Derivation of Batch Back Propagation

May 8, 2017

Variables:

- $X_{ij}, \quad 0 \le i < N, \quad 0 \le j < K^{(0)}$
  Inputs: N row of input samples with $K+1$ features. The first feature is always 1 for bias.
- $W_{ij}^{(l)}, \quad 0 \le l < L, \quad 0 \le i < K^{(l)}, \quad 0 \le j < K^{(l+1)}$
  Network Weights: Where l is the layer in the network. j is the number of neurons in that layer, i is number of inputs from previous layer. Note $K_0 = K+1$ from the input.
- $A^l = Z^l W^l, \quad 0 \le l < L$
  Activation of the l-th layer. Note that $Z^0 = X$
- $Z^l = f^{l-1}(A^{l-1}), \quad 1 \le l < L$
  Output for layer l.
- $f^l(A^l) = f^l(Z^l W^l)$
  Filter function for layer l
- $\tilde{Y}_{ij}, \quad 0 \le i < N, \quad 0 \le j < K^{(L)}$
  Note $Y = Z^L$
- $T_{ij}, \quad 0 \le i < N, \quad 0 \le j < K^{(L)}$
  Target value used to compared with Y
- $E(\tilde{Y}, T) = E(Z^{(L)})$
  Error function

Note that Z goes from $Z^{(0)} \ldots Z^{(L)}$, f goes from $f^{(0)} \ldots f^{(L-1)}$, A goes from $A^{(0)} \ldots A^{(L-1)}$ and W goes from $W^{(0)} \ldots W^{(L-1)}$

Derivation:

For each layer l by chain rule:

$$\frac{\partial E}{\partial W_{ij}^l} = \sum_{s,t} \frac{\partial E}{\partial A_{st}^l} \frac{\partial A_{st}^l}{\partial W_{ij}^l} \tag{1}$$

Since

$$A_{st}^l = \sum_k Z_{sk}^l W_{kt}^l \tag{2}$$

$$\frac{\partial A_{st}^l}{\partial W_{ij}^l} = Z_{sk}^l \delta_i^k \delta_j^t \tag{3}$$

$$= Z_{si}^l \delta_j^t \tag{4}$$

Substitute (4) into (1), we have

$$\frac{\partial E}{\partial W_{ij}^{l}} = \sum_{s,t} \frac{\partial E}{\partial A_{st}^{l}} Z_{si}^{l} \delta_j^t \tag{5}$$

$$= \sum_{s} \frac{\partial E}{\partial A_{sj}^{l}} Z_{si}^{l} \tag{6}$$

Also,

$$\frac{\partial E}{\partial A_{sj}^{l}} = \sum_{u,v} \frac{\partial E}{\partial A_{uv}^{l+1}} \frac{\partial A_{uv}^{l+1}}{\partial A_{sj}^{l}} \tag{7}$$

$$= \sum_{u,v} \frac{\partial E}{\partial A_{uv}^{l+1}} \frac{\partial}{\partial A_{sj}^{l}} \left(\sum_{p} \left[f^l(A^l) \right]_{up} W_{pv}^{l+1} \right) \tag{8}$$

$$= \sum_{u,v} \frac{\partial E}{\partial A_{uv}^{l+1}} \left(\sum_{p} \left[\partial_A f^l(A^l) \right]_{up} \delta_s^u \delta_j^p W_{pv}^{l+1} \right) \tag{9}$$

$$= \sum_{u,v} \frac{\partial E}{\partial A_{uv}^{l+1}} \left(\left[\partial_A f^l(A^l) \right]_{uj} \delta_s^u W_{jv}^{l+1} \right) \tag{10}$$

$$= \sum_{v} \frac{\partial E}{\partial A_{sv}^{l+1}} \left(\left[\partial_A f^l(A^l) \right]_{sj} W_{jv}^{l+1} \right) \tag{11}$$

In matrix form:

For base case, we have:

$$X = Z^0 \tag{12}$$

$$Y = Z^L = f^{L-1}(A^{L-1}) \tag{13}$$

$$E(Y) = E(Z^L) = E(f^{L-1}(A^{L-1})) \tag{14}$$

$$\frac{\partial E}{\partial A^{L-1}} = \frac{\partial E(Y)}{\partial Y} \partial_A f^{L-1}(A^{L-1}) \tag{15}$$

For recursive case ($0 \leq l \leq L-2$):

$$\frac{\partial E}{\partial W^l} = (Z^l)^T \frac{\partial E}{\partial A^l} \tag{16}$$

$$\frac{\partial E}{\partial A^l} = \frac{\partial E}{\partial A^{l+1}} \left(\partial_A f^l(A^l) W^{l+1} \right)^T \tag{17}$$

$$= \frac{\partial E}{\partial A^{l+1}} \left(\partial_A f^l(Z^l W^l) W^{l+1} \right)^T \tag{18}$$

To update W:

$$W^l \leftarrow W^l - \eta_l \frac{\partial E}{\partial W^l}$$