¿Por qué necesitamos modelos no lineales?

x_1	x_2	$y = x_1$ xor x_2
0	0	0
0	1	1
1	0	1
1	1	0

$n = 4$ (filas) ; $m = 2$ (columnas) ; label

* Modelo: $\hat{y}_i = x_{i,1} w_1 + x_{i,2} w_2 + b$ → ¿parámetro desconocido? → w_1, w_2, b

($\hat{y}_i$: escalar)

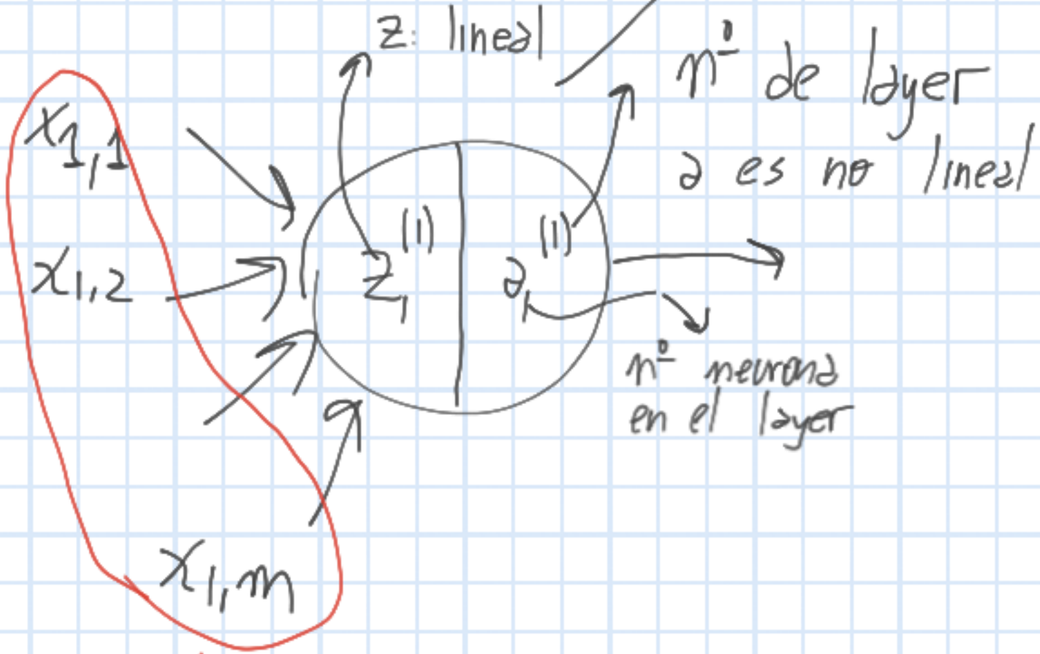
* Loss function $J(\bar{w}) = \frac{1}{4} \sum_{i=1}^{4} (y_i - \hat{y}_i)^2$; $\bar{w} = [w_1 \;\; w_2 \;\; b]^T$

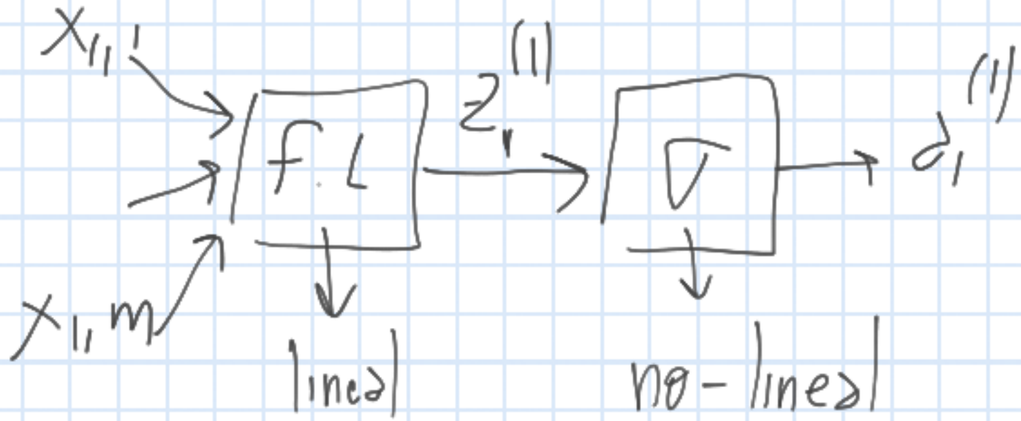
* Optimizador: → $\nabla_{\bar{w}} (J) = \bar{0}$ → $\begin{bmatrix} \partial J/\partial w_1 \\ \partial J/\partial w_2 \\ \partial J/\partial b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ → $\bar{w} = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix} = (X^T X)^{-1} X^T Y$

$= \begin{bmatrix} 0 \\ 0 \\ 1/2 \end{bmatrix}$

Redes Neuronales

* Neurona





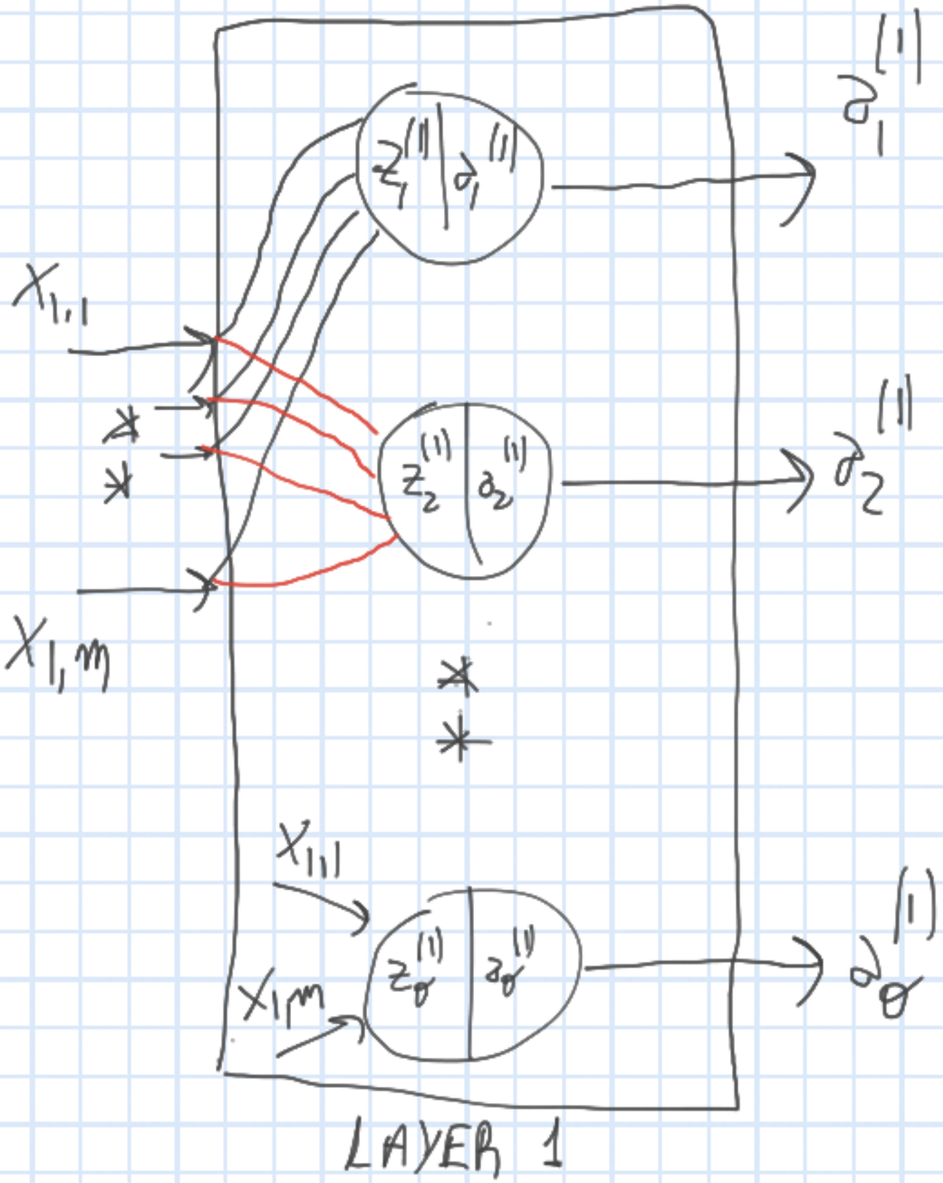
$$z_1^{(1)} = w_{1,1}^{(1)} x_{1,1} + \dots + w_{1,m}^{(1)} x_{1,m} + b_1^{(1)}$$

$$a_1^{(1)} = \nabla(z_1^{(1)}) \rightarrow \text{"activación"}$$

- sigmoid
- ReLU
- tanh

} SOFTMAX
...

* Definición de Layer

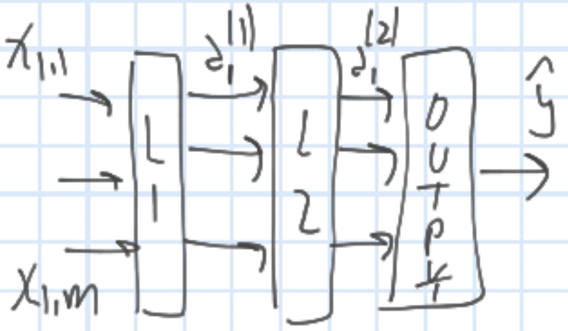


DATASET

$$\begin{cases} x_{1,1} \cdots x_{1,m} \rightarrow \bar{x}_1 \\ x_{n,1} \cdots x_{n,m} \end{cases}$$

features

Feedforward



$$a_1^{(1)} = \sigma(z_1^{(1)}) = \sigma(x_{1,1} w_{1,1} + \cdots + x_{1,m} w_{1,m} + b_1^{(1)})$$

$$\vdots$$

$$a_\theta^{(1)} = \sigma(z_\theta^{(1)}) = \sigma(x_{1,1} w_{\theta,1} + \cdots + x_{1,m} w_{\theta,m} + b_\theta^{(1)})$$

$$\Downarrow$$

$$\bar{A}^{(1)} = \sigma(\bar{Z}^{(1)})$$

$$\bar{Z}^{(1)} = W^{(1)} \cdot \bar{X}_1 + \bar{b}^{(1)}$$

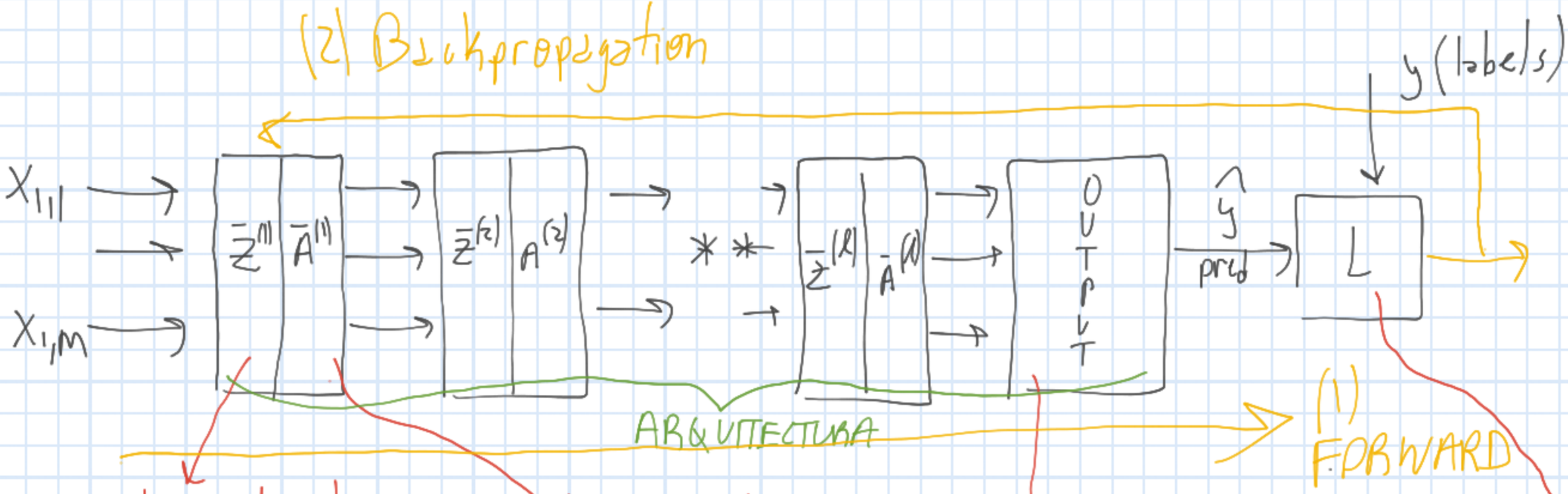
$\bar{Z}^{(1)} \in \mathbb{R}^{\theta \times 1}$ $W^{(1)} \in \mathbb{R}^{\theta \times m}$ $\bar{X}_1 \in \mathbb{R}^{m \times 1}$ $\bar{b}^{(1)} \in \mathbb{R}^{\theta \times 1}$

¿parametros?

$(\theta * m) + \theta \times 1$

$\theta * (m+1)$

(2) Backpropagation



layer lineal

layer convolucional
layer recurrente
layer embedding
layer 2tHention

sigmoid
ReLU
tanh

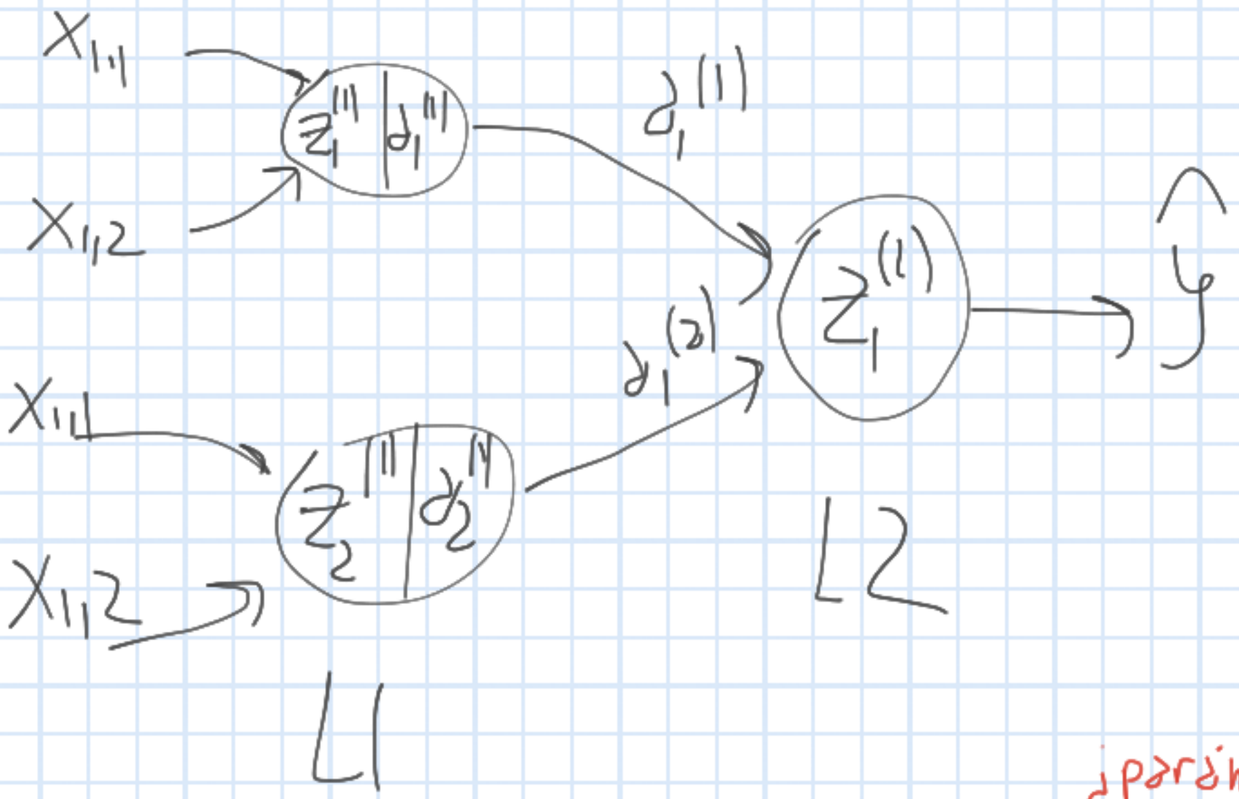
sigmoid (clasif binaria)
SOFTMAX (clasif multi-clase)

SIN SALIDA (regresion)

BCE
CE

MSE

* Resolver XOR con redes neu.



¿parámetros?

3+3+3=9

* Forward

$$z_{1,1}^{(1)} = x_{1,1}\, w_{1,1}^{(1)} + x_{1,2}\, w_{1,2}^{(1)} + b_1^{(1)}$$

$$a_{1,1}^{(1)} = \sigma\left(z_{1,1}^{(1)}\right) = \frac{1}{1+e^{-z_{1,1}^{(1)}}}$$

$$z_{1,2}^{(1)} = x_{1,1}\, w_{2,1}^{(1)} + x_{1,2}\, w_{2,2}^{(1)} + b_2^{(1)}$$

$$a_{1,2}^{(1)} = \sigma\left(z_{1,2}^{(1)}\right)$$

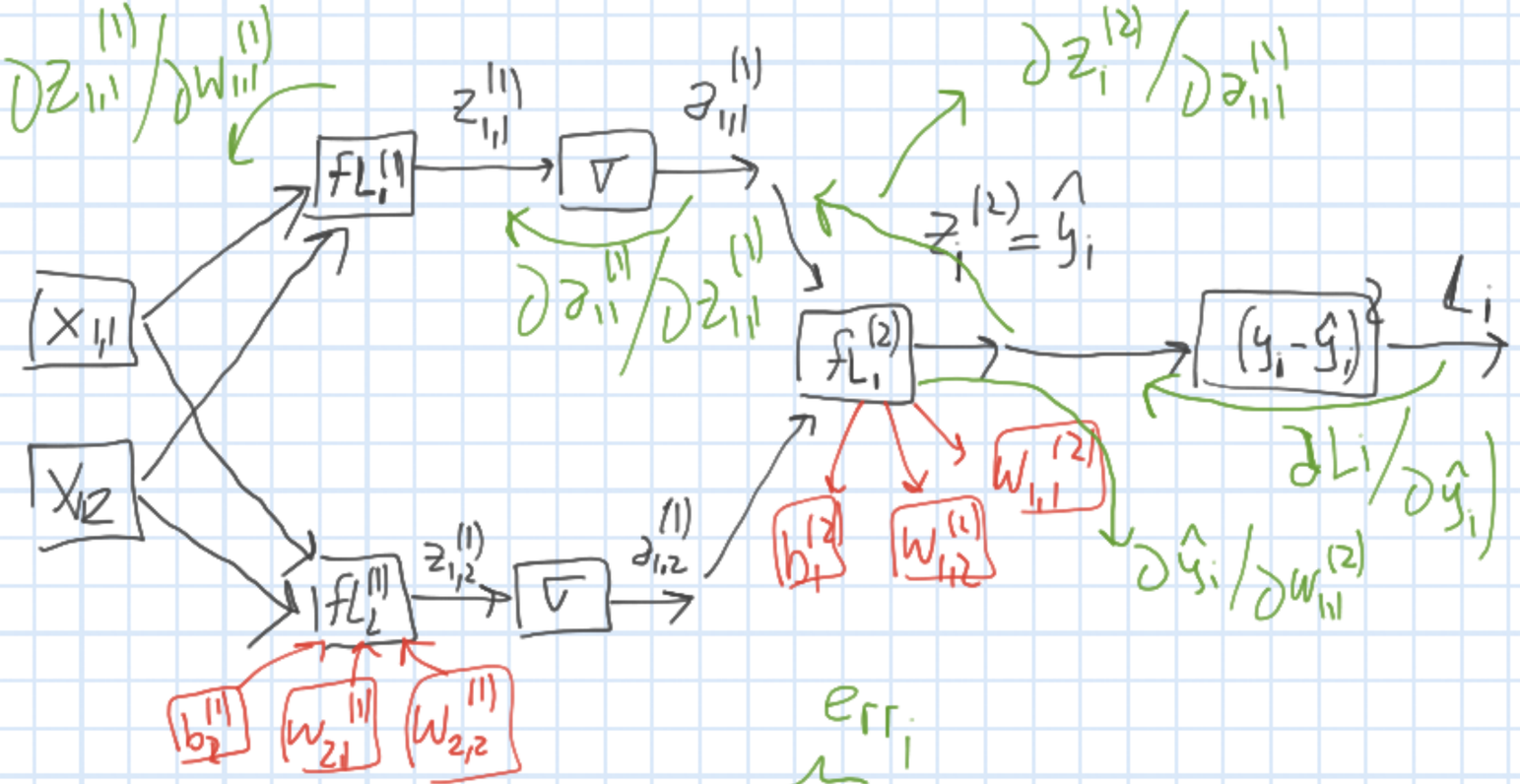
$$z_{1,1}^{(2)} = a_{1,1}^{(1)}\, w_{1,1}^{(2)} + a_{1,2}^{(1)}\, w_{2,1}^{(2)} + b_1^{(2)}$$

$$\hat{y}_1 = z_{1,1}^{(2)}$$

$$L_i = \left(y_i - \hat{y}_i\right)$$

* Backpropagation

$$\left.\begin{aligned}
&\frac{\partial L}{\partial w_{1,1}^{(1)}} = \\
&\vdots \\
&\frac{\partial L}{\partial w_{2,1}^{(1)}} = \\
&\frac{\partial L}{\partial b_1^{(1)}} = \\
&\partial L / \partial b_2^{(1)} = \\
&\vdots \\
&\partial L / \partial w_{1,2}^{(1)} =
\end{aligned}\right\} \text{9 ecuaciones}$$



$$\frac{\partial L_i}{\partial w_{1,1}^{(2)}} = \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w_{1,1}^{(L)}} = \underbrace{2\,(y_i - \hat{y}_i)\,(-1)}_{err_i}\; \underbrace{a_{1,1}}_{z}$$

$$\frac{\partial L_i}{\partial w_{1,1}^{(1)}} = \frac{\partial L_i}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_1^{(1)}} \frac{\partial z_1^{(1)}}{\partial a_1^{(1)}} \frac{\partial a_1^{(1)}}{\partial w_{1,1}^{(1)}} = -2\, err_i\, w_{1,1}^{(2)}\, \sigma(z_{1,1}^{(1)})\,(1 - \sigma(z_{1,1}^{(1)}))\, x_{1,1}$$

$$\bar{W} = \begin{pmatrix} w_{111}^{(1)} \\ w_{112}^{(1)} \\ w_{211}^{(1)} \\ w_{212}^{(1)} \\ b_1^{(1)} \\ b_2^{(1)} \\ w_{111}^{(2)} \\ w_{112}^{(2)} \\ b_1^{(2)} \end{pmatrix} = \text{Iniciar en valores random}$$

Iterar usando SGD

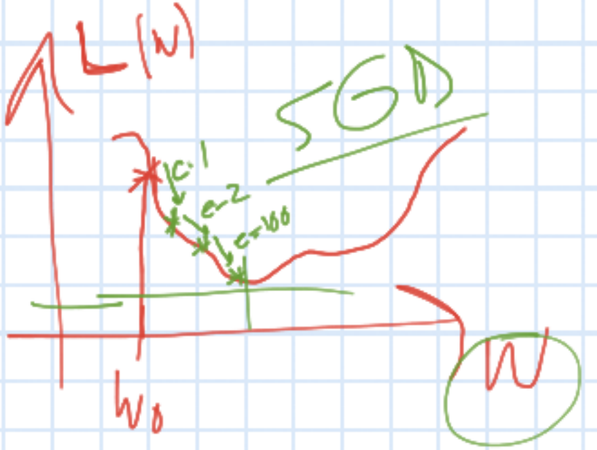
$$\bar{W}_{t+1} \leftarrow \bar{W}_t - \alpha \nabla_{\bar{W}}(J)$$

learning rate

calculamos

$$\begin{pmatrix} w_{111}^{(1)} \\ \\ b_1^{(2)} \end{pmatrix}_{t+1} \leftarrow \begin{pmatrix} w_{111}^{(1)}{}_t - \alpha \partial L / \partial w_{111}^{(1)} \\ \vdots \end{pmatrix}_t$$

Algoritmo SGD



* Iniciar 9 params random → $w_{11}^{(1)}, w_{12}^{(1)}, \dots$
* for epoch in range(100): → itero por cada muestra
    for i in range(4):
        (1) Forward $X_1 = X[1,1]$; $X_2 = X[i,2]$; $y_i = Y[i]$ → $\hat{y}_i$
        (2) err = $y_i - \hat{y}_i$
        (3) Backpropagation → 9 der. parciales
        (4) Update pesos → $\bar{w} = \bar{w} - \alpha \nabla(3)$
    (5) Calcular $MSE = \frac{1}{4}\Sigma(y-\hat{y})^2$ →