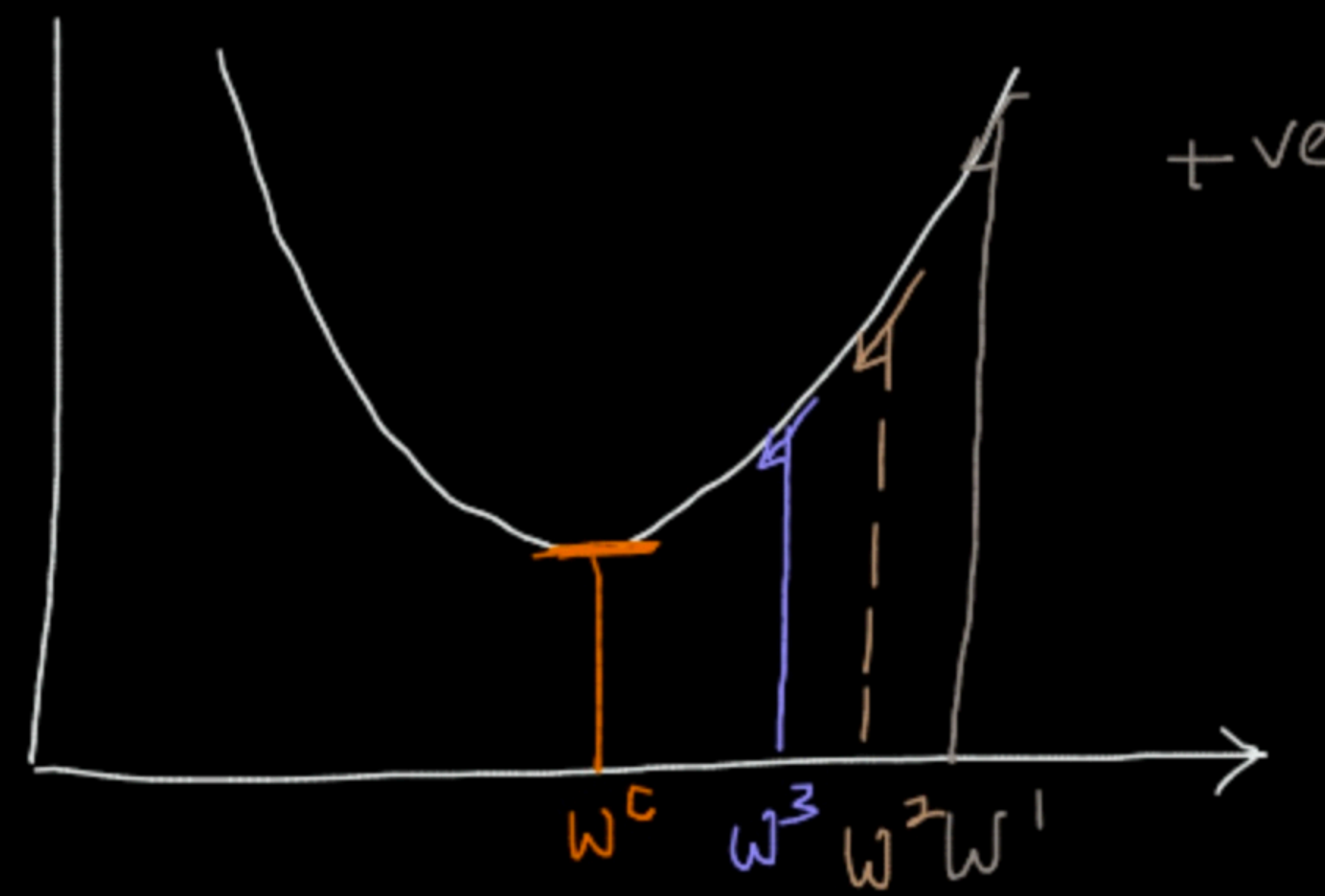
Gradient Descent Algorithm



Loss = $(y - \hat{y})^2$

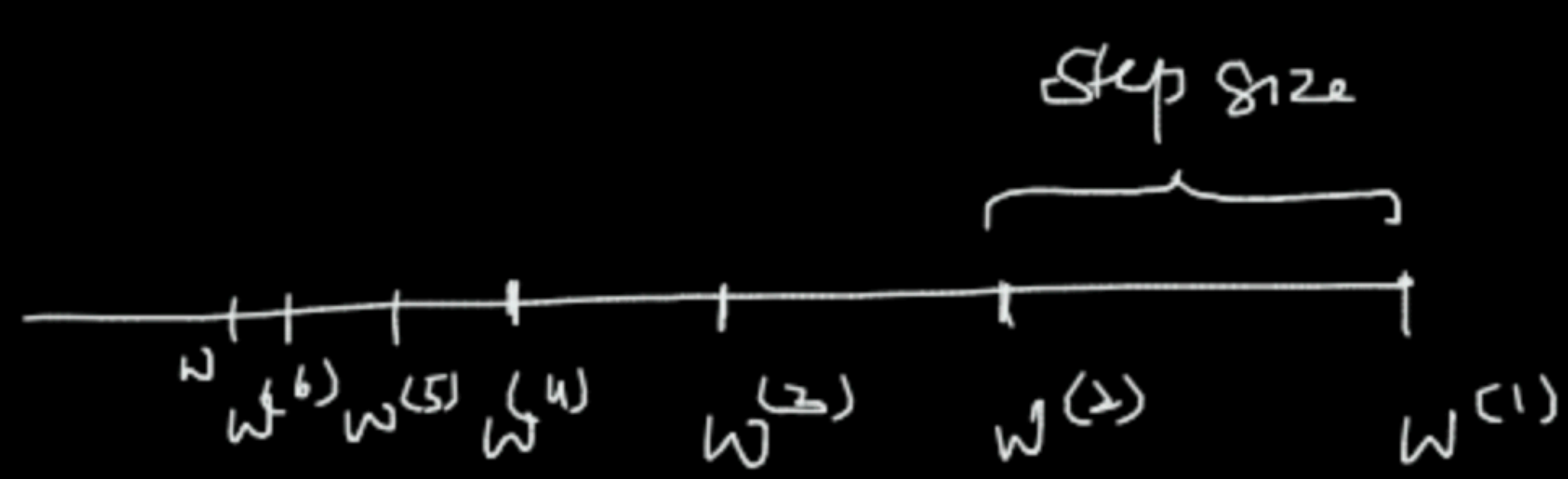
$\hat{y} = w_1 x_1 + w_2 x_2 + \quad + b$

$\frac{\partial L}{\partial w_1}$ $\frac{\partial L}{\partial w_2}$ $\frac{\partial L}{\partial w_3}$

$w_1^{new} = w_1^{old} + \lambda\left[-\frac{\partial L}{\partial w_1}\right]_{w_1^{old}}$

$w_2^{new} = w_2^{old} + \lambda\left[-\frac{\partial L}{\partial w_2}\right]_{w_2^{old}}$

$w^{new} = w^{old} \quad \underbrace{\qquad}_{\text{Step size}} \to 0$

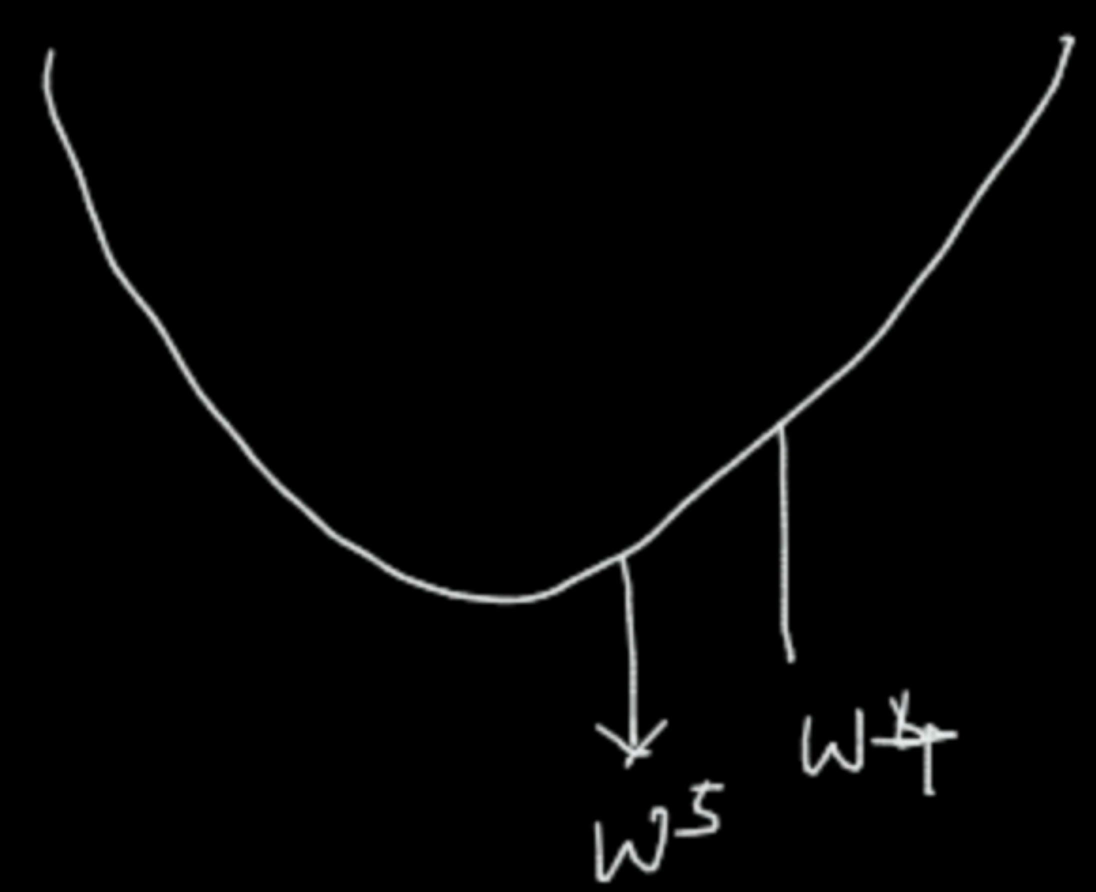


λ - learning rate

Step size $\left[\begin{matrix} \partial L / \partial w \\ \lambda \end{matrix} \right\}$ optimizer optimizers

1 Momentum — Gradient

$$w^{new} = w^{old} + \left(\frac{\partial L}{\partial w}\right)$$



2. Nestrov Momentum

$w^5 \rightarrow$

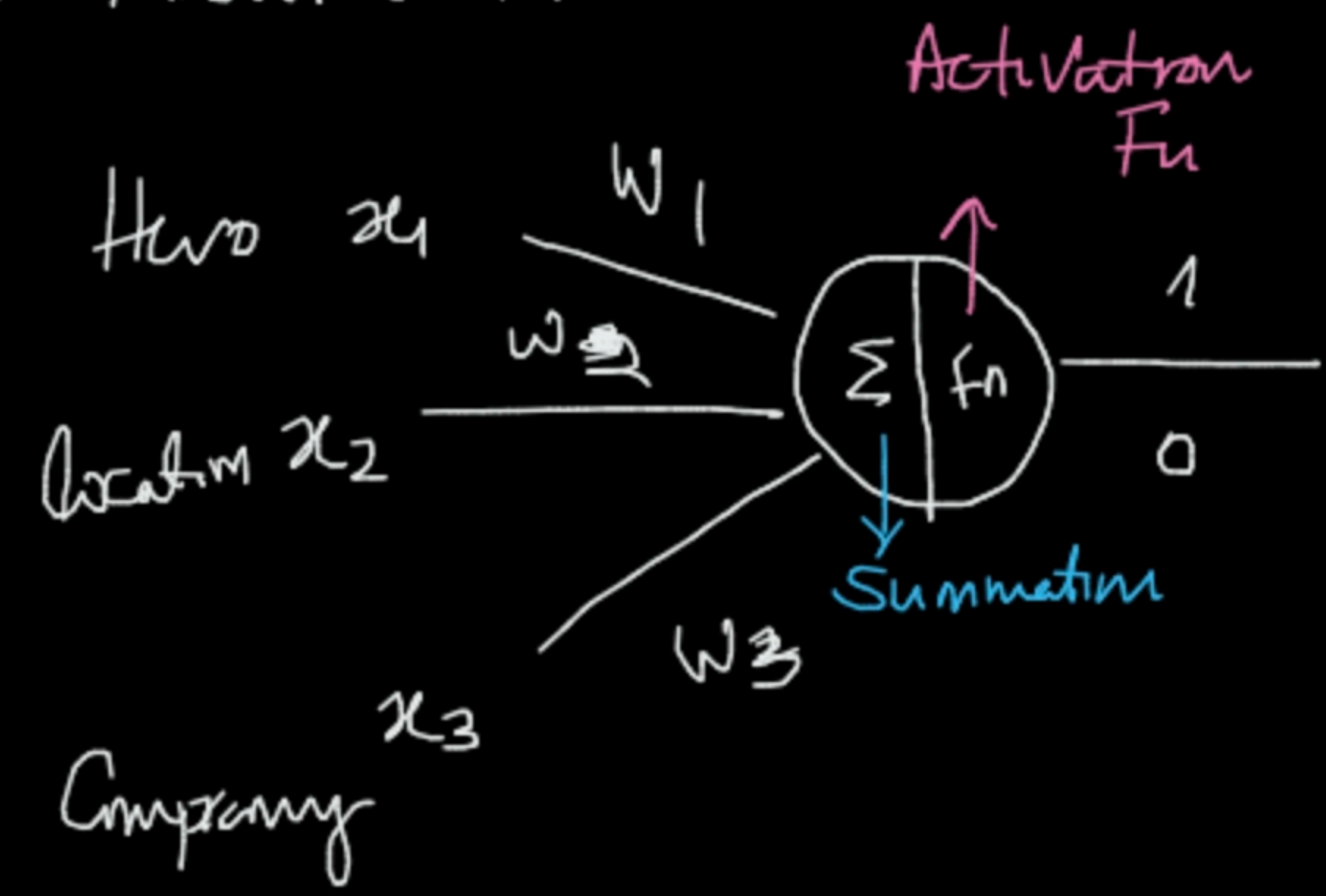
$w^6 \rightarrow$

$$w^5 = w_4^{old} - \lambda \underbrace{\left[\frac{\partial L}{\partial w_4} + \frac{\partial L}{\partial w^3} + \frac{\partial L}{\partial w^2}\right]}_{\text{Larger step size}}$$

Neural Networks

— Artificial Neurons

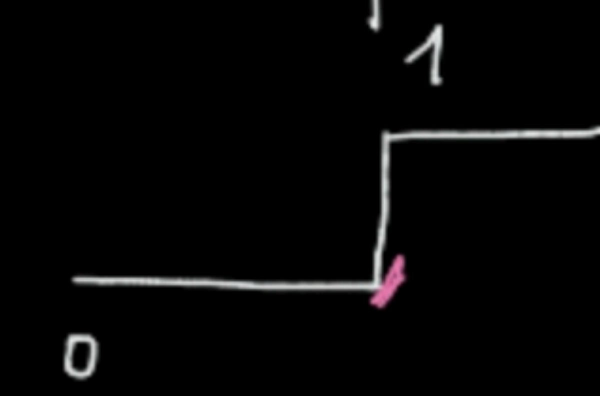
— Artificial Neural Networks



Perceptron

Activation Fn

fn → step fn



$$W_1 x_1 + W_2 x_2 + W_3 x_3 \geq b \rightarrow 1$$
$$\rightarrow 0$$

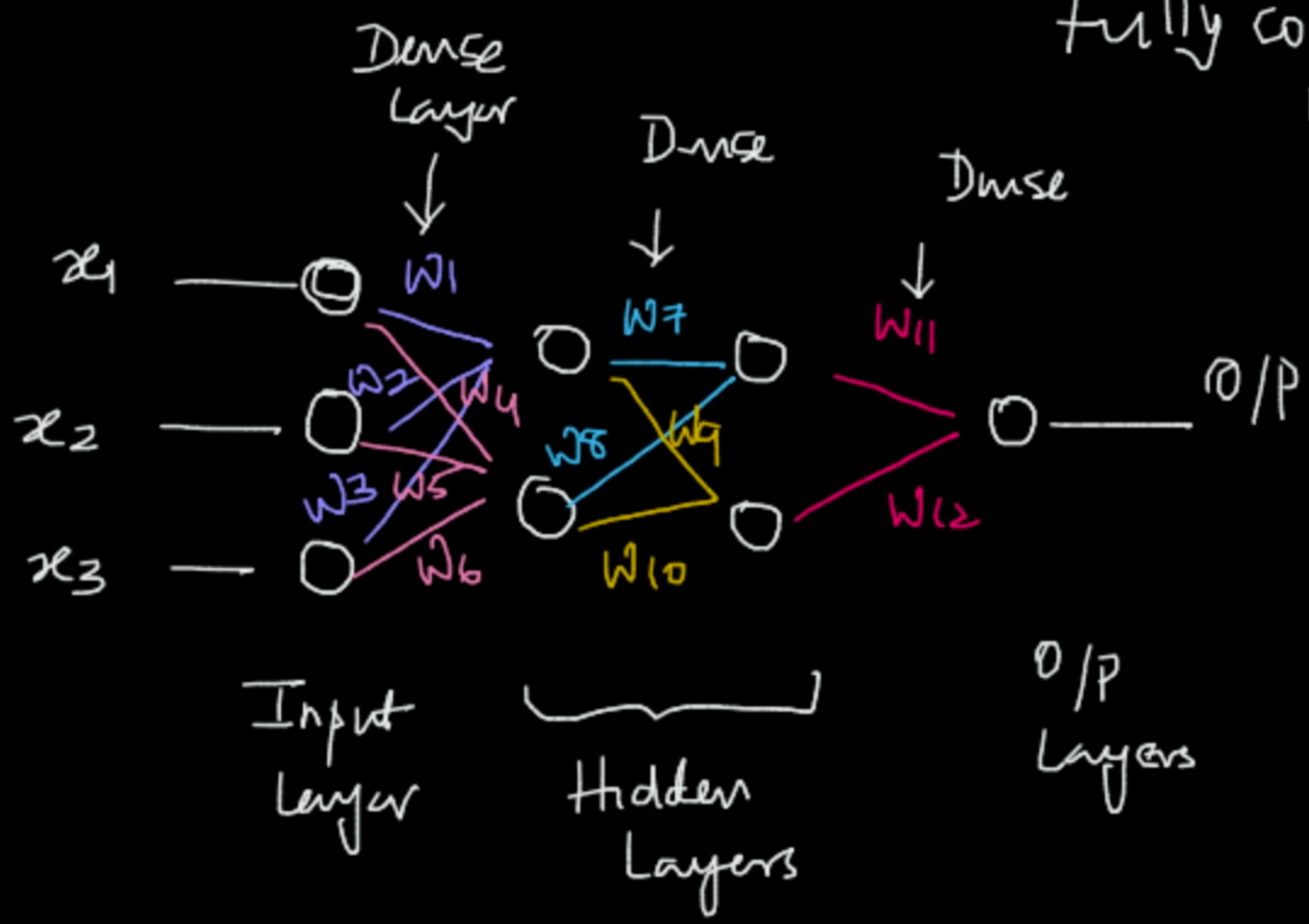
$$fn(W_1 x_1 + W_2 x_2 + W_3 x_3) \rightarrow 1$$
$$\rightarrow 0$$

↓

Activation fn → Step Fn

Fully connected / Dense Layers
Network

→ Artificial Neural Network (ANN)



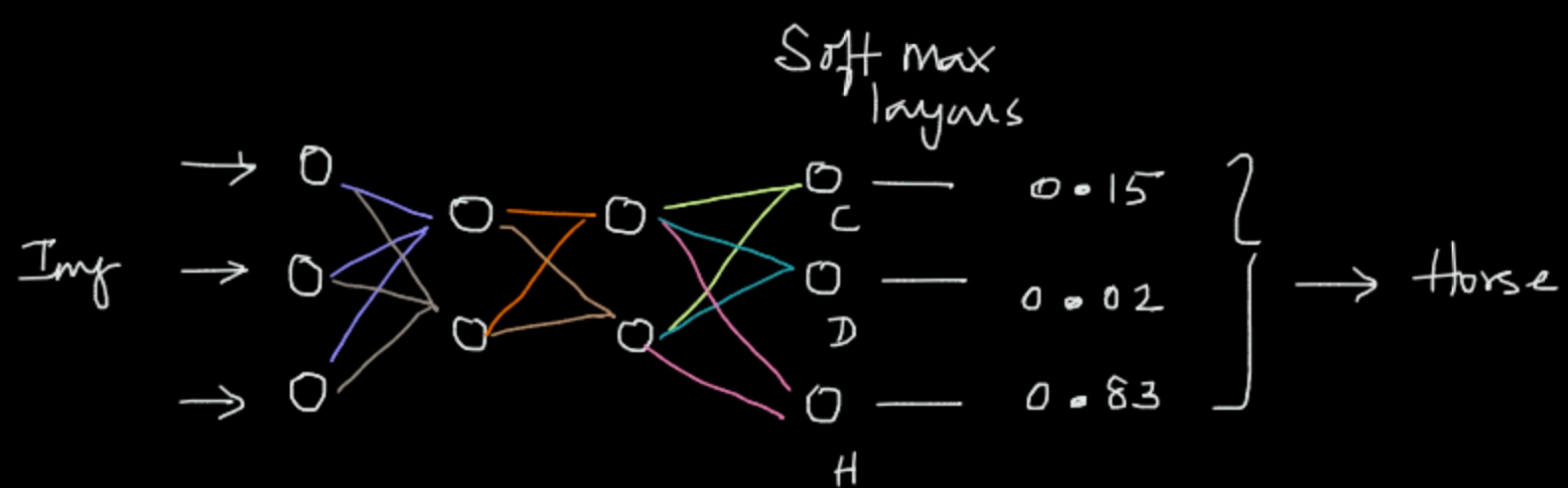
→ 12 Weights / parameters

x_1	x_2	x_3	y
			Regression
			Binary clf

Multiclass clf

→ Dog
→ Cat
→ Horse

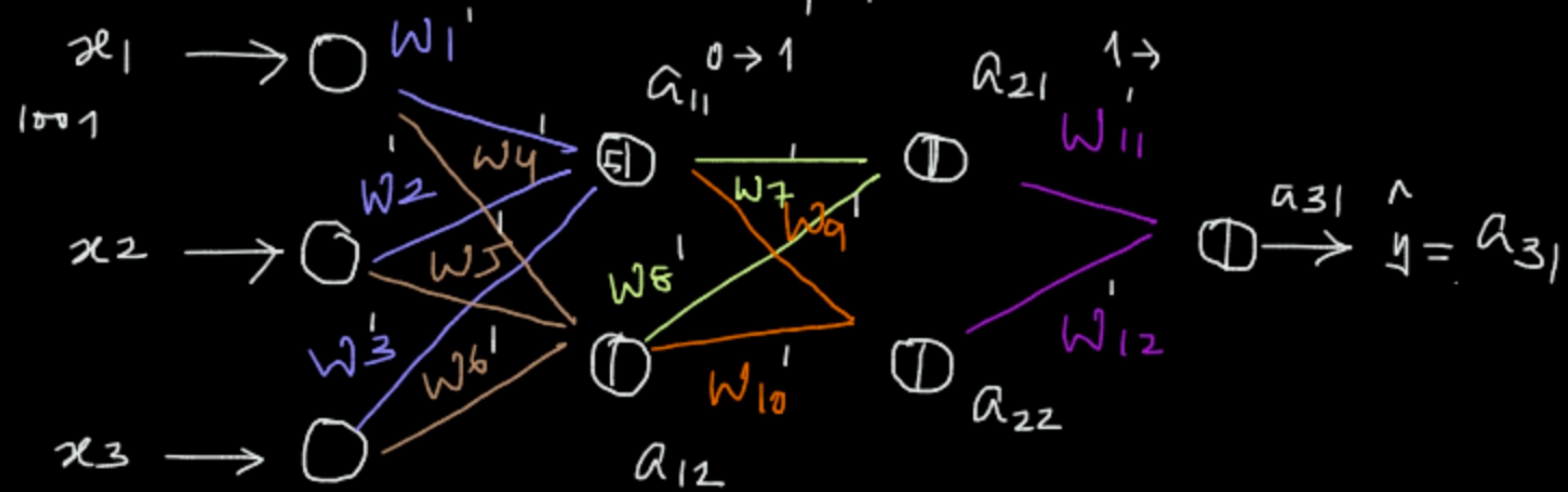
3 o/p



→ 0
→ 1
→ 2
→ 9

10 o/ps

— Forward Propagation →



← Backward propagation —

$$a_{11} = fn(w_1' x_1 + w_2' x_2 + w_3' x_3)$$

$$a_{12} = fn(w_4' x_1 + w_5' x_2 + w_6' x_3)$$

$$a_{21} = fn(w_7' a_{11} + w_8' a_{12})$$

$$a_{22} = fn(w_9' a_{11} + w_{10}' a_{12})$$

$$a_{31} = fn(w_{11}' a_{21} + w_{12}' a_{22})$$

$$\text{loss} = (y - \hat{y})^2$$

$$L = (y - a_{31})^2 \downarrow$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial L}{\partial a_{31}} \frac{\partial a_{31}}{\partial w_{11}}$$

$$\frac{\partial L}{\partial w_7} = \frac{\partial L}{\partial a_{31}} \frac{\partial a_{31}}{\partial a_{21}} \frac{\partial a_{21}}{\partial w_7}$$

— Gradient clipping

↳ Exploding gradient

$$\left(\frac{\partial L}{\partial w_1}\right) = \underset{0.2}{\overset{2}{\left(\frac{\partial L}{\partial a_{31}}\right)}} \underset{0.1}{\overset{3}{\left(\frac{\partial a_3}{\partial a_2}\right)}} \underset{0.01}{\overset{1}{\left(\frac{\partial a_{21}}{\partial a_1}\right)}} \underset{0.02}{\overset{2}{\left(\frac{\partial a_{11}}{\partial w_1}\right)}} \rightarrow 24$$

$$= 0.000004$$

↳ Vanishing Gradients

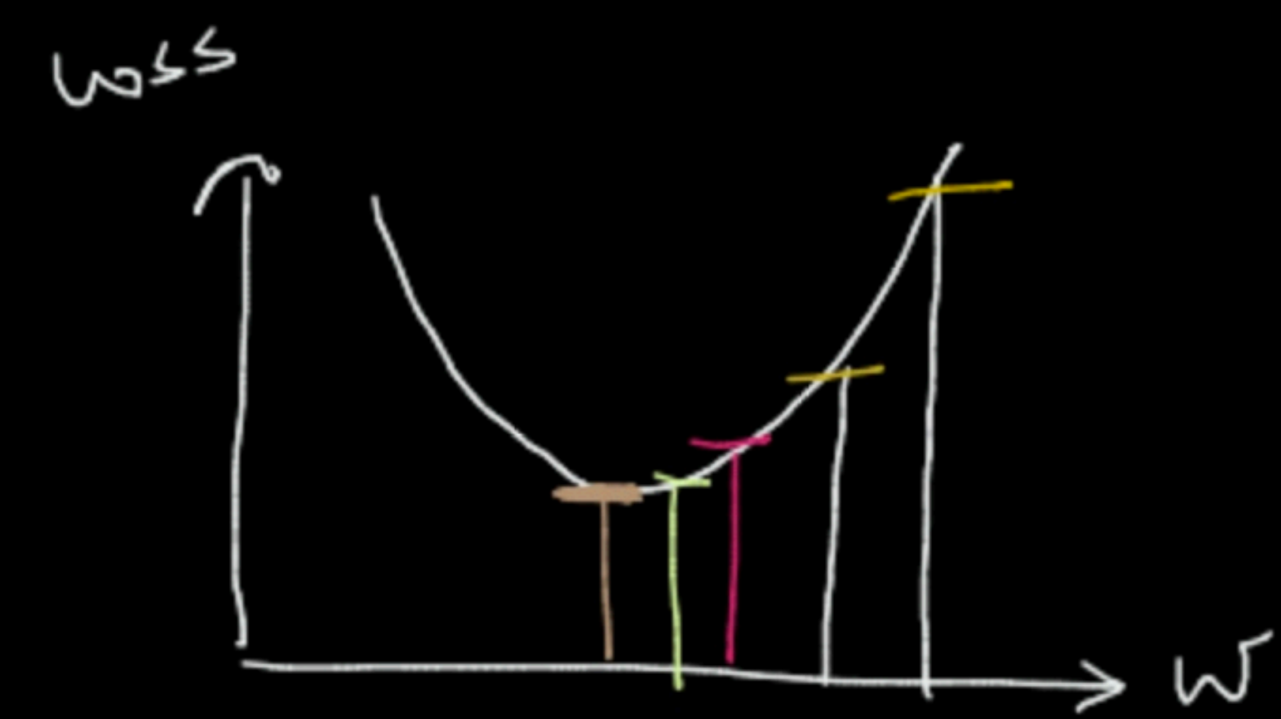
$$w_1^{(1)} \text{———} w_1^{(2)}$$

— Never Reach the min loss

Step 1 Randomly Choose 12 Weights

Step 2 Find the gradient of the loss fn w r t the weights

$$\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \frac{\partial L}{\partial w_3} \qquad \frac{\partial L}{\partial w_{12}}$$



Step 3 Simultaneously update weights

$$w_1^{new} = w_1^{old} + -\lambda \left[\frac{\partial L}{\partial w_1}\right]_{w_1^{old}}$$

$$w_2^{new} = w_2^{old} + -\lambda \left[\frac{\partial L}{\partial w_2}\right]_{w_2^{old}}$$

$$\vdots$$

$$w_{12}^{new} = w_{12}^{old} + -\lambda \left[\frac{\partial L}{\partial w_{12}}\right]_{w_{12}^{old}}$$

	x_1	x_2	x_3	y	$\hat{y}$	loss $(y-\hat{y})^2$	$\hat{y}$	loss $(y-\hat{y})^2$	$\hat{y}$	loss $(y-\hat{y})^2$	$\hat{y}$
1	1000	3	5	47	20	27^2	23	24^2	30	17^2	45

Step 4 Repeat Steps 2 & 3 until convergence
↳ all weights are Stable

6 / 6

3Blue1Brown

Composite Fn $\rightarrow$ $y = f(g(h(x)))$

$$\frac{dy}{dx} = \left(\frac{dy}{df}\right)\left(\frac{df}{dg}\right)\left(\frac{dg}{dh}\right)\left(\frac{dh}{\partial x}\right)$$

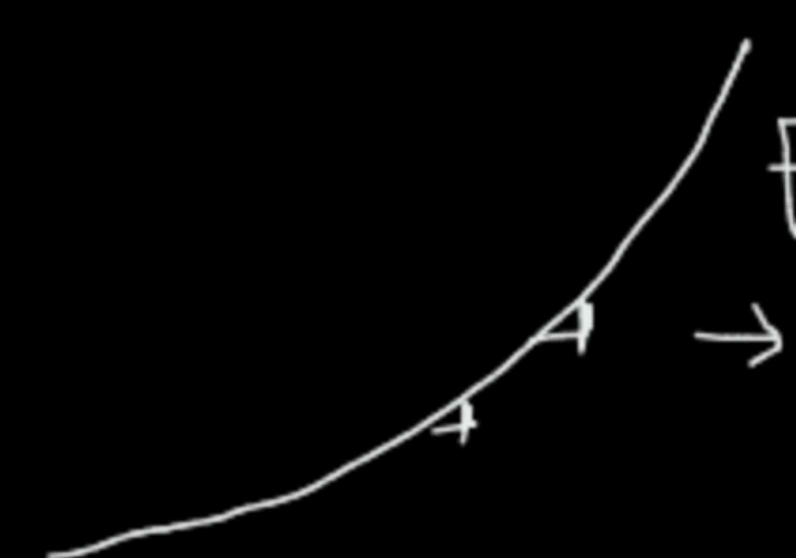
Chain Rule of differentiation

— product of multiple gradients

I/P $b-\Delta \rightarrow 0$

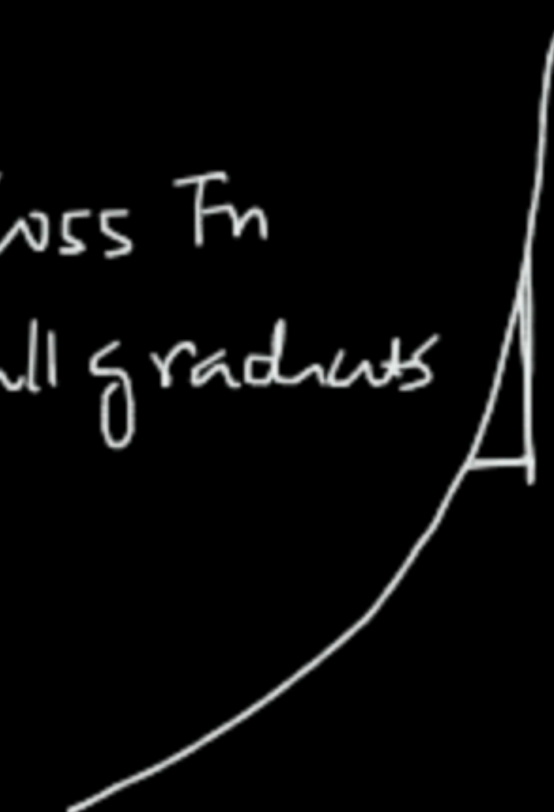
$b \rightarrow 1$

↳ sigmoid



Flat loss Fn

$\rightarrow$ small gradients



Steep loss Fn

$\rightarrow$ large gradients

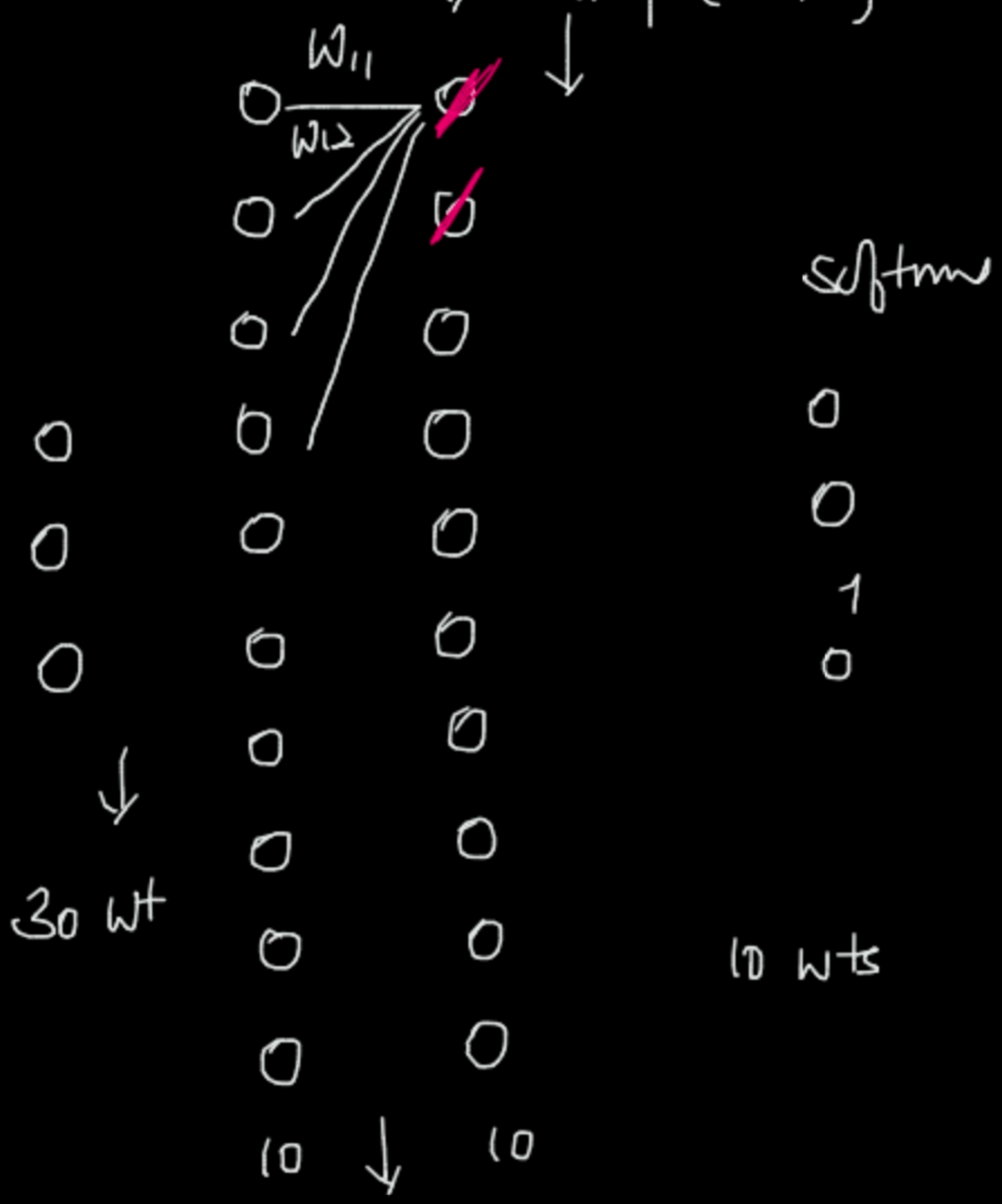
$$\underline{g(x)} = \frac{1}{1+e^{-x}}$$

Overfitting → Regularize

$$\hat{y} = w_1x_1 + w_2x_2 + \cancel{w_3x_3} \qquad + w_m x_m + b$$

drop (20%)

drop (20%)



140 wts

↳ 120 wts →

100 wts

1HL 1dL 1HL 1dL

- Activation Fn
  - Step Fn
  - Sigmoid Fn
  - tanh
  - ReLu
  - Leaky Relu
  - linear
- optimizers
  - Momentum
  - Nesterov Momentum
  - Adam
  - RMSprop
  - AdaGrad
- Weight Initialization
  - Normal
  - uniform
  - Zero
- Network
  - No of hidden layers
  - No of units in HL
- Layers
  - Dense layer
  - Soft max
  - dropout
- Regularization
  - L1 Reg
  - L2 Reg
- Loss Fn
  - Regression
    - squared loss
  - Classification
    - Binary cross entropy (2 class)
    - Categorical Cross entropy (Multiclass)

Gradient Descent
- learning rate
- Batch size
- Epochs