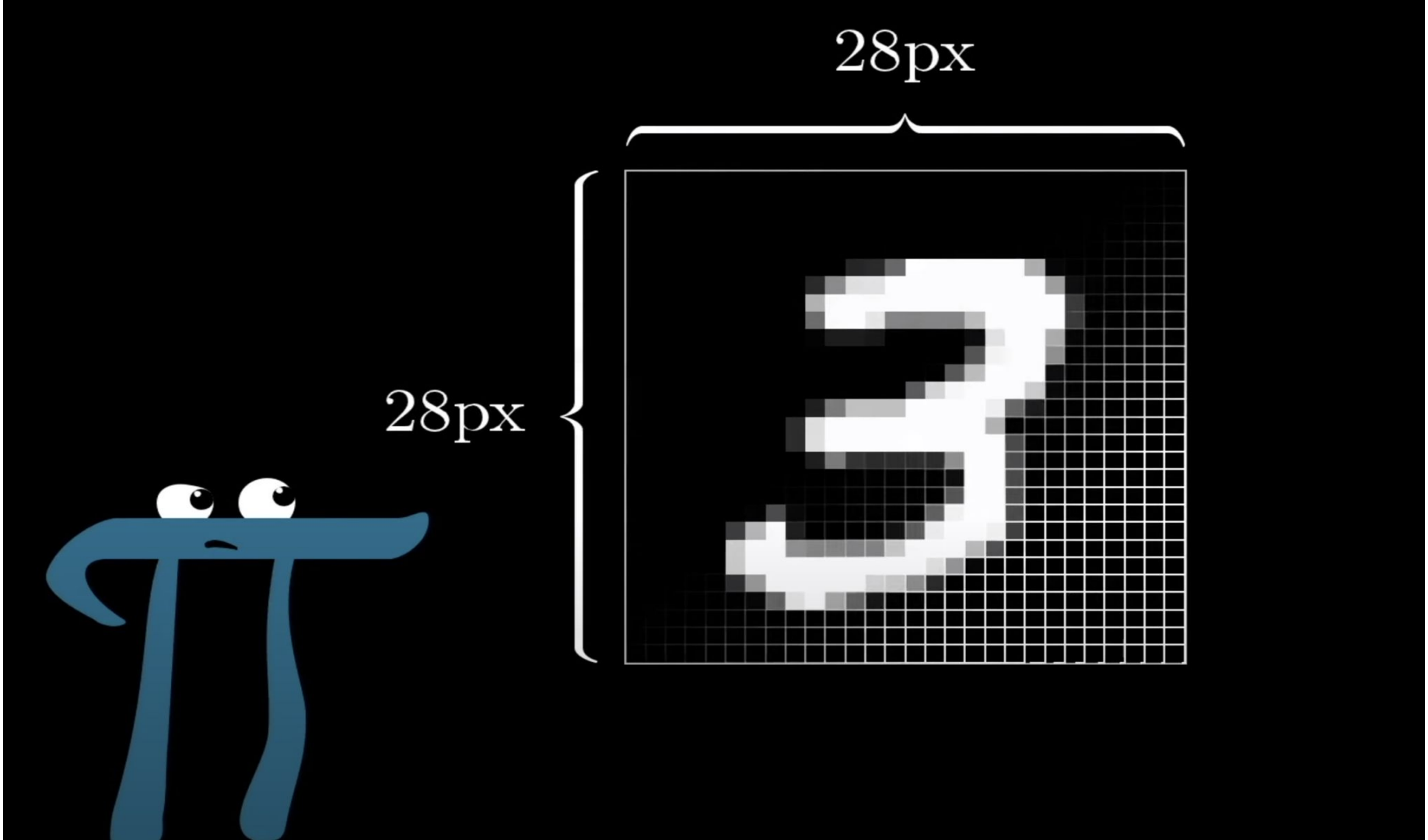
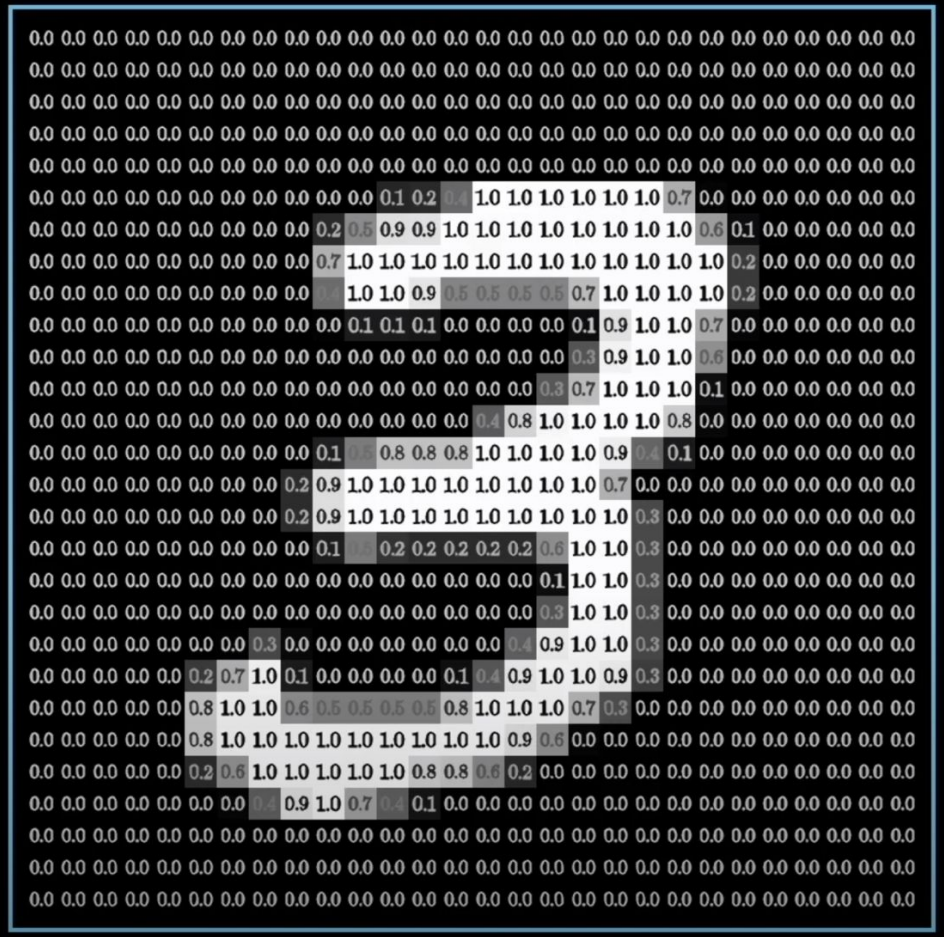
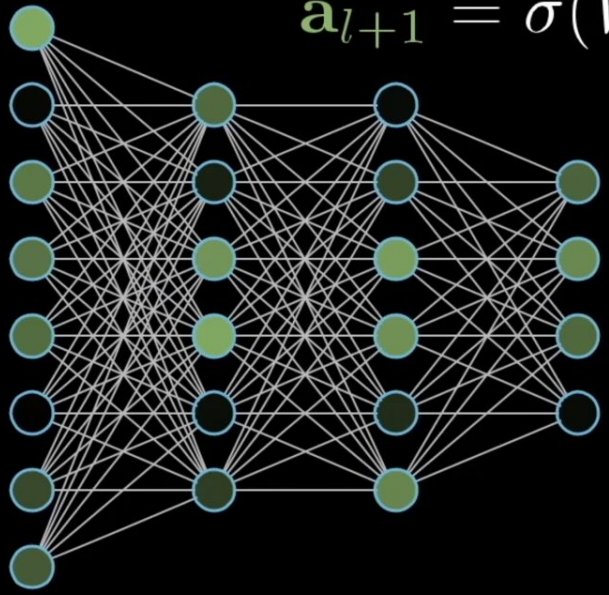

28px
28px



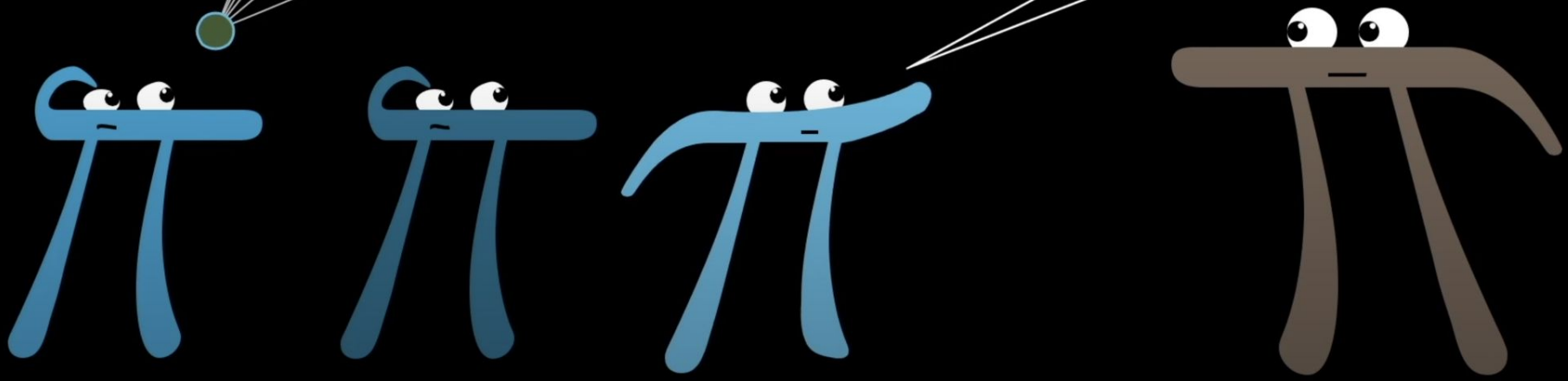
0
1
2
3 ?
4
5
6
7
8
9

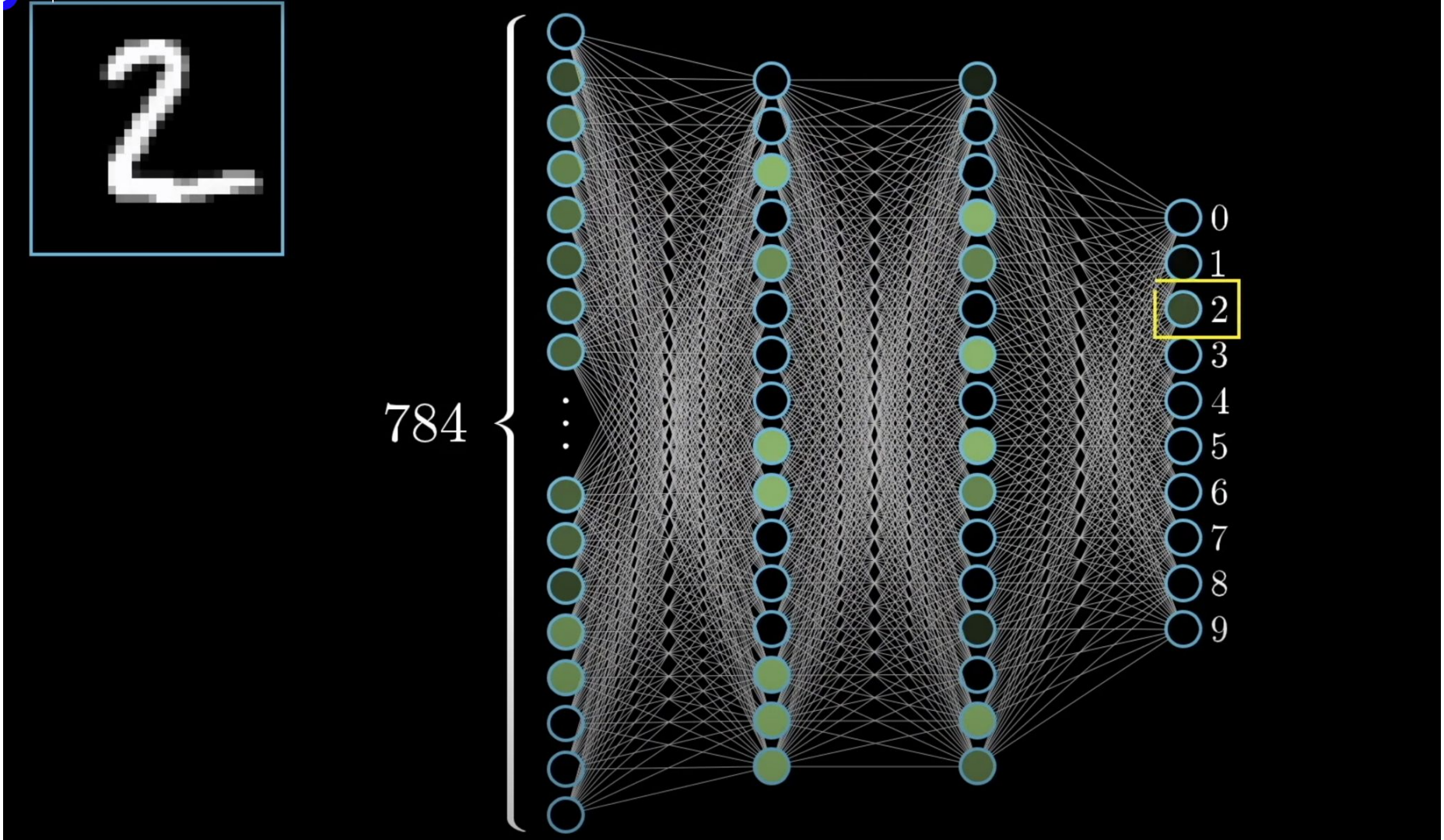
$$\mathbf{a}_{l+1} = \sigma(W_l \mathbf{a}_l + b_l)$$

Machine learning
Neural network



Why the layers?

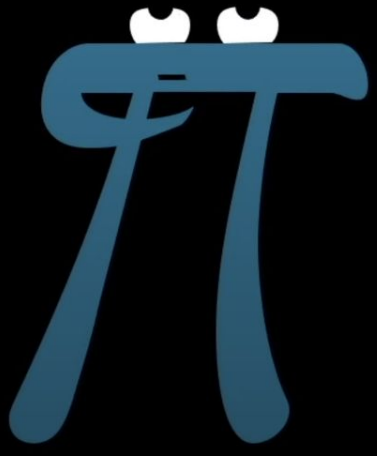



784
0
1
2
3
4
5
6
7
8
9

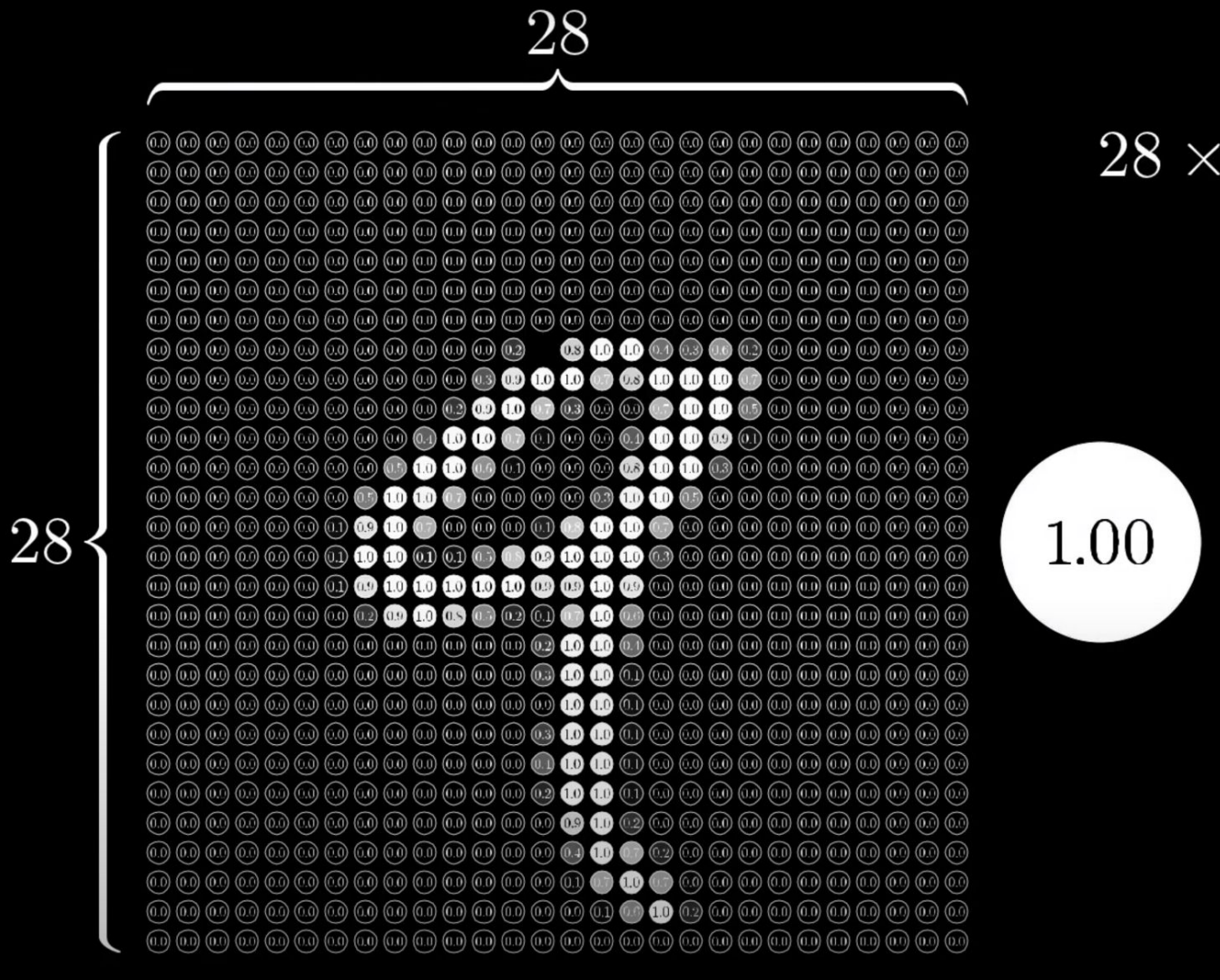
Neural network

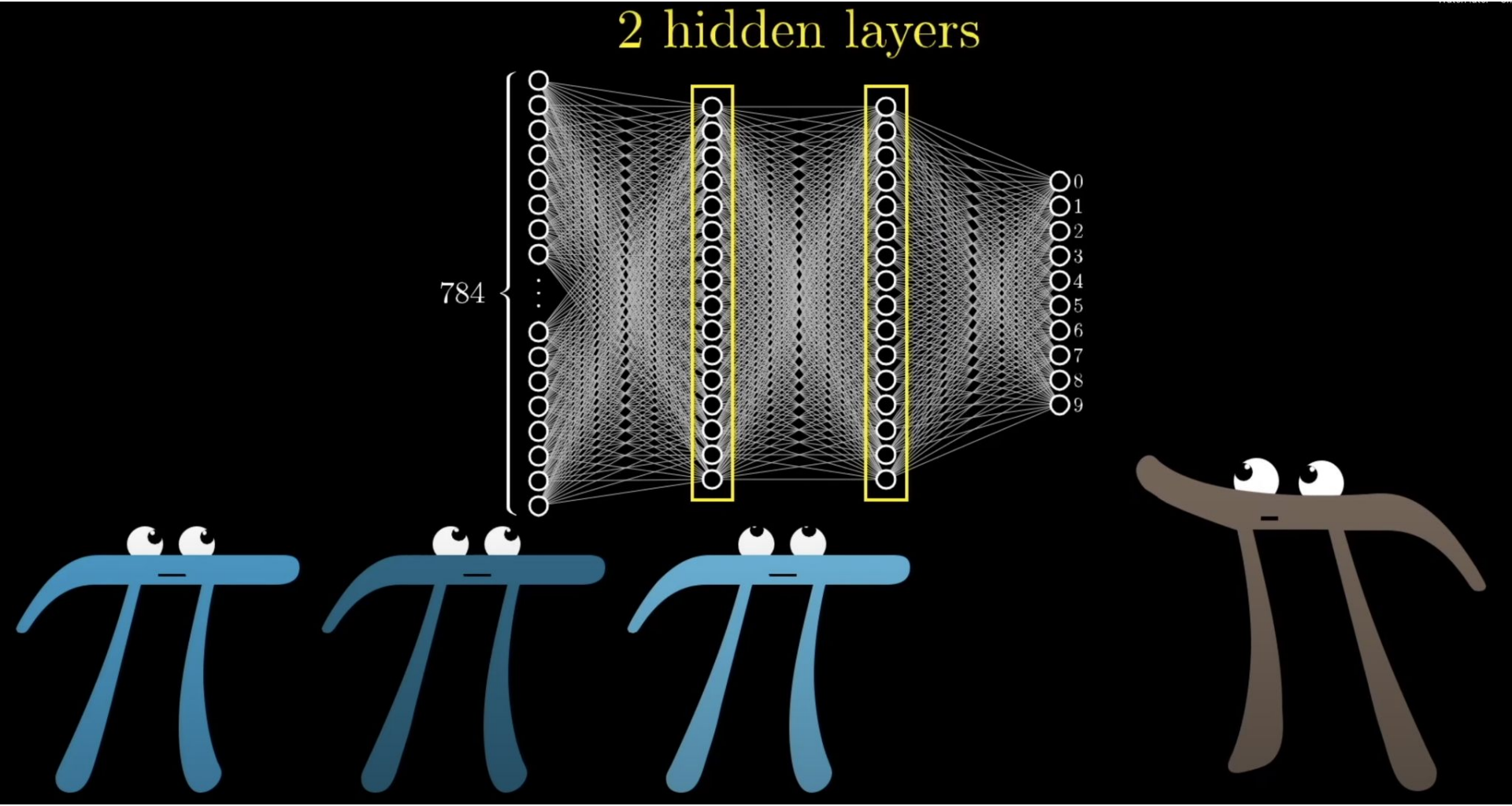
What are the neurons?

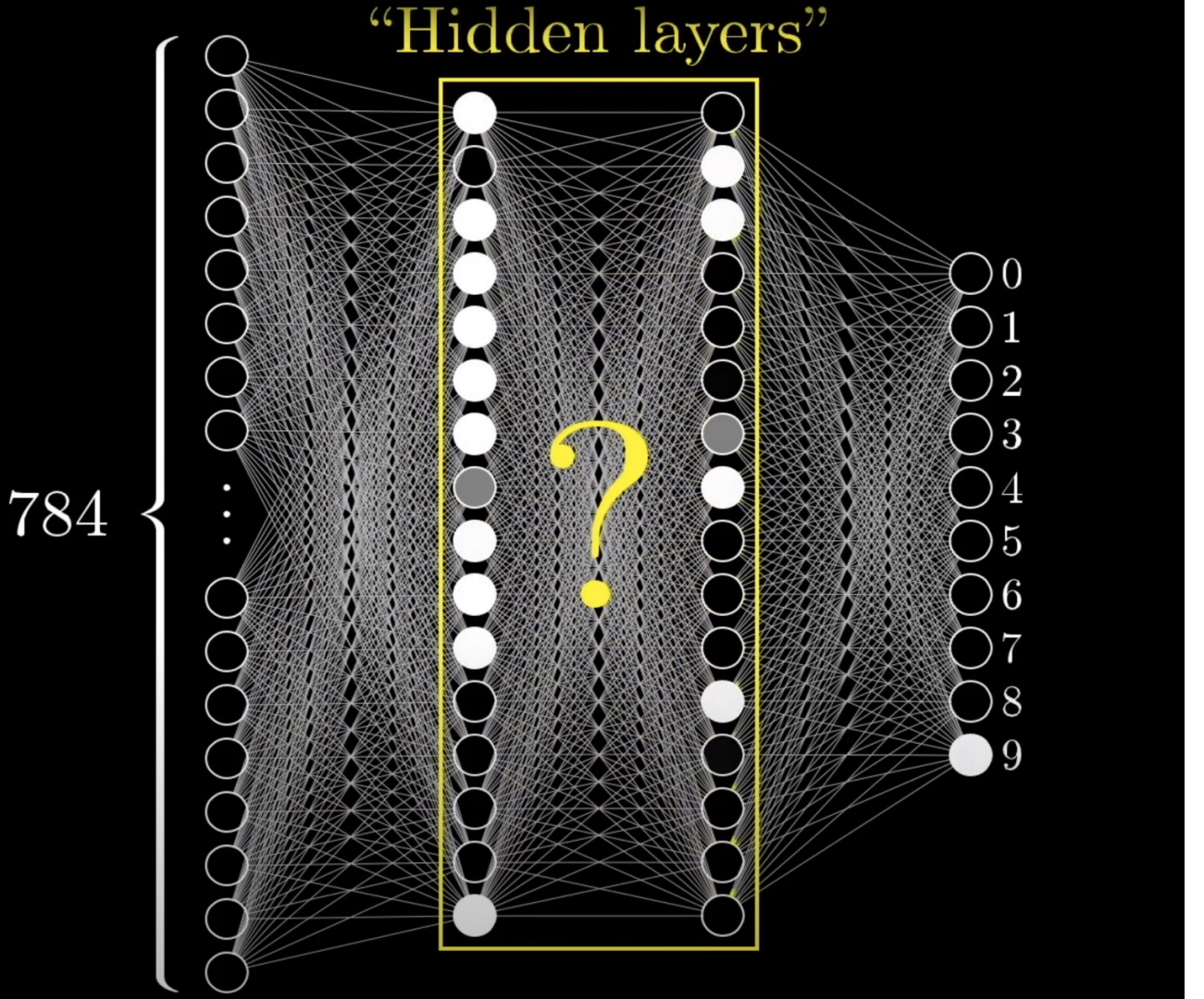
How are they connected?

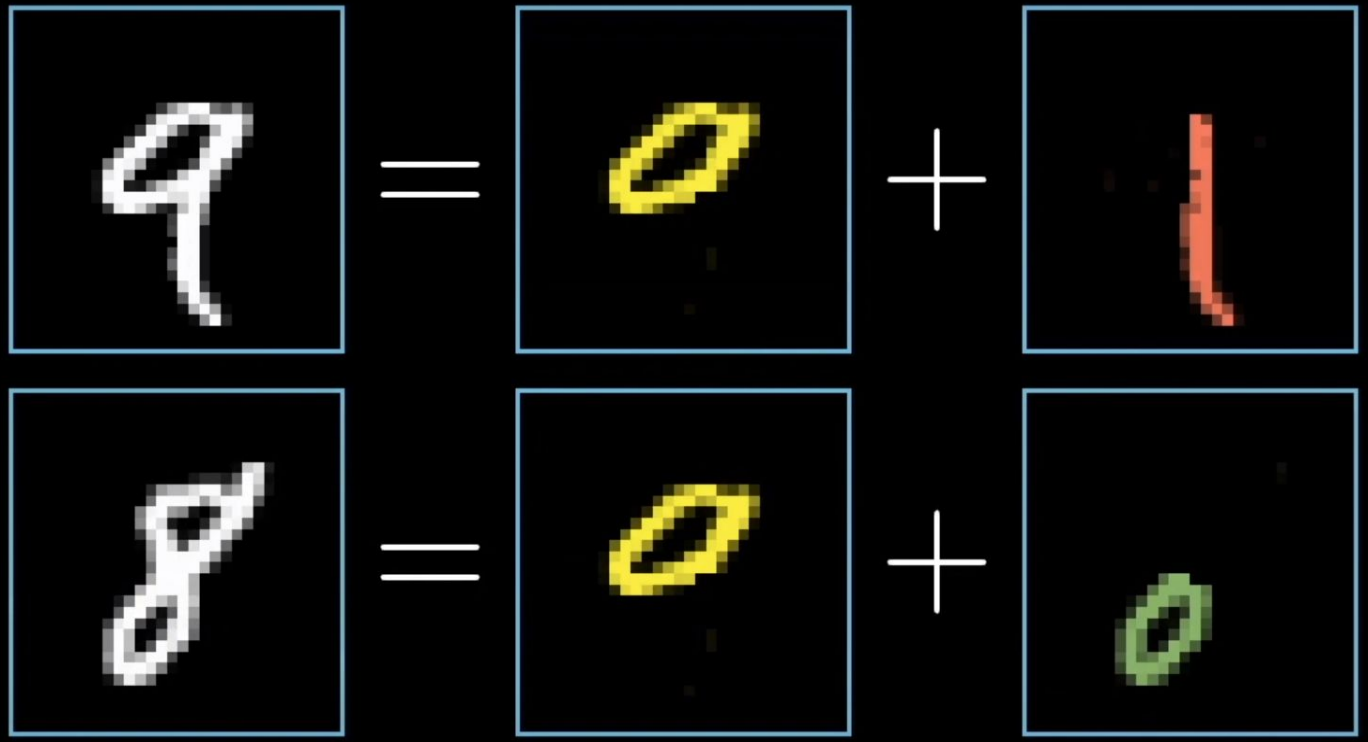


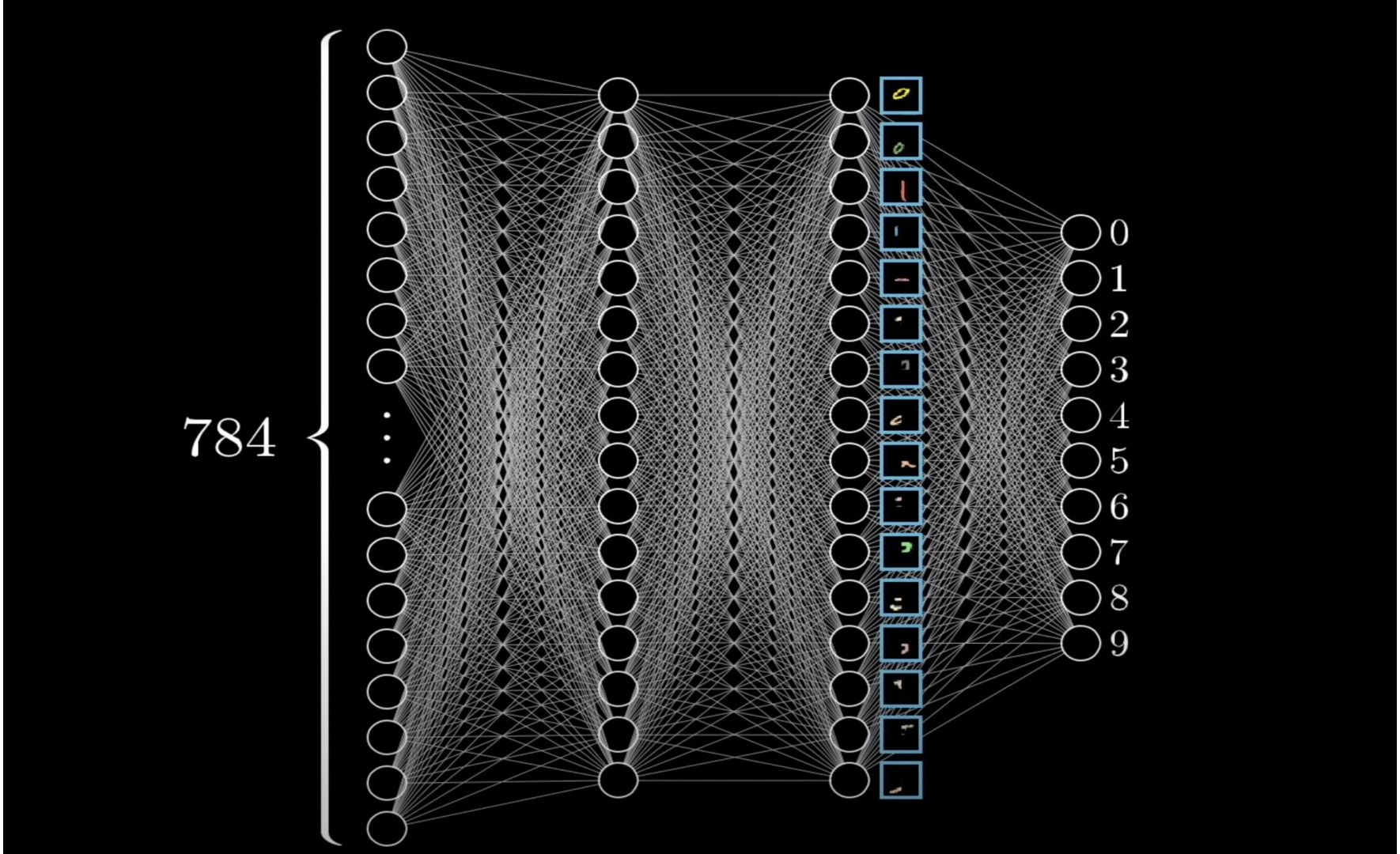
Neuron $\rightarrow$ Thing that holds a number


28
28
$28 \times 28 = 784$
1.00


2 hidden layers
784
0
1
2
3
4
5
6
7
8
9

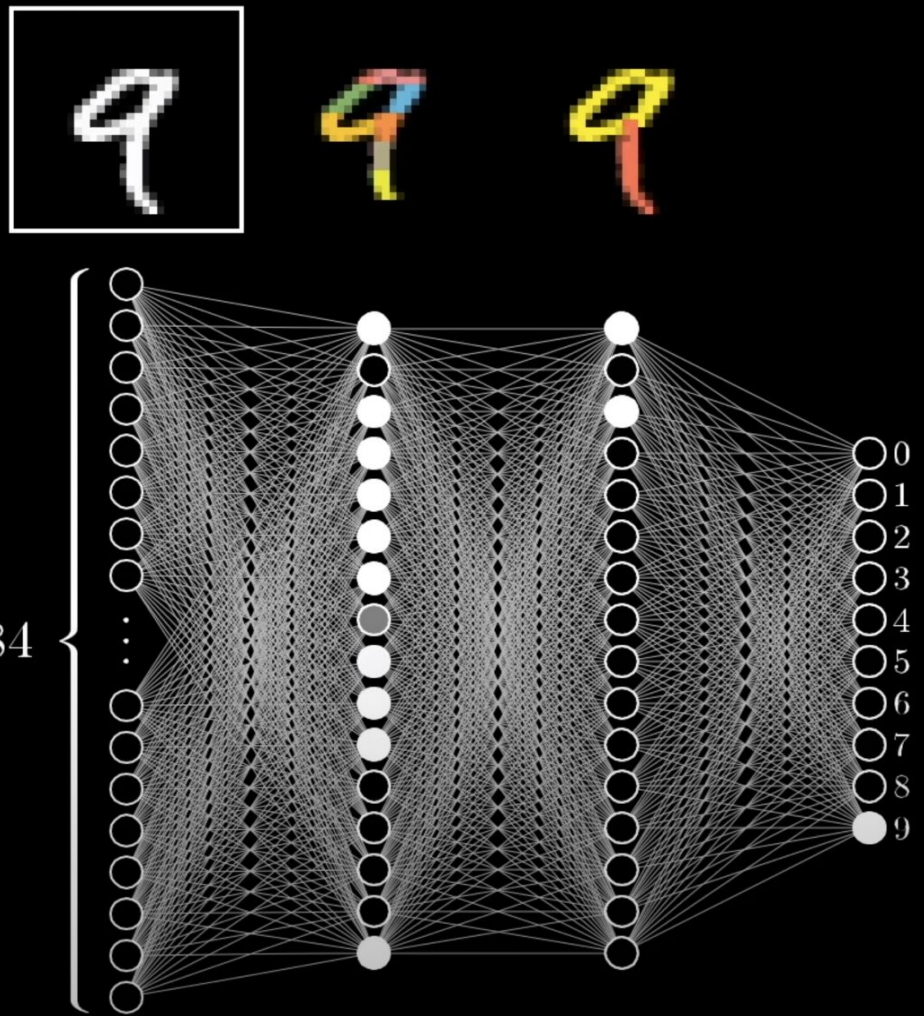

"Hidden layers"
784
?
0
1
2
3
4
5
6
7
8
9

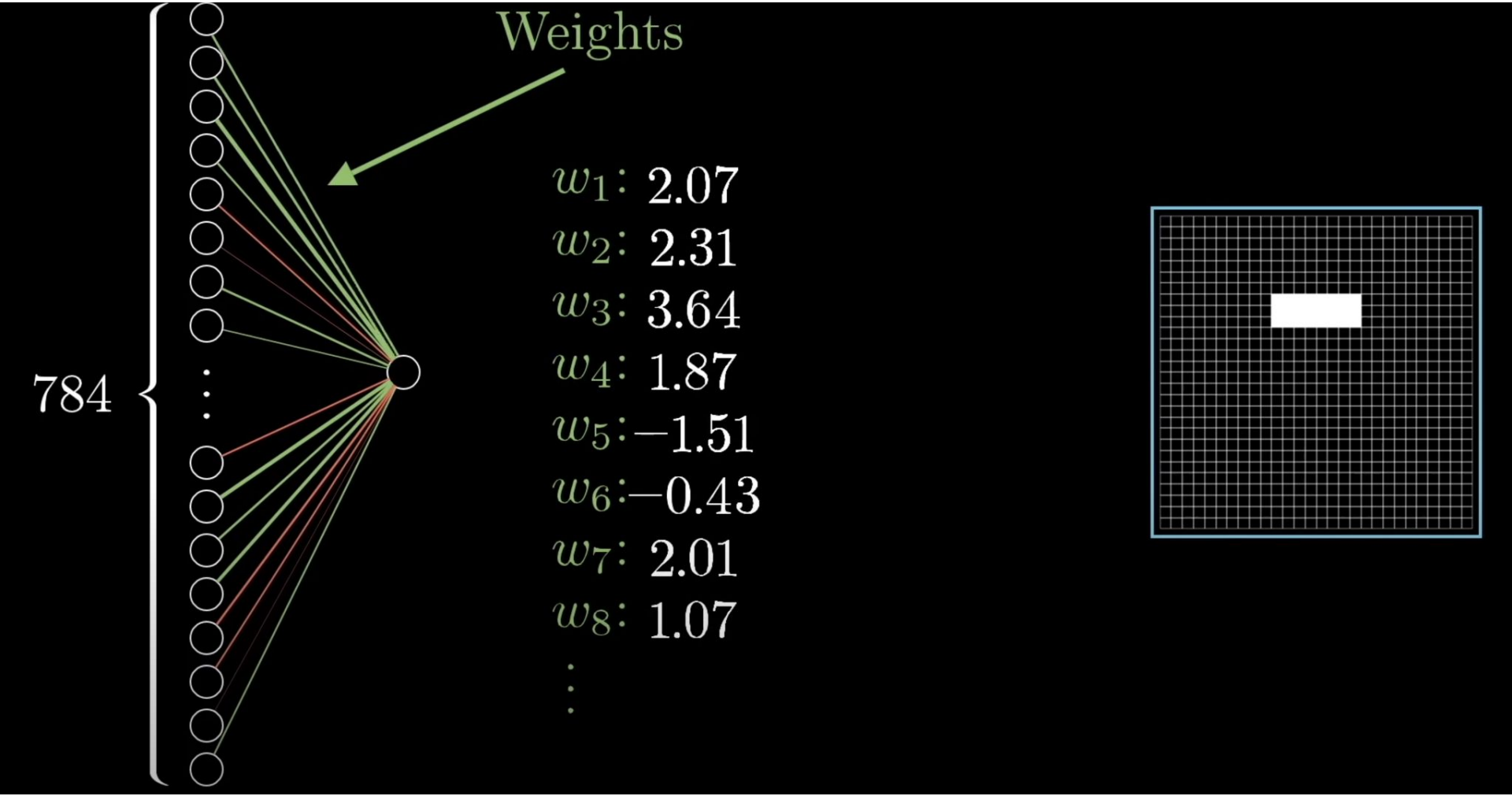


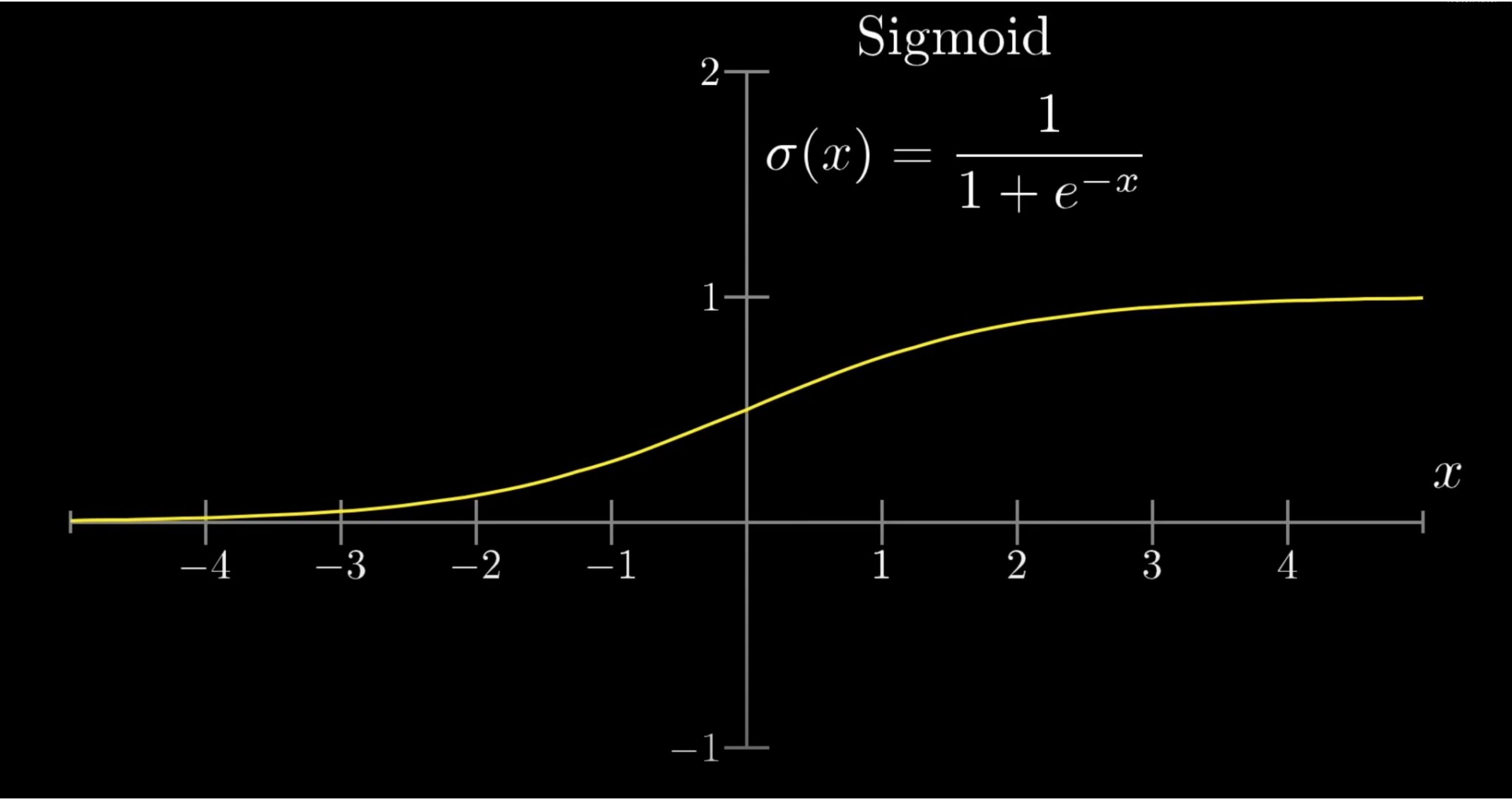

784
0
1
2
3
4
5
6
7
8
9

We'll get back to this

Does the network actually do this?




Weights
784
w_1: 2.07
w_2: 2.31
w_3: 3.64
w_4: 1.87
w_5: −1.51
w_6: −0.43
w_7: 2.01
w_8: 1.07

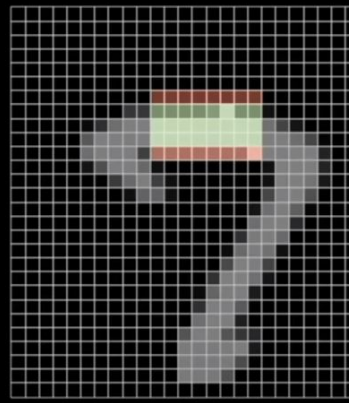

Sigmoid
$\sigma(x) = \frac{1}{1 + e^{-x}}$
2
1
−1
x
−4
−3
−2
−1
1
2
3
4

Sigmoid

How positive is this?

$$\sigma(w_1 a_1 + w_2 a_2 + w_3 a_3 + \cdots + w_n a_n)$$

784



Sigmoid

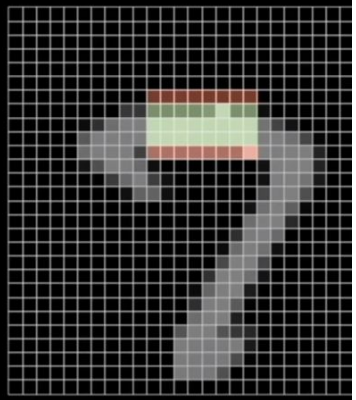
How positive is this?

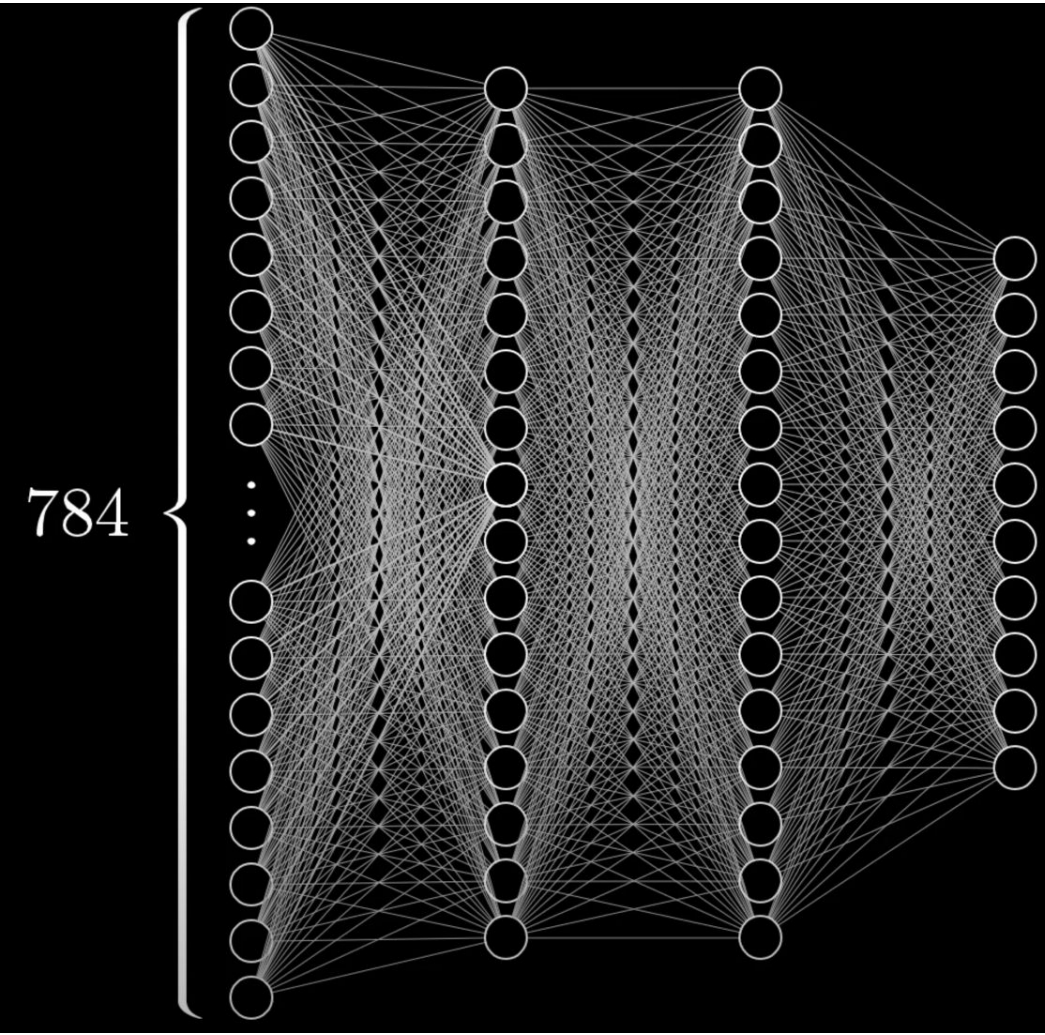
$$\sigma(w_1 a_1 + w_2 a_2 + w_3 a_3 + \cdots + w_n a_n \boxed{-10})$$

"bias"

784

Only activate meaningfully when weighted sum > 10



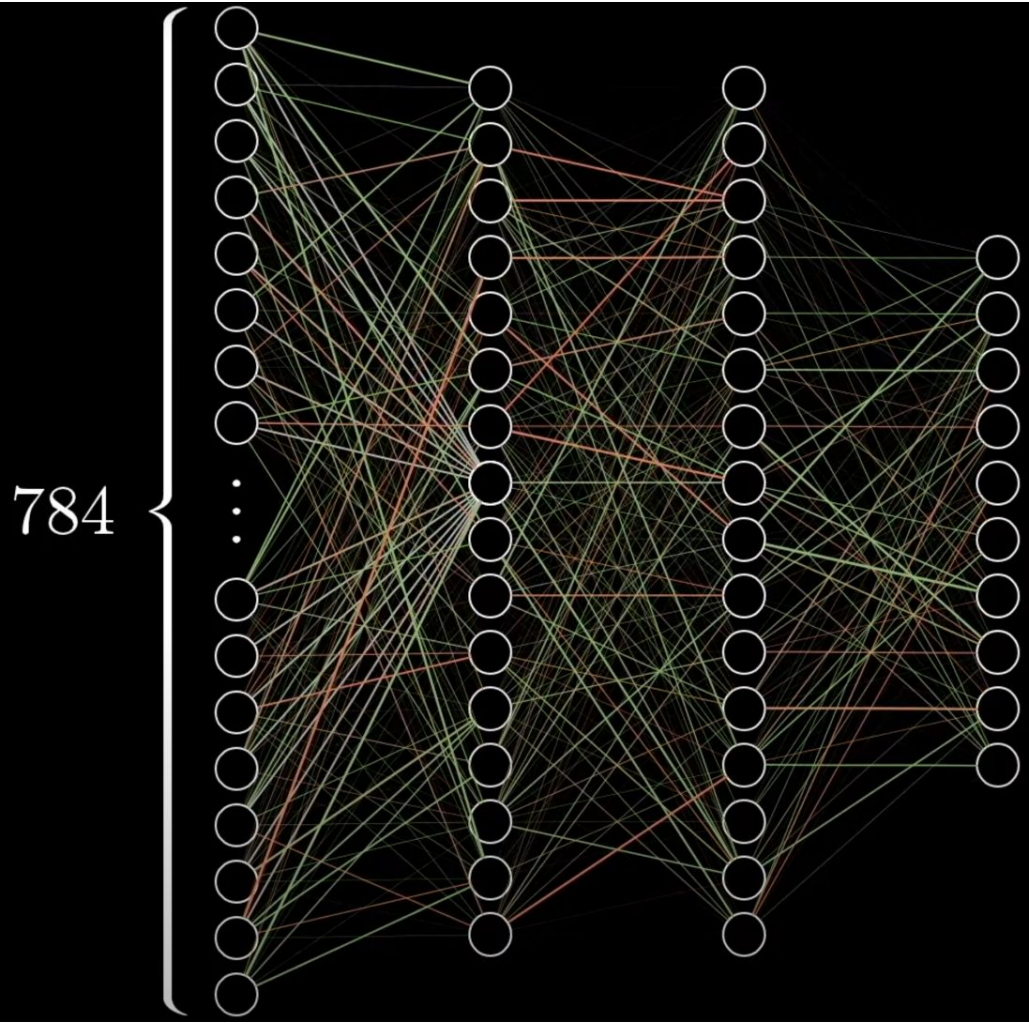


$$784 \times 16 + 16 \times 16 + 16 \times 10$$

weights

$$16 + 16 + 10$$

biases

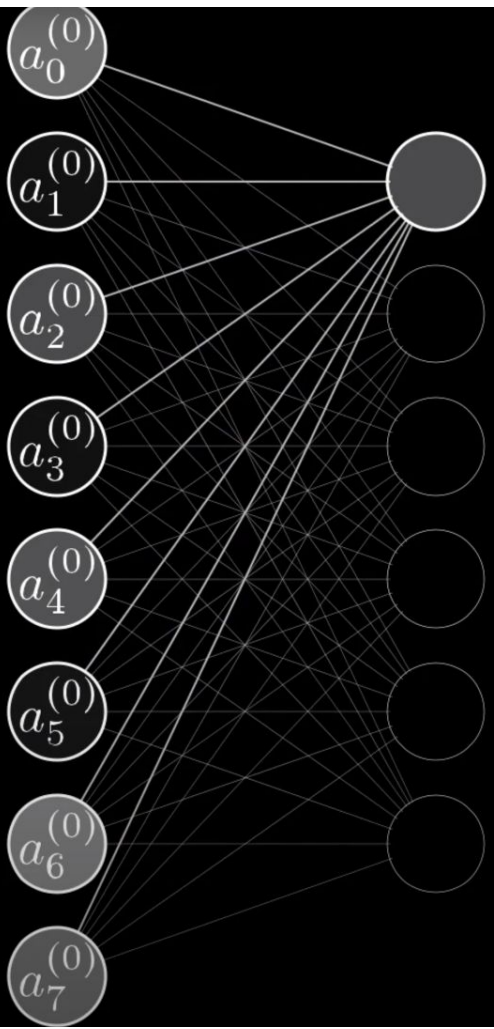


$784 \times 16 + 16 \times 16 + 16 \times 10$
weights

$16 + 16 + 10$
biases

13,002

Learning → Finding the right weights and biases

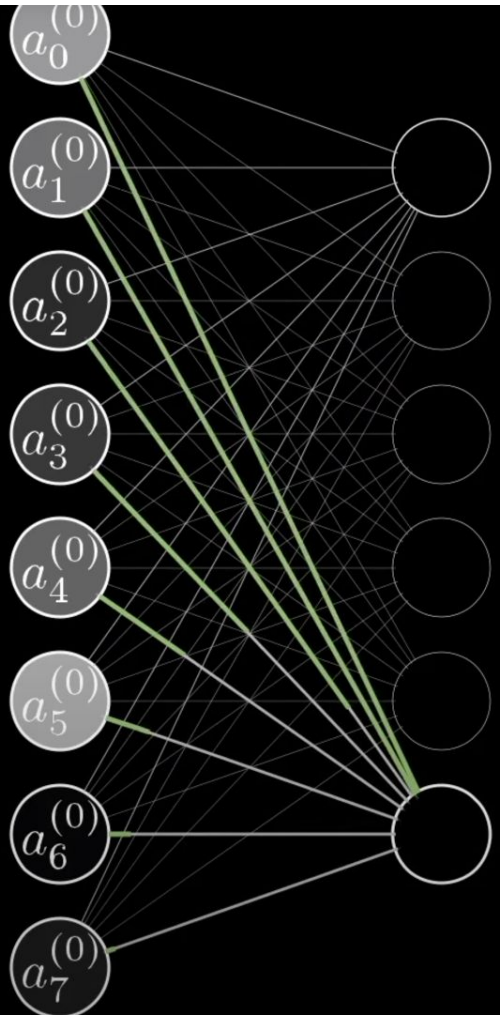


Sigmoid

$$a_0^{(1)} = \sigma\left(w_{0,0}\ a_0^{(0)} + w_{0,1}\ a_1^{(0)} + \cdots + w_{0,n}\ a_n^{(0)} + b_0 \right)$$

Bias

$$\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix}$$

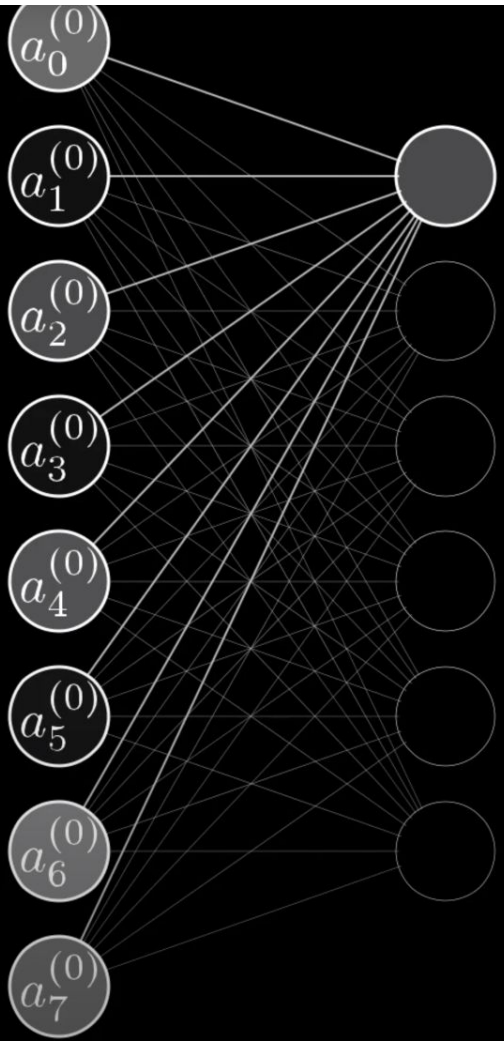


Sigmoid

$$a_0^{(1)} = \sigma\left(w_{0,0}\, a_0^{(0)} + w_{0,1}\, a_1^{(0)} + \cdots + w_{0,n}\, a_n^{(0)} + b_0 \right)$$

Bias

$$\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} = \begin{bmatrix} ? \\ ? \\ \vdots \\ ? \end{bmatrix}$$

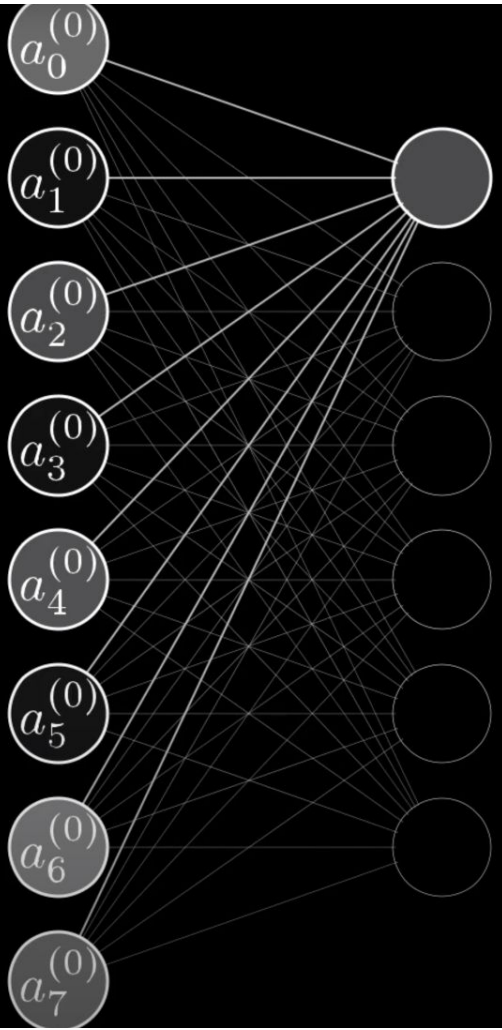


Sigmoid

$$a_0^{(1)} = \sigma\left(w_{0,0}\, a_0^{(0)} + w_{0,1}\, a_1^{(0)} + \cdots + w_{0,n}\, a_n^{(0)} + b_0 \right)$$

Bias

$$\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

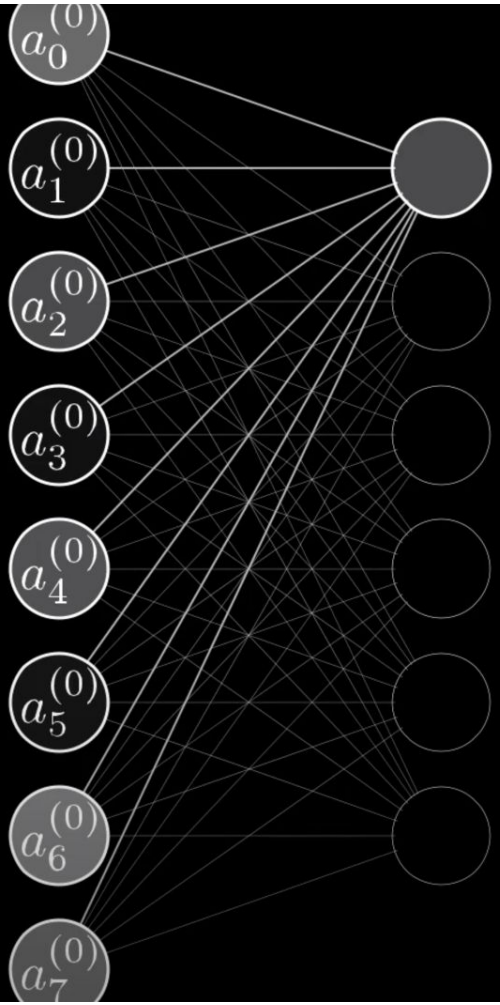


Sigmoid

$$a_0^{(1)} = \sigma\left(w_{0,0}\, a_0^{(0)} + w_{0,1}\, a_1^{(0)} + \cdots + w_{0,n}\, a_n^{(0)} + b_0 \right)$$

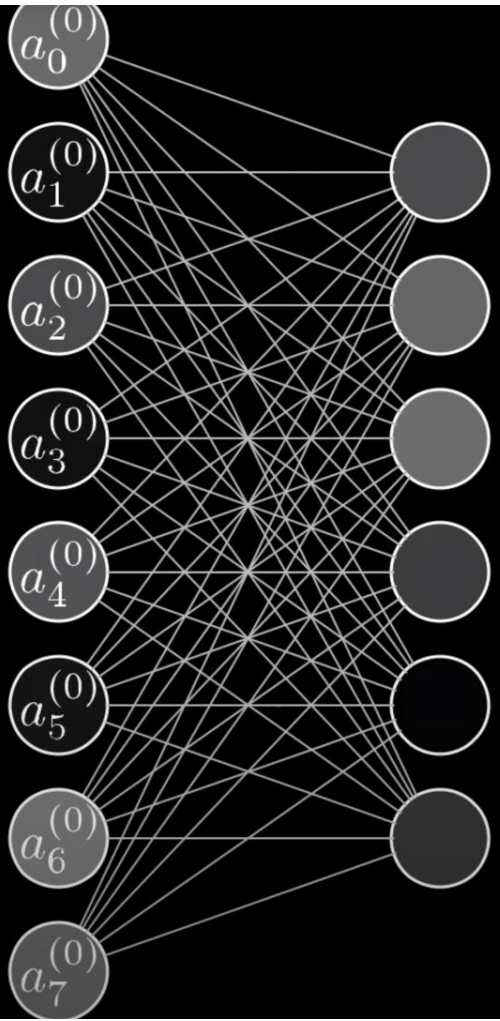
Bias

$$\sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \right)$$



$$\sigma\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \begin{bmatrix} \sigma(x) \\ \sigma(y) \\ \sigma(z) \end{bmatrix}$$

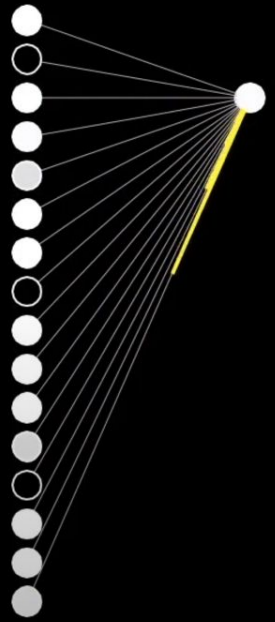
$$\sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}\right)$$



$$\mathbf{a}^{(1)} = \sigma(\mathbf{W}\mathbf{a}^{(0)} + \mathbf{b})$$

$$\sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}\right)$$

Neuron



Function

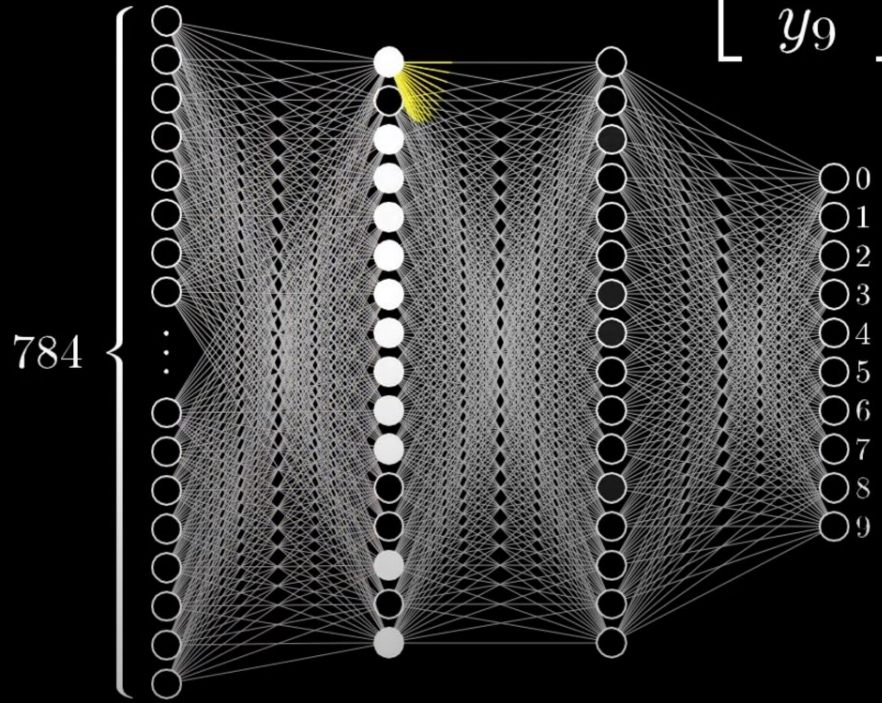
~~Thing that holds a number~~

$$f(a_0, \ldots, a_{783}) = \begin{bmatrix} y_0 \\ \vdots \\ y_9 \end{bmatrix}$$

Network

↓

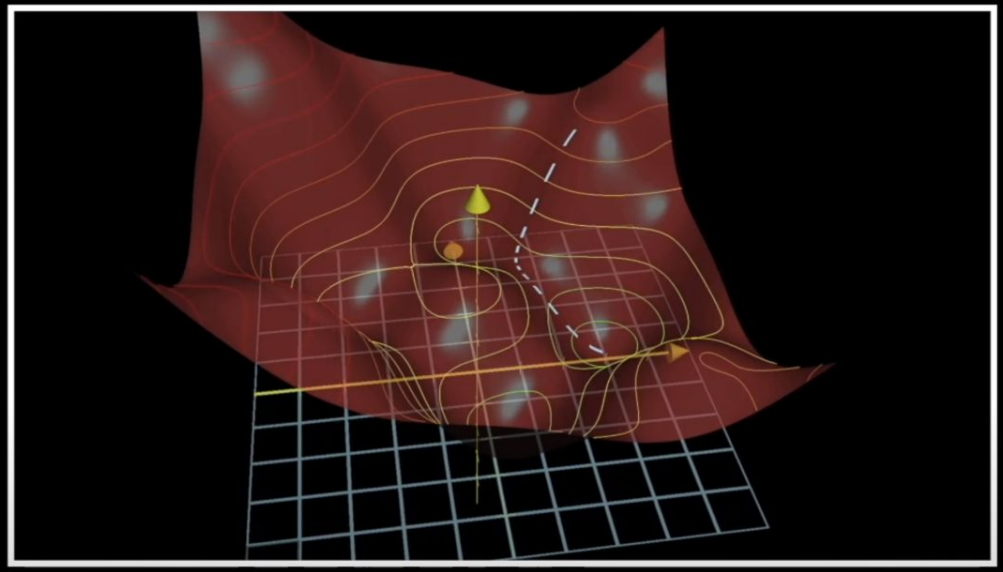
Function

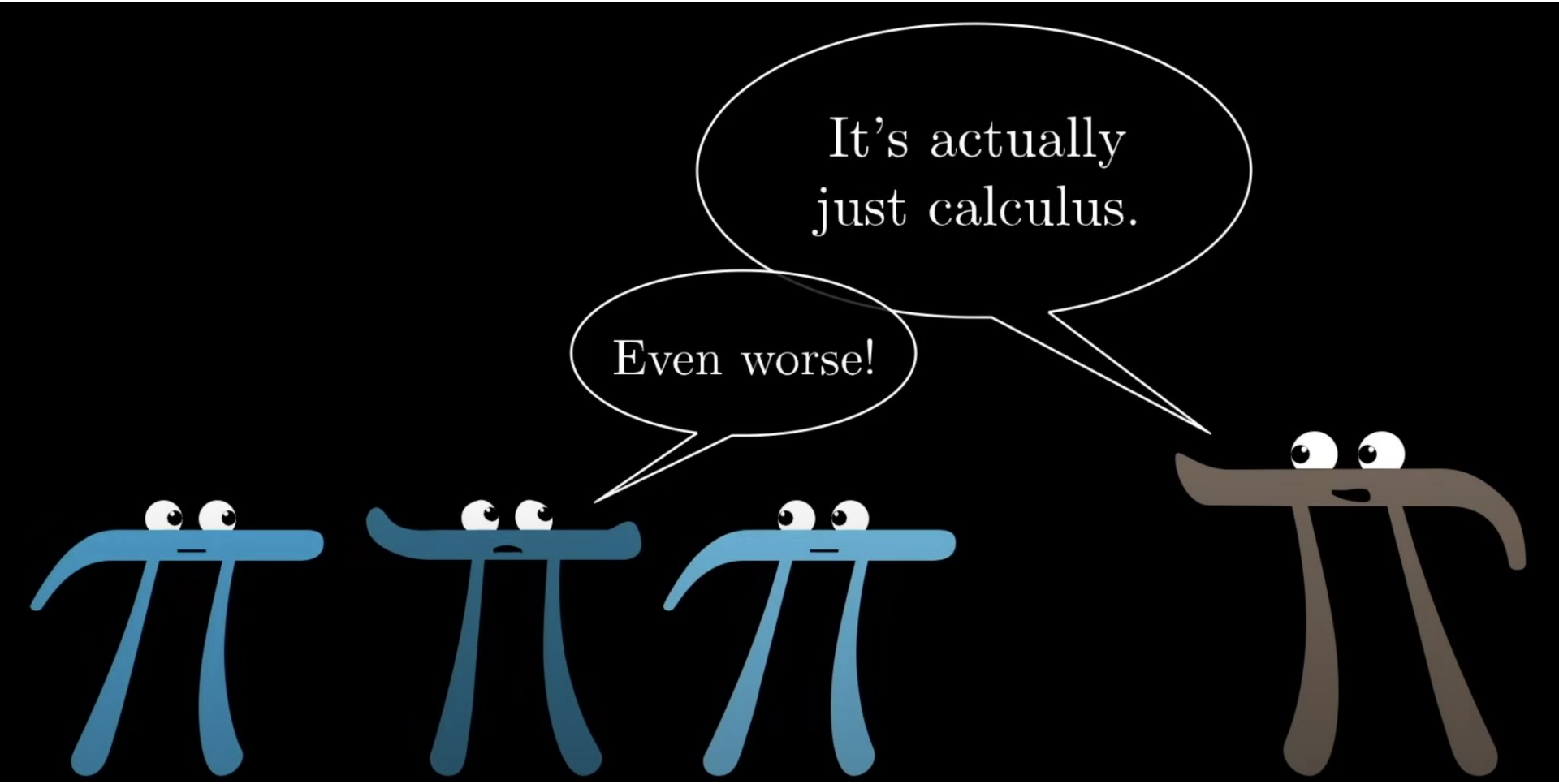


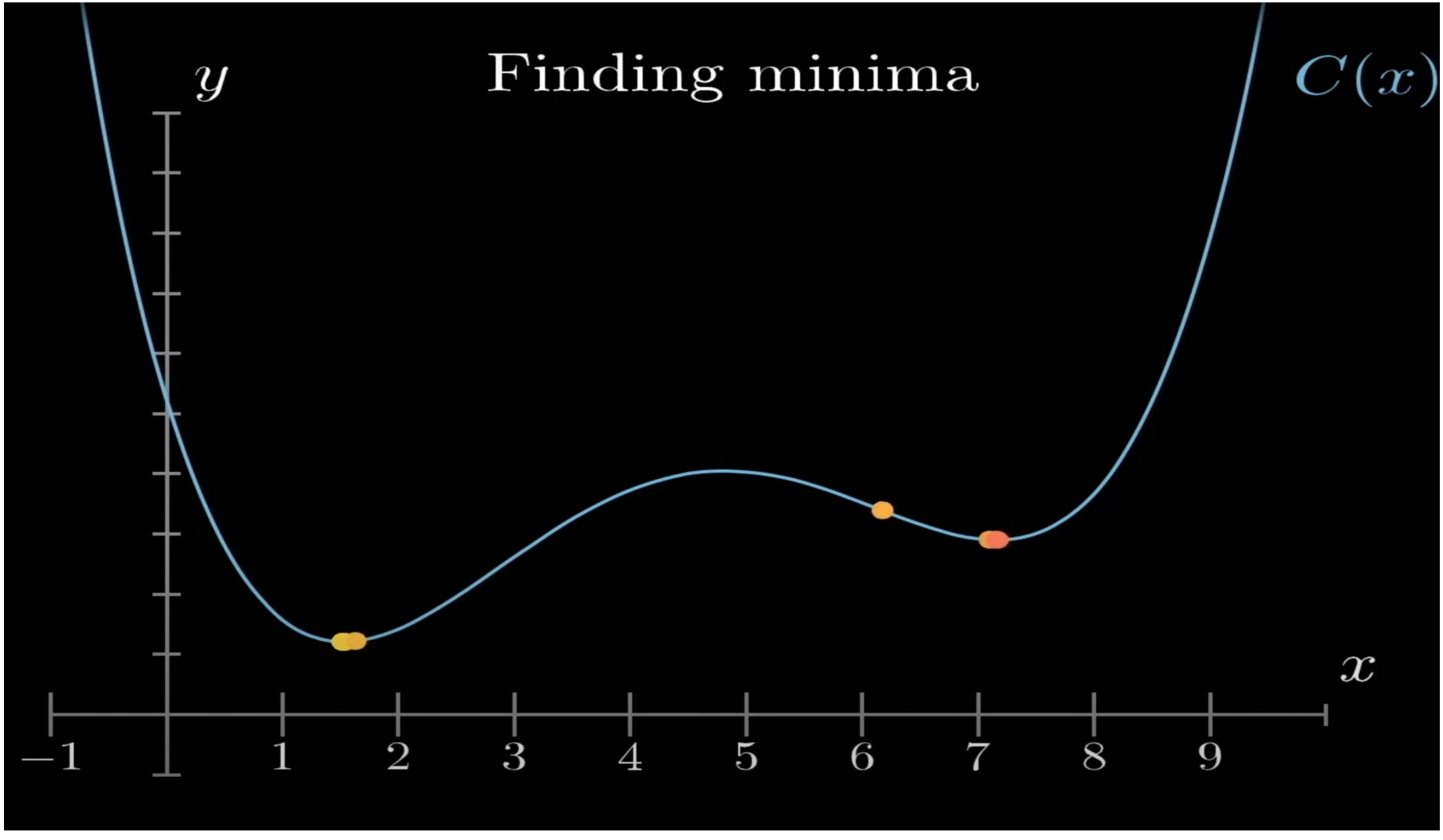
~~Thing that holds a number~~

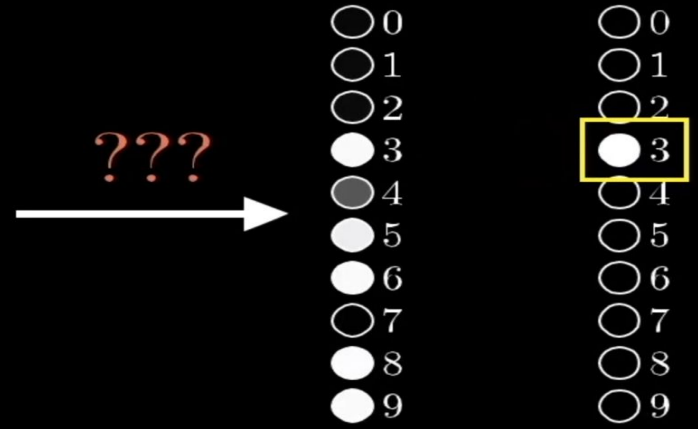
Plan

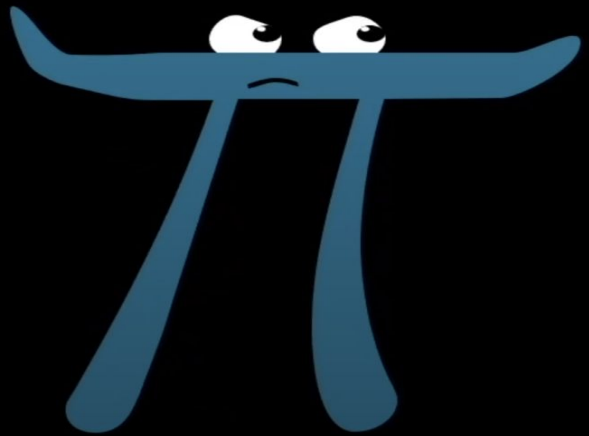
- Recap
- Gradient descent
- Analyze this network
- Where to learn more
- Research corner

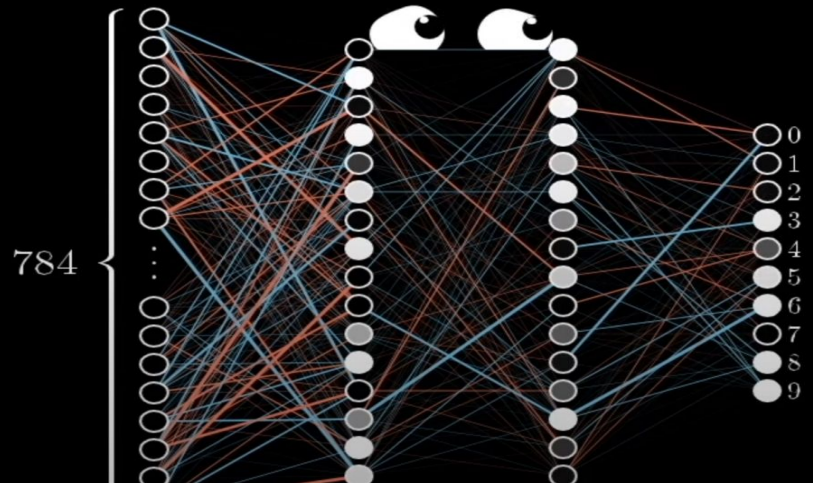


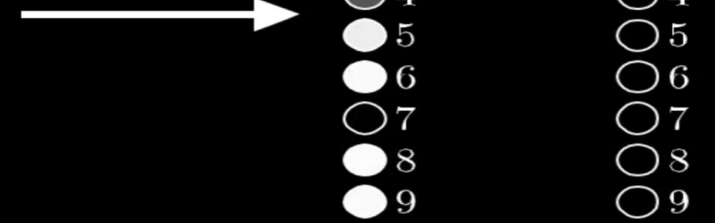

It's actually
just calculus.
Even worse!

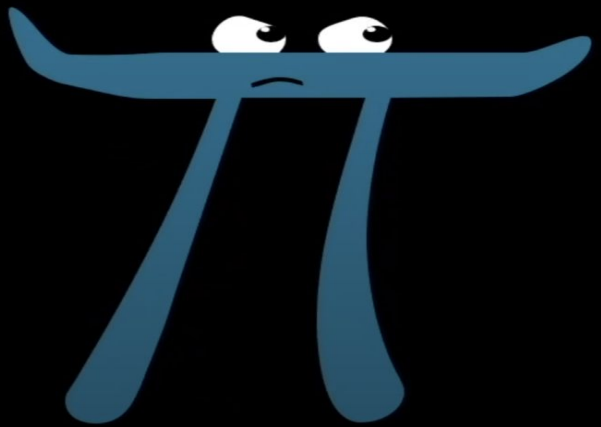

Finding minima
y
$C(x)$
x
−1
1
2
3
4
5
6
7
8
9

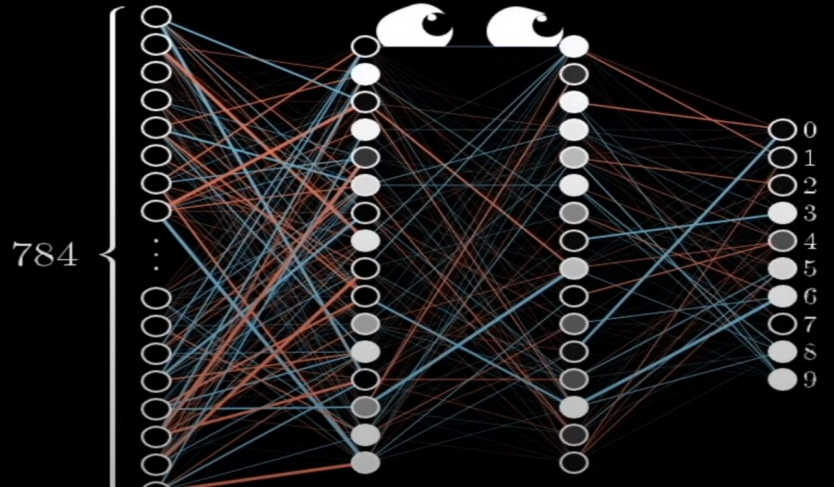

???
0
1
2
3
4
5
6
7
8
9
0
1
2
3
4
5
6
7
8
9




784
0
1
2
3
4
5
6
7
8
9


???
0
1
2
3
4
5
6
7
8
9
0
1
2
3
4
5
6
7
8
9

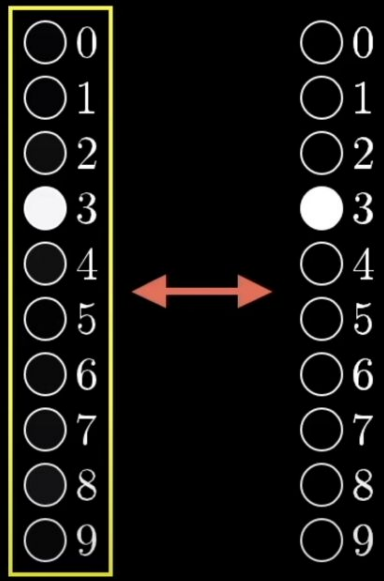



784
0
1
2
3
4
5
6
7
8
9

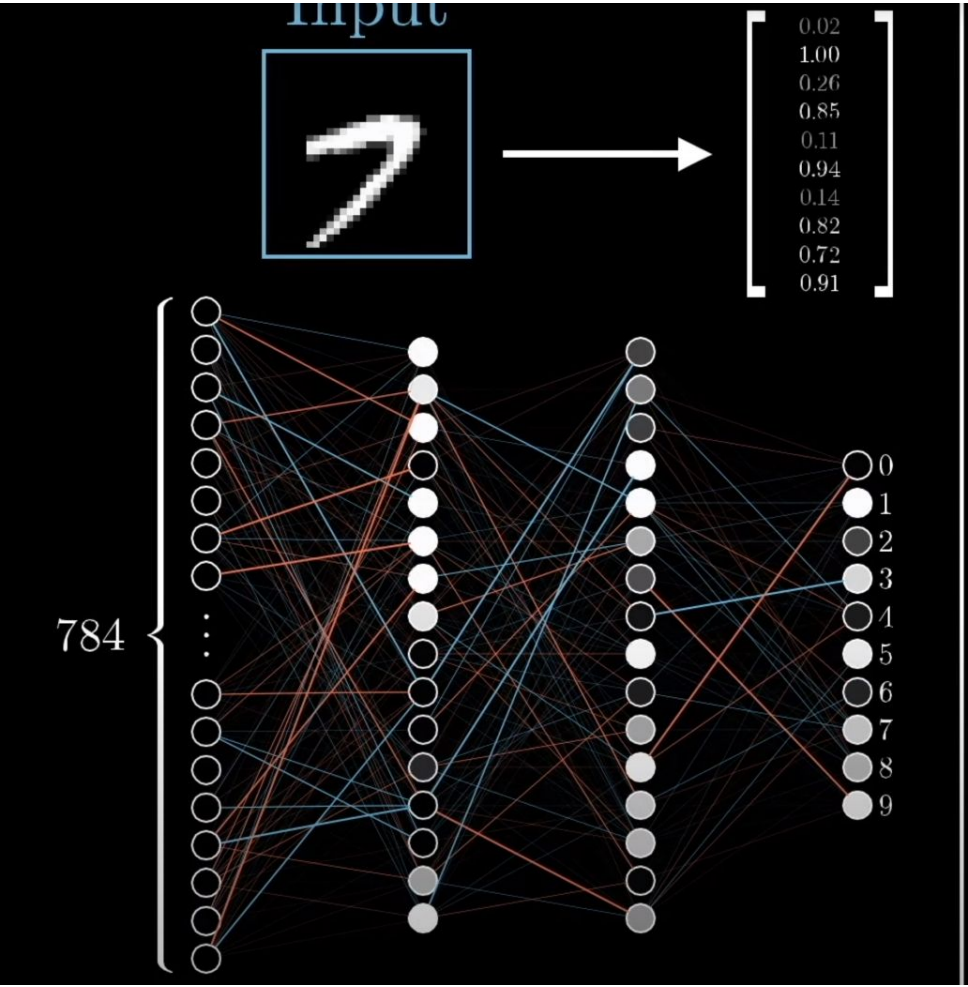
Cost of 

$$
0.03 \left\{ \begin{array}{rl}
0.0006 \leftarrow & (0.02 - 0.00)^2 + \\
0.0007 \leftarrow & (0.03 - 0.00)^2 + \\
0.0039 \leftarrow & (0.06 - 0.00)^2 + \\
0.0009 \leftarrow & (0.97 - 1.00)^2 + \\
0.0055 \leftarrow & (0.07 - 0.00)^2 + \\
0.0004 \leftarrow & (0.02 - 0.00)^2 + \\
0.0022 \leftarrow & (0.05 - 0.00)^2 + \\
0.0033 \leftarrow & (0.06 - 0.00)^2 + \\
0.0072 \leftarrow & (0.08 - 0.00)^2 + \\
0.0018 \leftarrow & (0.04 - 0.00)^2
\end{array} \right.
$$

What's the "cost" of this difference?



Utter trash



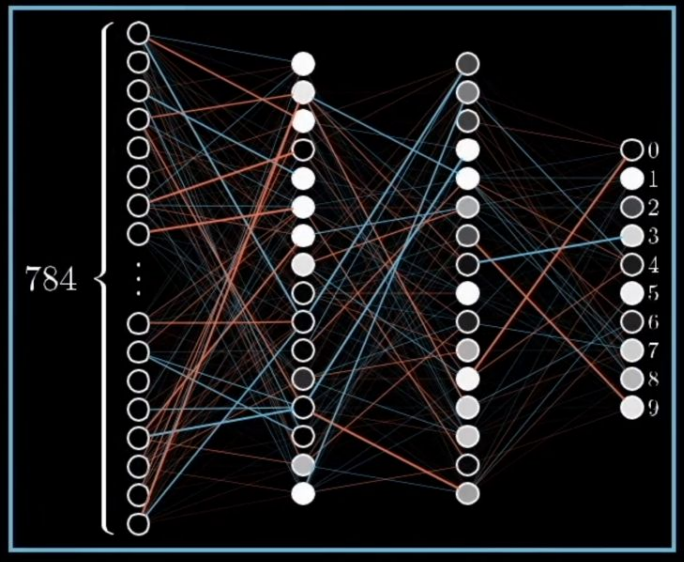
Neural network function

Input: 784 numbers (pixels)

Output: 10 numbers

Parameters: 13,002 weights/biases

Input



Cost: 5.4

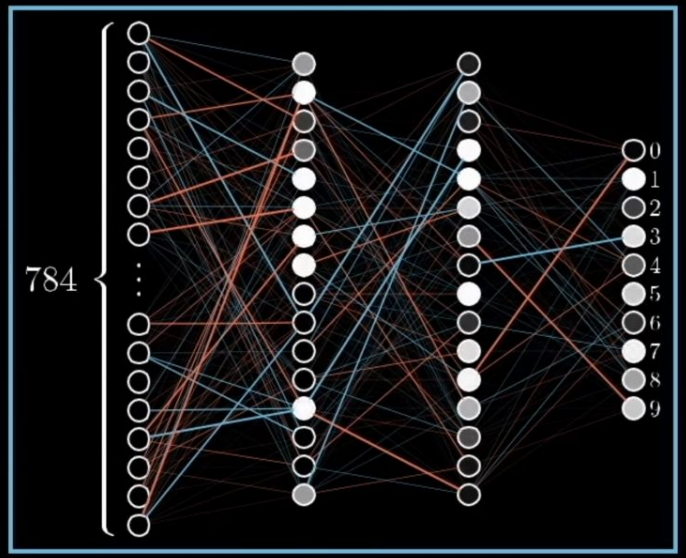
Cost function

Input: 13,002 weights/biases

Output: 1 number (the cost)

Parameters:

Input



Cost: 5.4

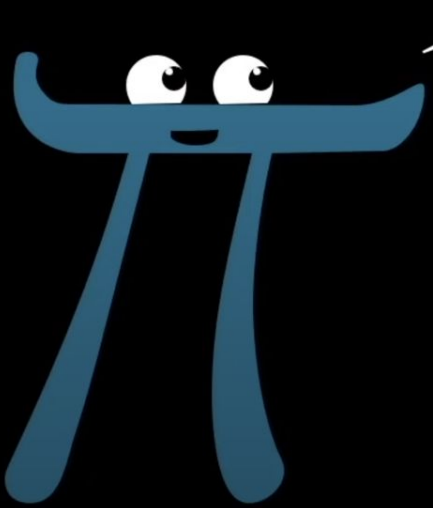
Cost function

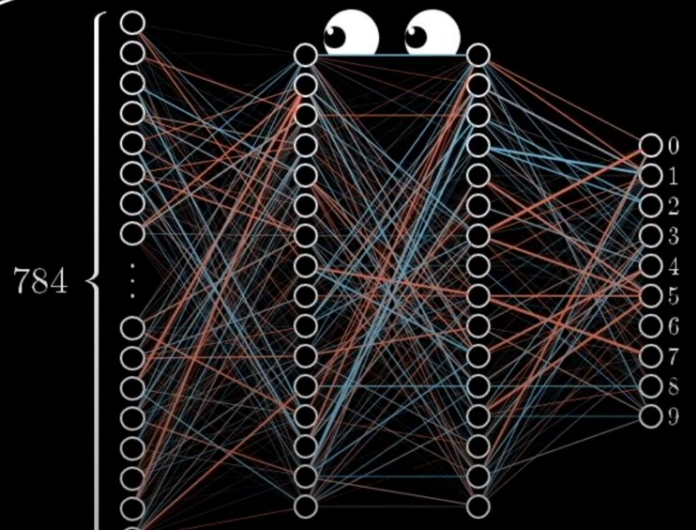
Input: 13,002 weights/biases

Output: 1 number (the cost)

Parameters: Many, many, many training examples

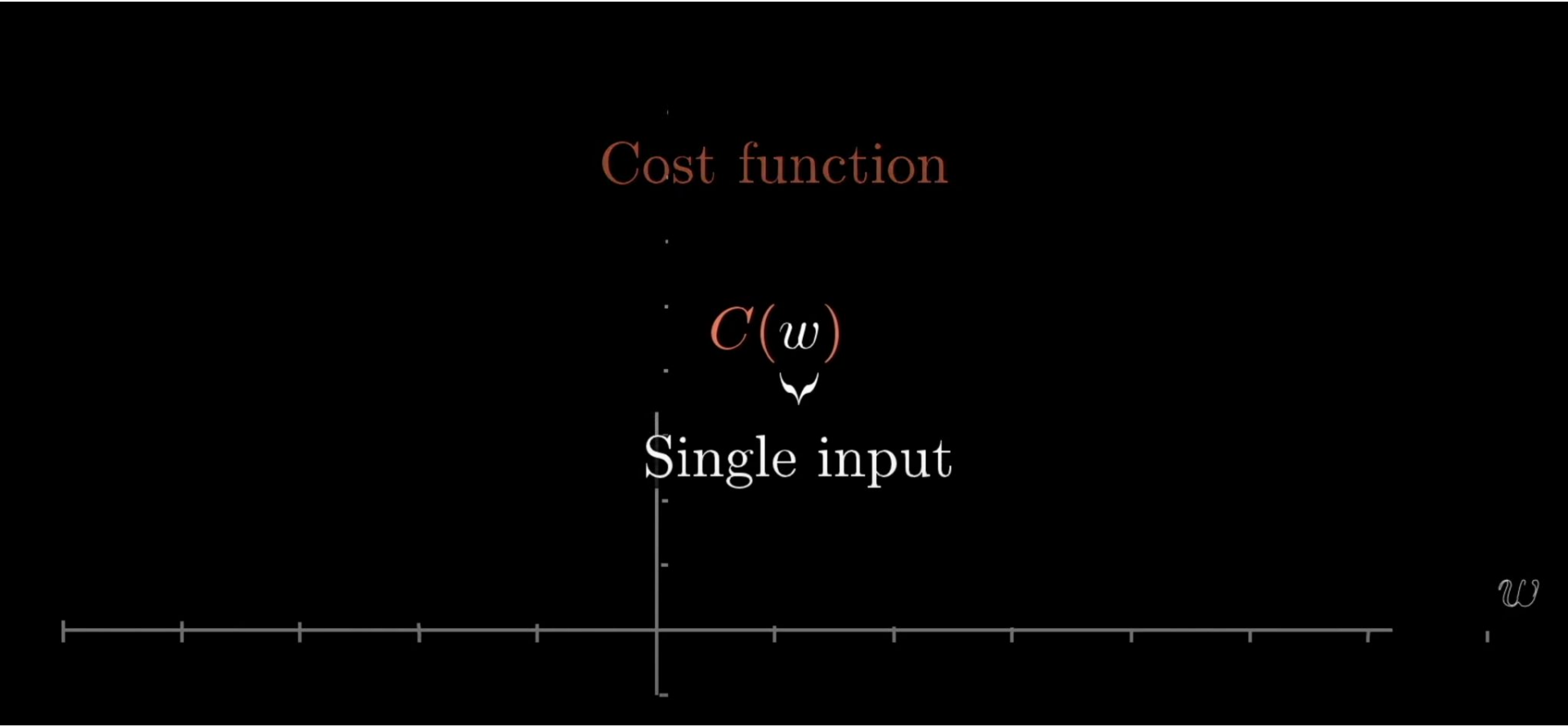
But we can do better!
Growth mindset!

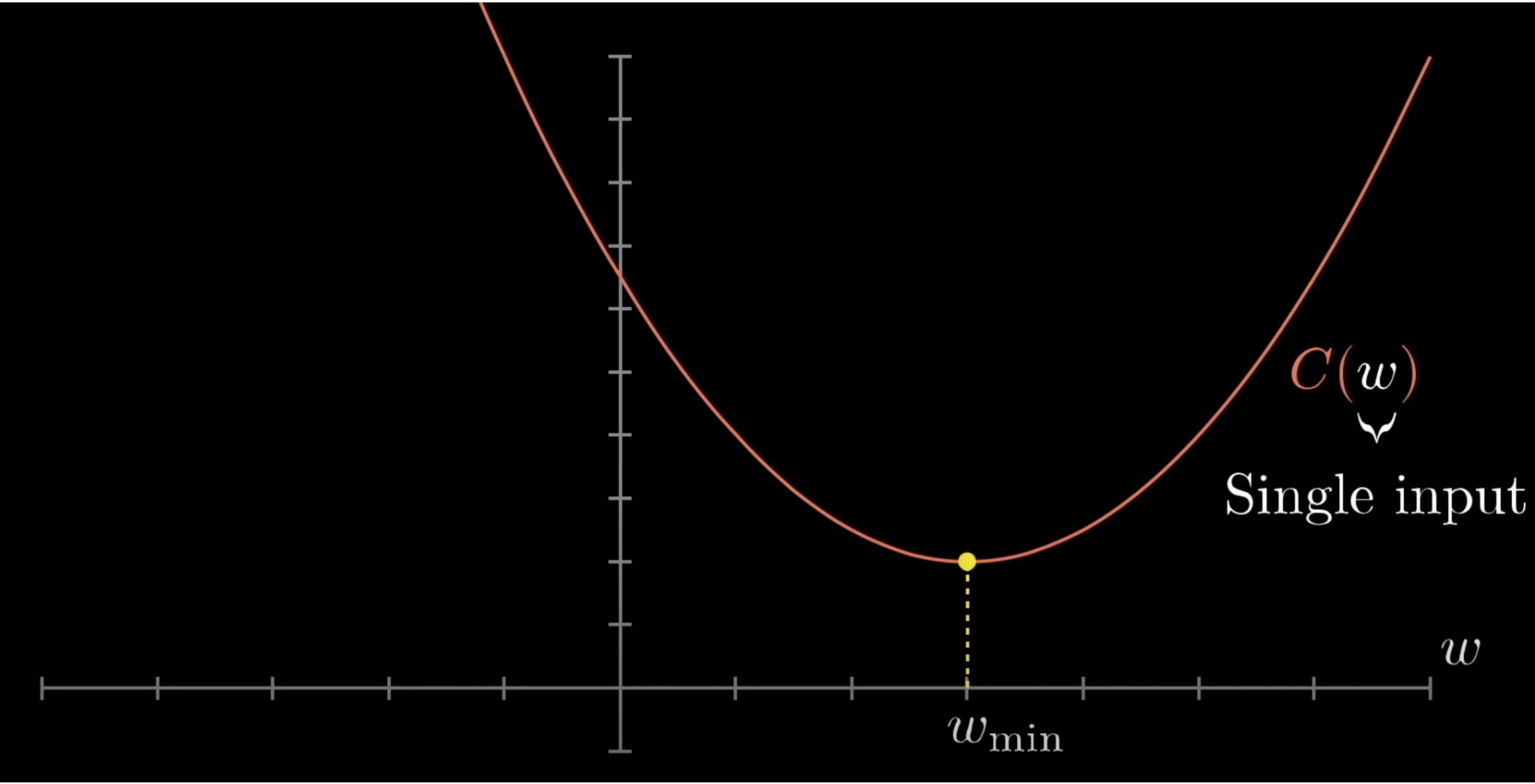


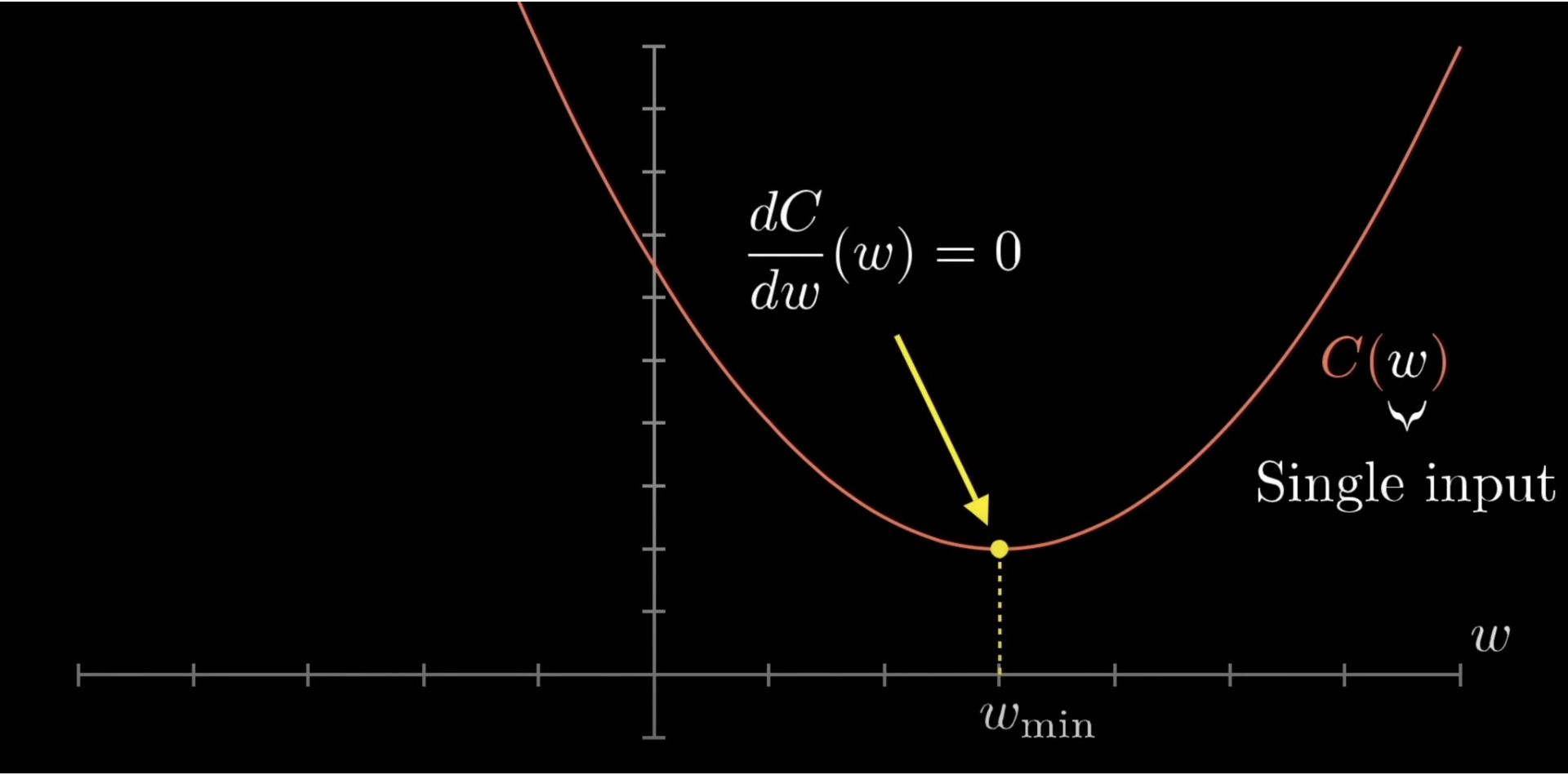


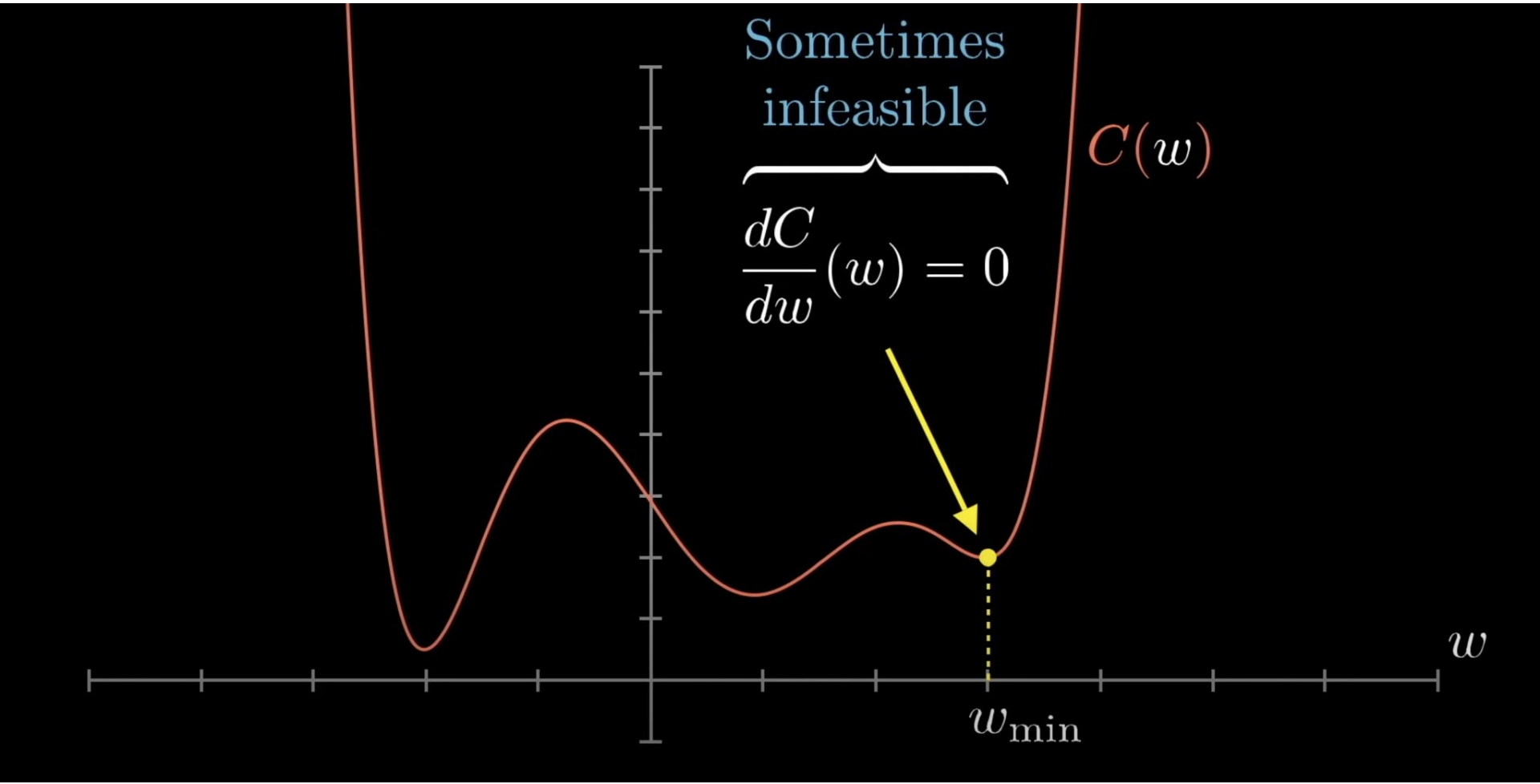
Cost function

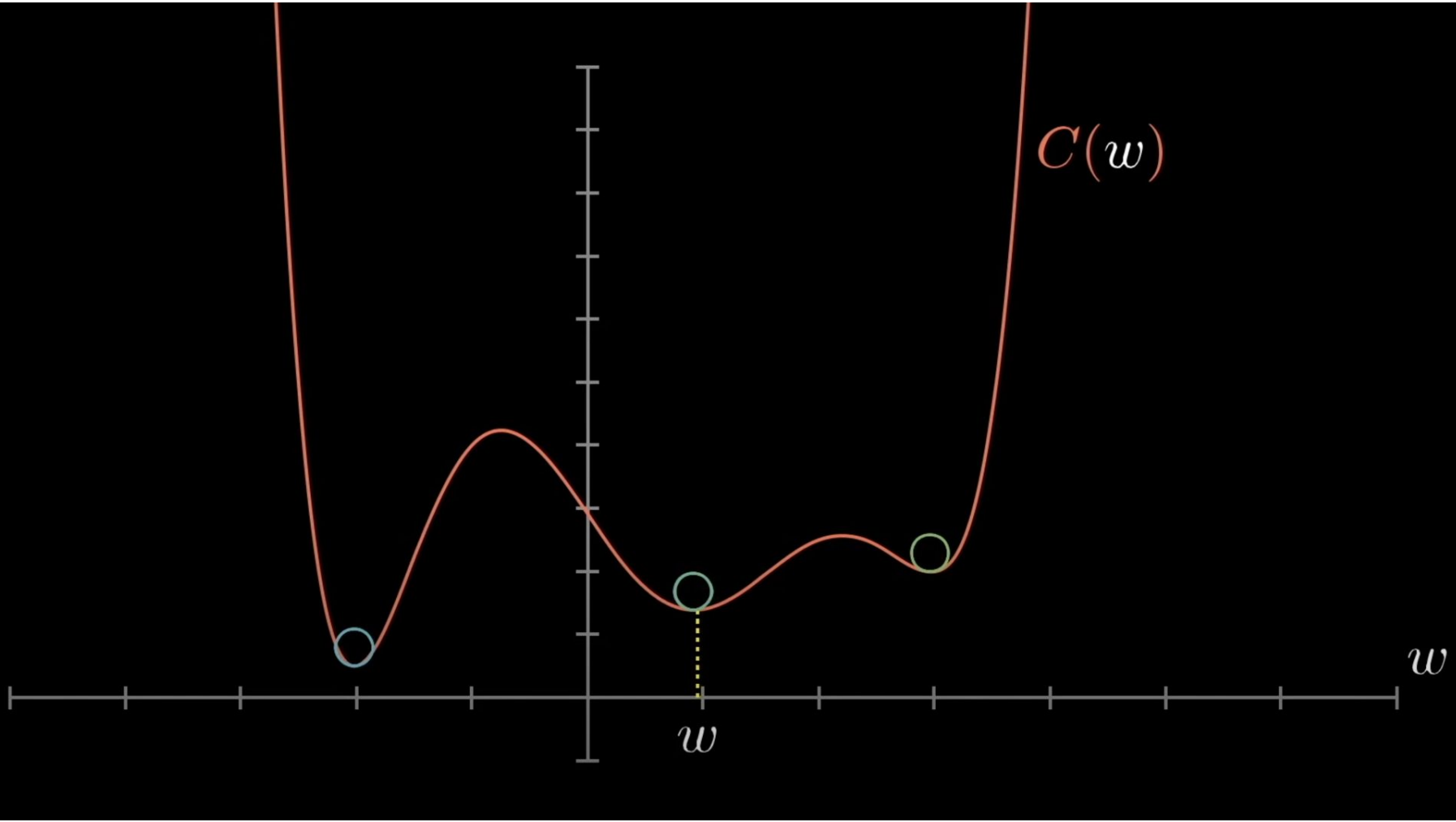
$$C(\underbrace{w_1, w_2, \ldots, w_{13{,}002}}_{\text{Weights and biases}})$$

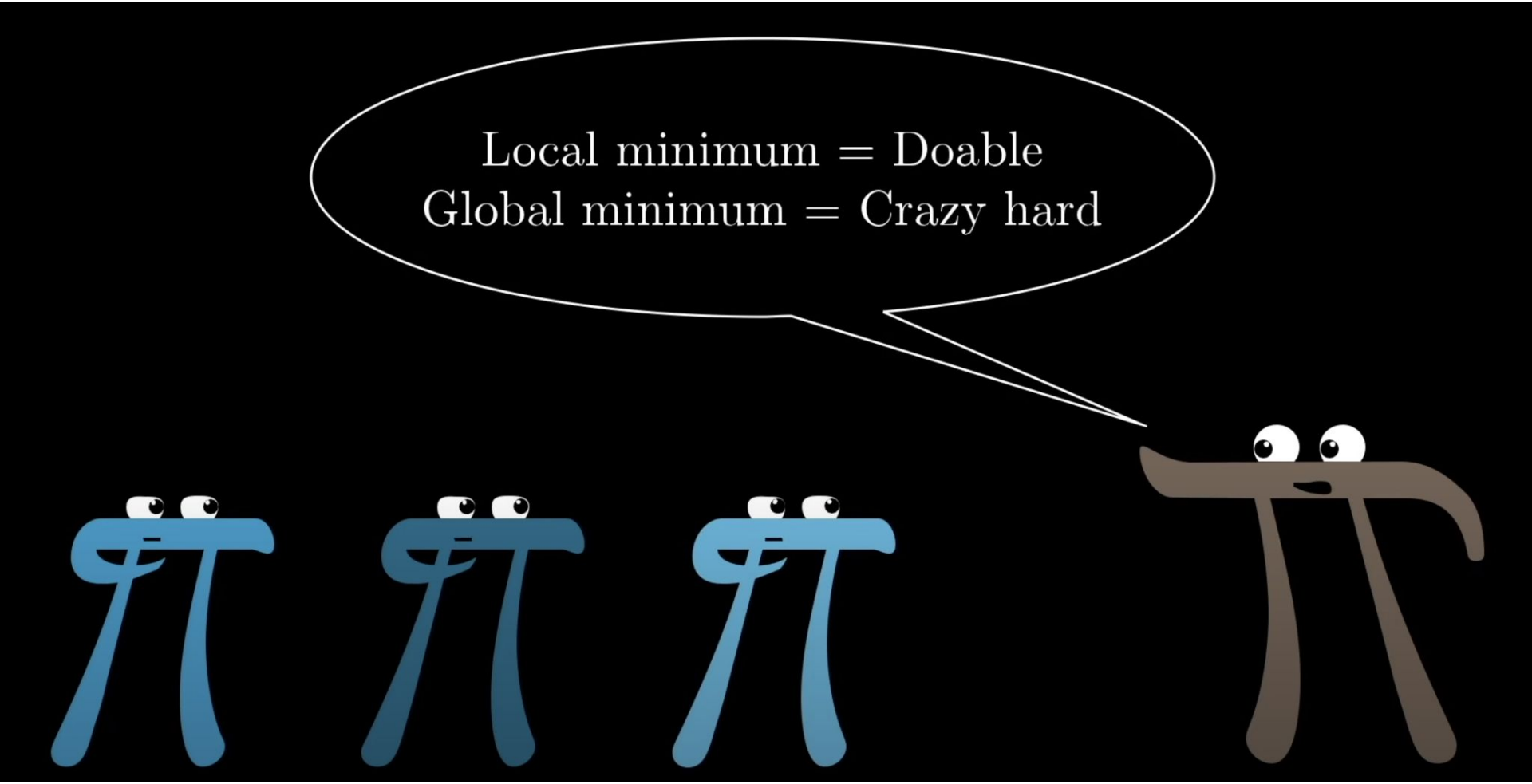

Cost function
$C(w)$
Single input
w


$C(w)$
Single input
w
w_{min}


$\frac{dC}{dw}(w) = 0$
$C(w)$
Single input
w
w_{min}


Sometimes
infeasible
$C(w)$
$\frac{dC}{dw}(w) = 0$
w
$w_{\min}$


$C(w)$
w
w

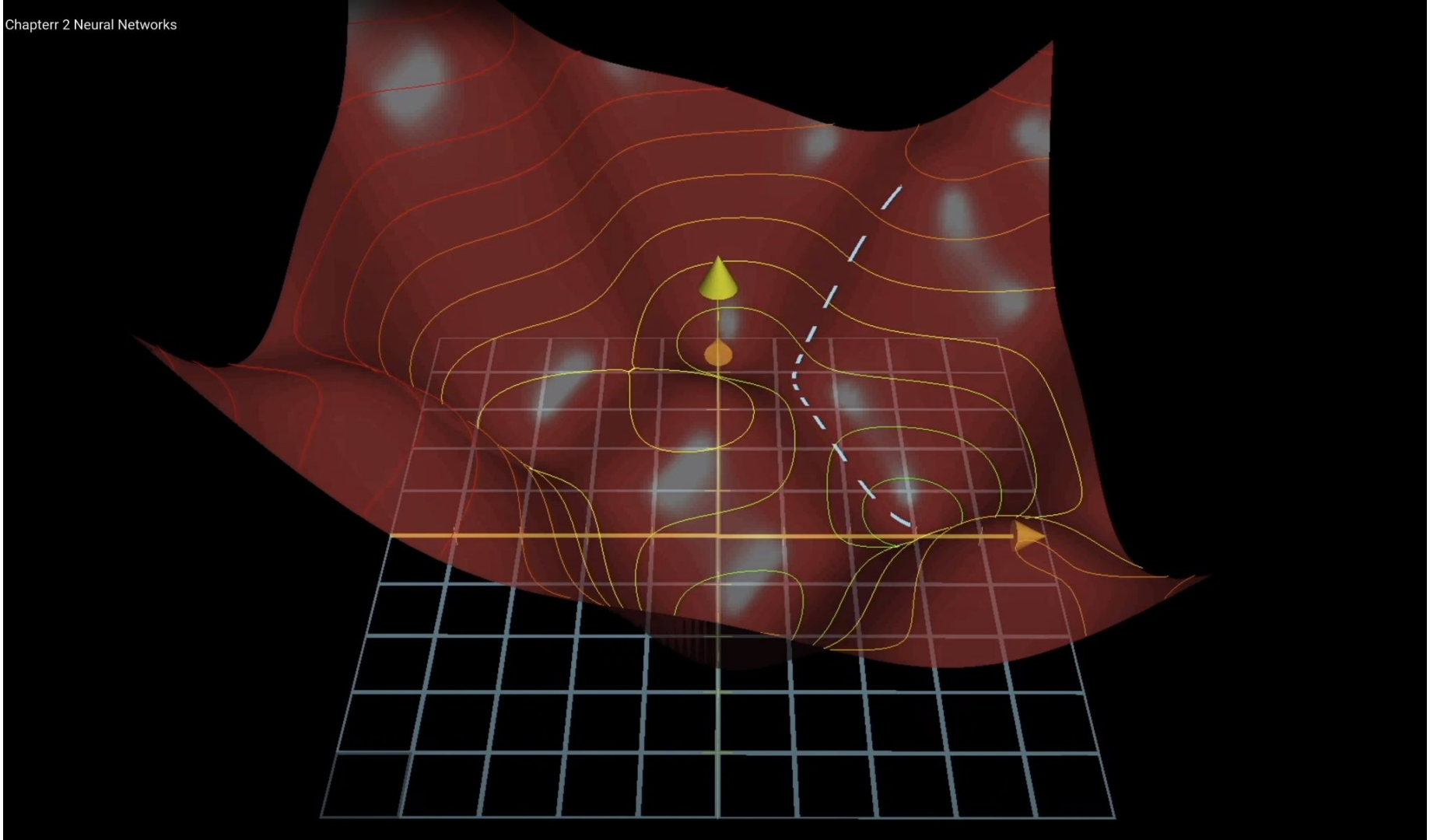

Local minimum = Doable
Global minimum = Crazy hard

13,002 weights and biases

$$\vec{\mathbf{W}} = \begin{bmatrix} 2.25 \\ -1.57 \\ 1.98 \\ \vdots \\ -1.16 \\ 3.82 \\ 1.21 \end{bmatrix}$$

How to nudge all weights and biases

$$-\nabla C(\vec{\mathbf{W}}) = \begin{bmatrix} 0.18 \\ 0.45 \\ -0.51 \\ \vdots \\ 0.40 \\ -0.32 \\ 0.82 \end{bmatrix}$$



$$\vec{\mathbf{W}} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_{13,000} \\ w_{13,001} \\ w_{13,002} \end{bmatrix}$$

$$-\nabla C(\vec{\mathbf{W}}) = \begin{bmatrix} 0.31 \\ 0.03 \\ -1.25 \\ \vdots \\ 0.78 \\ -0.37 \\ 0.16 \end{bmatrix}$$

w_0 should increase
w_1 should increase
w_2 should decrease

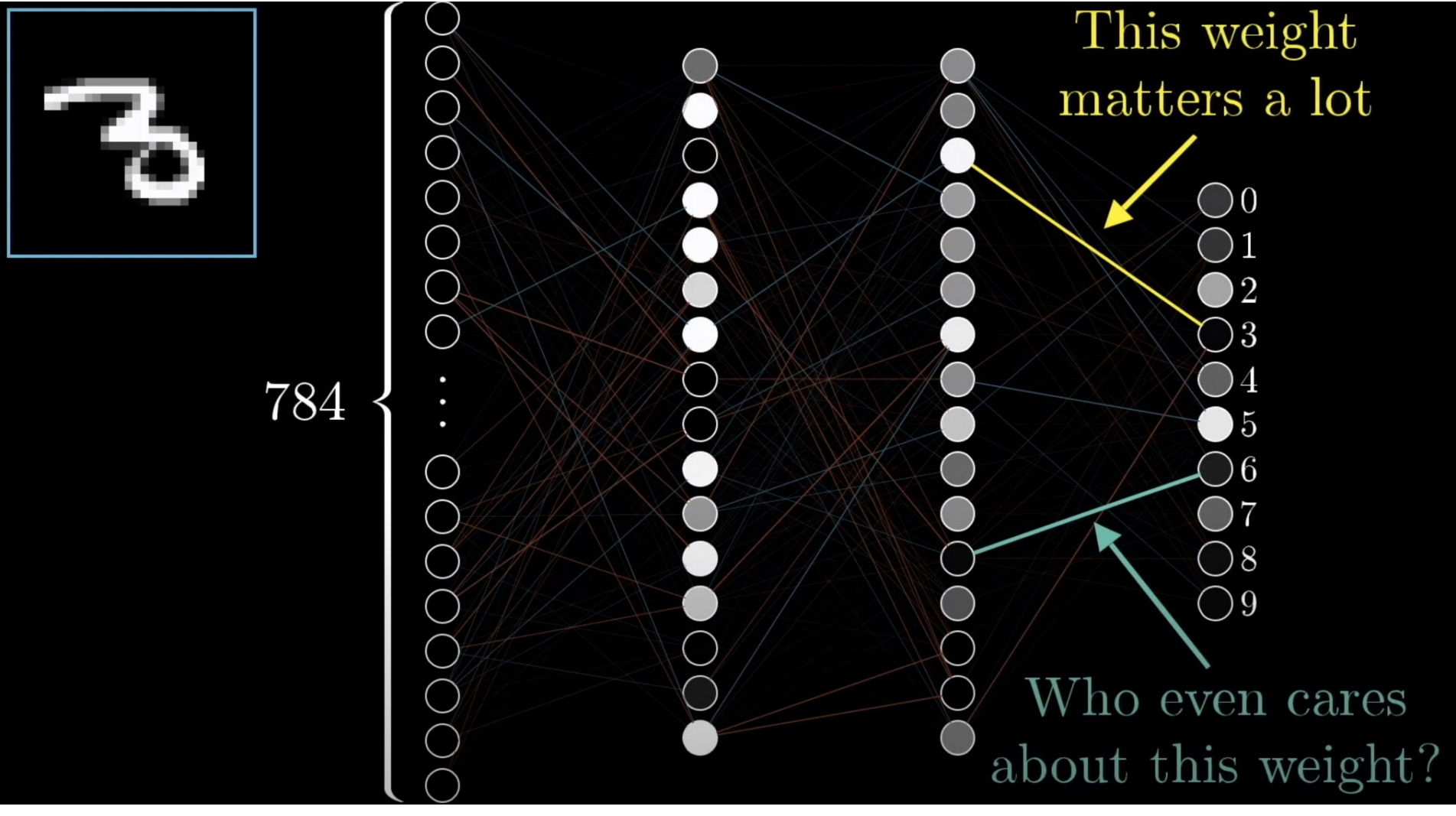
$w_{13,000}$ should increase
$w_{13,001}$ should decrease
$w_{13,002}$ should increase

$$\vec{\mathbf{W}} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_{13,000} \\ w_{13,001} \\ w_{13,002} \end{bmatrix}$$

$$-\nabla C(\vec{\mathbf{W}}) = \begin{bmatrix} 0.31 \\ 0.03 \\ -1.25 \\ \vdots \\ 0.78 \\ -0.37 \\ 0.16 \end{bmatrix}$$

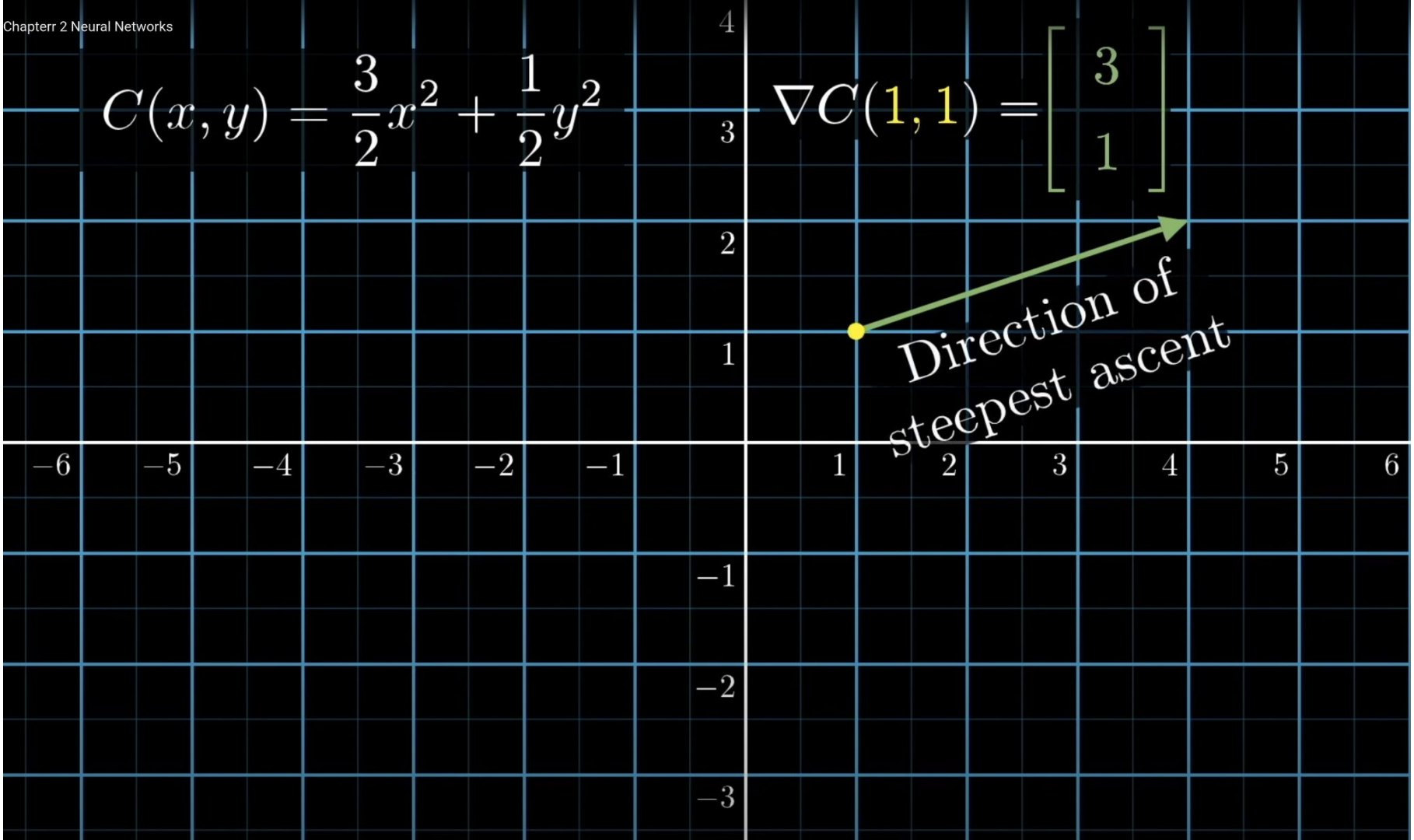
w_0 should increase somewhat
w_1 should increase a little
w_2 should decrease a lot

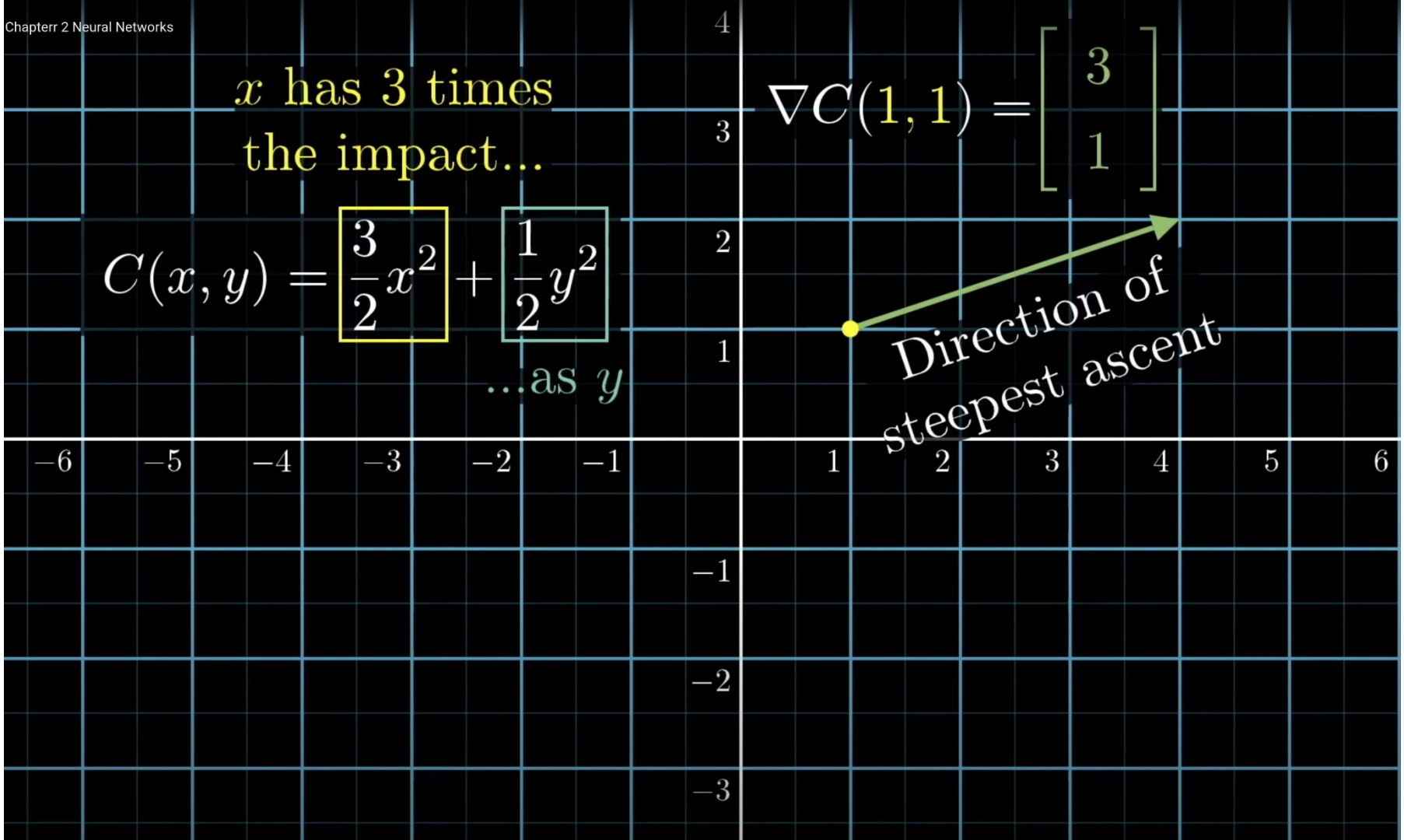
$w_{13,000}$ should increase a lot
$w_{13,001}$ should decrease somewhat
$w_{13,002}$ should increase a little


784
This weight matters a lot
0
1
2
3
4
5
6
7
8
9
Who even cares about this weight?

$$C(x, y) = \frac{3}{2}x^2 + \frac{1}{2}y^2$$

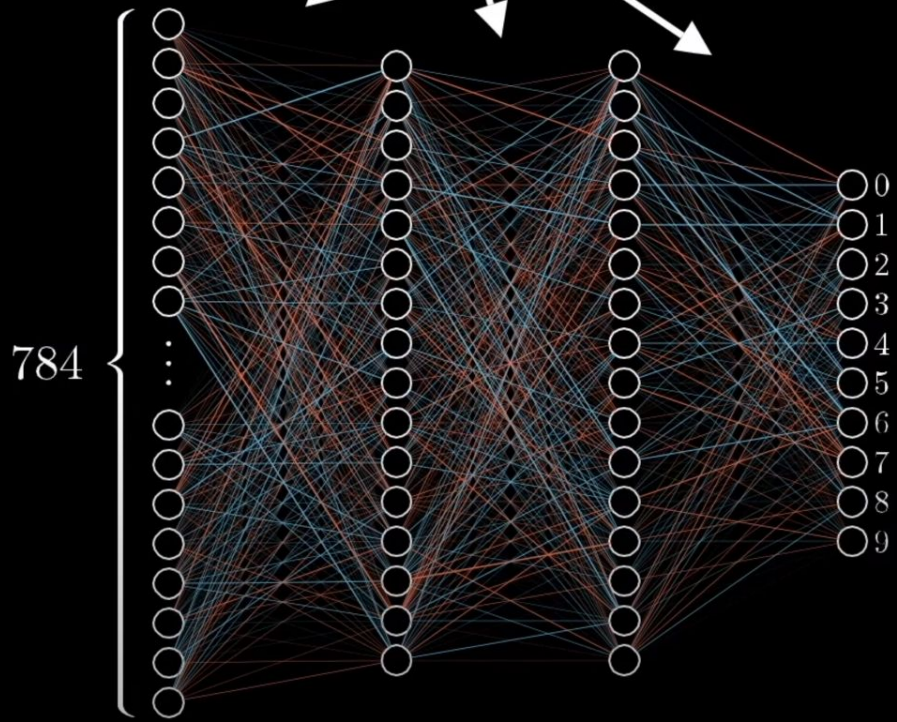
$$\nabla C(1, 1) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

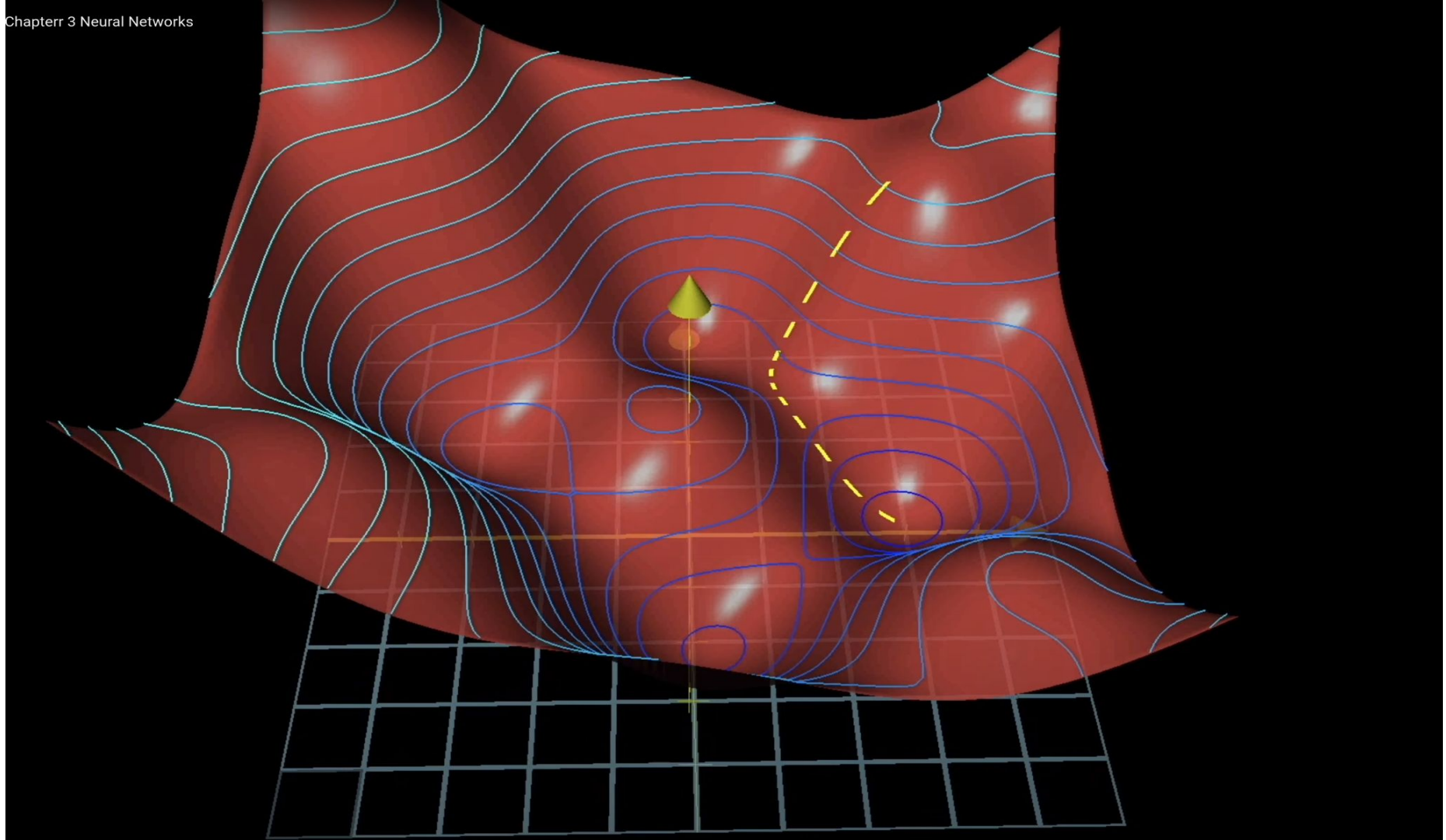



x has 3 times the impact...
$C(x, y) = \frac{3}{2}x^2 + \frac{1}{2}y^2$
...as y
$\nabla C(1, 1) = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$
Direction of steepest ascent
4
3
2
1
−1
−2
−3
−6
−5
−4
−3
−2
−1
1
2
3
4
5
6

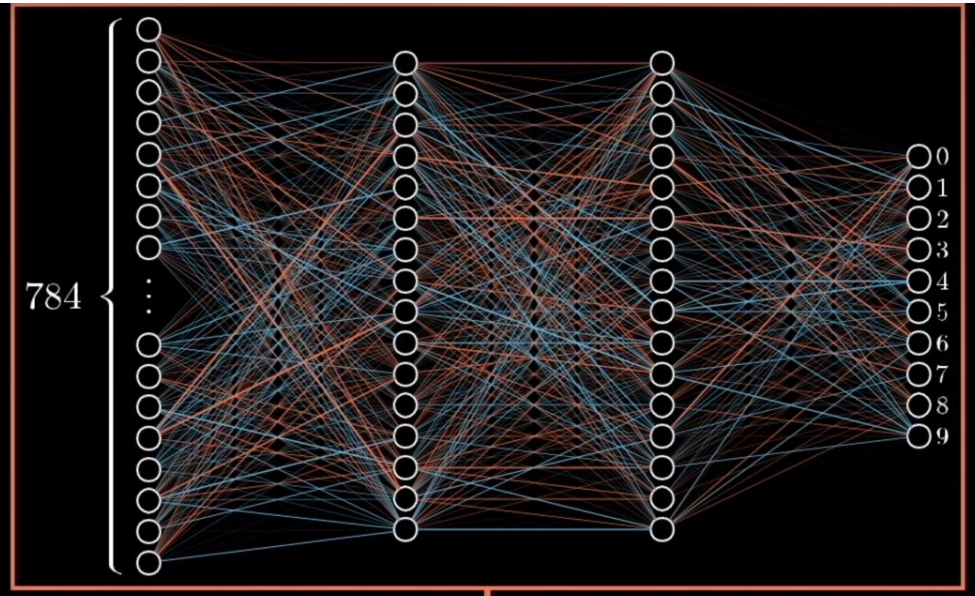
$$-\nabla C(\underbrace{\ldots}_{\text{All weights and biases}}) = \begin{bmatrix} 0.12 \\ 0.79 \\ -0.67 \\ 0.01 \\ \vdots \\ 1.49 \\ 1.59 \\ -1.53 \\ -1.15 \end{bmatrix}$$

Change by some small multiple of $-\nabla C(\ldots)$

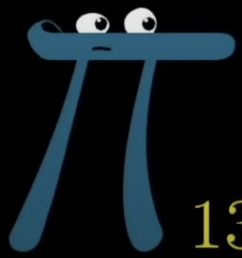




$$-\nabla C(\underbrace{\ldots}_{\text{All weights and biases}}) = \begin{bmatrix} 0.16 \\ 0.72 \\ -0.93 \\ \vdots \\ 0.04 \\ 1.64 \\ 1.52 \end{bmatrix}$$



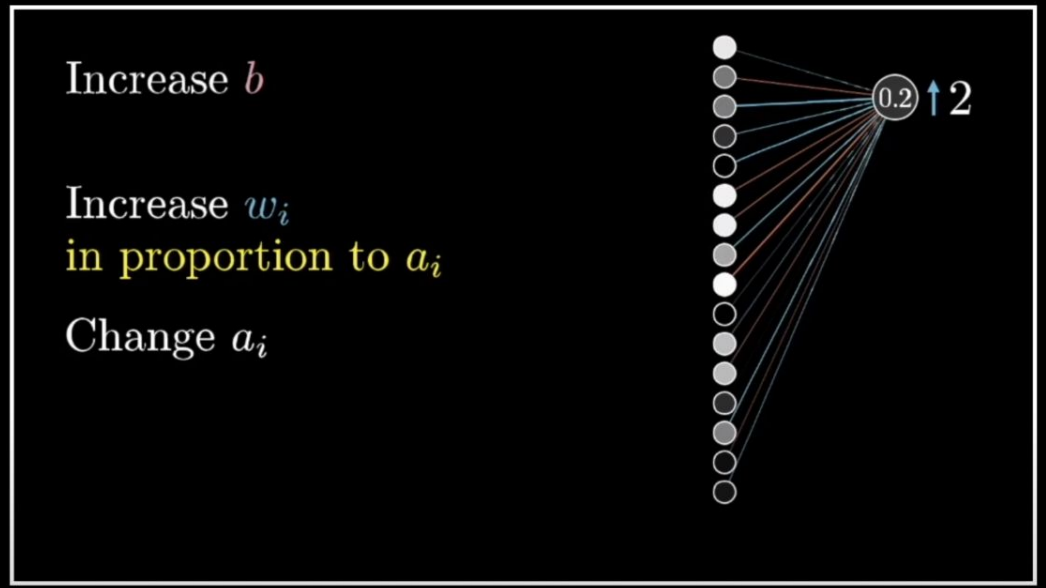
$$C(w_0, w_1, \ldots, w_{13{,}001}) = 2.85$$



Direction in 13,002 dimensions?!?

Plan

- Recap
- Intuitive walkthrough
- Derivatives in computational graphs



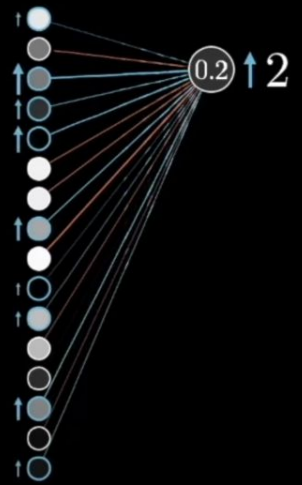
Plan

- Recap
- Intuitive walkthrough
- Derivatives in computational graphs

Increase b

Increase w_i
in proportion to a_i

Change a_i

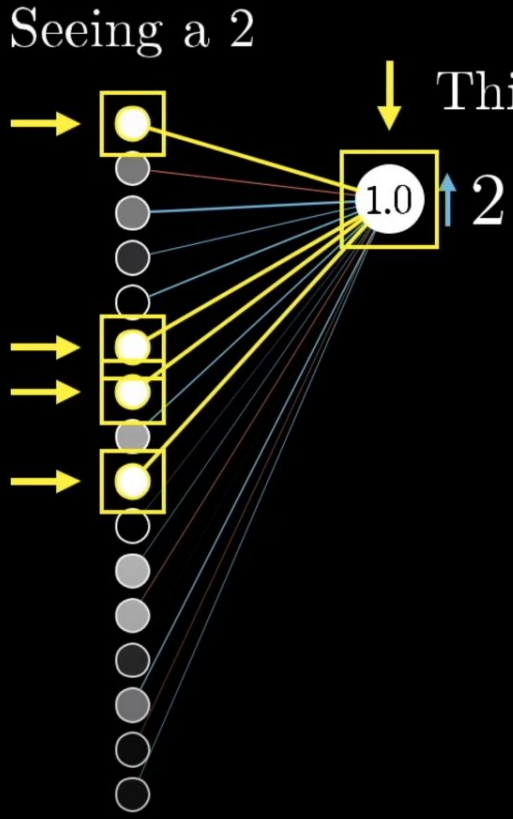


$$0.2 = \sigma(w_0 a_0 + w_1 a_1 + \cdots + w_{n-1} a_{n-1} + b)$$

Increase b

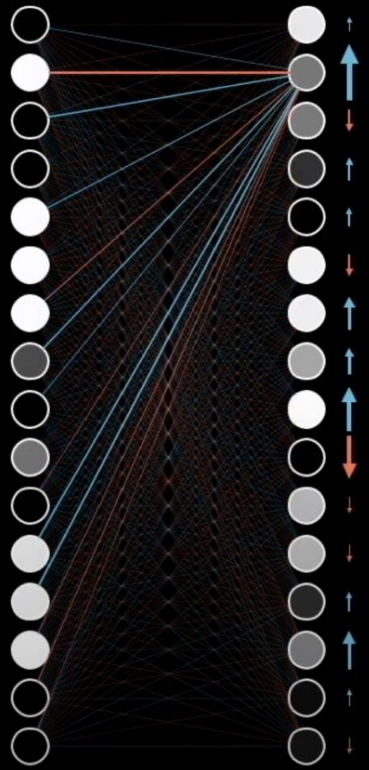
Increase w_i
in proportion to a_i

Change a_i

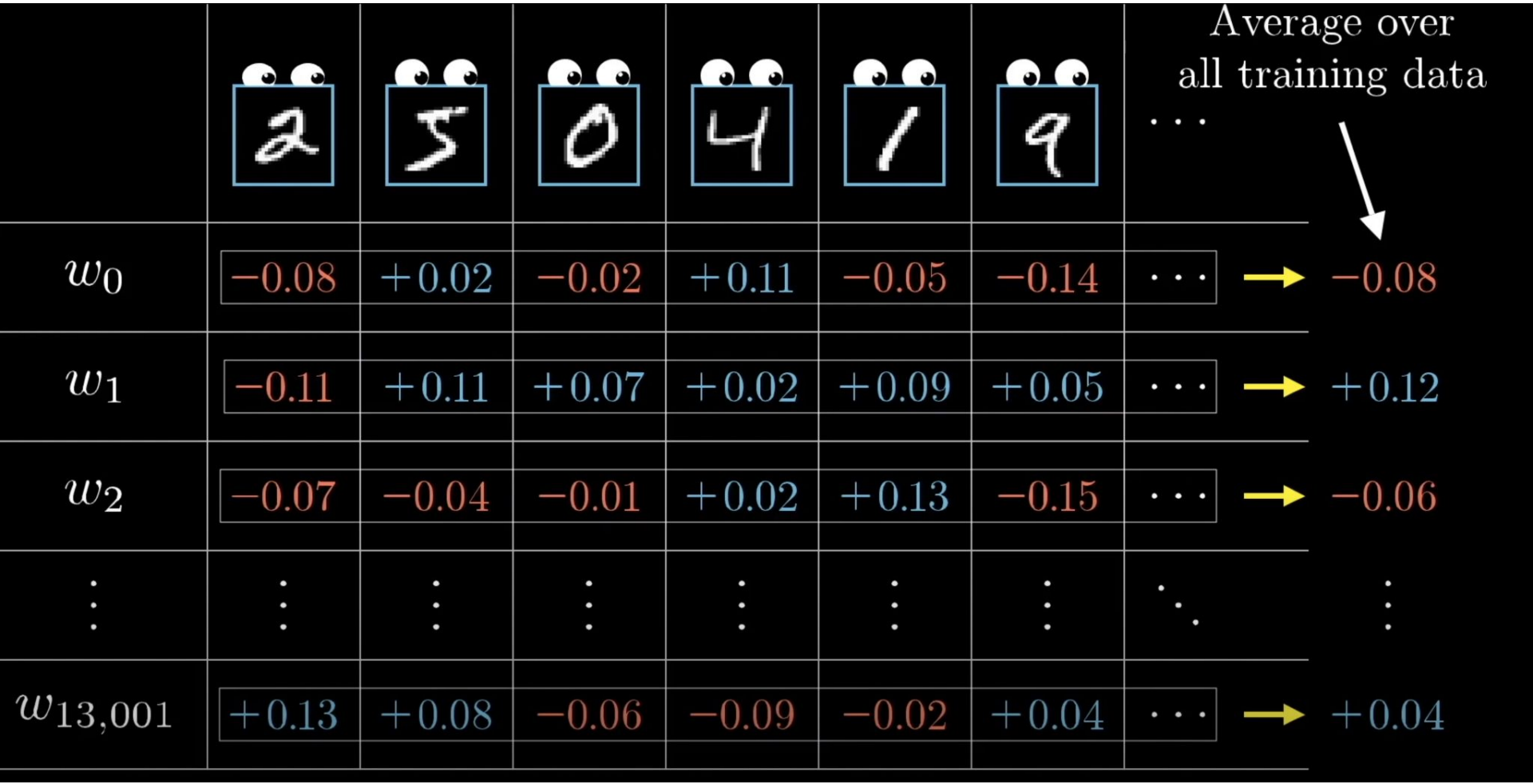


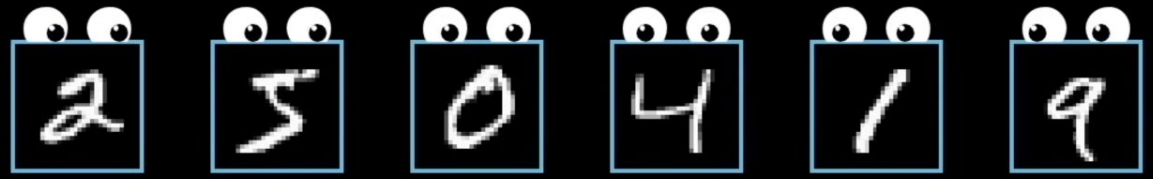
"Neurons that fire together wire together"

Propagate backwards



							$\cdots$
w_0	-0.08	$+0.02$	-0.02	$+0.11$	-0.05	-0.14	$\cdots$
w_1	-0.11	$+0.11$	$+0.07$	$+0.02$	$+0.09$	$+0.05$	$\cdots$
w_2	-0.07	-0.04	-0.01	$+0.02$	$+0.13$	-0.15	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
$w_{13,001}$	$+0.13$	$+0.08$	-0.06	-0.09	-0.02	$+0.04$	$\cdots$


Average over all training data
2
5
0
4
1
9
w_0 −0.08 +0.02 −0.02 +0.11 −0.05 −0.14 → −0.08
w_1 −0.11 +0.11 +0.07 +0.02 +0.09 +0.05 → +0.12
w_2 −0.07 −0.04 −0.01 +0.02 +0.13 −0.15 → −0.06
$w_{13,001}$ +0.13 +0.08 −0.06 −0.09 −0.02 +0.04 → +0.04



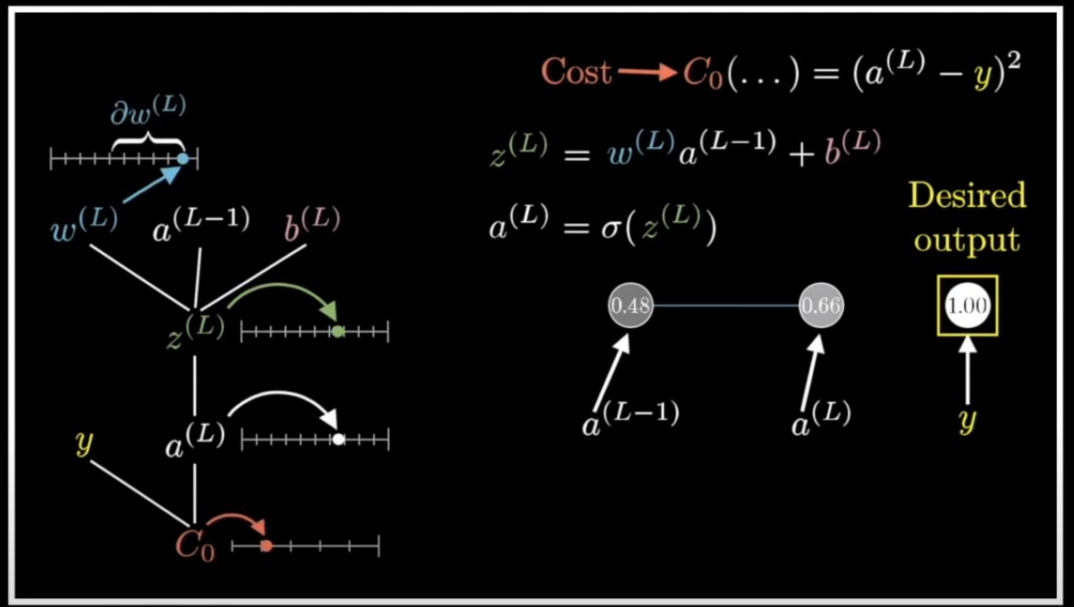
Average over all training data

$$-\nabla C(w_1, w_2, \ldots w_{13{,}001}) =$$

$$\begin{bmatrix} -0.08 \\ +0.12 \\ -0.06 \\ \vdots \\ +0.04 \end{bmatrix}$$

Plan

- Recap
- Intuitive walkthrough
- Derivatives in computational graphs

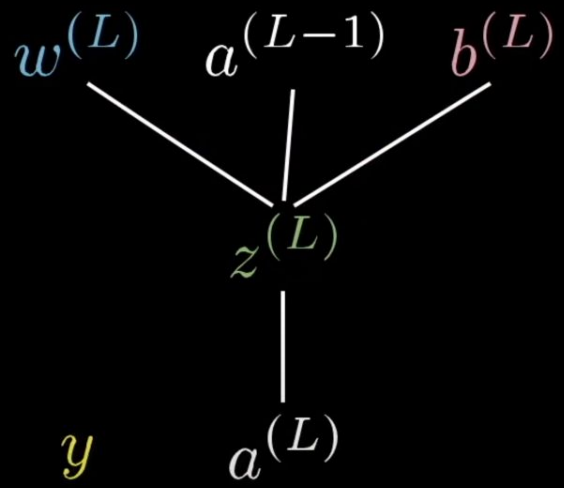


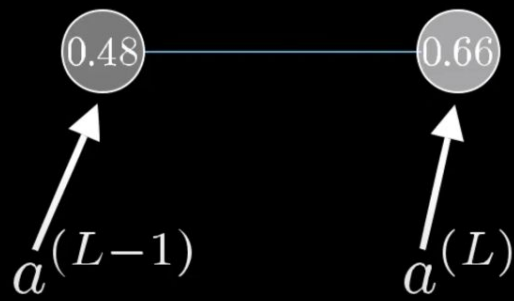
$$C(w_1, b_1, w_2, b_2, w_3, b_3)$$

$$\text{Cost} \longrightarrow C_0(\dots) = (a^{(L)} - y)^2$$

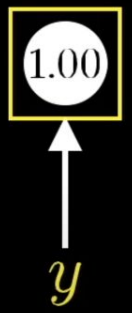
$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$





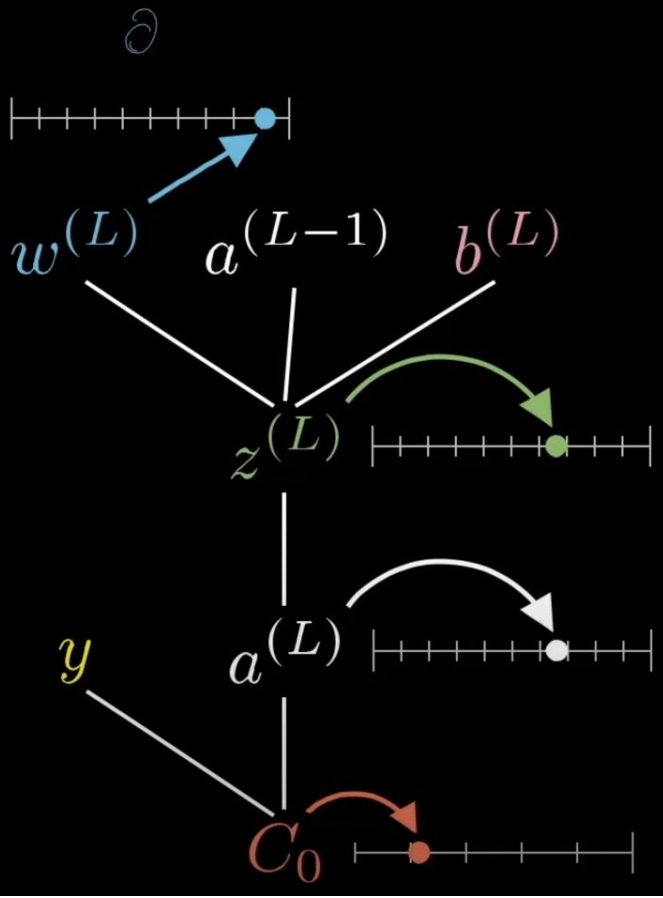
Desired output

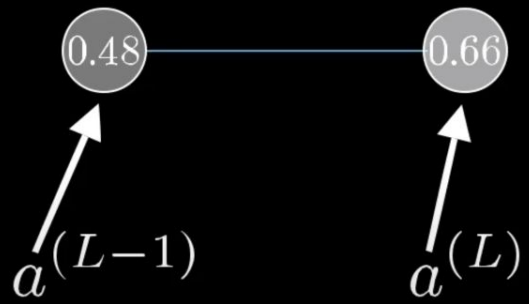


$$\text{Cost} \longrightarrow C_0(\dots) = (a^{(L)} - y)^2$$

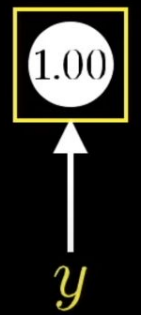
$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$





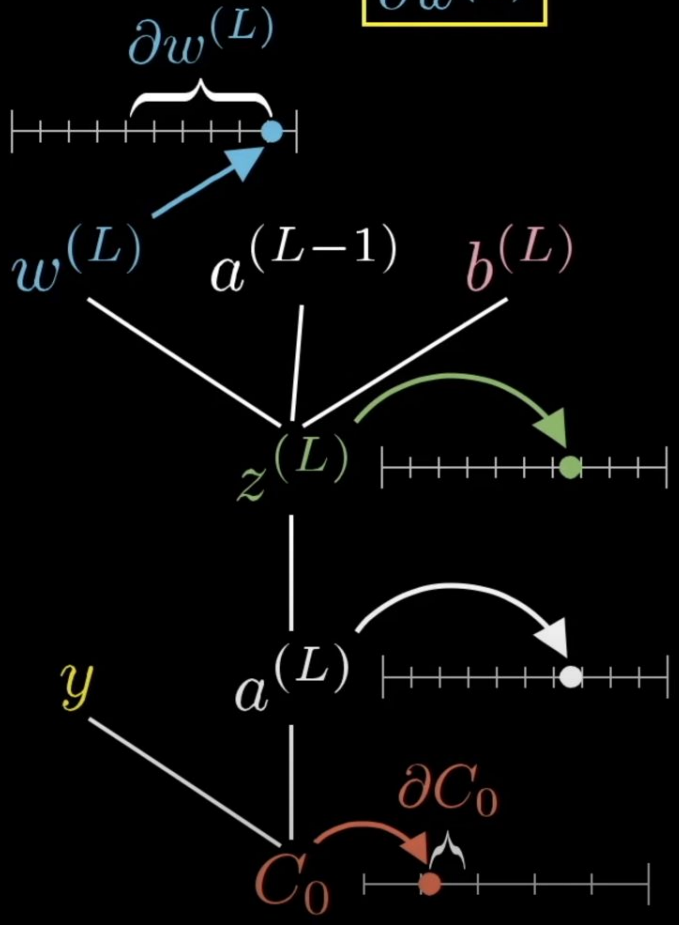
Desired output



$$\frac{\partial C_0}{\partial w^{(L)}}$$

What we want

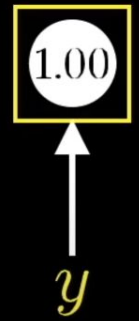
$$C_0(\dots) = (a^{(L)} - y)^2$$



$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$

Desired output

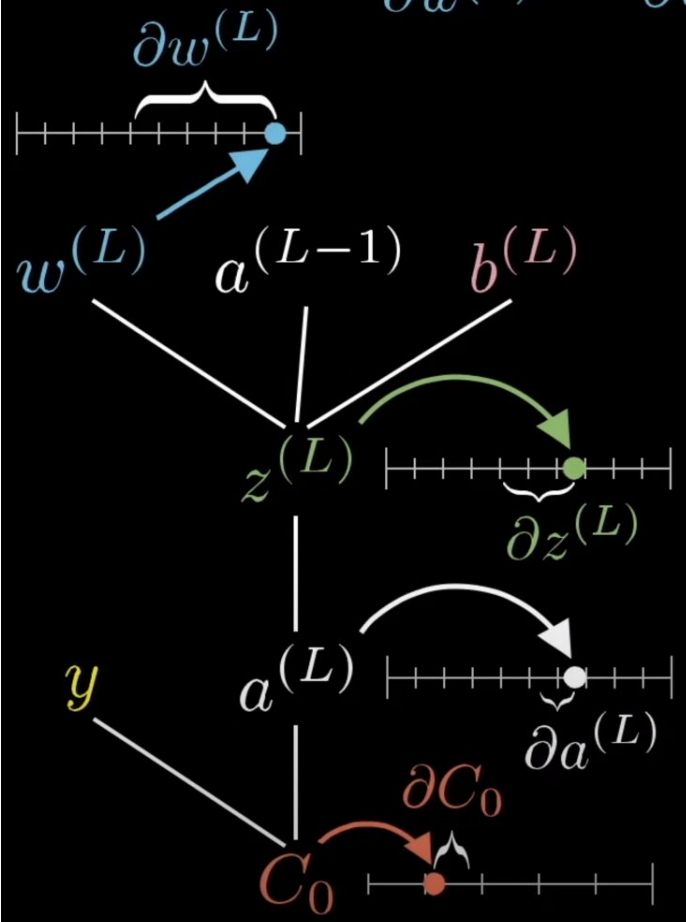


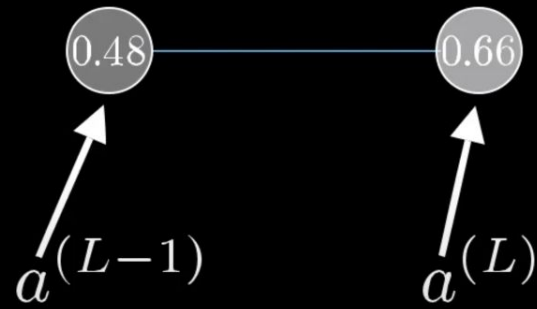
$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}}$$

$$C_0(\dots) = (a^{(L)} - y)^2$$

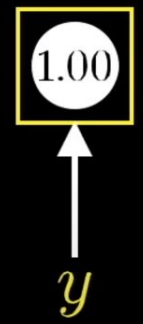
$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$



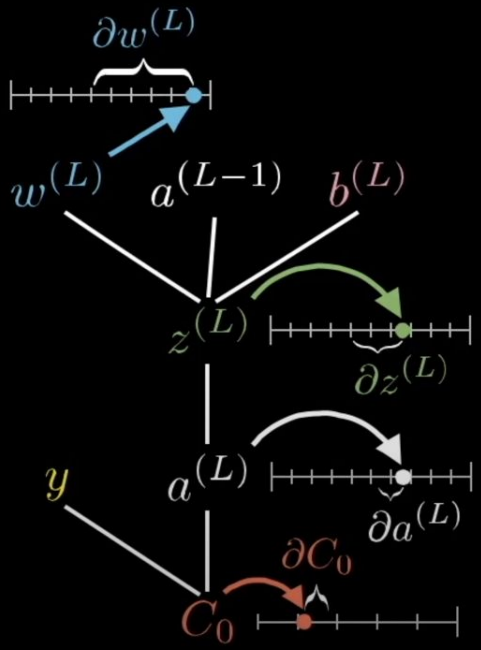


Desired output



$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}}$$

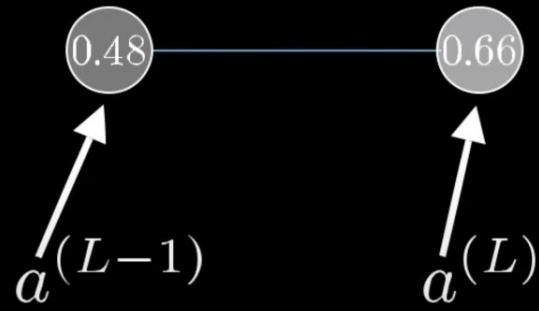
Chain rule



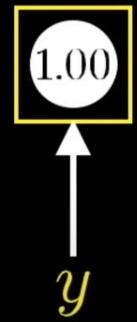
$$C_0(\dots) = (a^{(L)} - y)^2$$

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$



Desired output



$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}}$$

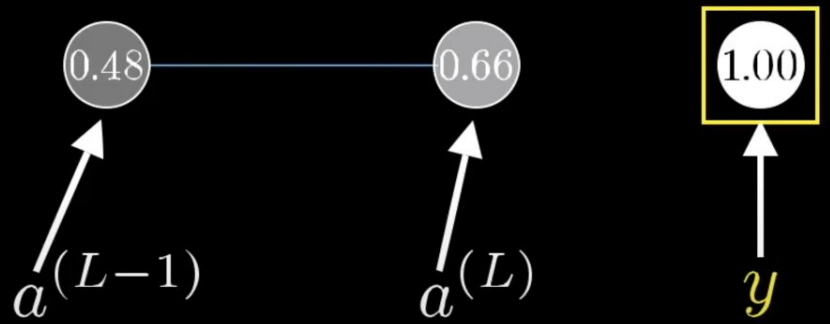
$$C_0 = (a^{(L)} - y)^2$$

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$\frac{\partial C0}{\partial a^{(L)}} = 2(a^{(L)} - y)$$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$

$$a^{(L)} = \sigma(z^{(L)})$$



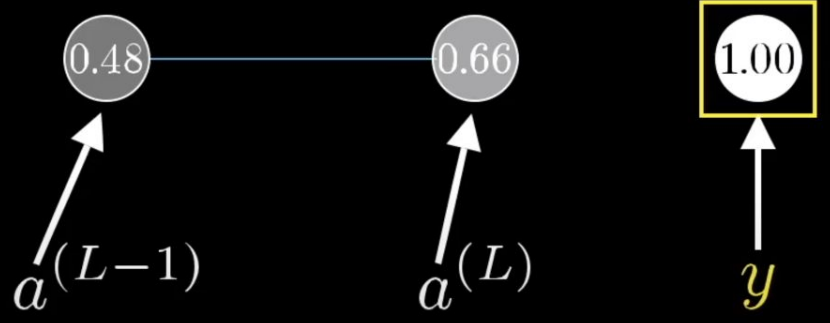
$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}}$$

$$C_0 = (a^{(L)} - y)^2$$

$$\frac{\partial C0}{\partial a^{(L)}} = 2(a^{(L)} - y)$$

$$a^{(L)} = \sigma(z^{(L)})$$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$



$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)}$$

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}} = a^{(L-1)} \sigma'(z^{(L)}) 2(a^{(L)} - y)$$

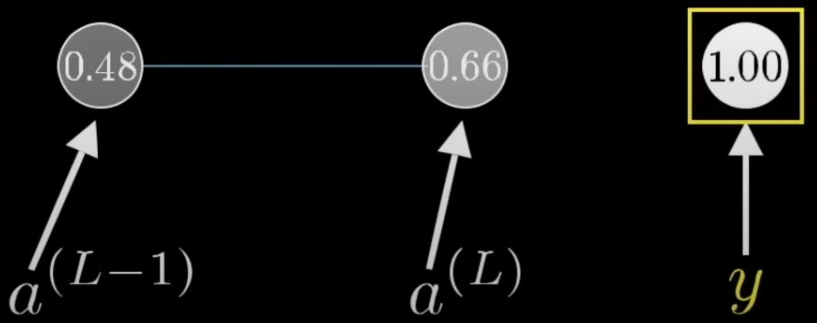
$$C_0 = (a^{(L)} - y)^2$$

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$

Average of all training examples

$$\frac{\partial C}{\partial w^{(L)}} = \overbrace{\frac{1}{n} \sum_{k=0}^{n-1} \frac{\partial C_k}{\partial w^{(L)}}}$$



$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}} = a^{(L-1)} \sigma'(z^{(L)}) 2(a^{(L)} - y)$$

$$C_0 = (a^{(L)} - y)^2$$

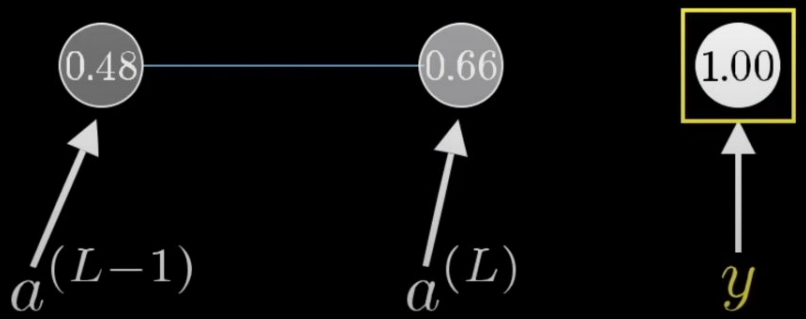
$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$

Average of all training examples

$$\underbrace{\frac{\partial C}{\partial w^{(L)}}} = \overbrace{\frac{1}{n} \sum_{k=0}^{n-1} \frac{\partial C_k}{\partial w^{(L)}}}$$

Derivative of full cost function



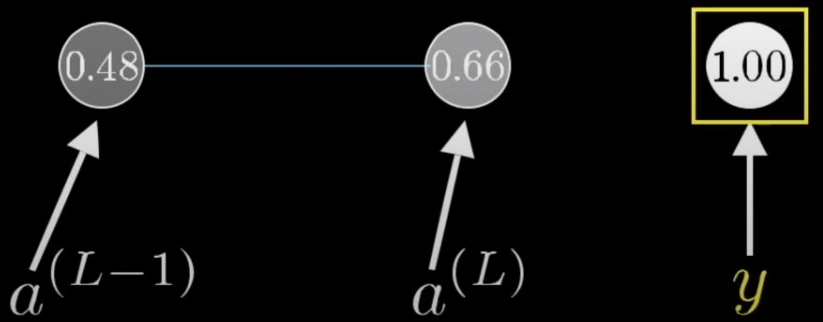
$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C0}{\partial a^{(L)}} = a^{(L-1)} \sigma'(z^{(L)}) 2(a^{(L)} - y)$$

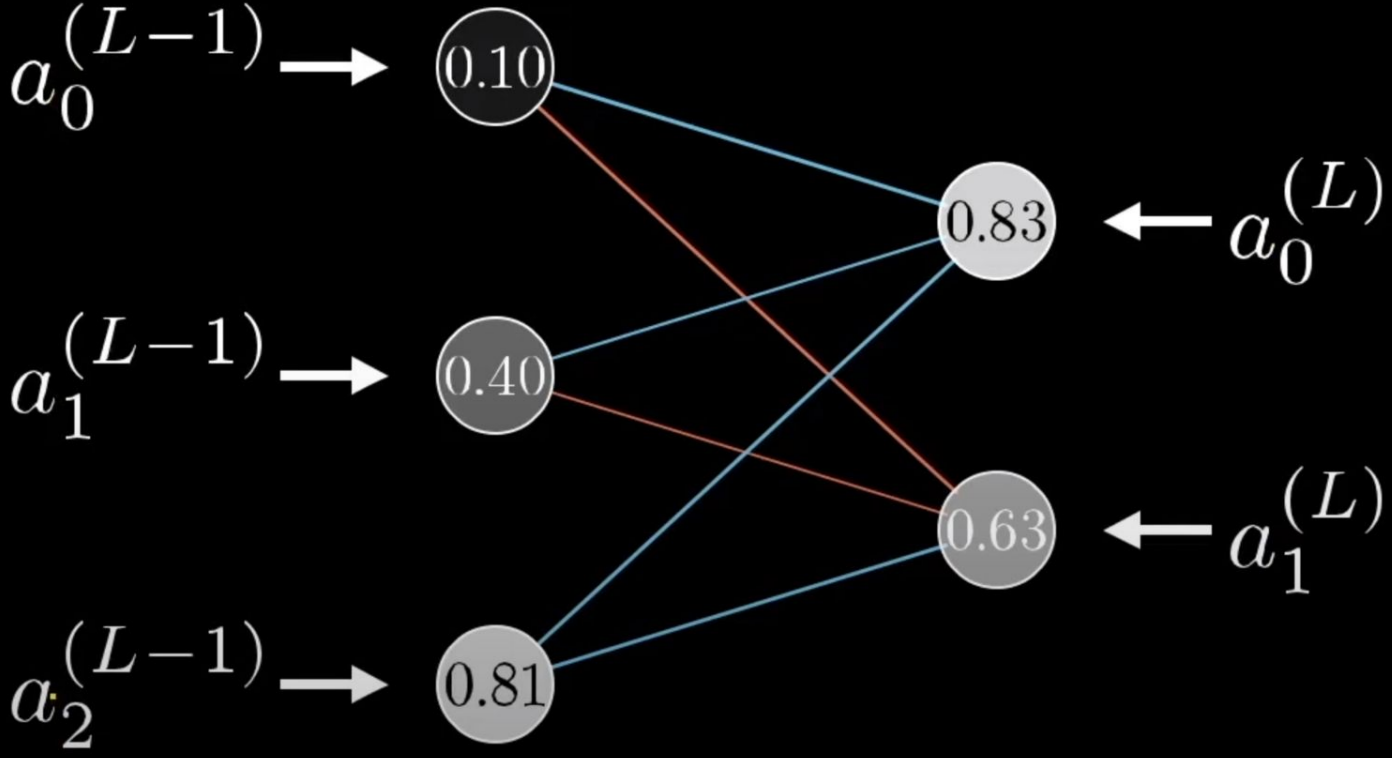
$$\nabla C = \begin{bmatrix} \frac{\partial C}{\partial w^{(1)}} \\ \frac{\partial C}{\partial b^{(1)}} \\ \vdots \\ \frac{\partial C}{\partial w^{(L)}} \\ \frac{\partial C}{\partial b^{(L)}} \end{bmatrix}$$

$$C_0 = (a^{(L)} - y)^2$$

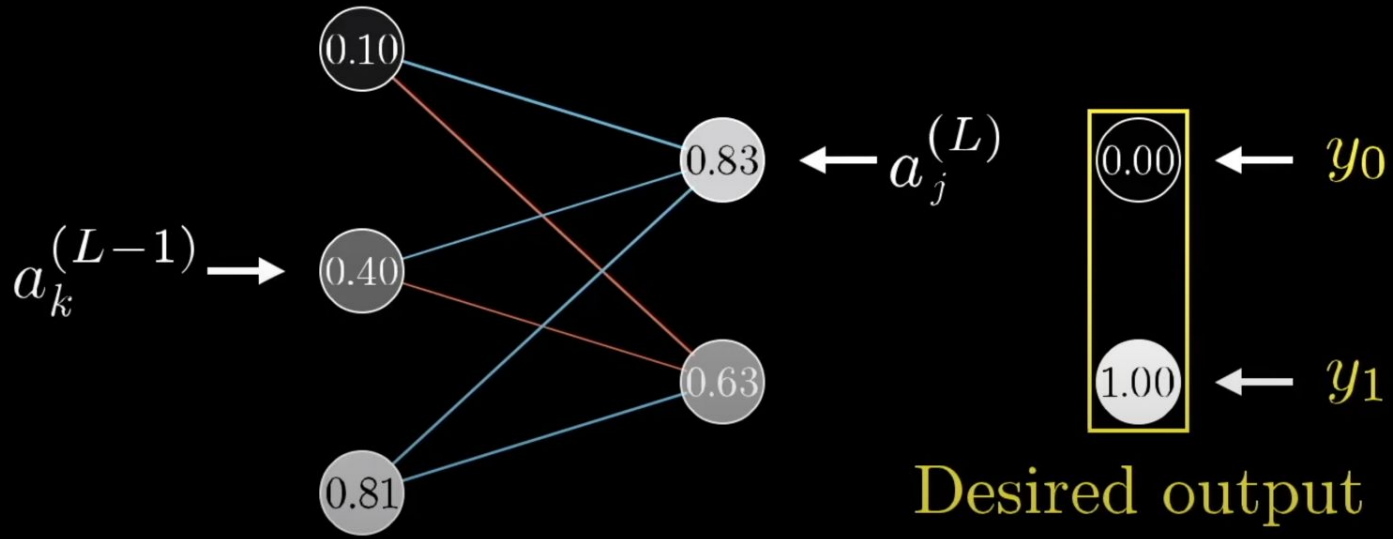
$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$



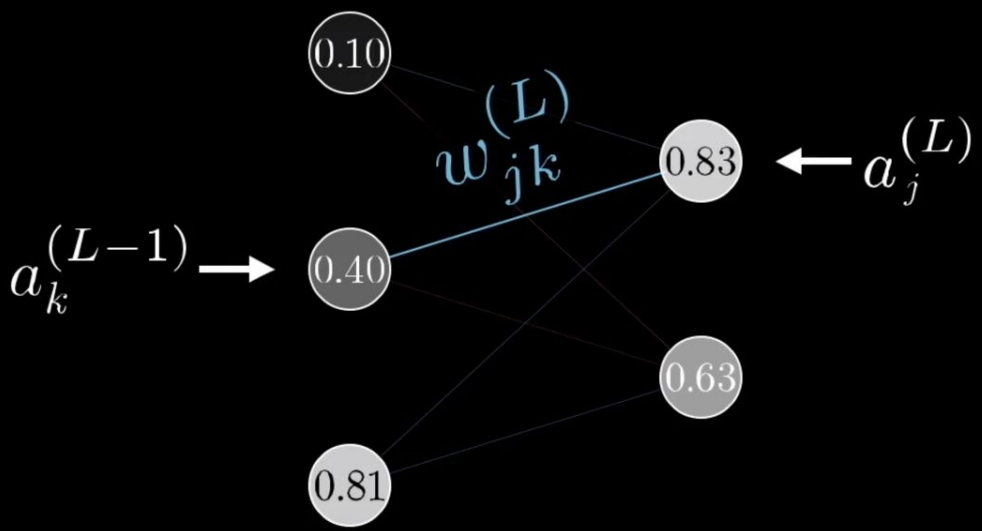

$a_0^{(L-1)}$
$a_1^{(L-1)}$
$a_2^{(L-1)}$
0.10
0.40
0.81
0.83
0.63
$a_0^{(L)}$
$a_1^{(L)}$

$$C_0 = \sum_{j=0}^{n_L-1} (a_j^{(L)} - y_j)^2$$



$$z_j^{(L)} = w_{j0}^{(L)} a_0^{(L-1)} + w_{j1}^{(L)} a_1^{(L-1)} + w_{j2}^{(L)} a_2^{(L-1)} + b_j^{(L)}$$

$$a_j^{(L)} = \sigma(z_j^{(L)})$$

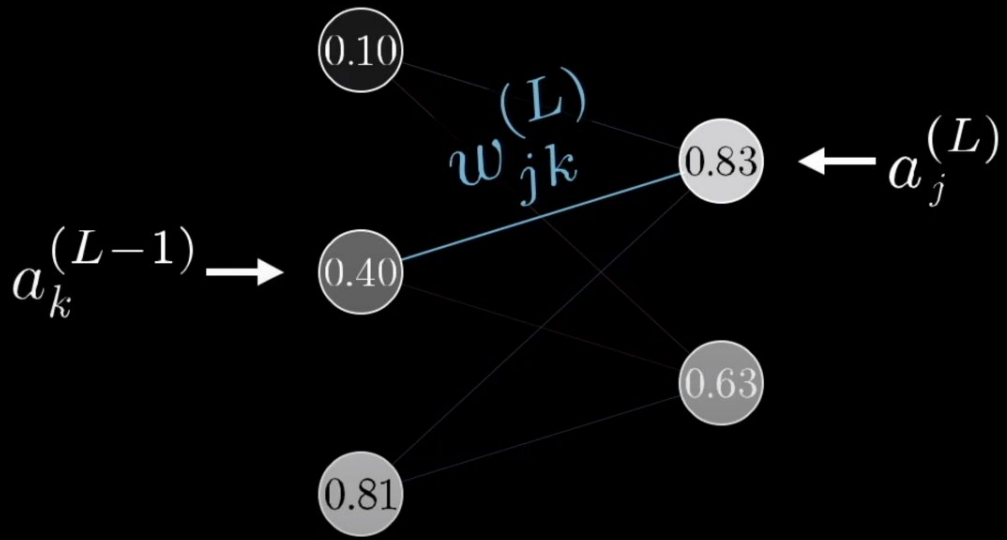


$$C_0 = \sum_{j=0}^{n_L - 1} (a_j^{(L)} - y_j)^2$$

$$\frac{\partial C_0}{\partial w_{jk}^{(L)}} = \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}$$

$$z_j^{(L)} = \cdots + w_{jk}^{(L)} a_k^{(L-1)} + \cdots$$

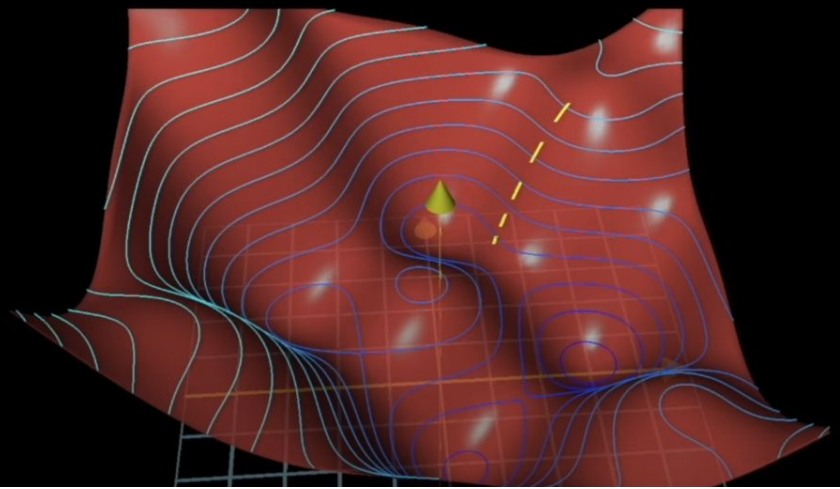
$$a_j^{(L)} = \sigma(z_j^{(L)})$$



$$C_0 = \sum_{j=0}^{n_L - 1} (a_j^{(L)} - y_j)^2$$

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = a_k^{(l-1)} \sigma'(z_j^{(l)}) \boxed{\frac{\partial C}{\partial a_j^{(l)}}}$$

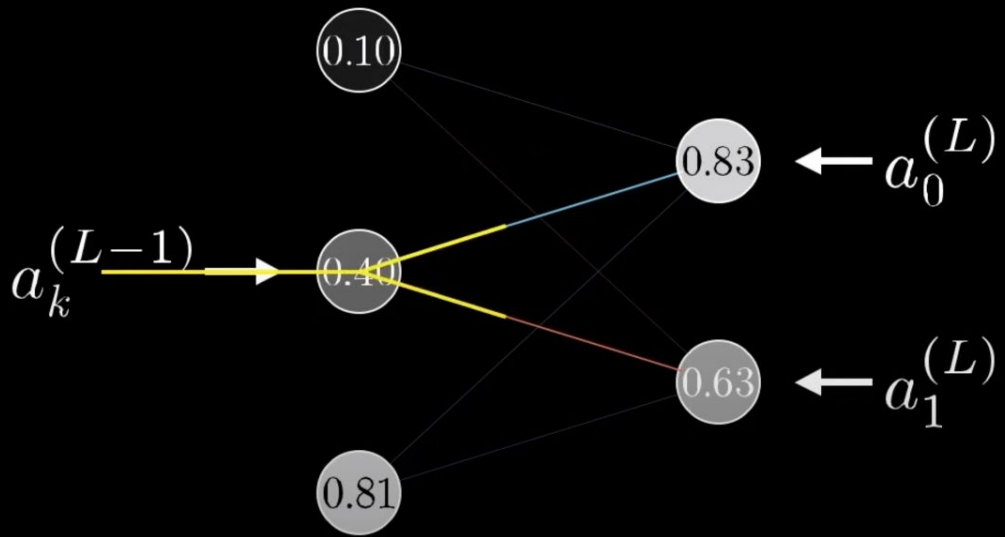
$$\nabla C \leftarrow \begin{cases} \\ \\ \\ \end{cases} \boxed{\begin{aligned} &\sum_{j=0}^{n_{l+1}-1} w_{jk}^{(l+1)} \sigma'(z_j^{(l+1)}) \frac{\partial C}{\partial a_j^{(l+1)}} \\ &\text{or} \\ &2(a_j^{(L)} - y_j) \end{aligned}}$$



$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{n_L-1} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}$$

$$z_j^{(L)} = \cdots + w_{jk}^{(L)} a_k^{(L-1)} + \cdots$$

$$a_j^{(L)} = \sigma(z_j^{(L)})$$



$$C_0 = \sum_{j=0}^{n_L-1} (a_j^{(L)} - y_j)^2$$

$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \underbrace{\sum_{j=0}^{n_L-1} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}}_{\text{Sum over layer L}}$$

$$z_j^{(L)} = \cdots + w_{jk}^{(L)} a_k^{(L-1)} + \cdots$$

$$a_j^{(L)} = \sigma(z_j^{(L)})$$

$$C_0 = \sum_{j=0}^{n_L-1} (a_j^{(L)} - y_j)^2$$