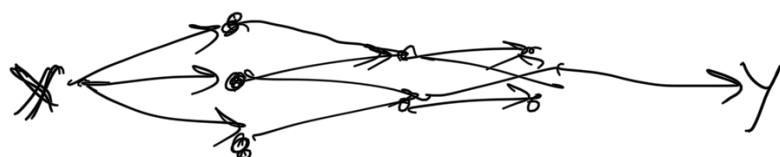


$$y = f(\mathbf{x})$$

Neural Networks.

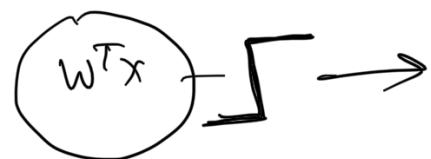


$$\underline{w^T x}$$

Thresholded
perception

$$\begin{aligned} w_1 x_1 \\ + w_2 x_2 \\ + w_3 x_3 \\ + w_4 x_4 \\ + w_5 x_5 \end{aligned}$$

$$\boxed{w^T x \geq 0} \rightarrow \begin{cases} 1 \\ -1 \end{cases}$$

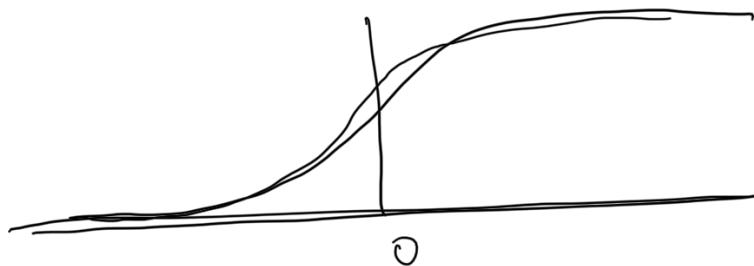


Unit

Layer

Sigmoid unit

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

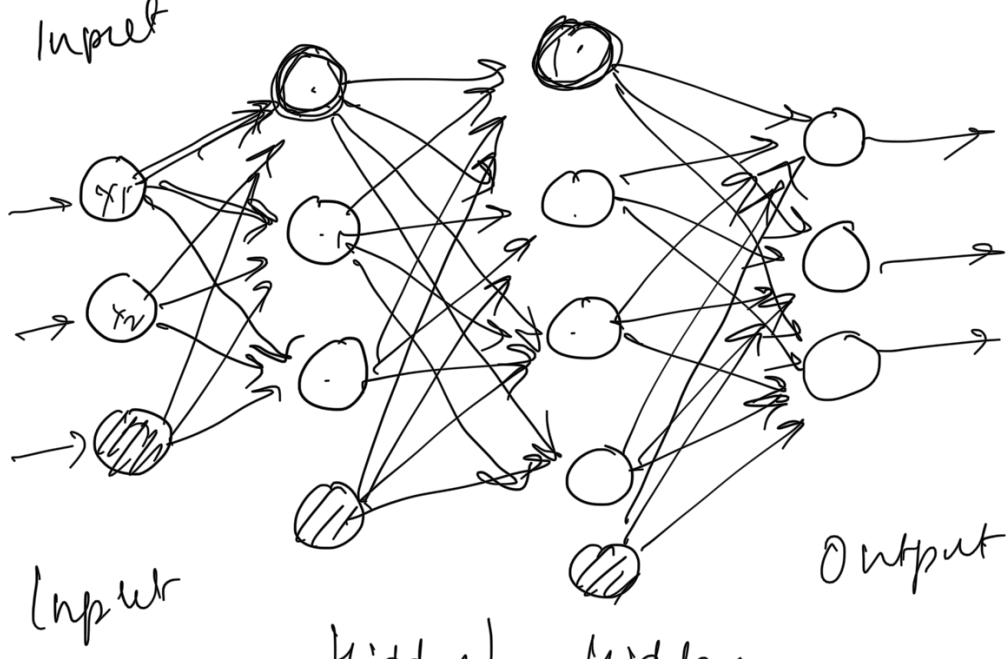


March 12 Friday

- PA 2 is out
- Mid-term next Friday

$$D=2$$

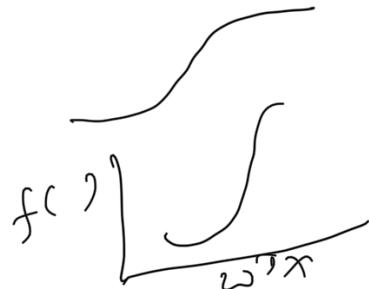
Input



Muadra' Muadra'

of units in a layer \equiv width
of a layer

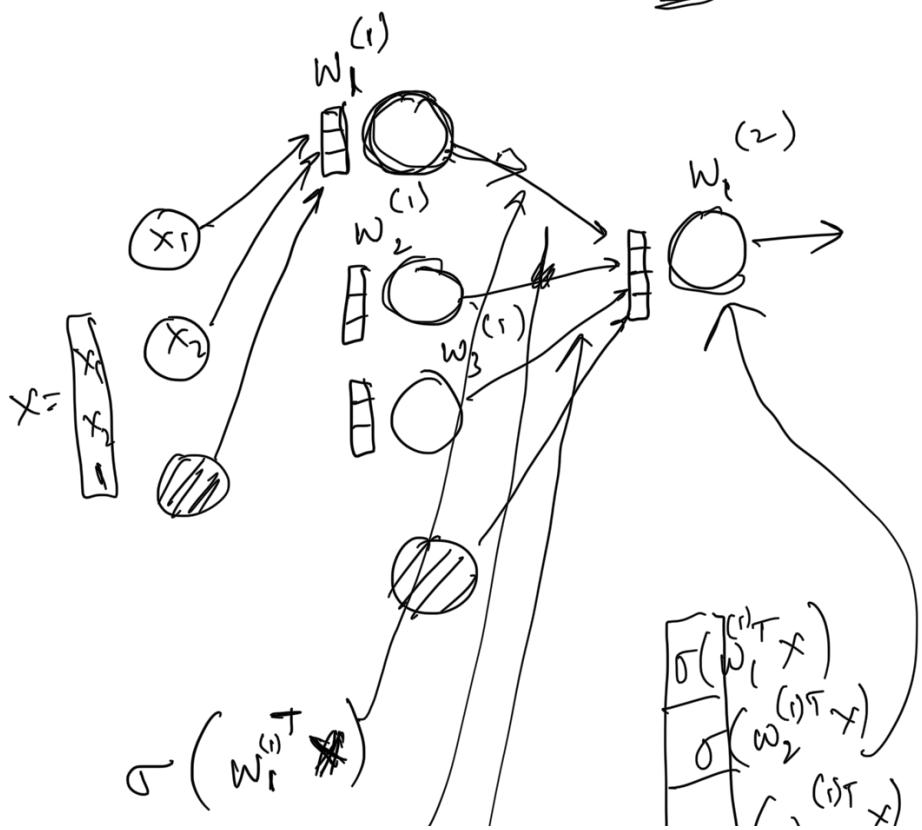
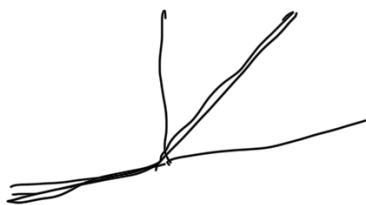
Sigmoid



tanh

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

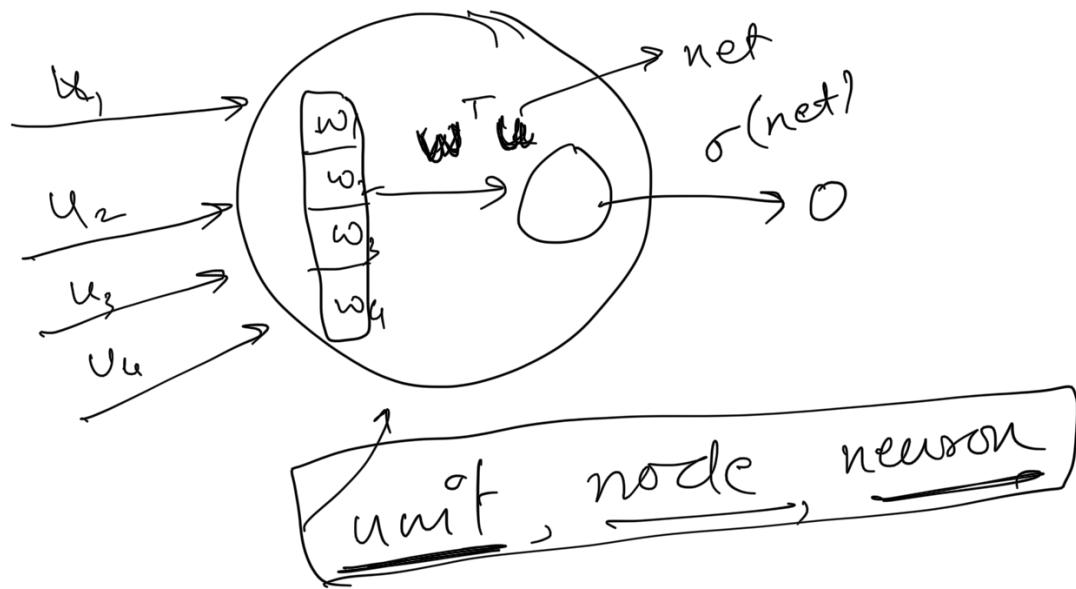
ReLU



$$\sigma(w_2^{(1)T}x)$$

$$\sigma(w_3^{(1)T}x)$$

$$\sigma(w_3^{(2)T}x)$$



Sigmoid

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

Tanh

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

Simple example

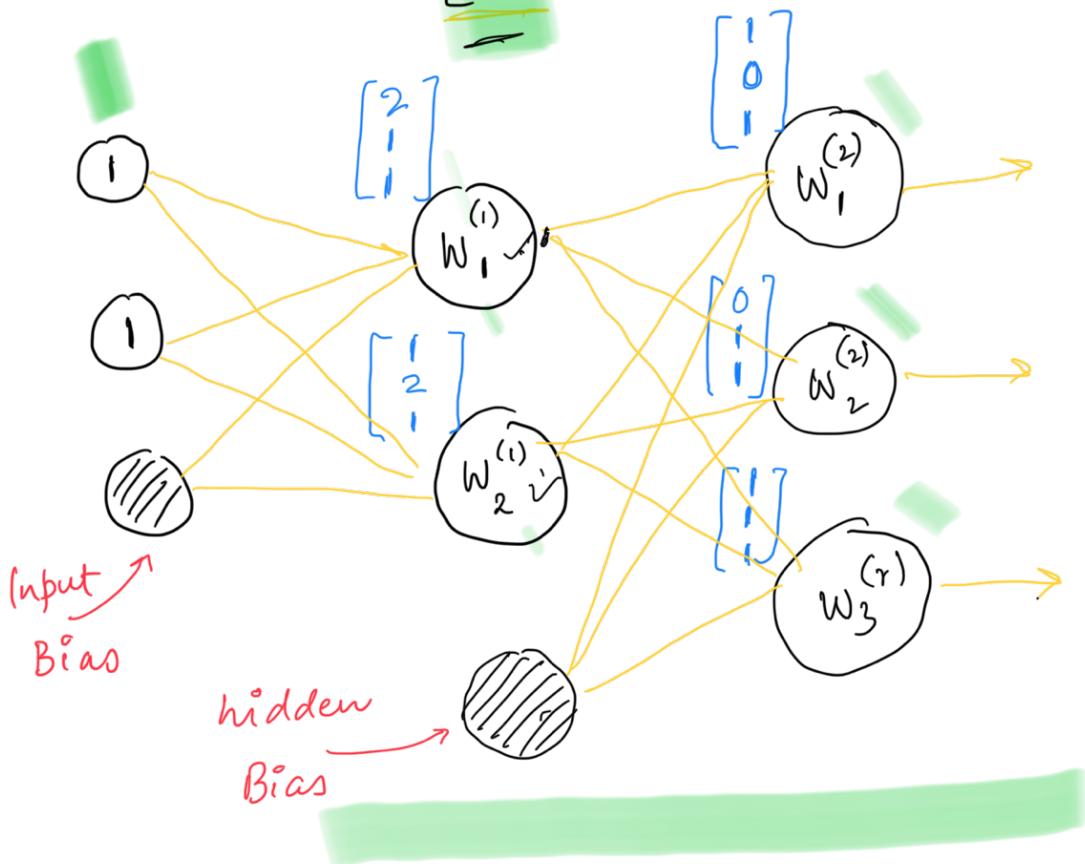
$D = 2$ (2-d data)

$M = 2$

$K = 3$ (3 outputs / classes)

Assume sigmoid activation

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



At hidden unit 1

$$\text{net}_1^{(1)} = \mathbf{w}_1^{(1)\top} \mathbf{x}$$

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

$$z_1 = \sigma(\text{net}_1^{(1)}) = \sigma(4)$$

$$= 0.98$$

At hidden unit 2

$$\text{net}_2^{(1)} = w_2^{(1)} X$$

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

$$z_2 = \sigma(\text{net}_2^{(1)}) = \sigma(4)$$

$$= 0.98$$

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

At Output unit 1

$$\text{net}_1^{(2)} = w_1^{(2)T} \underline{\mathbf{z}}$$

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

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

$$= 0.88$$

At output unit 2

$$\text{net}_2^{(2)} = w_2^{(2)T} \underline{\mathbf{z}}$$

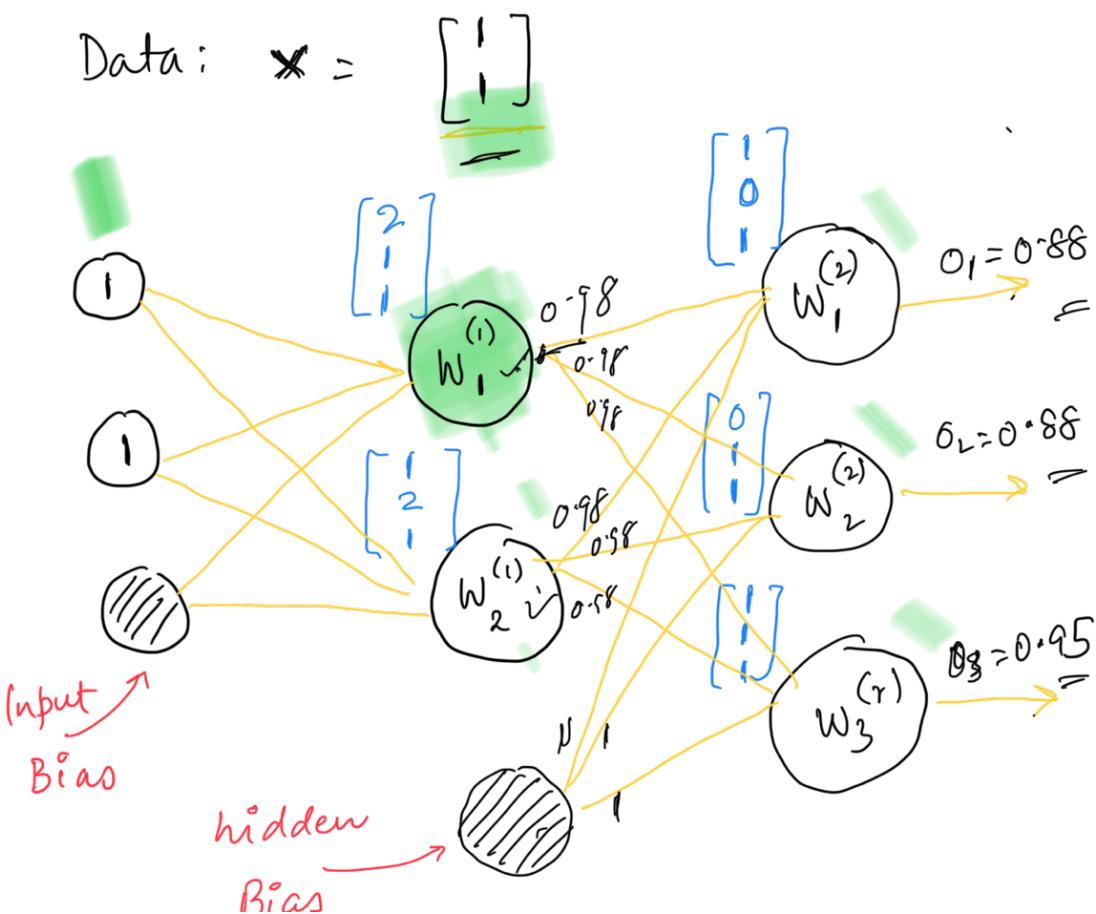
$$\text{net}_2 = \sum z_i w_{i2} = [0 \ 1 \ 1] \begin{bmatrix} 0.98 \\ 0.98 \\ 1 \end{bmatrix} = 1.98$$

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

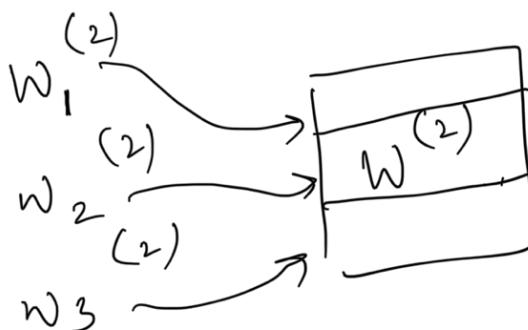
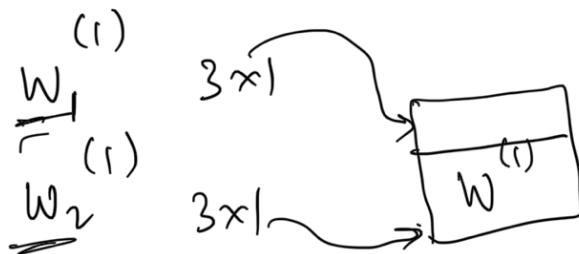
At output unit 3

$$\text{net}_3^{(2)} = w_3^{(2)T} z = [1 \ 1 1] \begin{bmatrix} 0.98 \\ 0.98 \\ 1 \end{bmatrix} = 2.96$$

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



softmax output : $\left[\frac{o_1}{o_1 + o_2 + o_3}, \frac{o_2}{o_1 + o_2 + o_3}, \frac{o_3}{o_1 + o_2 + o_3} \right]$

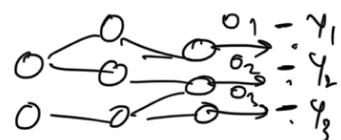


$$\sigma \left(\underbrace{W^{(1)} \times}_{3 \times 3} \underbrace{x}_{3 \times 1} \right) = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \bar{z} = \begin{bmatrix} 2 \\ z_2 \\ 1 \end{bmatrix}$$

$$\sigma \left(\underbrace{W^{(2)} \bar{z}}_{3 \times 2} \right) = \begin{bmatrix} o_1 \\ o_2 \\ o_3 \end{bmatrix}$$

$$J(W^{(1)}, \dots, W^{(L)}, \dots)$$

$$= \sum_i^N J_i$$



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

o_{il} → output for the
ith training example
at output unit l

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

y_{il} → true output for the
ith example at
lth output unit.

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	1	0
---	---	---	---

1 →

1	1	0	0
---	---	---	---

3 →

1	0	0	1
---	---	---	---

Condition + Descent

Gradient Descent

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

$$\boxed{\partial J}$$

Wednesday March 17

Notation

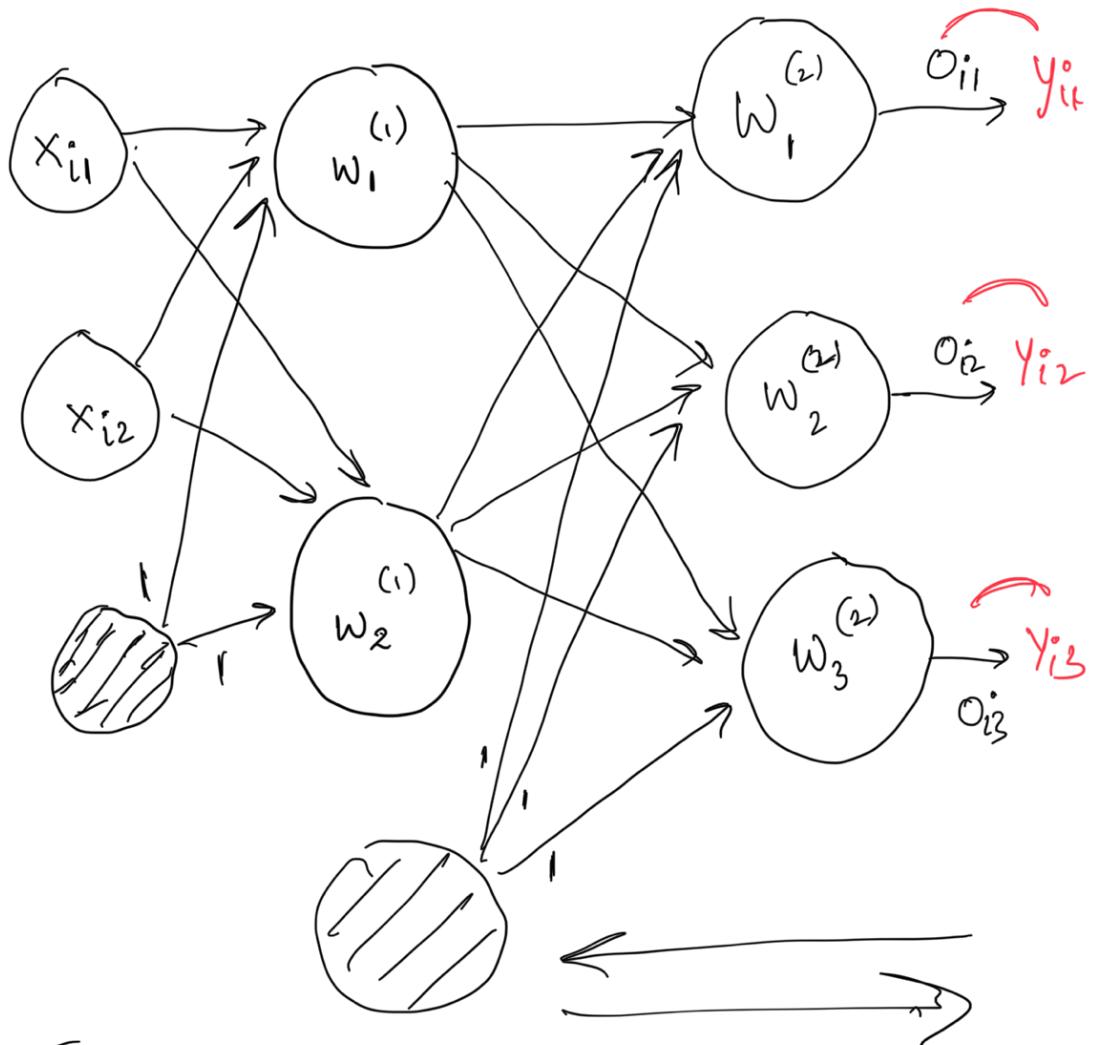
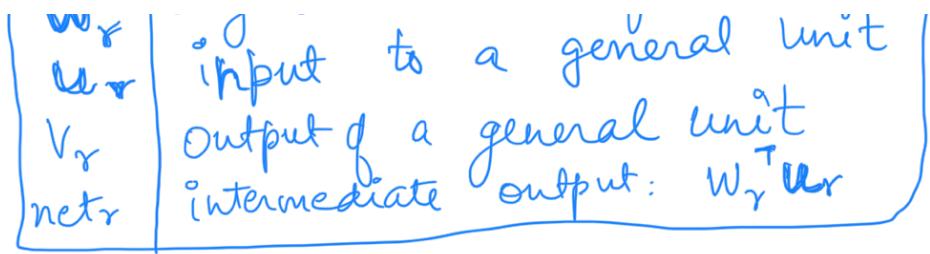
Subscripts

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

Variables

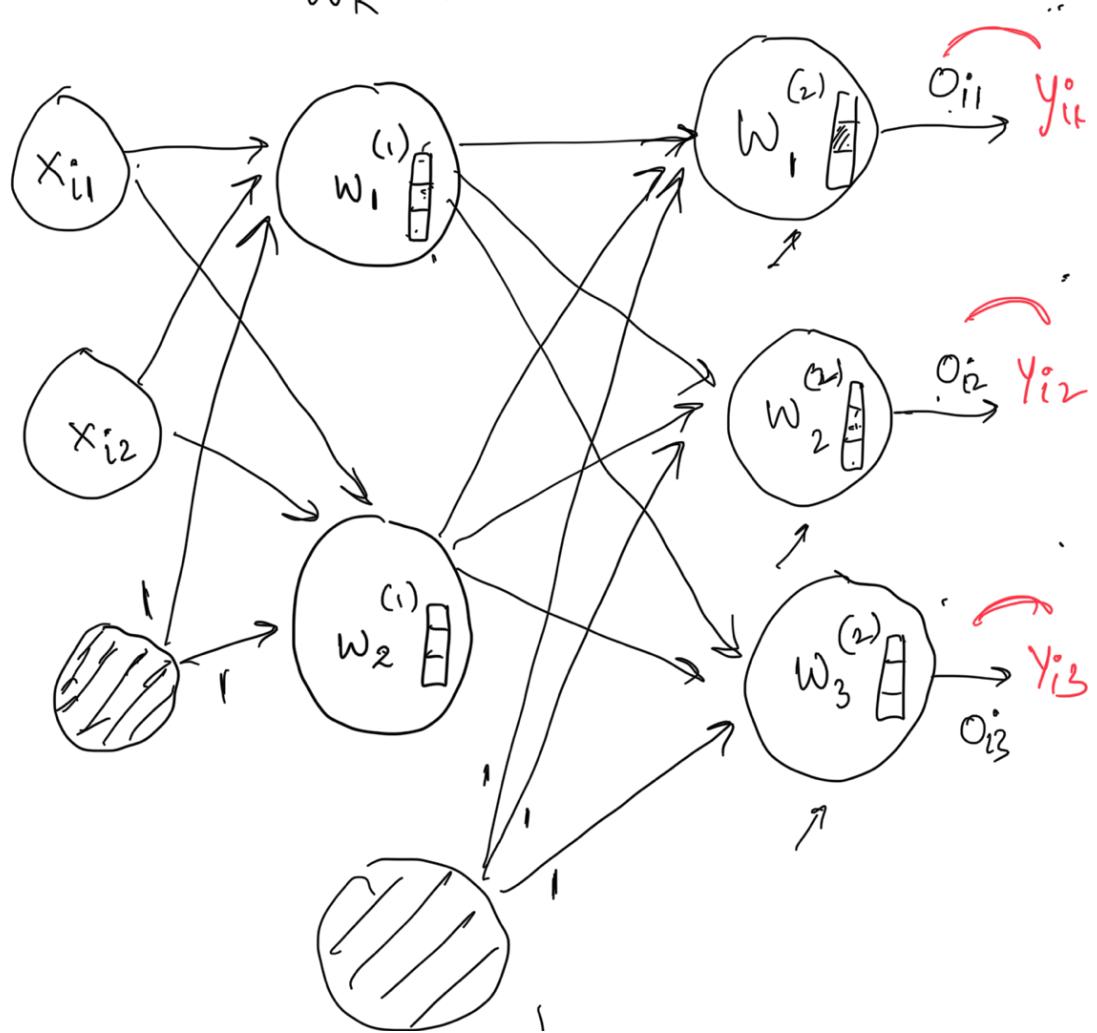
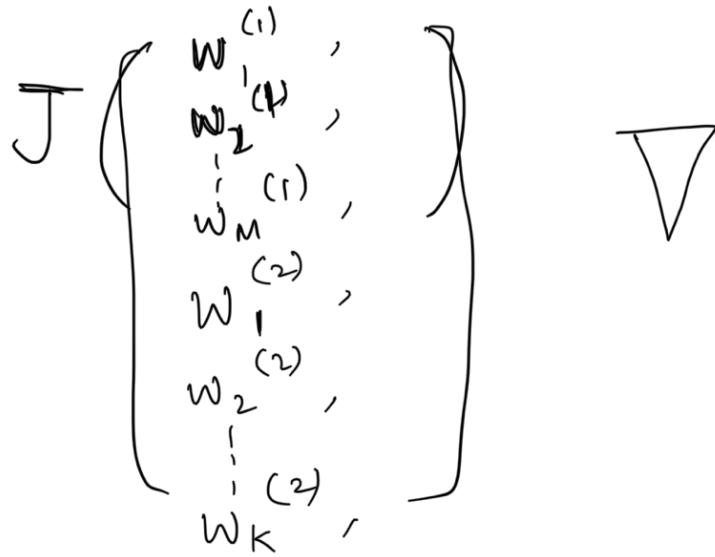
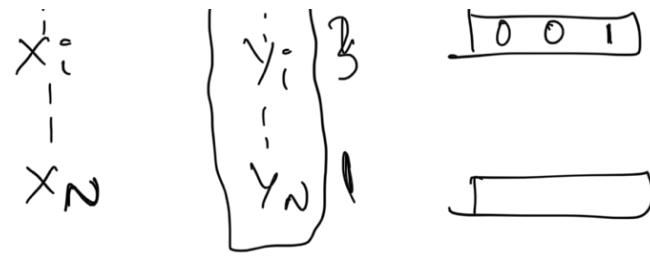
$D+1$	x_i	Input vector for i^{th} training example
$D+1$	$w_j^{(1)}$	Weight vector at j^{th} hidden unit
$M+1$	$w_l^{(2)}$	Weight vector at l^{th} output unit
	z_j	Output of the j^{th} hidden unit
	o_l	Output of the l^{th} output unit
K	y_i	True output for i^{th} training example

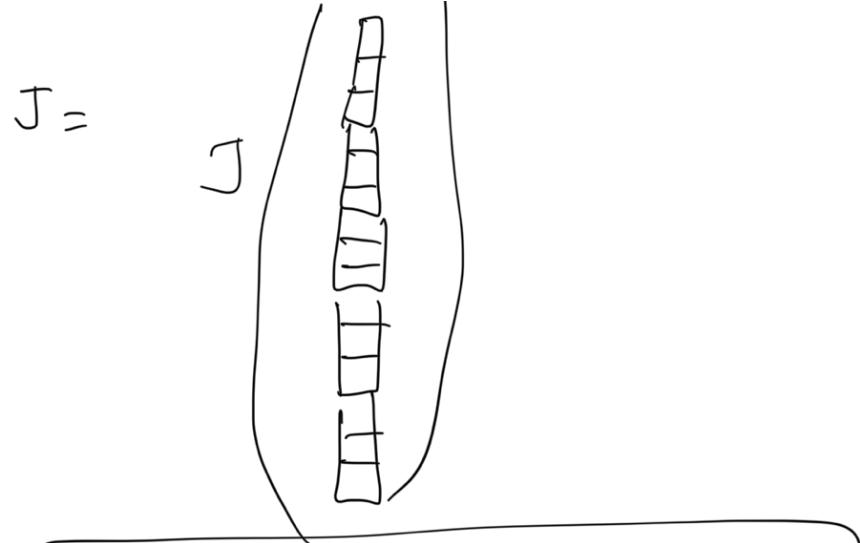
\vdots A general unit weight vector



Training data

x_1	y_1	1	1 0 0
x_2	y_2	1	1 0 0
x_3	y_3	2	0 1 0
:	:		





$$\frac{\partial J}{\partial \underline{w_{jp}^{(1)}}}$$

$$\frac{\partial J}{\partial \underline{\underline{w_{lj}^{(2)}}}}$$

scalar

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

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

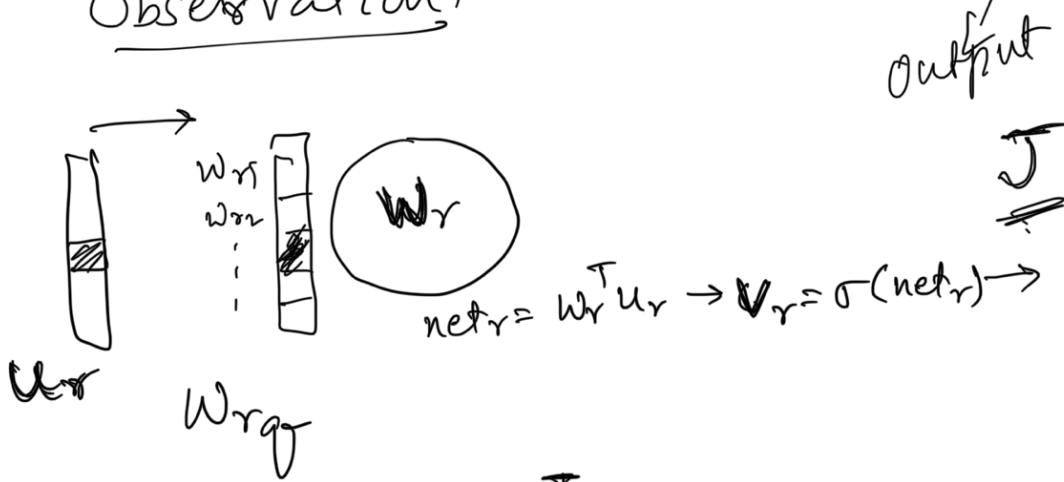
Drop i = Assume we have

only 1 training example.

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

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

Observation



$$net_r = w_r^T u_r$$

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

$$\frac{\partial J}{\partial w_{rq}} = \frac{\partial J}{\partial net_r} \left(\frac{\partial net_r}{\partial w_{rq}} \right)$$

chain rule of limits

... \dots

$$\frac{\partial \text{net}_r}{\partial w_{rq}} = u_{rq} /$$

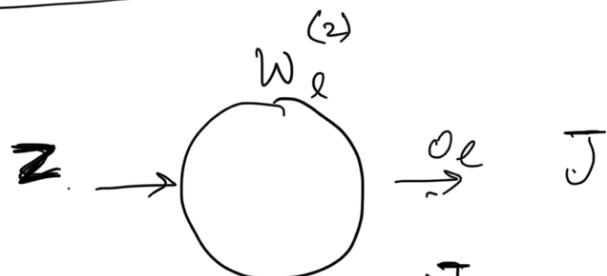
derivar

$$\frac{\partial J}{\partial w_{rq}} = \frac{\partial J}{\partial \text{net}_r} u_{rq}$$

Observation 2

for l^{th} output unit :

$$\frac{\partial J}{\partial \text{net}_e} = \frac{\partial J}{\partial o_e} \frac{\partial o_e}{\partial \text{net}_e}$$



$$o_e = \sigma(\text{net}_e)$$

$$J = \frac{1}{2} \sum_{k=1}^K (y_k - o_e)^2$$

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

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

$$J = \frac{1}{2} \sum_{e=1}^E (y_e - o_e)^2$$

$$\frac{\partial J}{\partial o_e} = \frac{1}{2} \sum_{e=1}^E 2(y_e - o_e) (-1)$$

$$= -(y_e - o_e) \quad (\#1)$$

What is $\frac{\partial o_e}{\partial \text{net}_e}$

$$o_e = \sigma(\text{net}_e)$$

$$= \frac{1}{1 + e^{-\text{net}_e}}$$

$$\frac{\partial o_e}{\partial \text{net}_e} = \frac{\partial}{\partial \text{net}_e} \left[\frac{1}{1 + e^{-\text{net}_e}} \right]$$

$$= - \frac{1}{(1 + e^{-\text{net}_e})^2} (-e^{-\text{net}_e})$$

$$\frac{\partial o_e}{\partial \text{net}_e} = \frac{e^{-\text{net}_e}}{(1 + e^{-\text{net}_e})^2} = \underline{\underline{o_e(1 - o_e)}} \quad (\#2)$$

Combining #1 & #2:

$$\frac{\partial J}{\partial \text{net}_e} = -o_e(1 - o_e)(y_e - o_e)$$

Let $\boxed{\delta_e = o_e(1-o_e)(y_e - o_e)}$

$$\begin{aligned}\frac{\partial J}{\partial w_{ej}^{(2)}} &= \frac{\partial J}{\partial \text{net}_j} z_j \\ &= -o_e(1-o_e)(y_e - o_e)z_j \\ &= -\delta_e z_j\end{aligned}$$

Update rule for output layer.

$$w_{ej}^{(2)} \leftarrow \underset{\text{old}}{w_{ej}^{(2)}} + \eta \delta_e z_j$$

Preview:

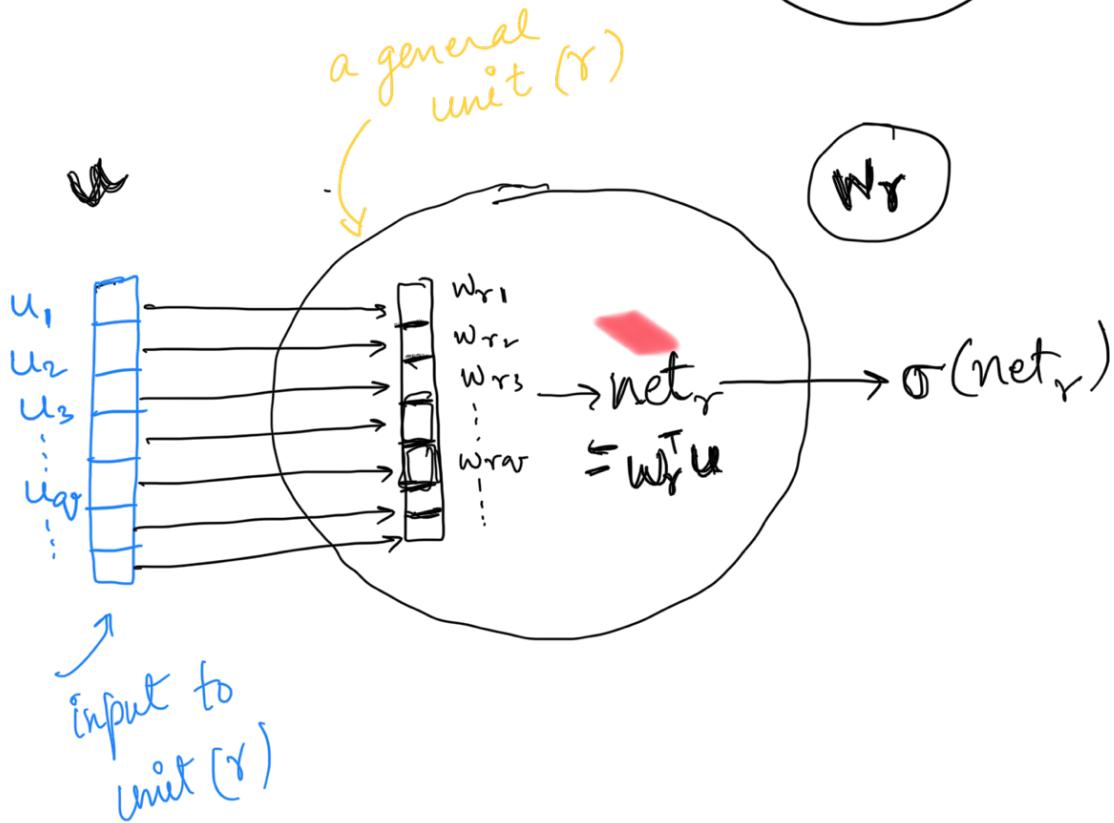
$$w_{jp}^{(1)} \leftarrow \underset{\text{old}}{w_{jp}^{(1)}} + \eta (\delta_j x_u)$$

δ_j = function of δ_e 's.

Monday Mar 22

What do we need?

$$\frac{\partial J}{\partial w_{rg}}$$



Observation 1

$$\frac{\partial J}{\partial w_{rg}} = \frac{\partial J}{\partial \text{net}_r} \frac{\partial \text{net}_r}{\partial w_{rg}} = \frac{\partial J}{\partial \text{net}_r} u_q$$

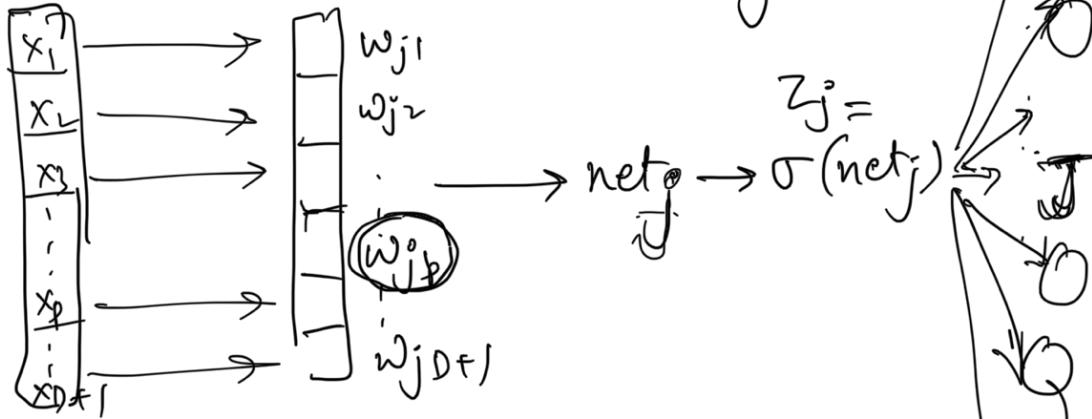
Observation 2

For weight on an output unit:

$$\frac{\partial J}{\partial w_{ej}} = -s_e z_j$$

$$\check{s}_e = (y_e - o_e) o_e (1 - o_e)$$

for a hidden unit w_j



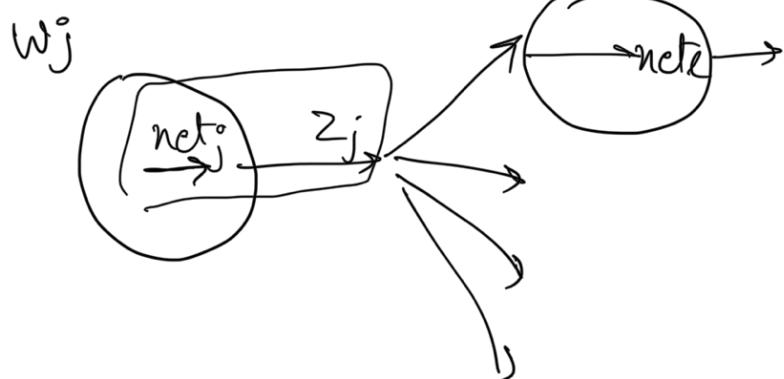
$$\frac{\partial J}{\partial \text{net}_j} = \sum_{l=1}^K \frac{\partial J}{\partial \text{net}_l} \frac{\partial \text{net}_l}{\partial \text{net}_j}$$

We already know: $\frac{\partial J}{\partial \text{net}_e} = -s_e$

Check wed derivation

$$\frac{\partial J}{\partial \text{net}_j} = - \sum_{l=1}^K s_e \frac{\partial \text{net}_e}{\partial \text{net}_j}$$

$$= - \sum_{l=1}^K \left[s_l \left(\frac{\partial \text{net}_l}{\partial z_j^o} \right) \left(\frac{\partial z_j^o}{\partial \text{net}_j} \right) \right]$$



$$z_j^o = \sigma(\text{net}_j)$$

$$\Rightarrow \boxed{\frac{\partial z_j^o}{\partial \text{net}_j} = z_j^o (1 - z_j^o)}$$

$$\frac{\partial J}{\partial \text{net}_j} = - \sum_{l=1}^K \left[s_l \left(\frac{\partial \text{net}_l}{\partial z_j^o} \right) \right] z_j^o (1 - z_j^o)$$

$$\frac{\partial \text{net}_e}{\partial z_j^o}$$

Recall that $\text{net}_e = \sum_{j=1}^{M+1} w_{ej} z_j^o$

$$\frac{\partial \text{net}_e}{\partial z_j^o} = w_{eo}$$

$$\overline{\frac{\partial z_j}{\partial z_j}} = -x_j$$

$$\Rightarrow \frac{\partial J}{\partial w_{netj}} = -z_j(1-z_j) \left(\sum_{l=1}^K \delta_l w_{ej} \right)$$

$$\frac{\partial J}{\partial w_{jp}} = -z_j(1-z_j) \left(\sum_{l=1}^K \delta_l w_{ej} \right) x_p$$

$$\delta_j = \alpha z_j(1-z_j) \left(\sum_{l=1}^K \delta_l w_{ej} \right)$$

$$\boxed{\frac{\partial J}{\partial w_{jp}} = -\delta_j x_p}$$

$$w_{jp} = w_{jp} + \eta \underline{\delta_j x_p}$$

$$w_{ej} = w_{ej} + \eta \underline{\delta_e z_j}$$

