5.3 For a K-output neural network where the conditional distribution of the target values is given by 5.192:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \mathcal{N}\left(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{W}), \mathbf{\Sigma}\right)$$

The likelihood function is given by:

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{t}_i|\mathbf{y}(\mathbf{x}_i, \mathbf{W}), \mathbf{\Sigma}\right)$$

The log-likelihood is given by:

$$\ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \sum_{i=1}^{N} \ln \mathcal{N}\left(\mathbf{t}_i|\mathbf{y}(\mathbf{x}_i, \mathbf{W}), \mathbf{\Sigma}\right)$$

$$= \sum_{i=1}^{N} \ln\left(\frac{1}{(2\pi)^{K/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))\right\}\right)$$

$$= \sum_{i=1}^{N}\left(-\frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))\right)$$

$$= -\frac{NK}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))$$

Maximizing this expression w.r.t $\mathbf{W}$ gives us:

$$\begin{aligned}
\mathbf{W}_{ML} &= \arg\max_{\mathbf{W}}\left(-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))\right) \\
&= \arg\min_{\mathbf{W}}\left(\sum_{i=1}^{N}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))\right) \\
\implies E(\mathbf{W}) &= \sum_{i=1}^{N}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1}(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))
\end{aligned}$$

Assuming that $\mathbf{\Sigma}$ is not known, if we maximize log-likelihood w.r.t $\mathbf{\Sigma}$ by taking the derivative, we get:

$$\frac{\partial \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W})}{\partial \mathbf{\Sigma}} = \frac{\partial}{\partial \mathbf{\Sigma}} \left(-\frac{NK}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{\Sigma}| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \mathbf{\Sigma}^{-1} (\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W})) \right)$$

The first terms becomes zero. For the second term, we use result 57 from The Matrix Cookbook, and for the third term, we use result 61 from The Matrix Cookbook, giving us:

$$= -\frac{N}{2} \mathbf{\Sigma}^{-1} + \frac{1}{2} \mathbf{\Sigma}^{-1} \left(\sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \right) \mathbf{\Sigma}^{-1}$$

Setting the derivative to $\mathbf{0}$, we get:

$$\frac{N}{2} \mathbf{\Sigma}^{-1} = \frac{1}{2} \mathbf{\Sigma}^{-1} \left(\sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T \right) \mathbf{\Sigma}^{-1}$$

$$\implies \mathbf{\Sigma}_{ML} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))(\mathbf{t}_i - \mathbf{y}(\mathbf{x}_i, \mathbf{W}))^T$$