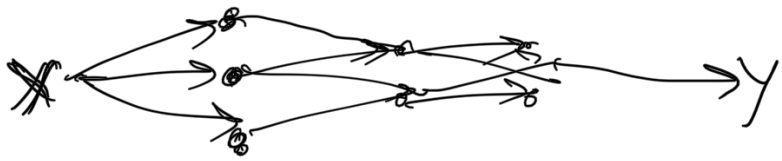
$$y = f(x)$$

Neural Networks.



$W^T X$

$w_1 x_1$

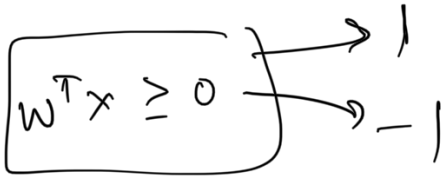
$+ w_2 x_2$

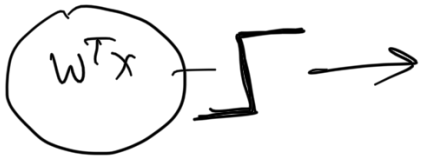
$+ w_3 x_3$

$+ w_4 x_4$

$+ w_5 x_5$

Thresholded perceptron



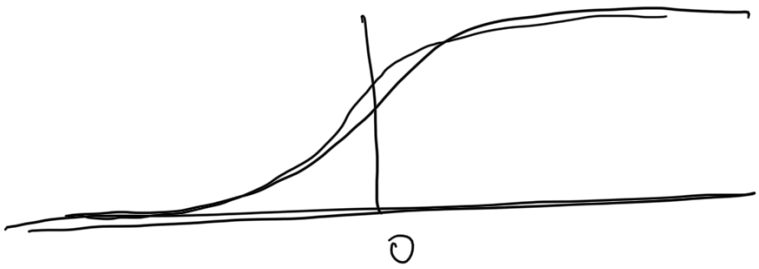


Unit

Layer

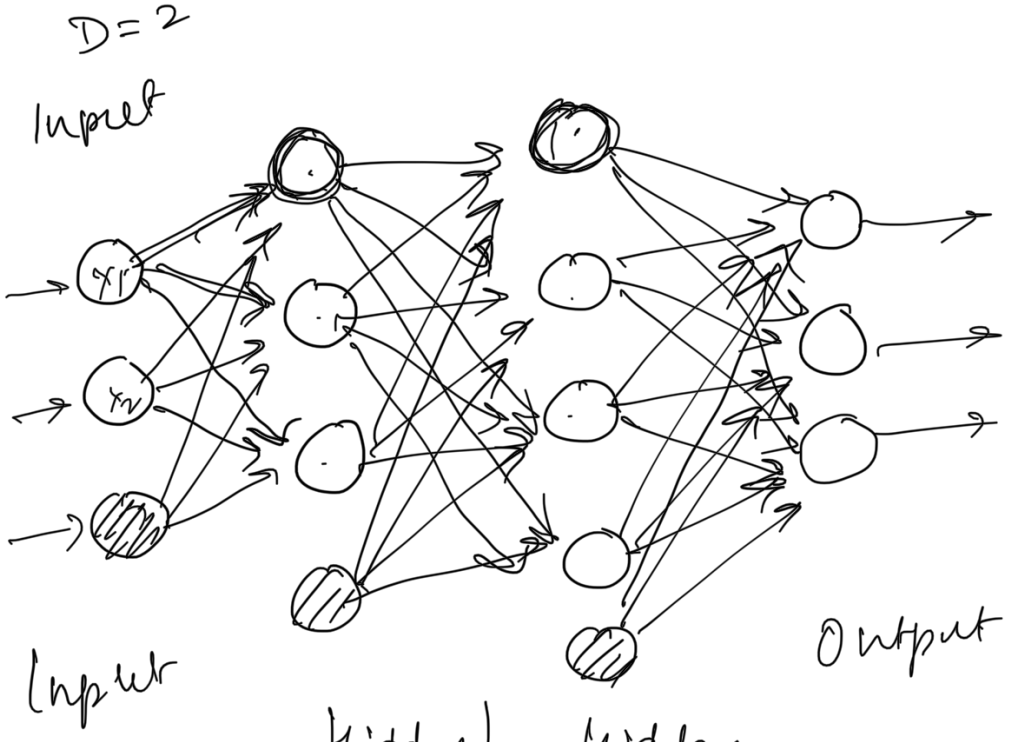
Sigmoid unit

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



March 12 Friday

- PA2 is out
- Mid-term next Friday

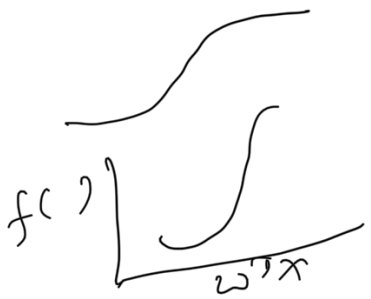


hidden layer hidden layer

# of units in a layer ≡ width of a layer

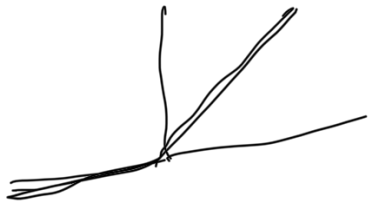
Sigmoid

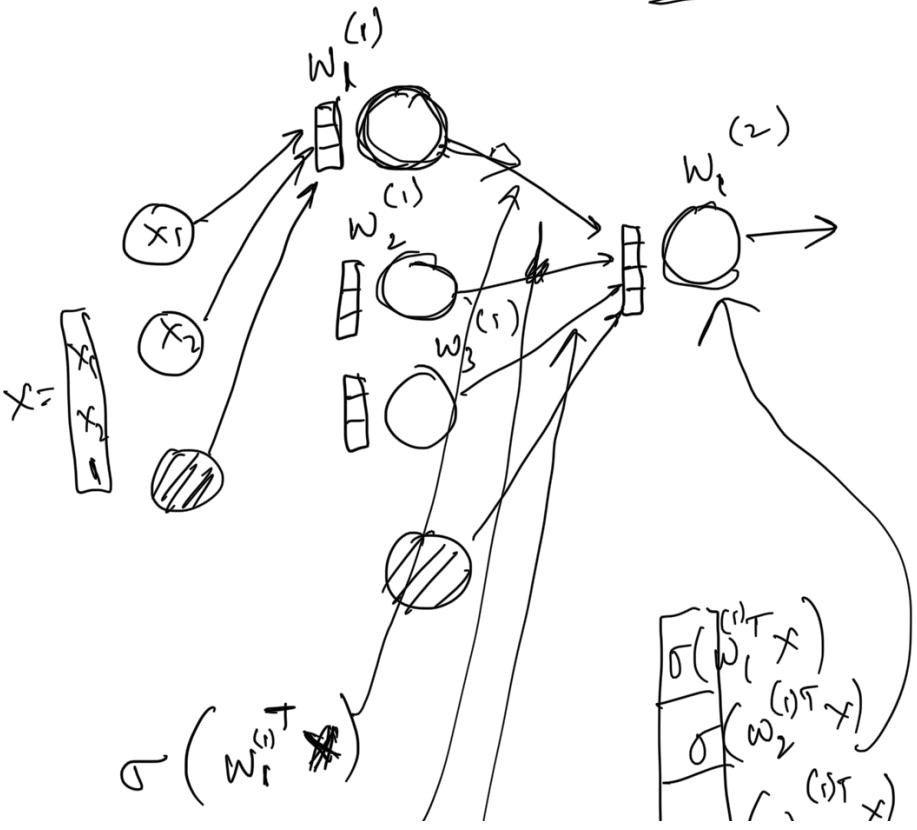
tanh

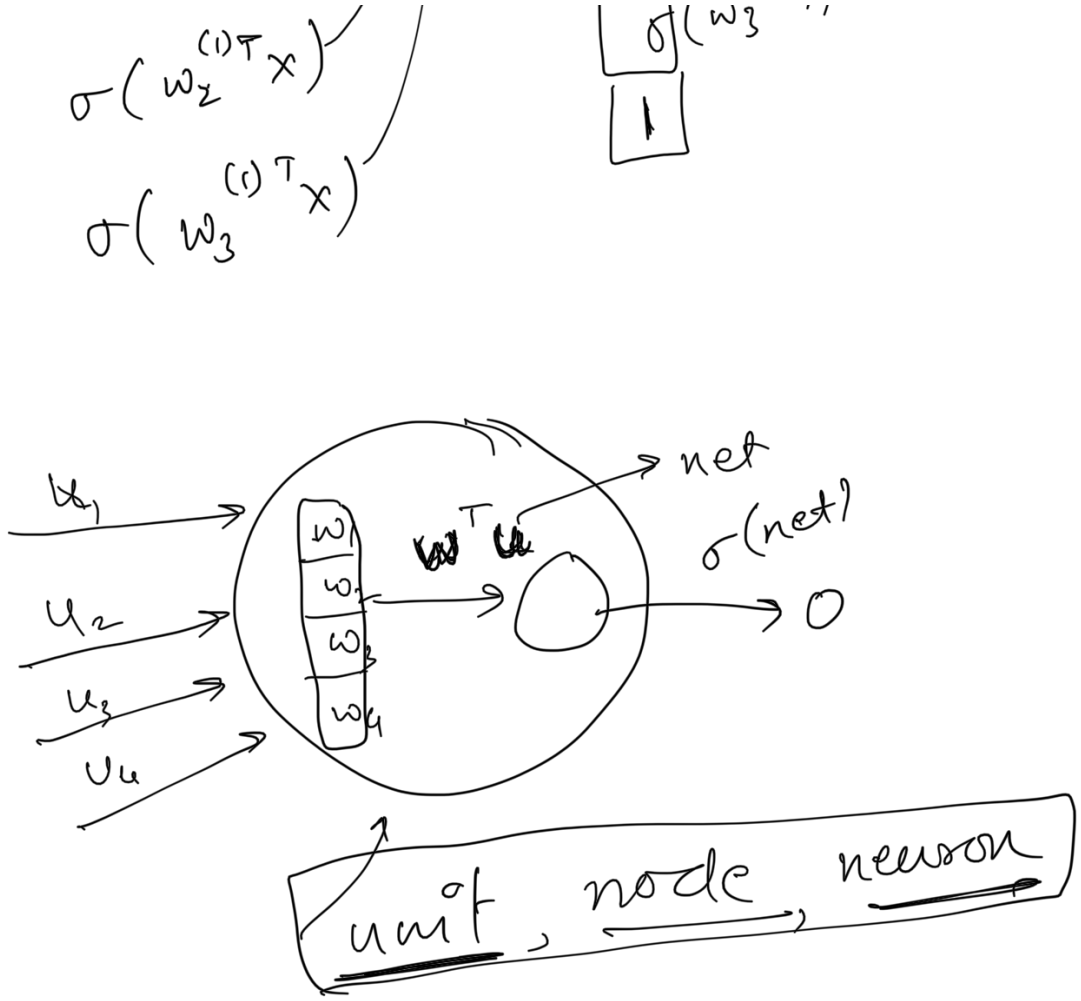


Relu

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







Sigmoid

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

Tanh

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

–

Simple example

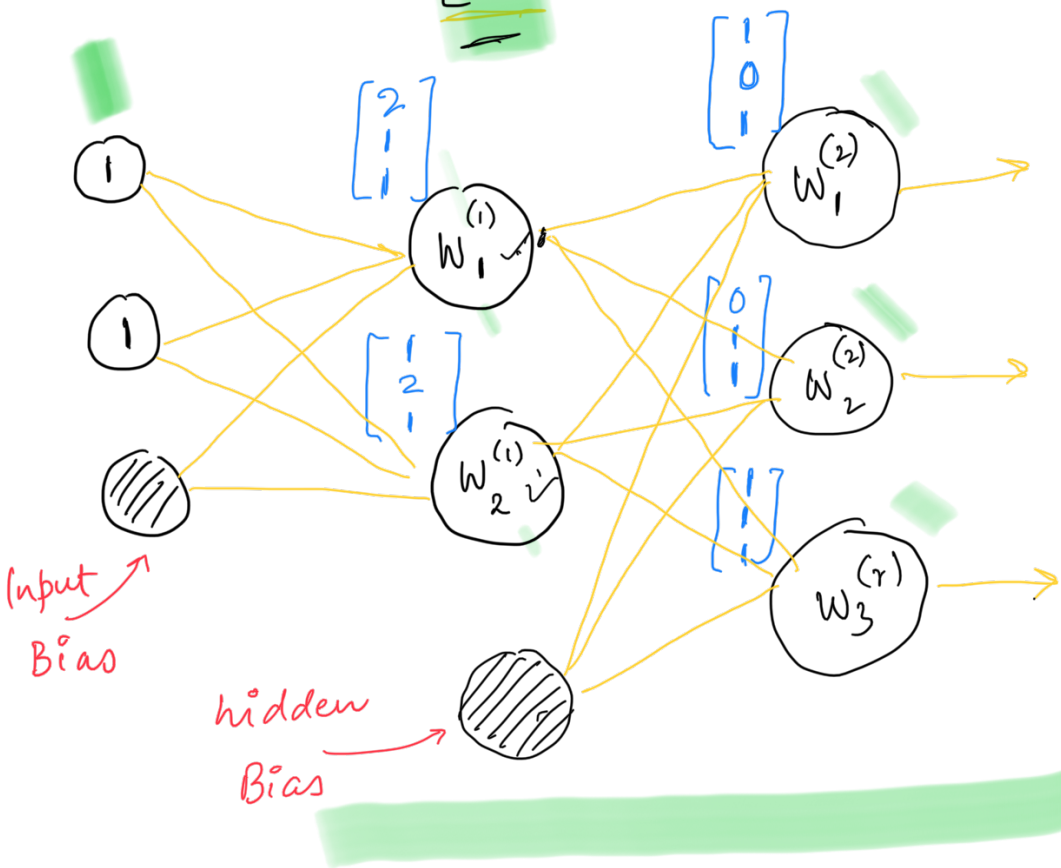
$D = 2$ (2-d data)

$M = 2$

$k = 3$ (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

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

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

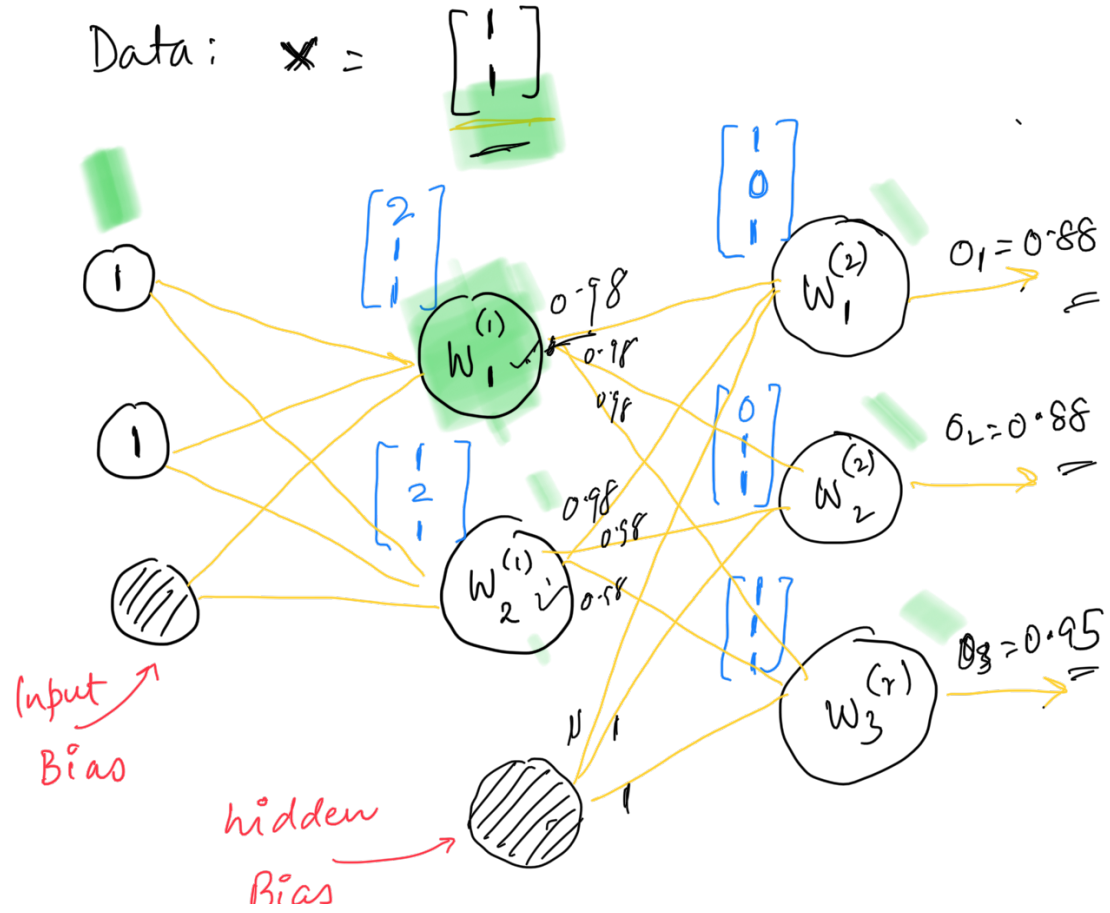
$$O_2 = \sigma(1.98) = 0.88$$

At output unit 3

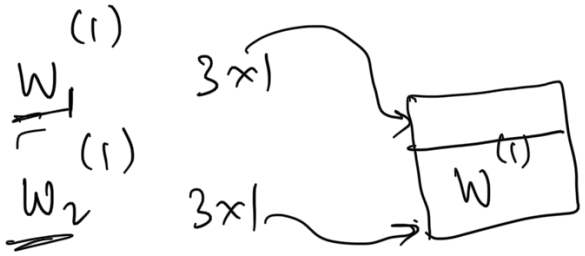
$$net_3^{(2)} = W_3^{(2)T} z = [1 \quad 1 \quad 1] \begin{bmatrix} 0.98 \\ 0.98 \\ 1 \end{bmatrix}$$

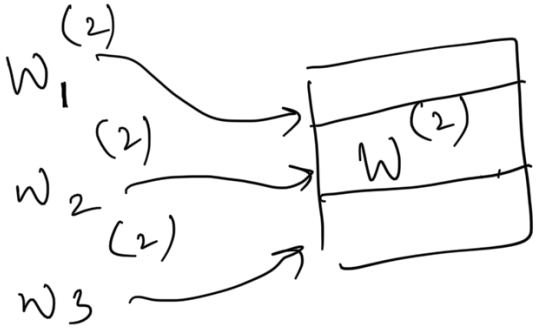
$$= 2.96$$

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



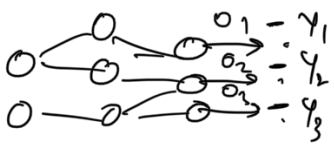
softmax output : $\left[\frac{O_1}{O_1+O_2+O_3}, \frac{O_2}{O_1+O_2+O_3}, \frac{O_3}{O_1+O_3+O_5} \right]$





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

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



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

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

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

Wednesday March 17

Notation

Subscripts

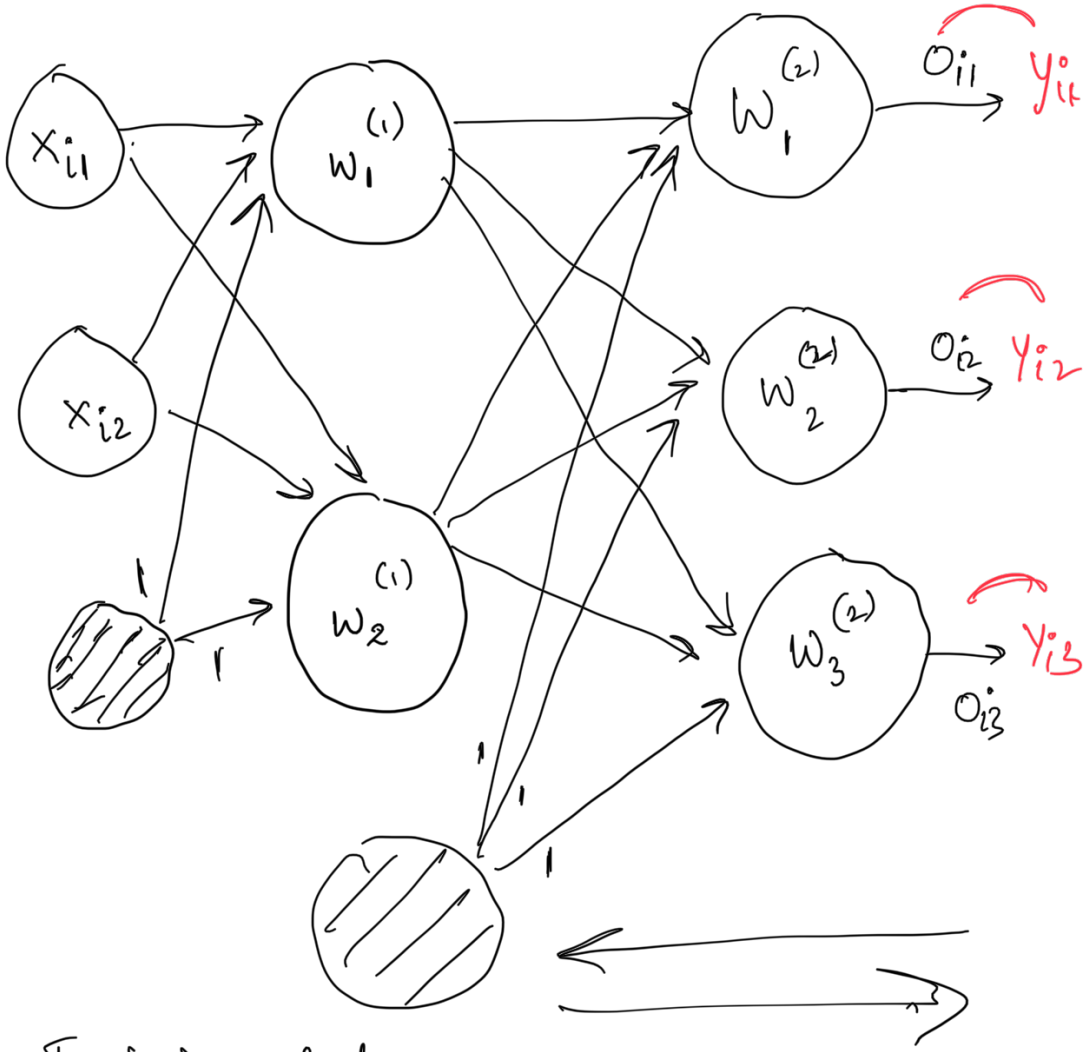
i	Training example	x_i
p	Feature	x_{ip}
j	Hidden layer unit	$w_j^{(1)}$
l	Output layer unit	$w_l^{(2)}$

Variables

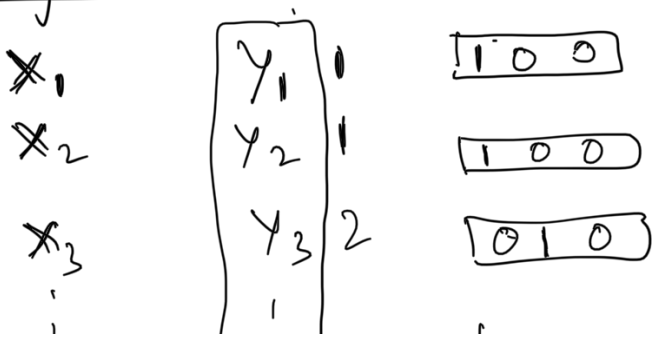
D+1	x_i	Input vector for i^{th} training example
D+1	$w_j^{(1)}$	weight vector at j^{th} hidden unit
M+1	$w_l^{(2)}$	weight vector at l^{th} output unit
	z_j	output of the j^{th} hidden unit
	o_l	output the l^{th} output unit
K	y_i	1-of-K true output for i^{th} training example

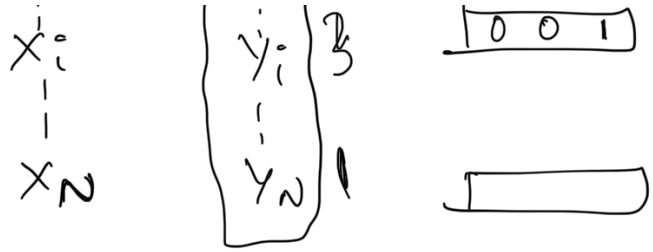
w a general unit weight vector

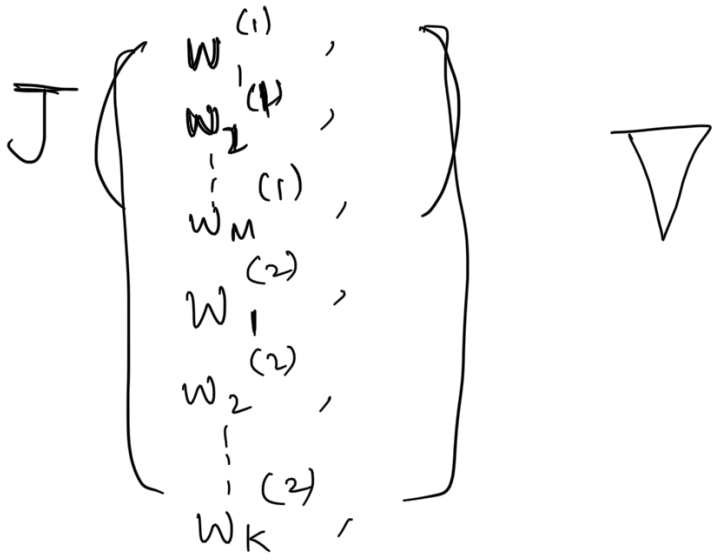
w_r	
u_r	input to a general unit
v_r	output of a general unit
net_r	intermediate output: $w_r^T u_r$

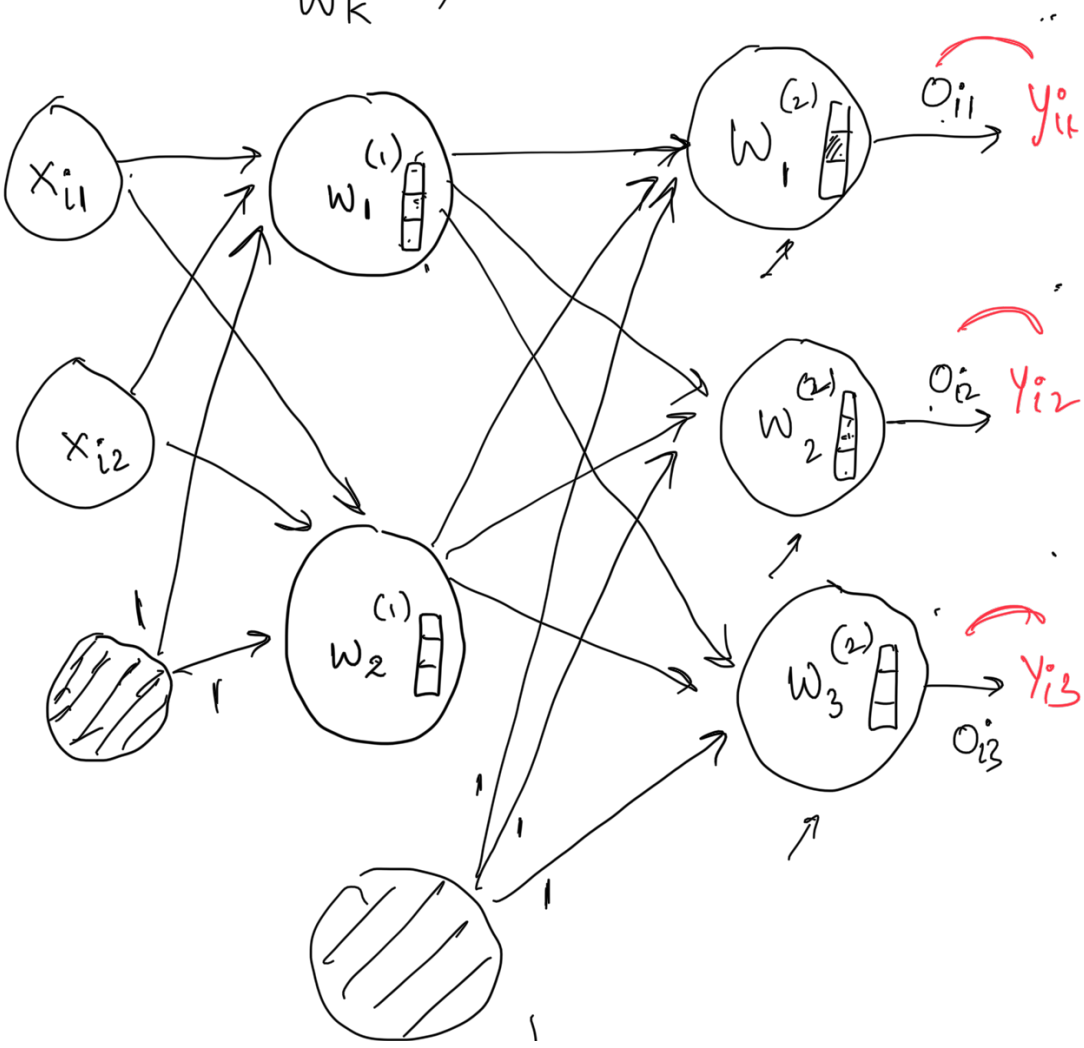


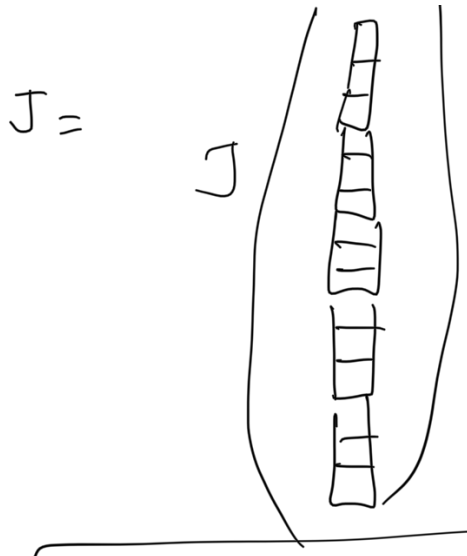
Training data




x_i
x_N
y_i
y_N
3
1
0 0 1


J
$w_1^{(1)}$
$w_2^{(1)}$
$w_M^{(1)}$
$w_1^{(2)}$
$w_2^{(2)}$
$w_K^{(2)}$


x_{i1}
x_{i2}
1
$w_1^{(1)}$
$w_2^{(1)}$
$w_1^{(2)}$
$w_2^{(2)}$
$w_3^{(2)}$
1
1
O_{i1}
y_{i1}
O_{i2}
y_{i2}
y_{i3}
O_{i3}



$$\frac{\partial J}{\partial w_{jp}^{(1)}} \qquad \frac{\partial J}{\partial w_{kj}^{(2)}}$$

scalar

$$w_{jp}^{(1)} \leftarrow w_{jp}^{(1)} - \eta \frac{\partial J}{\partial w_{jp}^{(1)}}$$

$$w_{kj}^{(2)} \leftarrow w_{kj}^{(2)} - \eta \frac{\partial J}{\partial w_{kj}^{(2)}}$$

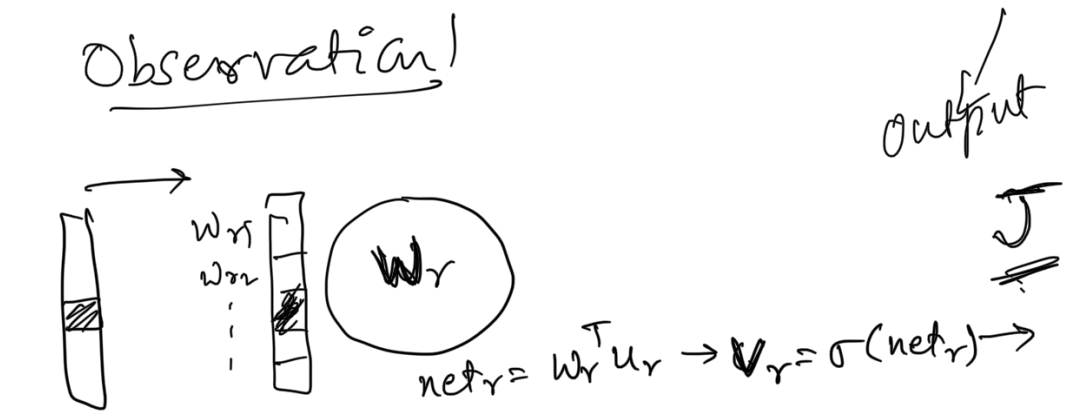
Drop i — Assume we have

only 1 training example.

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

$$\boxed{J = \frac{1}{2} \sum_{\ell=1}^{K} (y_\ell - o_\ell)^2}$$

Observation!



$$net_r = w_r^T u_r$$

$$= \sum_q w_{rq} u_{rq}$$

$$\frac{\partial J}{\partial w_{rq}} = \frac{\partial J}{\partial net_r} \frac{\partial net_r}{\partial w_{rq}}$$

← chain rule of

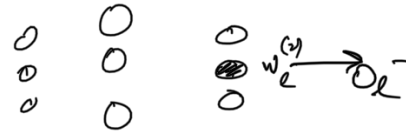
$$\frac{\partial net_r}{\partial w_{rq}} = u_{rq}$$

derivative

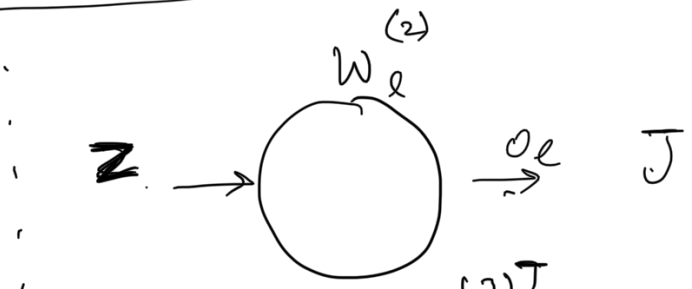
$$\frac{\partial J}{\partial w_{rq}} = \frac{\partial J}{\partial net_r} u_{rq}$$

Observation 2

For l^{th} output unit :



$$\frac{\partial J}{\partial net_l} = \frac{\partial J}{\partial o_l} \frac{\partial o_l}{\partial net_l}$$



$$net_l = w_l^{(2)T} z$$

$$o_l = \sigma(net_l)$$

$$J = \frac{1}{2} \sum_{l=1}^{K} (y_l - o_l)^2$$

What is $\frac{\partial J}{\partial o_l}$?

$$J = \frac{1}{2}\left[(y_1 - o_1)^2 + (y_2 - o_2)^2 + (y_3 - o_3)^2\right.$$

$$2\left[\cdots + (y_k - o_k)^2 + \cdots\right]$$

$$\frac{\partial J}{\partial o_k} = \frac{1}{2} 2 (y_k - o_k)(-1)$$

$$= -(y_k - o_k) \quad \text{(\#1)}$$

What is $\frac{\partial o_k}{\partial net_k}$

$$o_k = \sigma(net_k)$$

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

$$\frac{\partial o_k}{\partial net_k} = \frac{\partial}{\partial net_k}\left[\frac{1}{1+e^{-net_k}}\right]$$

$$= -\frac{1}{(1+e^{-net_k})^2}\left(-e^{-net_k}\right)$$

$$\frac{\partial o_k}{\partial net_k} = \frac{e^{-net_k}}{(1+e^{-net_k})^2} = o_k(1-o_k) \quad \text{(\#2)}$$

Combining #1 & #2:

$$\frac{\partial J}{\partial net_k} = -o_k(1-o_k)(y_k - o_k)$$

Let $\boxed{\delta_k = o_k(1-o_k)(y_k - o_k)}$

$$\frac{\partial J}{\partial w_{kj}^{(2)}} = \frac{\partial J}{\partial net_j} z_j$$

$$= -o_k(1-o_k)(y_k - o_k) z_j$$

$$= -\delta_k z_j$$

Update rule for output layer.

$$\boxed{w_{kj}^{(2)} \leftarrow w_{kj}^{(2)}_{old} + \eta \delta_k z_j}$$

Preview:

$$\boxed{w_{jp}^{(1)} \leftarrow w_{jp}^{(1)}_{old} + \eta \delta_j x_p}$$

δ_j = function of δ_k s.