

Stochastic Gradient Descent Rule

Update w_{ji} , one training example d at a time, with

$$\Delta w_{ji} = -\eta \frac{\partial E_d(\vec{w})}{\partial w_{ji}}$$

where

$$E_d(\vec{w}) = \frac{1}{2} \sum_{k \in \text{outputs}} (t_k - o_k)^2$$

is error summed over all output units

If $net_j = \sum_i w_{ji} x_{ji}$ is weighted sum of inputs for j

$$\begin{aligned} \frac{\partial E_d}{\partial w_{ji}} &= \frac{\partial E_d}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} \\ &= \frac{\partial E_d}{\partial net_j} x_{ji} \quad (WHY?) \end{aligned}$$

Output Units

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \frac{\partial o_j}{\partial net_j}$$

$$\begin{aligned}\frac{\partial E_d}{\partial o_j} &= \frac{\partial}{\partial o_j} \frac{1}{2} \sum_{k \in \text{outputs}} (t_k - o_k)^2 \\ &= \frac{\partial}{\partial o_j} \frac{1}{2} (t_j - o_j)^2 \\ &= -(t_j - o_j)\end{aligned}$$

$$\begin{aligned}\frac{\partial o_j}{\partial net_j} &= \frac{\partial \sigma(net_j)}{\partial net_j} \\ &= o_j(1 - o_j)\end{aligned}$$

$$\frac{\partial E_d}{\partial net_j} = -(t_j - o_j)o_j(1 - o_j)$$

Hidden Units

$$\begin{aligned}\frac{\partial E_d}{\partial net_j} &= \sum_{k \in \text{Downstream}(j)} \frac{\partial E_d}{\partial net_k} \frac{\partial net_k}{\partial net_j} \\ &= \sum_{k \in \text{Downstream}(j)} -\delta_k \frac{\partial net_k}{\partial net_j} \\ &= \sum_{k \in \text{Downstream}(j)} -\delta_k \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j} \\ &= \sum_{k \in \text{Downstream}(j)} -\delta_k w_{kj} \frac{\partial o_j}{\partial net_j} \\ &= \sum_{k \in \text{Downstream}(j)} -\delta_k w_{kj} o_j (1 - o_j)\end{aligned}$$

Therefore

$$\begin{aligned}\delta_j &= o_j(1 - o_j) \sum_{k \in \text{Downstream}(j)} \delta_k w_{kj} \\ \Delta w_{ji} &= \eta \delta_j x_{ji}\end{aligned}$$