We consider a neuron network of L layers. The activation states of neurons in the $l+1$ layer is determined by the activation states of neurons in the l layer as follows:

$$a_i^{l+1} = \sigma\left(\sum_{j=1}^{n^l} w_{ij}^l a_j^l + b_i^l\right) = \sigma(z_i^l), \qquad l = 1, ..., L-1$$

where n^l is the number of neurons in layer l. The network input is given by the activation states of the neurons in the first layer, $a_1^1, ..., a_{n_1}^1$. The network parameters are the weights w_{ij}^l and biases b_i^l. It is convenient to define the "weighted inputs" $z_i^l = \sum_j w_{ij}^l a_j^l + b_i^l$.

The cost is defined as a function of the activation states of the last layer,

$$C = C(a_1^L, ..., a_{n_L}^L)$$

Since in turn $a_1^L, ..., a_{n_L}^L$ are functions of the activation states of the previous layer, and so on until the first layer, we have the following recursion:

$$\frac{\partial C}{\partial a_j^l} = \sum_i \frac{\partial C}{\partial a_i^{l+1}} \sigma'(z_i^l) w_{ij}^l$$

Finally, since $a_1^{l+1}, ..., a_{n_{l+1}}^{l+1}$ are functions of the network parameters b_i^l, w_{ij}^l, we obtain:

$$\begin{aligned} \frac{\partial C}{\partial b_i^l} &= \frac{\partial C}{\partial a_i^{l+1}} \sigma'(z_i^l) \\ \frac{\partial C}{\partial w_{ij}^l} &= \frac{\partial C}{\partial a_i^{l+1}} \sigma'(z_i^l) a_j^l \end{aligned}$$

Combining these equations with the recursion above, we obtain first that $\frac{\partial C}{\partial a_j^l} = \sum_i \frac{\partial C}{\partial b_i^l} w_{ij}^l$, and therefore

$$\begin{aligned} \frac{\partial C}{\partial b_i^l} &= \sum_k \frac{\partial C}{\partial b_k^{l+1}} w_{ki}^{l+1} \sigma'(z_i^l) \\ \frac{\partial C}{\partial w_{ij}^l} &= \sum_k \frac{\partial C}{\partial b_k^{l+1}} w_{ki}^{l+1} \sigma'(z_i^l) a_j^l = \frac{\partial C}{\partial b_i^l} a_j^l \end{aligned}$$

The backpropagation algorithm consists of iterating this recursion to compute the gradient with respect to all the parameters. This iteration begins from the gradient with respect to the last layer parameters:

$$\begin{aligned} \frac{\partial C}{\partial b_i^{L-1}} &= \frac{\partial C}{\partial a_i^L} \sigma'(z_i^{L-1}) \\ \frac{\partial C}{\partial w_{ij}^{L-1}} &= \frac{\partial C}{\partial a_i^L} \sigma'(z_i^{L-1}) a_j^{L-1} = \frac{\partial C}{\partial b_i^{L-1}} a_j^{L-1} \end{aligned}$$

which can be computed directly from the form of the cost function.

As in Nielsen's book, it is convenient to define

$$\delta_i^l = \frac{\partial C}{\partial z_i^l} = \frac{\partial C}{\partial a_i^{l+1}} \sigma'(z_i^l), \qquad l = 1, ..., L-1$$

Then

$$\begin{aligned} \frac{\partial C}{\partial b_i^l} &= \delta_i^l \\ \frac{\partial C}{\partial w_{ij}^l} &= \delta_i^l a_j^l \end{aligned}$$

and

$$\delta_i^{l-1} = \sum_k \delta_k^l w_{ki}^l$$