notation

E is error ($\equiv$ loss)
o is output, of a neuron, after activation
s is sum of weight * underlying layer outputs, before activation
w is weight
o^2 is output of second layer
o_i^2 is output of node i in layer 2
w_{ji}^2 is weight from node j in layer 1 to node i in layer 2
layers are arranged as: 0 is input layer, then layer 1, layer 2 etc
$a(x)$ is activation function
y_i^* is label i, ie the ground truth for node i, in final output layer

overall

$$\frac{\partial E}{\partial w_{ji}^{l-1}} = \frac{\partial E}{\partial o_i}\frac{\partial o_i}{\partial s_i}\frac{\partial s_i}{\partial w_{ji}}$$

$$= \frac{\partial \text{loss}}{\partial o_i^l}\frac{\partial \text{activation}}{\partial s_i^l} o_j^{l-1}$$

$$= \frac{\partial \text{loss}}{\partial s_i^l} o_j^{l-1}$$

Recursion:

$$\frac{\partial E}{\partial s_i^{l-1}} = \sum_k \frac{\partial E}{\partial s_k^l}\frac{\partial s_k^l}{\partial o_i^{l-1}}\frac{\partial o_i^{l-1}}{\partial s_i^{l-1}}$$

$$= \frac{\partial \text{activation}_i^{l-1}}{\partial s_i^{l-1}} \sum_k (\text{loss from } l)_k w_{ik}^l$$

Alternatively,

$$\frac{\partial E}{\partial s_i^l} = \frac{\partial \text{activation}_i^l}{\partial s_i^l} \sum_k (\text{loss from } l+1)_k w_{ik}^{l+1}$$

Can also recurse on $\frac{\partial E}{\partial o_i^l}$: $\frac{\partial E}{\partial o_i^{l-1}} = \sum_k \frac{\partial E}{\partial o_k^l}\frac{\partial o_k^l}{\partial s_k^l}\frac{\partial s_k^l}{\partial o_i^{l-1}} = \sum_k (\text{loss from } l)_k \frac{\partial \text{activation}_k^l}{\partial s_k^l} w_{ik}^l$

loss

Squared error

$$E = \sum_i \frac{1}{2}(o_i - y_i^*)^2$$

$$\frac{\partial E}{\partial o_i} = o_i - y_i^*$$

Cross-entropy

$$E = -\sum_i (y_i^* \log o_i + (1 - y_i^*)\log(1 - o_i))$$

$$\frac{\partial E}{\partial o_i} = \frac{o_i - y_i^*}{o_i(1 - o_i)}$$

Multinomial cross-entropy

$$E = -\sum_i y_i^* \log o_i$$

$$\frac{\partial E}{\partial o_i} = -\frac{y_i^*}{o_i}$$

activation

sigmoid

$$o_i = \sigma(s_i)$$

$$\frac{\partial o_i}{\partial s_i} = o_i(1 - o_i)$$

tanh

$$o_i = \tanh(s_i)$$

$$\frac{\partial o_i}{\partial s_i} = 1 - (o_i)^2$$

relu

$$o_i = \begin{cases} s_i & \text{when } s_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial o_i}{\partial s_i} = \begin{cases} 1 & \text{when } o_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

linear

$$o_i = s_i$$

$$\frac{\partial o_i}{\partial s_i} = 1$$

softmax

$$o_i = \frac{\exp s_i}{\sum_k \exp s_k} = \frac{\exp(s_i - \max_j s_j)}{\sum_k \exp(s_k - \max_j s_j)}$$

$$\frac{\partial o_i}{\partial s_j} = o_i(\delta_{i,j} - o_j)$$