$$\hat{y} = \sigma(W^{(2)}\vec{x}^{(1)} + b^{(2)}) = \sigma(W^{(2)}\sigma(W^{(1)}\vec{x}^{(0)} + \vec{b}^{(1)}) + b^{(2)})$$

$$\text{with } W^{(2)} = \begin{bmatrix} w_1^{(2)} & w_2^{(2)} & w_3^{(2)} \end{bmatrix},\ \vec{b}^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \end{bmatrix},\ \sigma(x) = \frac{1}{1+e^{-x}}$$

$$\text{and } W^{(1)} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \\ w_{31}^{(1)} & w_{32}^{(1)} \end{bmatrix}$$

$$\implies \hat{y} = \sigma(w_1^{(2)}x_1^{(1)} + w_2^{(2)}x_2^{(1)} + w_3^{(2)}x_3^{(1)} + b^{(2)}) \text{ and}$$

$$\begin{aligned} x_1^{(1)} &= \sigma(w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2 + b_1^{(1)}) \\ x_2^{(1)} &= \sigma(w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2 + b_2^{(1)}) \\ x_3^{(1)} &= \sigma(w_{31}^{(1)}x_1 + w_{32}^{(1)}x_2 + b_3^{(1)}) \end{aligned}$$

To simplify the calculation of the gradiënt, I will use $\vec{z}^{(1)}$ and $\vec{z}^{(2)}$ that are defined as:

$$\vec{z}^{(2)} = W^{(2)}\vec{x}^{(1)} + b^{(2)} \text{ and } \vec{z}^{(1)} = W^{(1)}\vec{x}^{(0)} + \vec{b}^{(1)}$$

I define the cost function to be optimised by the neural network as:

$$\text{Cost}(\vec{C}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(\hat{y}(\vec{C},i) - y_i)^2$$

This means that for every data point, we obtain an individual cost $\text{Cost}(\vec{C},i)$ which we average out to calculate the total cost. This means that for calculating the gradient, we calculate the gradient for every point (or only a select few in case of SGD) and average all these vectors to find the next step.

The chosen activation function has a very useful property for calculating the derivative. Namely, $\sigma'(x) = \sigma(x)(1-\sigma(x))$

$$\Delta_{b^{(2)}} = \frac{\partial \text{Cost}(\vec{C},i)}{\partial b^{(2)}} = (\hat{y}(\vec{C},i) - y_i)\sigma'(W^{(2)}\vec{x}^{(1)} + b^{(2)}) = (\hat{y}(\vec{C},i) - y_i)\sigma(z^{(2)})(1-\sigma(z^{(2)}))$$

$$\Delta_{w_j^{(2)}} = \frac{\partial \text{Cost}(\vec{C},i)}{\partial w_j^{(2)}}(\hat{y}(\vec{C},i) - y_i)\sigma'(z^{(2)})\frac{\partial z^{(2)}}{\partial w_j^{(2)}} = \Delta_{b^{(2)}} \cdot x_j^{(1)}$$

$$\Delta_{b_j^{(1)}} = \frac{\partial \text{Cost}(\vec{C},i)}{\partial b_j^{(1)}} = (\hat{y}(\vec{C},i) - y_i)\sigma'(z^{(2)})\frac{\partial z^{(2)}}{\partial x_j^{(1)}}\frac{\partial x_j^{(1)}}{\partial b_j^{(1)}} = \Delta_{b^{(2)}} \cdot w_j^{(2)}\sigma'(\vec{z}_j^{(1)}) = \Delta_{b^{(2)}} \cdot w_j^{(2)}\sigma(\vec{z}_j^{(1)})(1-\sigma(\vec{z}_j^{(1)}))$$

$$\Delta_{w_{jl}^{(1)}} = (\hat{y}(\vec{C},i) - y_i)\sigma'(z^{(2)})\frac{\partial z^{(2)}}{\partial x_j^{(1)}}\frac{\partial x_j^{(1)}}{\partial w_{jl}^{(1)}} = \Delta_{b_j^{(1)}} \cdot x_l$$

which can be concluded to the following table:

$\Delta_{b^{(2)}}$	$(\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i)$
$\Delta_{w_j^{(2)}}$	$\Delta_{b^{(2)}} \cdot x_j^{(1)}$
$\Delta_{b_j^{(1)}}$	$\Delta_{w_j^{(2)}} \cdot w_j^{(2)}(1-x_j^{(1)})$
$\Delta_{w_{jl}^{(1)}}$	$\Delta_{b_j^{(1)}} \cdot x_l$