NNDL Problem Set 2

Divya Sai Sindhuja Vankineni

03 February 2025

Question 1.

Solution:

a) In a ML model, parameters are internal variables that are learned during training to minimize the error on the training data. Weights and biases are examples of parameters. Whereas, hyperparameters are external variables that are decided before training the model. They control the training process and are not learned from the data. Learning rate and batch size are examples of hyperparameters.

For arriving at optimal parameters that reduce the error, we use training dataset to train the model. Whereas, to decide optimal hyperparameters, we need to evaluate the model performance using another dataset that is not a training dataset. This results in forming a validation dataset that evaluates the model during training so that we can tune the hyperparameters accordingly.

b)

Item	P or HP?
A weight matrix **w**	P, because weights are learned during training to minimize the error.
The learning rate	HP, as it is decided before training and controls how the model updates parameters.
A bias term **b**	P, as bias term is learned during training to finally arrive at the optimal model.
The minibatch size	HP, because it controls how much data the model processes at once so chosen before the training.
The non-linear activation function	HP, as it is used to transform the input data at each layer so chosen before training.
The optimizer	HP, because it controls how the model's parameters are updated during training. So it has to be decided before training.

c) Number of Epochs is a hyperparameter that decides how many times the entire training dataset is passed through the model during training. It should be set before training and decides the training duration.

Momentum is another hyperparameter that is set within optimizers before training to control the acceleration of gradients in the correct direction so as to avoid local minima.

Question 2.

Solution:

a) The deep learning classification model seems to be **overfitting**. This conclusion can be drawn from the given two statements. The validation accuracy increases rapidly during first few epochs and then decreases suggesting that the model learned well during the first 5 epochs and then started to memorize the data, decreasing in its ability to generalize unseen (validation) data. Next, the training loss decreases to zero during the first 10 epochs indicating it is perfectly fitting the training data, leading to memorization and thus overfitting.

One possible modification to improve model generalization would be to employ **early stopping with patience**, i.e., a hyperparameter that decides how many epochs to wait before stopping. The goal is to stop training when the validation accuracy starts to decrease so that it prevents overfitting.

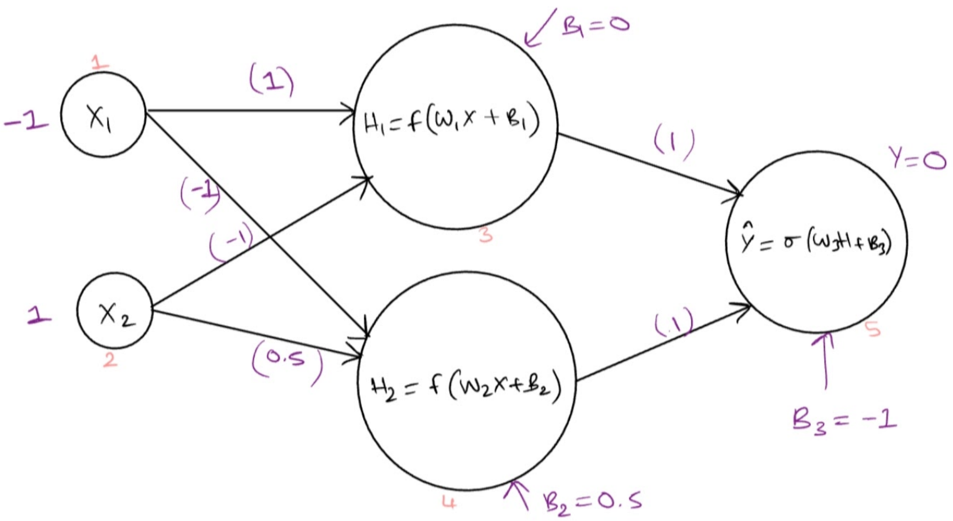
b) The deep learning classification model seems to be **underfitting**. This conclusion can be drawn from the given two statements. The validation accuracy, though trained over 100 epochs, is still slowly increasing implying slow learning of the model to generalize data. Next, the training loss decreases gradually during the entire training run but never approaches zero. This indicates that the model is not learning enough to capture the full complexity of data, leading to underfitting.

Underfitting occurs when the model is too simple to generalize the complex data. So increasing the **number of layers or neurons per layer** in the deep learning model will increase it's capacity to learn complex patterns from data to prevent underfitting.

Question 3.

Solution:

(3)(a)



Given, $W_1=[1,-1]$ $W_2=[-1,0.5]$ $W_3=[1,1]$

$B_1=0$ $B_2=0.5$ $B_3=-1$

(i) Forward Pass :

$$H_1=f(W_1X+B_1) = f([1,-1][-1,1]+0)$$

$$H_1 = f(-1-1) = f(-2)$$

$$H_1 = \max(-2, -0.2)$$

$$H_1 = -0.2$$

$$H_2 = f(W_2X+B_2) = f([-1,0.5][-1,1]+0.5)$$

$$H_2 = f(1+0.5+0.5) = f(2)$$

$$H_2 = \max(2, 0.2)$$

$$H_2 = 2$$

$$\hat{Y} = \sigma(W_3 H + B_3)$$

$$\hat{y} = \sigma\left([1,1][-0.2,2] + (-1)\right)$$

$$\hat{y} = \sigma\left(-0.2+2-1\right) = \sigma(0.8)$$

$$\boxed{\hat{y} = \frac{1}{1+e^{-0.8}} \approx 0.6899}$$

(ii) Binary Cross-Entropy Loss:

Binary Cross Entropy Loss $= -\left(Y\log(\hat{Y}) + (1-Y)\log(1-\hat{Y})\right) = E$

For $Y = 0$: Loss $= -\log(1-0.6899)$

$$\Rightarrow -\log(0.3101)$$

$$\boxed{L : \text{loss} \approx 1.1714}$$

(iii) Backward Propagation :

$$\frac{\partial L}{\partial \hat{Y}} = \frac{\partial\left(-(Y\log\hat{Y}) + (1-Y)\log(1-\hat{Y})\right)}{\partial \hat{Y}}$$

$$= \frac{\hat{Y}-Y}{\hat{Y}(1-\hat{Y})} = \frac{\hat{Y}}{\hat{Y}(1-\hat{Y})} = \frac{1}{1-\hat{Y}}$$

$$\frac{\partial L}{\partial \hat{Y}} \Rightarrow \frac{1}{1-Os} = \frac{1}{1-0.6899} = \frac{1}{0.3101} \approx 3.224$$

For $\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W_3} = \left(\frac{1}{1-\hat{y}}\right) \cdot \frac{\partial(\sigma(W_3H+B_3))}{\partial W_3}$

$$\Rightarrow \left(\frac{1}{1-\hat{y}}\right) \cdot \sigma(W_3H+B_3)(1-\sigma(W_3H+B_3)) \cdot H$$

$$\Rightarrow \left(\frac{1}{1-\hat{y}}\right)(\hat{y})(1-\hat{y}) \cdot H$$

$$\nabla W_3 \Rightarrow [\hat{y} . H_1 , \hat{y} . H_2] = [0.6899(-0.2), 0.6899(2)]$$

$$\boxed{\nabla W_3 = [-0.13798, 1.3798]}$$

For $\frac{\partial L}{\partial B_3} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial B_3} = \left(\frac{1}{1-\hat{y}}\right) \cdot \sigma(W_3H+B_3)(1-\sigma(W_3H+B_3)) . (1)$

$$\Rightarrow \left(\frac{1}{1-\hat{y}}\right)(\hat{y})(1-\hat{y}) = \hat{y}$$

$$\boxed{\nabla B_3 = 0.6899}$$

For $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial H_2} \cdot \frac{\partial H_2}{\partial W_2} = \left(\frac{1}{1-\hat{y}}\right)(\hat{y})(1-\hat{y})(1) \frac{\partial(f(W_2X+B_2))}{\partial W_2}$

$$\Rightarrow \hat{y} . X \quad [\because W_2X+B_2 \geq 0]$$

$$\Rightarrow (0.6899)[-1, 1]$$

$$\boxed{\nabla W_2 = [-0.6899, 0.6899]}$$

For $\dfrac{\partial L}{\partial B_2} = \dfrac{\partial L}{\partial \hat{y}} \cdot \dfrac{\partial \hat{y}}{\partial H_2} \cdot \dfrac{\partial H_2}{\partial B_2} = (\hat{y})(w_3)\dfrac{\partial (f(w_2x+B_2))}{\partial B_2}$

$\Rightarrow (\hat{y})(1)(1)$

$$\boxed{\nabla B_2 = 0.6899}$$

For $\dfrac{\partial L}{\partial H_2} = \dfrac{\partial L}{\partial \hat{y}} \cdot \dfrac{\partial \hat{y}}{\partial H_2} = \hat{y}\,(1)$

$$\boxed{\nabla H_2 = 0.6899}$$

For $\dfrac{\partial L}{\partial H_1} = \dfrac{\partial L}{\partial \hat{y}} \cdot \dfrac{\partial \hat{y}}{\partial H_1} = (\hat{y})(1)$

$$\boxed{\nabla H_1 = 0.6899}$$

For $\dfrac{\partial L}{\partial w_1} = \dfrac{\partial L}{\partial \hat{y}} \cdot \dfrac{\partial \hat{y}}{\partial H_1} \cdot \dfrac{\partial H_1}{\partial w_1} = \hat{y} \cdot \dfrac{\partial (f(w_1x+B_1))}{\partial w_1}$

$\Rightarrow \hat{y}.\,(0.1)\,[-1,1] \qquad [\because w_1x+B_1 < 0]$

$$\boxed{\nabla w_1 = [-0.06899,\ 0.06899]}$$

For $\dfrac{\partial L}{\partial B_1} = \dfrac{\partial L}{\partial \hat{y}} \cdot \dfrac{\partial \hat{y}}{\partial H_1} \cdot \dfrac{\partial H_1}{\partial B_1} = \hat{y}.\,\dfrac{\partial (f(w_1x+B_1))}{\partial B_1}$

$\Rightarrow \hat{y}\,(0.1)$

$$\boxed{\nabla B_1 = 0.06899}$$

for $\frac{\partial L}{\partial X_1} = \frac{\partial L}{\partial H_1} \cdot \frac{\partial H_1}{\partial X_1} + \frac{\partial L}{\partial H_2} \cdot \frac{\partial H_2}{\partial X_1}$

$$\Rightarrow 0.6899 \left[\frac{\partial}{\partial X_1} \left(f(W_1 X + B_1) \right) + \frac{\partial \left(f(W_2 X + B_2) \right)}{\partial X_1} \right]$$

$$\Rightarrow 0.6899 \left[(0.1)(1) + (1)(-1) \right]$$

$$\Rightarrow 0.6899 (-0.9)$$

$$\boxed{\nabla X_1 = -0.62}$$

for $\frac{\partial L}{\partial X_2} = \frac{\partial L}{\partial H_1} \cdot \frac{\partial H_1}{\partial X_2} + \frac{\partial L}{\partial H_2} \cdot \frac{\partial H_2}{\partial X_2}$

$$\Rightarrow 0.6899 \left[(0.1)(-1) + (1)(0.5) \right]$$

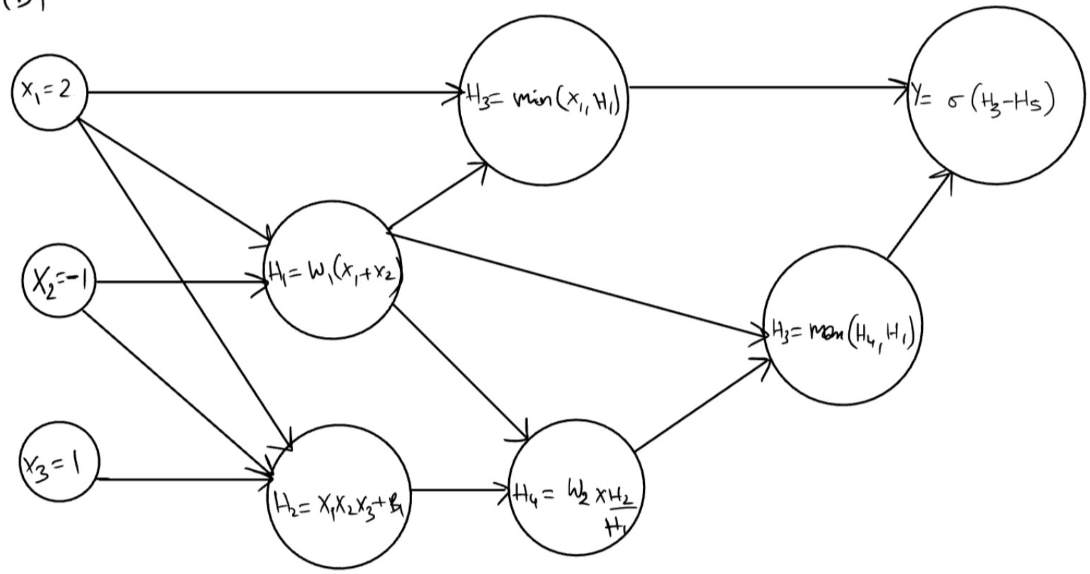
$$\boxed{\nabla X_2 \Rightarrow 0.276}$$

Given table filled with obtained values:

$\hat{Y}$	L	$\triangledown \mathbf{W_3}$	$\triangledown \mathbf{B_3}$	$\triangledown \mathbf{W_2}$	$\triangledown \mathbf{B_2}$	$\triangledown H_2$
0.6899	1.1714	[-0.13798, 1.3798]	0.6899	[-0.6899, 0.6899]	0.6899	0.6899

$\triangledown \mathbf{W_1}$	$\triangledown \mathbf{B_1}$	$\triangledown H_1$	$\triangledown \mathbf{X_1}$	$\triangledown \mathbf{X_2}$
[-0.06899, 0.06899]	0.06899	0.6899	-0.62	0.276

3) (b)



Given, $W_1 = 1.5$

$B_1 = 1$ And, $Y = 1$

$W_2 = -3$

(i) Forward Pass :

$$H_1 = W_1(X_1+X_2) = 1.5(2-1) = 1.5$$

$$H_2 = X_1X_2X_3 + B_1 = 2(-1)(1)+1 = -1$$

$$H_3 = \min(X_1, H_1) = \min(2, 1.5) = 1.5$$

$$H_4 = W_2 \times \frac{H_2}{H_1} = -3\times\frac{(-1)}{1.5} = 2$$

$$H_5 = \max(H_4, H_1) = \max(2, 1.5) = 2$$

$$\hat{Y} = \sigma(H_3 - H_5) = \sigma(1.5-2) = \sigma(-0.5)$$

$$\therefore \boxed{\hat{Y} = \frac{1}{1+e^{+0.5}} \approx 0.3775}$$

(ii) Binary Cross-Entropy Loss:

Binary Cross Entropy Loss $= -\left(Y\log(\hat{Y}) + (1-Y)\log(1-\hat{Y})\right) = E$

For $Y = 1$: Loss $= -Y\log(\hat{Y}) = -1\log(0.3775)$

$\Rightarrow 0.9744$

$$\boxed{L : \text{Loss} = 0.9744}$$

(iii) Backward Propagation:

$$\frac{\partial L}{\partial \hat{Y}} = \frac{\hat{Y}-Y}{\hat{Y}(1-\hat{Y})} = \frac{\hat{Y}-1}{\hat{Y}(1-\hat{Y})} \quad [\text{Taking } Y=1]$$

$$\frac{\partial L}{\partial \hat{Y}} = \frac{-1}{\hat{Y}}$$

For $\frac{\partial L}{\partial H_5} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial H_5} = \left(\frac{-1}{\hat{Y}}\right) \cdot \frac{\partial(\sigma(H_3-H_5))}{\partial H_5}$

$$\Rightarrow \frac{-1}{\hat{Y}} \cdot \sigma(H_3-H_5)\left(1-\sigma(H_3-H_5)\right)$$

$$\Rightarrow \frac{-1}{\hat{Y}} \cdot \hat{Y}.(1-\hat{Y}) \Rightarrow \hat{Y}-1 = 0.3775-1$$

$$\boxed{\nabla H_5 = 0.6225}$$

For $\frac{\partial L}{\partial H_4} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial H_5} \cdot \frac{\partial H_5}{\partial H_4} = \frac{\partial L}{\partial H_5} \cdot \frac{\partial (\max(H_4, H_1))}{\partial H_4}$

$\Rightarrow (0.6225)(1)$

$\boxed{\nabla H_4 = 0.6225}$

For $\frac{\partial L}{\partial H_3} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial H_3} = \left(\frac{-1}{\hat{y}}\right) \cdot \frac{\partial (\sigma(H_3 - H_5))}{\partial H_3}$

$\Rightarrow \frac{-1}{\hat{y}} \cdot \sigma(H_3 - H_5)(1 - \sigma(H_3 - H_5))$

$\Rightarrow \frac{-1}{\hat{y}} \cdot \hat{y}(1 - \hat{y}) = \hat{y} - 1 = 0.3775 - 1$

$\boxed{\nabla H_3 = -0.6225}$

For $\frac{\partial L}{\partial H_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial H_5} \cdot \frac{\partial H_5}{\partial H_4} \cdot \frac{\partial H_4}{\partial H_2} = \frac{\partial L}{\partial H_4} \cdot \frac{\partial}{\partial H_2}\left(\frac{W_2 H_2}{H_1}\right)$

$\Rightarrow (0.6225)\left(\frac{W_2}{H_1}\right) \Rightarrow 0.6225 \times \frac{(-3)}{1.5}$

$\boxed{\nabla H_2 = -1.245}$

For $\frac{\partial L}{\partial H_1} = \frac{\partial L}{\partial H_3} \cdot \frac{\partial H_3}{\partial H_1} + \frac{\partial L}{\partial H_5} \cdot \frac{\partial H_5}{\partial H_1} + \frac{\partial L}{\partial H_4} \cdot \frac{\partial H_4}{\partial H_1}$

$$\Rightarrow (0.6225)\left[-\frac{\partial(\min(x_1, H_1))}{\partial H_1} + \frac{\partial(\max(H_4, H_1))}{\partial H_1} + \frac{\partial}{\partial H_1}\left(\frac{W_2 H_2}{H_1}\right)\right]$$

$$\Rightarrow 0.6225\left[-1 + 1 + \left(-\frac{W_2 H_2}{H_1^2}\right)\right]$$

$$\Rightarrow 0.6225\left(-\frac{(-3)(-1)}{(1.5)^2}\right) \Rightarrow 0.6225\,(-1.33)$$

$$\boxed{\nabla H_1 = -0.828}$$

For $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial H_4} \cdot \frac{\partial H_4}{\partial W_2} = (0.6225)\frac{\partial}{\partial W_2}\left(\frac{W_2 H_2}{H_1}\right)$

$$\Rightarrow (0.6225)\frac{H_2}{H_1} = \frac{-0.6225}{1.5}$$

$$\boxed{\nabla W_2 = -0.415}$$

For $\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial H_1} \cdot \frac{\partial H_1}{\partial W_1} = (-0.828) \cdot \frac{\partial(w_1(x_1 + x_2))}{\partial w_1}$

$$\Rightarrow (-0.828)(x_1 + x_2) = (-0.828)(2 - 1)$$

$$\boxed{\nabla W_1 = -0.828}$$

For $\frac{\partial L}{\partial B_1} = \frac{\partial L}{\partial H_2} \cdot \frac{\partial H_2}{\partial B_1} = (-1.245) \frac{\partial}{\partial B_1} (X_1 X_2 X_3 + B_1)$

$$\boxed{\nabla B_1 = -1.245}$$

For $\frac{\partial L}{\partial X_3} = \frac{\partial L}{\partial H_2} \cdot \frac{\partial H_2}{\partial X_3} = (-1.245) \cdot \frac{\partial (X_1 X_2 X_3 + B_1)}{\partial X_3}$

$\Rightarrow (-1.245)(X_1)(X_2) = -1.245\,(2)(-1)$

$$\boxed{\nabla X_3 = 2.49}$$

For $\frac{\partial L}{\partial X_2} = \frac{\partial L}{\partial H_1} \cdot \frac{\partial H_1}{\partial X_2} + \frac{\partial L}{\partial H_2} \cdot \frac{\partial H_2}{\partial X_2}$

$\Rightarrow (-0.828) \frac{\partial}{\partial X_2} (W_1(X_1 + X_2)) + (-1.245) \frac{\partial (X_1 X_2 X_3 + B_1)}{\partial X_2}$

$\Rightarrow (-0.828)(W_1) + (-1.245)(X_1 X_3)$

$\Rightarrow (-0.828)(1.5) - 1.245\,(2)(1)$

$\Rightarrow -1.24 \quad -2.49$

$$\boxed{\nabla X_2 = -3.73}$$

$$\text{For } \frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial H_3} \cdot \frac{\partial H_3}{\partial x_1} + \frac{\partial L}{\partial H_2} \cdot \frac{\partial H_2}{\partial x_1} + \frac{\partial L}{\partial H_1} \cdot \frac{\partial H_1}{\partial x_1}$$

$$= (0.6225) \frac{\partial}{\partial x_1}(\min(x_1, H_1)) + (-1.245) \frac{\partial}{\partial x_1}(x_1 x_2 x_3 + B_1)$$

$$+ (-0.828) \cdot \frac{\partial}{\partial x_1}(w_1 (x_1 + x_2))$$

$$\Rightarrow (0.6225)(1) - (1.245)(x_2 x_3) + (-0.828)(w_1)$$

$$\Rightarrow 0.6225 + 1.245 - 1.24$$

$$\boxed{\nabla x_1 = 0.6275}$$

Given table filled with obtained values:

$\hat{Y}$	L	$\bigtriangledown H_5$	$\bigtriangledown H_4$	$\bigtriangledown H_3$	$\bigtriangledown H_2$	$\bigtriangledown H_1$
0.3775	0.9744	0.6225	0.6225	-0.6225	-1.245	-0.828

$\bigtriangledown \mathbf{W_2}$	$\bigtriangledown \mathbf{W_1}$	$\bigtriangledown \mathbf{B_1}$	$\bigtriangledown X_3$	$\bigtriangledown X_2$	$\bigtriangledown X_1$
-0.415	-0.828	-1.245	2.49	-3.73	0.6275

(4) (a) Given,

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

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

To verify:

$$\tanh(x) = 2\sigma(2x) - 1$$

Verification:

$$R.H.S \rightarrow 2\sigma(2x) - 1$$

$$\Rightarrow 2\left(\frac{1}{1+e^{-2x}}\right) - 1$$

$$\Rightarrow \frac{2 - 1 - e^{-2x}}{1+e^{-2x}}$$

$$\Rightarrow \frac{1 - e^{-2x}}{1+e^{-2x}}$$

$$\Rightarrow \frac{1 - \frac{1}{e^{2x}}}{1 + \frac{1}{e^{2x}}}$$

$$R.H.S = \frac{e^{2x}-1}{e^{2x}+1} = \tanh(x) = L.H.S.$$

Hence, proved.

(4)
(b) For 3 (a) graph, $\hat{y} = \sigma(W_3 H + B_3)$

where, $W_3 = [1,1]$, $H = [-0.2, 2]$ and $B_3 = -1$

Replacing σ with tanh function:

$$\hat{y} = \tanh(W_3 H + B_3) = \tanh(0.8) \quad [\text{From } 3a]$$

$$\hat{y} = 2\sigma(2(0.8)) - 1 \quad [\text{From eq. } (2)]$$

$$\hat{y} = 2\sigma(1.6) - 1$$

$$\hat{y} = 2(0.832) - 1$$

$$\hat{y} = 0.664$$

Loss using MSE:

Taking $Y = -1$: $MSE(Y, \hat{y}) = \frac{1}{2}(0.664 + 1)^2$

$$MSE(Y, \hat{y}) = \frac{1}{2}(1.664)^2$$

$$\therefore \text{Loss} = 1.384$$

The range of sigmoid function lies in [0, 1] so choosing Y = 0 and 1, attributes to taking the binary equivalents of the model's output. Whereas, the range of tanh lies in [-1, 1], so choosing Y = 0 here doesn't work out. Instead we have to take Y = -1 as the negative equivalent for tanh function so as to naturally align with the expected behaviour of the model.

For 3(b) graph, $\hat{y} = \sigma(H_3 - H_5)$

where, $H_3 = 1.5$ and $H_5 = 2$

Replacing σ with tanh function :

$$\hat{y} = \tanh(1.5-2) = \tanh(-0.5)$$

$$\hat{y} = 2\sigma(2(-0.5)) - 1$$

$$\hat{y} = 2\sigma(-1) - 1$$

$$\hat{y} = 2(0.269) - 1$$

$$\hat{y} = -0.462$$

Loss using MSE :

Taking $Y = 1$:

$$MSE(Y, \hat{y}) = \frac{1}{2}(-0.462-1)^2$$

$$\therefore \text{Loss} = 1.068$$

4(c) Backpropagation of the New loss for 3(a)

For $\frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}}\left(\frac{(\hat{Y}-Y)^2}{2}\right) = \hat{Y} - Y = \tanh(W_3H+B_3) - Y$

For $\frac{\partial \hat{Y}}{\partial Z} = \frac{\partial \tanh(Z)}{\partial Z} = 1 - \tanh^2(Z) \quad [Z = W_3H + B_3]$

So, $$\frac{\partial L}{\partial Z} = \left(\tanh(W_3H + B_3) - Y\right)\left(1 - \tanh^2(W_3H + B_3)\right)$$

For $$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial W_3} = \frac{\partial L}{\partial z} \cdot \frac{\partial (W_3H + B_3)}{\partial W_3}$$

$$= \left(\tanh(W_3H + B_3) - Y\right)\left(1 - \tanh^2(W_3H + B_3)\right)(H)$$

For $$\frac{\partial L}{\partial B_3} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial B_3} = \frac{\partial L}{\partial z} \cdot \frac{\partial (W_3H + B_3)}{\partial B_3}$$

$$= \left(\tanh(W_3H + B_3) - Y\right)\left(1 - \tanh^2(W_3H + B_3)\right)$$