1. Network structure

The network has n layers. Layer 1 is the input layer, and layer n is the output layer. Each layer i has m_i nodes.

Each node (except in the first layer) has an input z_{ij}. Here i indicates the layer number and j indicates the node number within the layer. Each node has an output y_{ij}.

2. Forward pass calculations

The outputs in the first layer are set to an input pattern, $y_{1i} = I_i$.

The inputs of the remaining nodes are calculated as $z_{ij} = \sum_k W_{jk}^{i-1} y_{(i-1)k}$. Here W^p is the matrix of weights that connects layer p to layer $p + 1$.

The outputs of the remaining nodes are calculated as $y_{ij} = f(z_{ij})$. f is called the activation function.

3. Backward pass calculations

The error of the network's response to an input pattern is $E = 0.5 \sum_i (y_{ni} - O_i)^2$. Here O_i is the targetted output pattern.

As a first step towards calculating the gradient of the error with respect to the weights, we calculate the delta terms, defined as $\delta_{ij} = \partial E / \partial z_{ij}$.

The deltas at the output layer n are given by

$$\delta_{ni} = \frac{\partial E}{\partial z_{ni}} = \frac{\partial E}{\partial y_{ni}} \frac{\partial y_{ni}}{\partial z_{ni}} = (y_{ni} - O_i) f'(z_{ni})$$

The deltas at earlier layers $1 < k < n$ are given by

$$\delta_{ki} = \frac{\partial E}{\partial z_{ki}} = \sum_j \frac{\partial E}{\partial z_{(k+1)j}} \frac{\partial z_{(k+1)j}}{\partial y_{ki}} \frac{\partial y_{ki}}{\partial z_{ki}} = \sum_j \delta_{(k+1)j} W_{ji}^k f'(z_{ki})$$

$$= f'(z_{ki}) \sum_j \delta_{(k+1)j} W_{ji}^k$$

That is, δ_{ki} is calculated as a weighted sum of the $\delta_{(k+1)j}$'s, using the same weights as in the forward pass.

Finally, we can use the deltas to find the error gradient with respect to the weights.

$$\frac{\partial E}{\partial W_{ij}^k} = \frac{\partial E}{\partial z_{(k+1)i}} \frac{\partial z_{(k+1)i}}{\partial W_{ij}^k} = \delta_{(k+1)i} y_{kj}$$