

NPTEL Week 6 Live Session

on Deep Learning (noc24_ee04)

A course offered by: Prof. Prabir Kumar Biswas, IIT Kharagpur

- Week 4 quiz solution (MLP, Backpropagation)
- Week 5 practice questions (Artificial Neural Nets)



By

Arka Roy

NPTEL PMRF TA

Prime Minister's Research Fellow

Department of Electrical Engineering, IIT Patna

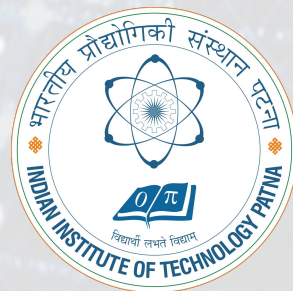
Web: <https://sites.google.com/view/arka-roy/home>

Powered by:



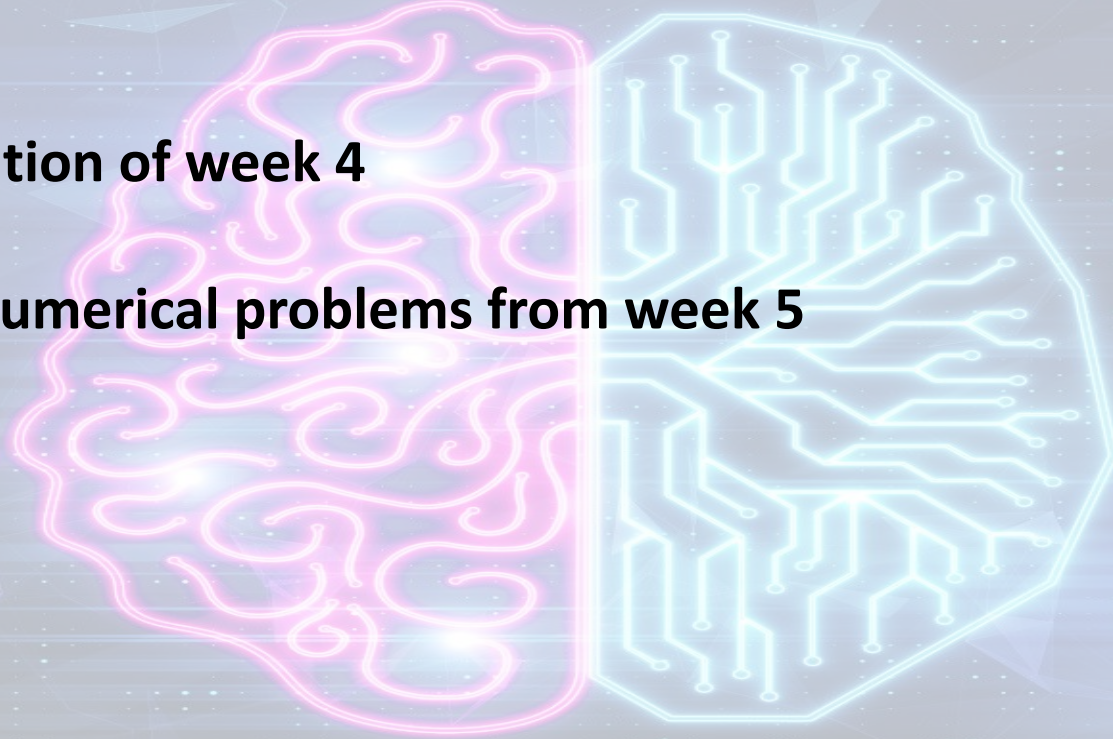
PMRF

Prime Minister's Research Fellows
Ministry of Education
Government of India

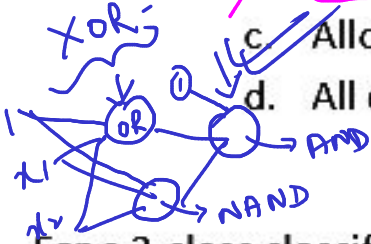
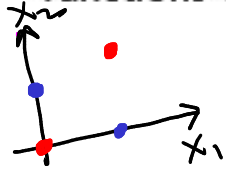


Content of the live session

1. Quiz solution of week 4
2. Solving numerical problems from week 5



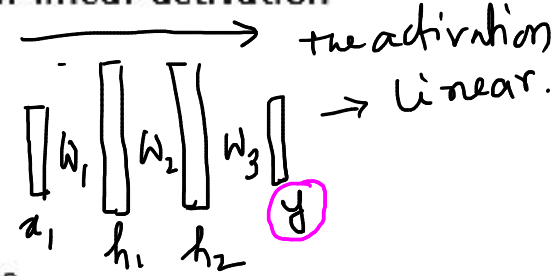
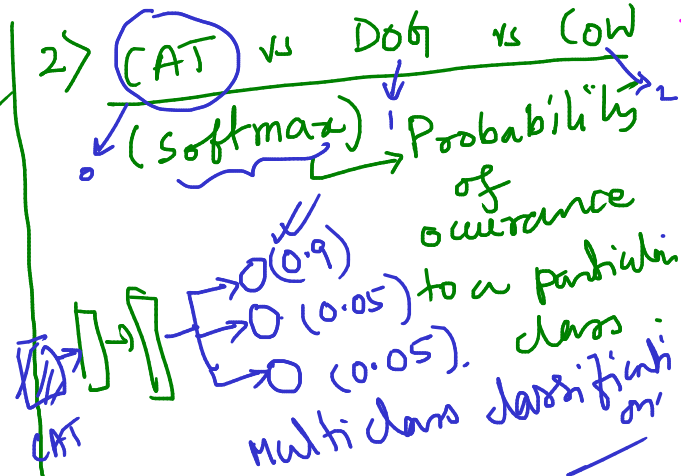
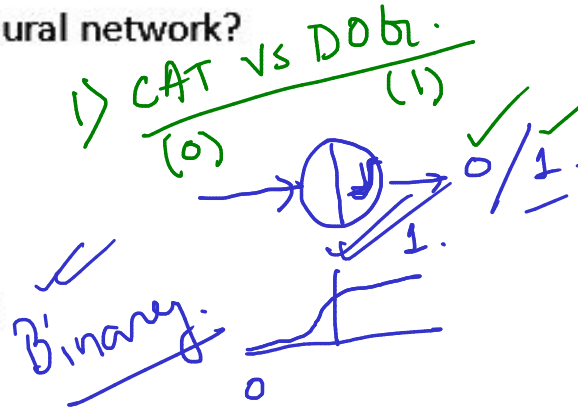
What is the main benefit of stacking multiple layers of neuron with non-linear activation functions over a single layer perceptron?



- a. ~~Reduces complexity of the network~~
- b. Reduce inference time during testing
- c. ~~Allows to create complex non-linear decision boundaries~~
- d. All of the above

For a 2-class classification problem, what is the minimum number of nodes required for the output layer of a multi-layered neural network?

- a. 2
- ~~b. 1~~
- c. 3
- d. None of the above



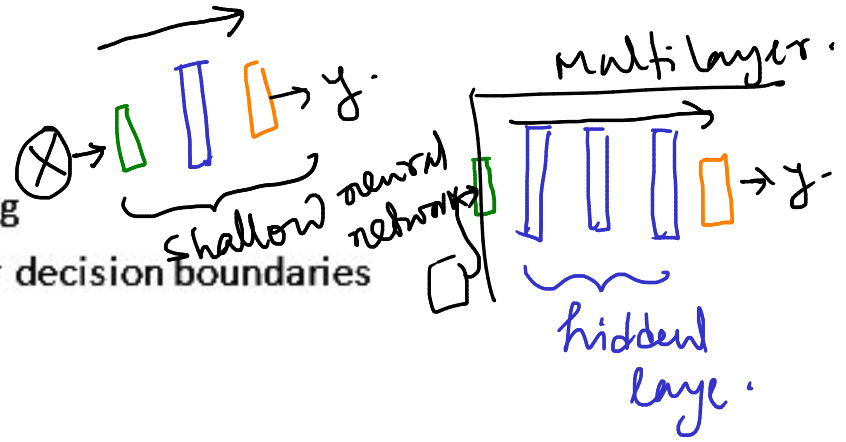
$$y = w_3 h_2$$

$$= w_3 w_2 h_1$$

$$= w_3 w_2 w_1 x_1$$

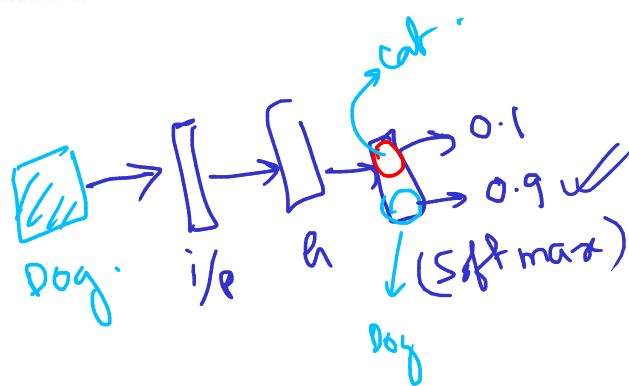
What is the main benefit of stacking multiple layers of neuron with non-linear activation functions over a single layer perceptron?

- Reduces complexity of the network
- Reduce inference time during testing
- Allows to create complex non-linear decision boundaries
- All of the above



For a 2-class classification problem, what is the minimum number of nodes required for the output layer of a multi-layered neural network?

- 2
- 1
- 3
- None of the above

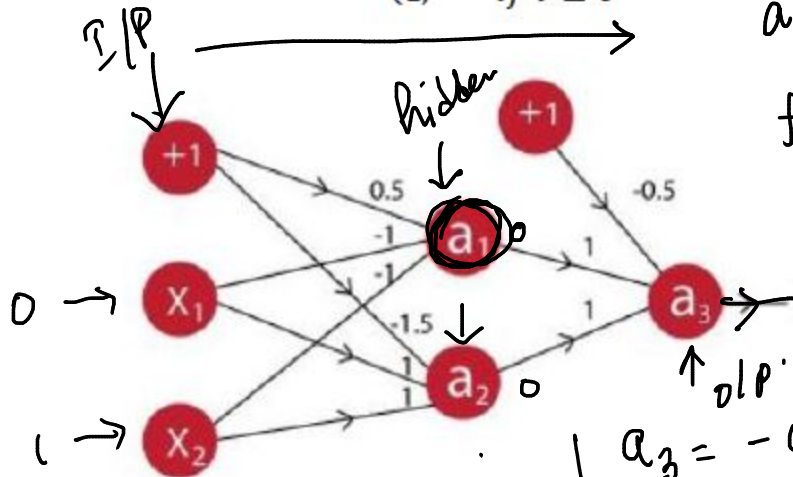


What will the output from node a_3 in the following neural network setup when the inputs are $(x_1, x_2) = (0, 1)$.

The activation function used in each of three nodes a_1 , a_2 and a_3 are zero-thresholding i.e.,

3

$$f(v) = \begin{cases} 0, & \text{if } v < 0 \\ 1, & \text{if } v \geq 0 \end{cases}$$



$$a_1 = (0.5 \times 1) + (-1 \times 0) + (1 \times -1) = -0.5$$

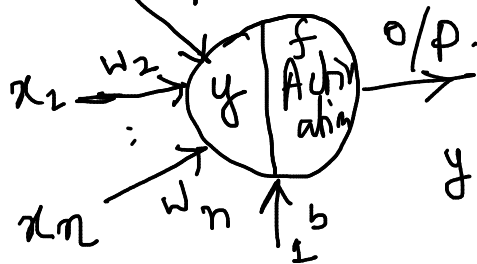
$$f(a_1) = f(-0.5) = 0.$$

$$f(a_2) = f(-1.5 + 0 + 1) = f(-0.5) = 0.$$

$$a_3 = -0.5 + 0 + 0 = -0.5.$$

$$f(a_3) = f(-0.5) = 0.$$

- a. -1
- ~~b. 0~~
- c. 1
- d. 0.5



$$y = \sum w_i x_i + b$$

$$= w_1 x_1 + w_2 x_2 + \dots + w_n x_n + b.$$

$$o/p = f(y) = f\left(\sum w_i x_i + b\right)$$

Which basic logic gate is implemented by the following neural network setup. The activation function used in each of three nodes a_1, a_2 and a_3 are zero-thresholding i.e.,

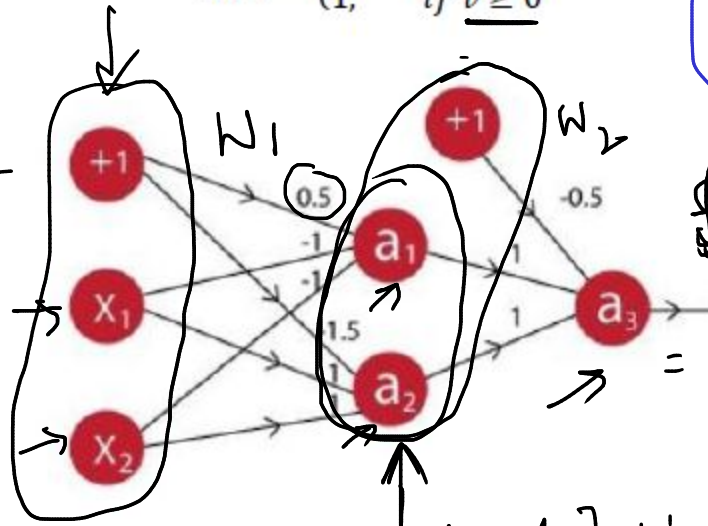
$$f(v) = \begin{cases} 0, & \text{if } v < 0 \\ 1, & \text{if } v \geq 0 \end{cases}$$

$$W_1 = \begin{bmatrix} 0.5 & -1.5 \\ -1 & 1 \\ -1 & 1 \end{bmatrix}_{3 \times 2}$$

$$W_2 = \begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix}$$

x_1	x_2	a_3
0	0	0
0	1	0
1	0	0
1	1	1

$$2^2 = 4$$



$$f(W_1^T X)$$

$$= f \left(\begin{bmatrix} 0.5 & -1 & -1 \\ -1.5 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \right)$$

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

$1 \rightarrow \text{bias}$
 $1 \rightarrow x_1$
 $1 \rightarrow x_2$

$$= f \left(\begin{bmatrix} 0.5 & -0.5 & -0.5 & -1.5 \\ -1.5 & -0.5 & -0.5 & 0.5 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

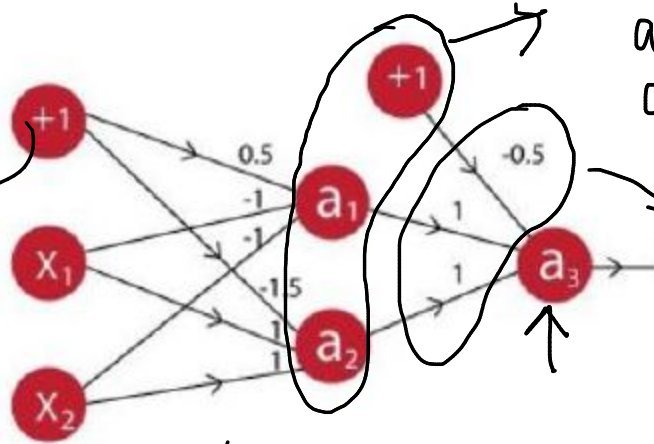
- a. AND
- b. NOR
- c. XNOR
- d. XOR

Which basic logic gate is implemented by the following neural network setup. The activation function used in each of three nodes a_1, a_2 and a_3 are zero-thresholding i.e.,

x_1	x_2	a_3
0	0	1
0	1	0
1	0	0
1	1	1

$$f(v) = \begin{cases} 0, & \text{if } v < 0 \\ 1, & \text{if } v \geq 0 \end{cases}$$

$$\text{Bias } a_1 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{3 \times 4} = H.$$



$$W = \begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix}_{3 \times 1}$$

$$f(W^t H) = f \left(\begin{bmatrix} -0.5 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

$$= f \left(\begin{bmatrix} 0.5 & -0.5 & -0.5 & 0.5 \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}$$

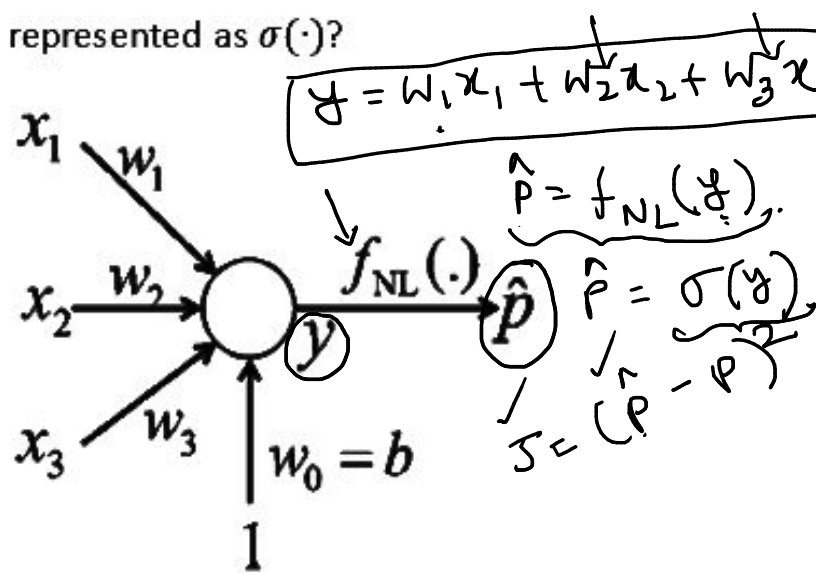
a. AND

b. NOR

~~c. XNOR~~

d. XOR

Find the gradient component $\frac{\partial J}{\partial w_1}$ for the network shown below if $J(\cdot) = \underbrace{(\hat{p} - p)^2}_{L_{loss}}$ is the loss function, p is the target and the non-linearity $f_{NL}(\cdot)$ is the sigmoid activation function represented as $\sigma(\cdot)$?



- $2\hat{p} \times (1 - \sigma(y)) \times x_1$
- ☒ $2(\hat{p} - p) \times \sigma(y) \times (1 - \sigma(y)) \times x_1$
- $2(\hat{p} - p) \times (1 - \sigma(y)) \times x_1$
- $2(1 - p) \times (1 - \sigma(y)) \times x_1$

$$J(\cdot) = (\hat{p} - p)^2$$

$$\frac{\partial J}{\partial w_1} = \underbrace{\frac{\partial J}{\partial \hat{p}}}_{2(\hat{p} - p)} \underbrace{\frac{\partial \hat{p}}{\partial y}}_{\sigma(y) \cdot (1 - \sigma(y))} \cdot \frac{\partial y}{\partial w_1}$$

$$\frac{\partial J}{\partial w_1} = (2 \cdot (\hat{p} - p) \cdot \sigma(y) \cdot (1 - \sigma(y))) x_1$$

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

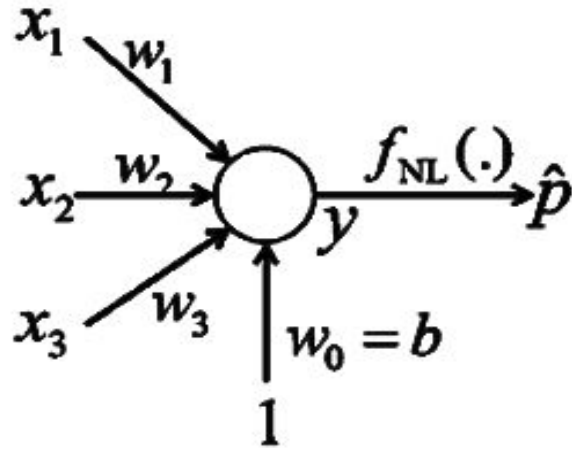
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$= \left(\frac{1}{1 + e^{-x}} \right) \cdot \frac{e^{-x}}{1 + e^{-x}}$$

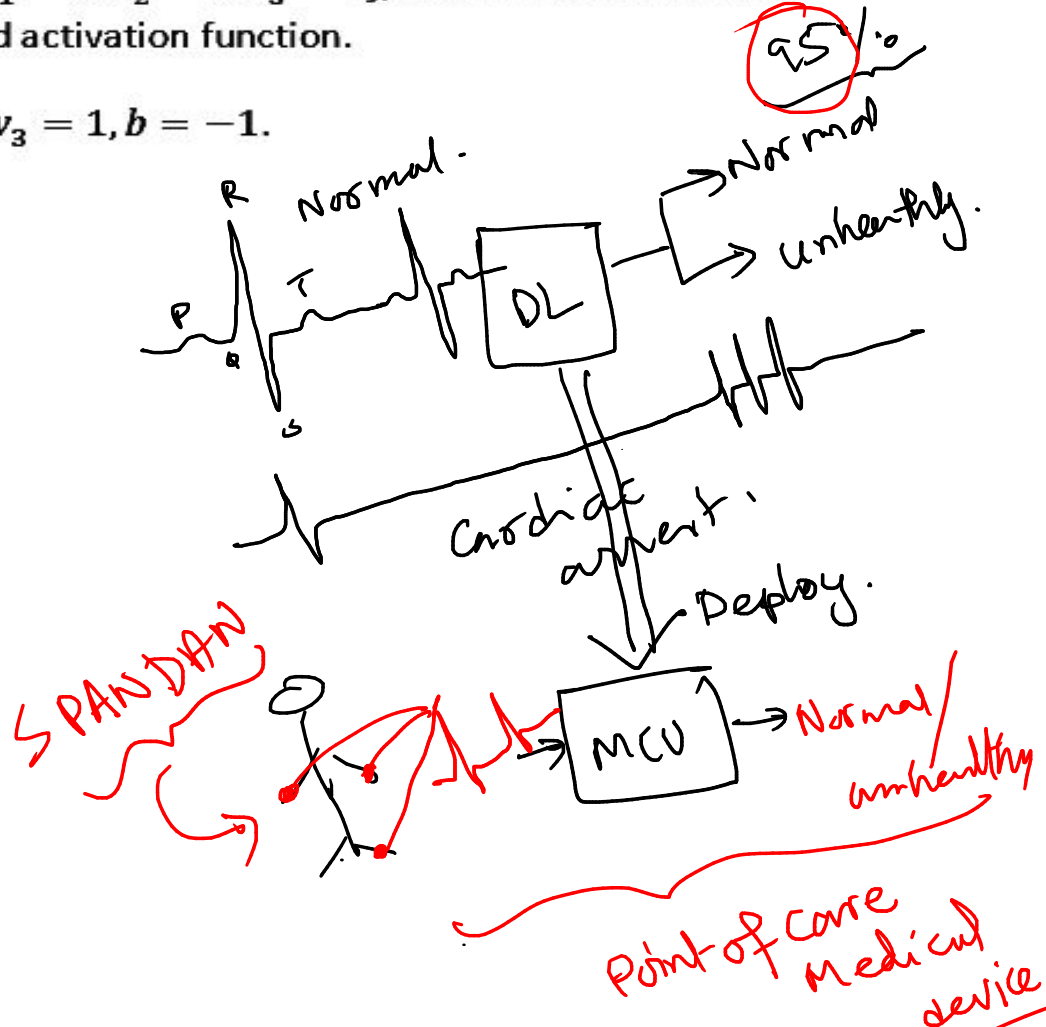
$$= \sigma(x) \cdot (1 - \sigma(x))$$

Find the output \hat{p} corresponding to input $\{x_1 = 1, x_2 = 1, x_3 = 0\}$, for the network shown below. The non-linearity $f_{NL}(\cdot)$ is the sigmoid activation function.

The weights are given as $w_1 = 2, w_2 = -1, w_3 = 1, b = -1$.

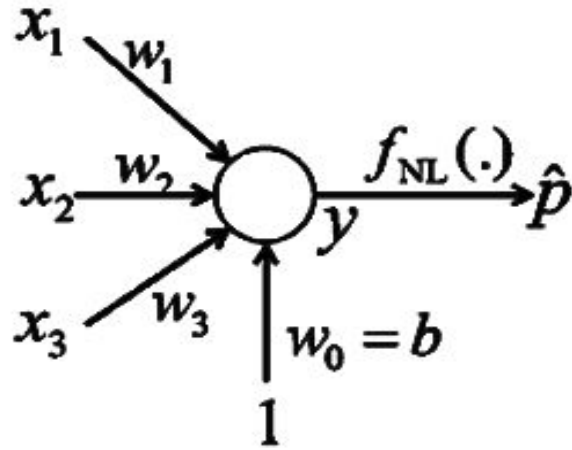


- a. 0
- b. 1
- c. 0.5
- d. 0.25



Find the output \hat{p} corresponding to input $\{x_1 = 1, x_2 = 1, x_3 = 0\}$, for the network shown below. The non-linearity $f_{NL}(\cdot)$ is the sigmoid activation function.

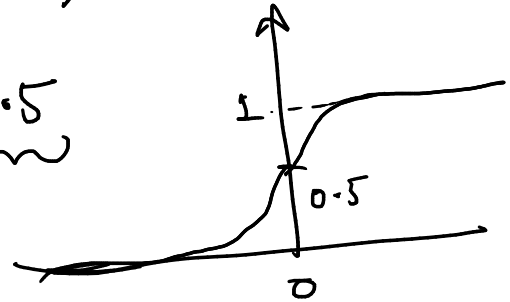
The weights are given as $w_1 = 2, w_2 = -1, w_3 = 1, b = -1$.



$$\hat{p} = \sigma(w_1 x_1 + w_2 x_2 + w_3 x_3 + b)$$

$$= \sigma(2 - 1 + 0 - 1)$$

$$= \sigma(0) = \underline{0.5}$$



a. 0

b. 1

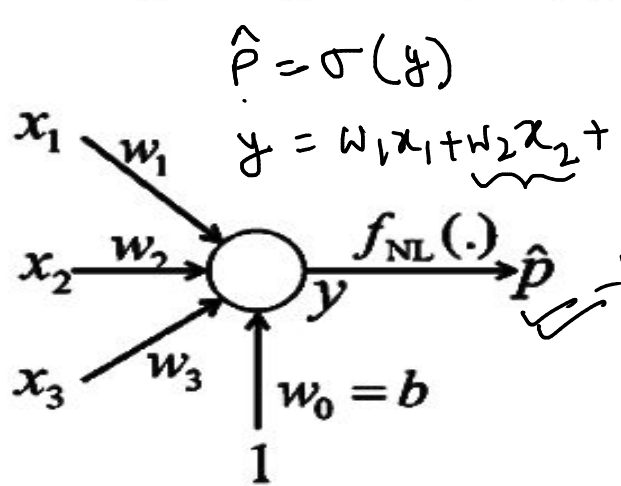
☒ c. 0.5

d. 0.25

Find the gradient component $\frac{\partial J}{\partial w_2}$ for the network shown below if $J(\cdot) = (\hat{p} - p)^2$ is the loss function, $p=1$ is the target and the non-linearity $f_{NL}(\cdot)$ is the sigmoid activation function represented as $\sigma(\cdot)$?

The input to the network is $\{x_1 = 1, x_2 = 1, x_3 = 0\}$

The weights are given as $w_1 = 2, w_2 = -1, w_3 = 1, b = -1$.



a. -0.5

b. -1

c. 0

☒ d. -0.25

$$\hat{p} = 0.5$$

$$\hat{p} = \sigma(y) \quad J = (\hat{p} - p)^2$$

$$y = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial \hat{p}} \cdot \frac{\partial \hat{p}}{\partial y} \cdot \frac{\partial y}{\partial w_2}$$

$$= 2(\hat{p} - p) \underbrace{\sigma(y)}_{\hat{p}} \cdot (1 - \sigma(y)) \cdot x_2$$

$$\boxed{\frac{\partial J}{\partial w_2} = 2(\hat{p} - p) \hat{p} (1 - \hat{p}) \cdot x_2}$$

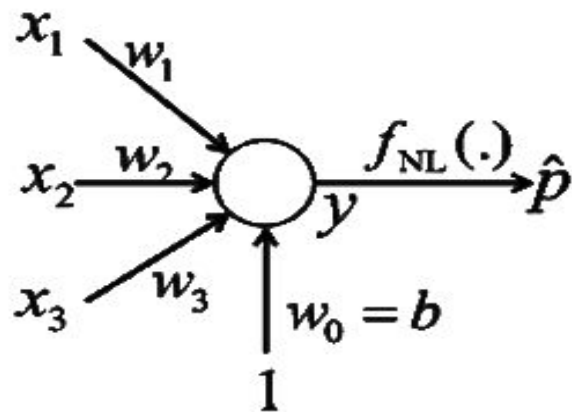
$$= 2\left(\frac{1}{2} - 1\right) \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) \cdot 1$$

$$= 2 \times -\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = -\frac{1}{4} = \underline{\underline{-0.25}}$$

What will be the updated value of w_2 after the first iteration from the current state of the network shown below if $J(\cdot) = (\hat{p} - p)^2$ is the loss function, $p=1$ is the target and the non-linearity $f_{NL}(\cdot)$ is the sigmoid activation function represented as $\sigma(\cdot)$?

The input to the network is $\{x_1 = 1, x_2 = 1, x_3 = 0\}$, the learning rate $\eta = 2$

The weights of the current state are given as $w_1 = 2, w_2 = -1, w_3 = 1, b = -1$.



- ☒ a. -0.5
☐ b. -1
☐ c. 0
☐ d. -0.25

$$w_2^{n+1} \leftarrow w_2^n - \eta \cdot \frac{\partial J}{\partial w_2}$$

$\eta = 2$

$$w_2^{(1)} = w_2^{(0)} - \eta \frac{\partial J}{\partial w_2}$$

$$w_2^{(1)} = (-1) - \left[2 \times -\frac{1}{2} \right]$$

$$= -1 + \frac{1}{2} = -\frac{1}{2} = \underline{\underline{-0.5}}$$

Suppose a neural network has 3 input nodes, x, y, z . There are 2 neurons, Q and F . $Q = x + y$ and $F = Q * z$. What is the gradient of F with respect to x, y and z ? Assume, $(x, y, z) = (-2, 5, -4)$.

- a. $(-4, 3, -3)$
- ☒ b. $(-4, -4, 3)$
- c. $(4, 4, -3)$
- d. $(3, 3, 4)$

$$Q = x + y$$

$$F = Q \cdot z$$

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial Q} \frac{\partial Q}{\partial x} = z \cdot 1 = -4$$

$$\frac{\partial F}{\partial y} = \frac{\partial F}{\partial Q} \frac{\partial Q}{\partial y} = z \cdot 1 = -4$$

$$\frac{\partial F}{\partial z} = Q = 3$$

Ans $(-4, -4, 3)$



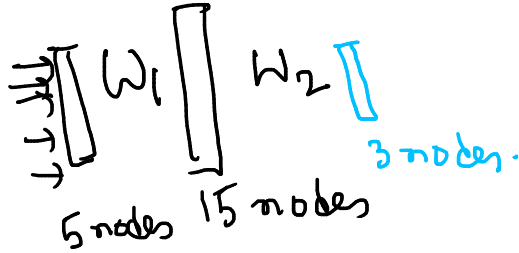
Suppose a fully-connected neural network has a single hidden layer with 15 nodes. The input is represented by a 5D feature vector and the number of classes is 3. Calculate the number of parameters of the network. Consider there are NO bias nodes in the network?

a. 225

b. 75

c. 78

~~d. 120~~



Size of $W_1 = (5 \times 15)$

Size of $W_2 = (15 \times 3)$.

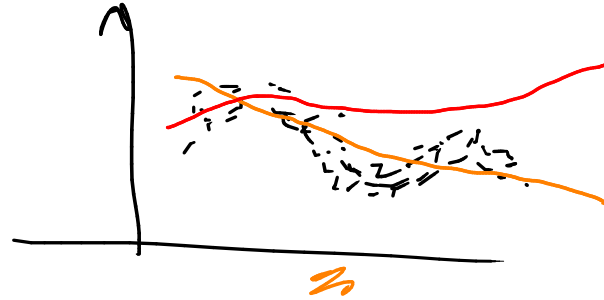
Total elements $W_1 = 75$

" " $W_2 = 45$

Total = 120

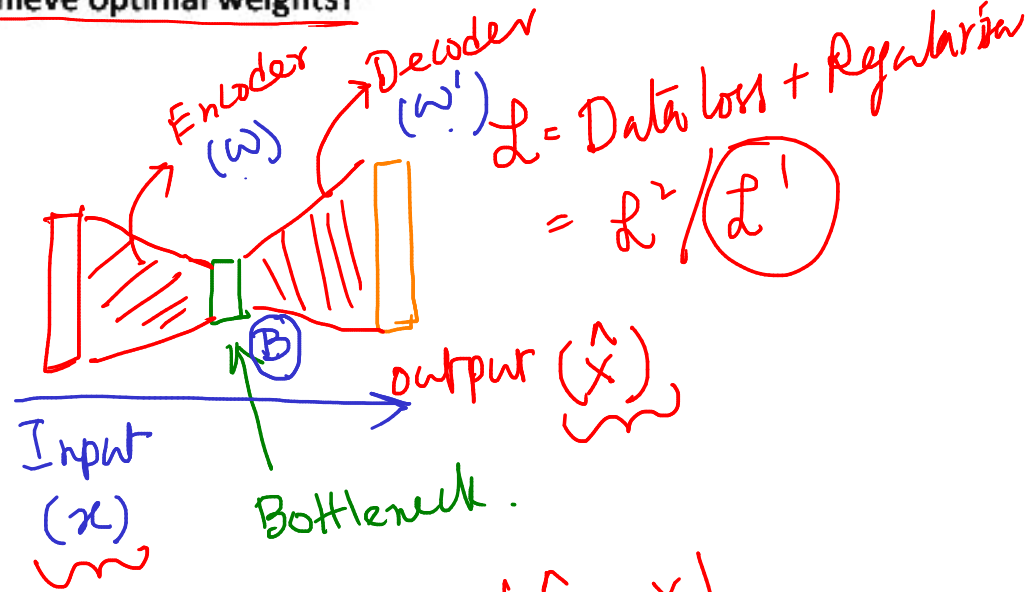
Which of the following is FALSE about PCA and Autoencoders?

- a. ~~PCA works well with non-linear data but Autoencoders are best suited for linear data~~ **FALSE**
- b. ~~Output of both PCA and Autoencoders is lossy~~
- c. ~~Both PCA and Autoencoders can be used for dimensionality reduction~~
- d. None of the above



Given input x and linear autoencoder (no bias) with random weights (W for encoder and W' for decoder), what mathematical form is minimized to achieve optimal weights?

- a. $|x - (W' \cdot W \cdot x)|$
- b. $|x - (W \cdot W' \cdot x)|$
- c. $|x - (W \cdot W \cdot x)|$
- d. $|x - (W' \cdot W' \cdot x)|$



$$\hat{x} = W' B$$

$$\hat{x} = W' W x$$

$$B = W x$$

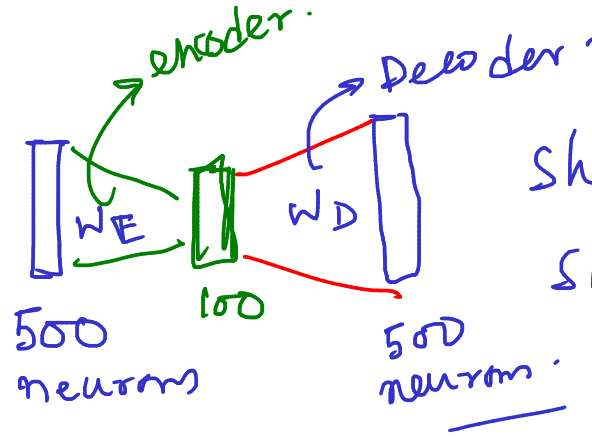
$$\hat{x} = W' B = W' W x$$

$$\mathcal{L}_1 = |\hat{x} - x|$$

$$= |x - \hat{x}|$$

$$\mathcal{L} = |x - (W' W x)|$$

A single hidden and no-bias autoencoder has 500 input neurons and 100 hidden neurons. What will be the number of parameters associated with this autoencoder?



$$\text{Shape } W_E = 500 \times 100$$

$$\text{Shape } W_D = 100 \times 500$$

$$\text{Param } W_E = 50000$$

$$W_D = 50000$$

$$\text{Total } W_E + W_D = 100000$$

When $\tanh(x) = T$ and $\text{sigmoid}(x) = S$ which of the following satisfies their relationship?

a. $T = \frac{2S+1}{2S^2-2S+1}$

b. $T = \frac{2S+1}{2S^2+2S+1}$

c. $T = \frac{2S-1}{2S^2-2S+1}$

d. $T = \frac{2S-1}{2S^2+1}$

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

$$\Rightarrow S(1 + e^{-x}) = 1$$

$$\Rightarrow 1 + Se^{-x} = 1$$

$$\Rightarrow e^{-x} = \left(\frac{1-S}{S} \right)$$

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

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

$$T = \frac{1 - \left(\frac{1-S}{S} \right)^2}{1 + \left(\frac{1-S}{S} \right)^2}$$

Which of the following two vectors can form the first two principal components?

a. $\{2; 3; 1\}$ and $\{3; 1; -9\}$

b. $\{2; 4; 1\}$ and $\{-2; 1; -8\}$

c. $\{2; 3; 1\}$ and $\{-3; 1; -9\}$


d. $\{2; 3; -1\}$ and $\{3; 1; -9\}$

eigen vectors - Covariance matrix.
symmetric.
orthogonal
 \vec{A}, \vec{B}
 $(N \times N)$

$$(a) \vec{A} = [2 \ 3 \ 1] \quad \vec{B} = [3 \ 1 \ -9]$$

$$(\vec{A} \cdot \vec{B}) = 6 + 3 - 9 = 0 \rightarrow \text{orthogonal}$$

$$(b) -4 + -8 = -12 \rightarrow \text{Non orthogonal}$$



$$\hat{i} \cdot \hat{j} = 0$$

$$= |\hat{i}| |\hat{j}| \cos 90^\circ$$

$$\hat{i} \cdot \hat{j} = 0 \rightarrow \text{orthogonal}$$