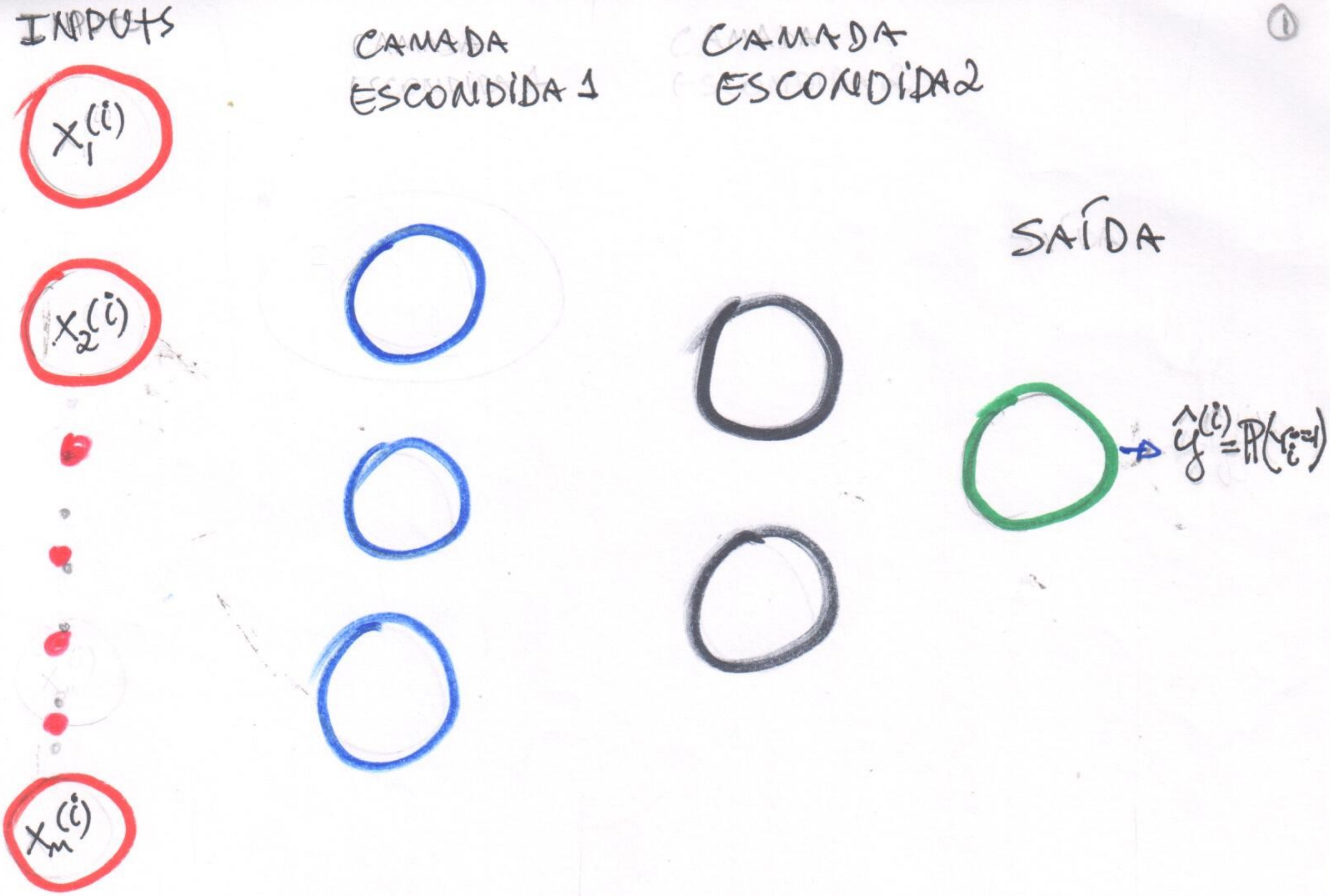
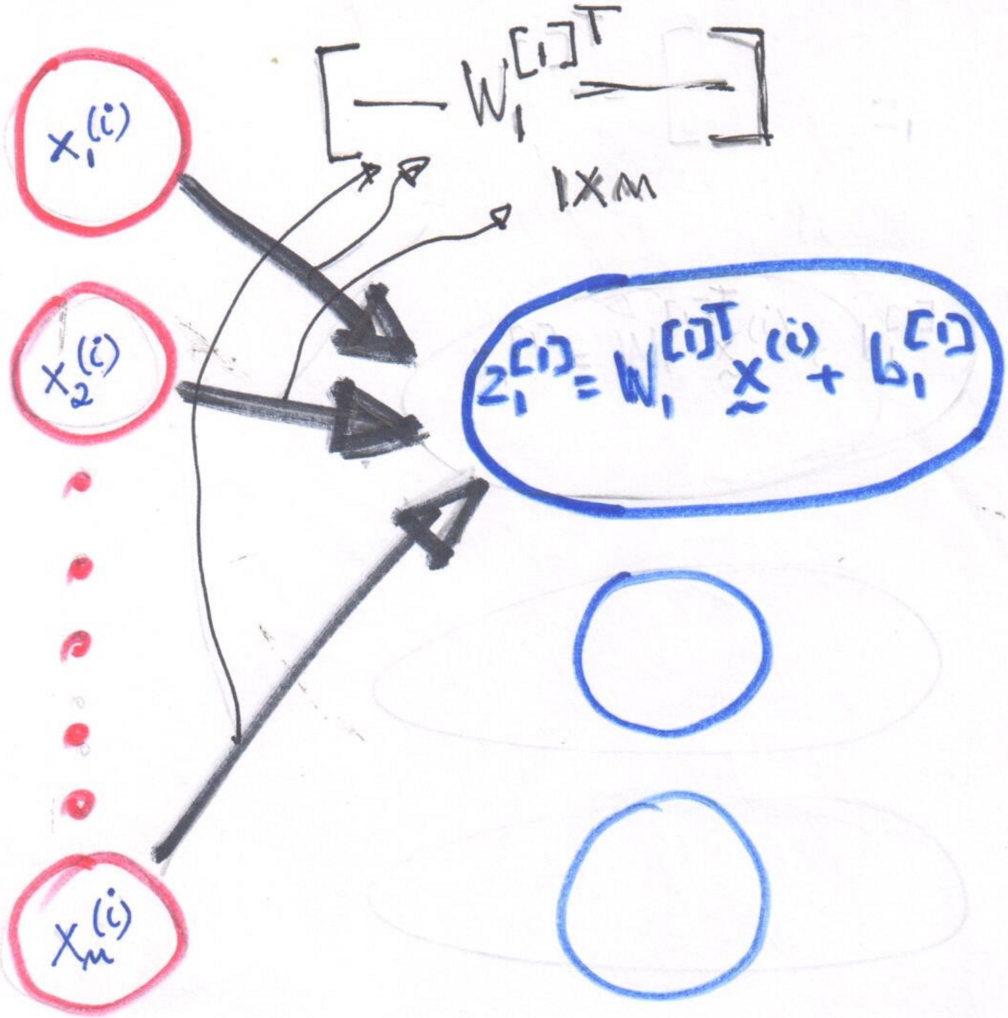
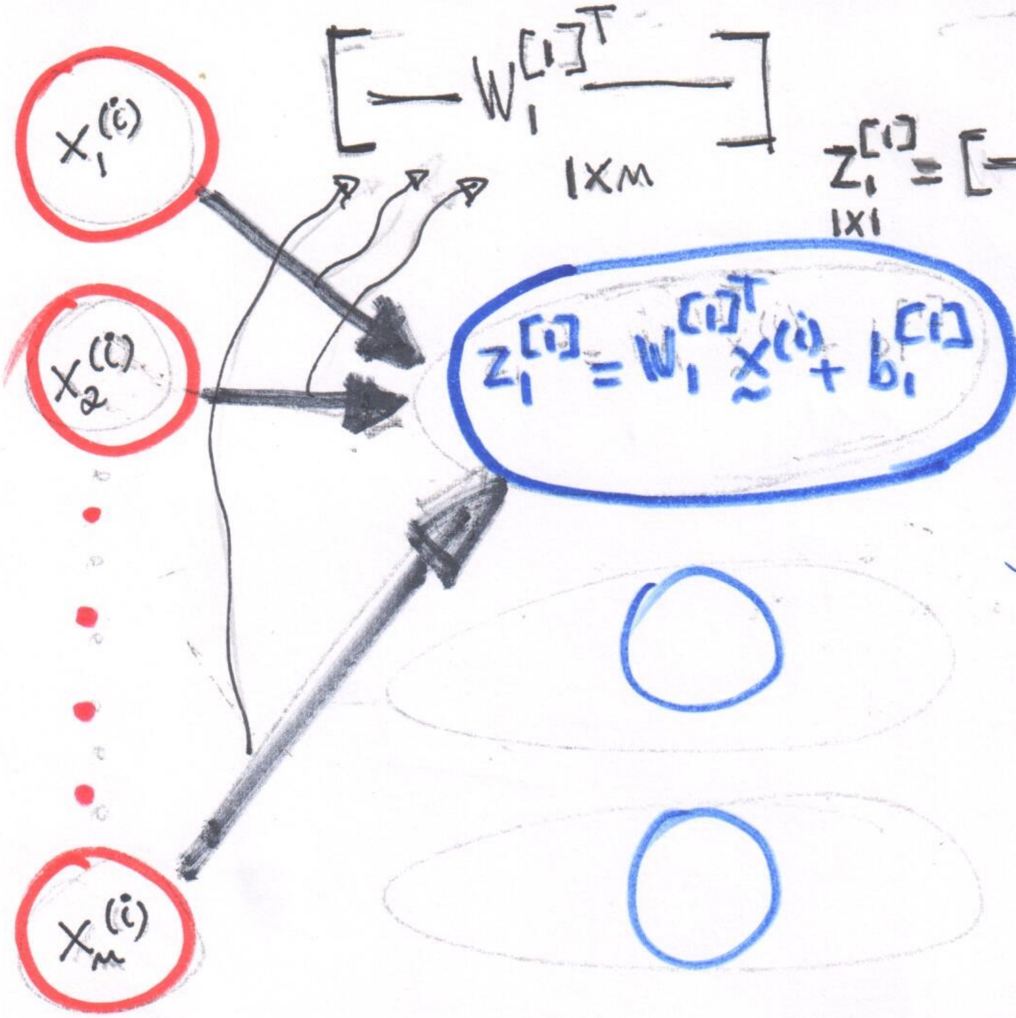

INPUTS
CAMADA ESCONDIDA 1
CAMADA ESCONDIDA 2
SAÍDA
$x_1^{(i)}$
$x_2^{(i)}$
$x_m^{(i)}$
$\hat{y}^{(i)} = \mathbb{P}(y_i = 1)$

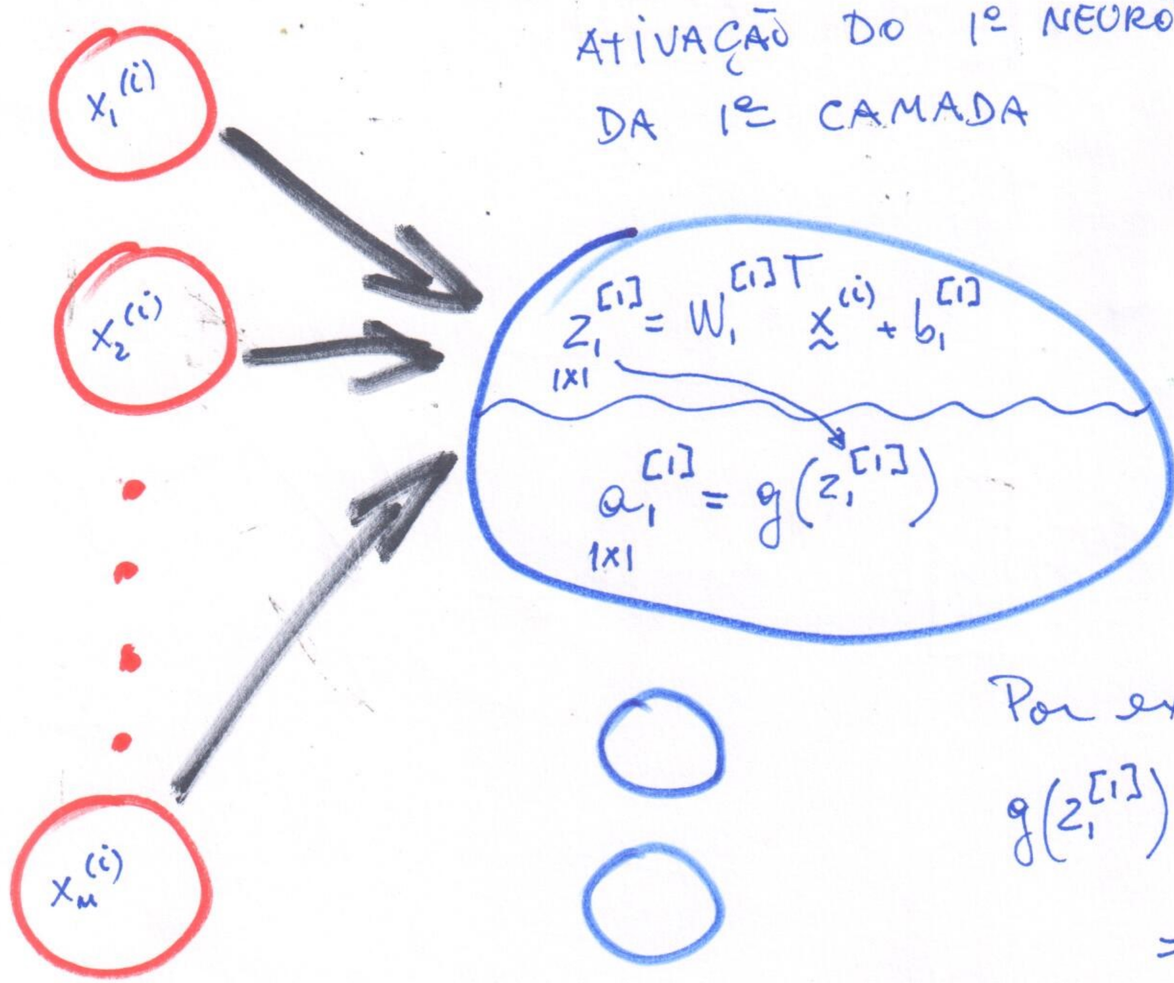

$[\text{—} W_1^{[1]T} \text{—}]$
$1 \times m$
$x_1^{(i)}$
$x_2^{(i)}$
$x_m^{(i)}$
$z_1^{[1]} = W_1^{[1]T} \underset{\sim}{x}^{(i)} + b_1^{[1]}$



$$\underset{1\times 1}{z_1^{[1]}} = \underset{1\times M}{\left[— W_1^{[1]T} — \right]} \cdot \underset{M\times 1}{\begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_M^{(i)} \end{bmatrix}} + \underset{1\times 1}{b_1^{[1]}}$$

$$\underbrace{\left[— W_1^{[1]} — \right] \text{ e } b_1^{[1]}}_{\text{PARÂMETROS DESCONHECIDOS}}$$

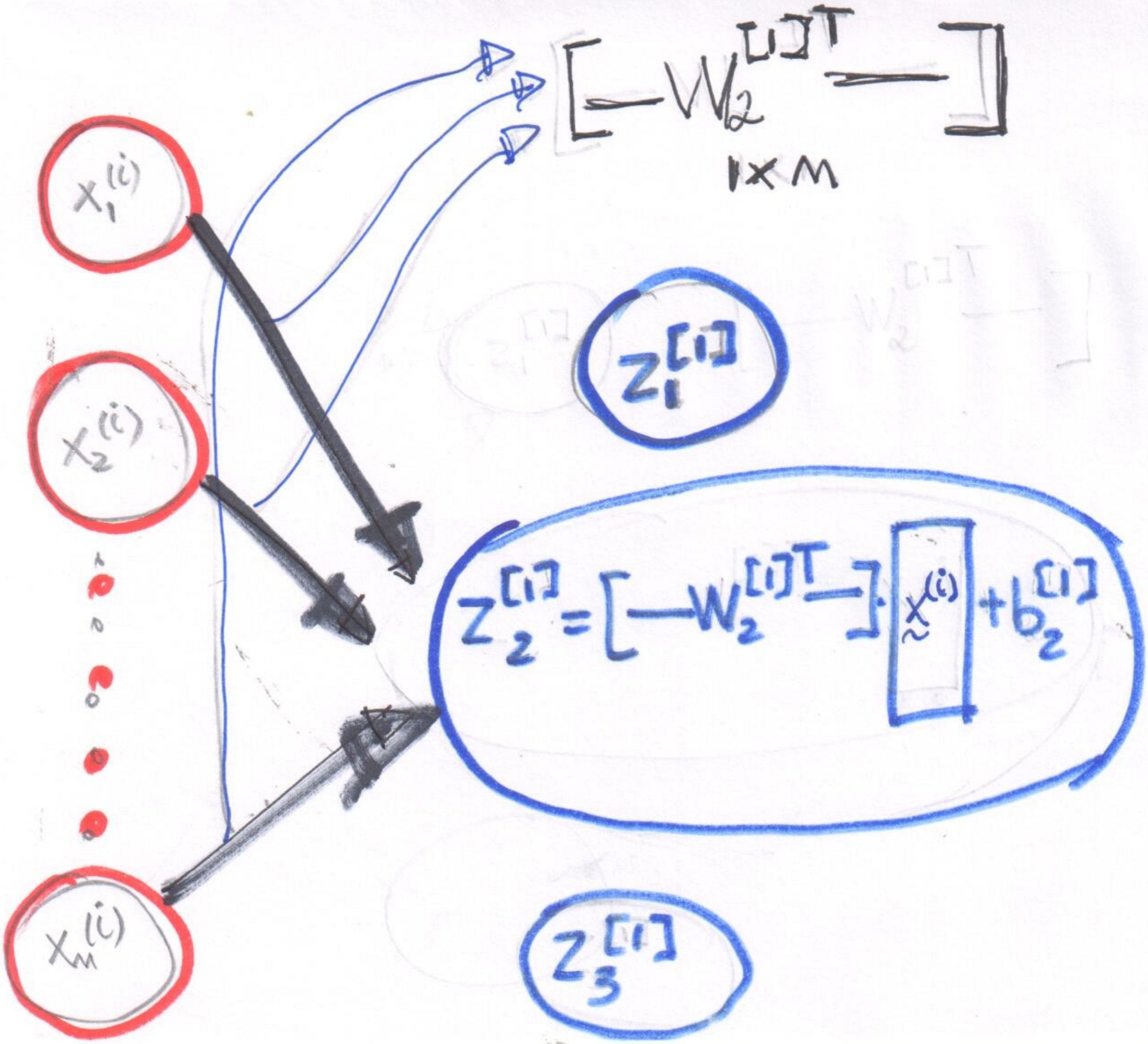
ATIVAÇÃO DO 1º NEURÔNIO DA 1ª CAMADA



NÃO ENVOLVE PARÂMETROS DESCONHECIDOS

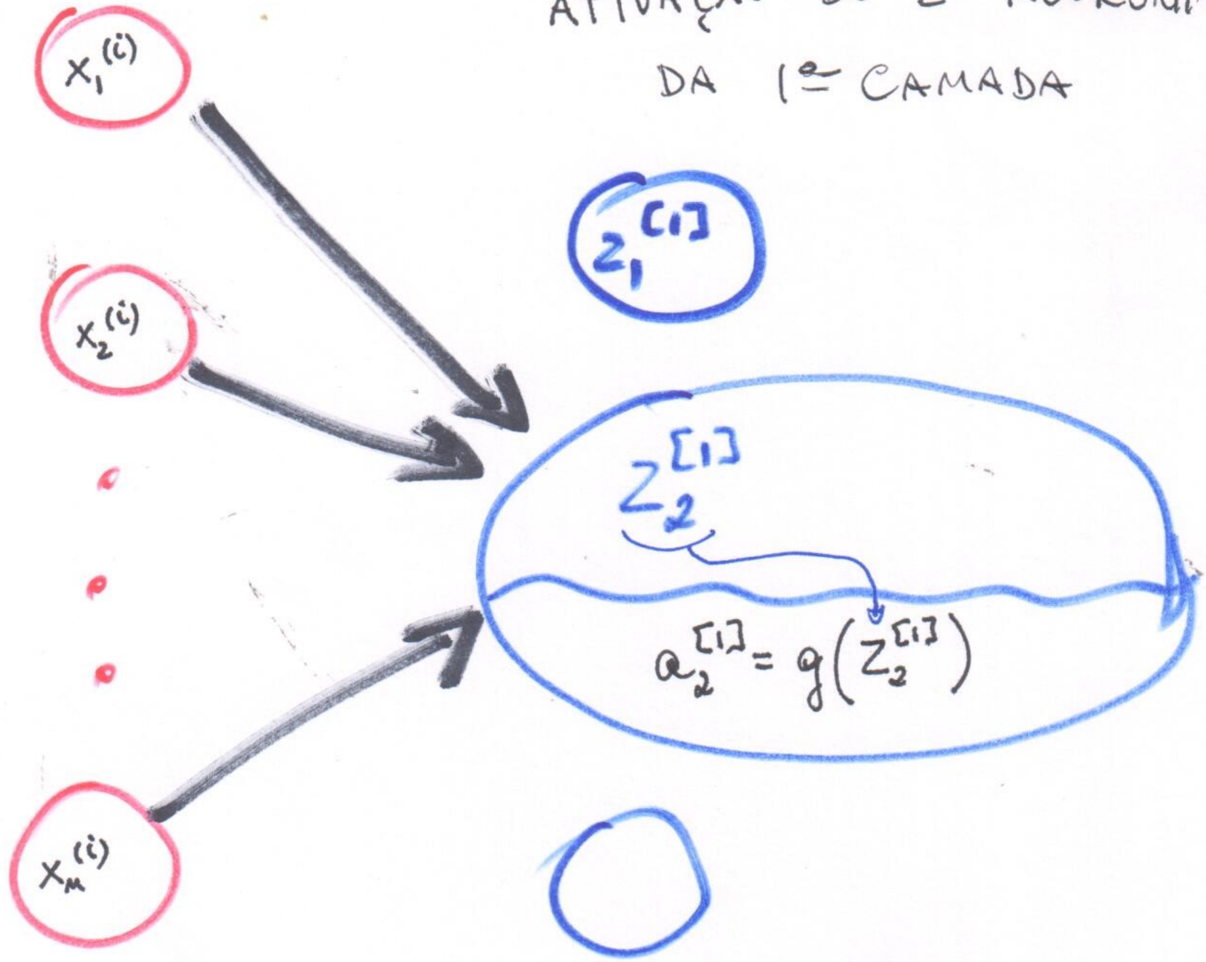
Por exemplo,

$$g(z_1^{[1]}) = ReLU(z_1^{[1]}) = \begin{cases} 0 \text{ se } z_1^{[1]} < 0 \\ z_1^{[1]} \text{ se } z_1^{[1]} \geq 0 \end{cases}$$

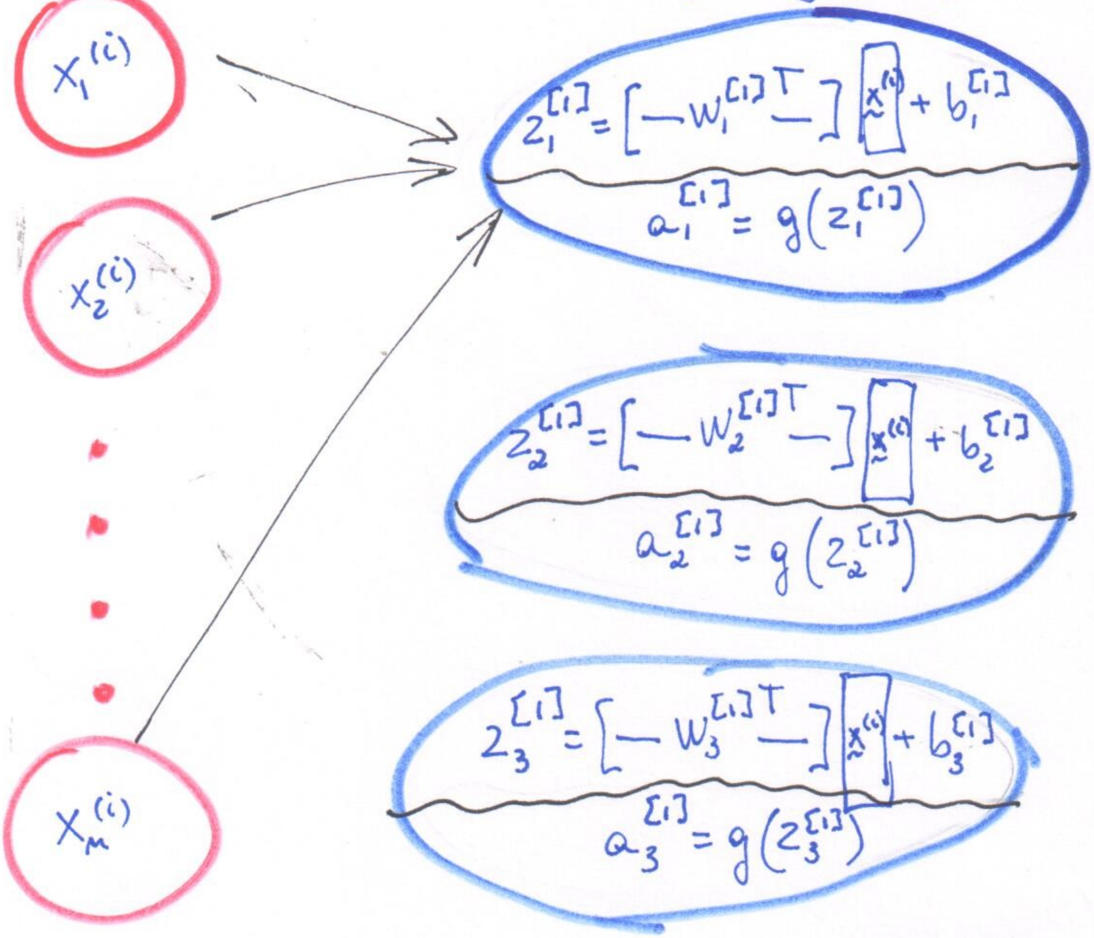


$\left[— W_2^{[1]T} — \right]$
$1 \times M$
$x_1^{(i)}$
$x_2^{(i)}$
$x_M^{(i)}$
$z_1^{[1]}$
$z_2^{[1]} = \left[— W_2^{[1]T} — \right] \cdot \left[\underset{\sim}{x}^{(i)} \right] + b_2^{[1]}$
$z_3^{[1]}$

ATIVAÇÃO DO 2º NEURÔNIO DA 1ª CAMADA

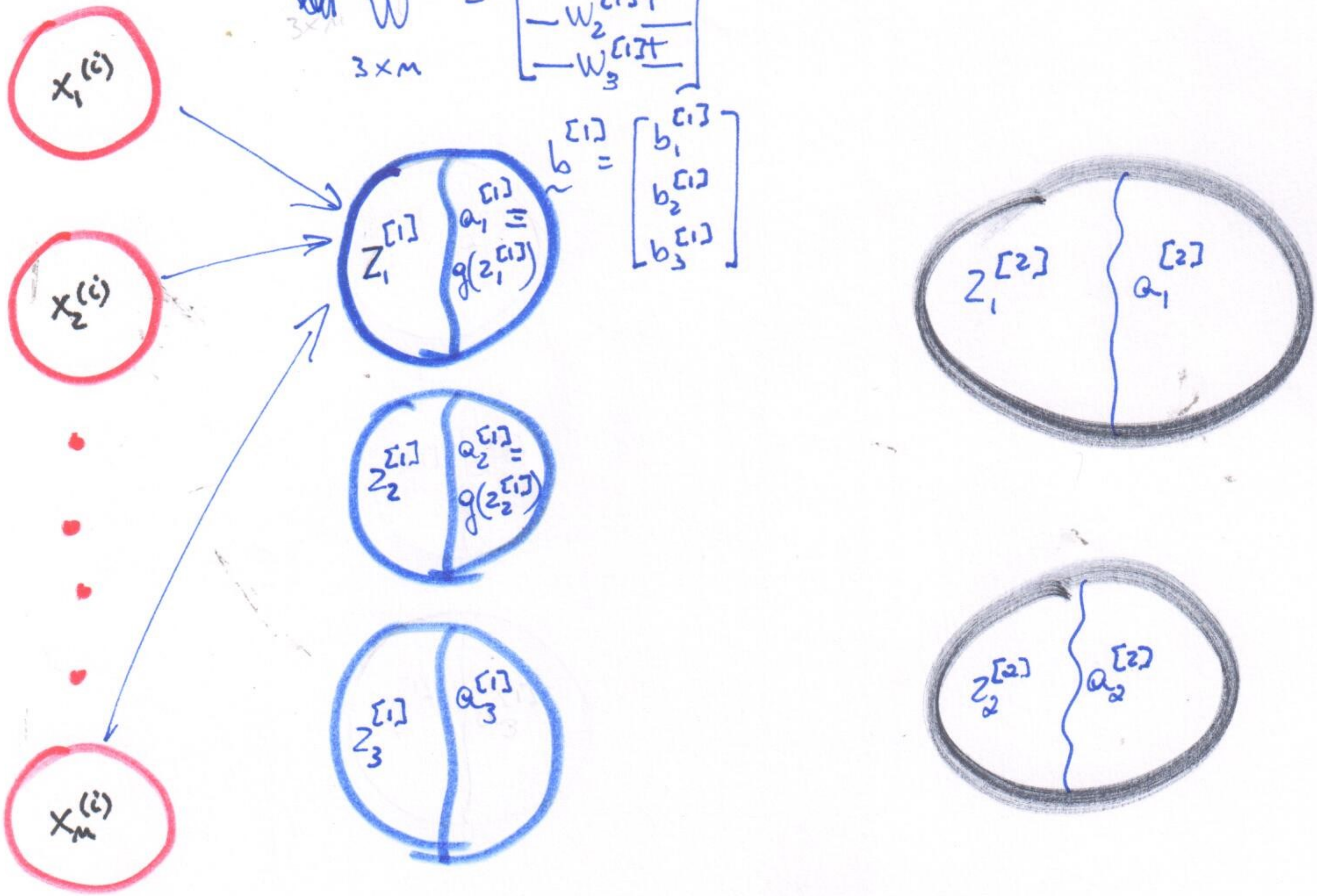


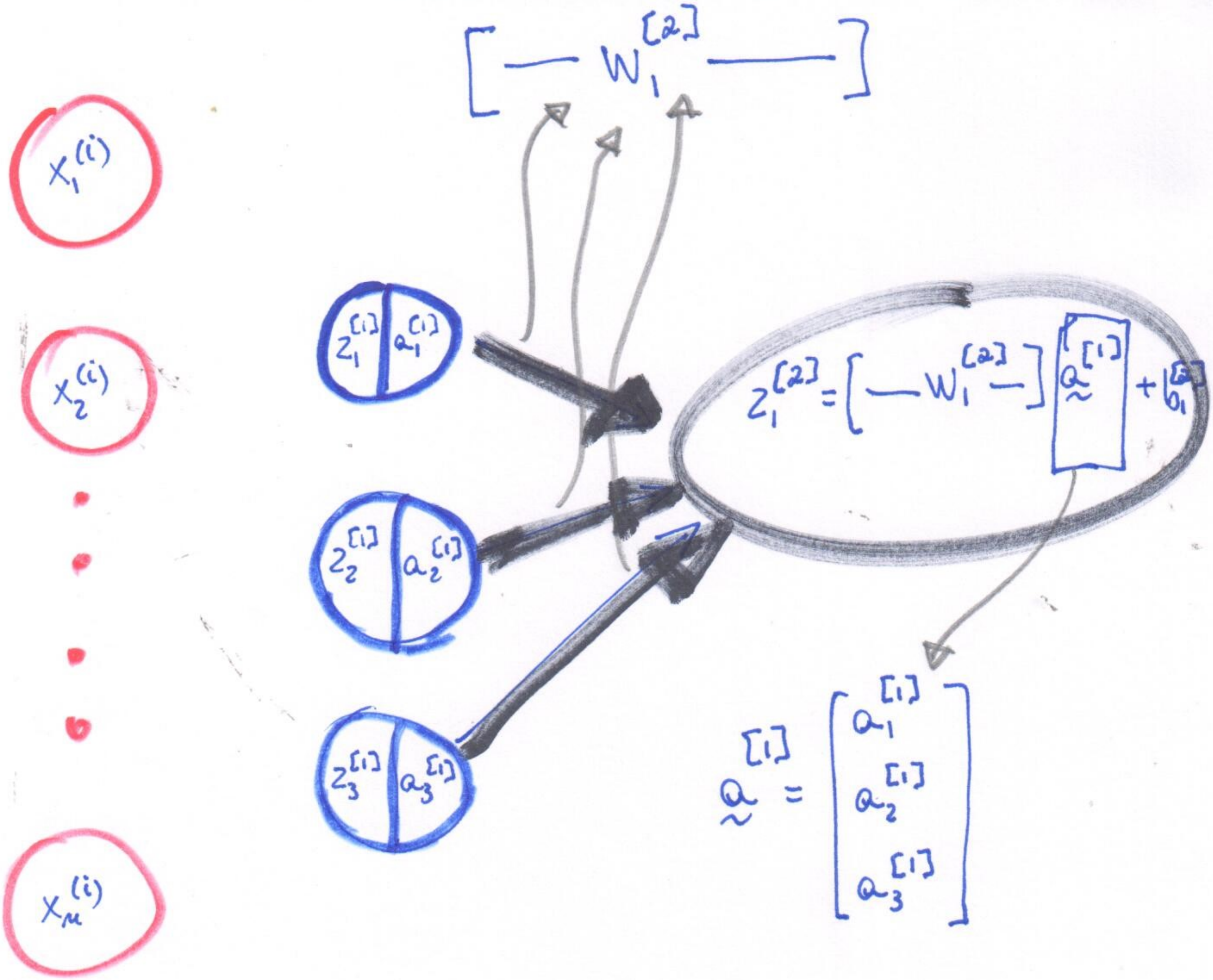
$$\underset{3\times m}{W^{[1]}} = \begin{bmatrix} — w_1^{[1]T} — \\ — w_2^{[1]T} — \\ — w_3^{[1]T} — \end{bmatrix} \qquad \underset{3\times 1}{b^{[1]}} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

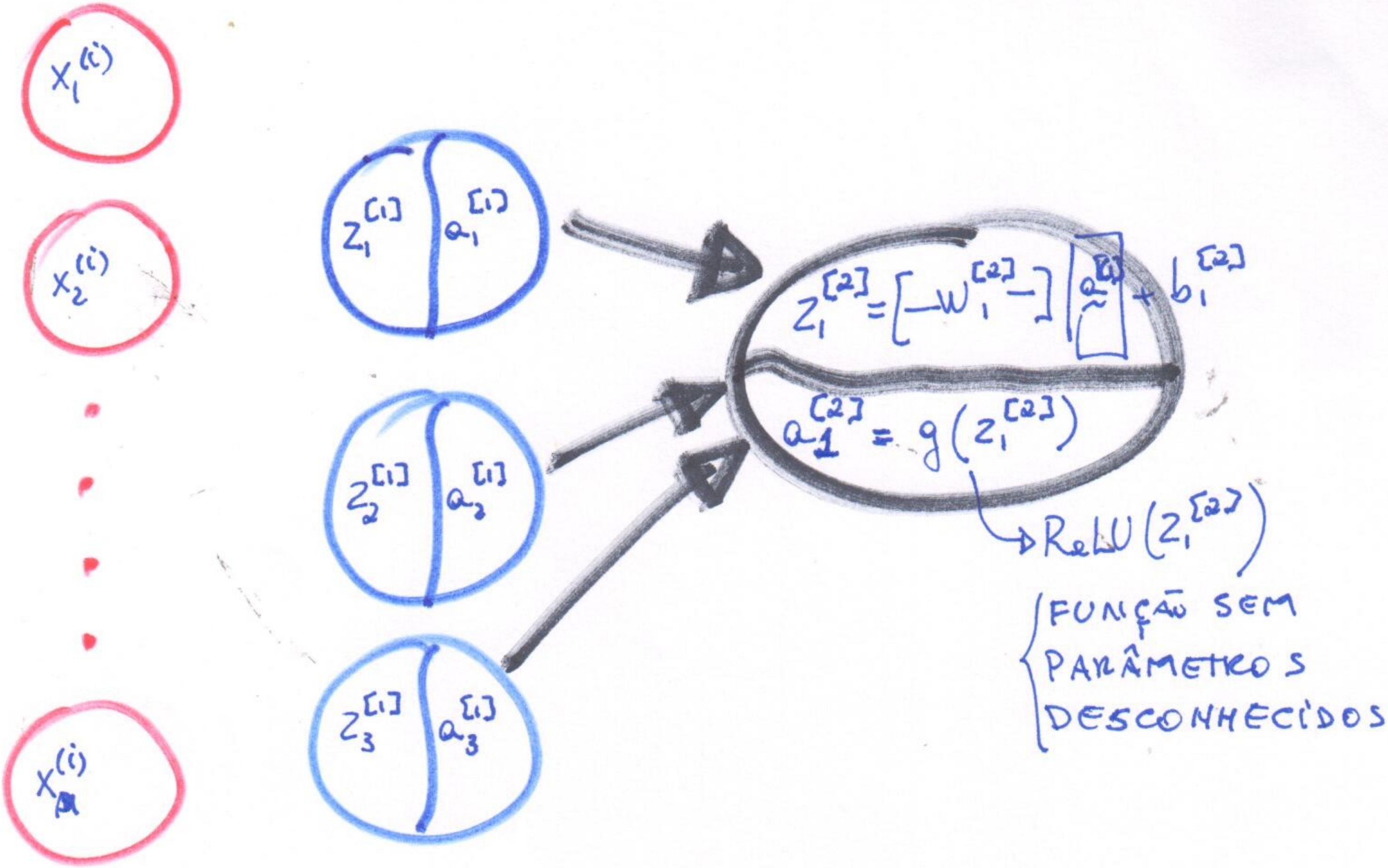


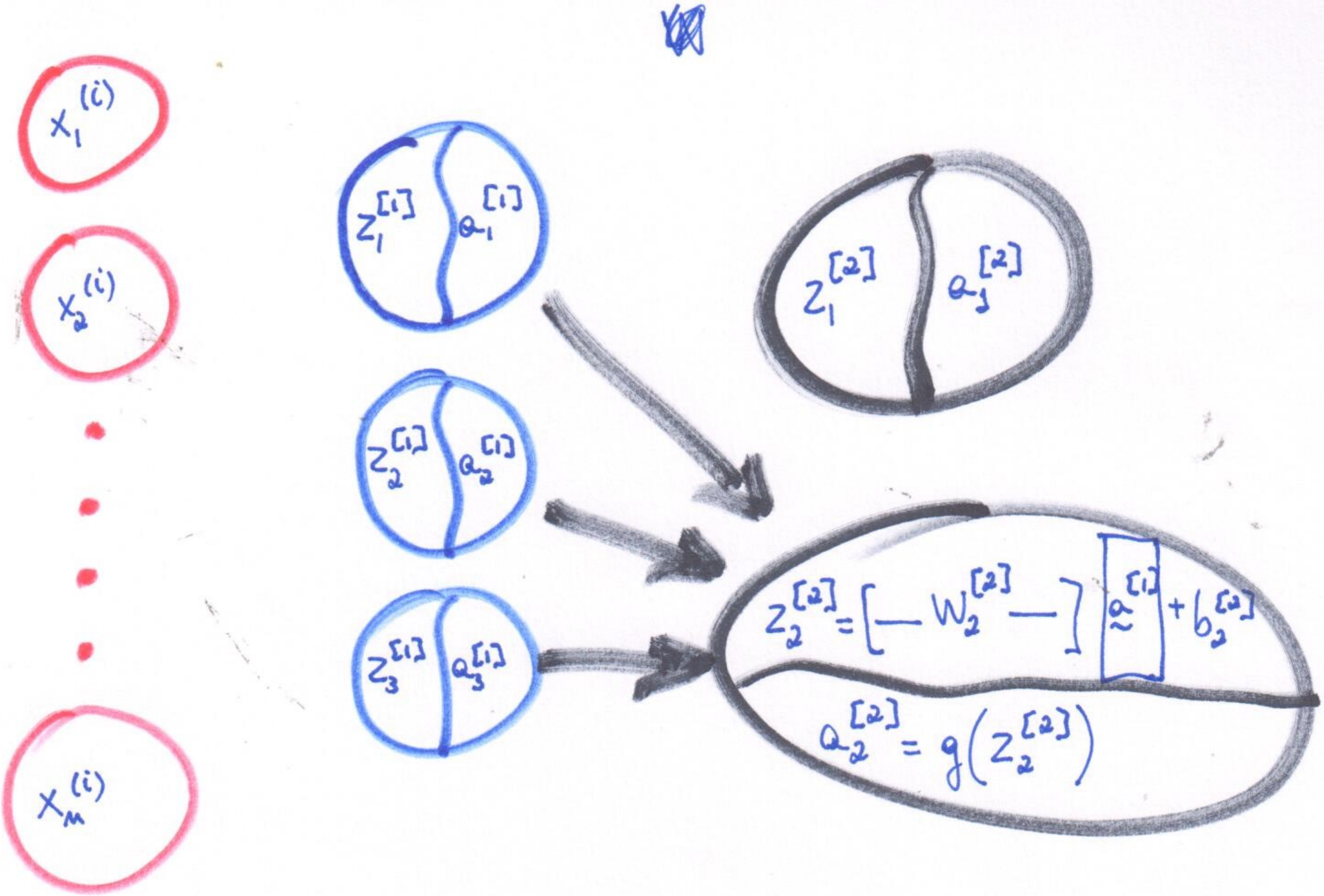
$$\underset{3\times m}{W^{[1]}} = \begin{bmatrix} — w_1^{[1]T} — \\ — w_2^{[1]T} — \\ — w_3^{[1]T} — \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$



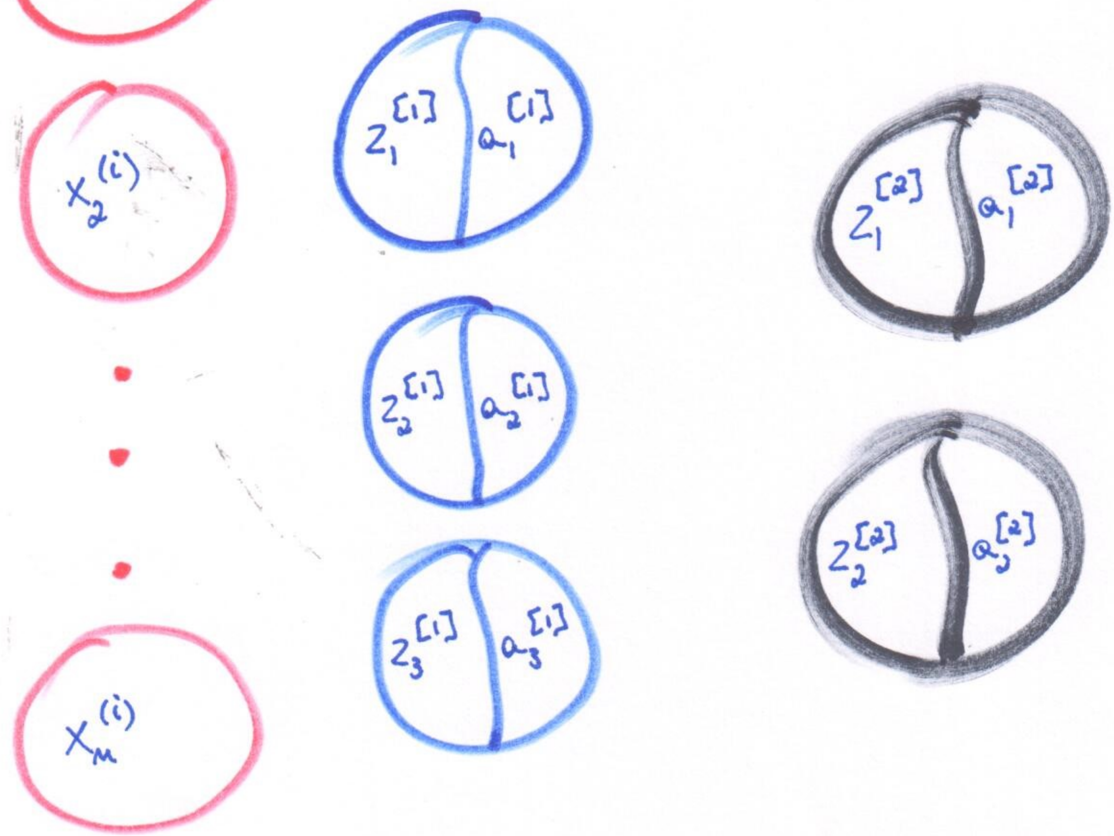

[— W_1^[2] —]
x_1^(i)
x_2^(i)
x_m^(i)
z_1^[1] | a_1^[1]
z_2^[1] | a_2^[1]
z_3^[1] | a_3^[1]
z_1^[2] = [— W_1^[2] —] a^[1] + b_1^[2]
a^[1] = [a_1^[1], a_2^[1], a_3^[1]]


$x_1^{(i)}$
$x_2^{(i)}$
$x_n^{(i)}$
$z_1^{[1]}$ | $a_1^{[1]}$
$z_2^{[1]}$ | $a_2^{[1]}$
$z_3^{[1]}$ | $a_3^{[1]}$
$z_1^{[2]} = [-w_1^{[2]}-] [\underset{\sim}{a}^{[1]}] + b_1^{[2]}$
$a_1^{[2]} = g(z_1^{[2]})$
$\rightarrow \text{ReLU}(z_1^{[2]})$
FUNÇÃO SEM
PARÂMETROS
DESCONHECIDOS

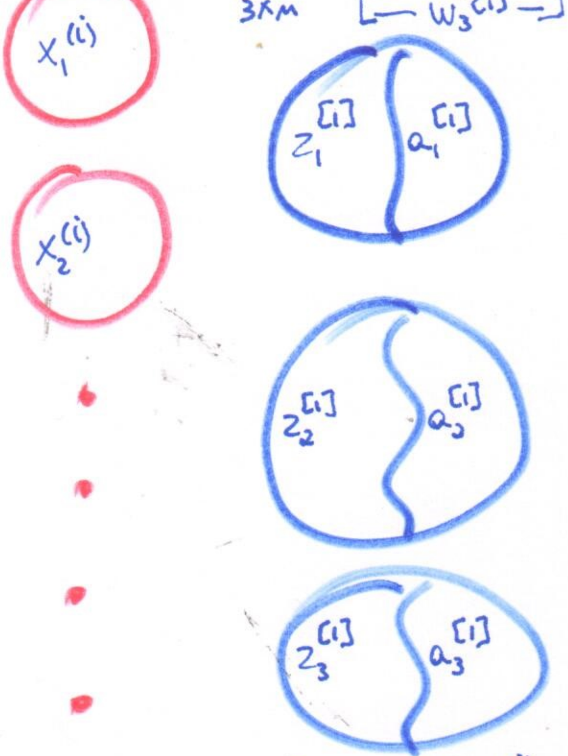


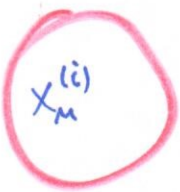
$x_1^{(i)}$
$x_2^{(i)}$
$x_m^{(i)}$
$z_1^{[1]}$ $a_1^{[1]}$
$z_2^{[1]}$ $a_2^{[1]}$
$z_3^{[1]}$ $a_3^{[1]}$
$z_1^{[2]}$ $a_1^{[2]}$
$z_2^{[2]} = \left[— W_2^{[2]} — \right] \boxed{\underset{\sim}{a}^{[1]}} + b_2^{[2]}$
$a_2^{[2]} = g\left(z_2^{[2]}\right)$

$$\underset{2\times 3}{W^{[2]}} = \begin{bmatrix} - w_1^{[2]} - \\ - w_2^{[2]} - \end{bmatrix} \qquad b^{[2]} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \end{bmatrix}$$



$$\underset{3\times m}{W^{[1]}} = \begin{bmatrix} - & w_1^{[1]} & - \\ - & w_2^{[1]} & - \\ - & w_3^{[1]} & - \end{bmatrix} \qquad \underset{\sim}{b}^{[1]} = \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{bmatrix}$$

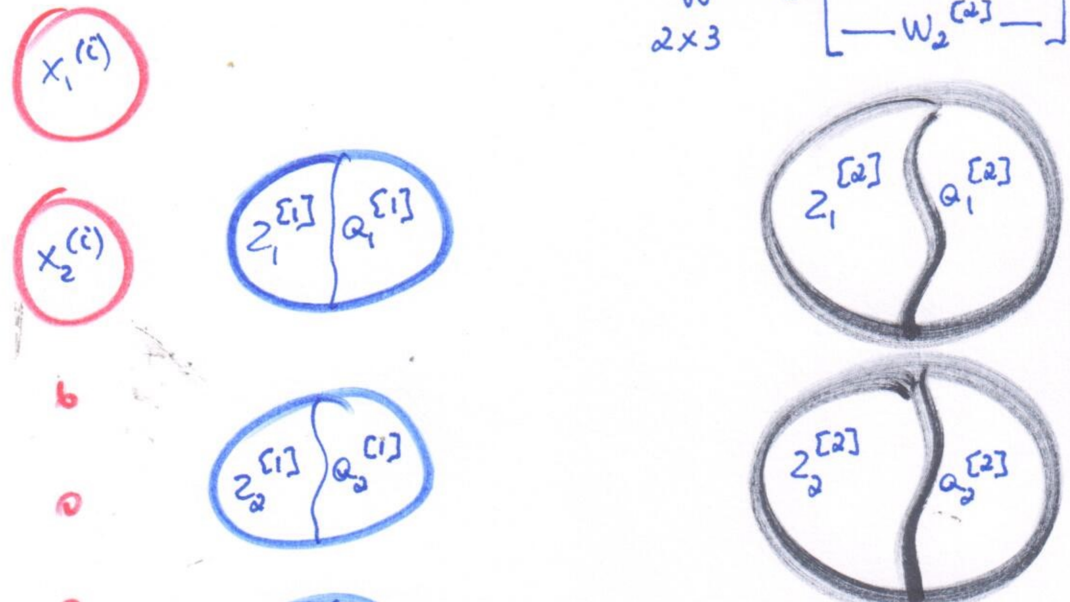


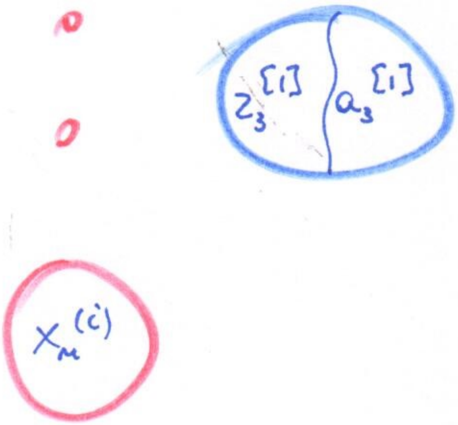


$$\underset{3\times 1}{\underset{\sim}{z}^{[1]}} = \underset{3\times m}{W^{[1]}} \cdot \underset{m\times 1}{\underset{\sim}{x}^{(i)}} + \underset{3\times 1}{b^{[1]}}$$

$$\underset{3\times 1}{\underset{\sim}{a}^{[1]}} = \begin{bmatrix} g(z_1^{[1]}) \\ g(z_2^{[1]}) \\ g(z_3^{[1]}) \end{bmatrix} = g\left(\underset{\sim}{z}^{[1]}\right)$$

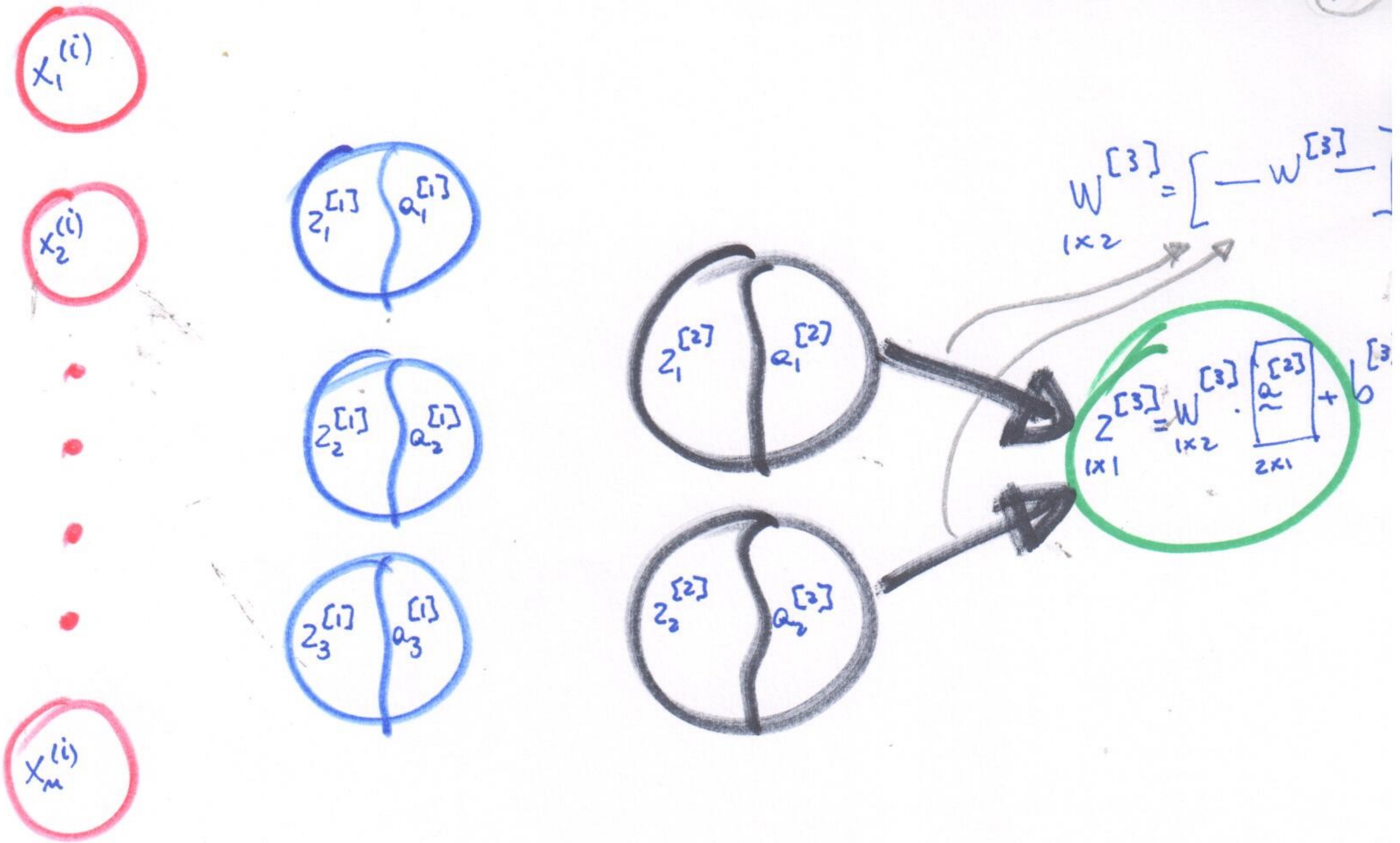
$$\underset{2\times 3}{W^{[2]}} = \begin{bmatrix} \text{---}\, w_1^{[2]} \,\text{---} \\ \text{---}\, w_2^{[2]} \,\text{---} \end{bmatrix} \qquad b^{[2]} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \end{bmatrix}$$



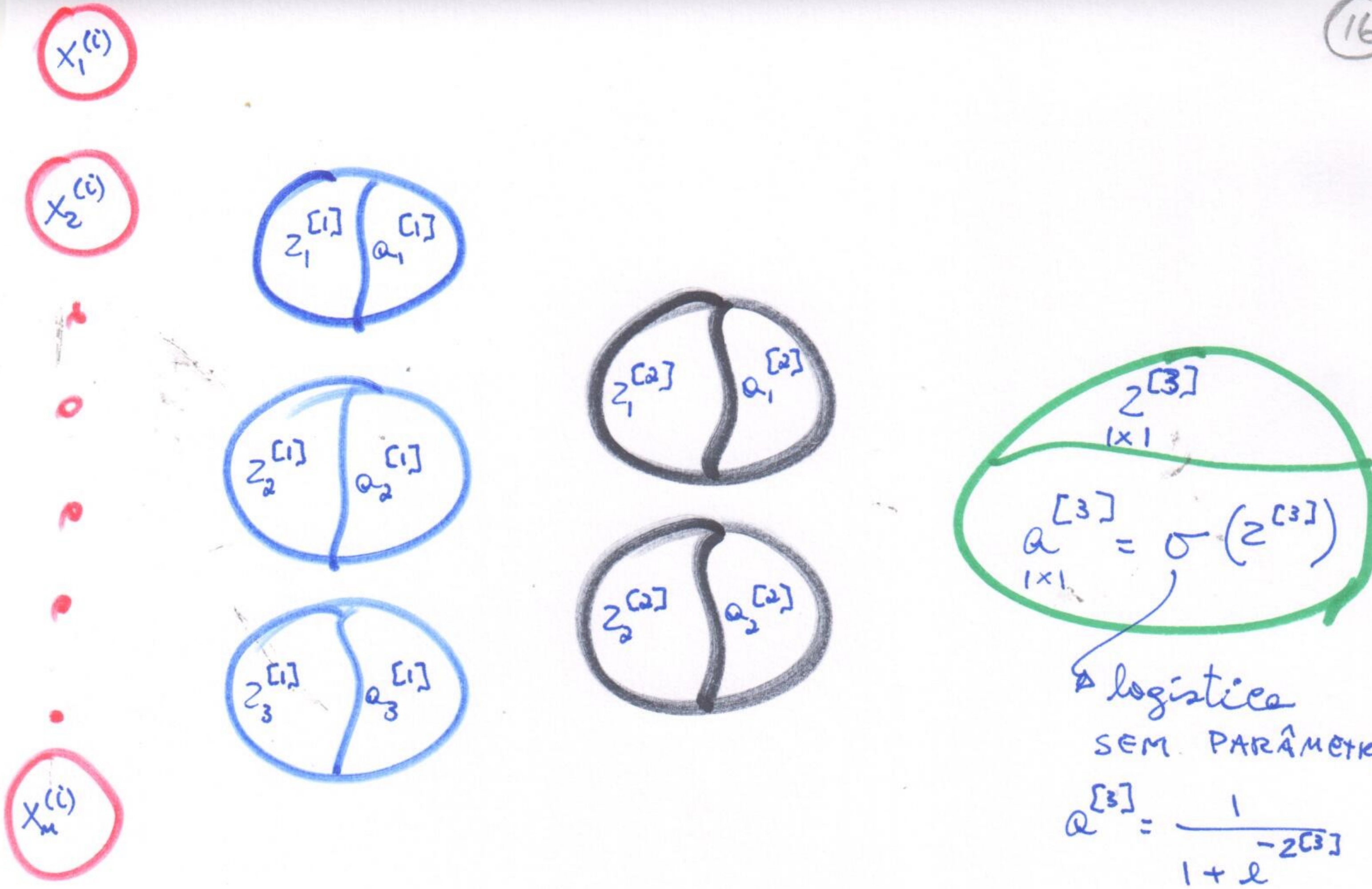


$$z^{[2]} = \begin{pmatrix} z_1^{[2]} \\ z_2^{[2]} \end{pmatrix} = \underset{2\times 3}{W^{[2]}} \cdot \underset{3\times 1}{a^{[1]}} + b^{[2]}$$

$$\underset{\sim}{a}^{[2]} = \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \end{bmatrix} = \begin{bmatrix} g(z_1^{[2]}) \\ g(z_2^{[2]}) \end{bmatrix}$$



$x_1^{(i)}$
$x_2^{(i)}$
$x_m^{(i)}$
$z_1^{[1]}$ $a_1^{[1]}$
$z_2^{[1]}$ $a_2^{[1]}$
$z_3^{[1]}$ $a_3^{[1]}$
$z_1^{[2]}$ $a_1^{[2]}$
$z_2^{[2]}$ $a_2^{[2]}$
$\underset{1\times 2}{W^{[3]}} = [\,— \; w^{[3]} \; —\,]$
$\underset{1\times 1}{z^{[3]}} = \underset{1\times 2}{W^{[3]}} \cdot \underset{2\times 1}{\boxed{\underset{\sim}{a}^{[2]}}} + b^{[3]}$



logística
SEM PARÂMETROS

$$a^{[3]} = \frac{1}{1 + e^{-z^{[3]}}}$$

$x_1^{(i)}$, $x_2^{(i)}$, $\vdots$, $x_m^{(i)}$

$$\underset{3\times m}{W^{[1]}} = \begin{bmatrix} — w_1^{[1]} — \\ — w_2^{[1]} — \\ — w_3^{[1]} — \end{bmatrix} \quad b^{[1]} = \begin{pmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{pmatrix}$$

$$\underset{2\times 3}{W^{[2]}} = \begin{bmatrix} — w_1^{[2]} — \\ — w_2^{[2]} — \end{bmatrix} \quad b^{[2]} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \end{bmatrix}$$

$$\underset{1\times 2}{W^{[3]}} = \begin{bmatrix} — w^{[3]} — \end{bmatrix} \quad \underset{1\times 1}{b^{[3]}}$$

$$z_1^{[1]} = [— w_1^{[1]} —] \boxed{\underset{\sim}{x}^{(i)}} + b_1^{[1]}$$
$$a_1^{[1]} = g(z_1^{[1]})$$

$$z_2^{[1]} = [— w_2^{[1]} —] \boxed{x^{(i)}} + b_2^{[1]}$$
$$a_2^{[1]} = g(z_2^{[1]})$$

$$z_3^{[1]} = [— w_3^{[1]} —] \cdot \boxed{x^{(i)}} + b_3^{[1]}$$
$$a_3^{[1]} = g(z_3^{[1]})$$

$$z_1^{[2]} = [— w_1^{[2]} —] \cdot \boxed{a^{[1]}} + b_1^{[2]}$$
$$a_1^{[2]} = g(z_2^{[1]})$$

$$z_2^{[2]} = [— w_2^{[2]} —] \cdot \boxed{a^{[1]}} + b_2^{[2]}$$
$$a_2^{[2]} = g(z_2^{[2]})$$

$$\underset{1\times 1}{z^{[3]}} = \underset{1\times 2}{W^{[3]}} \underset{2\times 1}{a^{[2]}} + b^{[3]}$$
$$\underset{1\times 1}{a^{[3]}} = \sigma(z^{[3]})$$

$$\underset{3\times 1}{z^{[1]}} = W^{[1]} x^{(i)} + b^{[1]}$$
$$\underset{3\times 1}{a^{[1]}} = g(\underset{\sim}{z}^{[1]})$$

$$\underset{\sim}{z}^{[2]} = \underset{2\times 3}{W^{[2]}} \underset{3\times 1}{a^{[1]}} + \underset{2\times 1}{b^{[2]}}$$

$$\underset{\sim}{a}^{[2]} = g(\underset{\sim}{z}^{[2]}) = \begin{pmatrix} g(z_1^{[2]}) \\ g(z_2^{[2]}) \end{pmatrix}$$

- Considere o item i
- $y^{(i)}$ = dado, não depende de parâmetros, é o número observado, não é uma função matemática
- $\hat{y}^{(i)} = \mathbb{P}(Y^{(i)} = 1 \mid \underset{\sim}{x}^{(i)}) \rightarrow$ depende dos pesos

$$\hat{y}^{(i)} = \hat{y}^{(i)}\Big(\underset{1\times 2}{W^{[3]}}, \underset{1\times 1}{b^{[3]}}, \underset{2\times 3}{W^{[2]}}, \underset{2\times 1}{b^{[2]}}, \underset{3\times n}{W^{[1]}}, \underset{3\times 1}{b^{[1]}}\Big)$$

= função matemática dos pesos e <u>biases</u>

- $$\Big(\underbrace{W^{[3]}, b^{[3]}, W^{[2]}, b^{[2]}, W^{[1]}, b^{[1]}}_{\text{PARÂMETROS "LIVRES"}}\Big) = \underset{\sim}{\theta}$$

- NOTE QUE $a^{[3]}, \underset{\sim}{a}^{[2]}, \underset{\sim}{a}^{[1]}, z^{[3]}, \underset{\sim}{z}^{[2]}, \underset{\sim}{z}^{[1]}$ SÃO <u>FUNÇÕES DETERMINÍSTICAS</u> (MATEMÁTICAS) DOS <u>PARÂMETROS LIVRES</u> E ~~DOS DADOS~~ D~~OS~~AS FEATURES. <u>**NÃO**</u> PRECISAMOS COLOCAR $\hat{y}^{(i)}$ COMO FUNÇÃO DE $\underset{\sim}{a}^{[3]}$, POR EXEMPLO

⊕ LOSS FUNCTION ≡ − log-likelihood

$$\oplus\ \mathcal{L} = \sum_{i=1}^{m} \mathcal{L}\left(y^{(i)}, \hat{y}^{(i)}\right) = \sum_{i=1}^{m} \mathcal{L}^{(i)}$$

$$= \sum_{i=1}^{m} -\left[\left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) + y^{(i)} \log\left(\hat{y}^{(i)}\right) \right]$$

$$= \sum_{i=1}^{m} - \begin{bmatrix} \log(1 - \hat{y}^{(i)}) & \text{se } y^{(i)} = 0 \\ \log(\hat{y}^{(i)}) & \text{se } y^{(i)} = 1 \end{bmatrix}$$

⊕ $\mathcal{L}$ é funcp dos parâmetros livres $\underset{\sim}{\theta} = \left(W^{[3]}, b^{[3]}, \dots, W^{[1]}, b^{[1]}\right)$

⊕ Para achar o mínimo de $\mathcal{L}$, use gradiente descendente:

$$\underset{\sim}{\theta}^{(t+1)} = \underset{\sim}{\theta}^{(t)} - \alpha \cdot \left(\nabla \mathcal{L}\right) = \underset{\sim}{\theta}^{(t)} - \alpha \begin{pmatrix} \partial\mathcal{L} / \partial W^{[3]} \\ \vdots \end{pmatrix}$$

• QUEREMOS $\dfrac{\partial \mathcal{L}}{\partial \theta_i}$ ONDE θ_i = i-ÉSIMO ELEMENTO DE

$$\underset{\sim}{\theta} = \left(\underset{1\times 2}{W^{[3]}},\ \underset{1\times 1}{b^{[3]}},\ \underset{2\times 3}{W^{[2]}},\ \underset{2\times 1}{b^{[2]}},\ \underset{3\times m}{W^{[1]}},\ \underset{3\times 1}{b^{[1]}} \right)$$

$$\underset{\sim}{\theta} = \left(\underset{1\times 2}{\left[- w^{[3]T} - \right]},\ \underset{1\times 1}{b^{[3]}},\ \begin{bmatrix} - w_1^{[2]T} - \\ - w_2^{[2]T} - \end{bmatrix},\ \left[b_1^{[2]},\ b_2^{[2]} \right],\ \underset{3\times m}{\begin{pmatrix} - w_1^{[1]} - \\ - w_2^{[1]} - \\ - w_3^{[1]} - \end{pmatrix}},\ \underset{3\times 1}{\begin{pmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \end{pmatrix}} \right)$$

VAMOS EMPILHAR ESTES ELEMENTOS FORMANDO UM LONGO VETOR (MOSTRADO AQUI COMO UMA LINHA):

$$\underset{\sim}{\theta} = \left(- w^{[3]T} -,\ b^{[3]},\ - w_1^{[2]T} -,\ - w^{[2]T} -,\ b_1^{[2]},\ b_2^{[2]},\ - w_1^{[1]} -,\ - w_2^{[1]} -,\ - w_3^{[1]} -,\ b_1^{[1]},\ b_2^{[1]},\ b_3^{[1]} \right)$$

- PRECISAMOS DA DERIVADA PARCIAL DE $\mathcal{L}$ COM RESPEITO A CADA ELEMENTO EM $\underset{\sim}{\theta} = (W^{[3]}, \dots, b^{[1]})$

- POR EXEMPLO, $\underset{1\times 1}{W^{[3]}}$

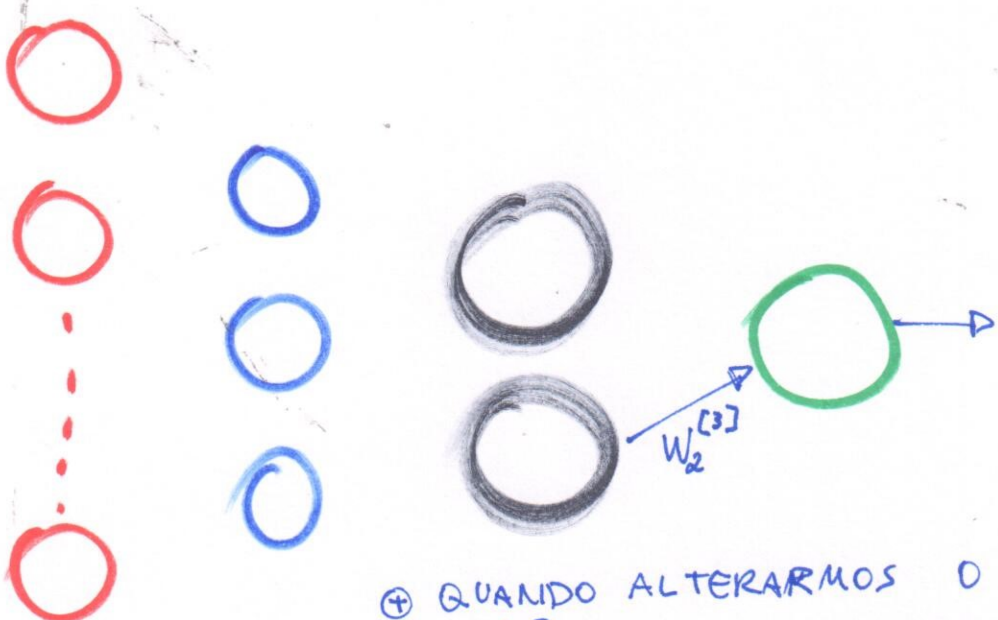
- QUEREMOS $\dfrac{\partial \mathcal{L}}{\partial W^{[3]}} = \dfrac{\partial}{\partial W^{[3]}} \sum_{i=1}^{m} \mathcal{L}(y^{(i)}, \hat{y}^{(i)})$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial W^{[3]}} \mathcal{L}(y^{(i)}, \hat{y}^{(i)}) = \sum_{i=1}^{m} \frac{\partial \mathcal{L}^{(i)}}{\partial W^{[3]}}$$

- VAMOS OBTER A DERIVADA PARCIAL ASSOCIADA COM CADA EXEMPLO E DEPOIS SOMAR TODOS.

- O QUE SIGNIFICA $\dfrac{\partial \mathcal{L}^{(i)}}{\partial \theta_i}$? QUANDO O PESO OU BIAS θ_i FOR LIGEIRAMENTE MODIFICADO, COMO A FUNÇÃO DE PERDA REAGE?

- VAMOS VOLTAR A OLHAR NOSSA REDE
- MAS AGORA SERÁ COMO UMA FUNÇÃO DOS PESOS W's E DOS BIASES b's

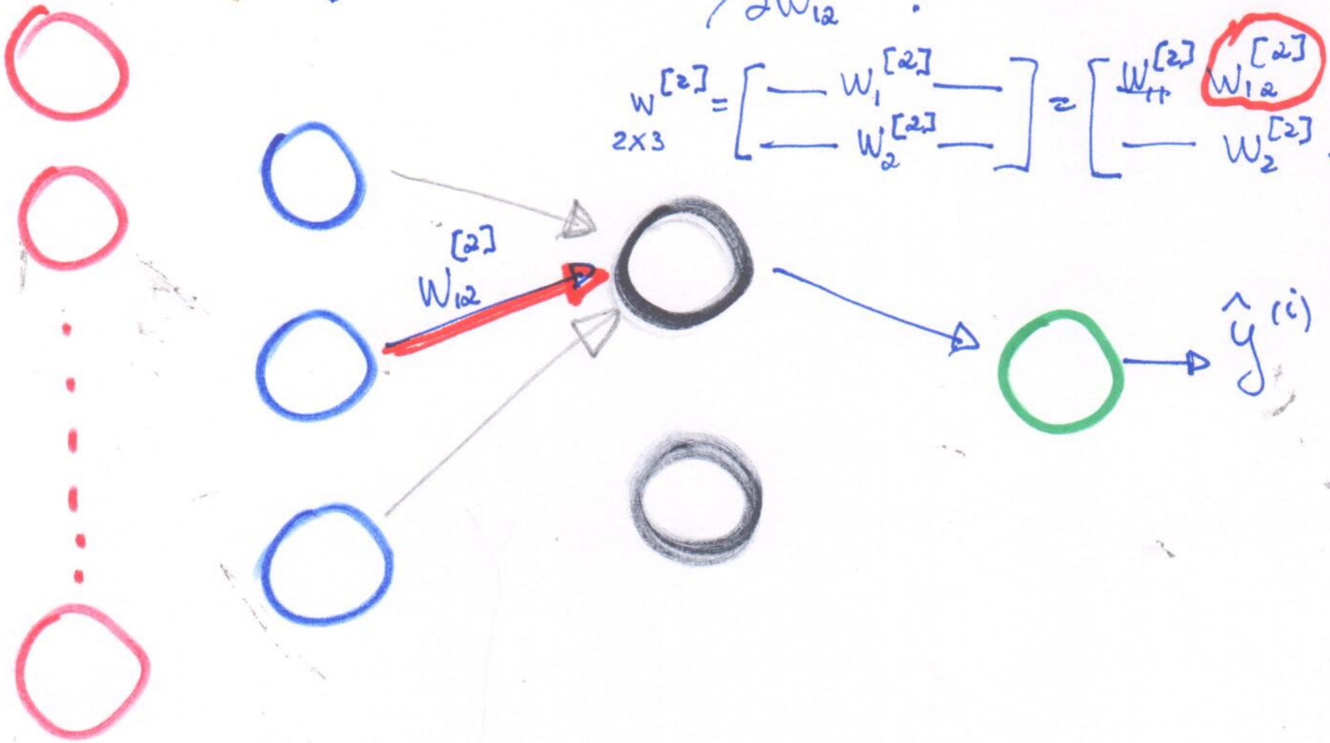


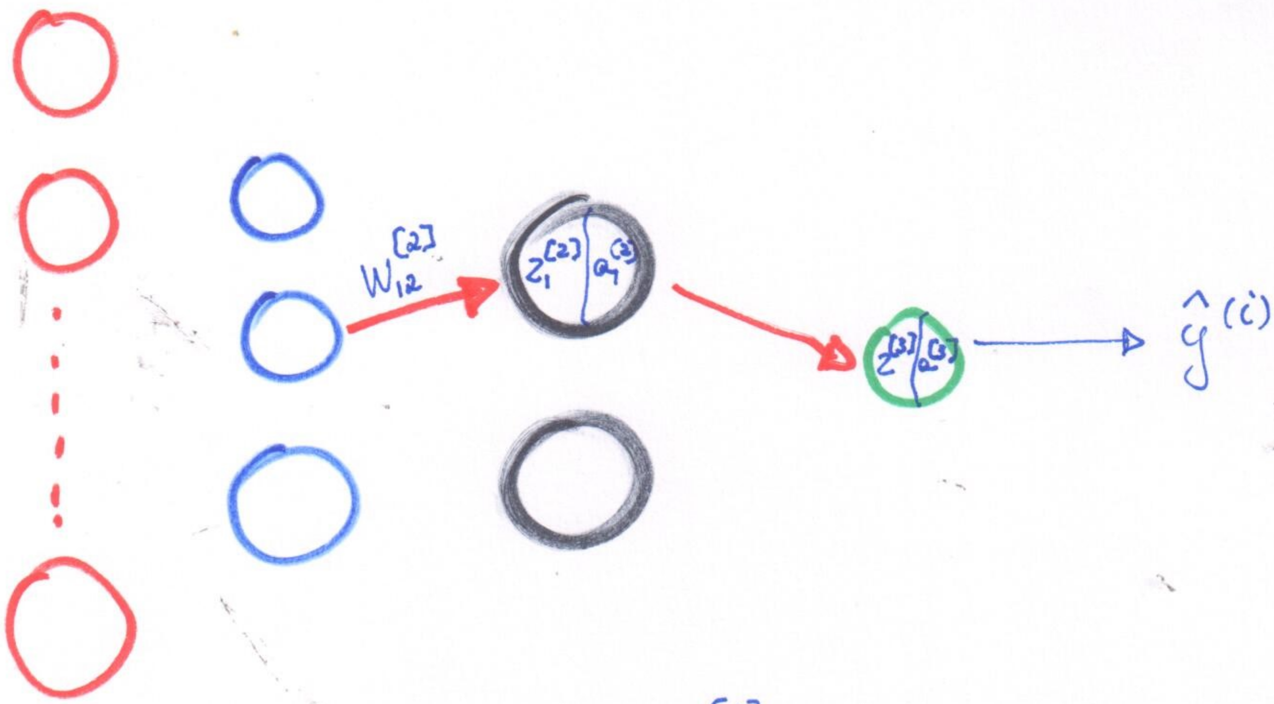
$$\hat{y}^{(i)} = \mathbb{P}\left(Y^{(i)} = 1 \,\middle|\, \underset{\sim}{x}^{(i)}, \text{PESOS } w\text{'s}, \text{BIASES } b\text{'s}\right)$$

- QUANDO ALTERARMOS O PESO $W_2^{[3]}$ PARA $W_2^{[3]} + \Delta$, QUAL O EFEITO EM $\hat{y}^{(i)}$?

- E QUANDO ALTERARMOS $W_{12}^{[2]}$ PARA $W_{12}^{[2]} + \Delta$?
- QUANTO É $\partial \mathcal{L} / \partial W_{12}^{[2]}$?

$$\underset{2\times 3}{W^{[2]}} = \begin{bmatrix} — W_1^{[2]} — \\ — W_2^{[2]} — \end{bmatrix} = \begin{bmatrix} W_{11}^{[2]} & W_{12}^{[2]} & W_{13}^{[2]} \\ — & W_2^{[2]} & — \end{bmatrix}$$





⊕ QUANDO O PESO $W_{12}^{[2]}$ FOR ALTERADO, O EFEITO EM $\hat{y}^{(i)}$ SERÁ SENTIDO ~~[illegible]~~ NO FINAL DO CAMINHO ACIMA:

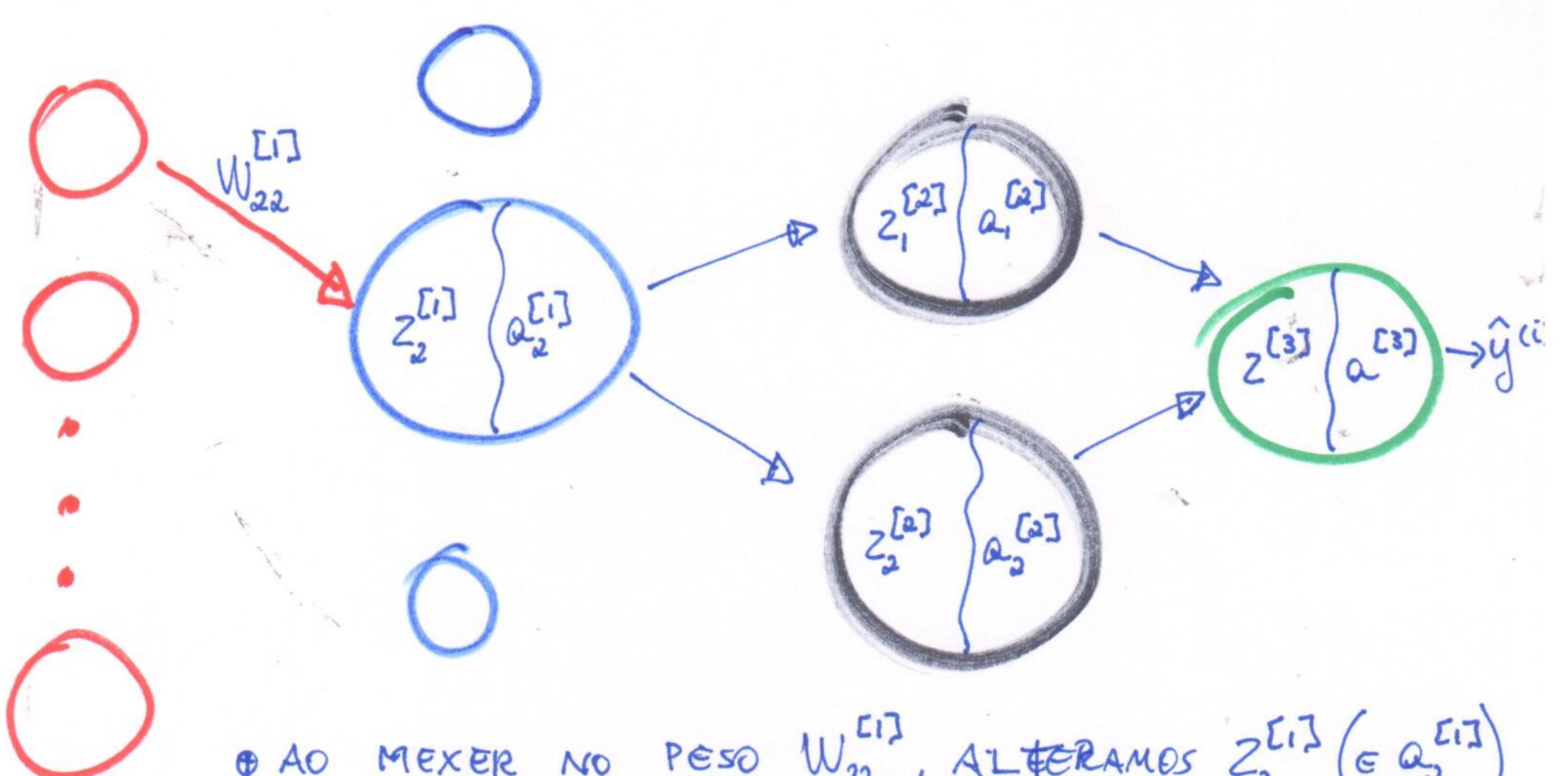
- PRIMEIRO ALTERAMOS $Z_1^{[2]}$ E DEPOIS $a_1^{[2]}$
- EM SEGUIDA, O $a_1^{[2]}$ ALTERADO AFETA $a^{[3]} = \hat{y}^{(i)}$

24

- ASSIM, PARA MEDIRMOS O IMPACTO EM $\hat{y}^{(i)}$ DE ALTERAR $W_{12}^{[2]}$ CALCULAMOS O SEU EFEITO AO LONGO DESTE CAMINHO DE DOIS PASSOS
- É O QUE A REGRA DA CADEIA VAI FAZER AUTOMATICAMENTE, COMO VEREMOS DAQUI A POUCO.
- O CASO MAIS INTERESSANTE É QUANDO ALTERAMOS UM PESO USADO NA <u>1ª</u> CAMADA:

$$W^{[1]} = \begin{bmatrix} \text{———} \; W_1^{[1]} \; \text{———} \\ W_{21}^{[1]} \quad W_{22}^{[1]} \quad \ldots \quad W_{2m}^{[1]} \\ \text{———} \; W_3^{[1]} \; \text{———} \end{bmatrix}$$

• QUEREMOS $\dfrac{\partial \mathcal{L}^{(i)}}{\partial W_{22}^{[1]}}$



• AO MEXER NO PESO $W_{22}^{[1]}$, ALTERAMOS $z_2^{[1]}$ (E $a_2^{[1]}$)

$$\Rightarrow \text{MEXER EM} \begin{cases} z_1^{[2]} \to a_1^{[2]} \\ z_2^{[2]} \to a_2^{[2]} \end{cases} \Rightarrow \text{MEXER EM } z^{[3]} \text{ POR}$$

<u>DOIS</u> CAMINHOS.

- VAMOS MOSTRAR COMO CADA UM DESTES TERMOS (DERIVADAS PARCIAIS) É OBTIDO. CONSULTAR DESENHO DA REDE NEURAL O TEMPO TODO
- VAMOS IGNORAR O ÍNDICE (i) DO ITEM PARA NÃO SOBRECARREGAR A NOTAÇÃO.
- A FUNÇÃO DE PERDA OU CUSTO PARA UM ITEM (i) (IGNORANDO O ÍNDICE) É

$$\mathcal{L} = -\left[(1-y)\log(1-\hat{y}) + y\log(\hat{y})\right]$$

DADOS

FUNÇÃO DOS PESOS E BIASES

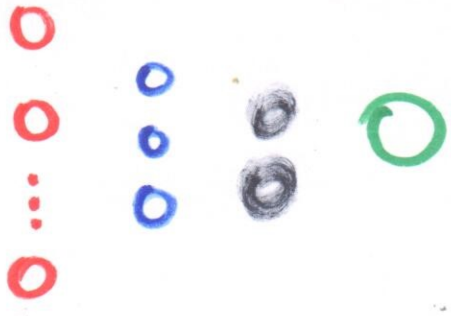
ASSIM, AO TOMARMOS AS DERIVADAS EM RELAÇÃO AOS PESOS E BIASES, $\begin{cases} \text{OS } y\text{'s SÃO CONSTANTES} \\ \text{OS } \hat{y}\text{'s SÃO DERIVÁVEIS} \end{cases}$

Definição e notação de derivada:

$$\mathcal{L} : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

$$\underset{\sim}{W}^{[3]} = \begin{pmatrix} w_1^{[3]} \\ w_2^{[3]} \end{pmatrix} \longrightarrow \mathcal{L}\left(\underset{\sim}{W}^{[3]}\right)$$

$$\underset{1\times 2}{\frac{\partial \mathcal{L}}{\partial \underset{\sim}{W}^{[3]}}} = \left(\frac{\partial \mathcal{L}}{\partial w_1^{[3]}} \ , \ \frac{\partial \mathcal{L}}{\partial w_2^{[3]}} \right)$$



$$\underset{1\times1}{z^{[3]}} = \left[— w^{[3]T} — \right] \cdot \boxed{a^{[2]}} + \underset{1\times1}{b^{[3]}}$$

$$= \left[w_1^{[3]} \;\; w_2^{[3]} \right] \cdot \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \end{bmatrix} + b^{[3]}$$

$$= w_1^{[3]} a_1^{[2]} + w_2^{[3]} a_2^{[2]} + b^{[3]}$$

$$a^{[3]} = \sigma(z^{[3]}) = \mathbb{P}(Y=1) = \frac{1}{1+e^{-z^{[3]}}} = \hat{y}$$

- VAMOS PRECISAR DO SEGUINTE:

$$\otimes \; \underset{1\times2}{\frac{\partial z^{[3]}}{\partial w^{[3]}}} = \left(\frac{\partial z^{[3]}}{\partial w_1^{[3]}} , \frac{\partial z^{[3]}}{\partial w_2^{[3]}} \right) = \left(a_1^{[2]} , a_2^{[2]} \right) = \underset{1\times2}{\left(a^{[2]}\right)^T}$$

$$\otimes \; \frac{\partial a^{[3]}}{\partial z^{[3]}} = \frac{\partial \sigma(z^{[3]})}{\partial z^{[3]}} = \sigma(z^{[3]})\left(1-\sigma(z^{[3]})\right) = \underset{1\times1}{\hat{y}(1-\hat{y})}$$

$\oplus$ PODEMOS AGORA CALCULAR $\underset{1\times2}{\dfrac{\partial \mathcal{L}}{\partial w^{[3]}}} \longrightarrow$

* $$\frac{\partial \mathcal{L}}{\partial w^{[3]}} = \frac{\partial}{\partial w^{[3]}}\left[-(1-y)\log(1-\hat{y}) + y\log(\hat{y})\right]$$

$$= -\left[(1-y)\frac{\partial}{\partial w^{[3]}}\left(\log(1-\hat{y})\right) + y\,\frac{\partial}{\partial w^{[3]}}\left(\log(\hat{y})\right)\right]$$

$$= -\left[\cancel{-}(1-y)\frac{1}{1-\hat{y}}\left(\frac{\partial \hat{y}}{\partial w^{[3]}}\right) + y\,\frac{1}{\hat{y}}\,\frac{\partial \hat{y}}{\partial w^{[3]}}\right]$$

$$= \left[\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right]\cdot\frac{\partial \hat{y}}{\partial w^{[3]}} = (*)$$

* $$\frac{\partial \hat{y}}{\partial w^{[3]}} = \frac{\partial a^{[3]}}{\partial w^{[3]}} = \frac{\partial \sigma(z^{[3]})}{\partial w^{[3]}} = \underbrace{\frac{\partial \sigma(z^{[3]})}{\partial z^{[3]}}}\ \underbrace{\frac{\partial z^{[3]}}{\partial w^{[3]}}} =$$

SLIDE ANTERIOR

$$= \underbrace{\hat{y}(1-\hat{y})}_{1\times 1}\cdot\underbrace{\left[a^{[2]T}\right]}_{1\times 2}$$

* SUBSTITUINDO EM (*) TEMOS:

$$\frac{\partial \mathcal{L}}{\partial w^{[3]}} = \underbrace{\left[\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right]\left(\hat{y}(1-\hat{y})\right)}\cdot\left[\underset{1\times 2}{a^{[2]T}}\right]$$

$$= \left[(1-y)\hat{y} - y(1-\hat{y})\right]\cdot\left[\underset{1\times 2}{a^{[2]T}}\right]$$

$$= (\hat{y}-y)\,\underbrace{a^{[2]T}}_{1\times 2}$$

$$= (\hat{y}-y)\cdot\left[a_1^{[2]}\ ,\ a_2^{[2]}\right]$$

PARA ENCONTRAR $\frac{\partial \mathcal{L}}{\partial b^{[3]}}$, OBSERVO QUE OS PRIMEI-
ROS PASSOS SÃO IDÊNTICOS:

(31)

$$\frac{\partial \mathcal{L}}{\partial b^{[3]}} = \frac{\partial}{\partial b^{[3]}}\left[-(1-y)\log(1-\hat{y}) + y\log(\hat{y})\right]$$

$$= -\left[(1-y)\frac{\partial}{\partial b^{[3]}}\left(\log(1-\hat{y})\right) + y\frac{\partial}{\partial b^{[3]}}\left(\log(\hat{y})\right)\right]$$

$$= -\left[-(1-y)\frac{1}{1-\hat{y}}\frac{\partial \hat{y}}{\partial b^{[3]}} + \frac{y}{\,}\frac{1}{\hat{y}}\frac{\partial \hat{y}}{\partial b^{[3]}}\right]$$

$$= \underbrace{\left[\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right]}_{\text{IGUAL AO CASO ANTERIOR}} \cdot \frac{\partial \hat{y}}{\partial b^{[3]}} = (**)$$

- PRECISAMOS ENCONTRAR $\frac{\partial \hat{y}}{\partial b^{[3]}}$

- LEMBRANDO QUE $\hat{y} = a^{[3]} = \mathbb{P}(Y=1 \mid \underline{x}) = \frac{1}{1+e^{-z^{[3]}}}$

  COM $z^{[3]} = \underbrace{w^{[3]T}}_{1\times 2} \cdot \underbrace{a^{[2]}}_{2\times 1} + \underbrace{b^{[3]}}_{1\times 1}$

- PORTANTO,

$$\frac{\partial \hat{y}}{\partial b^{[3]}} = \frac{\partial \hat{y}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial b^{[3]}}$$

$$= \hat{y}(1-\hat{y}) \cdot 1$$

- ASSIM, SUBSTITUINDO EM (**) E MANIPULANDO COMO ANTES, TEMOS $\frac{\partial \mathcal{L}}{\partial b^{[3]}} = (\hat{y}-y)$

* OBTIVEMOS AS PRIMEIRAS DERIVADAS PARCIAIS DOS PARÂMETROS DA REDE NEURAL:

$$\underset{\sim}{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \hline \theta_3 \\ \hline \theta_4 \\ \theta_5 \\ \theta_6 \\ \hline \vdots \end{bmatrix} = \begin{bmatrix} W^{[3]} \\ \hline b^{[3]} \\ \hline W_1^{[2]} \\ \hline \vdots \end{bmatrix} \Longrightarrow \begin{bmatrix} \dfrac{\partial \mathcal{L}}{\partial W^{[3]}} \\ \hline \dfrac{\partial \mathcal{L}}{\partial b^{[3]}} \\ \hline \vdots \end{bmatrix} = \begin{bmatrix} (\hat{y} - y) \cdot \underset{\sim}{a}^{[2]}_{2\times 1} \\ \hline (\hat{y} - y) \\ \hline \end{bmatrix}$$

⊕ PRÓXIMO PASSO: COMO A VARIAÇÃO DA MATRIZ DO PESOS $\underset{2\times 3}{W^{[2]}} = \begin{bmatrix} — w_1^{[2]} — \\ — w_2^{[2]} — \end{bmatrix}$ IMPACTA FUNCIONALMENTE A PREDIÇÃO $\hat{y}$ (E PORTANTO, A PERDA $\mathcal{L}$)

⊕ VAMOS OLHAR POR PARTES.

⊕ COMO A <u>1ª LINHA</u> DE $W^{[2]}$ IMPACTA $\hat{y}$?

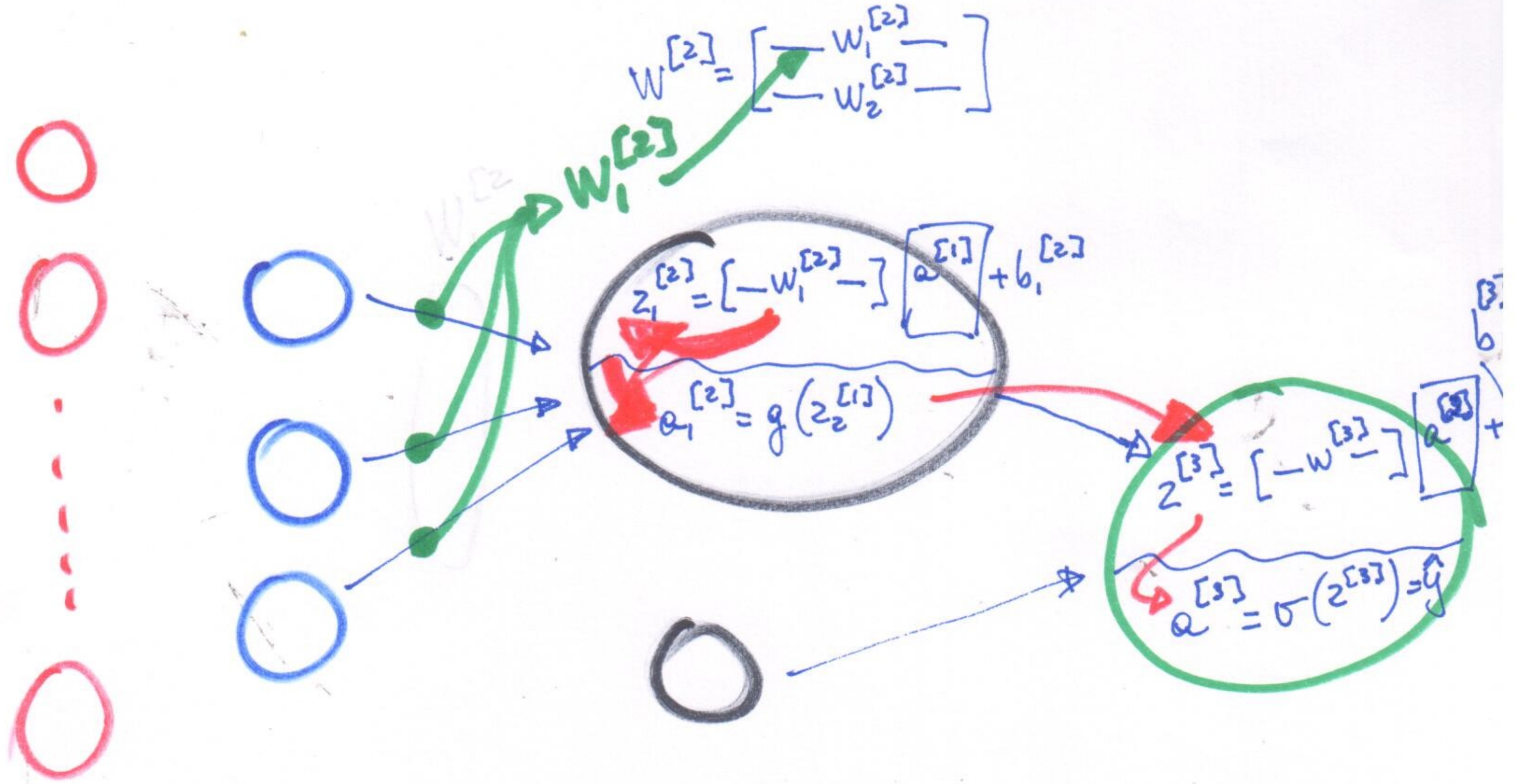
⊕ 1ª LINHA É $\underset{1\times 3}{\begin{bmatrix} — w_1^{[2]} — \end{bmatrix}}$

⊕ VEJA NA REDE A CADEIA DO IMPACTOS SUCESSIVOS AO ALTERAR $w_1^{[2]}$ ATÉ ATINGIR A PREDIÇÃO $\hat{y}$

$$— w_1^{[2]} — \longrightarrow z_1^{[2]} = \underset{1\times 3}{\begin{bmatrix} — w_1^{[2]} — \end{bmatrix}} \cdot \underset{3\times 1}{\boxed{a^{[1]}}} + b_1^{[2]} \longrightarrow a_1^{[2]} = g(z_1^{[2]})$$

$$a^{[3]} = \sigma(z^{[3]}) = \hat{y} \longleftarrow z^{[3]} = \begin{bmatrix} — w^{[3]} — \end{bmatrix} \boxed{a^{[2]}} + b^{[3]}$$

TODO O VETOR



$W^{[2]} = \begin{bmatrix} - w_1^{[2]} - \\ - w_2^{[2]} - \end{bmatrix}$
$w_1^{[2]}$
$z_1^{[2]} = [- w_1^{[2]} -] [a^{[1]}] + b_1^{[2]}$
$a_1^{[2]} = g(z_2^{[1]})$
$b^{[3]}$
$z^{[3]} = [- w^{[3]} -] [a^{[2]}] +$
$a^{[3]} = \sigma(z^{[3]}) = \hat{y}$

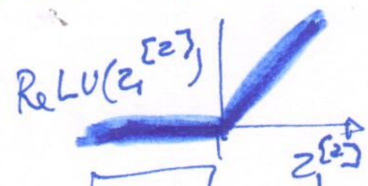
* USAMOS A REGRA DA CADEIA, NO SENTIDO INVERSO (BACK PROPAGATION) DO CAMINHO MOSTRADO:

$$\underset{3\times 1}{\frac{\partial \mathcal{L}}{\partial w_1^{[2]}}} = \underbrace{\frac{\partial \mathcal{L}}{\partial z^{[3]}}}_{\substack{\hat{y}-y \\ \text{JÁ OBTIDO}}} \cdot \underbrace{\frac{\partial z^{[3]}}{\partial a_1^{[2]}}} \cdot \frac{\partial a_1^{[2]}}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial w_1^{[2]}}$$

* COMO $z^{[3]} = \left[\text{—} \underset{1\times 2}{w^{[3]T}} \text{—} \right] \left[\underset{2\times 1}{a^{[2]}} \right] + b^{[3]} = w_1^{[3]} a_1^{[2]} + w_2^{[3]} a_2^{[2]} + b^{[3]}$

ENTÃO $\frac{\partial z^{[3]}}{\partial a_1^{[2]}} = w_1^{[3]}$ = 1º ELEMENTO DO VETOR-LINHA $W^{[3]T}$

* $\frac{\partial a_1^{[2]}}{\partial z_1^{[2]}} = \frac{\partial \, ReLU(z_1^{[2]})}{\partial z_1^{[2]}} = \begin{cases} 0 & \text{se } z_1^{[2]} < 0 \\ 1 & \text{se } z_1^{[2]} \geq 0 \end{cases}$



* COMO $z_1^{[2]} = \left[\text{—} \underset{1\times 3}{w_1^{[2]}} \text{—} \right] \cdot \boxed{\underset{3\times 1}{a^{[1]}}} + b_1^{[2]}$ temos $\frac{\partial z_1^{[2]}}{\partial w_1^{[2]}} = \boxed{\underset{3\times 1}{\underset{\sim}{a}^{[1]}}}$

* JUNTANDO TODOS ESTES PEDAÇOS, TEMOS A DERIVADA PARCIAL DESEJADA:

$$\underbrace{\frac{\partial \mathcal{L}}{\partial w_1^{[2]}}}_{3\times 1} = \underbrace{\frac{\partial \mathcal{L}}{\partial z^{[3]}}} \cdot \underbrace{\frac{\partial z^{[3]}}{\partial a_1^{[2]}}} \cdot \underbrace{\frac{\partial a_1^{[2]}}{\partial z_1^{[2]}}} \cdot \underbrace{\frac{\partial z_1^{[2]}}{\partial w_1^{[2]}}}$$

$$= \underbrace{(\hat{y} - y)}_{1\times 1} \cdot \underbrace{W_1^{[3]}}_{1\times 1} \cdot \underbrace{\begin{pmatrix} 0 \text{ SE } z_1^{[2]} < 0 \\ 1 \text{ SE } z_1^{[2]} \geq 0 \end{pmatrix}}_{1\times 1} \cdot \underbrace{\underset{\sim}{a}^{[1]}}_{3\times 1}$$

DE MODO ANÁLOGO, OBTEMOS

$$\underbrace{\frac{\partial \mathcal{L}}{\partial w_2^{[2]}}}_{3\times 1} = \underbrace{\frac{\partial \mathcal{L}}{\partial z^{[3]}}}_{\hat{y} - y} \cdot \underbrace{\frac{\partial z^{[3]}}{\partial a_2^{[2]}}}_{W_2^{[3]}} \cdot \underbrace{\frac{\partial a_2^{[2]}}{\partial z_2^{[2]}}}_{\begin{pmatrix} 0 \text{ SE } z_2^{[2]} \leq 0 \\ 1 \text{ SE } z_2^{[2]} \geq 0 \end{pmatrix}} \cdot \underbrace{\frac{\partial z_2^{[2]}}{\partial w_2^{[2]}}}_{\underset{\sim}{a}^{[1]}}$$

• ASSIM CONSEGUIMOS ENCONTRAR MAIS ELEMENTOS DO VETOR GRADIENTE DO $\mathcal{L}(\underset{\sim}{\theta})$

38

$$
\underset{\sim}{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \\ \theta_8 \\ \theta_9 \\ \vdots \end{bmatrix} = \begin{bmatrix} w^{[3]} \\ b^{[3]} \\ w_1^{[2]} \\ w_2^{[2]} \\ \vdots \end{bmatrix} \longrightarrow \begin{bmatrix} \partial \mathcal{L} / \partial w^{[3]} \\ \partial \mathcal{L} / \partial b^{[3]} \\ \partial \mathcal{L} / \partial w_1^{[2]} \\ \partial \mathcal{L} / \partial w_2^{[2]} \\ \vdots \end{bmatrix} = (\hat{y} - y) \begin{bmatrix} \underset{\sim}{a}^{[2]} \\ 1 \\ w_1^{[3]} \begin{pmatrix} 0 \text{ se } z_1^{[2]} < 0 \\ 1 \text{ se } z_1^{[2]} \geq 0 \end{pmatrix} \underset{\sim}{a}^{[1]} \\ w_2^{[3]} \begin{pmatrix} 0 \text{ se } z_2^{[2]} < 0 \\ 1 \text{ se } z_2^{[2]} \geq 0 \end{pmatrix} \cdot \underset{\sim}{a}^{[1]} \\ \vdots \end{bmatrix}
$$

⊕ GRADIENTE DESCENDENTE PARA ESTES PARÂMETROS FICA

$$w_1^{[2]} \leftarrow w_1^{[2]} + \alpha \frac{\partial L}{\partial w_1^{[2]}} = w_1^{[2]} + \alpha \left((\hat{y}-y) \underbrace{g'(z_1^{[2]})}_{0 \text{ ou } 1} \cdot w_1^{[3]} \cdot \underbrace{\underset{\sim}{a}_1^{[1]}}_{3\times 1} \right)$$

$$w_2^{[2]} \leftarrow w_2^{[2]} + \alpha \frac{\partial L}{\partial w_2^{[2]}} = w_2^{[2]} + \alpha \left((\hat{y}-y) \underbrace{g'(z_2^{[2]})}_{0 \text{ ou } 1} \cdot w_2^{[3]} \cdot \underbrace{\underset{\sim}{a}_1^{[1]}}_{3\times 1} \right)$$

⊕ PODEMOS ATUALIZAR TODA A MATRIZ $\underset{2\times 3}{W^{[2]}}$ AO INVÉS DE ATUALIZARMOS UMA LINHA DE CADA VEZ:

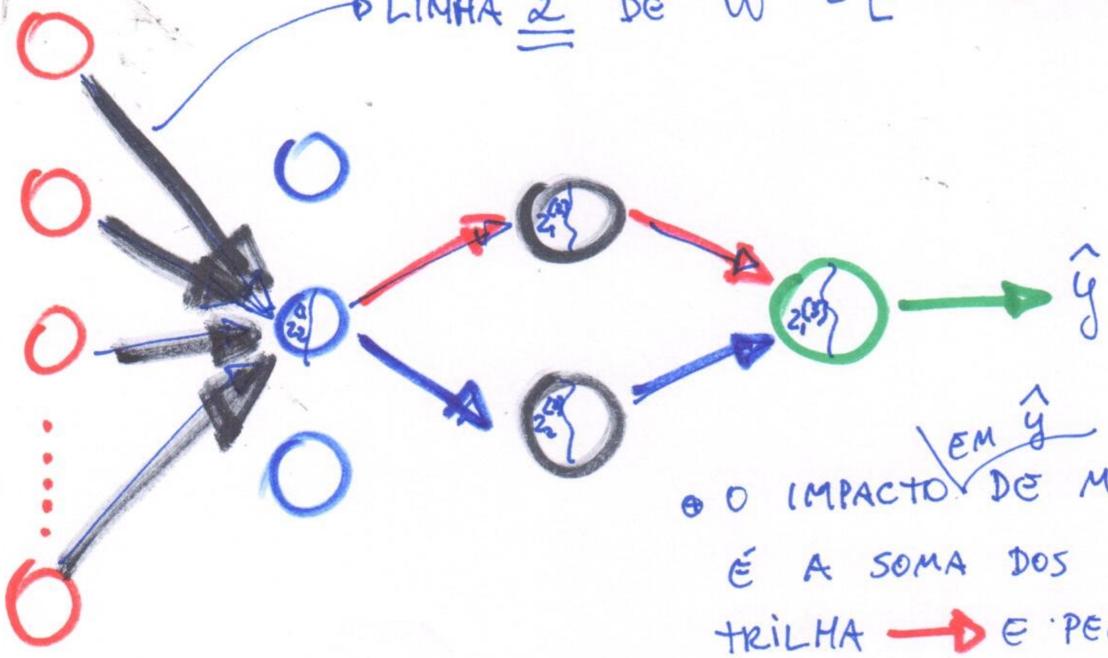
$$\boxed{\underset{2\times 3}{W^{[2]}}} \leftarrow \boxed{W^{[2]}} + \alpha \boxed{\frac{\partial L}{\partial W^{[2]}}} = \boxed{W^{[2]}} + \alpha(\hat{y}-y)\left(\underset{2\times 1}{W^{[3]T}} \odot \underbrace{g'(z^{[2]})}_{2\times 1}\right) \underset{1\times 3}{a^{[1]}}$$

HADAMARD PRODUCT
ELEMENTWISE PROD.

⊕ A ATUALIZAÇÃO DE $W^{[1]}$ É UM POUCO MAIS COMPLICADA POIS ELA ESTÁ MAIS PROFUNDA NA REDE.

⊕ AO FAZER UMA PEQUENA PERTURBAÇÃO NAS LINHAS DA MATRIZ $\underset{3\times m}{W^{[1]}} = \begin{bmatrix} — w_1^{[1]} — \\ — w_2^{[1]} — \\ — w_3^{[1]} — \end{bmatrix}$ TEREMOS MAIS DE UM CAMINHO PARA PERTURBAR $\hat{y}$ (e $\mathcal{L}$)

LINHA $\underline{\underline{2}}$ DE $W^{[1]} = [— w_2^{[1]} —]$



⊕ O IMPACTO EM $\hat{y}$ DE MEXER EM $[— w_2^{[1]} —]$ É A SOMA DOS IMPACTOS PELA TRILHA → E PELA TRILHA →

$$
\begin{tikzcd}
 & & & z_1^{[2]} \arrow[r] & a_1^{[2]} \arrow[dr] & & \\
W_1^{[1]} \arrow[r] & z_1^{[1]} \arrow[r] & a_1^{[1]} \arrow[ur] \arrow[dr] & & & z^{[3]} \arrow[r] & a^{[3]} = \hat{y} \\
 & & & z_2^{[2]} \arrow[r] & a_2^{[2]} \arrow[ur] & &
\end{tikzcd}
$$

1ª LINHA
(1º NEURÔNIO)

$$
\begin{tikzcd}
 & & & z_1^{[2]} \arrow[r] & a_1^{[2]} \arrow[dr] & & \\
W_2^{[1]} \arrow[r] & z_2^{[1]} \arrow[r] & a_2^{[1]} \arrow[ur] \arrow[dr] & & & z^{[3]} \arrow[r] & a^{[3]} = \hat{y} \\
 & & & z_2^{[2]} \arrow[r] & a_2^{[2]} \arrow[ur] & &
\end{tikzcd}
$$

$$
\begin{tikzcd}
 & & & z_1^{[2]} \arrow[r] & a_1^{[2]} \arrow[dr] & & \\
W_3^{[1]} \arrow[r] & z_3^{[1]} \arrow[r] & a_3^{[1]} \arrow[ur] \arrow[dr] & & & z^{[3]} \arrow[r] & a^{[3]} = \hat{y} \\
 & & & z_2^{[2]} \arrow[r] & a_2^{[2]} \arrow[ur] & &
\end{tikzcd}
$$

CONSIDERANDO A 1ª LINHA $W_1^{[1]}$ DE $W^{[1]}$ TEMOS:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W_1^{[1]}} &= \left[\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right] \frac{\partial \hat{y}}{\partial W_1^{[1]}} \\
&= \underbrace{\left[\frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}}\right] \frac{\partial \hat{y}}{\partial z^{[3]}}}_{(\hat{y}-y)} \left(\frac{\partial z^{[3]}}{\partial a_1^{[2]}} \frac{\partial a_1^{[2]}}{\partial z_1^{[2]}} \frac{\partial z_1^{[2]}}{\partial a_1^{[1]}} \underbrace{\frac{\partial a_1^{[1]}}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial W_1^{[1]}}}_{='s} + \frac{\partial z^{[3]}}{\partial a_2^{[2]}} \frac{\partial a_2^{[2]}}{\partial z_2^{[2]}} \frac{\partial z_2^{[2]}}{\partial a_1^{[1]}} \frac{\partial a_1^{[1]}}{\partial z_1^{[1]}} \cdot \frac{\partial z_1^{[1]}}{\partial W_1^{[1]}} \right) \\
&= (\hat{y}-y)\left(\underbrace{W_1^{[3]}}_{1\times 1} \underbrace{g'(z_1^{[2]})}_{1\times 1} \cdot \underbrace{W_{1,1}^{[2]}}_{1\times 1} + \underbrace{W_2^{[3]}}_{1\times 1} \underbrace{g'(z_2^{[2]})}_{1\times 1} \underbrace{W_{2,1}^{[2]}}_{1\times 1} \right) g'(z_1^{[1]}) \cdot \underbrace{x^{(i)}}_{M\times 1}
\end{aligned}
$$

ANÁLOGO PARA $\frac{\partial \mathcal{L}}{\partial W_2^{[1]}}$ e $\frac{\partial \mathcal{L}}{\partial W_3^{[1]}}$.

COM ISTO TEMOS O VETOR GRADIENTE $\frac{\partial \mathcal{L}}{\partial \cancel{\theta}} \in$