Forward & BackPropgation Derivation

-Pradeek Narang

5-12-2018

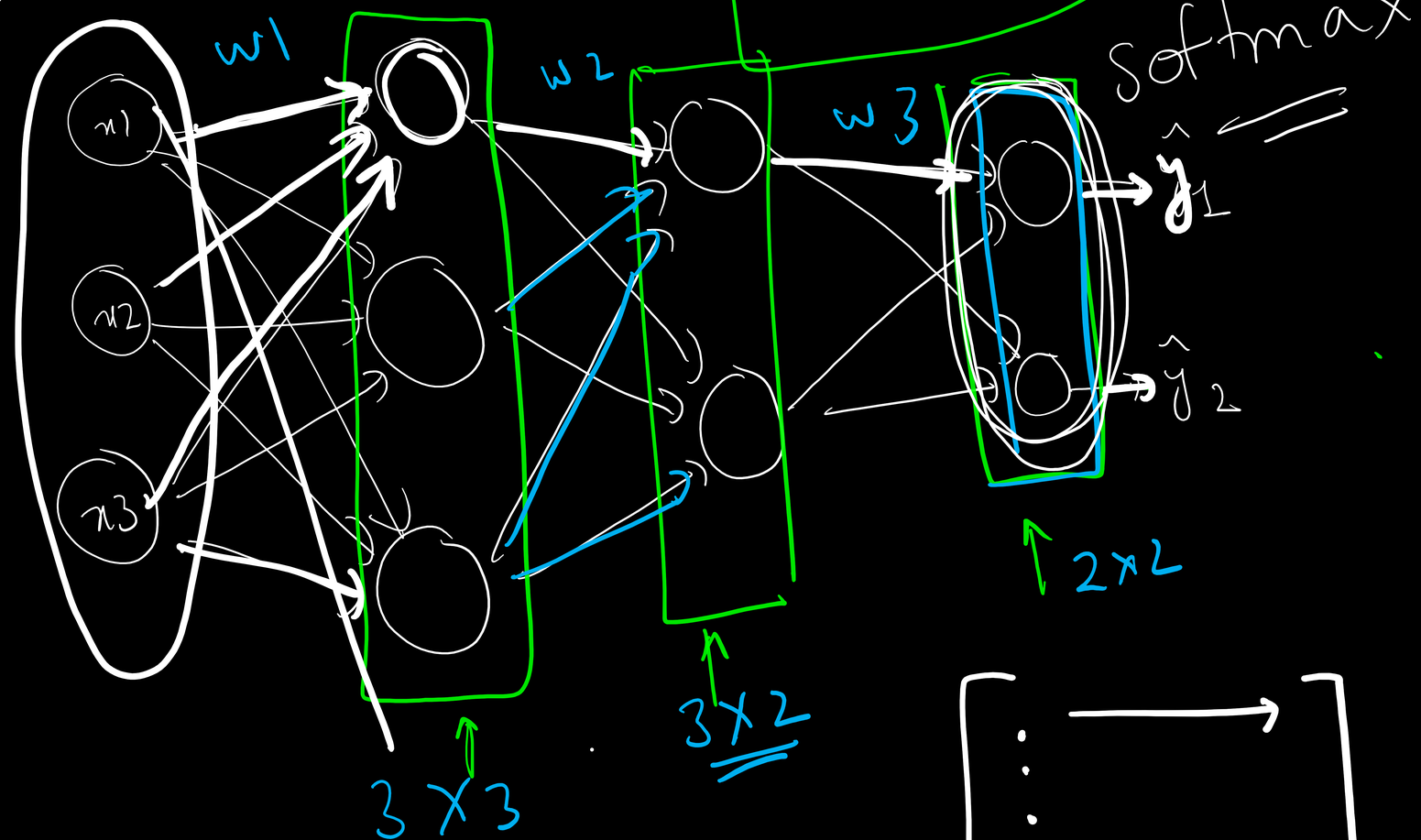
Deep Learning

Data

Code

one example

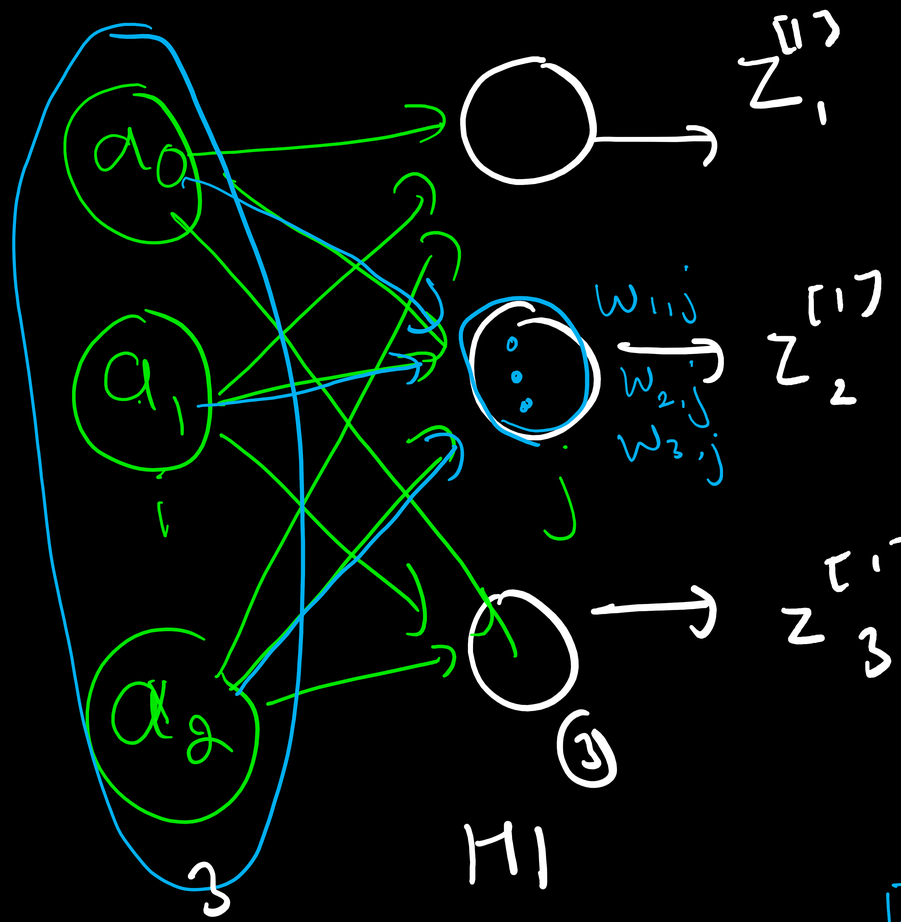
$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{3\times 1}$$



$$\begin{bmatrix} \vdots & \longrightarrow \\ \vdots & \end{bmatrix}_{3\times 4}$$

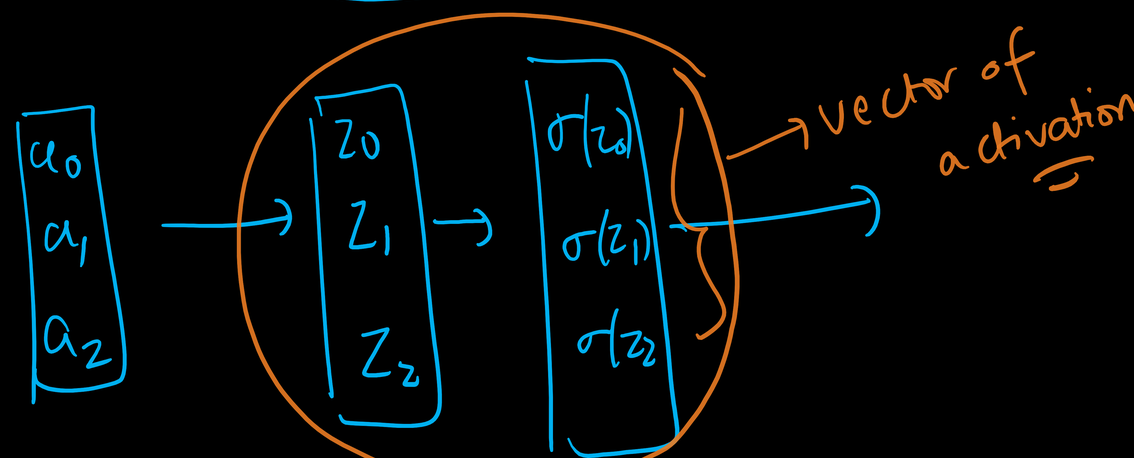
$$z = \sum w_i x_i + b$$

for one input example



$$W^1 = \begin{bmatrix} w^1_{11} & w^1_{12} & w^1_{13} \\ w^1_{21} & w^1_{22} & w^1_{23} \\ w^1_{31} & w^1_{32} & w^1_{33} \end{bmatrix}$$ (3×3)

$$z_j^{[l]} = \sum_i w_{i,j}^{[l]} a_i^{[l-1]} + b$$



$\Downarrow$

$z^{[l]}$

Vector Notation?

$$Z^{[l]} = [w^l]^T \boxed{a^{l-1}} + b^l$$

$(l, l-1) \times (l-1, 1)$

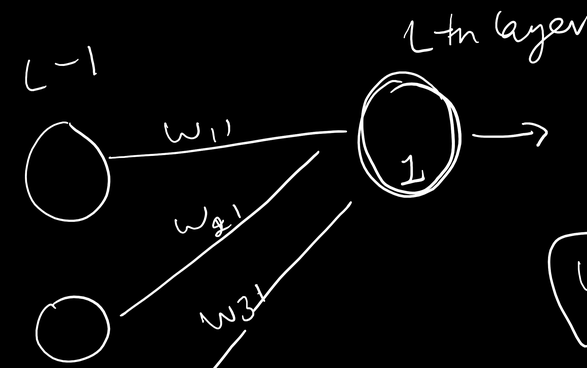
$(l, 1)$

$(l, 1)$

$(w^l)^T \to (l, l-1)$

$\therefore$ l outputs

$$(w^l)^T = \begin{bmatrix} w^l_{11} & w^l_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \end{bmatrix} \begin{bmatrix} a_1^{l-1} \\ a_2^{l-1} \\ a_3^{l-3} \end{bmatrix} = \begin{bmatrix} z_1^{[l]} \\ z_2^{[l]} \\ z_3^{[l]} \end{bmatrix} = Z^{[l]}$$



w_{ij}

$w_{11} a_1 + w_{21} a_2 + w_{31} a_3$

$$a^{[1]} = g_1(z^{[1]}) = \sigma(z^{[1]}) = \begin{bmatrix} \sigma(z_1^{[1]}) \\ \sigma(z_2^{[1]}) \\ \sigma(z_3^{[1]}) \end{bmatrix}$$ ← Output after hidden layer

Forward Propagation over 1 example

$$\begin{aligned}
z_1 &= (w^{[1]})^T a^{[0]} + b^{[1]} \\
a_1 &= \sigma(z_1) \\
z_2 &= (w^{[2]})^T a^{[1]} + b^{[2]} \\
a_2 &= \sigma(z_2) \\
z_3 &= (w^3)^T a^{[2]} + b^{[3]} \\
\hat{y} &= \text{Softmax}(z_3)
\end{aligned}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{(n,1)} \qquad X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ \leftarrow & x^{(i)} & \rightarrow & \\ & & & \end{bmatrix}_{m \times n}$$

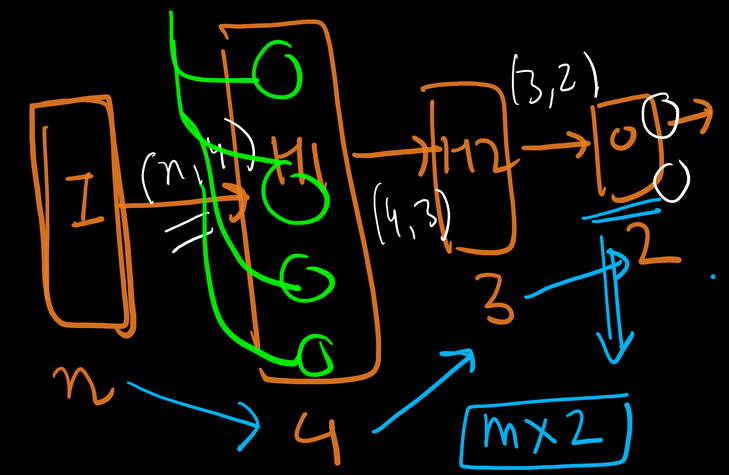
$$A^{[0]} = X$$

$$z_1 = \underset{(m,n)}{A^{[0]}} \cdot \underset{(n,4)}{W^{[1]}} + \underset{(1,4)}{b^{[1]}}$$

$$(m,4)$$

$$X = \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & x_3^2 & & \\ & & & & \end{bmatrix}_{m \times n} \qquad A^{[0]} = X$$

$$W^{[1]} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ | & | & | & & | \end{bmatrix}_{n \times 4}$$



w_{ij} =

10	19
20	30

$m \times 2$

x_j^i = j^{th} feature for i^{th} example

$$Z^{[1]} = \begin{bmatrix} Z_1^{[1](1)} & Z_2^{[1](1)} & Z_3^{[1](1)} & Z_4^{[1](1)} \\ Z_1^{[1](2)} & Z_2^{[1](2)} & Z_3^{[1](2)} & Z_4^{[1](2)} \\ \vdots & & & \end{bmatrix}$$

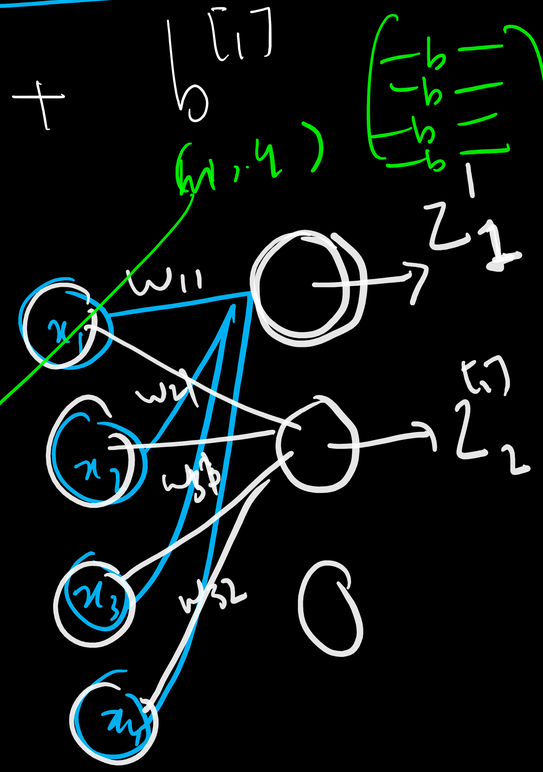
$Z_1^{[1](1)} = x_1^1 w_{11} + x_2^1 w_{21} + \dots + x_n^1 w_{n1} + b_1$

$Z_j^{[l](i)}$: l = layer, i = #example, j = #neuron

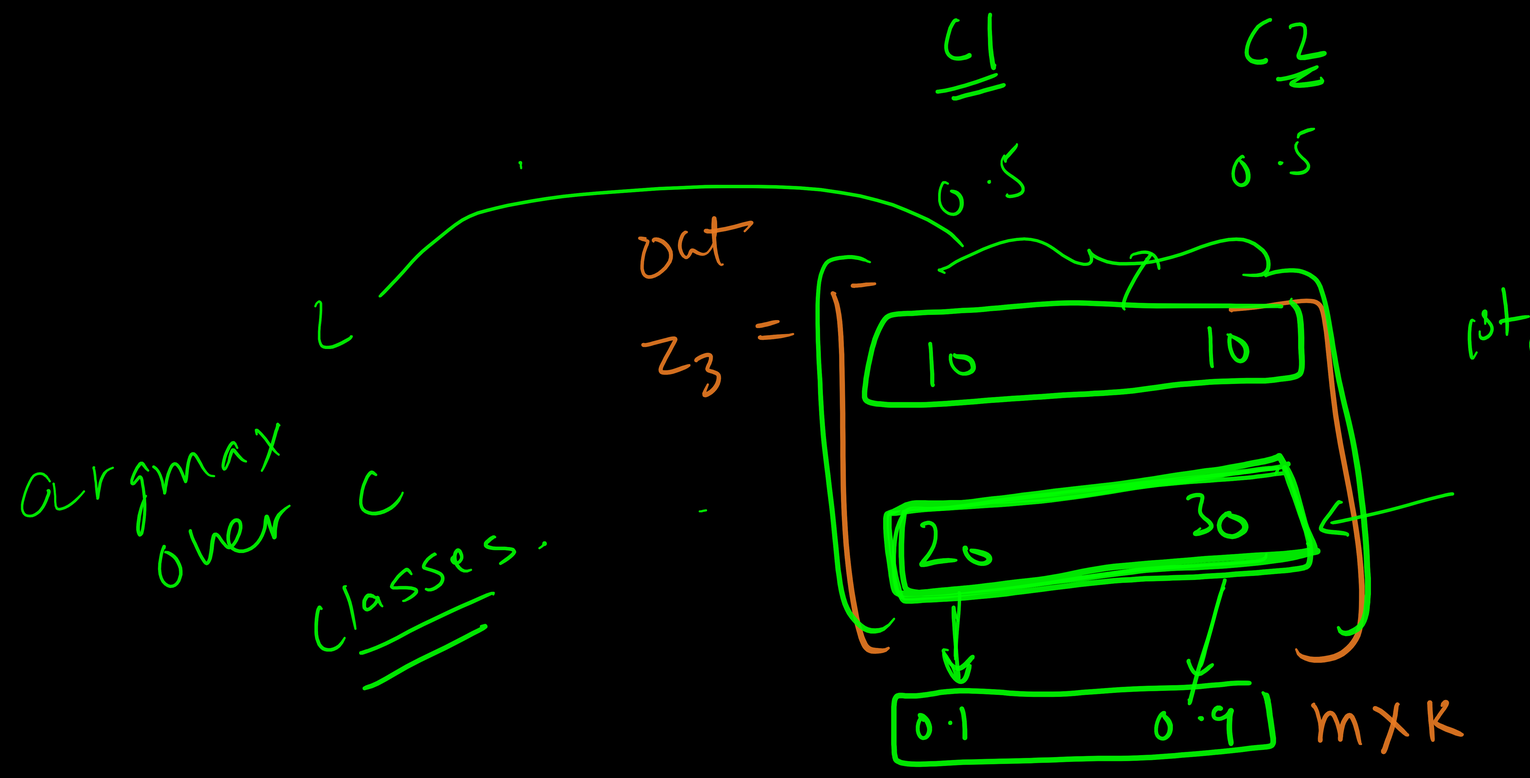
$Z^{[1]} = W^{[1]} X + b^{[1]}$

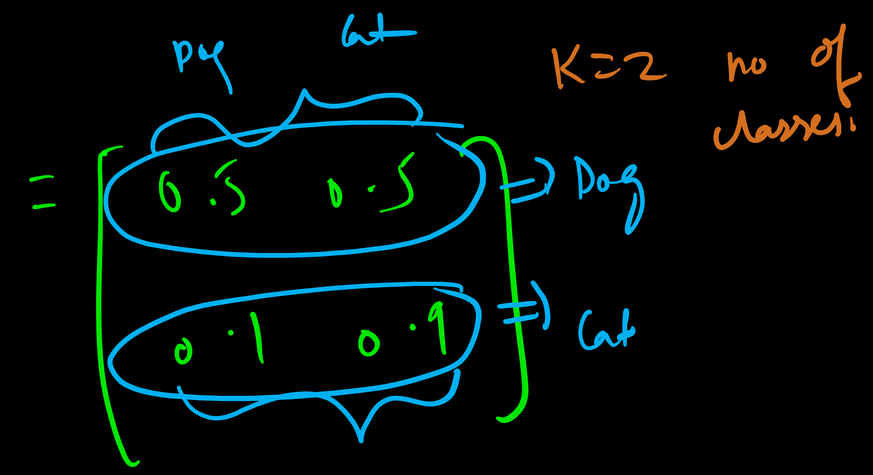
$W^{[1]}$: $n \times 4$ (4 neurons)

$b^{[1]}$: 4×1



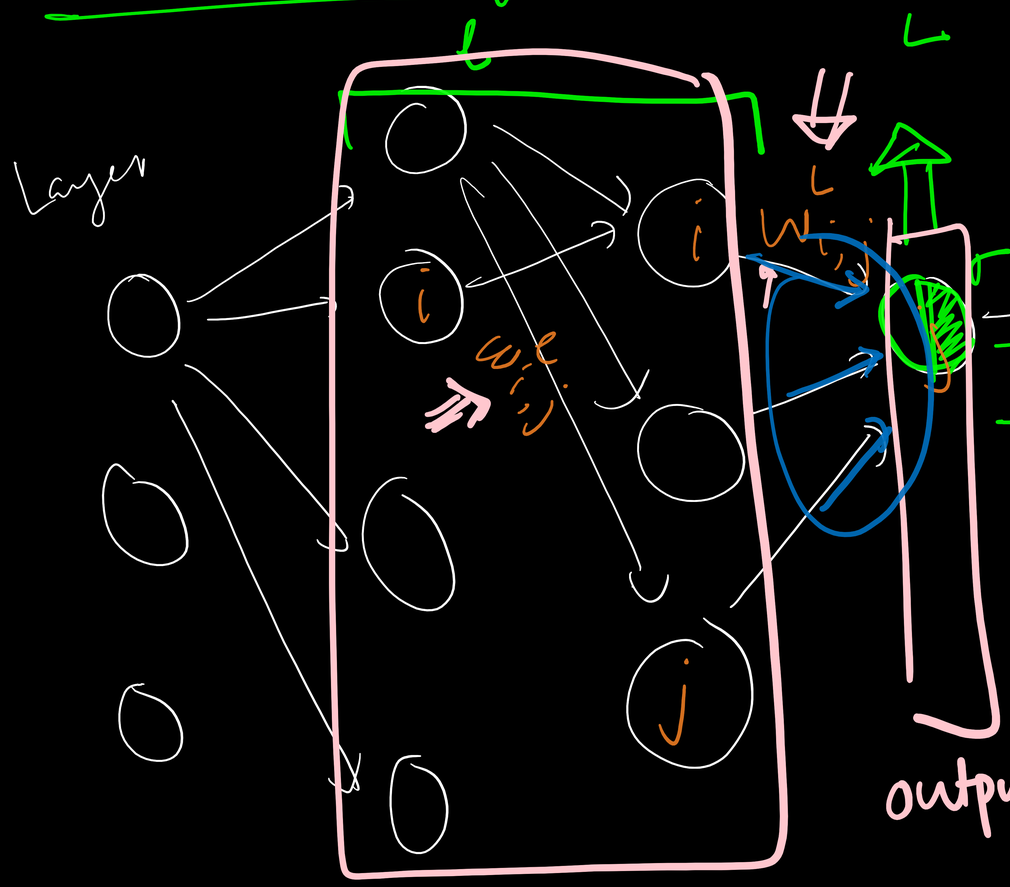
$$A^{[1]} = \sigma\begin{bmatrix} | & | & | & | \\ Z^{(1)} & & & \\ | & | & | & | \end{bmatrix}$$





PART-3

BackPROPAgation



L : last layer

l : hidden layer

0.6 → 1 Dog 1

0.3 → 0 Cat 0

ŷ

y

hidden neurons

For simplification —

$$MSE = L(\hat{y}, y) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

Case-1 output layer.

$$\frac{\partial L}{\partial W_{i,j}^{L}} = ?$$

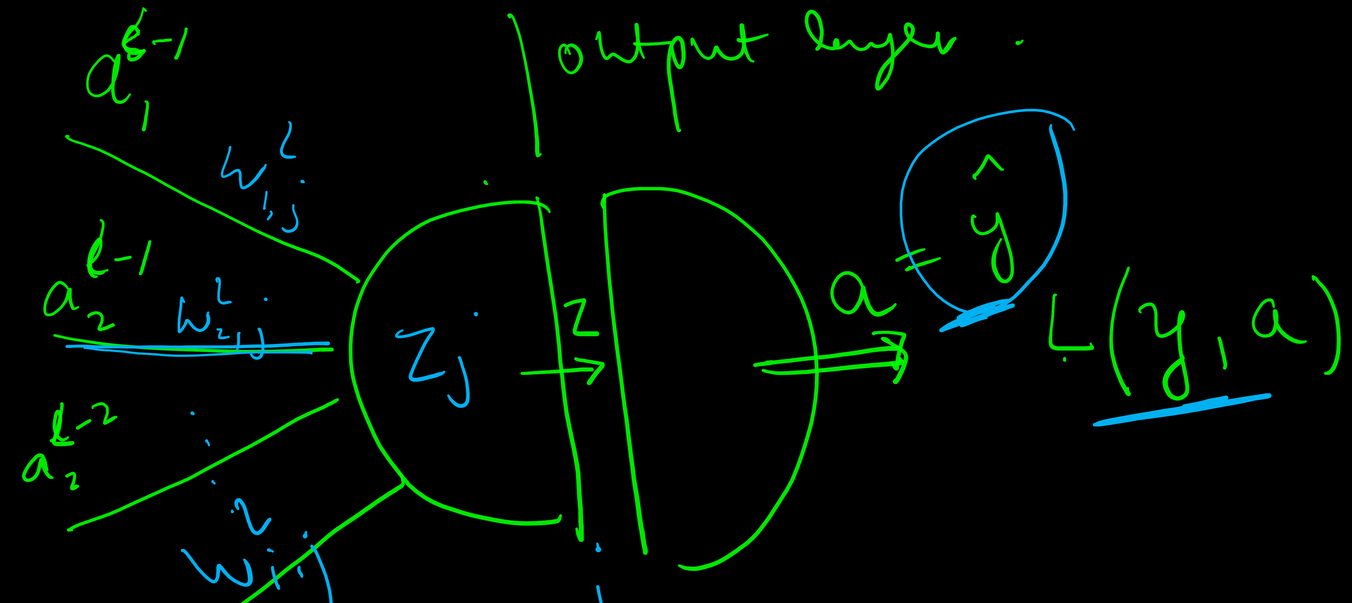
$$\frac{\partial L}{\partial b^{L}} = ?$$

Case-2 hidden layer.

$$\frac{\partial L}{\partial W_{i,j}^{l}} = ?$$

$$\frac{\partial L}{\partial b^{l}} = ?$$

Case-1

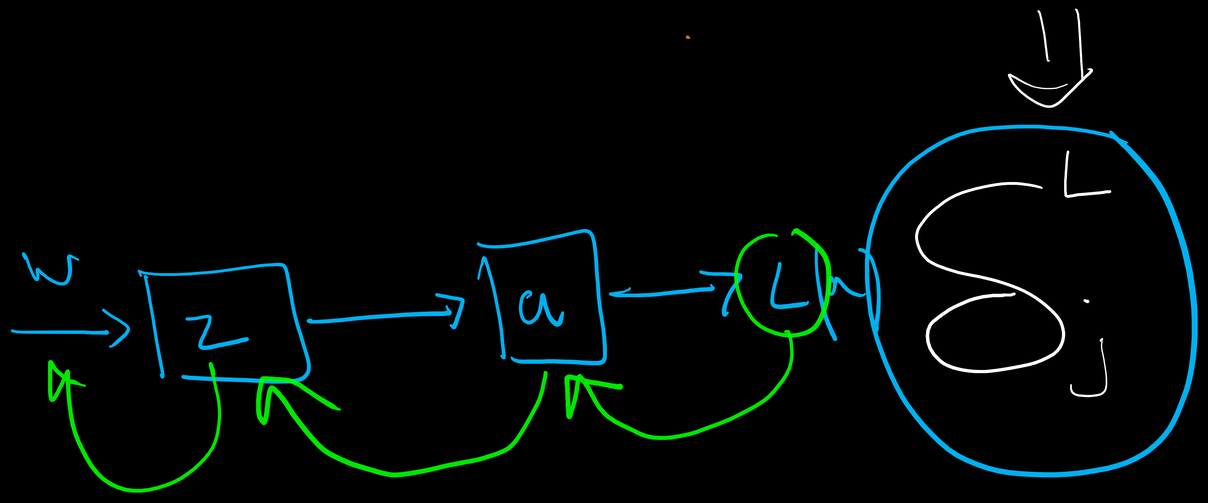


$w_{ij} \rightarrow z_j \rightarrow a_j \rightarrow L(a, y)$

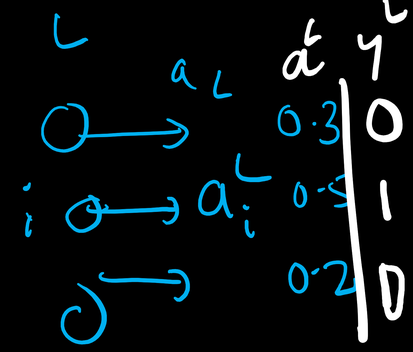
$$z_j = \sum_i^{l} w_{ij}^{l} \, a_i^{l-1} + b_j^{l}$$

$$\frac{\partial L}{\partial w_{ij}} = \underbrace{\frac{\partial L}{\partial a_j}}_{①} \underbrace{\frac{\partial a_j}{\partial z_j}}_{②} \underbrace{\frac{\partial z_j}{\partial w_{ij}}}_{③}$$

Chain Rule



$$\frac{\partial L}{\partial z_j}$$



① $\begin{cases} \frac{\partial L}{\partial a_j} = (y_j - a_j) \end{cases}$

② $\begin{cases} \frac{\partial a_j}{\partial z_j} = \sigma'(z_j) \end{cases}$

③ $\begin{cases} \frac{\partial z_j}{\partial w_{i,j}} = a_i^{l-1} \end{cases}$

① $L = \frac{1}{2} \sum_i (y_i^L - a_i^L)^2$

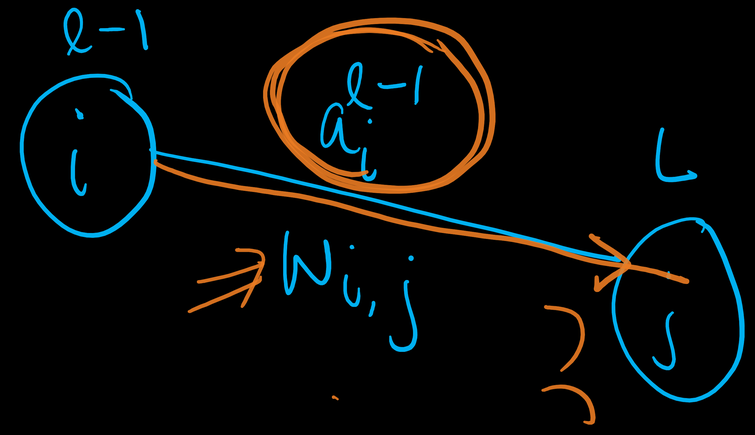
$\frac{\partial L}{\partial a_j} = \frac{1}{2} \cdot 2 (y_j - a_j)^{2-1} \cdot (-1)$

② $a_j = \sigma(z_j)$

$\frac{\partial a_j}{\partial z_j} = \sigma'(z_j)$

$= (1 - \sigma(z_j))\,\sigma(z_j)$

③ $z_j^l = \sum_i w_{i,j}^l a_i^{l-1} + b_j^l$



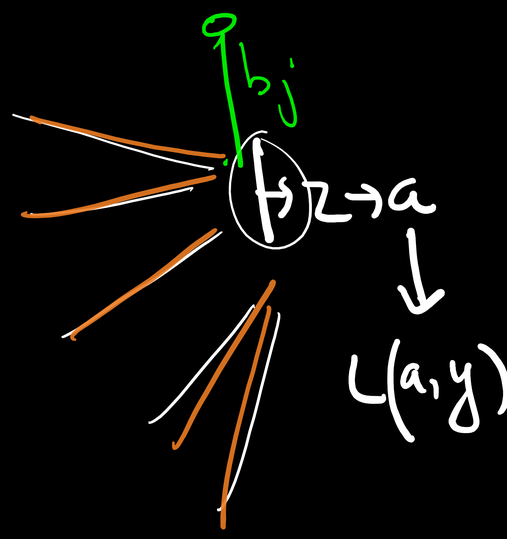
$$\frac{\partial z_j}{\partial w_{ij}} = a_i^{l-1}$$

$$\frac{\partial z_j}{\partial b_j} = 1$$

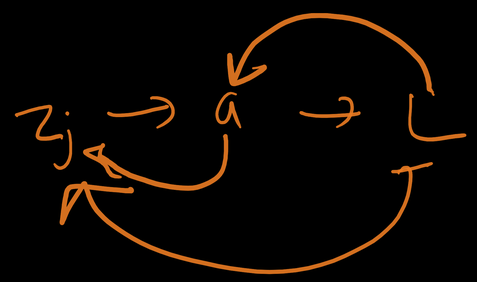
$$\frac{\partial L}{\partial w_{ij}} = \underbrace{(y_j + a_j)\sigma'(z_j)}_{\delta_j^l} a_i^{l-1}$$

$$\Rightarrow \boxed{\frac{\partial L}{\partial w_{ij}}} = \delta_j^l \, a_i^{l-1}$$

$$\Rightarrow \frac{\partial L}{\partial b_j} = \frac{\partial L}{\partial a_j}\left(\frac{\partial a_j}{\partial z_j}\cdot\frac{\partial z_j}{\partial b_j}\right)$$



$$\frac{\partial L}{b_j} = \delta_j^L \cdot 1$$



$$\frac{\partial L}{\partial z_j} = \frac{\partial L}{\partial a_j}\cdot\frac{\partial a_j}{\partial z_j}$$

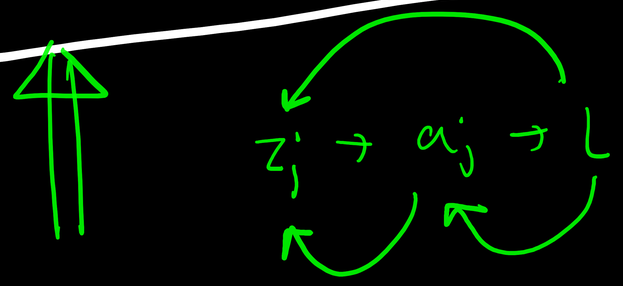
For output layer

$$\textcircled{1}\ \frac{\partial L}{\partial w_{ij}} = \delta_j^L \cdot a_i^{L-1}$$

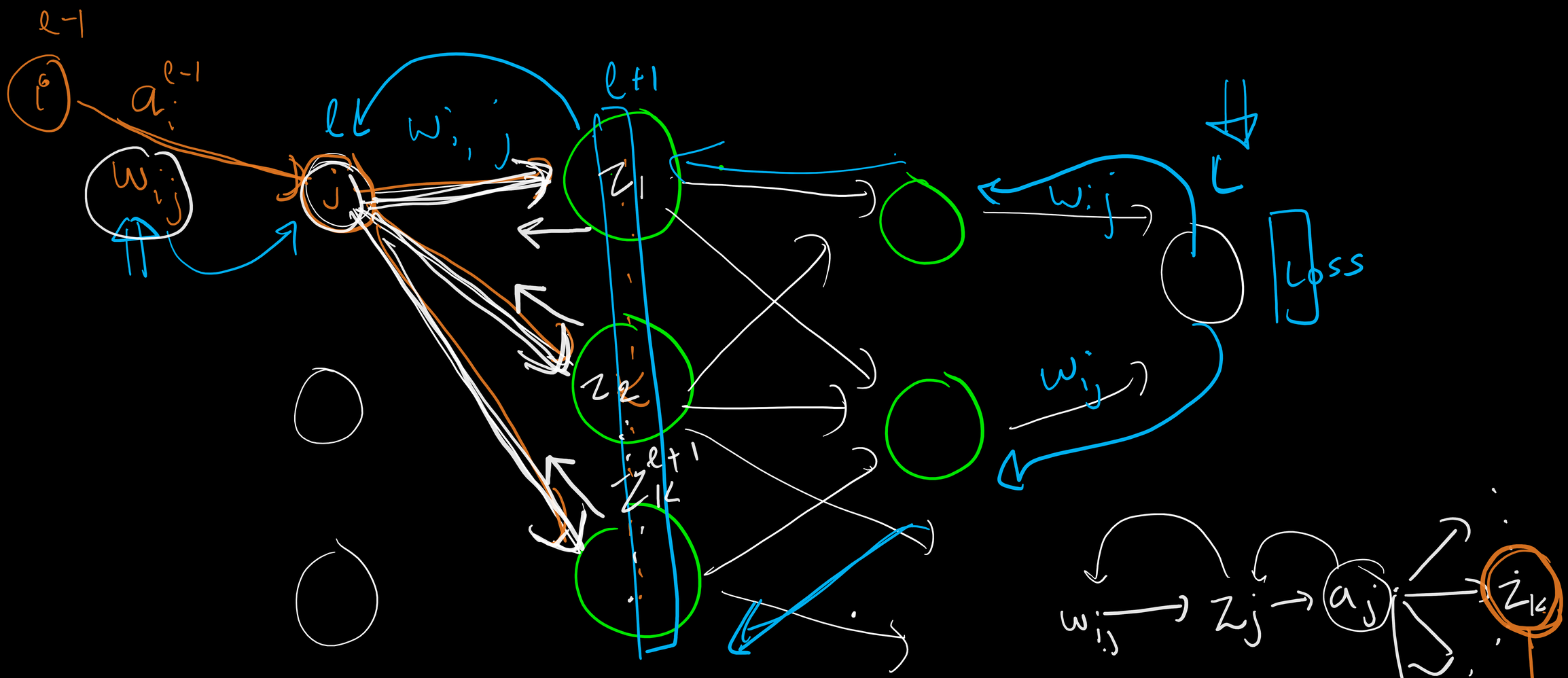
$$\textcircled{2}\ \frac{\partial L}{\partial b_j} = \delta_j^L \cdot 1$$

where -

$$\delta_j^L = (a_j - y_i) \odot \sigma'(z_j)$$



L



$$\frac{\partial L}{\partial w_{ij}^{l}} = \sum_k \frac{\partial L}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}}$$

$$\frac{\partial L}{\partial w_{ij}^{\ell}} = \sum_K \left(\delta_K^{\ell+1} \cdot w_{jk}\right) \sigma'(z_j^{\ell}) \, a_i^{\ell-1}$$

$$z_j = \sum_i w_{ij} \cdot a_i^{\ell-1}$$

$$\frac{\partial z_j}{\partial w_{ij}} = a_i^{\ell-1}$$

$$a_j = \sigma(z_j)$$

$$\frac{\partial a_j}{\partial z_j} = \sigma'(z_j)$$

$$z_K^{\ell+1} = \sum_j w_{jk} \cdot a_j^{\ell}$$

$$\frac{\partial z_K^{\ell+1}}{\partial a_j^{\ell}} = w_{jk}$$

$$\delta_j^{\ell} = \sum_K \left(\delta_K \cdot w_{jk}\right) \sigma'(z_j^{\ell})$$

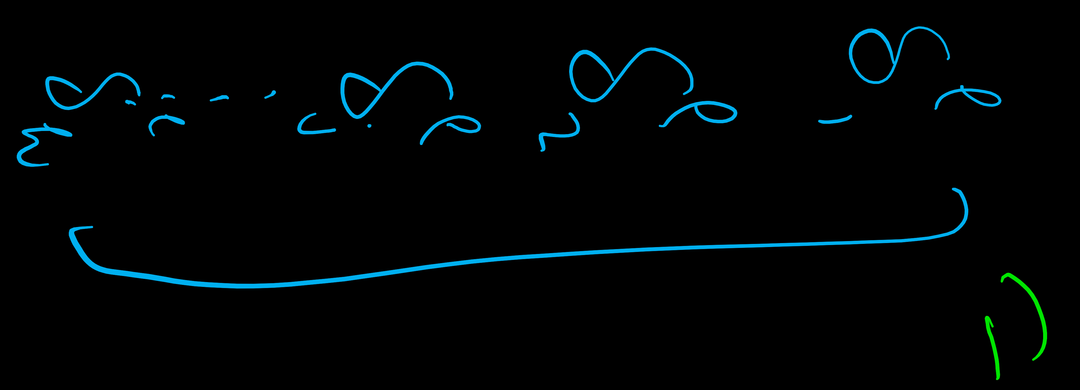
Hidden Neurons

① $\frac{\partial L}{\partial w_{ij}^{l}} = \delta_j^{l} \, a_i^{l-1}$

$\delta_j^{l} = \sum_k \left(w_{jk}^{l+1} \cdot \delta_k^{l+1} \right) \sigma'(z_j^{l})$

② $\frac{\partial L}{\partial b_j^{l}} = \delta_j^{l}$

Matrix / vector



$$\frac{\partial L}{\partial w_{i,j}}$$

$$\frac{\partial L}{\partial W^{l}} = \underbrace{a^{l-1}}_{(\ell-1,1)} \cdot \underbrace{\left(\delta^{l}\right)^{T}}_{(1\times\ell)}$$

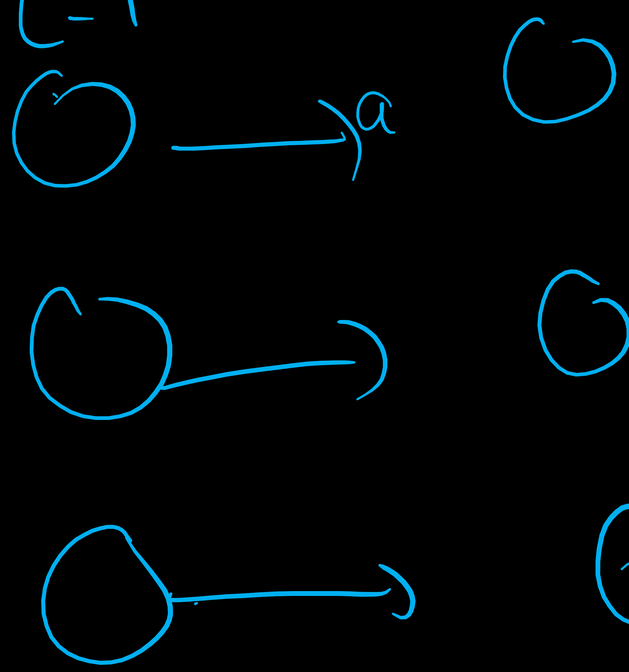
correct

$\delta^{l}_{(\ell,1)}$

$(\ell-1,\ell) \Longleftrightarrow (\ell-1,h)$

$W^{l} = \begin{bmatrix} dw & dw & \cdots \\ \vdots & \vdots & \\ \cdots & \cdots & \end{bmatrix}_{(l-1,\,l)}$

$L-1$



$\begin{bmatrix} a_1^{l-1} \\ a_2^{l-1} \\ a_3^{l-1} \\ \vdots \end{bmatrix} \begin{bmatrix} \delta_1^{l} & \delta_2^{l} & \delta_3^{l} & \cdots & \delta_l^{l} \end{bmatrix}$

$= \begin{bmatrix} a_1^{l-1}\delta_1^{l} & a_1^{l-1}\delta_2^{l} & \cdots \\ a_2^{l-1}\delta_1^{l} & \cdots & \\ \vdots & & \end{bmatrix}$

$\frac{\partial L}{\partial w_{i,j}^{l}}$

$$\frac{\partial L}{\partial b_j^e} = \delta_j^e \rightarrow 0 \qquad b = b - \eta \frac{\partial L}{\partial b}$$

$$\rightarrow 0$$

$$\rightarrow 0$$

$$\frac{\partial L}{\partial b^\ell} = \delta^L = \begin{bmatrix} b_1^e \\ b_2^e \\ b_3^e \end{bmatrix}$$

$$\frac{\partial L}{\partial b} = \begin{bmatrix} \delta_1^L \\ \delta_2^L \\ \delta_3^L \\ \vdots \end{bmatrix}$$

vector

vector

vector/matrix
notation

$$\frac{\partial L}{\partial W^l} = a^{l-1} (\delta^l)^T$$

matrix

$$\frac{\partial L}{\partial b^l} = \delta^l$$

vector

for both
output
hidden
layers.

$$\delta^L = (\underbrace{a^L - y^L}) \odot \underbrace{\sigma'(z^L)}$$

$$\delta^l = (\underbrace{w^{l+1}}_{\cdot} \cdot \delta^{(l+1)}) \odot \underbrace{\sigma'(z^l)}$$