2.5 Proof of the four fundamental equations (optional)

We'll now prove the four fundamental equations (BP1)–(BP4). All four are consequences of the chain rule from multivariable calculus. If you're comfortable with the chain rule, then I strongly encourage you to attempt the derivation yourself before reading on.

Let's begin with Equation (BP1), which gives an expression for the output error, δ^l. To prove this equation, recall that by definition

$$\delta^L_j = \frac{\partial C}{\partial z^L_j}. \tag{2.14}$$

Applying the chain rule, we can re-express the partial derivative above in terms of partial derivatives with respect to the output activations,

$$\delta^L_j = \sum_k \frac{\partial C}{\partial a^L_k}\frac{\partial a^L_k}{\partial z^L_j}, \tag{2.15}$$

where the sum is over all neurons k in the output layer. Of course, the output activation a^L_k of the k-th neuron depends only on the weighted input z^L_j for the j-th neuron when $k = j$. And so $\partial a^L_k/\partial z^L_j$ vanishes when $k \neq j$. As a result we can simplify the previous equation to

$$\delta^L_j = \frac{\partial C}{\partial a^L_j}\frac{\partial a^L_j}{\partial z^L_j}. \tag{2.16}$$

Recalling that $a^L_j = \sigma(z^L_j)$ the second term on the right can be written as $\sigma'(z^L_j)$, and the equation becomes

$$\delta^L_j = \frac{\partial C}{\partial a^L_j}\sigma'(z^L_j), \tag{2.17}$$

which is just (BP1), in component form. Next, we'll prove (BP2), which gives an equation for the error δ^l in terms of the error in the next layer, δ^{l+1}. To do this, we want to rewrite $\delta^l_j = \partial C/\partial z^l_j$ in terms of $\delta^{l+1}_k = \partial C/\partial z^{l+1}_k$. We can do this using the chain rule,

$$\delta^l_j = \frac{\partial C}{\partial z^l_j} = \sum_k \frac{\partial C}{\partial z^{l+1}_k}\frac{\partial z^{l+1}_k}{\partial z^l_j} = \sum_k \frac{\partial z^{l+1}_k}{\partial z^l_j}\delta^{l+1}_k, \tag{2.18}$$

where in the last line we have interchanged the two terms on the right-hand side, and substituted the definition of δ^{l+1}_k. To evaluate the first term on the last line, note that

$$z^{l+1}_k = \sum_j w^{l+1}_{kj} a^l_j + b^{l+1}_k = \sum_j w^{l+1}_{kj}\sigma(z^l_j) + b^{l+1}_k. \tag{2.19}$$

Differentiating, we obtain

$$\frac{\partial z^{l+1}_k}{\partial z^l_j} = w^{l+1}_{kj}\sigma'(z^l_j). \tag{2.20}$$