Derivation of the Backprop Learning Rule

David S. Touretzky
15-486/782: Artificial Neural Networks

1. Unit Activation Equation

$$net_k = \sum_j y_j \cdot w_{jk} \tag{1}$$

$$y_k = f(net_k) \tag{2}$$

The transfer function $f(\cdot)$ can be any smooth, differentiable, nonlinear function. Originally the logistic function $(1+\exp(-x))^{-1}$ was used, but many today favor tanh because its range is $[-1,+1]$ instead of $[0,1]$, which gives better learning behavior.

2. Error Measure

Error E is summed over all patterns and all output units. The summation over patterns is left implicit below. d_k is the desired output value for unit k on the present pattern, and y_k is the actual output produced by unit k.

$$E = \frac{1}{2}\sum_k (d_k - y_k)^2 \tag{3}$$

3. Error of the Output Layer

δ_k is the gradient of the error with respect to unit k's input. It is backpropagated to the preceding layer to calculate δ_j, and also used to calculate the weight update Δw_{jk}.

$$\frac{\partial E}{\partial y_k} = (y_k - d_k) \tag{4}$$

$$\begin{aligned}
\delta_k &= \frac{\partial E}{\partial net_k} && (5)\\
&= \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial net_k} && (6)\\
&= (y_k - d_k) \cdot f'(net_k) && (7)
\end{aligned}$$

4. Backpropagated Error for Hidden Units

We back-propagate the error through the w_{jk} connections to calculate the error signal for hidden unit j.

$$\frac{\partial E}{\partial y_j} = \sum_k \left(\frac{\partial E}{\partial net_k} \cdot \frac{\partial net_k}{\partial y_j} \right) \tag{8}$$

$$= \sum_k (\delta_k \cdot w_{jk}) \tag{9}$$

$$\delta_j = \frac{\partial E}{\partial net_j} \tag{10}$$

$$= \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial net_j} \tag{11}$$

$$= \frac{\partial E}{\partial y_j} \cdot f'(net_j) \tag{12}$$

5. Weight Update

We update the weights by the negative of the error gradient (because we want error to decrease), scaled by a learning rate η.

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial net_k} \cdot \frac{\partial net_k}{\partial w_{jk}} \tag{13}$$

$$= \delta_k \cdot y_j \tag{14}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}} \tag{15}$$

$$= \delta_j \cdot y_i \tag{16}$$

$$\Delta w_{jk} = -\eta \cdot \frac{\partial E}{\partial w_{jk}} \tag{17}$$

$$\Delta w_{ij} = -\eta \cdot \frac{\partial E}{\partial w_{ij}} \tag{18}$$