AND

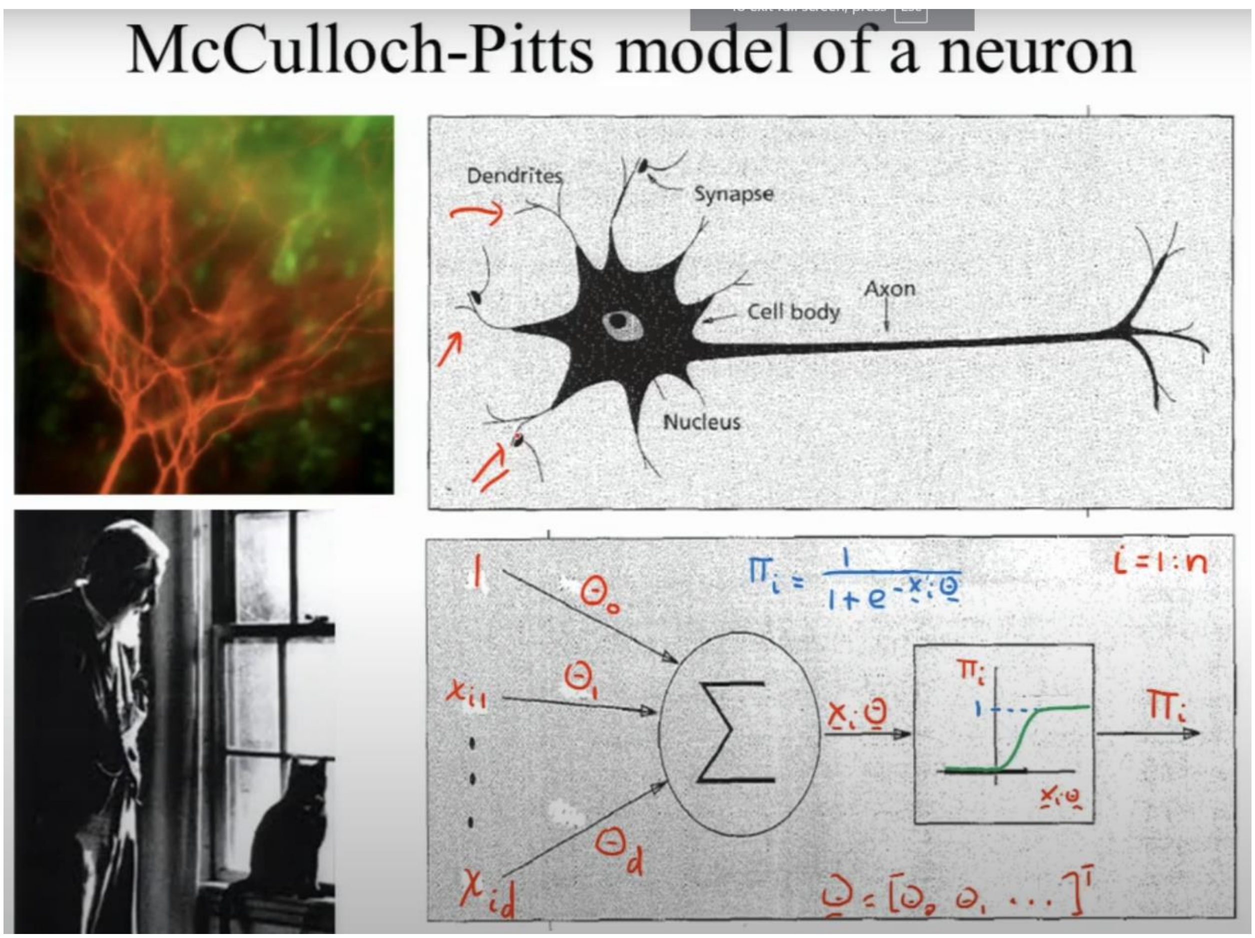
x1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1

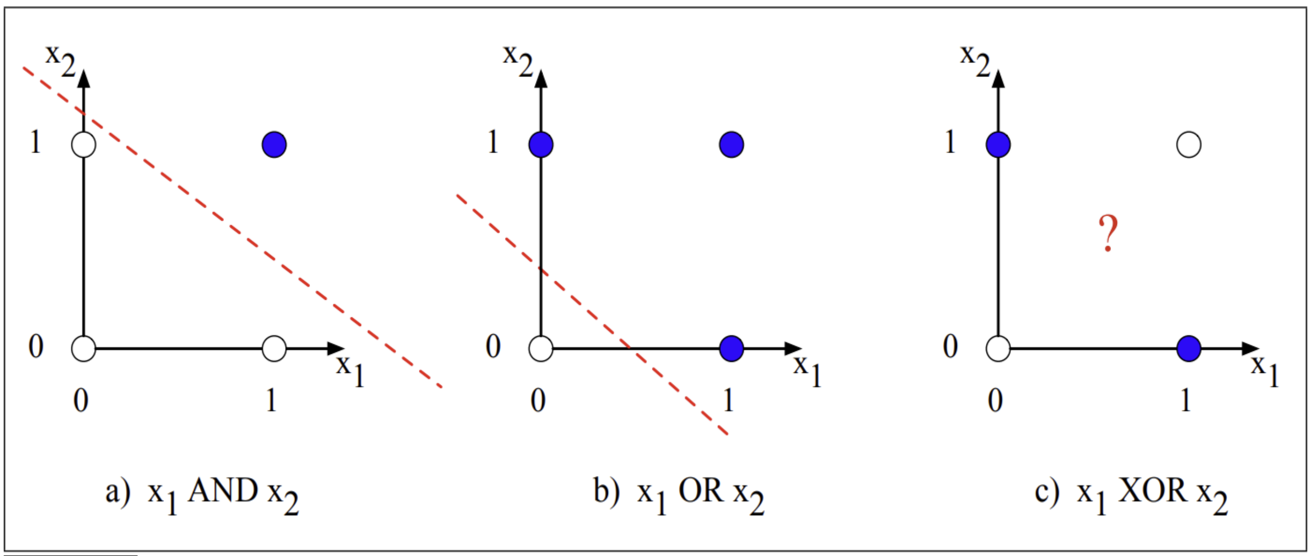
OR

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	1

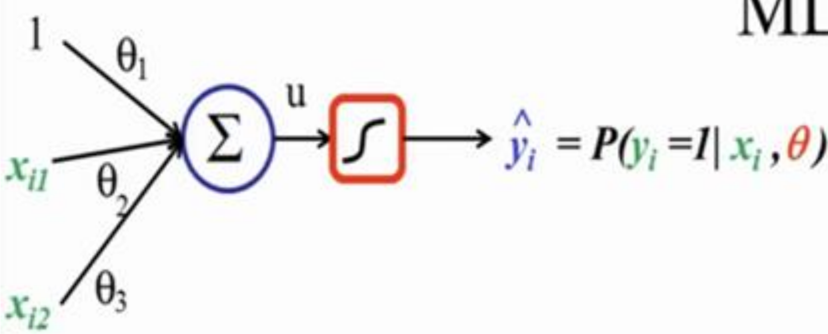
XOR

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0




x2
1
0
0
1
x1
a) x1 AND x2
b) x1 OR x2
?
c) x1 XOR x2

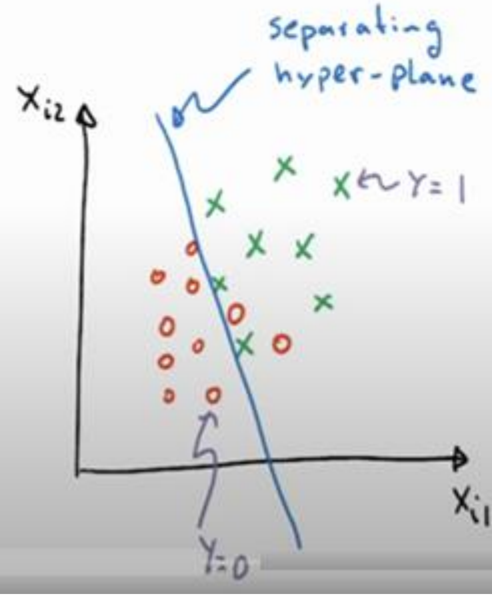
MLP – 1 neuron



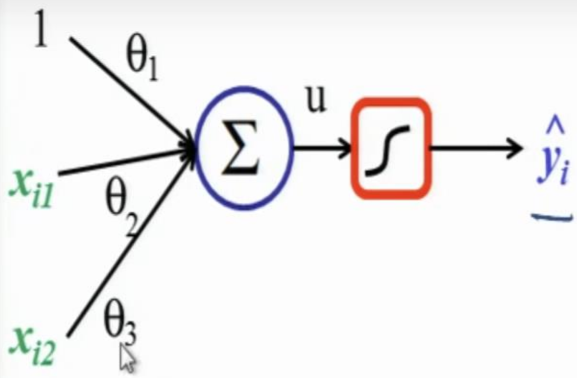
We are given the data $\{x_i, y_i\}_{i=1}^{n}$

eg.

	x_{i1}	x_{i2}	y_i
i=1	0.2	6	0
i=2	0.3	22	1
i=3	0.6	-0.6	1
i=4	-0.4	58	0
⋮			



MLP – 1 neuron



$\hat{y}_i = P(y_i = 1 | x_i, \theta)$

$0 < \hat{y}_i < 1$

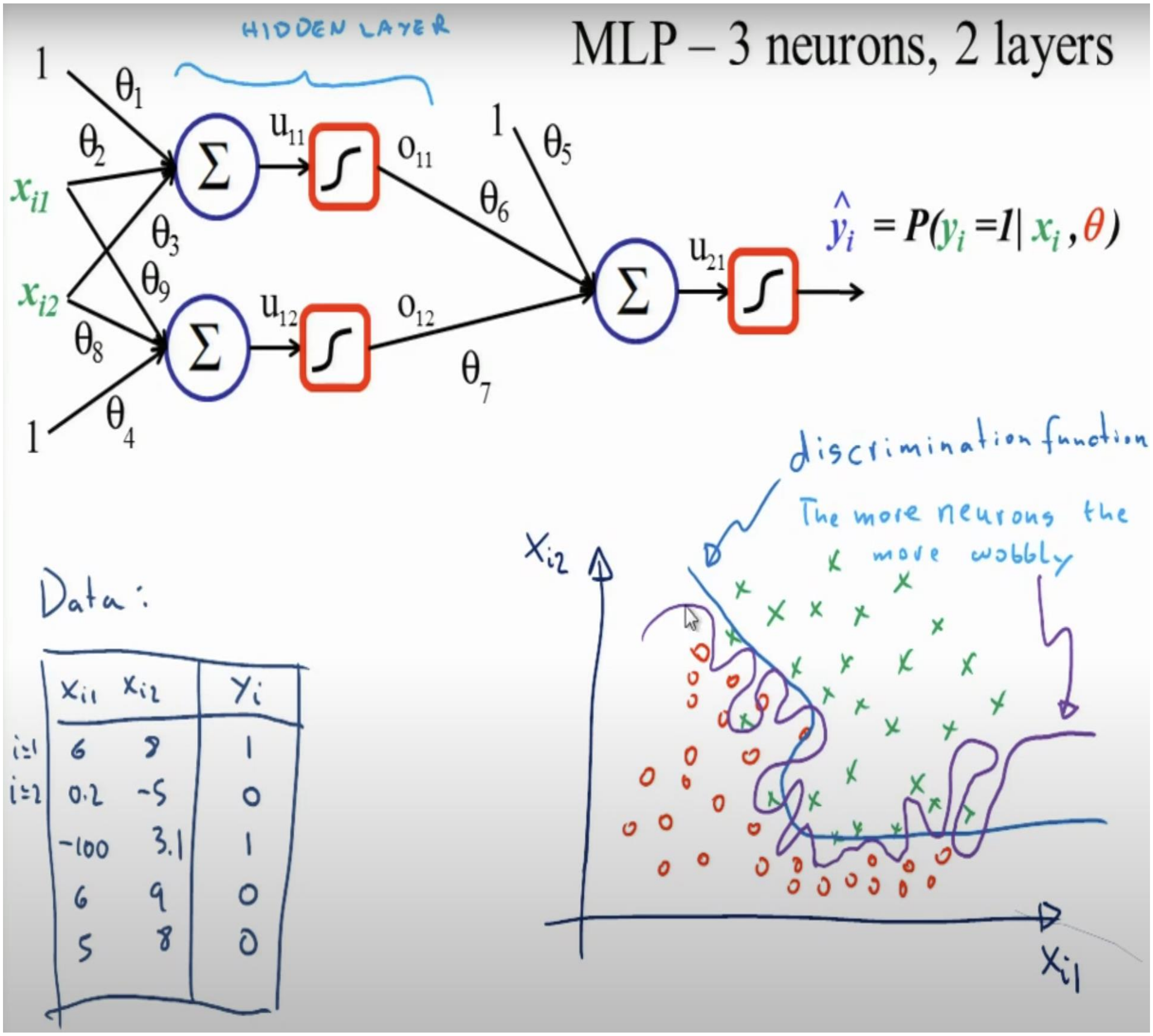
$$\boxed{u = \theta_1 + \theta_2 x_{i1} + \theta_3 x_{i2}}$$

$$\hat{y}_i = \frac{1}{1+e^{-\boxed{u}}} = \frac{1}{1+e^{-\theta_1-\theta_2 x_{i1}-\theta_3 x_{i2}}} = P(y_i = 1 | \underline{x}_i, \underline{\theta})$$

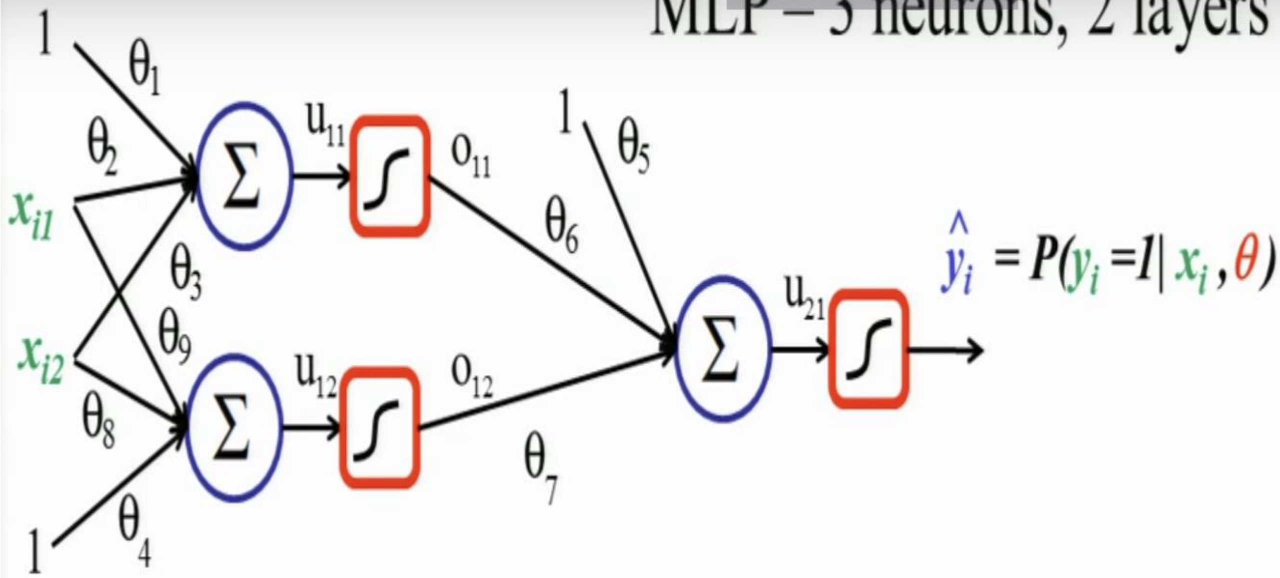
$$P(y_i | \underline{x}_i, \underline{\theta}) = \underline{\hat{y}_i}^{\,y_i} (1-\hat{y}_i)^{1-y_i} = \begin{cases} \hat{y}_i & \text{when } y_i = 1 \\ 1-\hat{y}_i & \text{otherwise} \end{cases}$$

For n independent observations (Bernoulli)

$$P(\underline{y} | \underline{\underline{x}}, \underline{\theta}) = \prod_{i=1}^{n} P(y_i | \underline{x}_i, \underline{\theta})$$


MLP – 3 neurons, 2 layers
HIDDEN LAYER
1
θ1
θ2
θ3
θ4
θ5
θ6
θ7
θ8
θ9
x_i1
x_i2
u11
u12
u21
o11
o12
ŷ_i = P(y_i =1| x_i , θ)
discrimination function
The more neurons the more wobbly
Data:
x_i1 x_i2 y_i
i=1 6 8 1
i=2 0.2 -5 0
-100 3.1 1
6 9 0
5 8 0
x_i2
x_i1

MLP – 3 neurons, 2 layers



$u_{11} = \Theta_1 + \Theta_2 x_{i1} + \Theta_3 x_{i2}$

$u_{12} = O_4 + \Theta_9 x_{i1} + \Theta_8 x_{i2}$

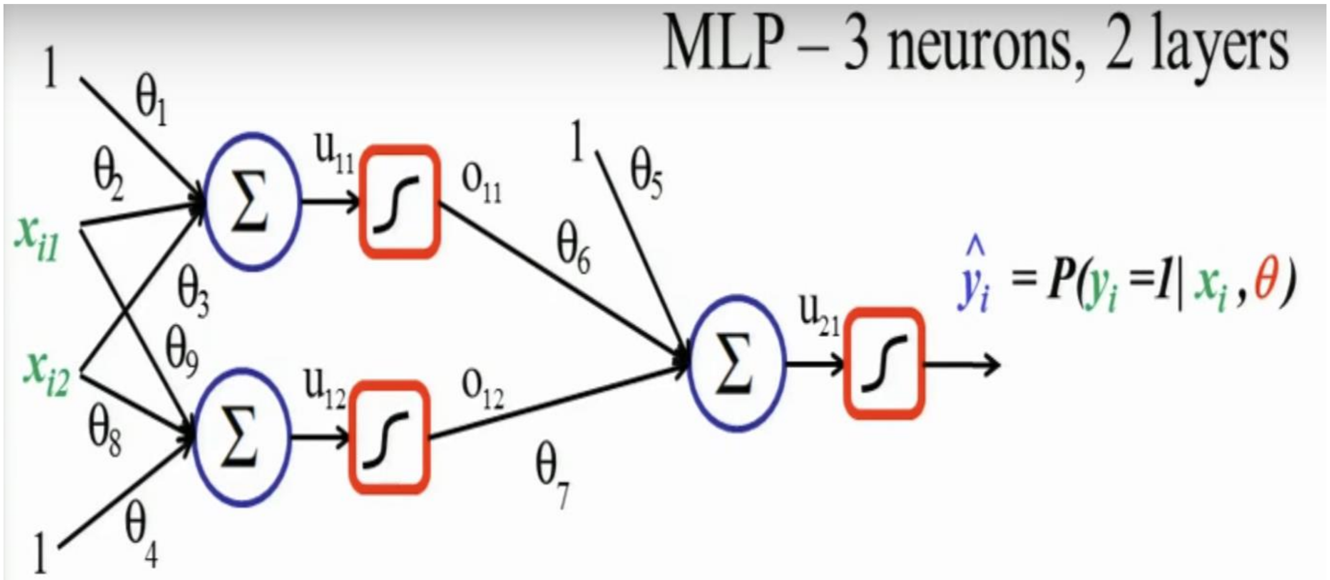
$O_{11} = 1/_{1+e^{-u_{11}}}$

$O_{12} = 1/_{1+e^{-u_{12}}}$

$\hat{y}_i = \dfrac{1}{1+e^{-u_{21}}}$

$u_{21} = \Theta_5 + \Theta_6 O_{11} + \Theta_7 O_{12}$

$$P(y_i | \underline{x}_i, \underline{\theta}) = \hat{y}_i^{\,y_i} (1-\hat{y}_i)^{1-y_i}$$



For n independent observations.

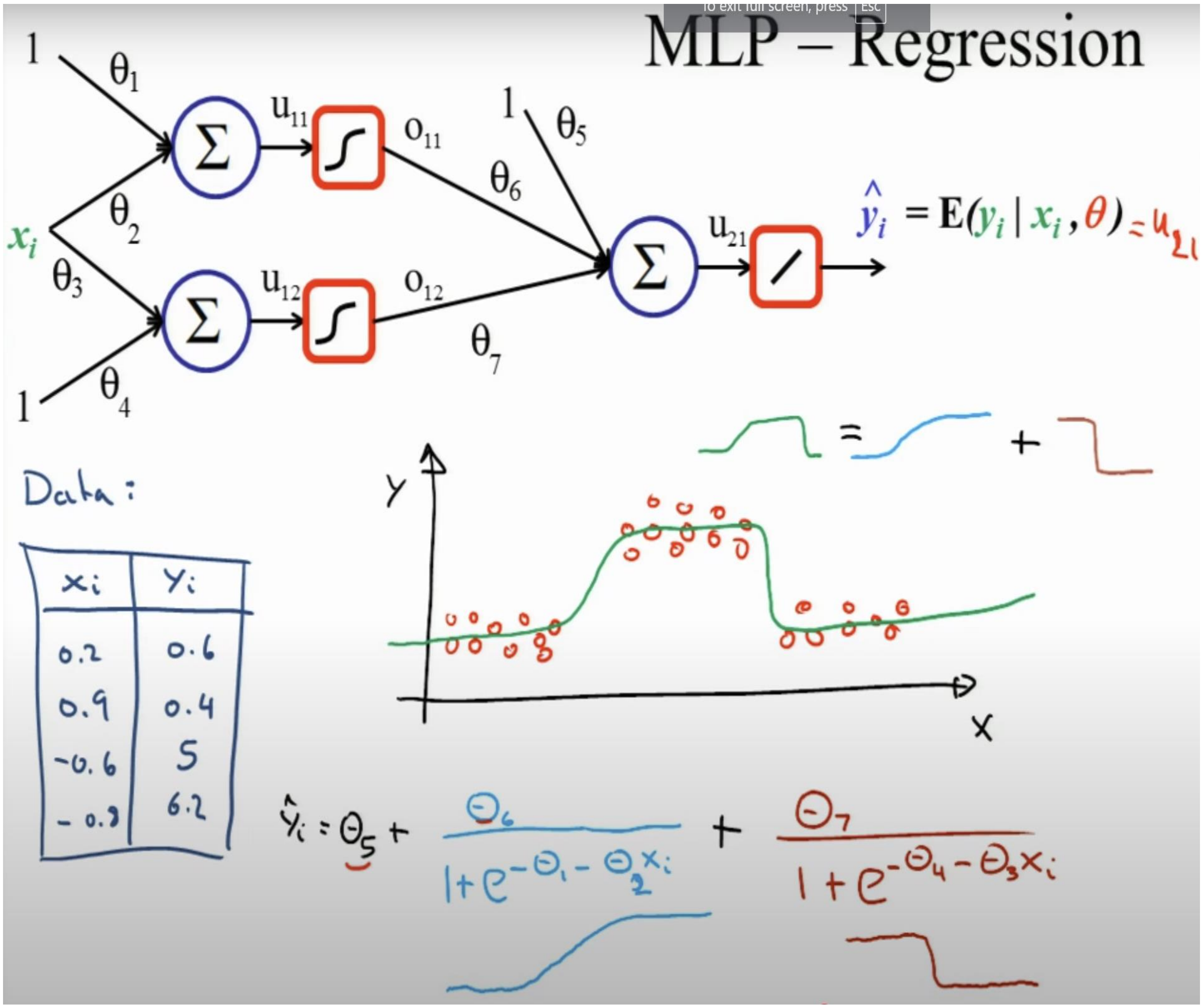
$$P(\underline{y} \mid \underline{x}, \underline{\theta}) = \prod_{i=1}^{n} \hat{y}_i^{\,y_i} (1-\hat{y}_i)^{1-y_i} = \prod_{i=1}^{n} P(y_i \mid x_i, \theta)$$

Cost:

$$C(\theta) = -\log P(\underline{y} \mid \underline{x}, \underline{\theta}) = -\sum_{i=1}^{n} y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i)$$

i.e minimize the cross-entropy error.

Cross-entropy measures uncertainty. By minimizing cross-entropy, we maximize the information gained about the data as the model learns



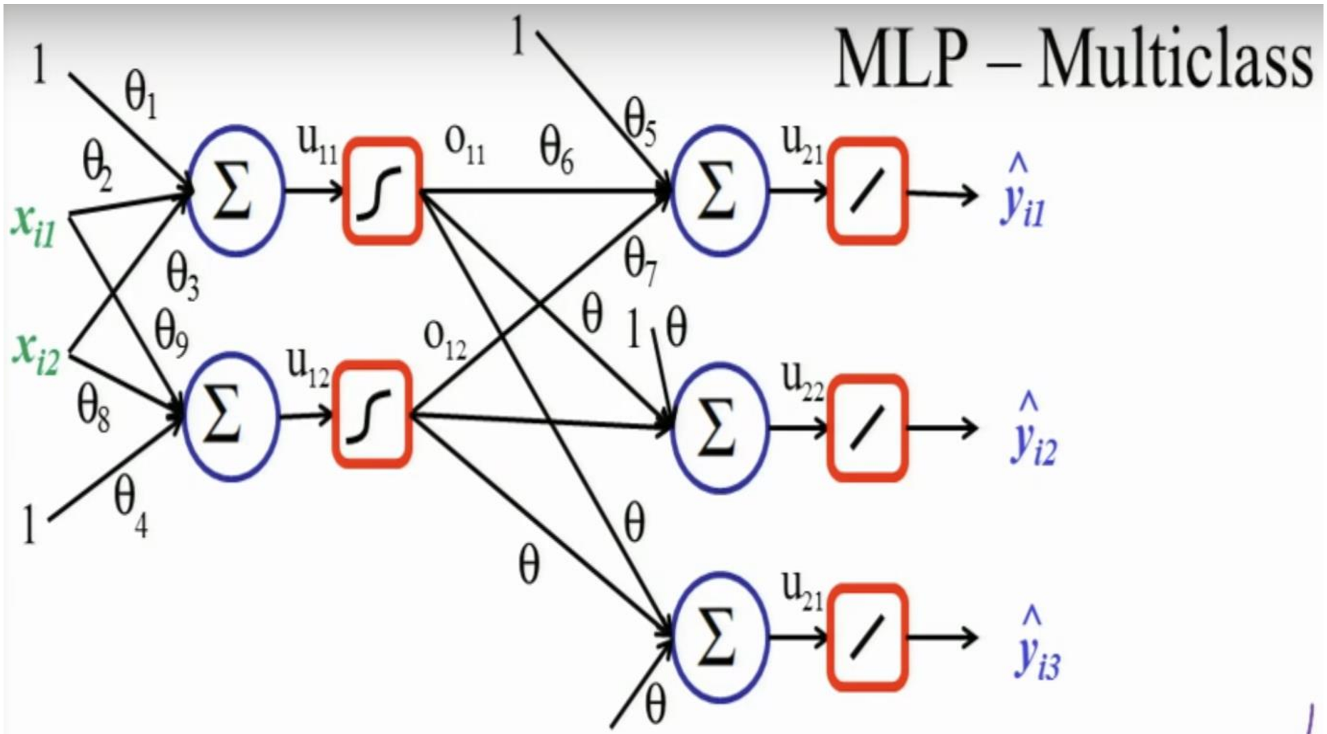
x_i	y_i
0.2	0.6
0.9	0.4
-0.6	5
-0.3	6.2

$$\hat{y}_i = \mathbf{E}(y_i \mid x_i, \theta) = u_{21}$$

$$\hat{y}_i = \Theta_5 + \frac{\Theta_6}{1+e^{-\Theta_1 - \Theta_2 x_i}} + \frac{\Theta_7}{1+e^{-\Theta_4 - \Theta_3 x_i}}$$

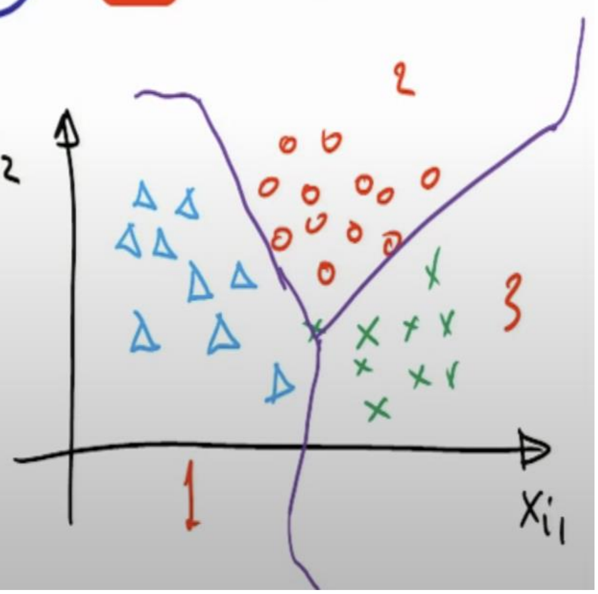
Theta five shifts the curve up and down. Theta six controls the height of the S-shaped curve; if theta six is large, the curve is tall, and vice versa. If theta six is negative, it flips the curve. Theta one shifts the curve left and right, while theta two controls its width, making it either wider or thinner.

MLP – Multiclass

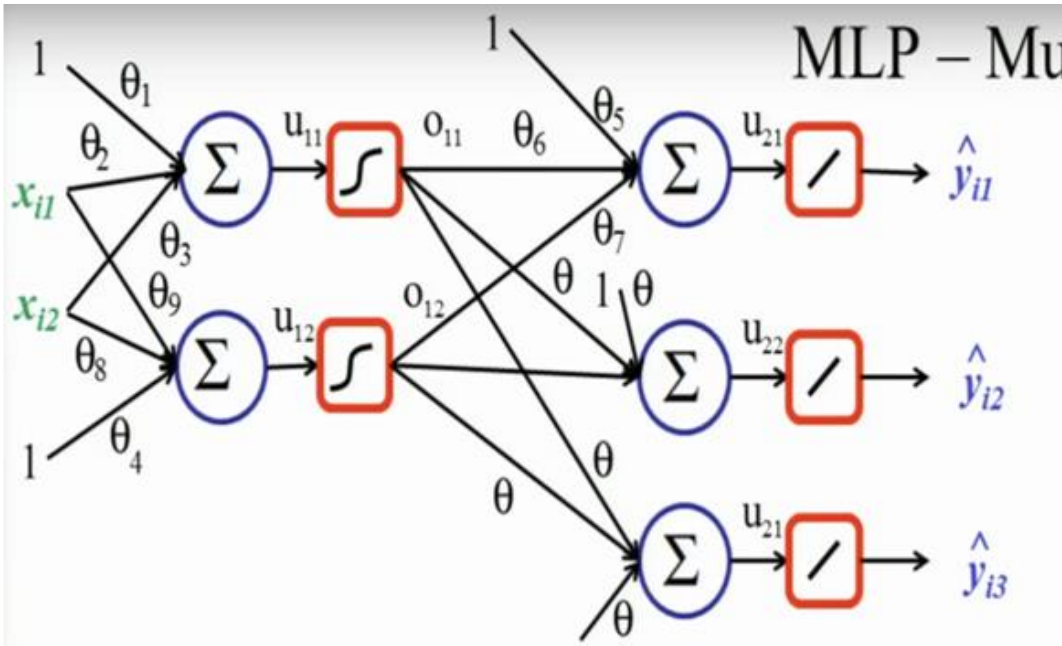


Data :

x_{i1}	x_{i2}	y_{i1}	y_{i2}	y_{i3}	
0.2	0.3	0	1	0	Class 2
-5	-6	1	0	0	Class 1
-20	4	1	0	0	" 1
42	6.8	0	0	1	Class 3



MLP – Multiclass



To get a probabilistic model, define: SOFTMAX

$$P(y_i = (0\,1\,0) \mid x_i, \theta) = P(y_i = 2 \mid x_i, \theta) = \frac{e^{\hat{y}_2}}{e^{\hat{y}_1} + e^{\hat{y}_2} + e^{\hat{y}_3}}$$

MLP – Multiclass

$$\mathbb{I}_2(y_i) = \begin{cases} 1 & y_i = 2 \\ 0 & \text{o.w.} \end{cases}$$

Then,

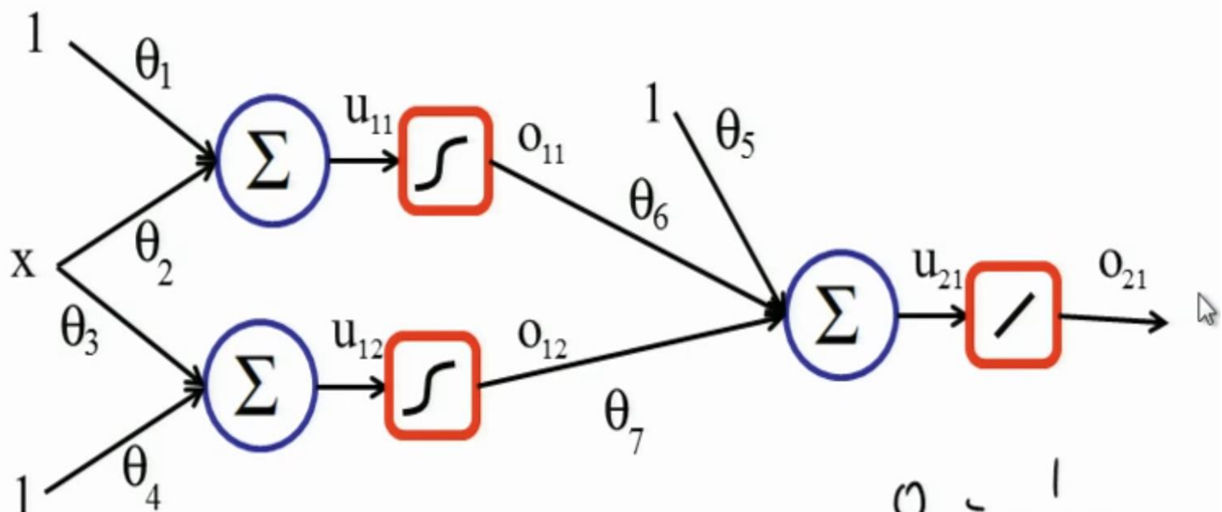
$$P(y_i \mid x_i, \theta) = \left[\frac{e^{\hat{y}_{i1}}}{\underbrace{e^{\hat{y}_{i1}} + e^{\hat{y}_{i2}} + e^{\hat{y}_{i3}}}_{\text{sum}}}\right]^{\mathbb{I}_1(y_i)} \left[\frac{e^{\hat{y}_{i2}}}{e^{\hat{y}_{i1}} + e^{\hat{y}_{i2}} + e^{\hat{y}_{i3}}}\right]^{\mathbb{I}_2(y_i)} \left[\frac{e^{\hat{y}_{i3}}}{e^{\hat{y}_{i1}} + e^{\hat{y}_{i2}} + e^{\hat{y}_{i3}}}\right]^{\mathbb{I}_3(y_i)}$$

$$= \begin{cases} e^{\hat{y}_{i1}}/\text{sum} & y_i = 1 \\ e^{\hat{y}_{i2}}/\text{sum} & y_i = 2 \\ e^{\hat{y}_{i3}}/\text{sum} & y_i = 3 \end{cases}$$

$\underline{\underline{\text{Cost}}}$:

$$C(\theta) = -\log P(\underset{\sim}{y} \mid \underset{\sim}{X}, \underset{\sim}{\theta}) = -\sum_{i=1}^{n} \sum_{j=1}^{3} \mathbb{I}_j(y_i) \log \frac{e^{\hat{y}_{ij}}}{\text{sum}}$$

Backpropagation



$$o_{11} = \frac{1}{1+e^{-u_{11}}}$$

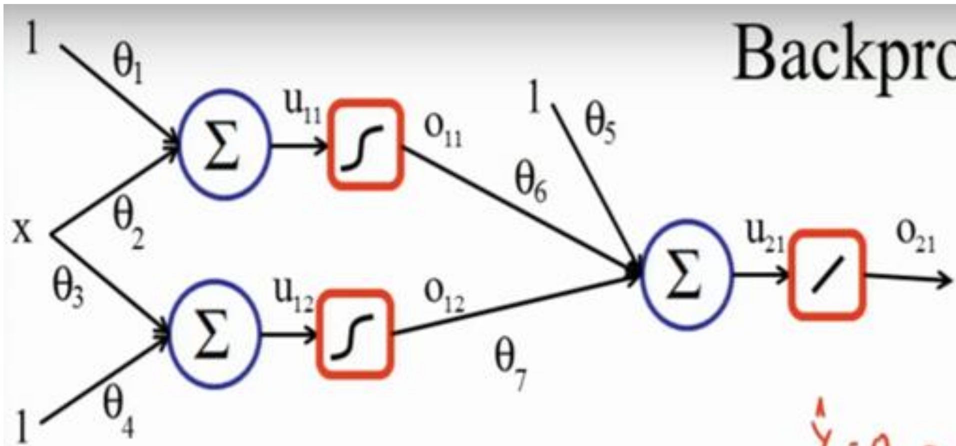
$$\hat{y} = o_{21} = u_{21} = \theta_5 + \theta_6 o_{11} + \theta_7 o_{12}$$

$$o_{12} = \frac{1}{1+e^{-u_{12}}}$$

$$u_{11} = \theta_1 + \theta_2 x$$

$$u_{12} = \theta_4 + \theta_3 x$$

Backpropagation



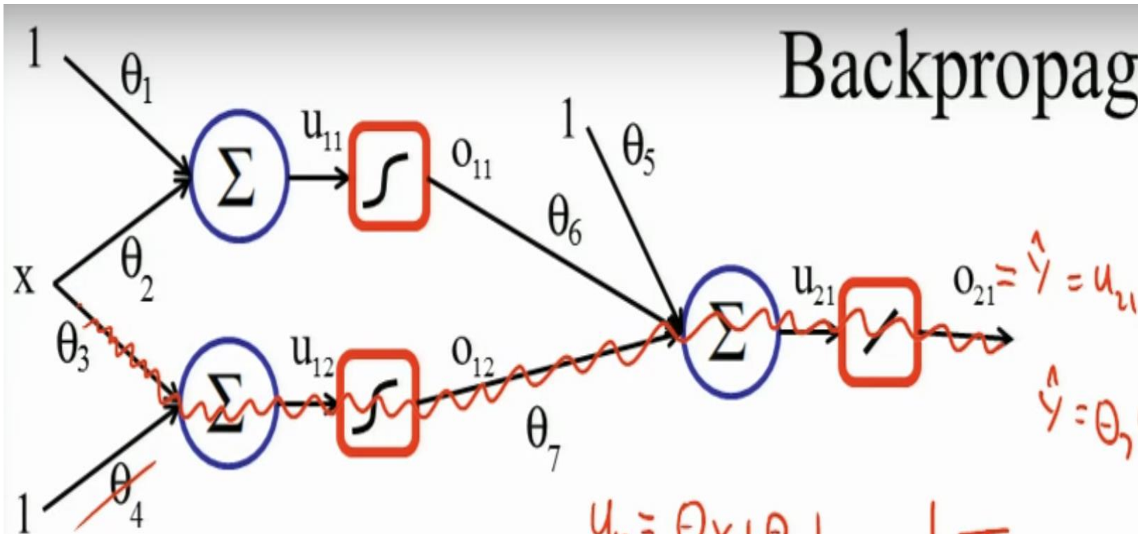
$$\hat{y} = o_{21} = \theta_5 + \theta_6 o_{11} + \theta_7 o_{12}$$

$$E(\theta) = \left(y_i - \hat{y}_i(x_i,\theta)\right)^2$$

$$\frac{\partial E(\theta)}{\partial \theta_j} = -2\left(y_i - \hat{y}_i(x_i,\theta)\right)\frac{\partial \hat{y}_i(x_i,\theta)}{\partial \theta_j}$$

$$\frac{\partial \hat{y}_i}{\partial \theta_5} = 1 \qquad \frac{\partial \hat{y}_i}{\partial \theta_6} = o_{11} \qquad \frac{\partial \hat{y}_i}{\partial \theta_7} = o_{12}$$

Backpropagation



$$\hat{y} = \theta_7 O_{12} + \theta_6 O_{11} + \theta_5$$

$$u_{12} = \theta_3 x + \theta_4 1$$

$$\frac{\partial \hat{y}}{\partial \theta_3} = \frac{\partial \hat{y}}{\partial O_{12}} \frac{\partial O_{12}}{\partial u_{12}} \frac{\partial u_{12}}{\partial \theta_3}$$

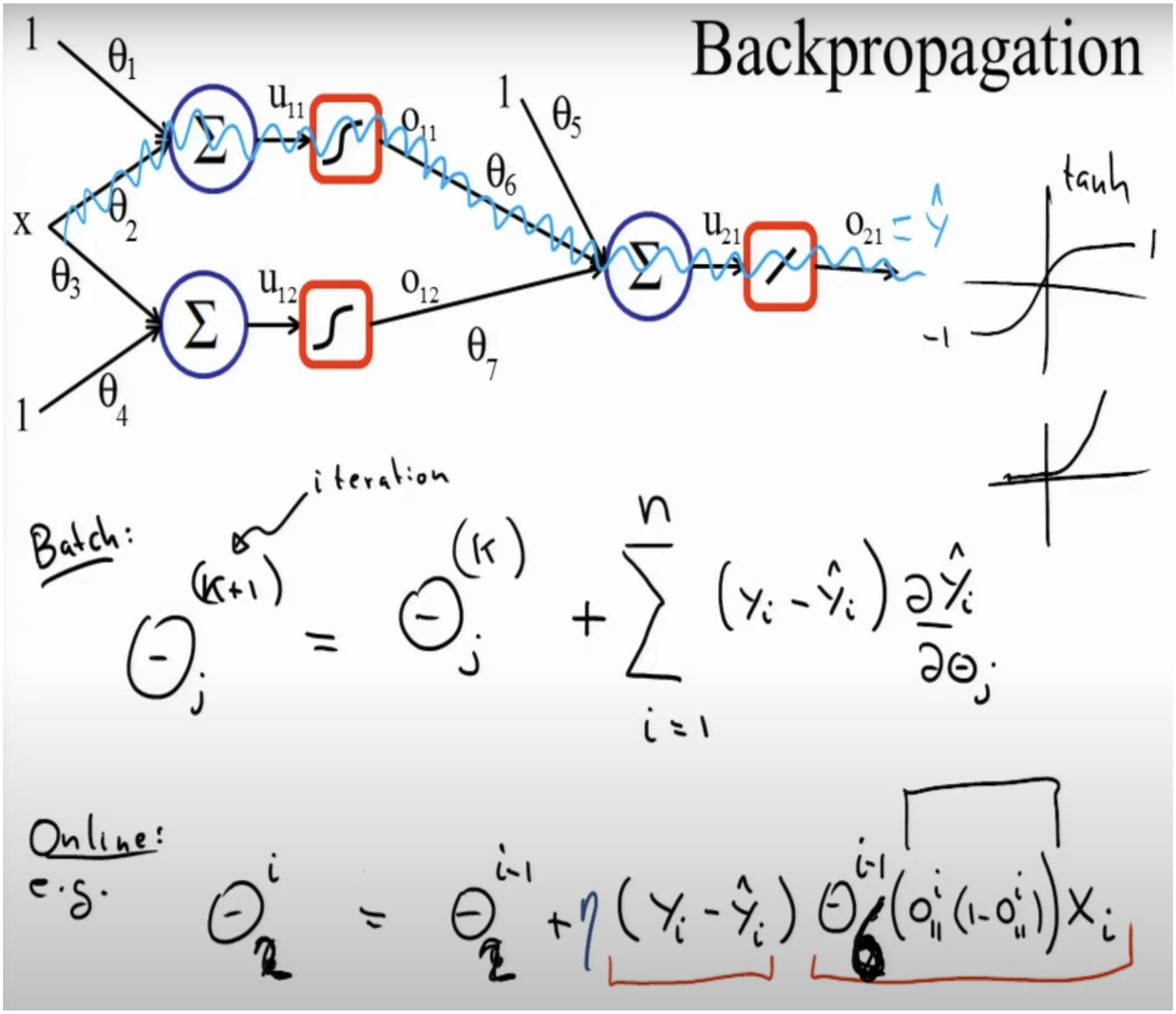
$$= \theta_7 \underbrace{O_{12}\left[1 - O_{12}\right] x}_{W}$$

Thm

$$\frac{\partial O_{12}}{\partial u_{12}} = O_{12}(1 - O_{12})$$

i.e.

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



Hyperbolic tangent function or rectified unit to deal with vanishing gradients as the network gets deeper.