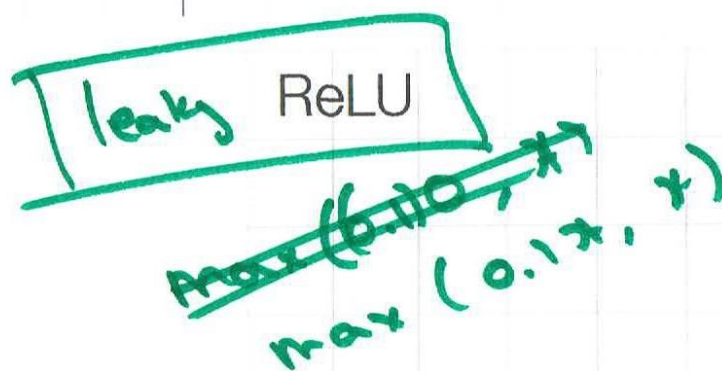
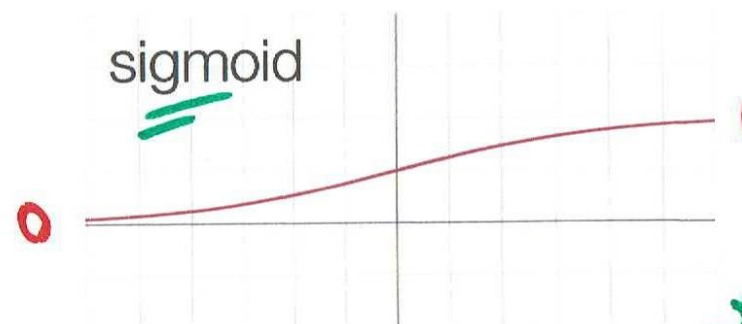
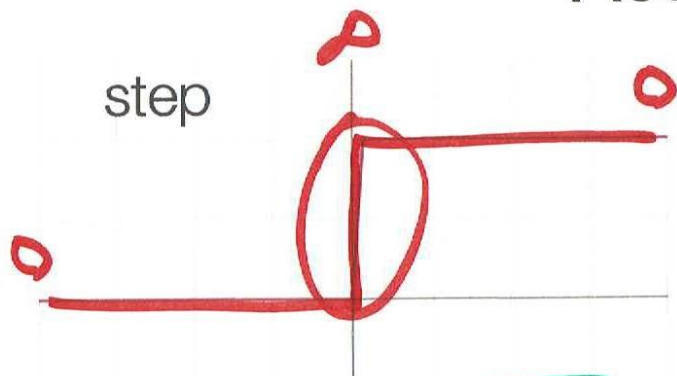
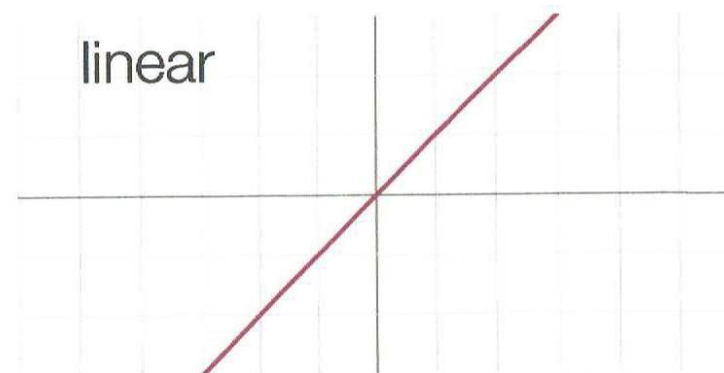
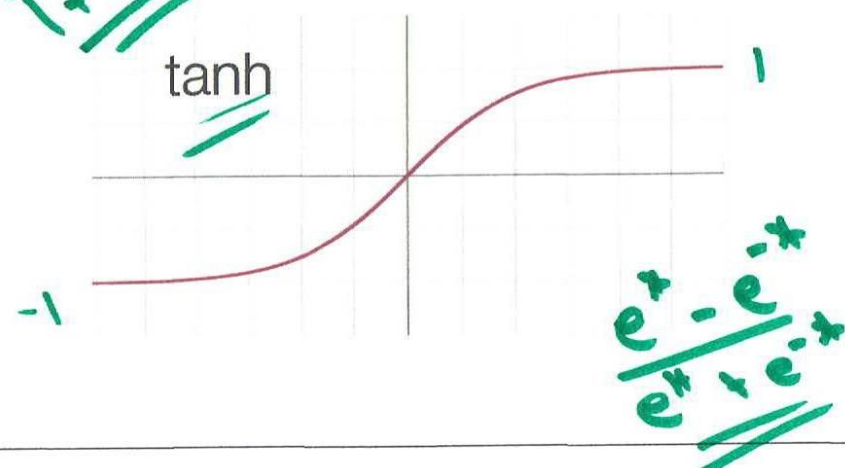
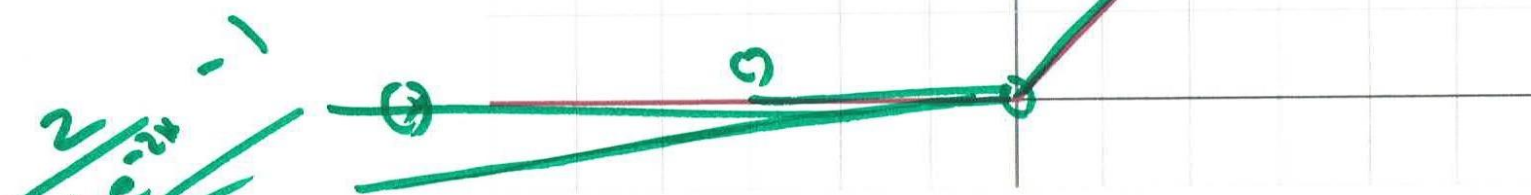


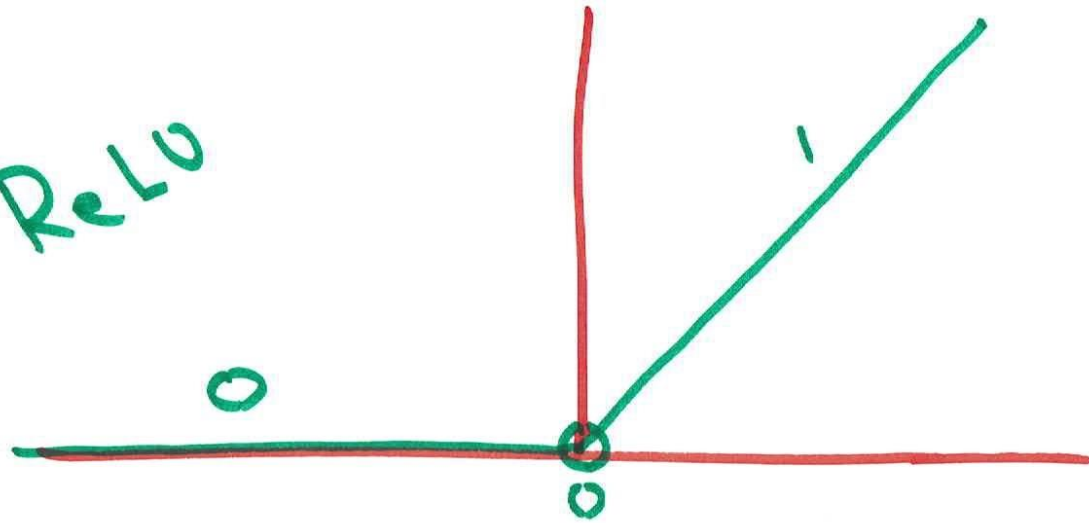
Activation Functions



$$\frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

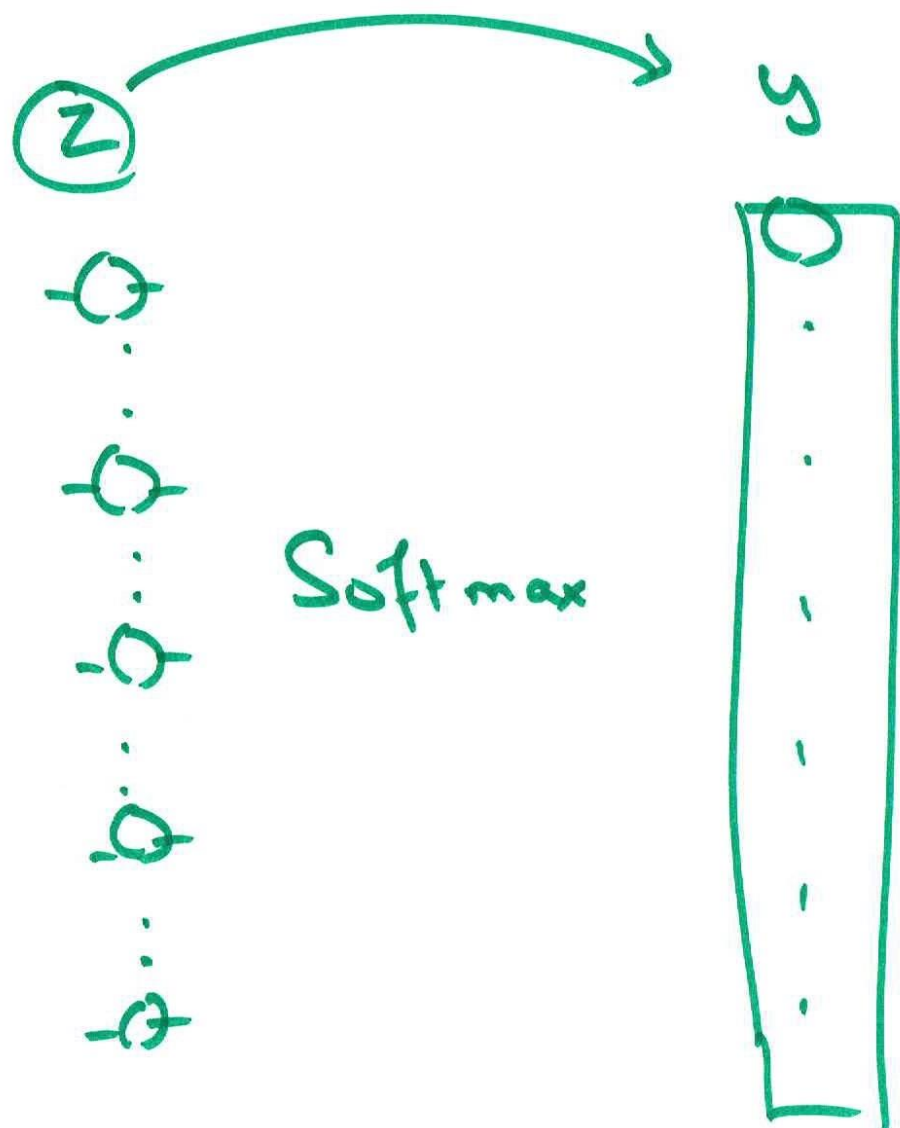


ReLU



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

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



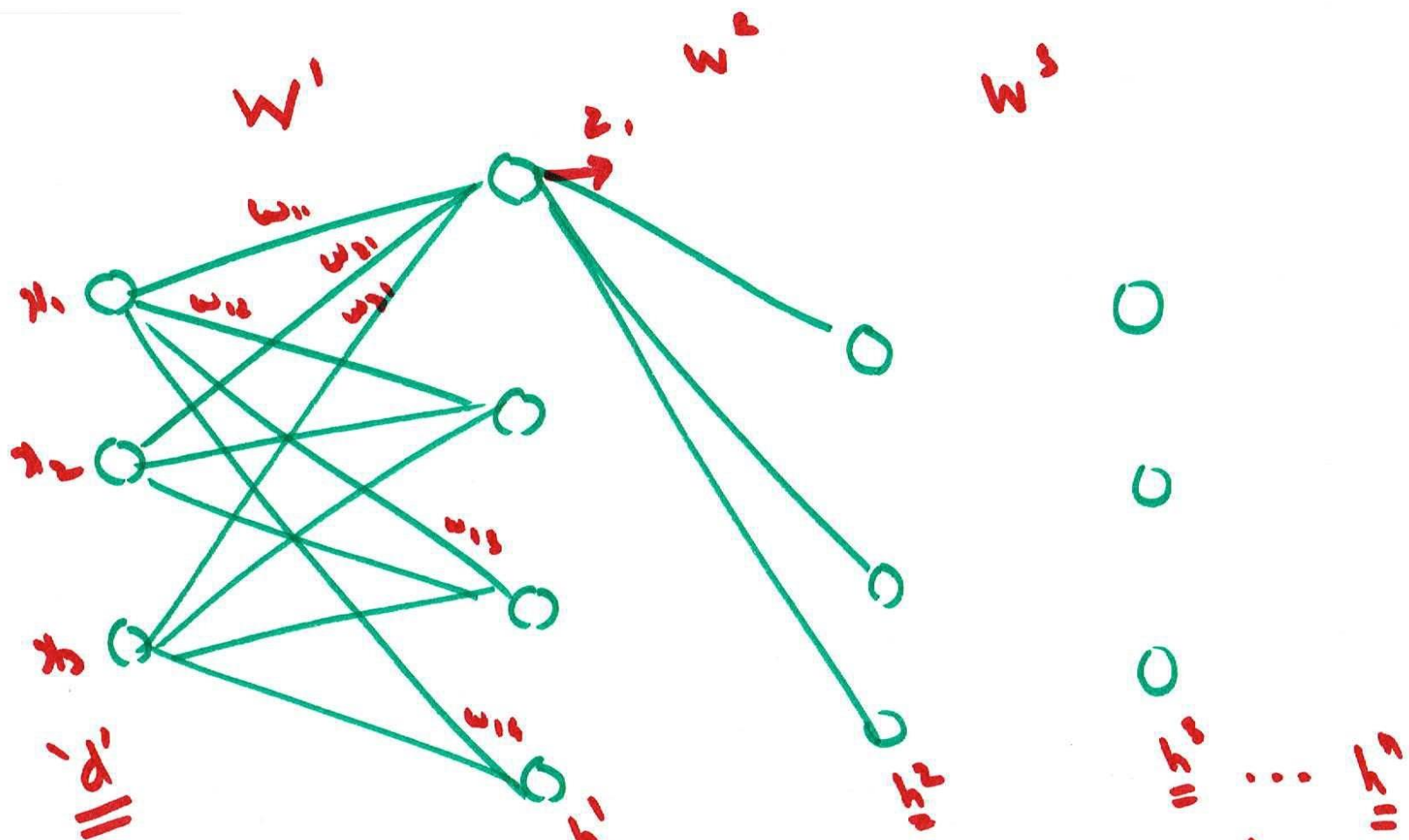
z

y

$$y_i = \frac{e^{z_i}}{\sum e^{z_i}}$$

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

$$= \omega^{\text{old}} - \frac{1}{N} \eta \sum \nabla_{\omega} l_i(\omega)$$



$$z_1 = f(w_{11}x_1 + w_{21}x_2 + w_{31}x_3 + b)$$

$$z_j = f\left(\sum_i w_{ji}x_i + b_j\right)$$

$$W' = \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & \dots & w_{mn} \end{pmatrix}$$

$d \times d$ matrix




Diagram illustrating the chain rule for the derivative of a nested sigmoid function:

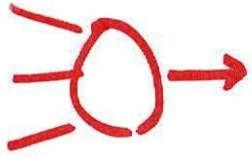
$$\hat{y} = f(f(f(W^3 f(W^2 f(W^1 x + b^1) + b^2) + b^3)))$$

The diagram shows the flow of the derivative from the output \hat{y} through the layers of weights and biases, ultimately reaching the input x_1 .

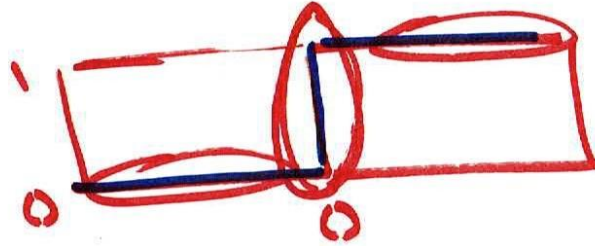
Key components and flow:

- Input:** x_1
- Layer 1 (Hidden):** $W^1 x + b^1$ (Derivative flow: $h^1 x_1$)
- Layer 2 (Hidden):** $W^2 f(W^1 x + b^1) + b^2$ (Derivative flow: $h^2 x_1$)
- Layer 3 (Hidden):** $W^3 f(W^2 f(W^1 x + b^1) + b^2) + b^3$ (Derivative flow: $h^3 x_1$)
- Output:** \hat{y} (Derivative flow: $h^4 x_1$)

The final result of the derivative calculation is $h^2 x_1$.

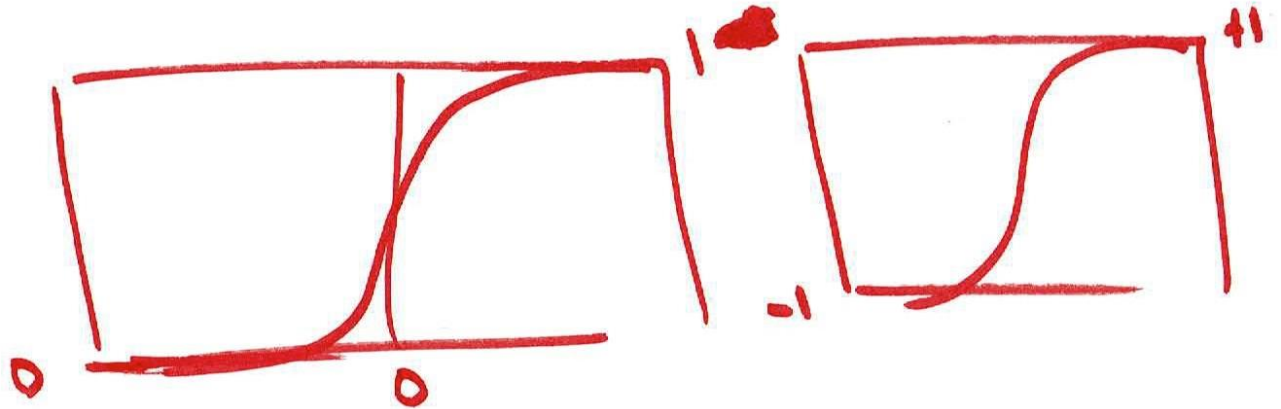


Step (\downarrow \downarrow \downarrow
 $w_1 x_1 + w_2 x_2 + w_3 x_3 + b$)



Sigmoid

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

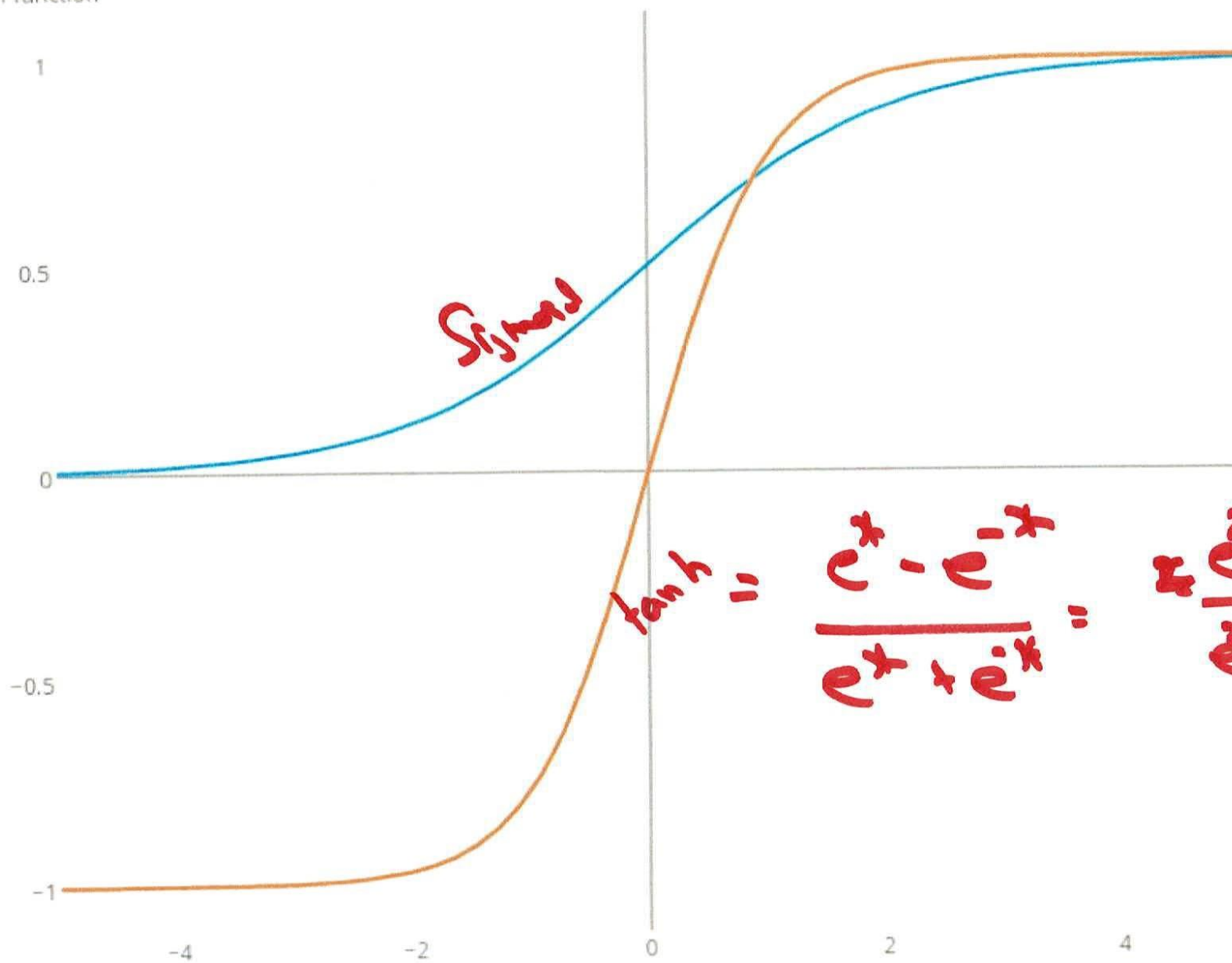


$$\boxed{2\sigma(x) = 1}$$

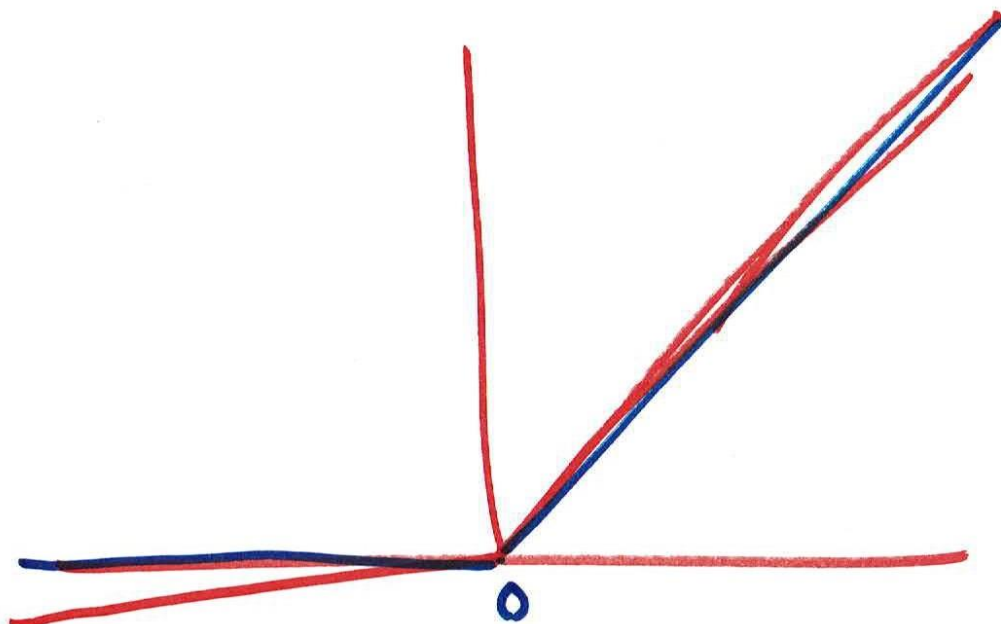
$$\boxed{\tanh = 2\sigma(2x) - 1}$$

↑

— Sigmoid function
— Tanh function



ReLU



$$\max(0.01x, x)$$

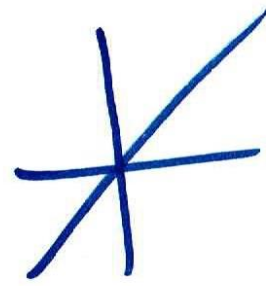
$$= \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$\begin{aligned} &\text{if } x \geq 0 \\ &\text{if } x < 0 \end{aligned}$$

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

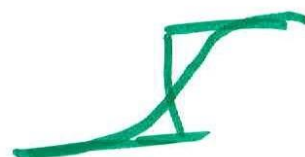
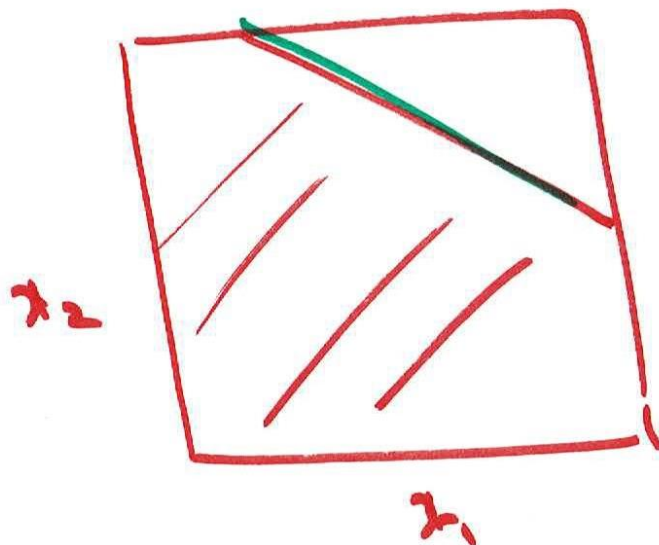
linear

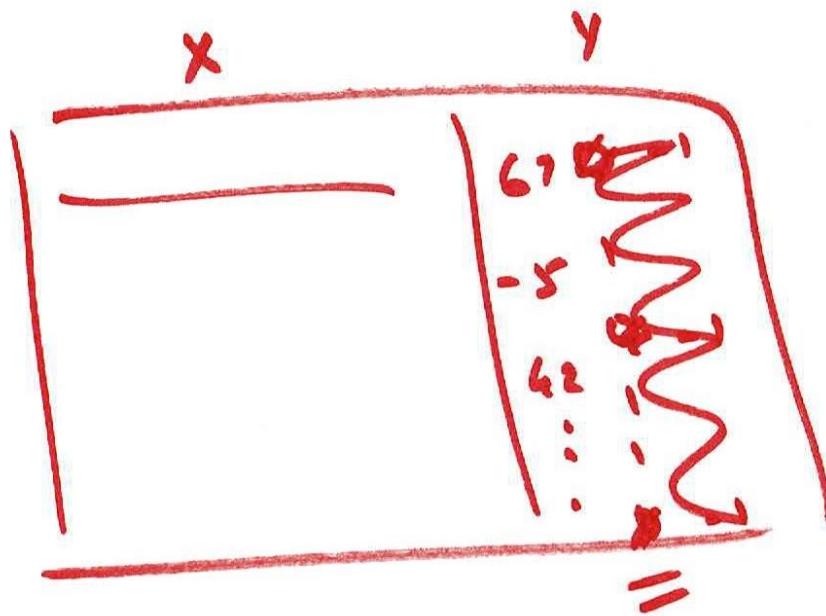
= λ



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

$$f(\omega_1 x_1 + \omega_2 x_2 + b)$$





Classification \rightarrow

Sigmoid, tanh
Softmax

Reg \rightarrow

linear

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

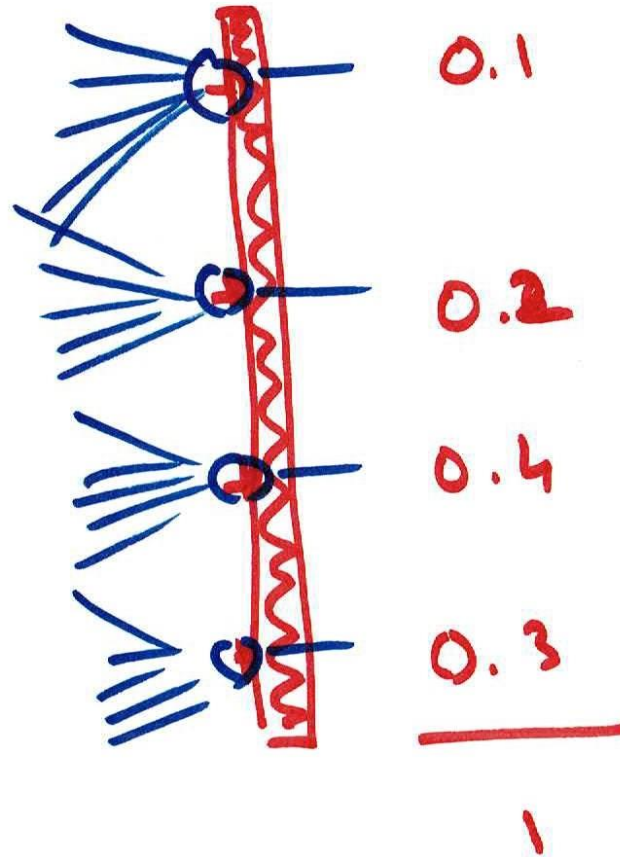
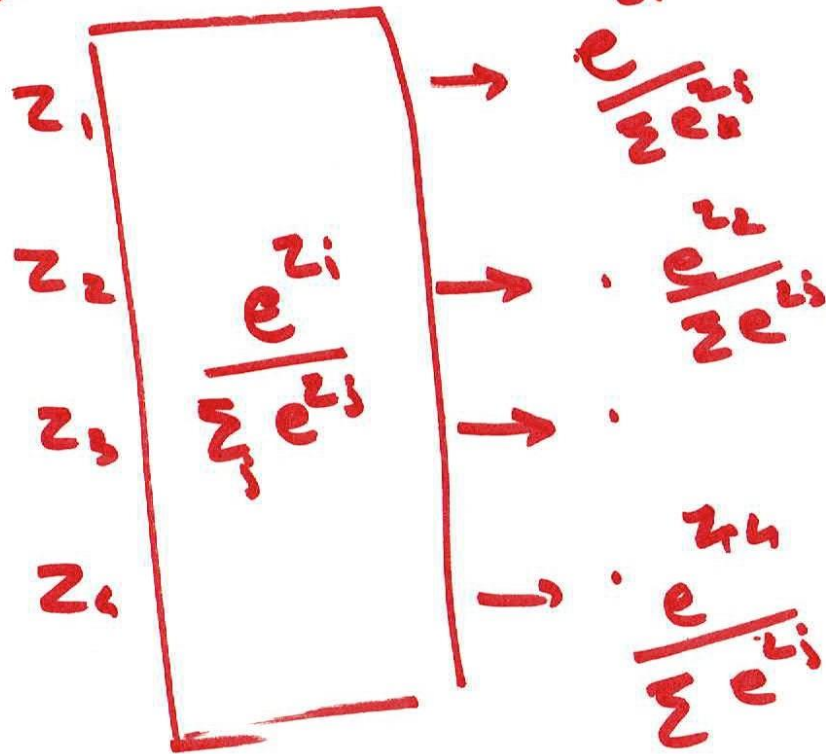
Output nodes

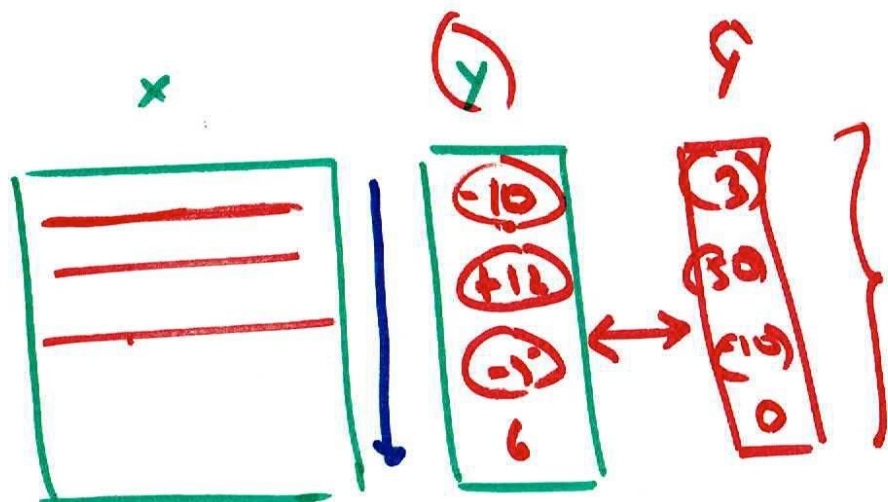
hidden layer

Sigmoid \checkmark
tanh \checkmark

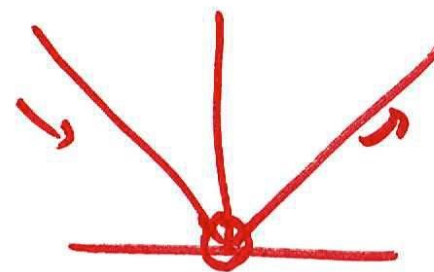
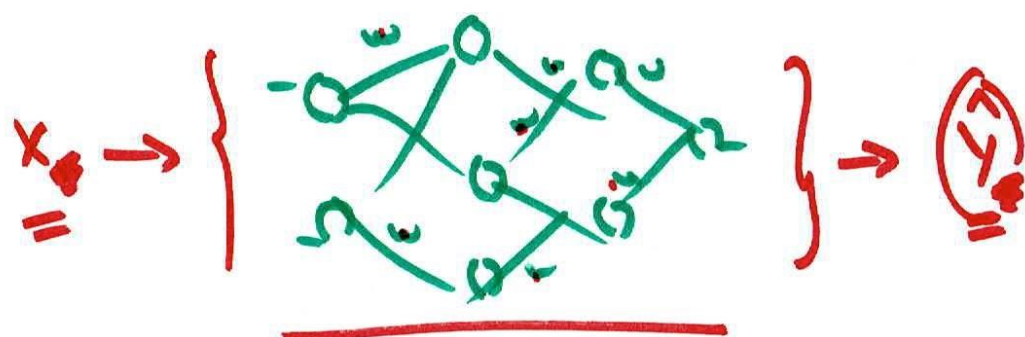
ReLU \checkmark
~~linear~~

Softmax





-10	6
$+12$	30
-5	19
6	0



Loss function $L(y, \hat{y})$

Re

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

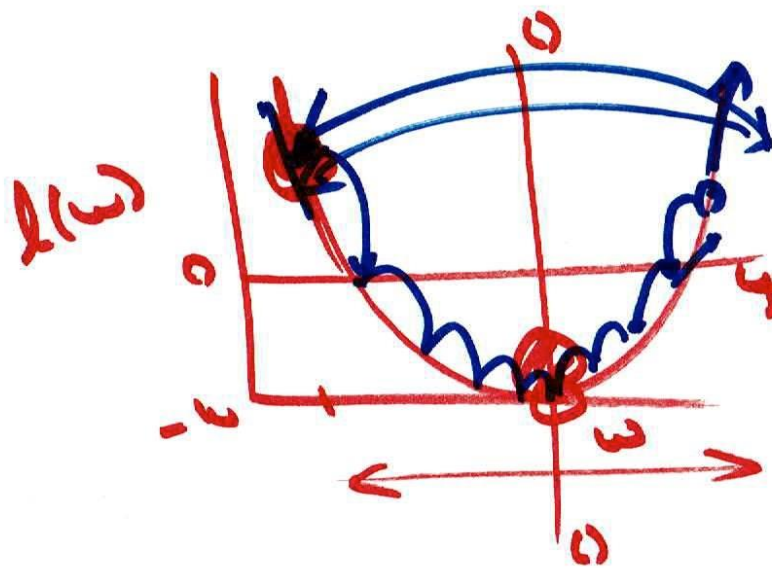
$$L(y, \hat{y}) = l(w)$$

L_2 loss
MSE
SSE

Classifier $L(y, \hat{y}) = - (y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i))$

Cross entropy loss

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



$$y = x^2 - 10 = -10$$

$$\frac{dy}{dx} = 2x = 0$$

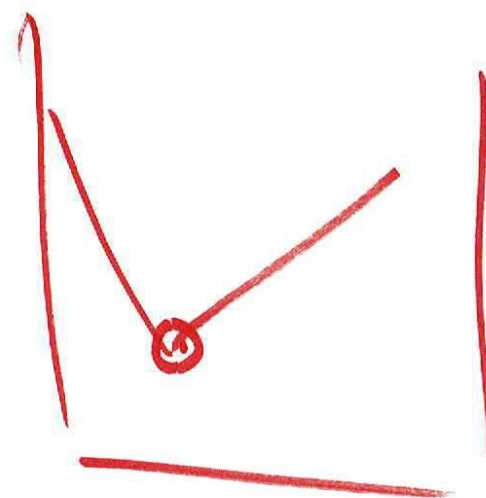
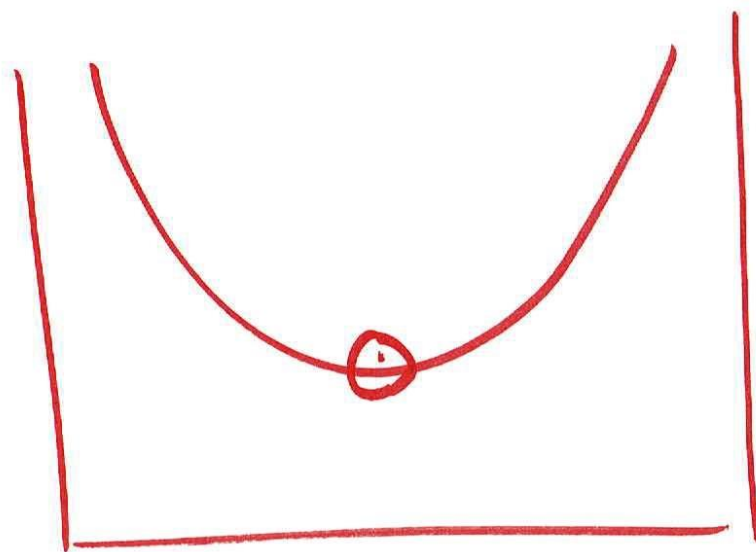
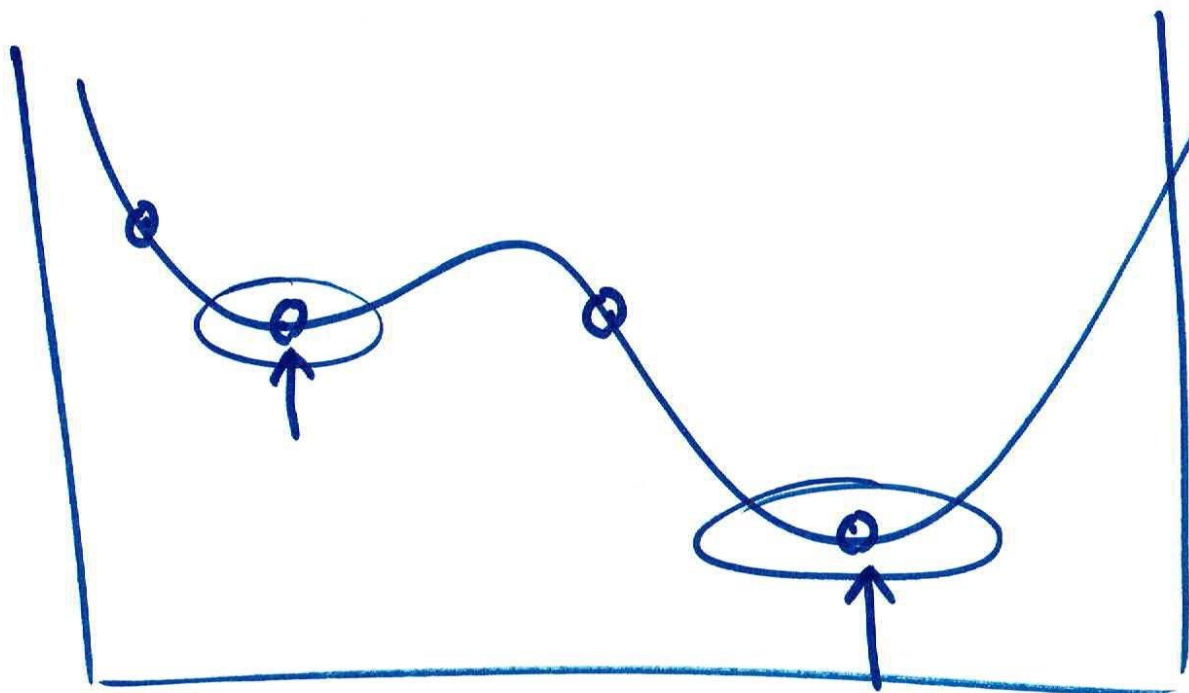
$$x = 0$$

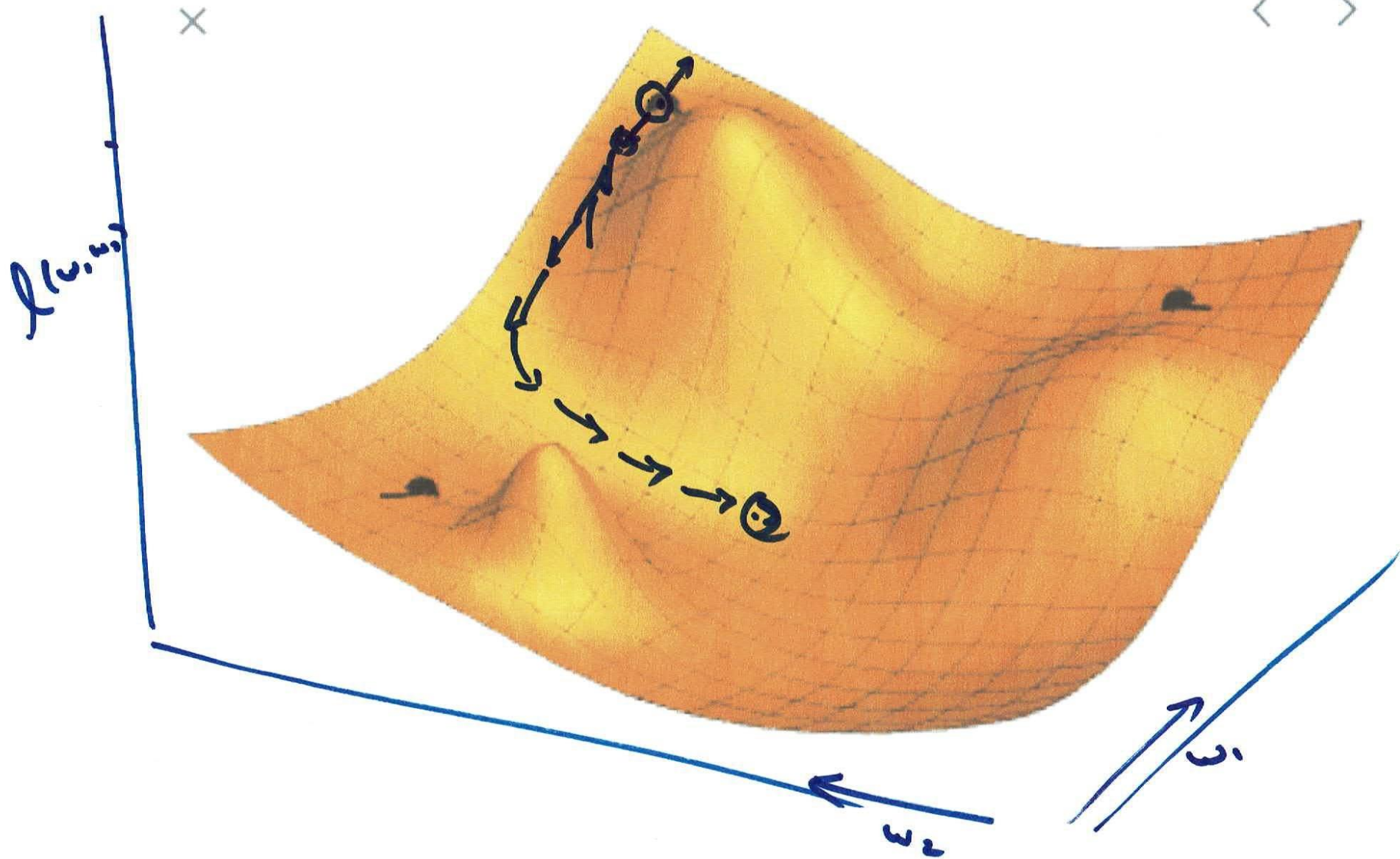
$$\left(\frac{dL}{dw} \right) = \underbrace{\hspace{10em}} = 0$$

$$\frac{dL}{dw}$$

$$w_{new} = w - \eta \nabla_w L$$

↑
learning rate





$$\omega^{\text{new}} = \omega^{\text{old}} - \frac{2 \nabla_{\omega} l(\omega)}{2}$$

$$= \omega^{\text{old}} - \frac{1}{N} \sum \nabla_{\omega} l_i(\omega) \leftarrow$$

SGD

$$\omega^{\text{new}} = \omega^{\text{old}} - \eta \nabla_{\omega} l_i(\omega) \leftarrow$$

$$\omega^{\text{new}} = \omega^{\text{old}} - \frac{1}{N} \eta \sum \nabla_{\omega} l_i(\omega)$$

over
a min
batch

Loss

$$L = \frac{1}{2} \sum_i (y_i - \underbrace{f(\dots f(w^2 f'(w'x + b')) + b^2))_i)^2$$

Function of w ($l(w)$)

Chain Rule

$$f(g(h(x)))$$

$$\frac{df}{dx} = \left(\frac{df}{dy} \cdot \frac{dy}{dh} \right) \cdot \frac{dh}{dx}$$

Back Propagation