Logistic Regression $\Rightarrow$ $P(y|x) = \cancel{\dfrac{P(y)\,P(x|y)}{P(x)}} = \dfrac{1}{1+\exp(-\underset{\sim}{X\theta})}$ $[0, 1]$

Posterior

$$\text{Odds} = \frac{p}{1-p} \in [0, +\infty) \qquad X\theta \in \mathbb{R} \quad (-\infty, +\infty)$$

$$\frac{p}{1-p} \neq X\theta \Rightarrow \log\left(\frac{p}{1-p}\right) \in (-\infty, +\infty)$$

$$\log\left(\frac{p}{1-p}\right) = X\theta \Rightarrow \exp\left(\log\left(\frac{p}{1-p}\right)\right) = \exp(X\theta) \Rightarrow p = \frac{1}{1+\exp(-X\theta)}$$

$$P(y|x) = \frac{1}{1+\exp(-X\theta)}$$

① ②
σ; Σ

$N(x|\mu,\sigma)$

Posterior

Generative models: NB $\Rightarrow$ $P(y|x) = \dfrac{p(x|y)\,p(y)}{p(x)}$

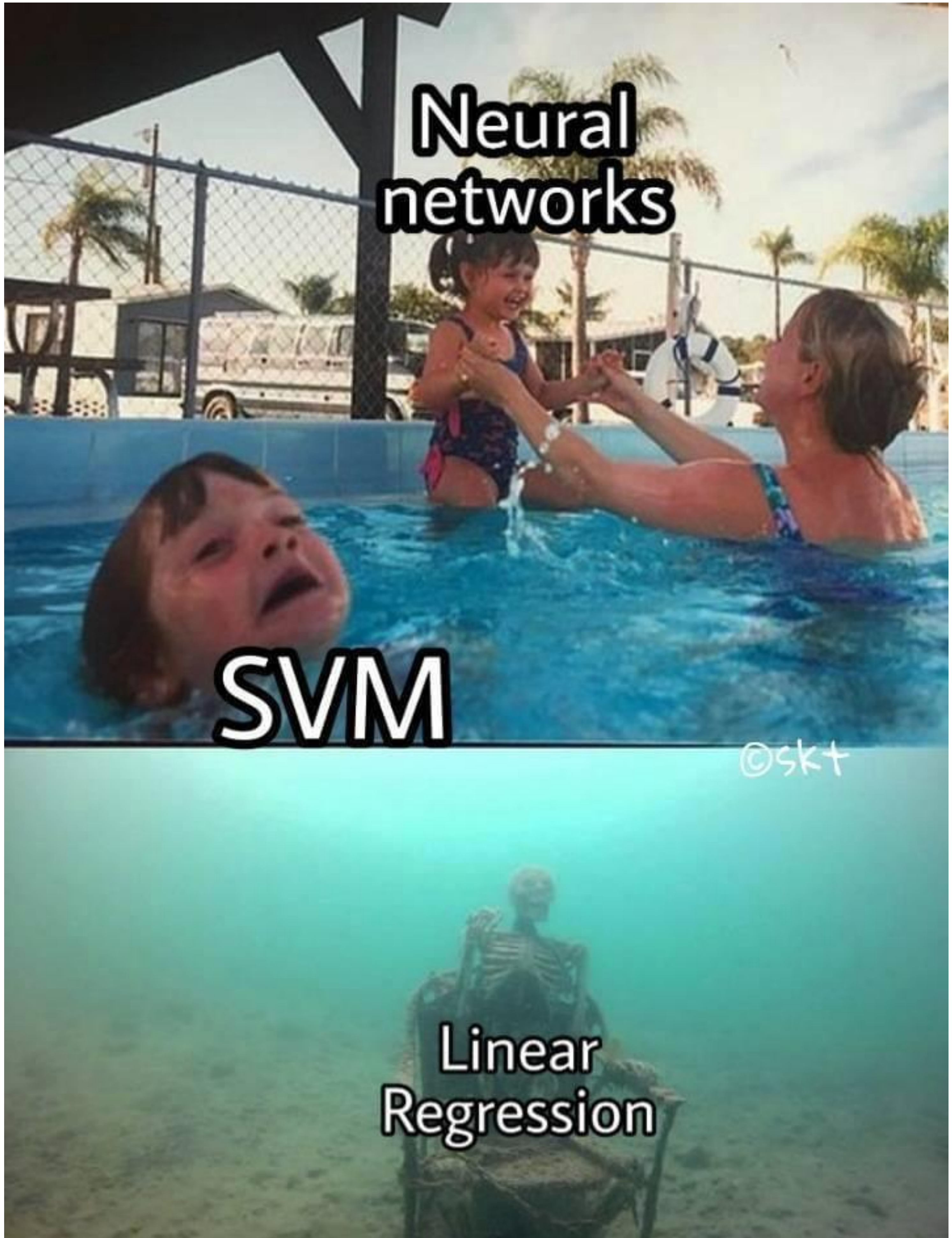
Discriminative models: Logistic regression $P(y|x) \overset{?}{=} \; = \dfrac{1}{1+\exp(-x\theta)}$

linear combination of features

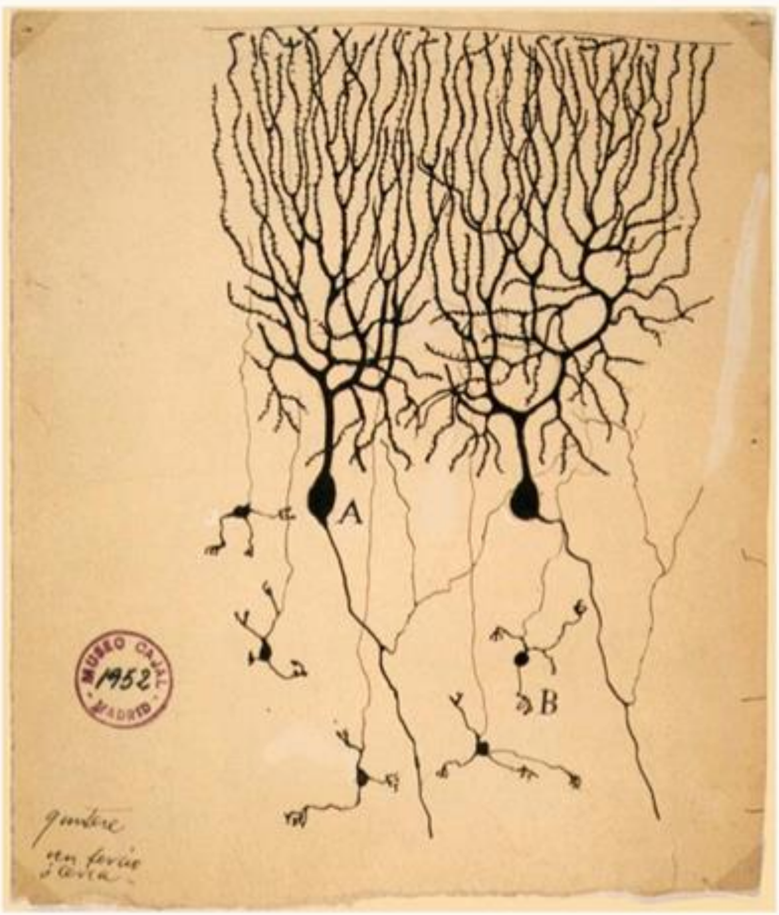
Neural Networks
Forward Pass and Back Propagation

Mahdi Roozbahani

Georgia Tech


Neural networks
SVM
©skt
Linear Regression

Inspiration from Biological Neurons



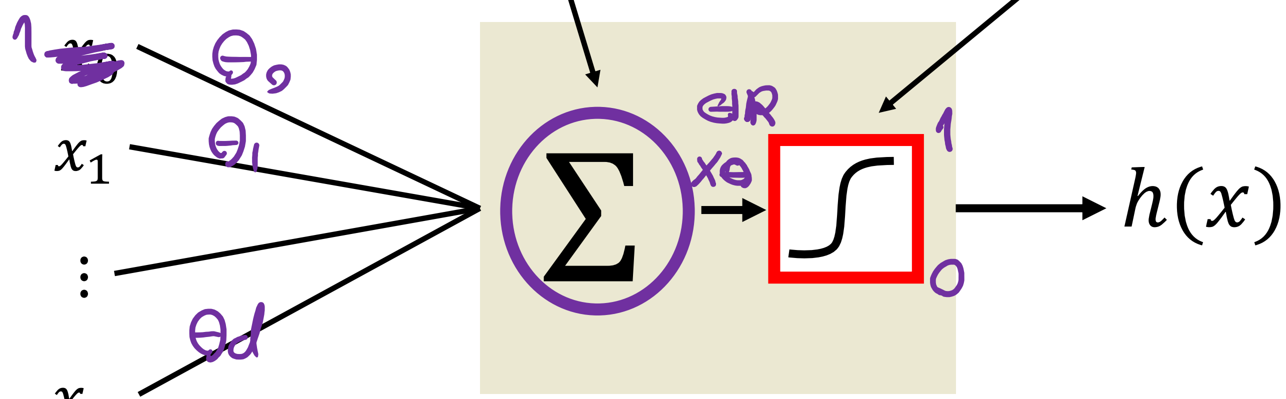
The first drawing of a brain cells by Santiago Ramón y Cajal in 1899

Neurons: core components of brain and the nervous system consisting of

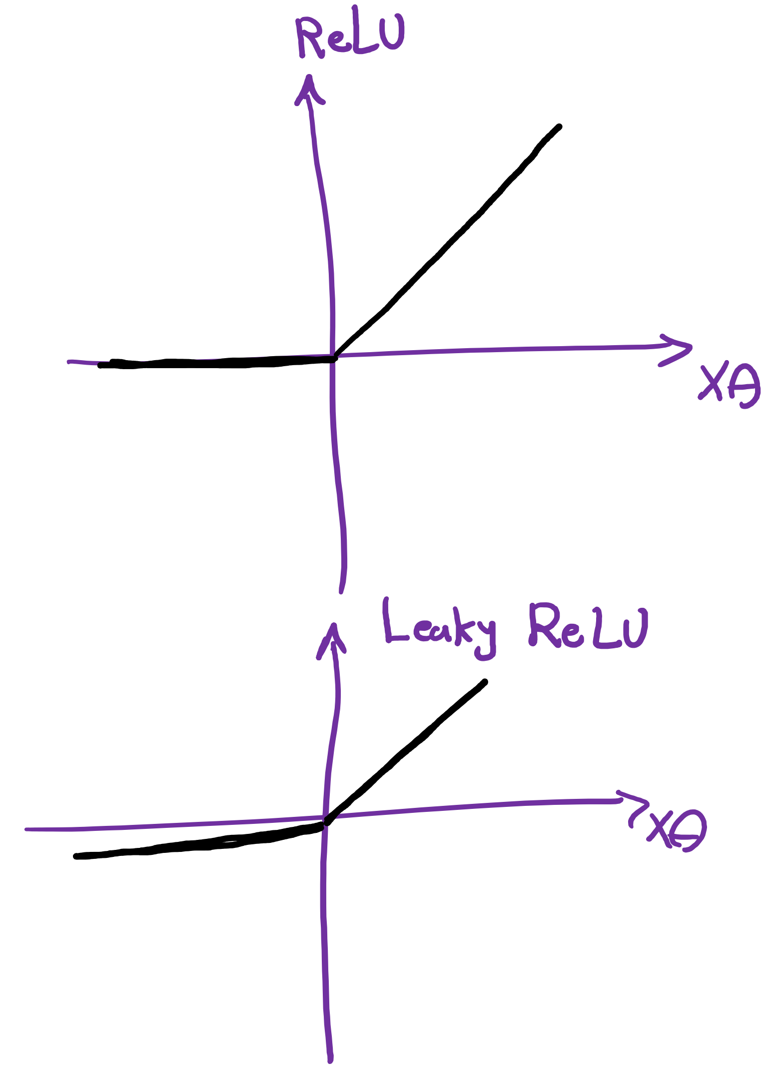
1. Dendrites that collect information from other neurons
2. An axon that generates outgoing spikes

A learning block

$\Sigma = \theta_0 + \theta_1 x_1 + \cdots = x\theta$



$h(x) = u(x\theta) \rightarrow$ logistic regression

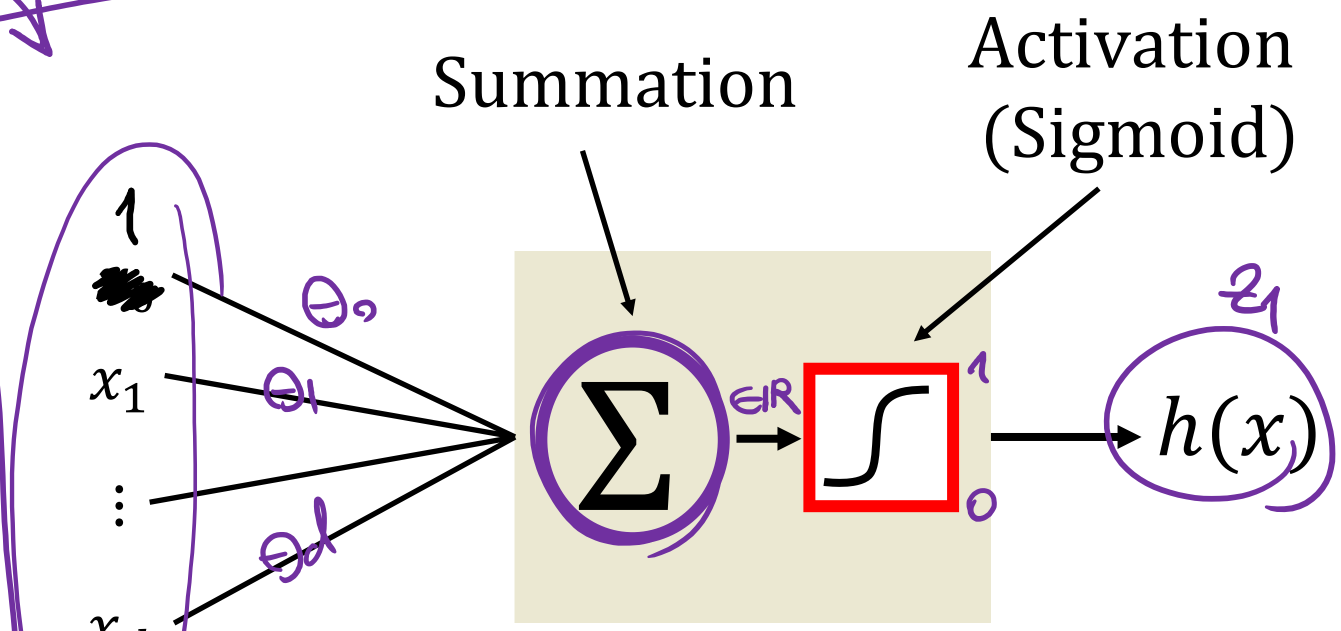


$$output = activation(x\theta + b)$$

Name of the neuron	Activation function: $activation(z)$
Linear unit *linear*	z
Threshold/sign unit *non-linear*	$\mathrm{sgn}(z)$
Sigmoid unit *non-linear*	$\dfrac{1}{1+\exp(-z)}$
Rectified linear unit (ReLU) *non-linear*	$\max(0, z)$
Tanh unit *non-linear*	$\tanh(z)$

$+\infty$, $-\infty$; $+1$, -1; 1, 0; $+1$, -1

$\rightarrow$ a learning block



$h(x) = u(x\theta) \rightarrow$ logistic regression

$$X = \begin{array}{c} x_1 \ \dots \ x_d \\ \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{n \times d} \end{array}$$

$\Sigma = X\theta \in \mathbb{R}$

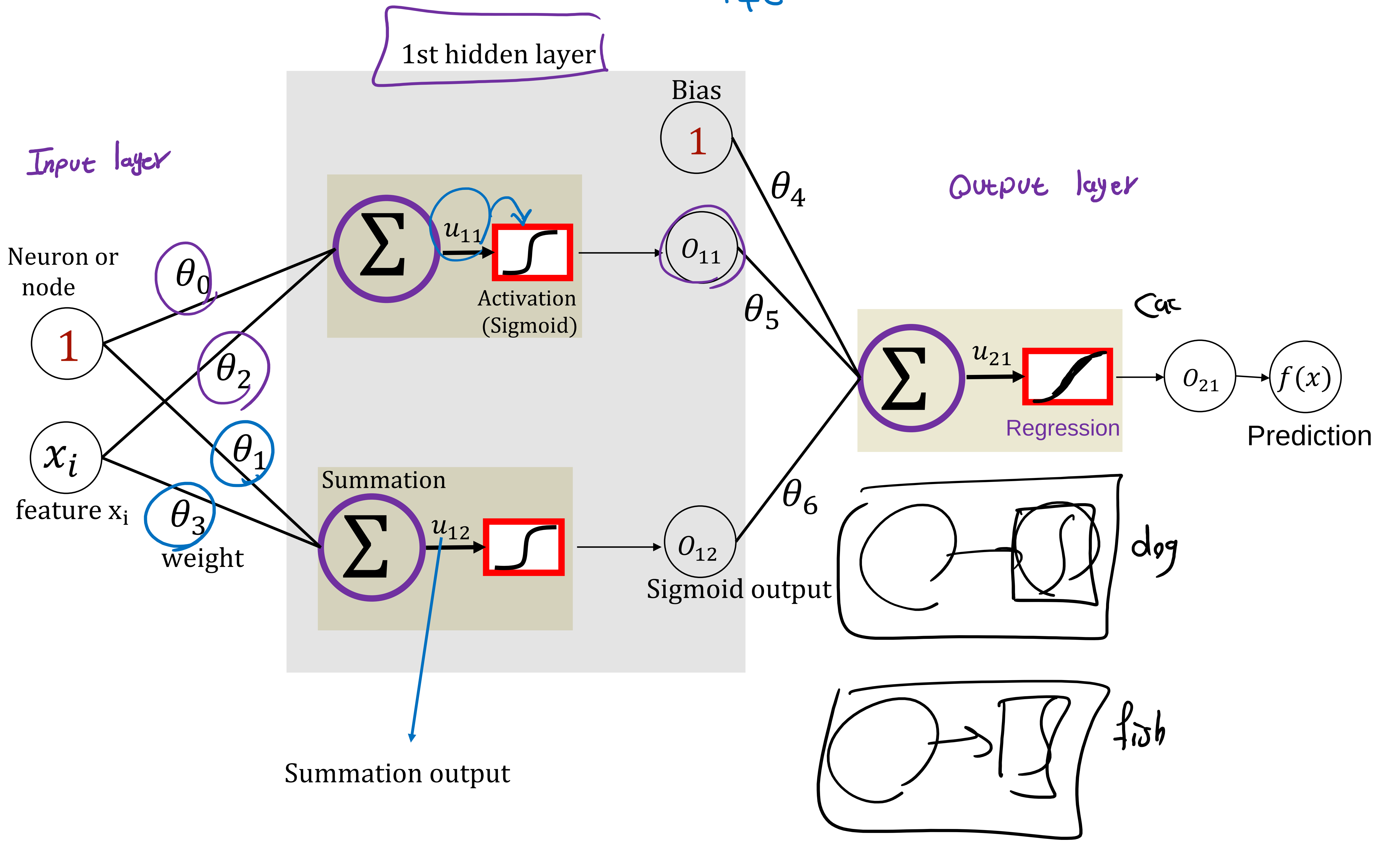
NN Regression

$\Sigma = u_{11} = X\theta$

$S = \dfrac{1}{1+e^{-u_{11}}}$

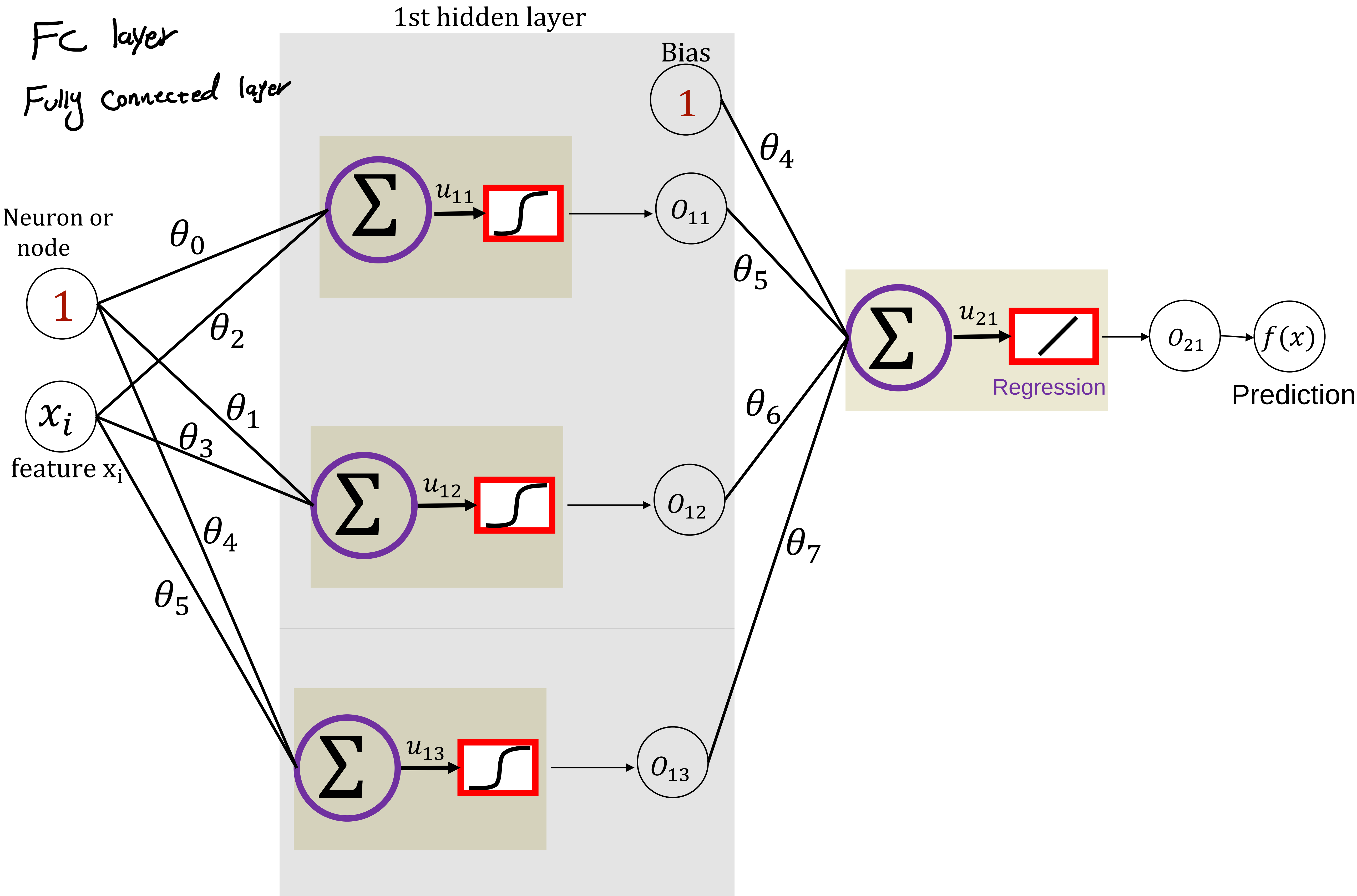
u_{21} O_{21}

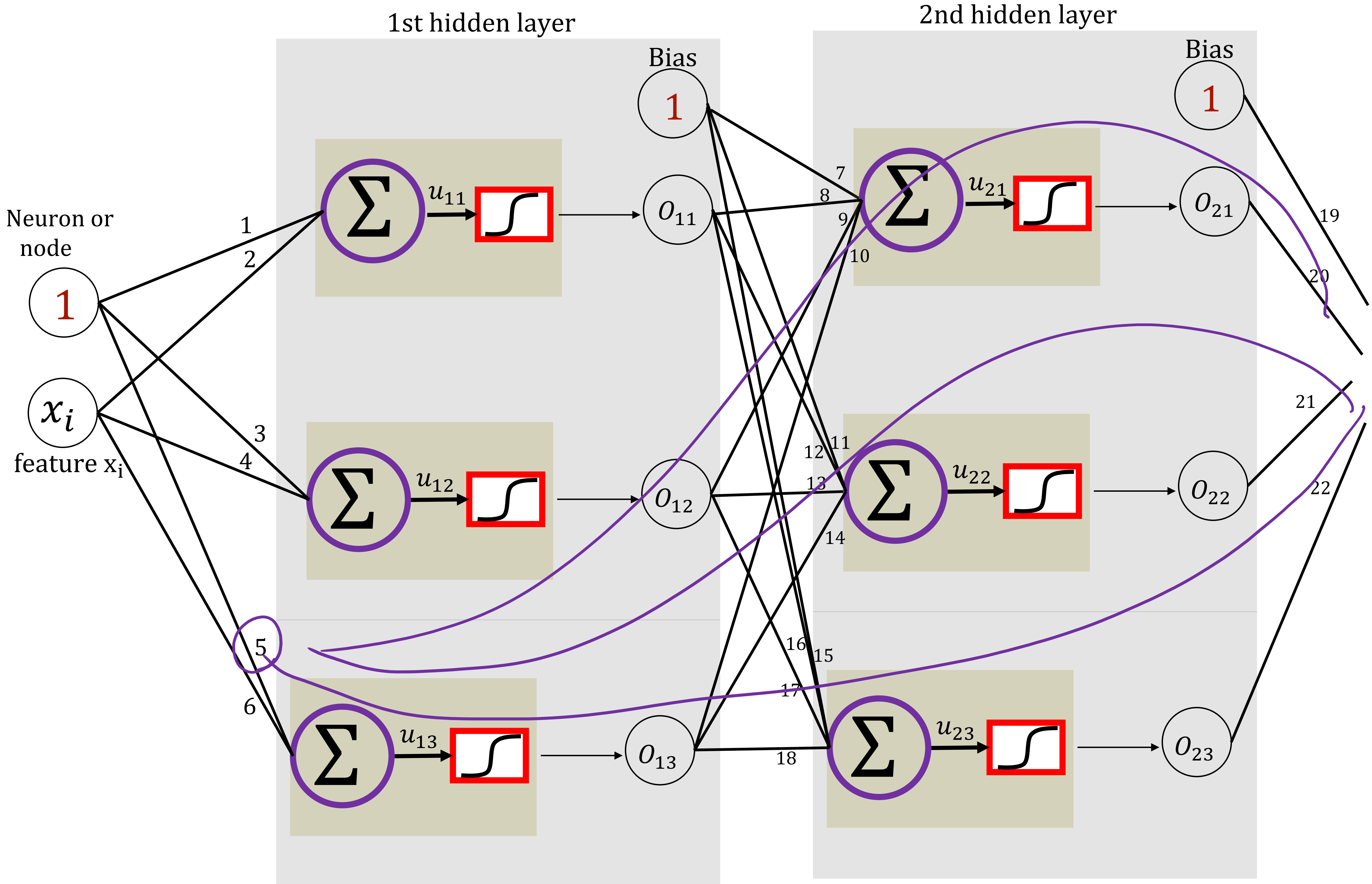
layer depth



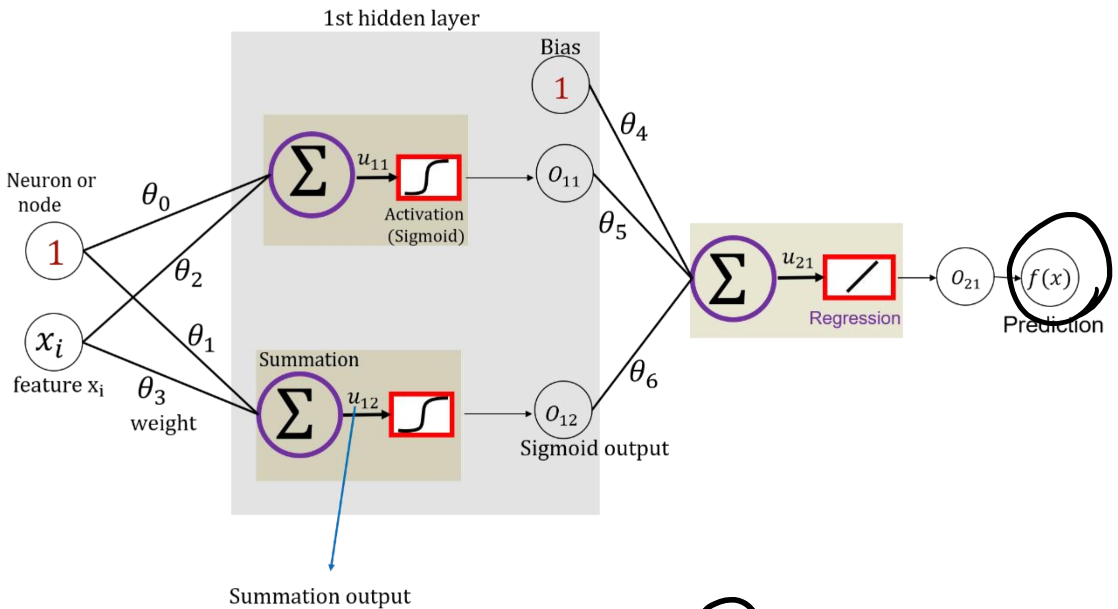
Fc layer

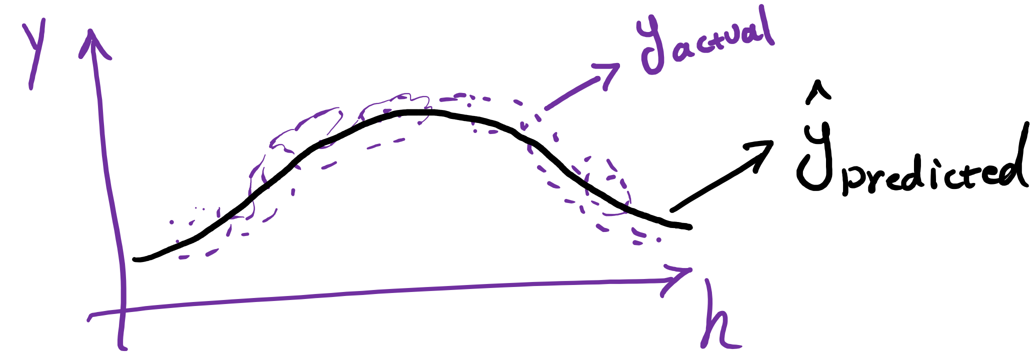
Fully connected layer





1st hidden layer
2nd hidden layer
Bias
Bias
Neuron or node
feature x_i
x_i
1
1
1
u_{11}
u_{12}
u_{13}
u_{21}
u_{22}
u_{23}
o_{11}
o_{12}
o_{13}
o_{21}
o_{22}
o_{23}
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22



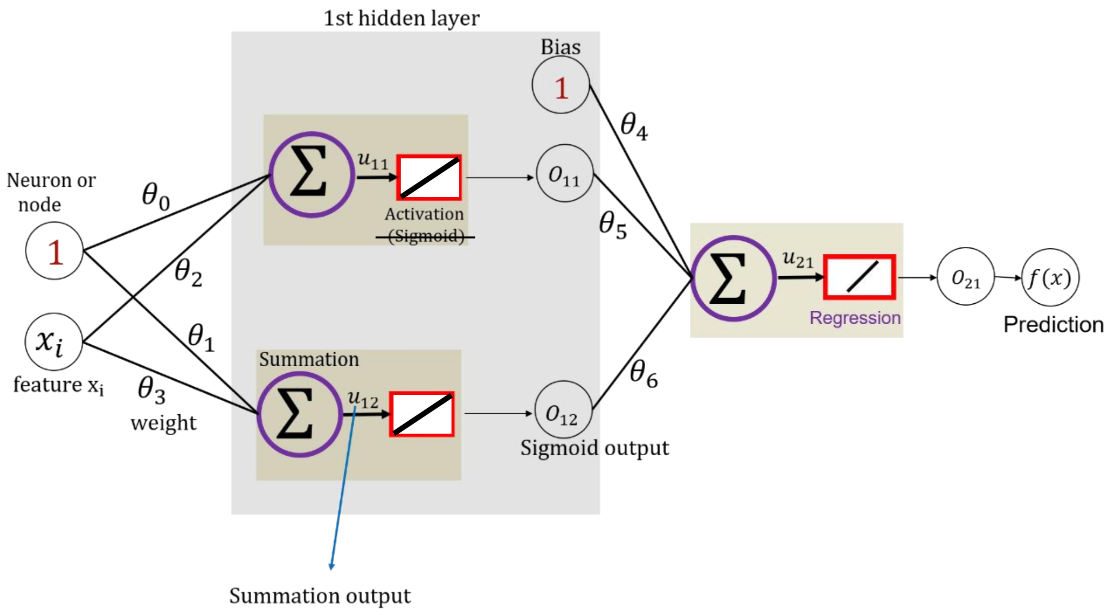


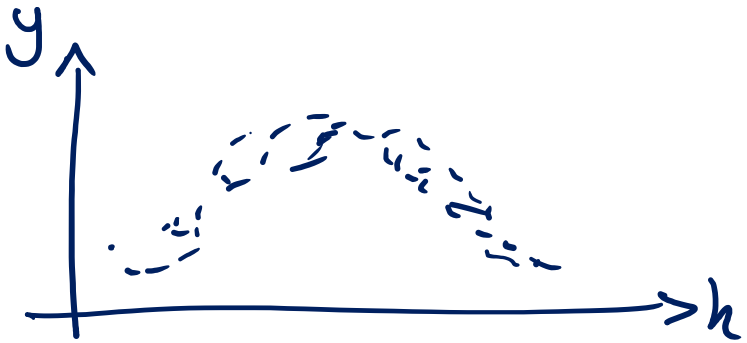
1. Initialize all the parameters $\Theta s \in \{\Theta_0, \dots, \Theta_6\}$
   ↳ Do not use zero

2. Forward Pass ⟶ U_s, O_s

3. Back-propagation ⟶ Optimize parameters or Θ_s

4. Check for convergence





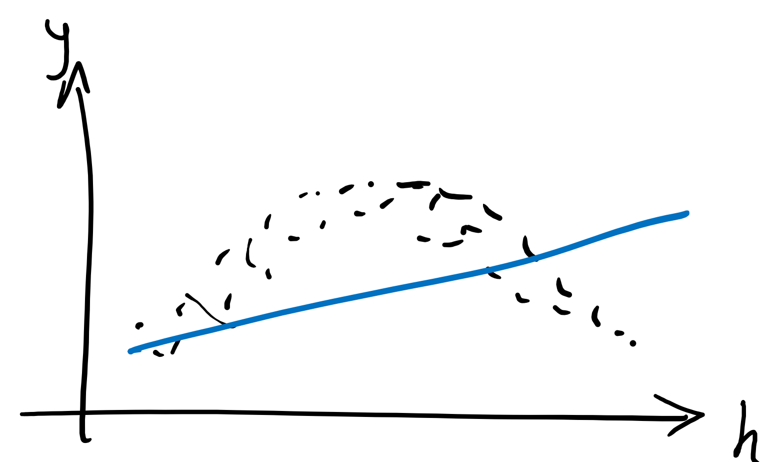
$$U_{11} = \Theta_0 + \Theta_2 x_i \Rightarrow O_{11} = U_{11} = \Theta_0 + \Theta_2 x_i$$

$$U_{12} = \Theta_1 + \Theta_3 x_i \Rightarrow O_{12} = U_{12} = \Theta_1 + \Theta_3 x_i$$

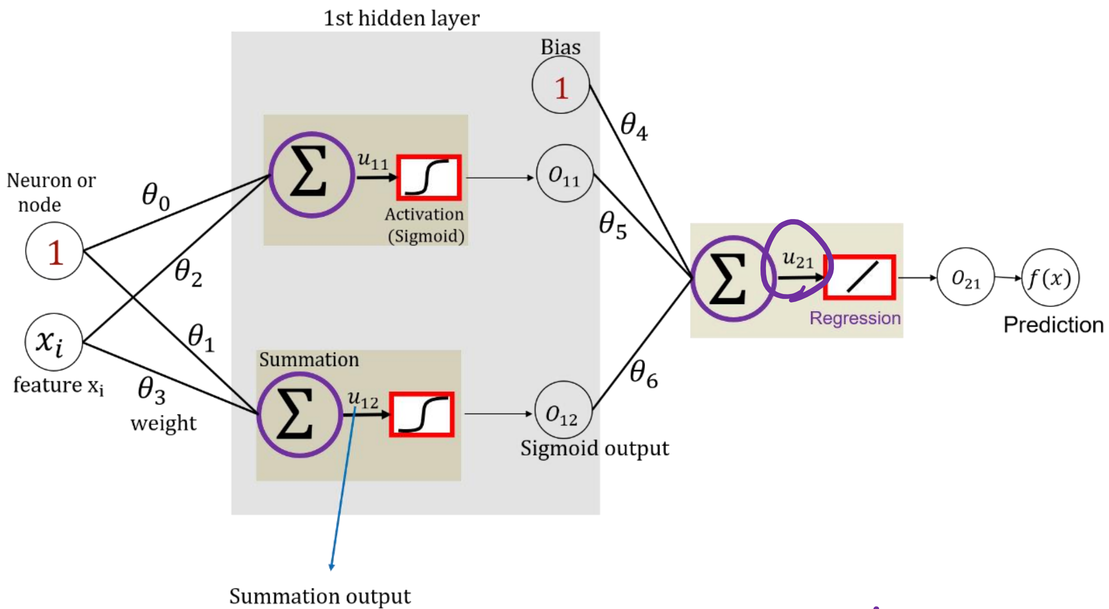
$$U_{21} = \Theta_4 + \Theta_5 O_{11} + \Theta_6 O_{12} = O_{21} = f(x)$$

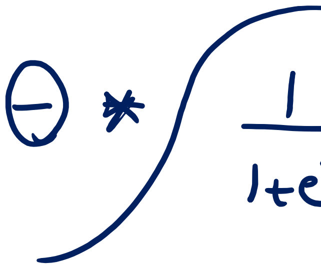
$$f(x) = \Theta_4 + \Theta_5 \Theta_0 + \Theta_5 \Theta_2 x_i + \Theta_6 \Theta_1 + \Theta_6 \Theta_3 x_i$$

$$f(x) = \underbrace{\Theta_4 + \Theta_5 \Theta_0 + \Theta_6 \Theta_1}_{\Theta_0} + \underbrace{(\Theta_5 \Theta_2 + \Theta_6 \Theta_3)}_{\Theta_1} x_i$$



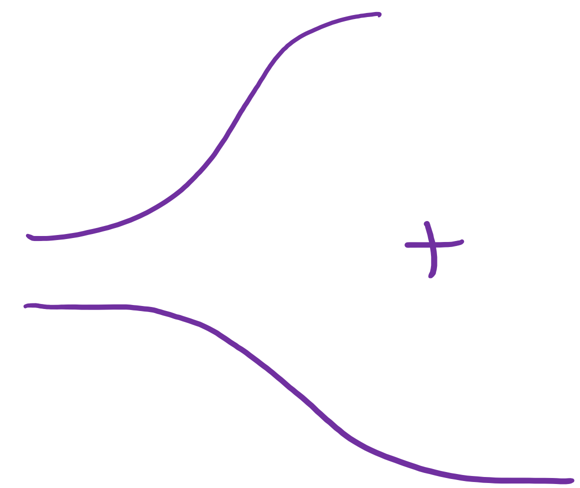
$$f(x) = \Theta_0 + \Theta_1 x_i$$



$\theta *$  $\frac{1}{1+e^{-u}}$

$$u_{11} = \theta_0 + \theta_2 x_i \Rightarrow O_{11} = \frac{1}{1+e^{-u_{11}}} = \frac{1}{1+e^{-(\theta_0+\theta_2 x_i)}}$$

$$u_{12} = \theta_1 + \theta_3 x_i \Rightarrow O_{12} = \frac{1}{1+e^{-u_{12}}} = \frac{1}{1+e^{-(\theta_1+\theta_3 x_i)}}$$



$$u_{21} = \theta_4 + \theta_5 O_{11} + \theta_6 O_{12} \Rightarrow O_{21} = u_{21} = f(x)$$

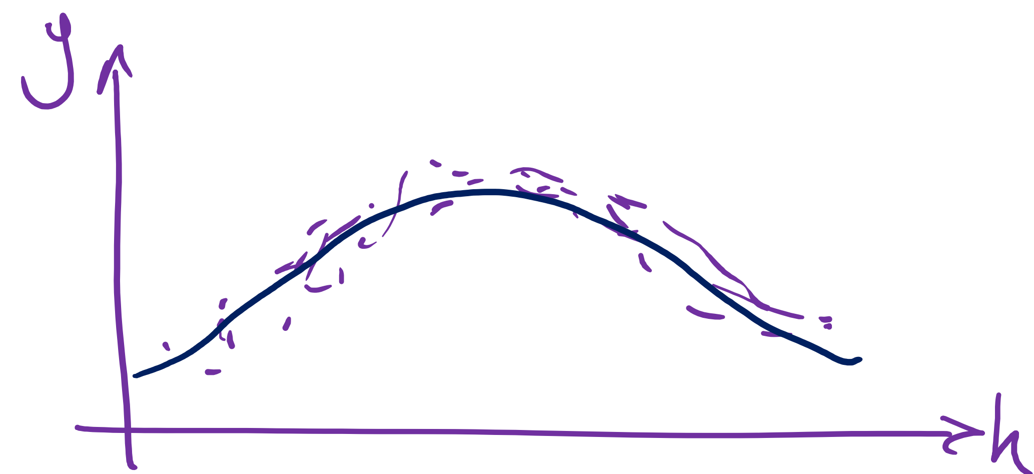
$$f(x) = \theta_4 + \frac{\theta_5}{1+e^{-(\theta_0+\theta_2 x_i)}} + \frac{\theta_6}{1+e^{-(\theta_1+\theta_3 x_i)}}$$

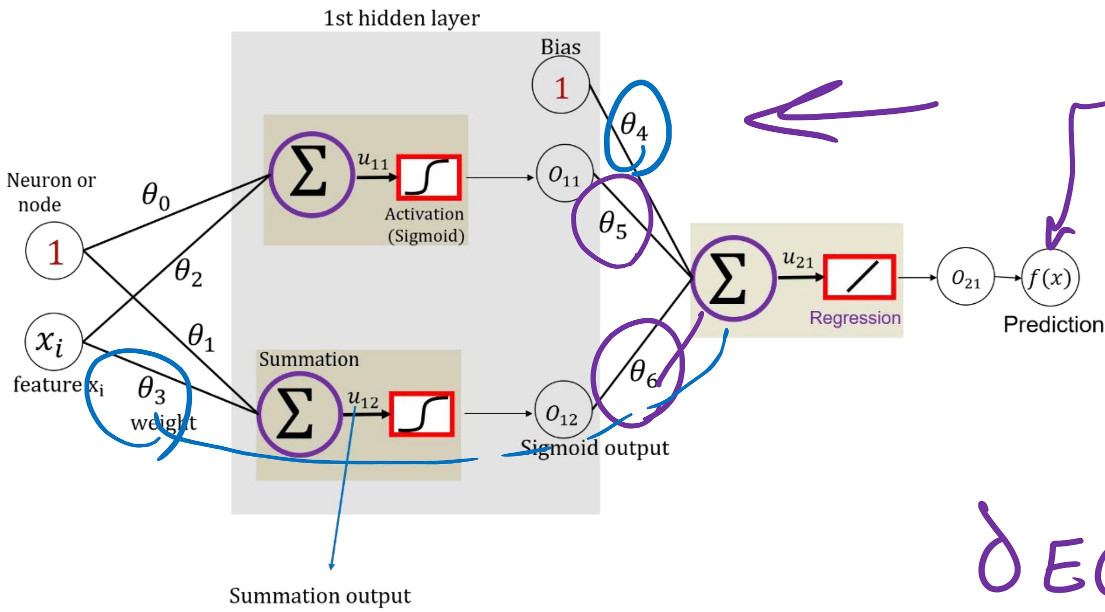
$\theta_5 \rightarrow$ squash or stretch in y direction

θ_4: Translation in y direction

θ_0: Translation in x direction

θ_2: squash or stretch in x direction





Back propagation

$$GD \Rightarrow E(\Theta) = L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} (y_a - \hat{y}_p)^2$$

$$SGD \Rightarrow E(\Theta) = L(\Theta) = \frac{1}{2} (y_a - f(x))^2$$

$$\frac{\partial E(\Theta)}{\partial \Theta} = \underbrace{-(y_a - f(x))}_{\Delta} \frac{\partial f(x)}{\partial \Theta} = \Delta \frac{\partial f(x)}{\partial \Theta}$$

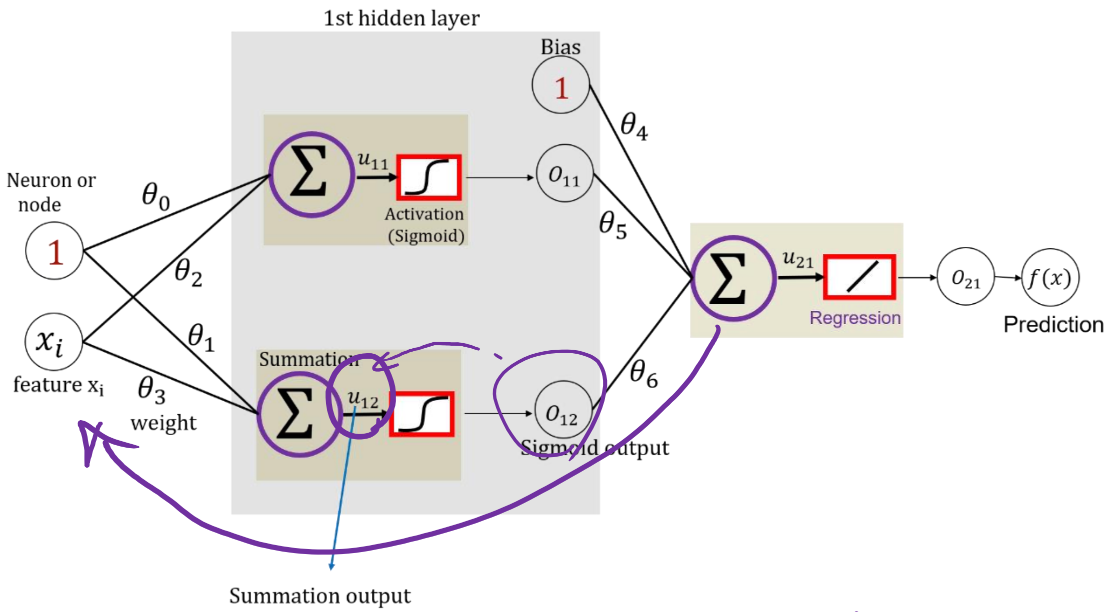
$\Theta_0, \Theta_1, \dots, \Theta_6$

$$f(x) = O_{21} = U_{21} = \Theta_4 + \Theta_5 O_{11} + \Theta_6 O_{12}$$

$$\Theta_6^{\{t+1\}} \leftarrow \Theta_6^{\{t\}} - \alpha \frac{\partial E(\Theta)}{\partial \Theta_6} = \Theta_6^{\{t\}} - \alpha \Delta \frac{\partial f(x)}{\partial \Theta_6} = \Theta_6^{\{t\}} - \alpha \Delta O_{12}$$

$$\Theta_5^{\{t+1\}} \leftarrow \Theta_5^{\{t\}} - \alpha \frac{\partial E(\Theta)}{\partial \Theta_5} = \Theta_5^{\{t\}} - \alpha \Delta \frac{\partial f(x)}{\partial \Theta_5} = \Theta_5^{\{t\}} - \alpha \Delta O_{11}$$

$$\Theta_4^{\{t+1\}} \leftarrow \Theta_4^{\{t\}} - \alpha \frac{\partial E(\Theta)}{\partial \Theta_4} = \Theta_4^{\{t\}} - \alpha \Delta \frac{\partial f(x)}{\partial \Theta_4} = \Theta_4^{\{t\}} - \alpha \Delta \times 1$$



$$f(x) = O_{21} = U_{21} = \theta_4 + \theta_5 O_{11} + \theta_6 O_{12}$$

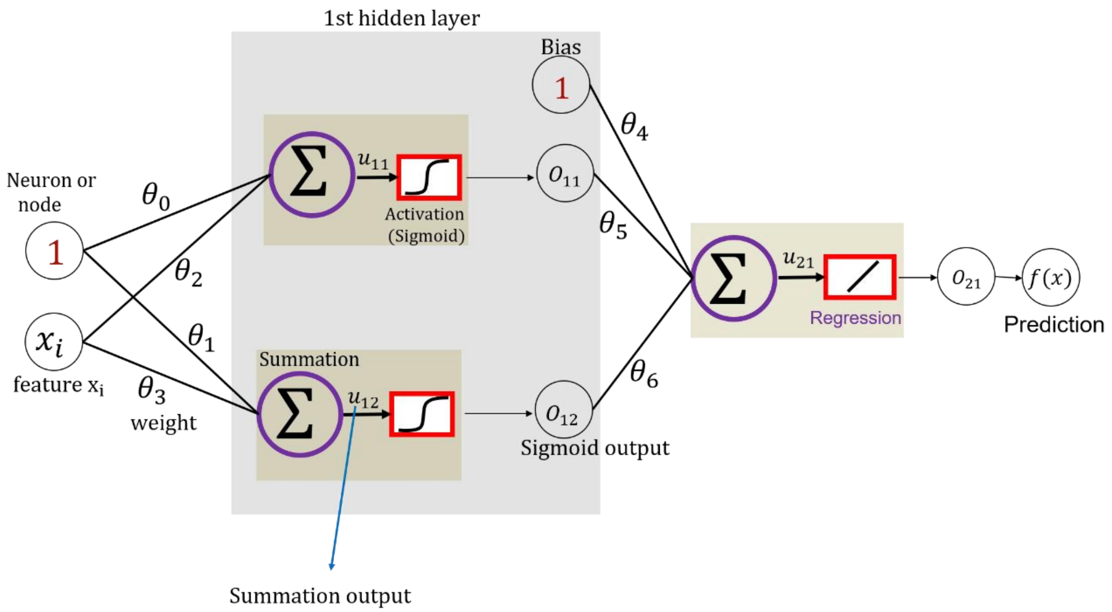
$$\nabla_{\theta} E(\theta) = \Delta \frac{\partial f(x)}{\partial \theta} \qquad U_{12} = \theta_1 + \theta_3 x_i$$

$$\frac{\partial f(x)}{\partial \theta_3} = \frac{\partial f(x)}{\partial O_{12}} \frac{\partial O_{12}}{\partial U_{12}} \frac{\partial U_{12}}{\partial \theta_3} = \theta_6 \; O_{12}[1 - O_{12}] \; x_i$$

$$O = \frac{1}{1+e^{-U}} = (1+e^{-U})^{-1} \Rightarrow \frac{\partial O}{\partial U} = -1 \times \frac{-e^{-U}}{(1+e^{-U})^2} = \frac{e^{-U}}{(1+e^{-U})^2} = O[1-O]$$

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

$$= O[1-O]$$



$$\Theta_3^{\{t+1\}} \leftarrow \Theta_3^{\{t\}} - \alpha \Delta \frac{\partial f(x)}{\partial \Theta_3}$$

$$\frac{\partial f(x)}{\partial \Theta_3} = \Theta_6 \, O_{12} [1 - O_{12}] \, X_i$$

Vanishing gradient