The Backpropagation Algorithm

1 Feedforward Neural Network

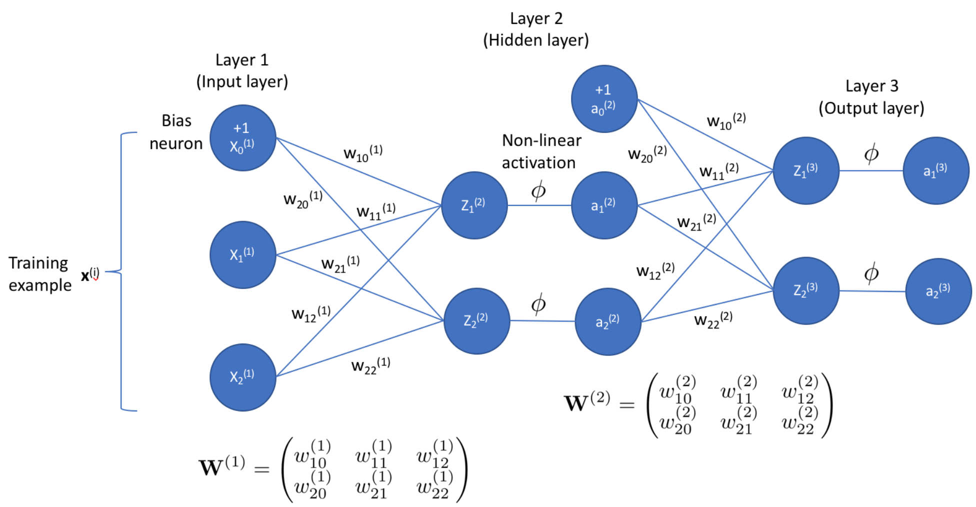

Figure 1:

Assume a cost function C, first begin backpropagation from the output layer L, where $L = 3$ from the given example and ignoring the weights on the bias neurons for now

$$\frac{\partial C}{\partial w_{11}^{(2)}} = \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{\partial w_{11}^{(2)}} \qquad \frac{\partial C}{\partial w_{12}^{(2)}} = \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{\partial w_{12}^{(2)}}$$

$$\frac{\partial C}{\partial w_{21}^{(2)}} = \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial w_{21}^{(2)}} \qquad \frac{\partial C}{\partial w_{22}^{(2)}} = \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial w_{22}^{(2)}}$$

In matrix form,

$$\frac{\partial C}{\partial \mathbf{W}^{(2)}} = \begin{pmatrix} \frac{\partial C}{\partial w_{11}^{(2)}} & \frac{\partial C}{\partial w_{12}^{(2)}} \\ \frac{\partial C}{\partial w_{21}^{(2)}} & \frac{\partial C}{\partial w_{22}^{(2)}} \end{pmatrix} = \begin{pmatrix} \frac{\partial C}{\partial a_1^{(3)}} \\ \frac{\partial C}{\partial a_2^{(3)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \\ \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \end{pmatrix} \odot \underbrace{\begin{pmatrix} \frac{\partial Z_1^{(3)}}{\partial w_{11}^{(2)}} & \frac{\partial Z_1^{(3)}}{\partial w_{12}^{(2)}} \\ \frac{\partial Z_2^{(3)}}{\partial w_{21}^{(2)}} & \frac{\partial Z_2^{(3)}}{\partial w_{22}^{(2)}} \end{pmatrix}}_{\frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(2)}}} = \delta^L \odot \frac{\partial \mathbf{Z}^{(3)}}{\partial \mathbf{W}^{(2)}}$$

Now for the hidden layer,

$$\begin{aligned}
\frac{\partial C}{\partial w_{11}^{(1)}} &= \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} \\
&+ \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} \\
&= \left(\frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \right) \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} \\
&= \left(\delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \right) \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial C}{\partial w_{12}^{(1)}} &= \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}} \\
&+ \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}} \\
&= \left(\delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \right) \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial C}{\partial w_{21}^{(1)}} &= \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}} \\
&+ \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}} \\
&= \left(\delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_2^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \right) \cdot \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial C}{\partial w_{22}^{(1)}} &= \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}} \\
&+ \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \cdot \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}} \\
&= \left(\delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_2^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \right) \cdot \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \cdot \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}}
\end{aligned}$$

In matrix form,

$$
\begin{aligned}
\frac{\partial C}{\partial \mathbf{W}^{(1)}} &= \begin{pmatrix} \frac{\partial C}{\partial w_{11}^{(1)}} & \frac{\partial C}{\partial w_{12}^{(1)}} \\ \frac{\partial C}{\partial w_{21}^{(1)}} & \frac{\partial C}{\partial w_{22}^{(1)}} \end{pmatrix} \\
&= \begin{pmatrix} \delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \\ \delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_2^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \\ \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}} \\ \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}} & \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}} \end{pmatrix} \\
&= \left[\begin{pmatrix} \delta_1^L & \delta_2^L \end{pmatrix} \begin{pmatrix} \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} & \frac{\partial Z_1^{(3)}}{\partial a_2^{(2)}} \\ \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} & \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \end{pmatrix} \right]^\top \odot \begin{pmatrix} \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \\ \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}} \\ \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}} & \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}} \end{pmatrix} \\
&= \left[\begin{pmatrix} \delta_1^L & \delta_2^L \end{pmatrix} \begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} \end{pmatrix} \right]^\top \odot \begin{pmatrix} \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \\ \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} X_1^{(1)} & X_2^{(1)} \\ X_1^{(1)} & X_2^{(1)} \end{pmatrix} \\
&\equiv \underbrace{\mathbf{W}^{(2)^\top} \delta^L \odot \sigma'(\mathbf{Z}^{(L-1)})}_{\delta^{(L-1)}} \odot \frac{\partial Z^{(2)}}{\partial \mathbf{W}^{(1)}}
\end{aligned}
$$

where

$$
\begin{aligned}
\delta^{(L-1)} &= \frac{\partial C}{\partial \mathbf{Z}^{(L-1)}} = (\mathbf{W}^{(L-1)})^\top \delta^L \odot \sigma'(\mathbf{Z}^{(L-1)}) \\
&\implies \boxed{\delta^l = \left((\mathbf{W}^l)^\top \delta^{(l+1)}\right) \odot \sigma'(\mathbf{Z}^{(l)})} \quad \text{for } 1 < l < L
\end{aligned}
$$

For the weights on the bias neurons,

$$
\frac{\partial C}{\partial w_{10}^{(2)}} = \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{\partial w_{10}^{(2)}} = \delta_1^L \cdot 1 \qquad \frac{\partial C}{\partial w_{20}^{(2)}} = \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial w_{20}^{(2)}} = \delta_2^L \cdot 1
$$

$$
\begin{aligned}
\frac{\partial C}{\partial w_{10}^{(1)}} &= \frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{10}^{(1)}} \\
&\quad + \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{10}^{(1)}} \\
&= \left(\frac{\partial C}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial Z_1^{(3)}} \cdot \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \frac{\partial C}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial Z_2^{(3)}} \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \right) \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{10}^{(1)}} \\
&= \left(\delta_1^L \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \delta_2^L \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \right) \cdot \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \cdot \frac{\partial Z_1^{(2)}}{\partial w_{10}^{(1)}} \\
&= \delta_1^{(L-1)}
\end{aligned}
$$

Equivalently, we have $\dfrac{\partial C}{\partial w_{20}^{(1)}} = \delta_2^{(L-1)}$. In particular, $\frac{\partial C}{\partial \mathbf{b}^{(l)}} = \delta^{(l+1)}$

Finally, let's combine the partial derivative of the weight matrix above with the partial derivatives wrt to the weights of the bias neurons

$$
\begin{aligned}
\frac{\partial C}{\partial \mathbf{W}^{(1)}} &= \begin{pmatrix} \frac{\partial C}{\partial w_{10}^{(1)}} & \frac{\partial C}{\partial w_{11}^{(1)}} & \frac{\partial C}{\partial w_{12}^{(1)}} \\ \frac{\partial C}{\partial w_{20}^{(1)}} & \frac{\partial C}{\partial w_{21}^{(1)}} & \frac{\partial C}{\partial w_{22}^{(1)}} \end{pmatrix} \\
&= \begin{pmatrix} \delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} \\ \delta_1^L \cdot \frac{\partial Z_1^{(3)}}{\partial a_2^{(2)}} + \delta_2^L \cdot \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \\ \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial Z_1^{(2)}}{\partial w_{10}^{(1)}} & \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}} \\ \frac{\partial Z_2^{(2)}}{\partial w_{20}^{(1)}} & \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}} & \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}} \end{pmatrix} \\
&= \left[\begin{pmatrix} \delta_1^L & \delta_2^{(L)} \end{pmatrix} \begin{pmatrix} \frac{\partial Z_1^{(3)}}{\partial a_1^{(2)}} & \frac{\partial Z_1^{(3)}}{\partial a_2^{(2)}} \\ \frac{\partial Z_2^{(3)}}{\partial a_1^{(2)}} & \frac{\partial Z_2^{(3)}}{\partial a_2^{(2)}} \end{pmatrix} \right]^\top \odot \begin{pmatrix} \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \\ \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \end{pmatrix} \odot \begin{pmatrix} \frac{\partial Z_1^{(2)}}{\partial w_{10}^{(1)}} & \frac{\partial Z_1^{(2)}}{\partial w_{11}^{(1)}} & \frac{\partial Z_1^{(2)}}{\partial w_{12}^{(1)}} \\ \frac{\partial Z_2^{(2)}}{\partial w_{20}^{(1)}} & \frac{\partial Z_2^{(2)}}{\partial w_{21}^{(1)}} & \frac{\partial Z_2^{(2)}}{\partial w_{22}^{(1)}} \end{pmatrix} \\
&= \underbrace{\left[\begin{pmatrix} \delta_1^L & \delta_2^{(L)} \end{pmatrix} \begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} \end{pmatrix} \right]^\top \odot \begin{pmatrix} \frac{\partial a_1^{(2)}}{\partial Z_1^{(2)}} \\ \frac{\partial a_2^{(2)}}{\partial Z_2^{(2)}} \end{pmatrix}}_{\delta^2} \odot \begin{pmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(1)} & X_2^{(1)} \end{pmatrix} \\
&= \begin{pmatrix} \delta_1^2 \\ \delta_2^2 \end{pmatrix} \odot \begin{pmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(1)} & X_2^{(1)} \end{pmatrix} \\
&= \begin{pmatrix} \delta_1^2 \\ \delta_2^2 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & a_1^{(1)} & a_2^{(1)} \end{pmatrix}}_{(\mathbf{a}^{(1)})^\top}
\end{aligned}
$$

where $\mathbf{a}^{(1)} = \begin{pmatrix} 1 \\ X_1^{(1)} \\ X_2^{(1)} \end{pmatrix}$

In summary, we have the following key backpropagation formulas

$$
\begin{aligned}
\delta^L &= \frac{\partial C}{\partial \mathbf{a}^L} \odot \sigma'(\mathbf{Z}^L) = \frac{\partial C}{\partial \mathbf{Z}^L} \\
\delta^l &= \left((\mathbf{W}^l)^\top \delta^{(l+1)} \right) \odot \sigma'(\mathbf{Z}^{(l)}) \quad \text{for } 1 < l < L \text{, where } \mathbf{W}^l \text{ has weights on bias units removed} \\
\frac{\partial C}{\partial \mathbf{b}^l} &= \delta^{l+1} \\
\frac{\partial C}{\partial w_{jk}^l} &= \mathbf{a}_k^l \delta_j^{(l+1)} \quad j \text{ indexes unit in } (l+1)\text{th layer, } k \text{ indexes unit in } l\text{th layer} \\
\implies \frac{\partial C}{\partial \mathbf{w}^l} &= \delta^{(l+1)} (\mathbf{a}^{(l)})^\top \quad \text{(Vectorised update)}
\end{aligned}
$$

Note: For $m > 1$ training examples, $\frac{\partial C}{\partial \mathbf{w}^l} = \frac{1}{m} \sum_{j=1}^{m} \frac{\partial C^{(j)}}{\partial \mathbf{w}^l}$ i.e. average gradient of all training examples.

2 Initialisation of Weight Matrix

Depending on the cost and/or activation function, weights of neural networks are typically initialized randomly to small values. For example, if we have a sigmoid activation or anything where $\phi(0) \neq 0$, then weights would "move" together during the gradient descent update since outputs from all units would be identical. If we use a *tanh* or *ReLU* activation or anything where $\phi(0) = 0$, then all outputs will be 0 and there would be no learning since all gradients are 0.