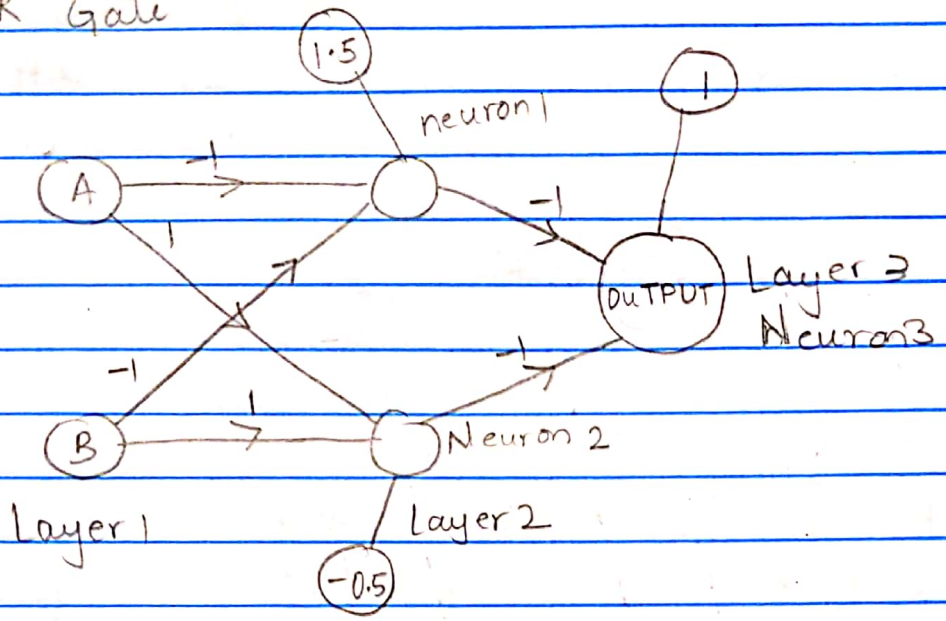
(i) One input layer → 2 Neurons

(ii) One output layer → 1 Neuron

② Activation function → Sigmoid

B	A	$\Sigma Wx+b$	$\sigma(\Sigma Wx+b)$	X
0	0	-0.5	0.376	0
0	1	0.5	0.622	1
1	0	0.5	0.622	1
1	1	1.5	0.817	1

③ XOR Gate

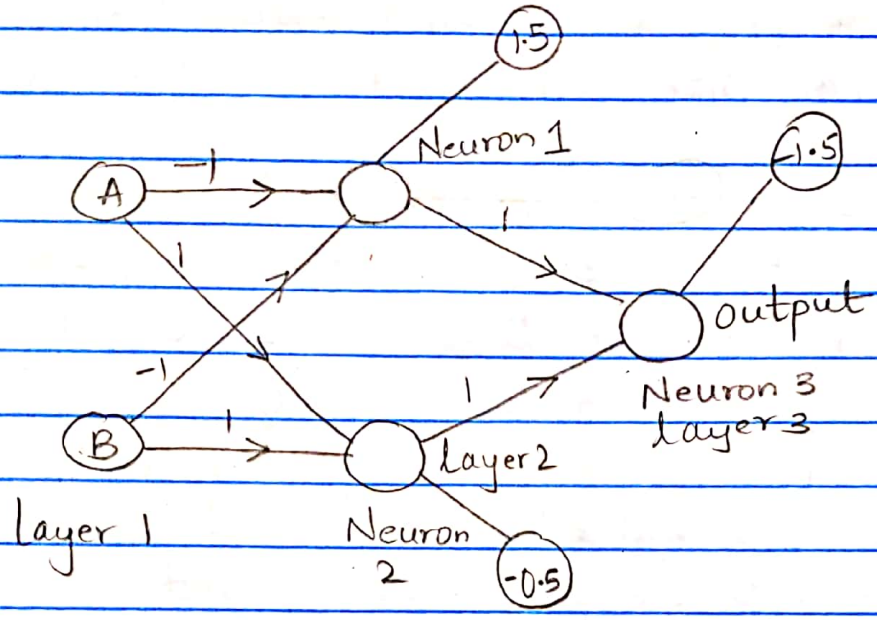


At layer 2, we use ReLU as the activation funtion
At layer 3 (output layer), we use sigmoid as activation function

Number of layers:- 1 input layer → 2 Neuron
1 Hidden layer → 2 Neurons
1 output layer → 1 Neuron

		NEURON ①		Neuron 2		Neuron 3		
A	B	$\Sigma W_n + b$	$\sigma(\Sigma W_n + b)$	$\Sigma W_n + b$	$\sigma(\Sigma W_n + b)$	$\Sigma W_n + b$	$\sigma(\Sigma W_n + b)$	X
0	0	1.5	1.5	-0.5	0	-0.5	0.378	0
0	1	0.5	0.5	0.5	0.5	0	0.5	1
1	0	0.5	0.5	0.5	0.5	0	0.5	1
1	1	-0.5	0	1.5	1.5	-0.5	0.378	0

④ XNOR Gate

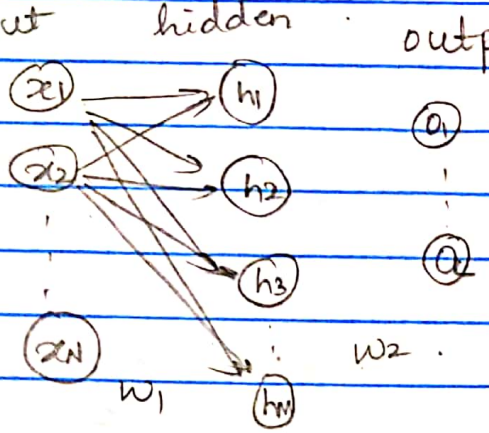


Number of layers- 1 input layer - 2 Neurons
1 Hidden layer - 2 Neuron
1 output layer - 1 Neuron.

- At layer 2, we use RELU as activation function
- At layer 3 (output layer), we use sigmoid as activation function

		Neuron 1		Neuron 2		Neuron 3		
A	B	$\Sigma Wx+b$	$\sigma(\Sigma Wx+b)$	$\Sigma Wx+b$	$\sigma(\Sigma Wx+b)$	$\Sigma Wx+b$	$\sigma(\Sigma Wx+b)$	X
0	0	1.5	1.5	-0.5	0	0	0.5	1
0	1	0.5	0.5	0.5	0.5	-0.5	0.378	0
1	0	0.5	0.5	0.5	0.5	-0.5	0.378	0
1	1	-0.5	0	1.5	1.5	0	0.5	1

Problem 3.



• Sigmoid for activation of hidden layer.

For the above neural network it is given,

$f_1 = xw_1 + b_1$ ——— ①

$a = \sigma(f_1)$ ——— ②

$f_2 = aw_2 + b_2$ ——— ③

$O = s(f_2)$ ——— ④

Cross entropy loss $(E(O)) = -\sum_i^K y_i \log o_i$ ——— ⑤.

(a) Loss gradient w.r.t f_2.

$$\frac{\partial E}{\partial f_2} = \frac{\partial E}{\partial O} \times \frac{\partial O}{\partial f_2} \quad \cdots \text{from chain rule.}$$

$$\frac{\partial E}{\partial o_i} = \frac{-y_i}{o_i}$$

$$\frac{\partial o_i}{\partial f_{2i}} = o_i(1-o_i) \quad \cdots \text{gradient of sigmoid.}$$

when $i = k$

$$\frac{\partial E}{\partial o_i} \times \frac{\partial o_i}{\partial f_{2i}} = \frac{-y_i}{o_i}(o_i)(1-o_i) = y_i o_i - y_i$$

When $i = k$

$$\frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial f_{2i}} = \frac{-y_k}{o_k}(-o_k o_i) = y_k o_i$$

$$\frac{\partial E}{\partial f_{2i}} = \sum_k \frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial f_{2i}} = \sum_k y_k o_i - y_i$$

Since y is one-hot vector $\sum_k y_k = 1$

$$\frac{\partial E}{\partial f_{2i}} = o_i - y_i \text{ , thus } \boxed{\frac{\partial E}{\partial f_2} = o - y}$$

(b) Loss gradient w.r.t x

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial f_2} \times \frac{\partial f_2}{\partial a} \times \frac{\partial a}{\partial f_1} \times \frac{\partial f_1}{\partial x} \quad \text{... from chain rule.}$$

$$\frac{\partial E}{\partial f_2} = o - y \quad \text{... calculated from above.}$$

$$\frac{\partial f_2}{\partial a} = w_2^T$$

$$\frac{\partial a}{\partial f_1} = a(1-a)$$

$$\frac{\partial E}{\partial x} = (o - y) * w_2^T \circ [a \circ (1 - a)] * w_1^T$$

$*$ → Matrix multiplication

$\circ$ → Element wise multiplication

Problem 4

(a) To get a new feature map F' with same size as the original F with stride $S=1$, the padding size must be 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

Feature map

-1	0.5	-2
2	0	1
0	1	1.5

Filter 1

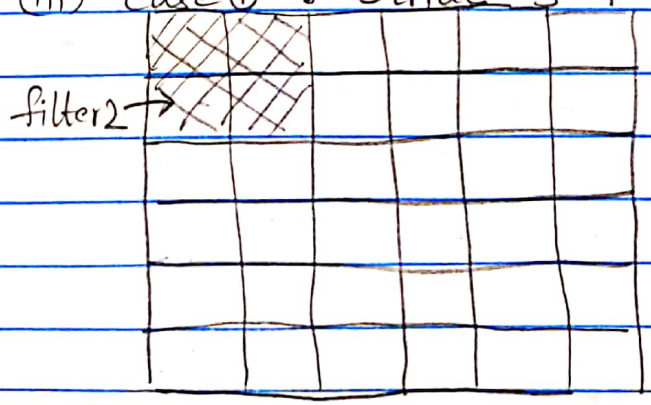
F' feature map $\Rightarrow$

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

(b)(i) With padding size $p=1$, the feature map dimension becomes 6×6.

(ii) The dimension of filter 2 is 2×2

(iii) Case ① : Stride $S=1$



$\Rightarrow$ We get a feature map with dimension 5×5.

(iv) Case ②: Stride $S=2$

With stride $s=2$, the new feature map dimension becomes 3×3

(v) From case ① & case ② we can see that dimension of 4×4 cannot be reached, which is the feature map actual dimension

(c) Given pooling size $=2$, Stride $s=2$

The feature map F' is given below.

On performing average pooling we get

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 →

F''

17·75	16·75
2	3·5

(d) In order to achieve feature map F'', the filter should be able to replicate the average pooling function. This is possible if the stride $S=2$ & filter is

$$\text{filter } 3 = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$$

$$F' \xrightarrow[S=2]{\text{filter } 3} F''$$