Stochastic GD

$$w_{ji}^{(m)}(k+1) = w_{ji}^{(m)}(k) - \alpha \boxed{\frac{\partial J}{\partial w_{ji}^{(m)}(k)}}$$

$$\frac{\partial J}{\partial w_{ji}^{(m)}(k)} = \frac{\partial J}{\partial n_i^{(m)}} \times \frac{\partial n_i^{(m)}}{\partial w_{ji}^{(m)}(k)}$$

$$\frac{\partial J}{\partial b_i^{(m)}(k)} = \frac{\partial J}{\partial n_i^{(m)}} \times \frac{\partial n_i^{(m)}}{\partial b_i^{(m)}(k)}$$

The second term in each of these equations can be easily computed, because

$$n_i^{(m)} = \sum_{j=1}^{S^{(m-1)}} w_{ji}^{(m)}(k) a_j^{(m-1)} + b_i^{(m)}(k)$$

$$\frac{\partial n_i^{(m)}}{\partial w_{ji}^{(m)}(k)} = a_j^{(m-1)}$$

$$\frac{\partial n_i^{(m)}}{\partial b_i^{(m)}} = 1$$

The most tricky part is

to calculate

$\dfrac{\partial J}{\partial n_i^{(m)}}$. Let

$$s_i^{(m)} = \frac{\partial J}{\partial n_i^{(m)}}$$

$$\underline{s}^{(m)} = \frac{\partial J}{\partial \underline{n}^{(m)}}$$

We call $s_i^{(m)}$ the sensitivity of J to changes to the ith element of $\underline{n}^{(m)}$;

therefore

$$\frac{\partial J}{\partial w_{ji}^{(m)}(k)} = s_i^{(m)} a_j^{(m-1)}$$

$$\frac{\partial J}{\partial b_i^{(m)}(k)} = s_i^{(m)}$$

and

$$w_{ji}^{(m)}(k+1) = w_{ji}^{(m)}(k) - \alpha s_i^{(m)} a_j^{(m-1)}$$

$$b_i^{(m)}(k+1) = b_i^{(m)}(k) - \alpha s_i^{(m)}$$

Backpropagating sensitivities

Using chain rule in matrix form

We calculate $\frac{\partial J}{\partial \underline{n}^{(m)}}$ based on $\frac{\partial J}{\underline{n}^{(m+1)}}$, because the only "easy" gradient is $\frac{\partial J}{\partial \underline{n}^{(M)}}$, i.e. at the last layer and we want to use it calculate $\dfrac{\partial J}{\partial \underline{n}^{(m)}}$ in hidden layers.

$$\frac{\partial J}{\partial n_i^{(m)}} = \sum_{j=1}^{S^{(m+1)}} \frac{\partial J}{\partial n_j^{(m+1)}} \frac{\partial n_j^{(m+1)}}{\partial n_i^{(m)}}$$

$$\left[\frac{\partial J}{\partial \underline{n}^{(m)}}\right]_i = \sum_{j=1}^{S^{(m+1)}} \left[\frac{\partial \underline{n}^{(m+1)}}{\partial \underline{n}^{(m)}}\right]_{ji} \left[\frac{\partial J}{\partial \underline{n}^{(m+1)}}\right]_j$$

$$\frac{\partial J}{\partial \underline{n}^{(m)}} = \left[\frac{\partial \underline{n}^{(m+1)}}{\partial \underline{n}^{(m)}}\right]^T \frac{\partial J}{\partial \underline{n}^{(m+1)}}$$

$$\underline{s}^{(m)} = \frac{\partial J}{\partial \underline{n}^{(m)}} = \left[\frac{\partial \underline{n}^{(m+1)}}{\partial \underline{n}^{(m)}}\right]^T \frac{\partial J}{\partial \underline{n}^{(m+1)}}$$

Now, we are going to calculate the Jacobian

$$\frac{\partial \underline{n}^{(m+1)}}{\partial \underline{n}^{(m)}} :$$

$$\frac{\partial n_i^{(m+1)}}{\partial n_j^{(m)}} = \frac{\partial\left(\sum_{l=1}^{S^{(m)}} w_{li}^{(m+1)} a_l^{(k)(m)} + b_i^{(m+1)}\right)}{\partial n_j^{(m)}}$$

$$= w_{ji}^{(m+1)}(k) \frac{\partial a_j^{(m)}}{\partial n_j^{(m)}}$$

$$= w_{ji}^{(m+1)}(k) \underbrace{\frac{\partial f^{(m)}(n_j^{(m)})}{\partial n_j^{(m)}}}_{\equiv \dot{f}^{(m)}(n_j^{(m)})}$$

So, the Jacobian matrix is

$$\frac{\partial \underline{n}^{(m+1)}}{\partial \underline{n}^{(m)}} = \left[W^{(m+1)}(k)\right]^T \dot{F}^{(m)}(\underline{n}^{(m)})$$

where $F^{(m)}(\underline{n}^{(m)})$ is defined as

$$\dot{F}^{(m)}(\underline{n}^{(m)}) = \begin{bmatrix} \dot{f}^{(m)}(n_1^{(m)}) & 0 & \cdots & 0 \\ 0 & \dot{f}^{(m)}(n_2^{(m)}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dot{f}^{(m)}(n_{S^{(m)}}^{(m)}) \end{bmatrix}$$

So, back to $\underline{s}^{(m)}$

$$\underline{s}^{(m)} = \left[\frac{\partial \underline{n}^{(m+1)}}{\partial \underline{n}^{(m)}}\right]^T \frac{\partial J}{\partial \underline{n}^{(m+1)}}$$

$$= \left\{ \left[W^{(m+1)}(k) \right]^T \dot{F}^{(m)}(\underline{n}^{(m)}) \right\}^T \underline{s}^{(m+1)}$$

$$= \dot{F}^{(m)}(\underline{n}^{(m)}) \, W^{(m+1)} \quad \underline{s}^{(m+1)}$$

Remember $s^{(M)} = \dfrac{\partial J}{\partial \underline{n}^{(M)}}$

$$= \frac{\partial (\underline{y}-\underline{a})^T (\underline{y}-\underline{a})}{\partial \underline{n}}$$

$$\frac{\partial J}{\partial n_i^{(M)}} = \frac{\partial \sum_{j=1}^{S^{(M)}} (y_j - a_j)^2}{\partial n_i^{(M)}}$$

$$= -2\,(y_i - a_i)\,\frac{\partial a_i}{\partial n_i^{(M)}}$$

$$= -2\,(y_i - a_i^{(M)})\underbrace{\left(\frac{\partial a_i^{(M)}}{\partial n_i^{(M)}}\right)}_{\dot{f}^{(M)}(n_i^{(M)})}$$

$$\Rightarrow \underline{S}^{(M)} = -2\,\dot{F}^{(M)}(\underline{n}^{(M)})\,(\underline{y} - \underline{a})$$

Summary :

$$w_{ji}^{(m)}(k+1) = w_{ji}^{(m)}(k) - \alpha \frac{\partial J}{\partial w_{ji}^{(m)}(k)}$$

$$b_i^{(m)}(k+1) = b_i^{(m)}(k) - \alpha \frac{\partial J}{\partial b_i^{(m)}(k)}$$

$$\frac{\partial J}{\partial w_{ji}(k)} = s_i^{(m)} \, a_j^{(m-1)}$$

$$\Rightarrow w_{ij}^{(m)}(k+1) = w_{ij}^{(m)}(k) - \alpha \frac{\partial J}{\partial w_{ij}^{(m)}(k)}$$

$$W^{(m)}(k+1) = W^{(m)}(k) - \alpha \, \underline{a}^{(m-1)} \, \underline{s}^{(m)T}$$

Update rule in matrix form

$$\frac{\partial J}{\partial b_i^{(m)}(k)} = \delta_i^{(m)}$$

$$\underline{b_i^{T(m)}}(k+1) = b_i^{T(m)}(k) - \alpha\, \delta^{T(m)}$$

Update rule for biases

Backpropagation for sensitivities:

$$\underline{s}^{(m)} = \dot{F}^{(m)}(n^{(m)})\, W^{(m+1)}(k)\, \underline{s}^{(m+1)}$$

$$\underline{s}^{(M)} = -2\,\dot{F}^{(M)}(\underline{n}^{(M)})(\underline{y}-\underline{a})$$

$$\underline{s}^{(M)} \rightarrow \underline{s}^{(M-1)} \rightarrow \cdots \rightarrow \underline{s}^{(1)}$$