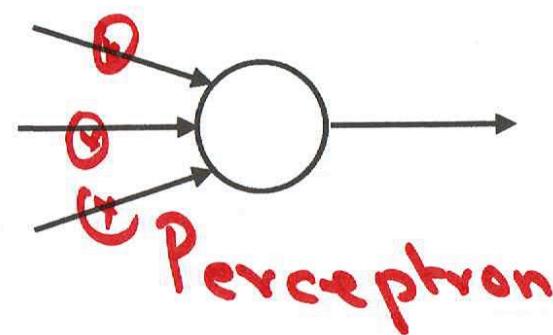
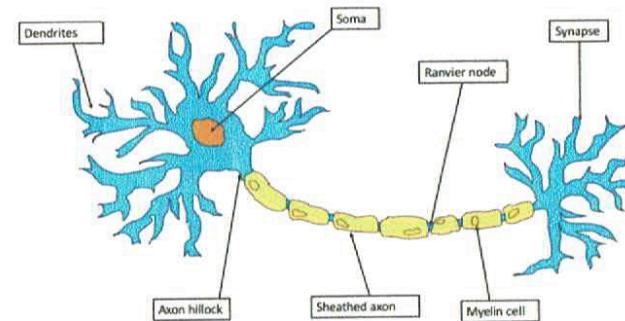
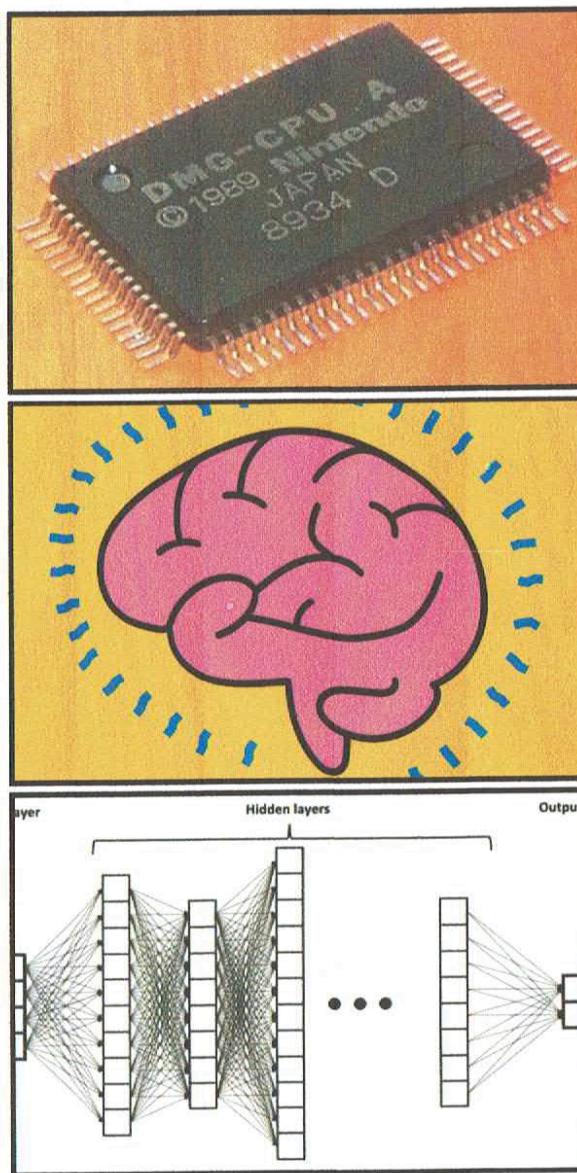


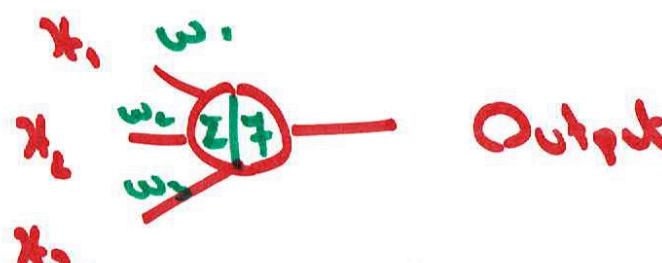
# Building towards a complex task!



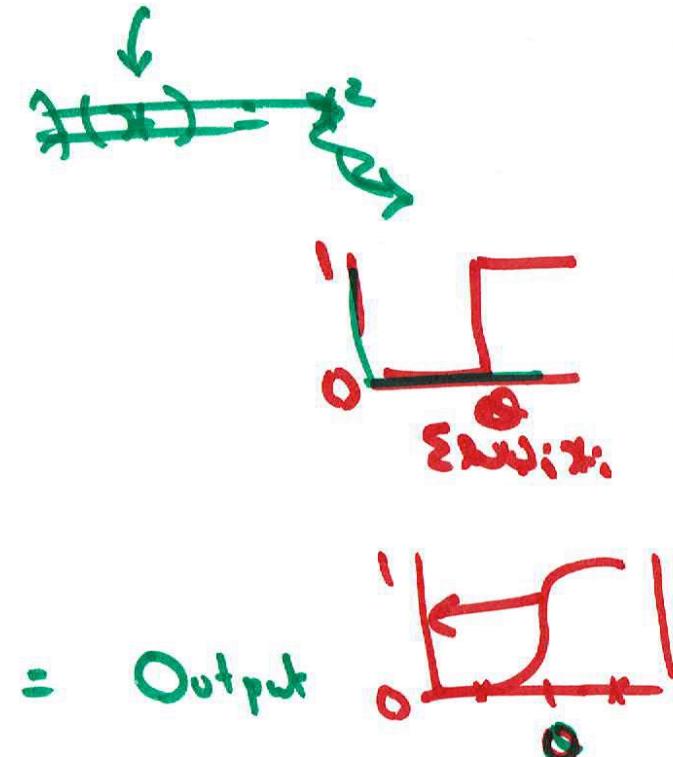
# Perceptron!

- What does it do?

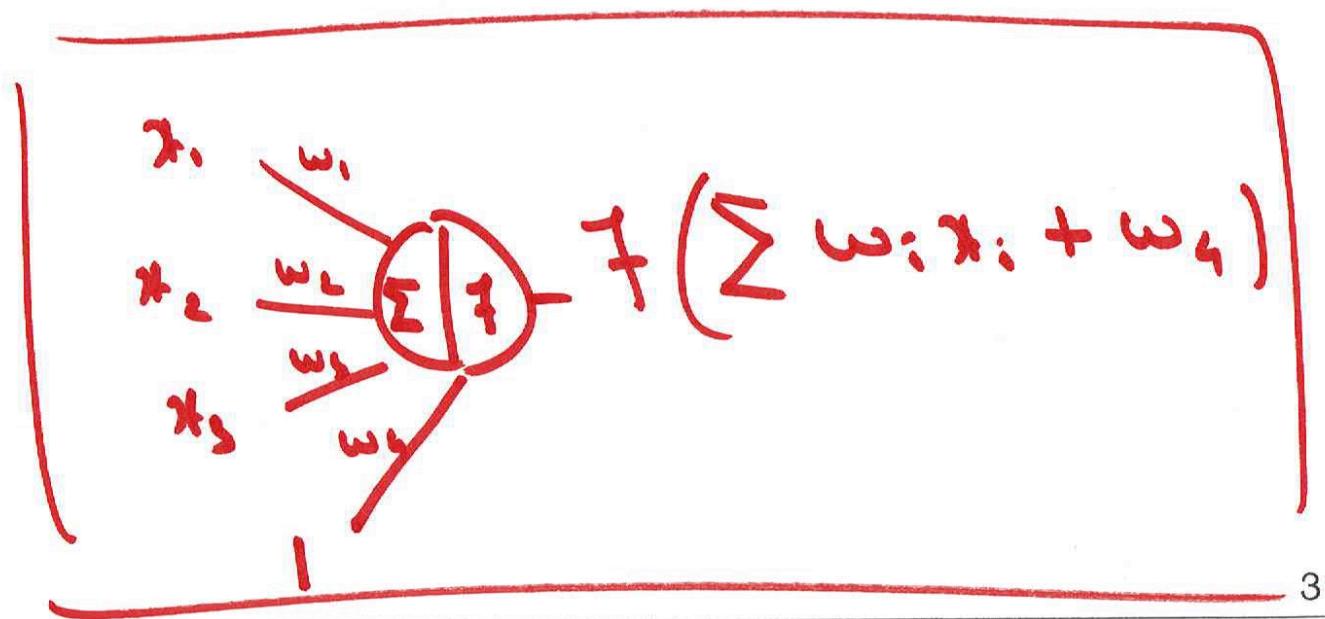
1958



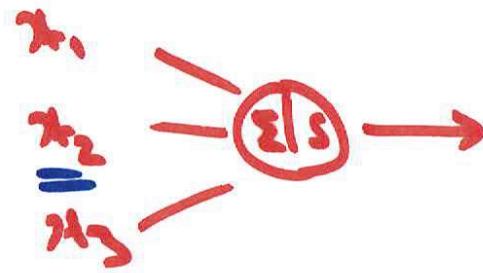
$$f(\sum w_i x_i) = \text{Output}$$



- Bias



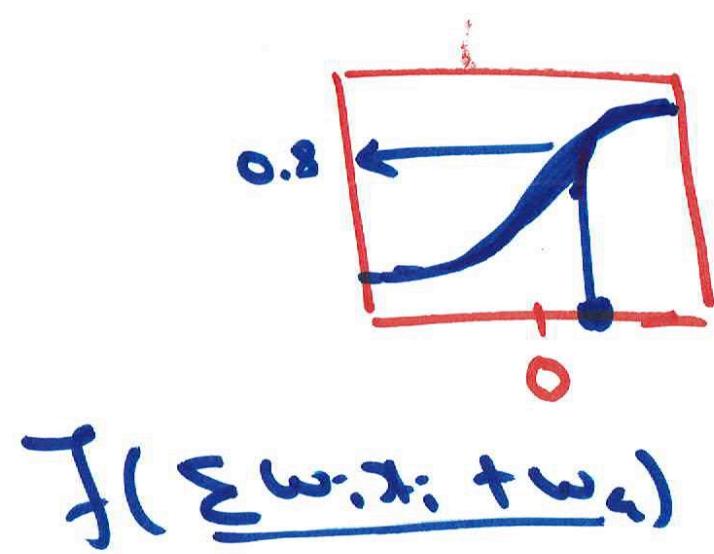
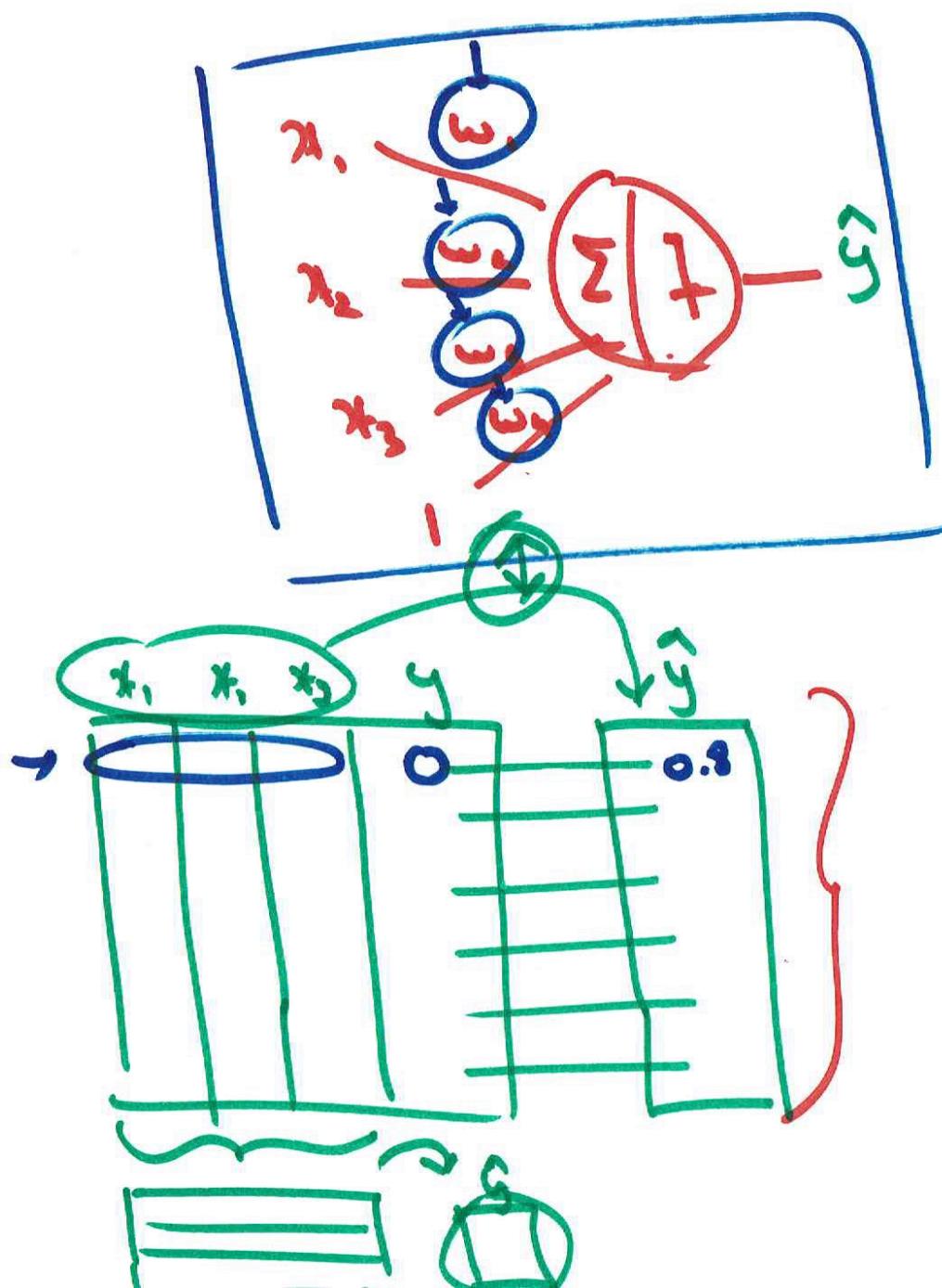
# Perception



$$\sum_{i=1}^n x_i \geq \theta \quad \hat{\uparrow} \quad 0$$
$$\sum_{i=1}^n x_i < \theta \quad \hat{\uparrow} \quad 0$$

$$\theta = 1 \quad \hat{\uparrow} \quad "0\&"$$

$$\theta = 3 \quad \hat{\uparrow} \quad "AND"$$

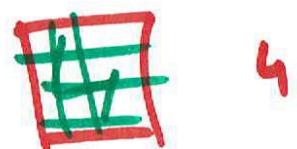
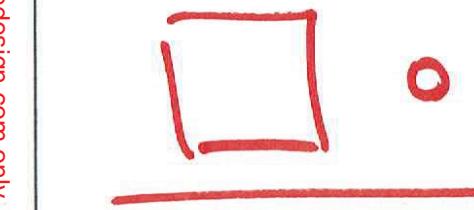
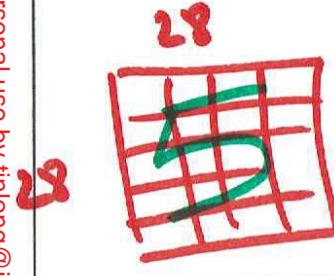


MNIST

## An Example

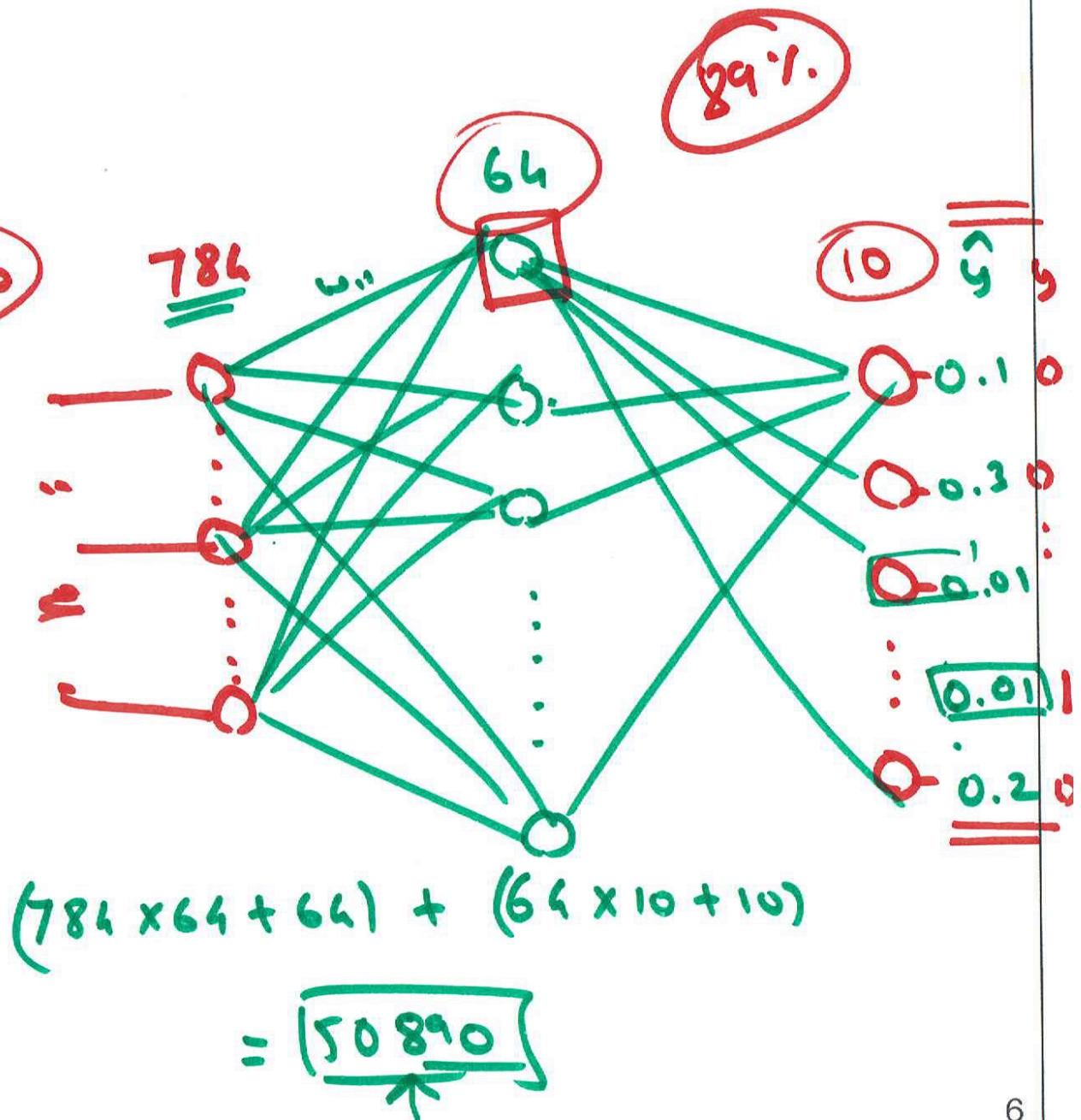
5 0 4 1 9 2 1 3 1 4  
= = = = = = = = = =

60000 train

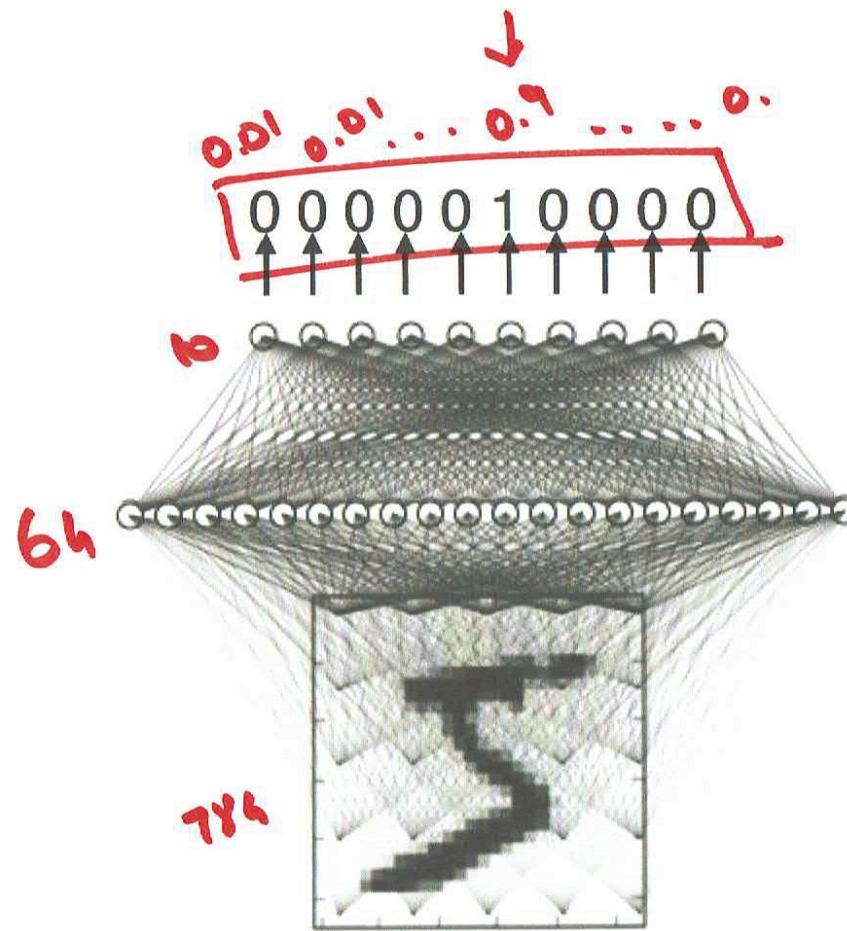


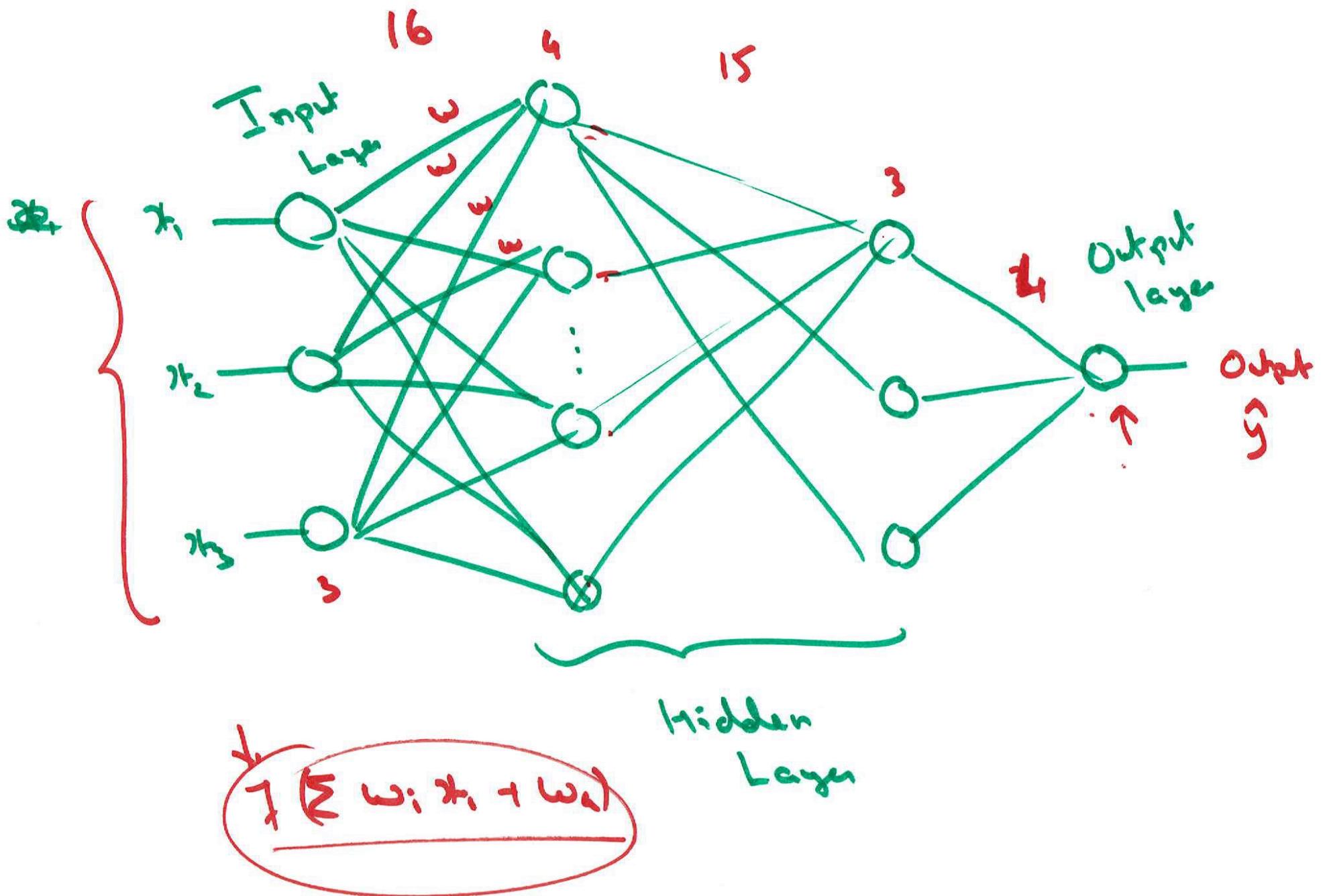
9

10000

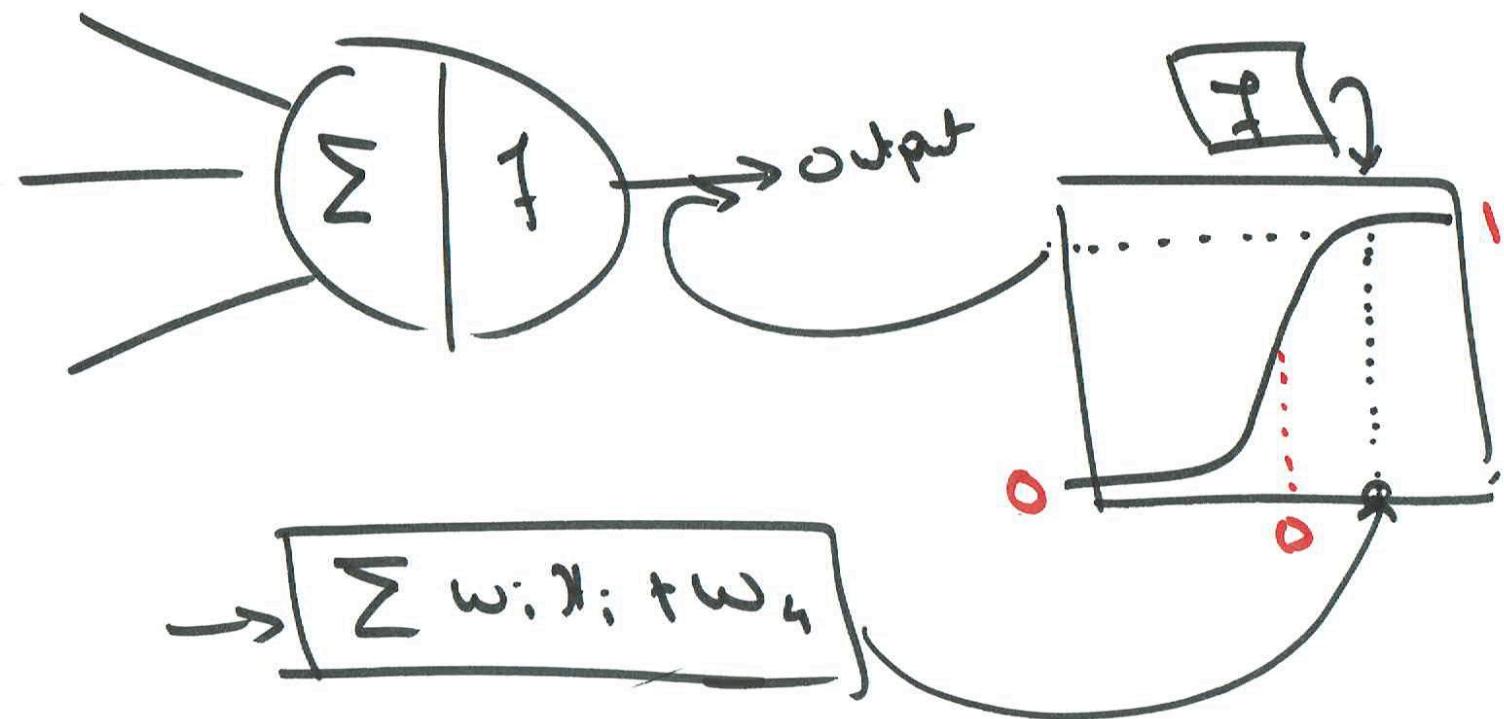


# An Example

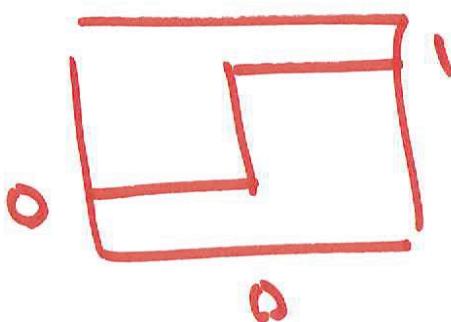




sigmoid

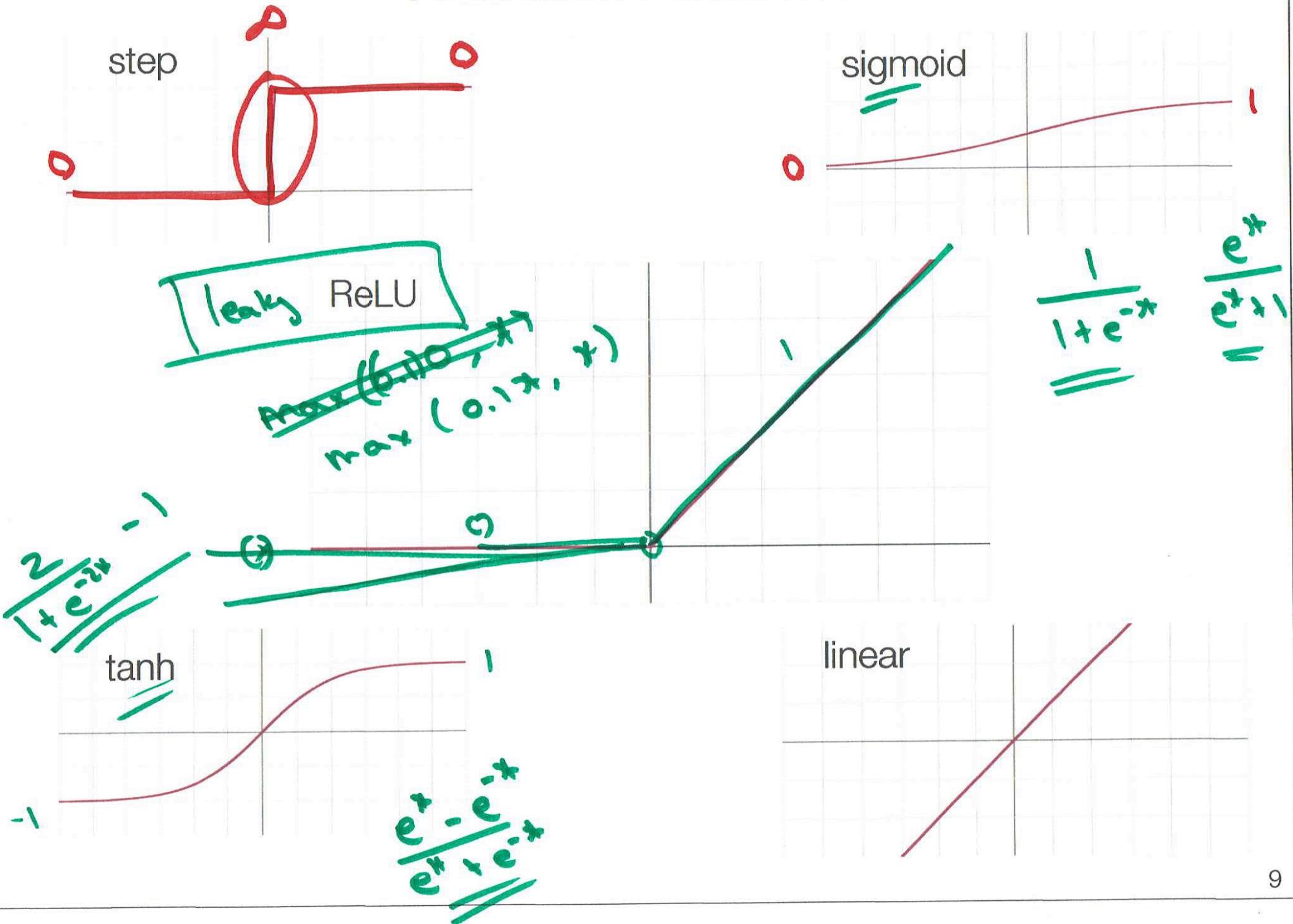


Step



$$\text{Output} = \text{step}(\Sigma w_i \cdot i + w_0)$$

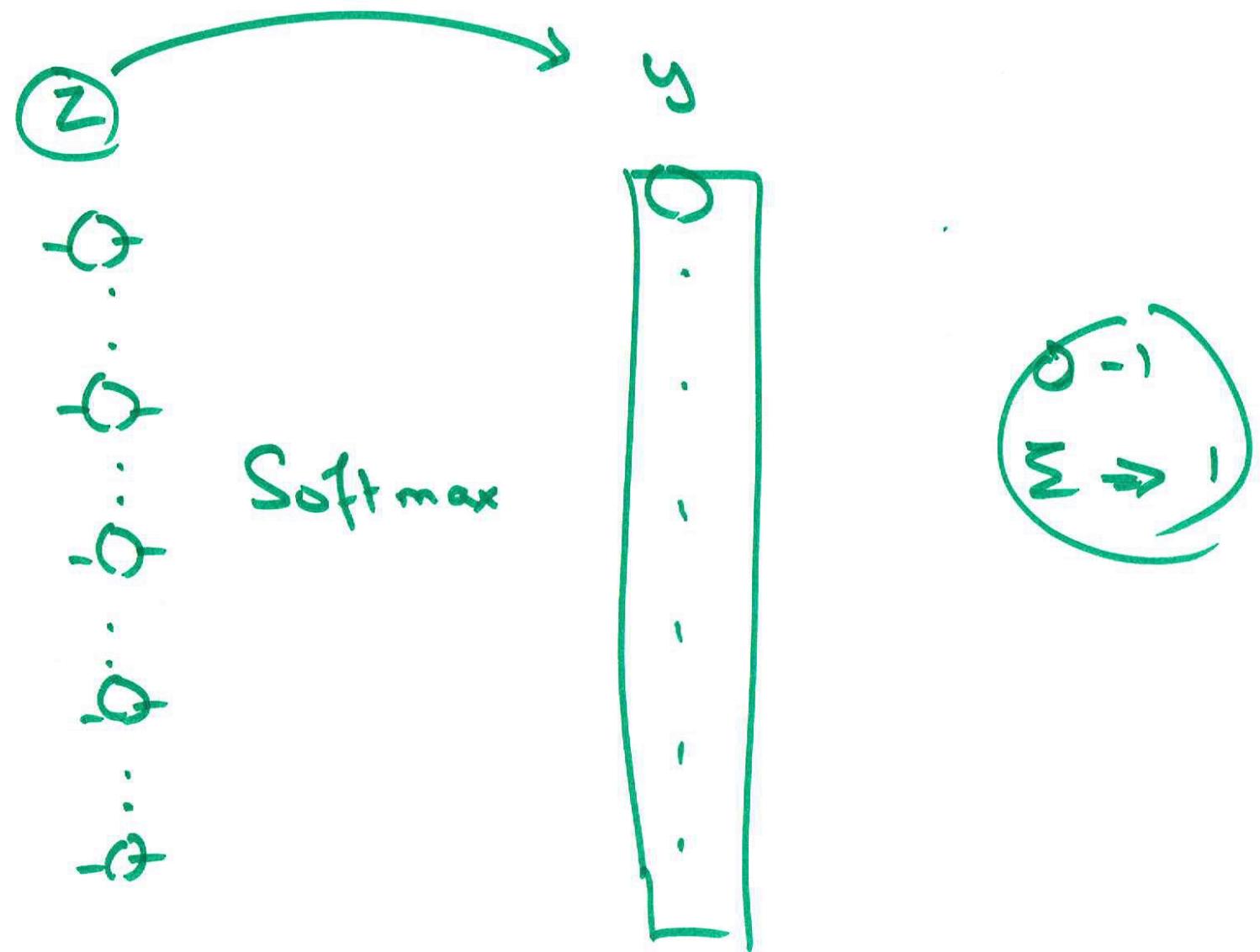
# Activation Functions



$\alpha$

$$f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$f(x) = \max(0, x)$$



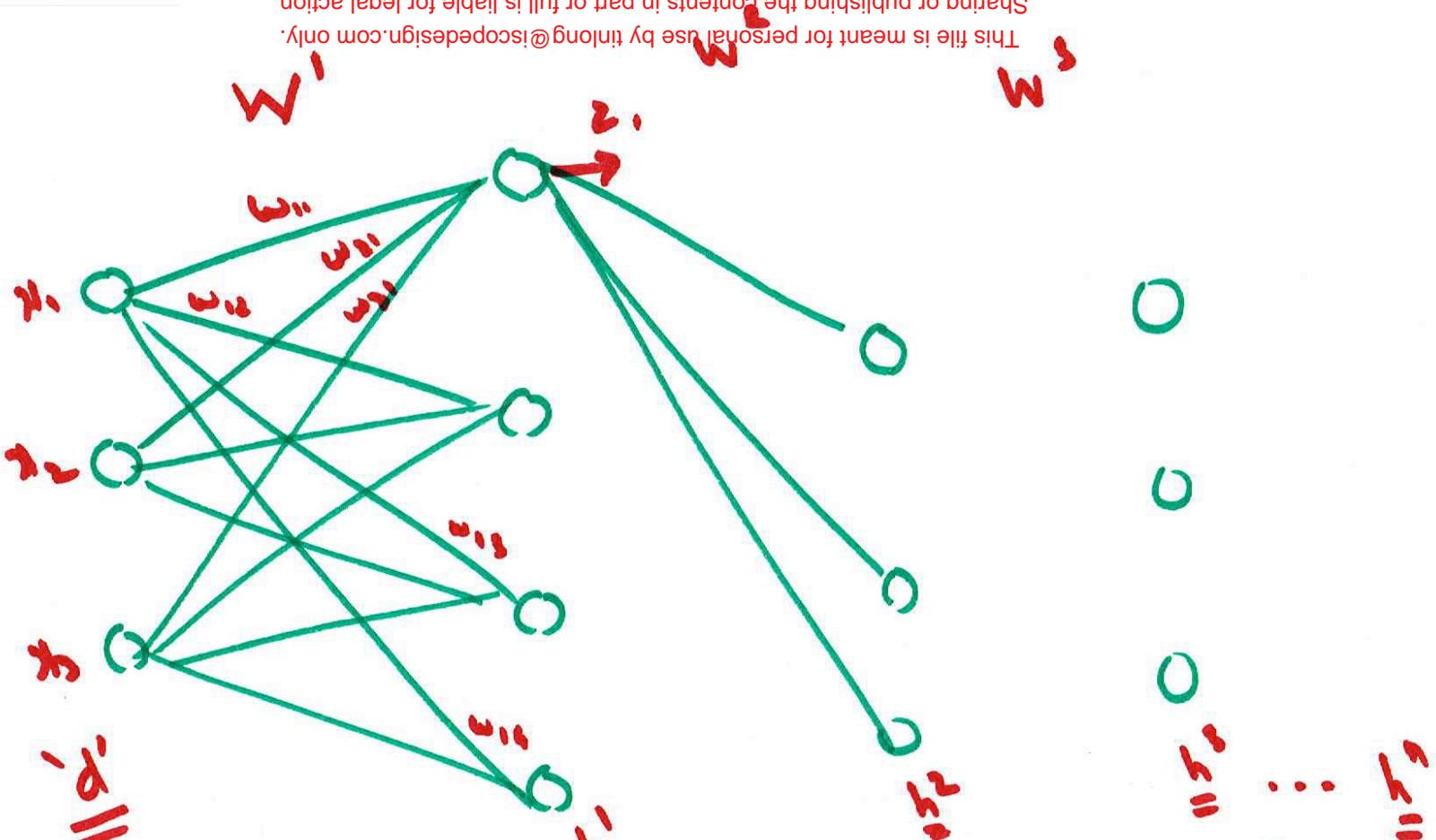
50

N

$$\dot{\gamma} = \frac{e|v|}{\mu}$$

$$\mathbf{z}^{\text{new}} = \mathbf{z}^{\text{old}} - \eta \nabla_{\mathbf{z}} l(\mathbf{z})$$

$$= \mathbf{z}^{\text{old}} - \frac{1}{N} \eta \sum_i \nabla_{\mathbf{z}} l_i(\mathbf{z})$$



$$z_1 = f(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4 + b)$$

$$z_j = f(\sum_i w_{ij}x_i + b)$$

$$W' = \begin{pmatrix} w_{00} & w_{01} & w_{02} & \dots & w_{0n} \\ w_{10} & w_{11} & w_{12} & \dots & w_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{d0} & w_{d1} & w_{d2} & \dots & w_{dn} \\ w_{00} & \dots & \dots & \dots & w_{dn} \end{pmatrix}$$

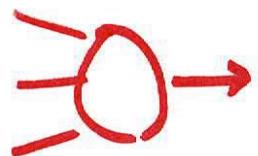
$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{pmatrix}$

$d \times d$  matrix

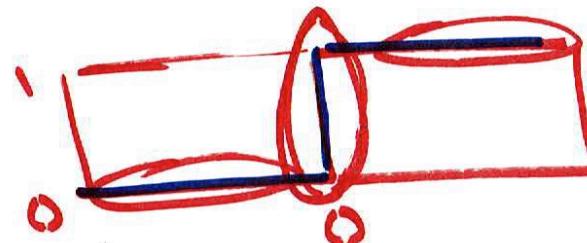
$$\hat{y} = f(-f(f(W^T)^T)(W^T f(W' x + b') + b^2) + b^3) \dots$$

Diagram illustrating the dimensions of the terms in the equation:

- $W^T$  is  $d \times d$ .
- $f(W^T)$  is  $d \times 1$ .
- $f(W' x + b')$  is  $d \times 1$ .
- $(W^T f(W' x + b'))$  is  $d \times d$ .
- $f(W^T)^T$  is  $d \times d$ .
- $(W^T f(W' x + b'))^T$  is  $d \times 1$ .
- $(W^T)^T f(W' x + b')$  is  $d \times 1$ .
- $(W^T)^T f(W^T)^T$  is  $d \times d$ .
- $(W^T)^T f(W^T)^T (W^T f(W' x + b'))$  is  $d \times d$ .
- $(W^T)^T f(W^T)^T (W^T f(W' x + b'))^T$  is  $d \times 1$ .
- $(W^T)^T f(W^T)^T (W^T f(W' x + b'))^T f(W' x + b')$  is  $d \times 1$ .
- $(W^T)^T f(W^T)^T (W^T f(W' x + b'))^T f(W' x + b') + b^2$  is  $d \times 1$ .
- $(W^T)^T f(W^T)^T (W^T f(W' x + b'))^T f(W' x + b') + b^2 + b^3$  is  $d \times 1$ .
- $\sqrt{\sum w_{ii} x_i + b}$  is  $d \times 1$ .

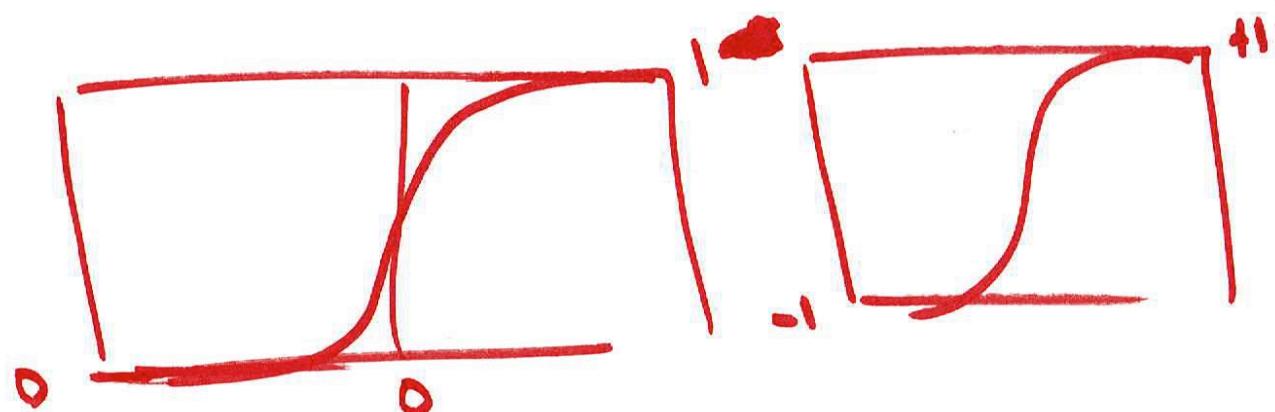


Step ( $\omega_1x_1 + \omega_2x_2 + b$ )



Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$
$$\therefore \frac{e^x}{e^x + 1}$$

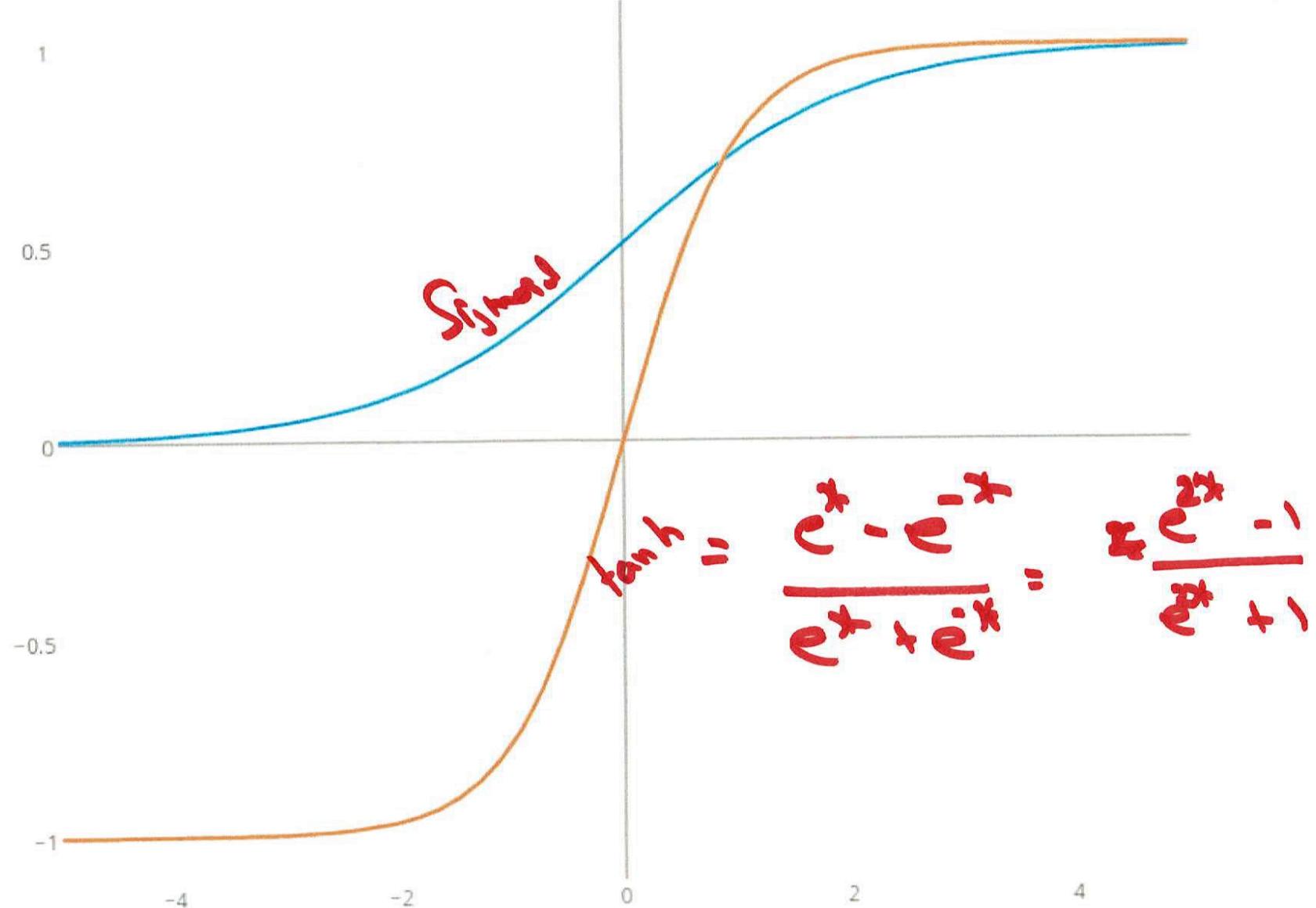


$$2\sigma(x) - 1$$

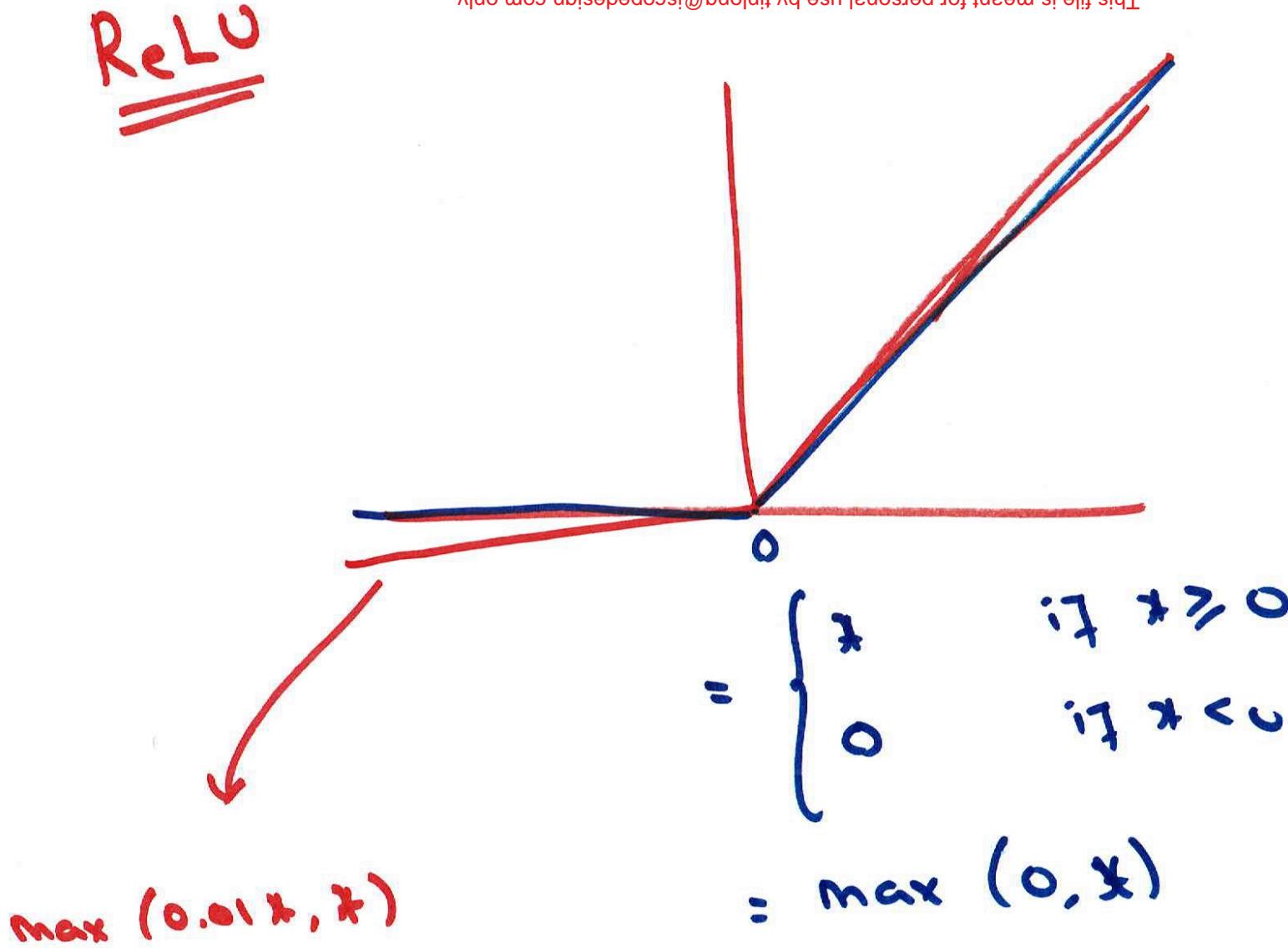
$$\tanh = 2\sigma(2x) - 1$$

- Sigmoid function
- Tanh function

Sharing or publishing the contents in part or full is liable for legal action.  
This file is meant for personal use by [tinlong@iscopedesign.com](mailto:tinlong@iscopedesign.com) only.



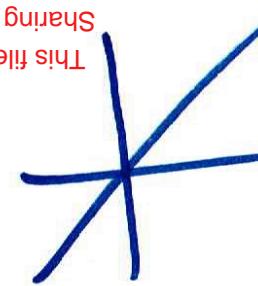
ReLU



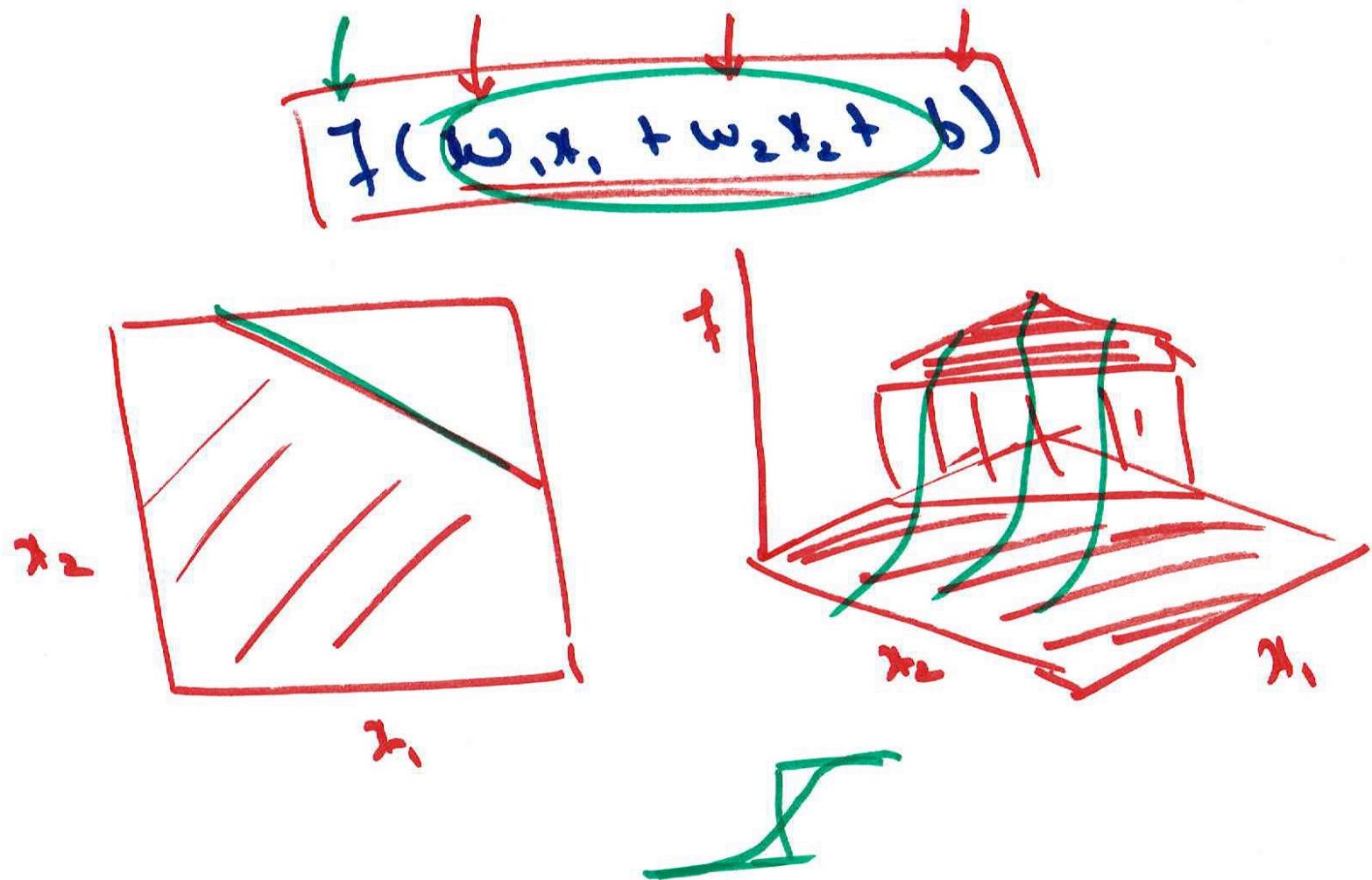
Sharing or publishing the contents in part or full is liable for legal action.  
This file is meant for personal use by [tinlong@iscapedesign.com](mailto:tinlong@iscapedesign.com) only.

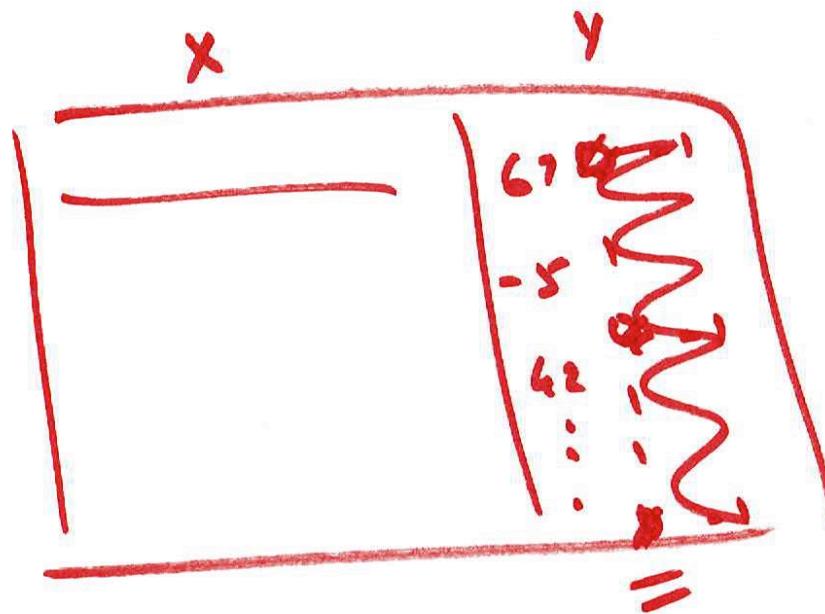
Linear

= \*



$$\text{Step } = f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$





Output nodes

Classification

Sigmoid, tanh  
Softmax

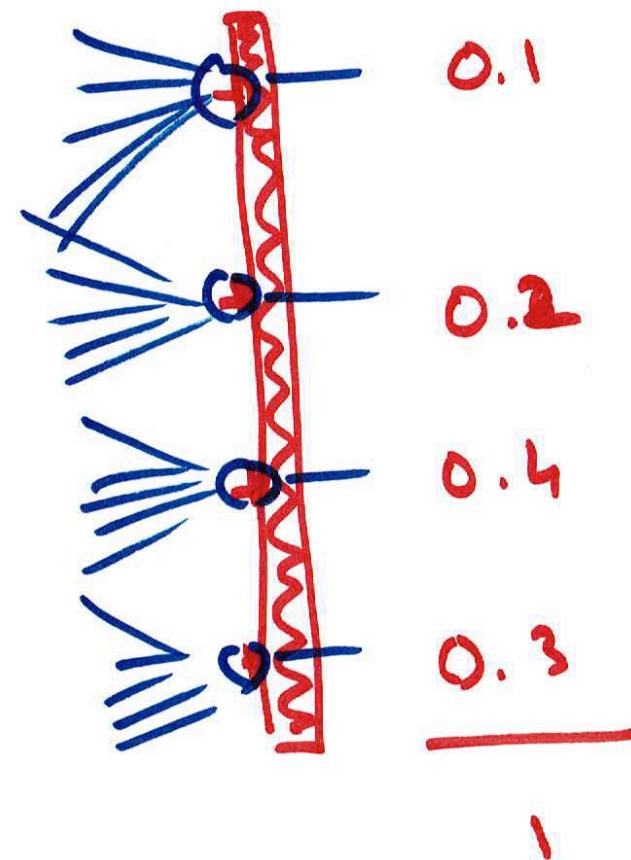
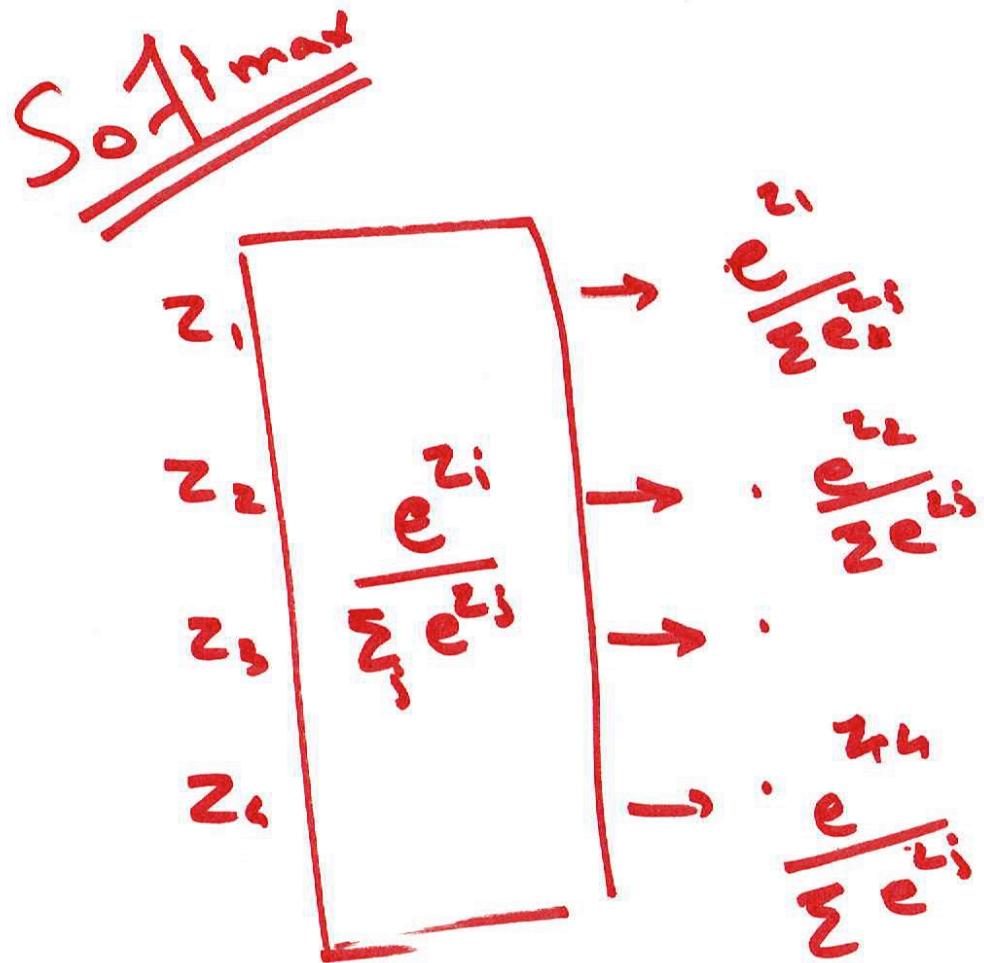
hidden layer

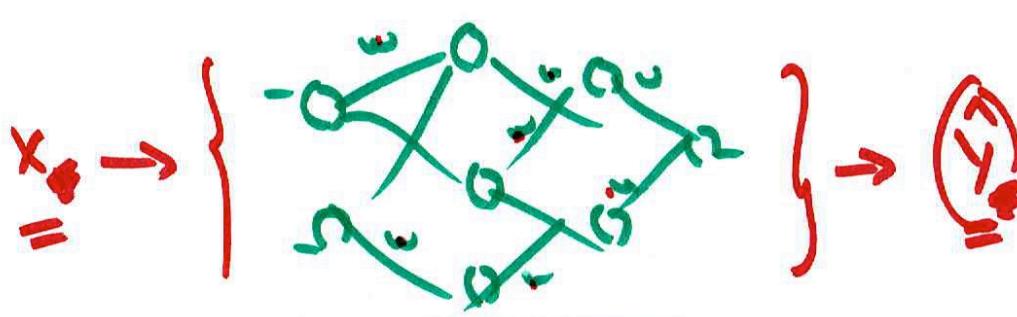
Sigmoid ✓  
tanh ✓

ReLU ✓

linear

$$\hat{a} + \hat{b} (a + b x)$$





Loss Function

$$L(y, \hat{y}) = \frac{1}{n} \sum_i (y_i - \hat{y}_i)^2$$

ReLU

$$L(y, \hat{y}) = L(\omega)$$

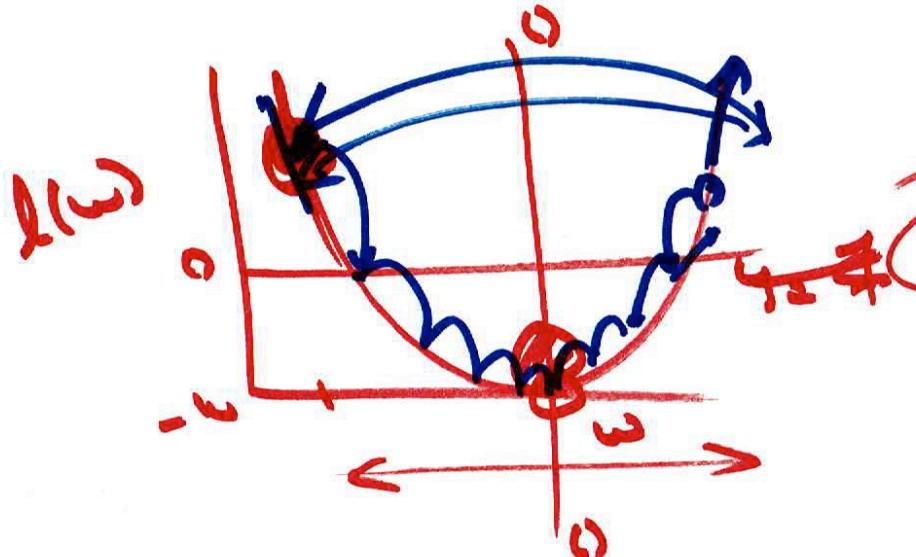
$L_2$  loss  
MSE  
SSE

Classification

$$L(y, \hat{y}) = - (y_1 \log(\hat{y}_1) + (1-y_1) \log(1-\hat{y}_1))$$

Cross entropy loss

How  $\min_{\underline{w}} \underline{L}(\underline{y}, \underline{\underline{w}})$  by changing  $w^1, w^2, \dots, w^n$



$$J(w) = w^2 - 10 = 0$$

$$\frac{dJ}{dw} = 2w = 0$$

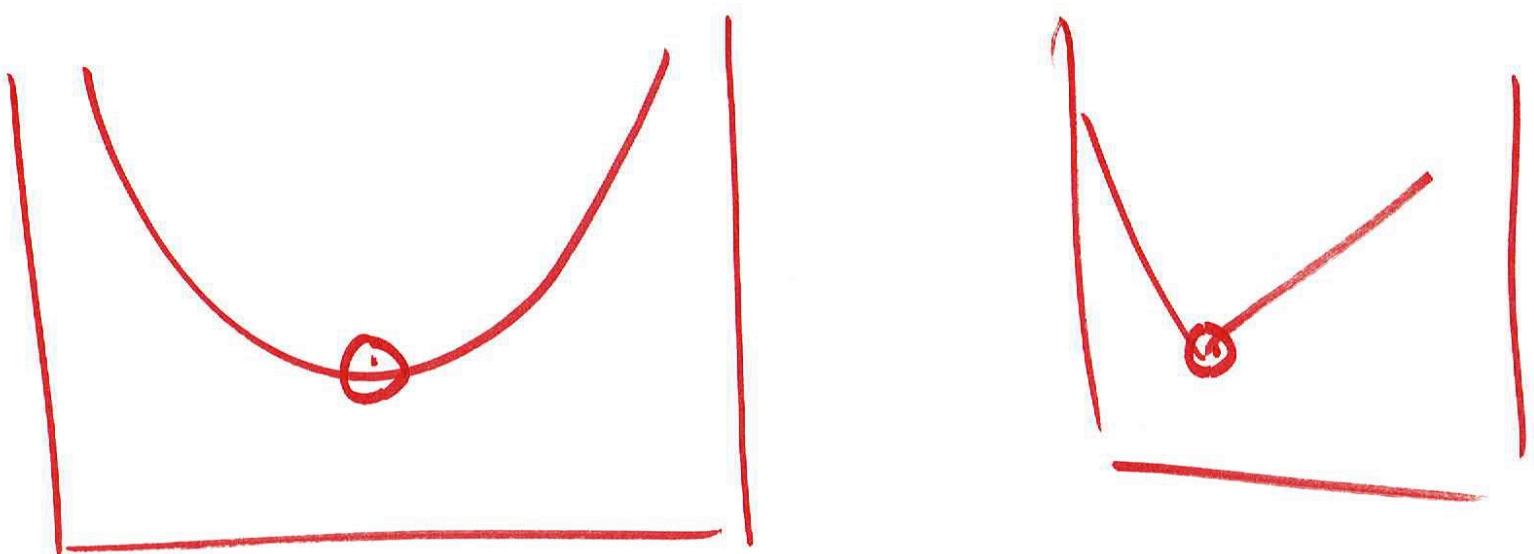
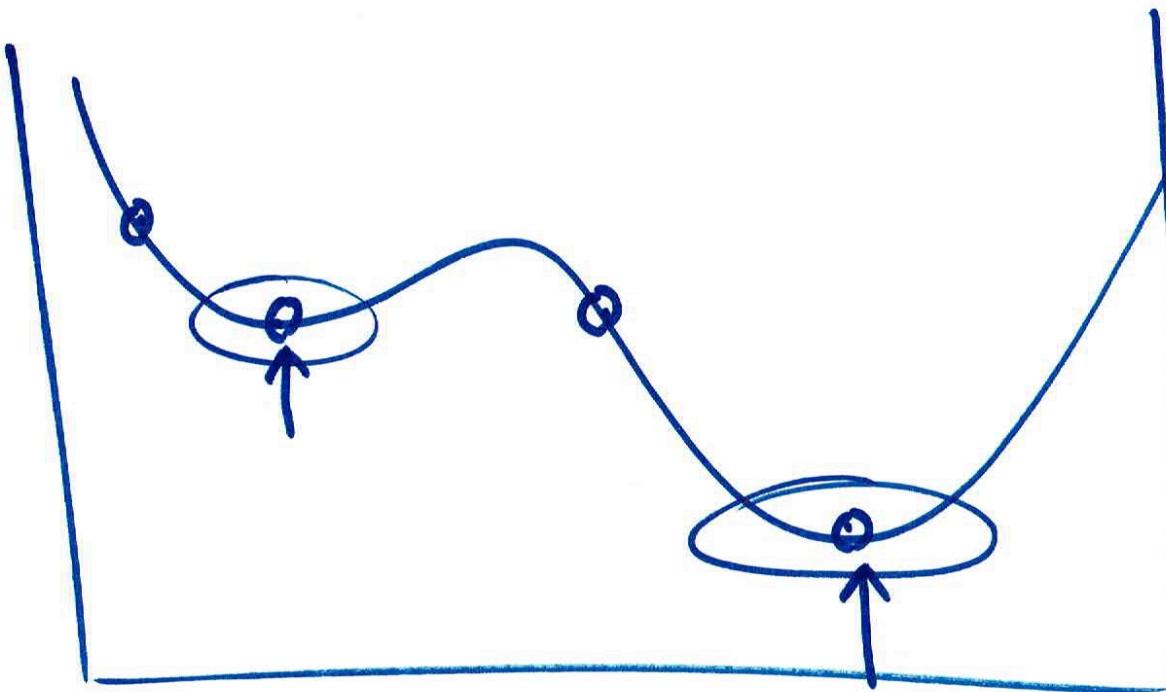
$$w = 0$$

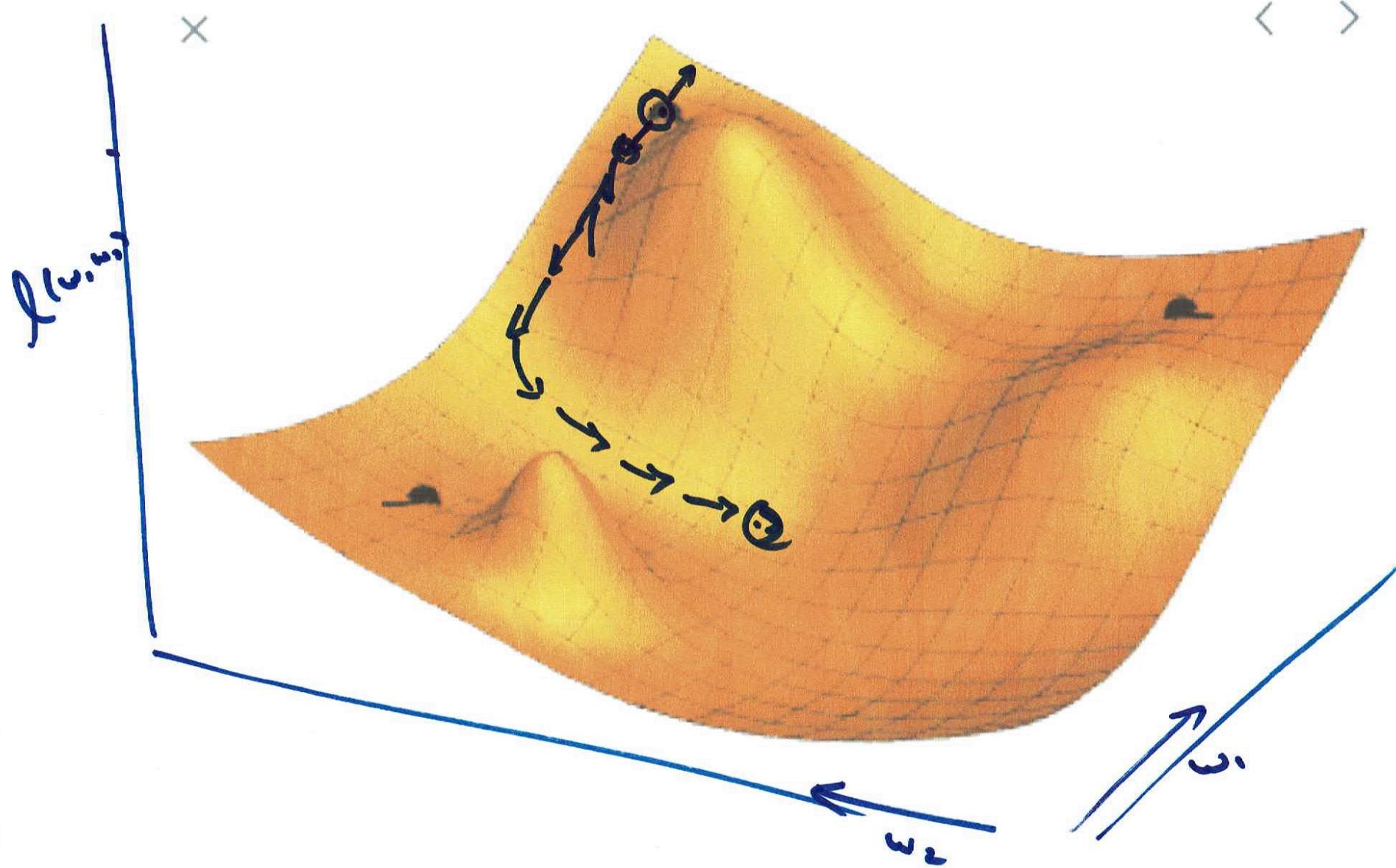
$$\frac{dL}{dw} = \boxed{\quad} = 0$$

$$\frac{dl}{dw}$$

$$w^{new} = w - \eta \nabla_w l$$

↑ learning rate





$$\hat{w} = w^{\text{old}} - \frac{\eta \nabla_w l(w)}{N}$$

$$= w^{\text{old}} - \frac{1}{N} \nabla_w \left( \sum_i \nabla_w l_i(w) \right)$$

(SAD)

$$w^{new} = w^{\text{old}} - \eta \nabla_w l_i(w) \leftarrow$$

$$w^{new} = w^{\text{old}} - \frac{1}{N} \eta \sum_{\text{over a min batch}} \nabla_w l_i(w)$$

Loss

$$L = \frac{1}{N} \sum \text{tr} \left( \underline{\underline{y}}_i - \underline{\underline{T}}^n \left( \dots \text{tr} \left( \underline{\underline{w}}^2 \underline{\underline{T}}^i \left( \underline{\underline{w}}^i \underline{\underline{x}} + \underline{\underline{b}}^i \right) + \underline{\underline{b}}^2 \right) \right) \right)$$

Chain Rule

$$\underline{\underline{T}}(\underline{\underline{g}}(\underline{\underline{h}}(\underline{\underline{x}})))$$

$$\frac{d\underline{\underline{T}}}{d\underline{\underline{x}}} = \boxed{\frac{d\underline{\underline{T}}}{d\underline{\underline{g}}}} \cdot \frac{d\underline{\underline{g}}}{d\underline{\underline{h}}} \cdot \frac{d\underline{\underline{h}}}{d\underline{\underline{x}}}$$

