1 Definitions

All indexing is from 0. L is the number of layers in the neural network, including the input and output layers (so the total number of hidden layers is $N-2$ and the total number of weight matrices is $N-1$). n^ℓ is the number of neurons in a given layer plus one. So, if there are 100 inputs, $n^0 = 101$. This extra "neuron" is always set to 1 to allow for biases to be handled succinctly. σ is the activation function and $\tilde{L}$ is the loss function, and I put no assumptions on them.

$$\begin{aligned}
o_i^\ell &= \sigma(a_i^\ell) && \text{(outputs)} \\
a_i^\ell &= \sum_{i=0}^{n^\ell - 1} o_j^{\ell-1} w_{ji}^{\ell-1} && \text{(activation)} \\
o_i^0 &= \text{NN inputs} \\
o_i^{L-1} &= \text{NN outputs} \\
o_0^\ell &= 1 && \text{(allows for biases)} \\
w_{0i}^\ell &= \text{biases} \\
w_{j0}^\ell &= \delta_{j0}^{(\text{kronecker})} \\
\tilde{L} &= \tilde{L}(o_1^{L-1}, \ldots, o_{n^{L-1}}^{L-1}) && \text{(Loss)}
\end{aligned}$$

$$\delta_i^\ell = -\frac{\partial \tilde{L}}{\partial a_i^\ell}$$

2 Forward Phase

$$\begin{aligned}
o_j^{\ell+1} &= \sigma\left(\sum_{i=0}^{n^\ell} o_i^\ell w_{ij}^\ell\right) && (1 \le i < n^{L-1}) \\
o_j^0 &= I_j^d && \text{(base case)}
\end{aligned}$$

where $o_0^\ell = 1$ always.

3 Backwards Phase Derivations

Theorem.

$$\frac{\partial \tilde{L}}{\partial w_{ij}^\ell} = \delta_j^{\ell+1} o_i^\ell$$

.

Proof. First, consider the definition of $a_k^{\ell+1} = \sum_m o_m^\ell w_{mk}^\ell$. Then its partial derivative with respect to some w_{ij}^ℓ is trivial:

$$\frac{\partial a_k^{\ell+1}}{\partial w_{ij}^\ell} = o_i^\ell \delta_{jk}^{(\text{kronecker})}$$

Now it's easy to compute:

$$\begin{aligned}
\frac{\partial \tilde{L}}{\partial w_{ij}^\ell} &= \sum_k \frac{\partial \tilde{L}}{\partial a_k^{\ell+1}} \frac{\partial a_k^{\ell+1}}{\partial w_{ij}^\ell} \\
&= \sum_k \delta_k^{\ell+1} o_i^\ell \delta_{jk}^{(\text{kronecker})} \\
&= \delta_j^{\ell+1} o_i^\ell
\end{aligned}$$

Theorem. (Backpropagation Formula)

$$\delta_i^\ell = \sigma'(a_i^\ell) \sum_k w_{ik}^\ell \delta_k^{\ell+1}$$

Proof. First note that, like for $\frac{\partial a_i^{\ell+1}}{\partial w_{ij}^\ell}$, we have:

$$\frac{\partial o_j^\ell}{\partial a_i^\ell} = \sigma'(a_i^\ell) \delta_{ij}^k$$

This implies a simple formula for $\frac{\partial a_j^{\ell+1}}{\partial a_i^\ell}$:

$$\begin{aligned}
\frac{\partial a_j^{\ell+1}}{\partial a_i^\ell} &= \frac{\partial}{\partial a_i^\ell} \sum_k o_k^\ell w_{ki}^\ell \\
&= sigma'(a_i^\ell) \delta_{ij}^k
\end{aligned}$$

¡++¿

Then we can expand:

$$
\begin{aligned}
\delta_i^\ell &= \frac{\partial \tilde{L}}{\partial a_i^\ell} \\
&= \sum_k \frac{\partial \tilde{L}}{\partial a_k^{\ell+1}} \frac{\partial a_k^{\ell+1}}{\partial a_i^\ell} \\
&= \sum_k \delta_k^{\ell+1} \frac{\partial a_k^{\ell+1}}{\partial a_i^\ell} \\
&= \sum_k \delta_k^{\ell+1} o_i^\ell \delta_{jk}^{(\text{kronecker})} \\
&= \delta_j^{\ell+1} o_i^\ell
\end{aligned}
$$

4 Backwards Phase

Base case:

$$
\begin{aligned}
\delta_j^{L-1} &= o_j^{L-1}(1 - o_j^{L-1})(o_j^{L-1} - D_j^d) && (0 \leq j < n^{L-1}) \\
\delta_j^\ell &= o_j^\ell (1 - o_j^\ell) \sum_k w_{jk}^\ell \delta_k^{\ell+1} && 0 \leq j < n^\ell
\end{aligned}
$$

Note that this is kind of funky for the term which would affect the biases, δ_0^ℓ. because $o_0^\ell = 1$, $\delta_0^\ell = 0$ always.

5 Stepping

$$
\Delta w_{ij}^\ell = -\alpha \delta_j^{\ell+1} o_i^\ell
$$

Note that Δw_{i0}^ℓ will always evaluate to zero. So this column is never changed. But Δ_{0j}^ℓ can evaluate to nonzero values, so the biases are indeed updated.