

5.20 The cross-entropy error function for the multi-class classification problem is given by 5.24:

$$E = - \sum_{n=1}^N \sum_{k=1}^K t_{kn} \ln y_k(\mathbf{x}_n, \mathbf{w})$$

$$\implies \nabla E = - \sum_{n=1}^N \sum_{k=1}^K \frac{t_{nk}}{y_{kn}} \nabla y_{kn}$$

where $y_{kn} = y_k(\mathbf{x}_n, \mathbf{w})$.

Since $y_{kn} = \frac{\exp(a_{kn})}{\sum_j \exp(a_{jn})}$, it is dependent on all a_{jn} s.

$$= - \sum_{n=1}^N \sum_{k=1}^K \frac{t_{kn}}{y_{kn}} \left(\begin{bmatrix} : & \frac{\partial a_{kn}}{\partial \mathbf{w}} & : \end{bmatrix} \right) \left(\begin{bmatrix} \cdots \\ \frac{\partial y_{kn}}{\partial a_{jn}} \\ \cdots \end{bmatrix} \right)$$

$$= - \sum_{n=1}^N \sum_{k=1}^K \frac{t_{kn}}{y_{kn}} \left(\sum_{j=1}^K \left(\frac{\partial y_{kn}}{\partial a_{jn}} \right) \left(\frac{\partial a_{jn}}{\partial \mathbf{w}} \right) \right)$$

Using 4.106, we know that:

$$\frac{\partial y_{kn}}{\partial a_{jn}} = y_{kn}(I_{kj} - y_{jn})$$

Also, let

$$\nabla a_{jn} = \frac{\partial a_{jn}}{\partial \mathbf{w}}$$

$$\implies \nabla E = - \sum_{n=1}^N \sum_{k=1}^K \frac{t_{kn}}{y_{kn}} \left(\sum_{j=1}^K y_{kn}(I_{kj} - y_{jn}) \nabla a_{jn} \right)$$

$$= - \sum_{n=1}^N \sum_{k=1}^K t_{kn} \left(\sum_{j=1}^K (I_{kj} - y_{jn}) \nabla a_{jn} \right)$$

$$\begin{aligned}
\Rightarrow \nabla \nabla E &= - \sum_{n=1}^N \sum_{k=1}^K t_{kn} \left(\sum_{j=1}^K \nabla (I_{kj} - y_{jn}) \nabla a_{jn} + \sum_{j=1}^K (I_{kj} - y_{jn}) \nabla \nabla a_{jn} \right) \\
&= - \sum_{n=1}^N \sum_{k=1}^K t_{kn} \left(- \sum_{j=1}^K \nabla y_{jn} \nabla a_{jn} + \sum_{j=1}^K (I_{kj} - y_{jn}) \nabla \nabla a_{jn} \right)
\end{aligned}$$

Neglecting the second term as per the Levenberg–Marquardt approximation, and using the result for ∇y_{jn} as obtained above, we get:

$$\mathbf{H} = \nabla \nabla E \simeq \sum_{n=1}^N \sum_{k=1}^K t_{kn} \left(\sum_{j=1}^K \left(\sum_{l=1}^K y_{jn} (I_{jl} - y_{ln}) \nabla a_{ln} \right) \nabla a_{jn} \right)$$

Since k is not connected to any of the rest of the expression, we can write this as:

$$\mathbf{H} = \nabla \nabla E \simeq \sum_{n=1}^N \left(\sum_{k=1}^K t_{kn} \right) \left(\sum_{j=1}^K \left(\sum_{l=1}^K y_{jn} (I_{jl} - y_{ln}) \nabla a_{ln} \right) \nabla a_{jn} \right)$$

Due to one-hot-encoding, $\sum_{k=1}^K t_{kn} = 1$, so the expression becomes:

$$\mathbf{H} \simeq \sum_{n=1}^N \sum_{j=1}^K \sum_{l=1}^K y_{jn} (I_{jl} - y_{ln}) \nabla a_{ln} \nabla a_{jn}$$