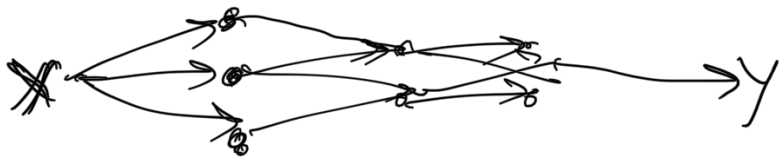
$$y = f(x)$$

Neural Networks.



$W^T X$

$w_1 x_1$

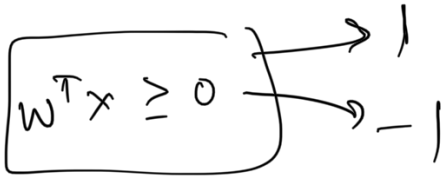
$+ w_2 x_2$

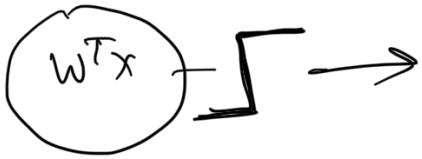
$+ w_3 x_3$

$+ w_4 x_4$

$+ w_5 x_5$

Thresholded perceptron



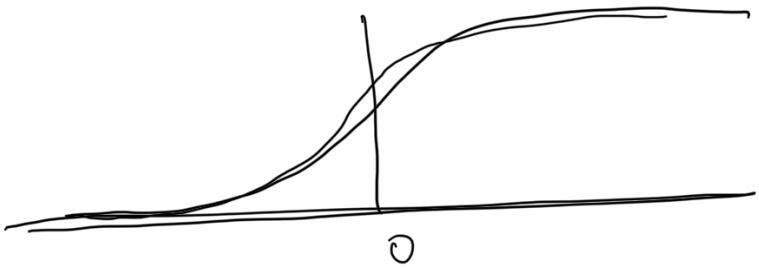


Unit

Layer

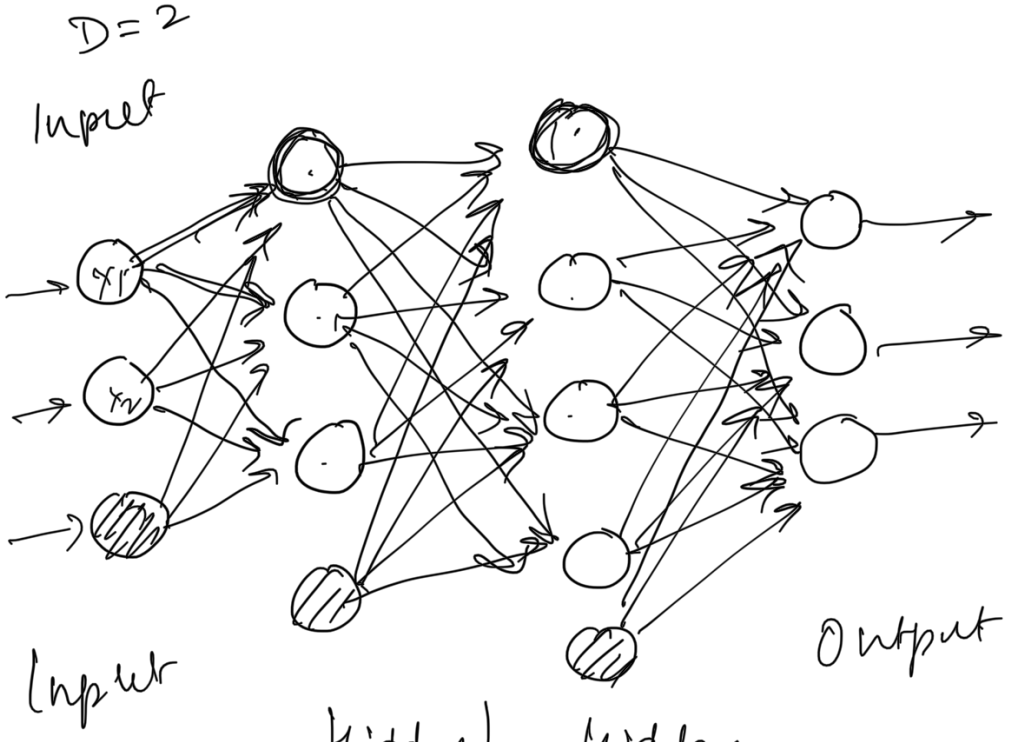
Sigmoid unit

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$



March 12 Friday

- PA2 is out
- Mid-term next Friday

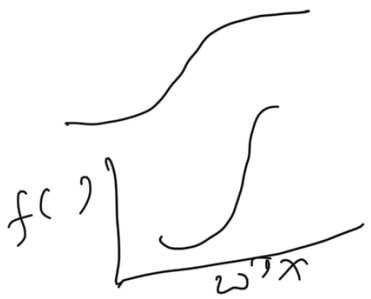


hidden layer hidden layer

# of units in a layer ≡ width of a layer

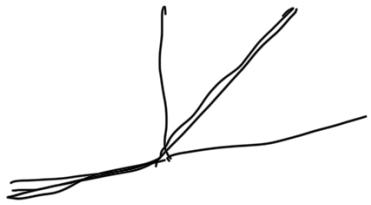
Sigmoid

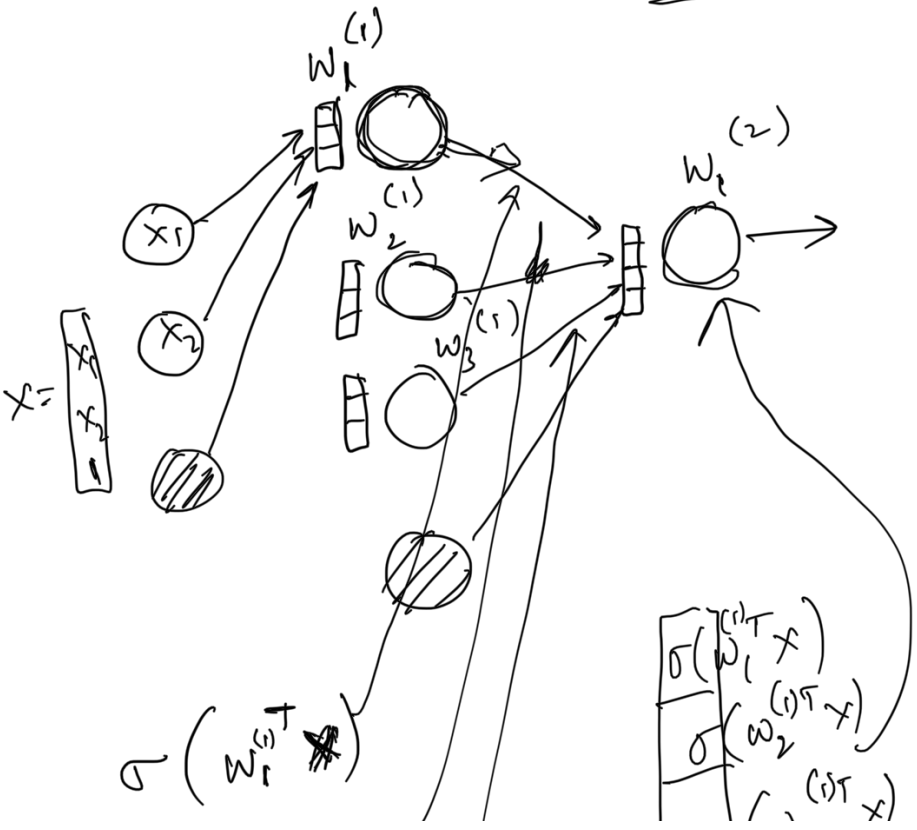
tanh

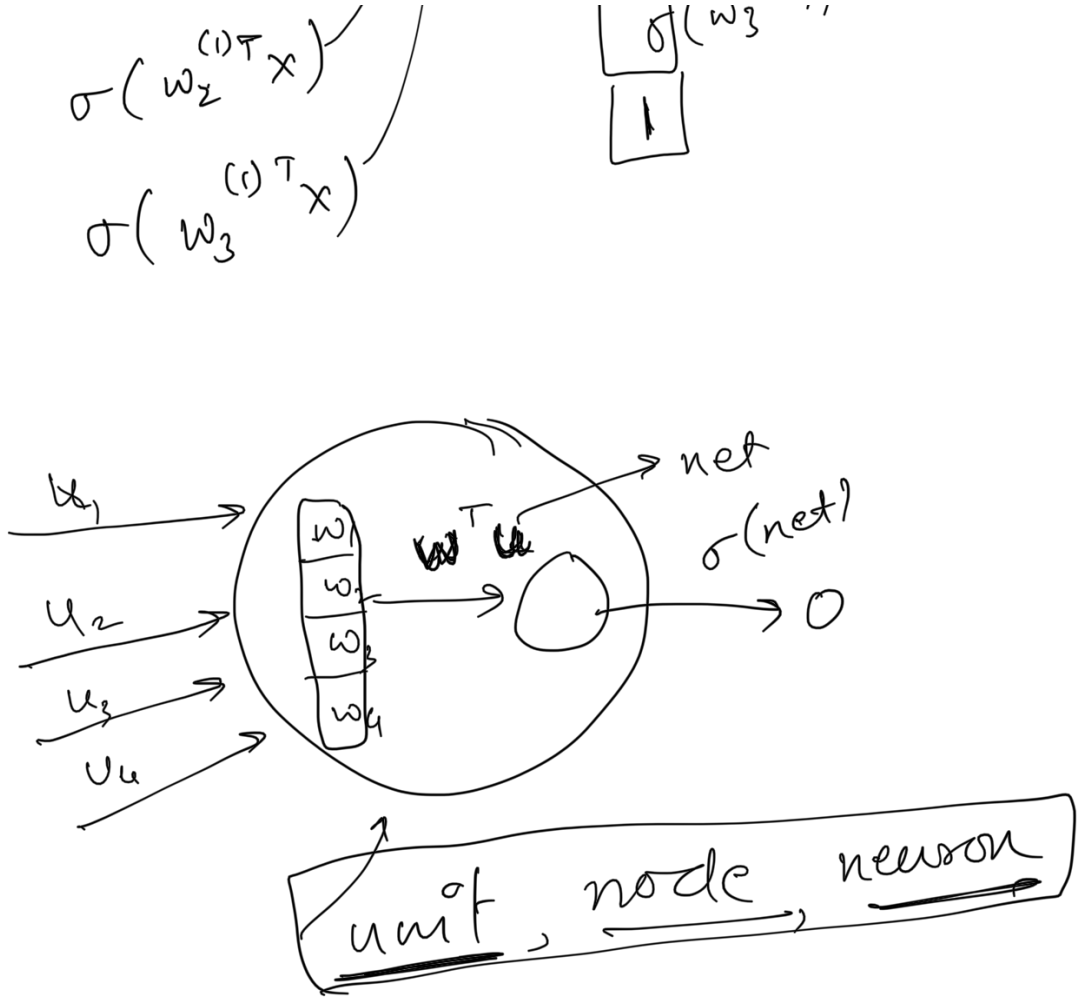


Relu

$\max(0, w^T x)$







Sigmoid

$$sigmoid(z) = \frac{1}{1+e^{-z}}$$

Tanh

$$tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$

–

Simple example

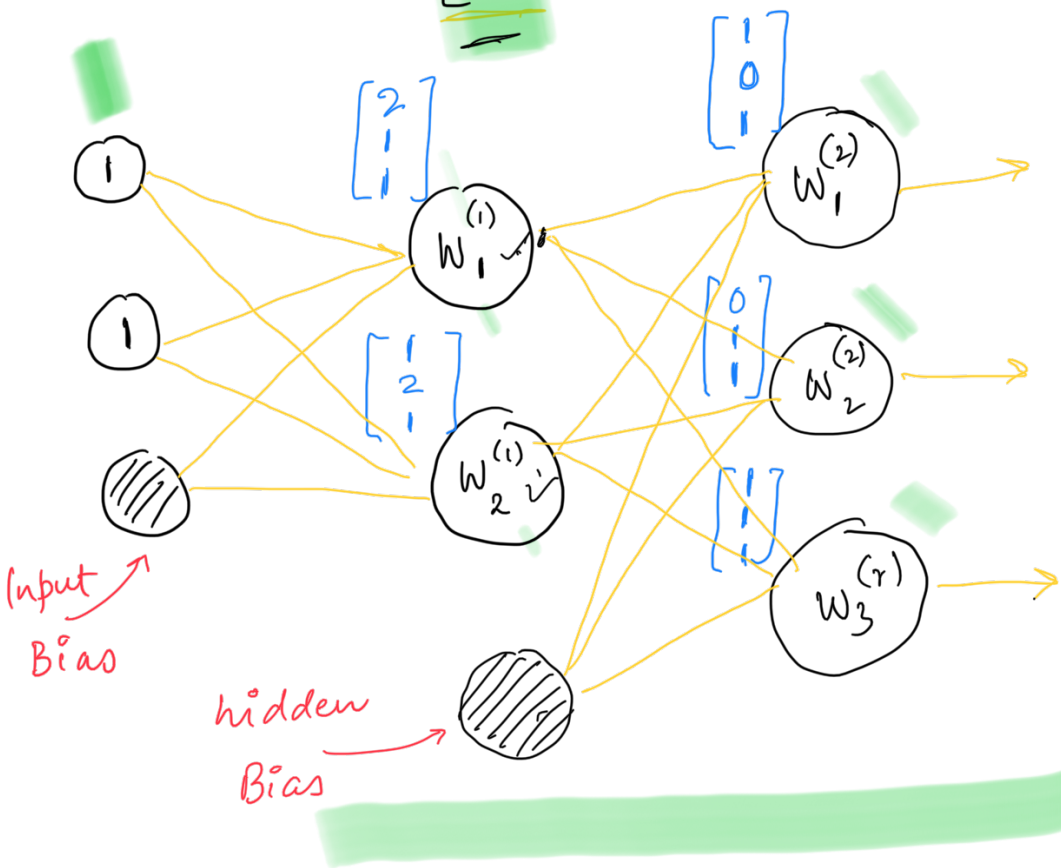
$D = 2$ (2-d data)

$M = 2$

$k = 3$ (3- outputs / classes)

Assume sigmoid activation

Data: $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$



At hidden unit 1

$$net_1^{(1)} = w_1^{(1)T} \mathbf{x}$$

$$= [2 \;\; 1 \;\; 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 4$$

$$z_1 = \sigma(net_1^{(1)}) = \sigma(4)$$

$$= 0.98$$

At hidden unit 2

$$net_2^{(1)} = w_2^{(1)} x$$

$$= [1 \;\; 2 \;\; 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 4$$

$$z_2 = \sigma(net_2^{(1)}) = \sigma(4)$$

$$= 0.98$$

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ 1 \end{bmatrix}$$

At output unit 1

$$net_1^{(2)} = w_1^{(2)T} \mathbf{z}$$

$$= [1 \;\; 0 \;\; 1] \begin{bmatrix} 0.98 \\ 0.98 \\ 1 \end{bmatrix} = 1.98$$

$$o_1 = \sigma(1.98) = \frac{1}{1+\exp(-1.98)}$$

$$= 0.88$$

At output unit 2

$$net^{(2)} = w^{(2)T} z$$

$$= [0 \quad 1 \quad 1] \begin{bmatrix} 0.98 \\ 0.98 \\ 1 \end{bmatrix} = 1.98$$

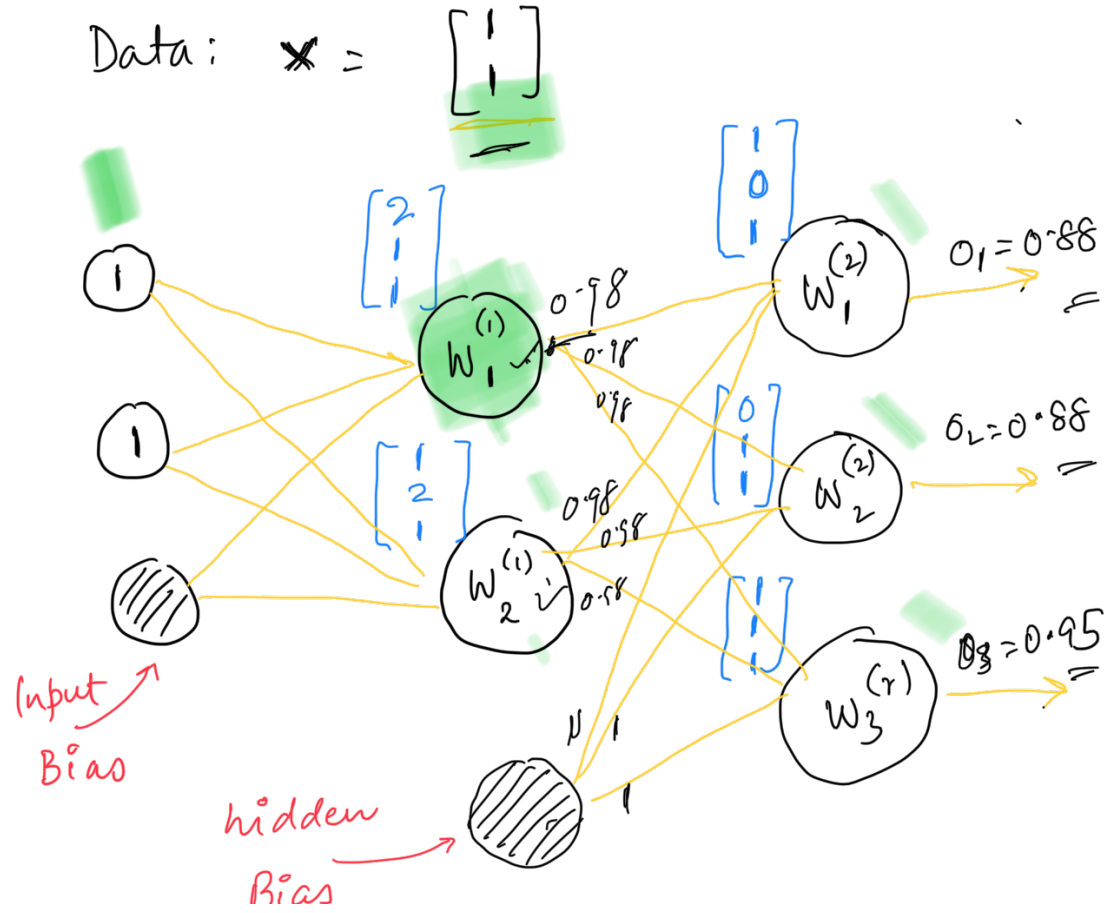
$$O_2 = \sigma(1.98) = 0.88$$

At output unit 3

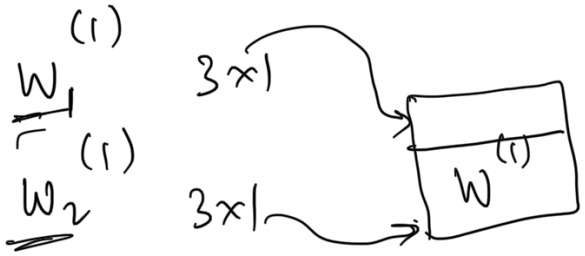
$$net_3^{(2)} = W_3^{(2)T} z = [1 \quad 1 \quad 1] \begin{bmatrix} 0.98 \\ 0.98 \\ 1 \end{bmatrix}$$

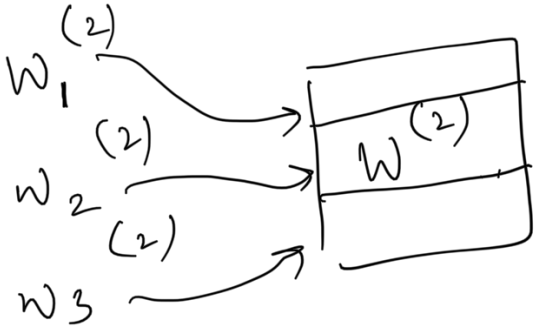
$$= 2.96$$

$$O_3 = \sigma(2.96) = 0.95$$



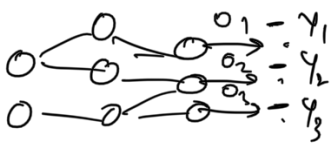
softmax output : $\left[\frac{O_1}{O_1+O_2+O_3}, \frac{O_2}{O_1+O_2+O_3}, \frac{O_3}{O_1+O_3+O_5} \right]$





$$\sigma\left(W^{(1)} X\right) = \begin{bmatrix} Z_{11} \\ Z_2 \end{bmatrix} \qquad \bar{Z} = \begin{bmatrix} 1 \\ Z_2 \\ 1 \end{bmatrix}$$

$$\sigma\left(W^{(2)} \bar{Z}\right) = \begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix}$$



$$J\left(W_1^{(1)}, \ldots, W^{(2)}, \ldots\right) = \sum_{i}^{N} J_i$$

$$J_i = \frac{1}{2}\sum_{l=1}^{K} (y_{il} - o_{il})^2 \qquad i = 1$$

o_{il} → output for the i^{th} training example at output unit l

y_{il} → true output for the i^{th} example at l^{th} output unit.

$$J = \frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{K} (y_{il} - o_{il})^2$$

E.g. 3 class classifier.

x D = 2	y
3.7, 4.8	2
3.4, 1.2	1

One-of-K encoding

Dummy encoding

let K = 3

2 → [0 | 1 | 0]

1 → [1 | 0 | 0]

3 → [0 | 0 | 1]

Gradient Descent

Gradient Descent

$J \rightarrow$ is a function of all the weights.

∂J