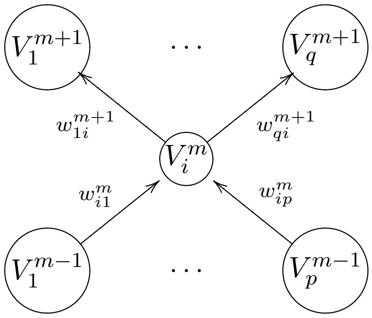
Error Back Propagation

The algorithm is described with respect to a single node in the network. The network is initialized by assigning small random values to each weight. It is trained by repeated epoches.



Temporary variables: $\delta_i^m,\ h_i^m$

Forward propagation:

Given the example $x_1, \ldots, x_n$ with the desired outputs $y_1, \ldots, y_k$, the values of all nodes are computed recursively as follows.

Initialization: $$V_i^0 = x_i \qquad i = 1, \ldots, n$$

Recursive step: $$h_i^m = \sum_{j=1}^{p} w_{ij}^m V_j^{m-1}, \quad V_i^m = g(h_i^m) \qquad \text{for all } i, m$$

This produces values for the top nodes V_i^M, $i = 1, \ldots k$. Produce the output: $O_i = V_i^M$, $i = 1, \ldots k$.

Error back propagation:

The values of the temporary variables δ_i^m are computed recursively as follows:

Initialization: $$\delta_j^M = g'(h_j^M)(y_j - V_j^M) \qquad j = 1, \ldots, k$$

Recursive step: $$\delta_i^m = g'(h_i^m) \sum_{j=1}^{q} w_{ji}^{m+1} \delta_j^{m+1}, \qquad \text{for all } i, m$$

Weights update:

$$w_{ij}^m = w_{ij}^m + \epsilon \delta_i^m V_j^{m-1}$$

Related formulas:

linear: $g(h) = h,$ $\qquad g'(h) = 1$

sigmoid: $g(h) = \frac{1}{1+e^{-2\beta h}},$ $\qquad g' = 2\beta g(1-g)$

hyperbolic tangent: $g(h) = \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}},$ $\qquad g'(h) = \beta(1 - g^2)$

ReLU: $g(h) = \max(0, h) = \begin{cases} 0 & h \leq 0 \\ h & h > 0 \end{cases},$ $\quad g'(h) = \begin{cases} 0 & h \leq 0 \\ 1 & h > 0 \end{cases}$

Correctness proof for BP

The Error-Back-Propagation algorithm updates the weights according to:

$$w_{ij}^m = w_{ij}^m + \epsilon \delta_i^m V_j^{m-1} \quad \text{where} \quad \delta_i^m = g'(h_i^m) \sum_{j=1}^{q} w_{ji}^{m+1} \delta_j^{m+1}$$

We need to show that it is a proper steepest-descent step: $w_{ij}^m = w_{ij}^m + \epsilon(-\frac{\partial E}{\partial w_{ij}^m})$
Here E is the network error:

$$E = \sum_j (y_j - O_j)^2 = \sum_j (y_j - g(h_j^M))^2 \tag{1}$$

The tough part is the following theorem:

Theorem: In BP: $\delta_i^m = -\frac{1}{2}\frac{\partial E}{\partial h_i^m}$ for $m = M, M-1, \ldots, 1$

Proof: for $m = M$ we have: $\delta_j^M = g'(h_j^M)(y_j - O_j)$

$$\frac{\partial E}{\partial h_i^M} = -2(y_i - g(h_i^M))g'(h_i^M) = -2\delta_i^M$$

Now assume that the theorem holds for $m+1$: $\frac{\partial E}{\partial h_i^{m+1}} = -2\delta_i^{m+1}$
Prove that the theorem holds for m: $\frac{\partial E}{\partial h_i^m} = -2\delta_i^m$
Observe that the values of h_j^{m+1} for all j completely determines E. Therefore, applying the chain rule:

$$\frac{\partial E}{\partial h_i^m} = \sum_j \frac{\partial E}{\partial h_j^{m+1}} \frac{\partial h_j^{m+1}}{\partial h_i^m}$$

But $h_j^{m+1} = \sum_i w_{ji}^{m+1} g(h_i^m)$ and therefore: $\frac{\partial h_j^{m+1}}{\partial h_i^m} = w_{ji}^{m+1} g'(h_i^m)$. Combining this with the inductive assumption we have:

$$\frac{\partial E}{\partial h_i^m} = \sum_j -2\delta_j^{m+1} w_{ji}^{m+1} g'(h_i^m) = -2g'(h_i^m)\sum_j w_{ji}^{m+1}\delta_j^{m+1} = -2\delta_i^m$$

end of proof
The other part that we need for the proof that BP is proper steepest descent is:

$$\frac{\partial E}{\partial w_{ij}^m} = \frac{\partial E}{\partial h_i^m} V_j^{m-1} \tag{2}$$

This follows from the fact that $h_i^m = \sum_j w_{ij}^m V_j^{m-1}$ so that $\frac{\partial h_i^m}{\partial w_{ij}^m} = V_j^{m-1}$. Now:

$$\frac{\partial E}{\partial w_{ij}^m} = \frac{\partial E}{\partial h_i^m} \frac{\partial h_i^m}{\partial w_{ij}^m} = \frac{\partial E}{\partial h_i^m} V_j^{m-1}$$

BP is exactly steepest descent if: $\delta_i^m V_j^{m-1} = -\frac{\partial E}{w_{ij}^m}$ up to a positive constant. Indeed, combining the theorem and the result in (2) we have:

$$\frac{\partial E}{\partial w_{ij}^m} = \frac{\partial E}{\partial h_i^m} V_j^{m-1} = -2\delta_i^m V_j^{m-1}$$