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_j^i. Here i indicates the layer number and j indicates the node number within the layer. Each node also has an output y_j^i.

2. Forward pass

The outputs of the first layer are set to an input pattern, $y_j^1 = I_j$.

The inputs of the remaining nodes are calculated as $z_j^i = \sum_k W_{jk}^{i-1} y_k^{i-1}$. 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 $y_j^i = f(z_j^i)$. f is called the activation function.

3. Backward pass

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

As a first step towards calculating the gradient of the error with respect to the weights W^p, we calculate the delta terms, defined as $\delta_j^i = \partial E / \partial z_j^i$.

The deltas at the output layer n are given by

$$\delta_i^n = \frac{\partial E}{\partial z_i^n} = \frac{\partial E}{\partial y_i^n} \frac{\partial y_i^n}{\partial z_i^n} = (y_i^n - O_i) f'(z_i^n)$$

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

$$\begin{aligned} \delta_i^k &= \frac{\partial E}{\partial z_i^k} = \sum_j \frac{\partial E}{\partial z_j^{k+1}} \frac{\partial z_j^{k+1}}{\partial y_i^k} \frac{\partial y_i^k}{\partial z_i^k} \\ &= \sum_j \delta_j^{k+1} W_{ji}^k f'(z_i^k) \\ &= f'(z_i^k) \sum_j \delta_j^{k+1} W_{ji}^k \end{aligned}$$

That is, δ_i^k is calculated as a weighted sum of the δ_j^{k+1}'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.

$$\frac{\partial E}{\partial W_{ij}^k} = \frac{\partial E}{\partial z_i^{k+1}} \frac{\partial z_i^{k+1}}{\partial W_{ij}^k} = \delta_i^{k+1} y_j^k$$