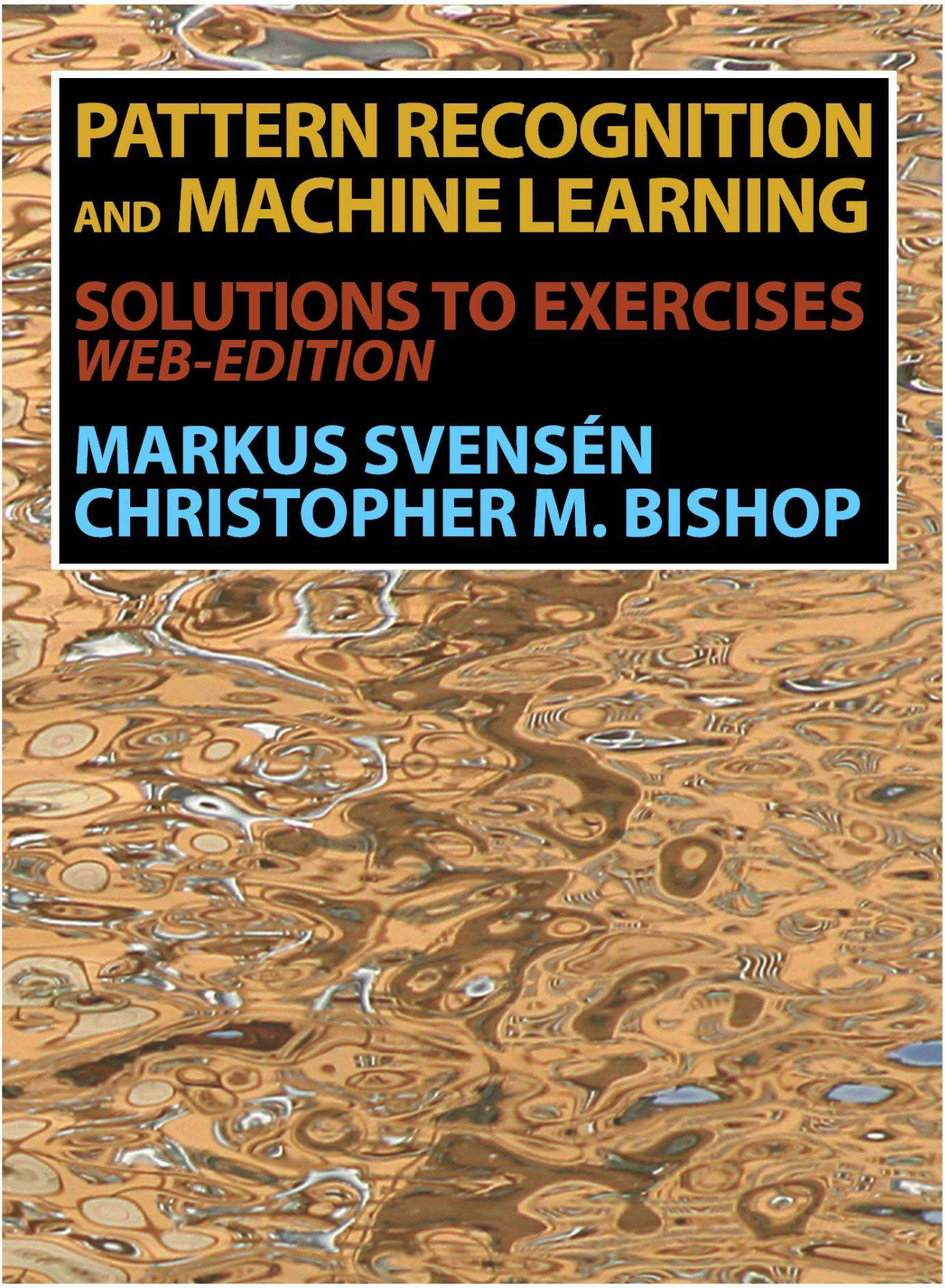

PATTERN RECOGNITION AND MACHINE LEARNING

SOLUTIONS TO EXERCISES
WEB-EDITION

MARKUS SVENSÉN
CHRISTOPHER M. BISHOP

Pattern Recognition and Machine Learning

Solutions to the Exercises: Web-Edition

Markus Svensén and Christopher M. Bishop

This is the solutions manual (web-edition) for the book *Pattern Recognition and Machine Learning* (PRML; published by Springer in 2006). It contains solutions to the www exercises. This release was created September 8, 2009. Future releases with corrections to errors will be published on the PRML web-site (see below).

The authors would like to express their gratitude to the various people who have provided feedback on earlier releases of this document. In particular, the "Bishop Reading Group", held in the Visual Geometry Group at the University of Oxford provided valuable comments and suggestions.

The authors welcome all comments, questions and suggestions about the solutions as well as reports on (potential) errors in text or formulae in this document; please send any such feedback to

`prml-fb@microsoft.com`

Further information about PRML is available from

http://research.microsoft.com/∼cmbishop/PRML

we obtain $\lambda = -N$. Eliminating λ then gives our final result for the maximum likelihood solution for h_k in the form

$$h_k = \frac{n_k}{N}\frac{1}{\Delta_k}.$$

Note that, for equal sized bins $\Delta_k = \Delta$ we obtain a bin height h_k which is proportional to the fraction of points falling within that bin, as expected.

Chapter 3 Linear Models for Regression

3.1 **NOTE**: In the 1st printing of PRML, there is a 2 missing in the denominator of the argument to the 'tanh' function in equation (3.102).

Using (3.6), we have

$$\begin{aligned} 2\sigma(2a) - 1 &= \frac{2}{1+e^{-2a}} - 1 \\ &= \frac{2}{1+e^{-2a}} - \frac{1+e^{-2a}}{1+e^{-2a}} \\ &= \frac{1-e^{-2a}}{1+e^{-2a}} \\ &= \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}} \\ &= \tanh(a) \end{aligned}$$

If we now take $a_j = (x - \mu_j)/2s$, we can rewrite (3.101) as

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}\left(2\sigma(2a_j) - 1 + 1\right) \\ &= u_0 + \sum_{j=1}^{M} u_j \tanh(a_j), \end{aligned}$$

where $u_j = w_j/2$, for $j = 1, \ldots, M$, and $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

3.4 Let

$$\begin{aligned}\widetilde{y}_n &= w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) \\ &= y_n + \sum_{i=1}^{D} w_i \epsilon_{ni}\end{aligned}$$

where $y_n = y(x_n, \mathbf{w})$ and $\epsilon_{ni} \sim \mathcal{N}(0, \sigma^2)$ and we have used (3.105). From (3.106) we then define

$$\begin{aligned}\widetilde{E} &= \frac{1}{2}\sum_{n=1}^{N} \{\widetilde{y}_n - t_n\}^2 \\ &= \frac{1}{2}\sum_{n=1}^{N} \{\widetilde{y}_n^2 - 2\widetilde{y}_n t_n + t_n^2\} \\ &= \frac{1}{2}\sum_{n=1}^{N} \left\{ y_n^2 + 2y_n \sum_{i=1}^{D} w_i\epsilon_{ni} + \left(\sum_{i=1}^{D} w_i \epsilon_{ni}\right)^2 \right. \\ &\qquad \left. -2t_n y_n - 2t_n \sum_{i=1}^{D} w_i \epsilon_{ni} + t_n^2 \right\}.\end{aligned}$$

If we take the expectation of $\widetilde{E}$ under the distribution of ϵ_{ni}, we see that the second and fifth terms disappear, since $\mathbb{E}[\epsilon_{ni}] = 0$, while for the third term we get

$$\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i \epsilon_{ni}\right)^2\right] = \sum_{i=1}^{D} w_i^2 \sigma^2$$

since the ϵ_{ni} are all independent with variance σ^2.

From this and (3.106) we see that

$$\mathbb{E}\left[\widetilde{E}\right] = E_D + \frac{1}{2}\sum_{i=1}^{D} w_i^2 \sigma^2,$$

as required.

3.5 We can rewrite (3.30) as

$$\frac{1}{2}\left(\sum_{j=1}^{M} |w_j|^q - \eta\right) \leqslant 0$$

where we have incorporated the $1/2$ scaling factor for convenience. Clearly this does not affect the constraint.

Employing the technique described in Appendix E, we can combine this with (3.12) to obtain the Lagrangian function

$$L(\mathbf{w}, \lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right)$$

and by comparing this with (3.29) we see immediately that they are identical in their dependence on $\mathbf{w}$.

Now suppose we choose a specific value of $\lambda > 0$ and minimize (3.29). Denoting the resulting value of $\mathbf{w}$ by $\mathbf{w}^\star(\lambda)$, and using the KKT condition (E.11), we see that the value of η is given by

$$\eta = \sum_{j=1}^{M}|w_j^\star(\lambda)|^q.$$

3.6 We first write down the log likelihood function which is given by

$$\ln L(\mathbf{W}, \boldsymbol{\Sigma}) = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)).$$

First of all we set the derivative with respect to $\mathbf{W}$ equal to zero, giving

$$0 = -\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}.$$

Multiplying through by $\boldsymbol{\Sigma}$ and introducing the design matrix $\boldsymbol{\Phi}$ and the target data matrix $\mathbf{T}$ we have

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{W} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{T}$$

Solving for $\mathbf{W}$ then gives (3.15) as required.

The maximum likelihood solution for $\boldsymbol{\Sigma}$ is easily found by appealing to the standard result from Chapter 2 giving

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))^{\mathrm{T}}.$$

as required. Since we are finding a joint maximum with respect to both $\mathbf{W}$ and $\boldsymbol{\Sigma}$ we see that it is $\mathbf{W}_{\mathrm{ML}}$ which appears in this expression, as in the standard result for an unconditional Gaussian distribution.

3.8 Combining the prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

and the likelihood

$$p(t_{N+1}|\mathbf{x}_{N+1},\mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left(-\frac{\beta}{2}(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right) \tag{94}$$

where $\boldsymbol{\phi}_{N+1} = \boldsymbol{\phi}(\mathbf{x}_{N+1})$, we obtain a posterior of the form

$$p(\mathbf{w}|t_{N+1},\mathbf{x}_{N+1},\mathbf{m}_N,\mathbf{S}_N) \propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) - \frac{1}{2}\beta(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right).$$

We can expand the argument of the exponential, omitting the $-1/2$ factors, as follows

$$\begin{aligned}
&(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) + \beta(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2 \\
&\quad = \mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N \\
&\qquad + \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}\mathbf{w} - 2\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}t_{N+1} + \mathrm{const} \\
&\quad = \mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}})\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta\boldsymbol{\phi}_{N+1}t_{N+1}) + \mathrm{const},
\end{aligned}$$

where const denotes remaining terms independent of $\mathbf{w}$. From this we can read off the desired result directly,

$$p(\mathbf{w}|t_{N+1},\mathbf{x}_{N+1},\mathbf{m}_N,\mathbf{S}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1},\mathbf{S}_{N+1}),$$

with

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}. \tag{95}$$

and

$$\mathbf{m}_{N+1} = \mathbf{S}_{N+1}(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta\boldsymbol{\phi}_{N+1}t_{N+1}). \tag{96}$$

3.10 Using (3.3), (3.8) and (3.49), we can re-write (3.57) as

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta) = \int \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w},\beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)\,\mathrm{d}\mathbf{w}.$$

By matching the first factor of the integrand with (2.114) and the second factor with (2.113), we obtain the desired result directly from (2.115).

3.15 This is easily shown by substituting the re-estimation formulae (3.92) and (3.95) into (3.82), giving

$$\begin{aligned}
E(\mathbf{m}_N) &= \frac{\beta}{2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N \\
&= \frac{N-\gamma}{2} + \frac{\gamma}{2} = \frac{N}{2}\,.
\end{aligned}$$

3.18 We can rewrite (3.79)

$$\begin{aligned}
&\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi w}\|^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\\
&\quad=\quad\frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi w}\right)+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\\
&\quad=\quad\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}\right)
\end{aligned}$$

where, in the last line, we have used (3.81). We now use the tricks of adding $\mathbf{0}=\mathbf{m}_{\mathbf{N}}^{\mathrm{T}}\mathbf{A}\mathbf{m}_{\mathbf{N}}-\mathbf{m}_{\mathbf{N}}^{\mathrm{T}}\mathbf{A}\mathbf{m}_{\mathbf{N}}$ and using $\mathbf{I}=\mathbf{A}^{-1}\mathbf{A}$, combined with (3.84), as follows:

$$\begin{aligned}
&\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}\right)\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\beta\mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}\right)\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A w}+\mathbf{w}^{\mathrm{T}}\mathbf{A w}+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)\\
&\quad=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)+\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N).
\end{aligned}$$

Here the last term equals term the last term of (3.80) and so it remains to show that the first term equals the r.h.s. of (3.82). To do this, we use the same tricks again:

$$\begin{aligned}
\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)&=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N+\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{m}_N\right)\\
&=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{A}\mathbf{A}^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta+\mathbf{m}_N^{\mathrm{T}}\left(\alpha\mathbf{I}+\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\mathbf{m}_N\right)\\
&=\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t}-2\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\beta+\beta\mathbf{m}_N^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{m}_N+\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right)\\
&=\frac{1}{2}\left(\beta(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)+\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right)\\
&=\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N\|^2+\frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N
\end{aligned}$$

as required.

3.20 We only need to consider the terms of (3.86) that depend on α, which are the first, third and fourth terms.

Following the sequence of steps in Section 3.5.2, we start with the last of these terms,

$$-\frac{1}{2}\ln|\mathbf{A}|.$$

From (3.81), (3.87) and the fact that that eigenvectors $\mathbf{u}_i$ are orthonormal (see also Appendix C), we find that the eigenvectors of $\mathbf{A}$ to be $\alpha+\lambda_i$. We can then use (C.47) and the properties of the logarithm to take us from the left to the right side of (3.88).

The derivatives for the first and third term of (3.86) are more easily obtained using standard derivatives and (3.82), yielding

$$\frac{1}{2}\left(\frac{M}{\alpha} + \mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right).$$

We combine these results into (3.89), from which we get (3.92) via (3.90). The expression for γ in (3.91) is obtained from (3.90) by substituting

$$\sum_i^M \frac{\lambda_i + \alpha}{\lambda_i + \alpha}$$

for M and re-arranging.

3.23 From (3.10), (3.112) and the properties of the Gaussian and Gamma distributions (see Appendix B), we get

$$\begin{aligned}
p(\mathbf{t}) &= \iint p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\beta)\,\mathrm{d}\mathbf{w}p(\beta)\,\mathrm{d}\beta \\
&= \iint \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\} \\
&\qquad \left(\frac{\beta}{2\pi}\right)^{M/2} |\mathbf{S}_0|^{-1/2} \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} \exp(-b_0\beta)\,\mathrm{d}\beta \\
&= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\} \\
&\qquad \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \beta^{a_0-1}\beta^{N/2}\beta^{M/2} \exp(-b_0\beta)\,\mathrm{d}\beta \\
&= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\}\,\mathrm{d}\mathbf{w} \\
&\qquad \exp\left\{-\frac{\beta}{2}\left(\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\right)\right\} \\
&\qquad \beta^{a_N-1}\beta^{M/2} \exp(-b_0\beta)\,\mathrm{d}\beta
\end{aligned}$$

where we have completed the square for the quadratic form in $\mathbf{w}$, using

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left[\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right] \\
\mathbf{S}_N^{-1} &= \beta\left(\mathbf{S}_0^{-1} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right) \\
a_N &= a_0 + \frac{N}{2} \\
b_N &= b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \sum_{n=1}^{N} t_n^2\right).
\end{aligned}$$

Now we are ready to do the integration, first over $\mathbf{w}$ and then β, and re-arrange the terms to obtain the desired result

$$\begin{aligned}
p(\mathbf{t}) &= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\mathbf{S}_0|\right)^{1/2}}(2\pi)^{M/2}|\mathbf{S}_N|^{1/2}\int \beta^{a_N-1}\exp(-b_N\beta)\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{N/2}}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}.
\end{aligned}$$

Chapter 4 Linear Models for Classification

4.2 For the purpose of this exercise, we make the contribution of the bias weights explicit in (4.15), giving

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\right\}, \tag{97}$$

where $\mathbf{w}_0$ is the column vector of bias weights (the top row of $\widetilde{\mathbf{W}}$ transposed) and $\mathbf{1}$ is a column vector of N ones.

We can take the derivative of (97) w.r.t. $\mathbf{w}_0$, giving

$$2N\mathbf{w}_0 + 2(\mathbf{X}\mathbf{W} - \mathbf{T})^{\mathrm{T}}\mathbf{1}.$$

Setting this to zero, and solving for $\mathbf{w}_0$, we obtain

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \tag{98}$$

where

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^{\mathrm{T}}\mathbf{1} \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{1}.$$

If we subsitute (98) into (97), we get

$$E_D(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})\right\},$$

where

$$\overline{\mathbf{T}} = \mathbf{1}\bar{\mathbf{t}}^{\mathrm{T}} \quad \text{and} \quad \overline{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}.$$

Setting the derivative of this w.r.t. $\mathbf{W}$ to zero we get

$$\mathbf{W} = (\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{T}} = \widehat{\mathbf{X}}^{\dagger}\widehat{\mathbf{T}},$$

where we have defined $\widehat{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$ and $\widehat{\mathbf{T}} = \mathbf{T} - \overline{\mathbf{T}}$.
Now consider the prediction for a new input vector $\mathbf{x}^{\star}$,

$$\begin{aligned} \mathbf{y}(\mathbf{x}^{\star}) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^{\star} + \mathbf{w}_0 \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^{\star} + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\ &= \bar{\mathbf{t}} - \widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^{\star} - \bar{\mathbf{x}}). \end{aligned} \tag{99}$$

If we apply (4.157) to $\bar{\mathbf{t}}$, we get

$$\mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{1} = -b.$$

Therefore, applying (4.157) to (99), we obtain

$$\begin{aligned} \mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}^{\star}) &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} + \mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^{\star} - \bar{\mathbf{x}}) \\ &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = -b, \end{aligned}$$

since $\mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = b(\mathbf{1} - \mathbf{1})^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$.

4.4 **NOTE**: In the 1st printing of PRML, the text of the exercise refers equation (4.23) where it should refer to (4.22).

From (4.22) we can construct the Lagrangian function

$$L = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1) + \lambda\left(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1\right).$$

Taking the gradient of L we obtain

$$\nabla L = \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} \tag{100}$$

and setting this gradient to zero gives

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m}_2 - \mathbf{m}_1)$$

form which it follows that $\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1$.

4.7 From (4.59) we have

$$\begin{aligned} 1 - \sigma(a) &= 1 - \frac{1}{1 + e^{-a}} = \frac{1 + e^{-a} - 1}{1 + e^{-a}} \\ &= \frac{e^{-a}}{1 + e^{-a}} = \frac{1}{e^{a} + 1} = \sigma(-a). \end{aligned}$$

The inverse of the logistic sigmoid is easily found as follows

$$\begin{aligned} y = \sigma(a) &= \frac{1}{1+e^{-a}} \\ \Rightarrow \quad \frac{1}{y} - 1 &= e^{-a} \\ \Rightarrow \quad \ln\left\{\frac{1-y}{y}\right\} &= -a \\ \Rightarrow \quad \ln\left\{\frac{y}{1-y}\right\} &= a = \sigma^{-1}(y). \end{aligned}$$

4.9 The likelihood function is given by

$$p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left\{p(\phi_n|\mathcal{C}_k)\pi_k\right\}^{t_{nk}}$$

and taking the logarithm, we obtain

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln p(\phi_n|\mathcal{C}_k) + \ln \pi_k\right\}. \tag{101}$$

In order to maximize the log likelihood with respect to π_k we need to preserve the constraint $\sum_k \pi_k = 1$. This can be done by introducing a Lagrange multiplier λ and maximizing

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

Setting the derivative with respect to π_k equal to zero, we obtain

$$\sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda = 0.$$

Re-arranging then gives

$$-\pi_k\lambda = \sum_{n=1}^{N} t_{nk} = N_k. \tag{102}$$

Summing both sides over k we find that $\lambda = -N$, and using this to eliminate λ we obtain (4.159).

4.12 Differentiating (4.59) we obtain

$$\begin{aligned}\frac{d\sigma}{da} &= \frac{e^{-a}}{(1+e^{-a})^2} \\ &= \sigma(a)\left\{\frac{e^{-a}}{1+e^{-a}}\right\} \\ &= \sigma(a)\left\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right\} \\ &= \sigma(a)(1-\sigma(a)).\end{aligned}$$

4.13 We start by computing the derivative of (4.90) w.r.t. y_n

$$\frac{\partial E}{\partial y_n} = \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \tag{103}$$

$$= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)}$$

$$= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \tag{104}$$

$$= \frac{y_n - t_n}{y_n(1-y_n)}. \tag{105}$$

From (4.88), we see that

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)\left(1-\sigma(a_n)\right) = y_n(1-y_n). \tag{106}$$

Finally, we have

$$\nabla a_n = \boldsymbol{\phi}_n \tag{107}$$

where ∇ denotes the gradient with respect to $\mathbf{w}$. Combining (105), (106) and (107) using the chain rule, we obtain

$$\begin{aligned}\nabla E &= \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n \\ &= \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n\end{aligned}$$

as required.

4.17 From (4.104) we have

$$\begin{aligned}\frac{\partial y_k}{\partial a_k} &= \frac{e^{a_k}}{\sum_i e^{a_i}} - \left(\frac{e^{a_k}}{\sum_i e^{a_i}}\right)^2 = y_k(1-y_k), \\ \frac{\partial y_k}{\partial a_j} &= -\frac{e^{a_k}e^{a_j}}{\left(\sum_i e^{a_i}\right)^2} = -y_k y_j, \qquad j \neq k.\end{aligned}$$

Combining these results we obtain (4.106).

4.19 Using the cross-entropy error function (4.90), and following Exercise 4.13, we have

$$\frac{\partial E}{\partial y_n} = \frac{y_n - t_n}{y_n(1-y_n)}. \tag{108}$$

Also

$$\nabla a_n = \boldsymbol{\phi}_n. \tag{109}$$

From (4.115) and (4.116) we have

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \Phi(a_n)}{\partial a_n} = \frac{1}{\sqrt{2\pi}} e^{-a_n^2}. \tag{110}$$

Combining (108), (109) and (110), we get

$$\nabla E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1-y_n)}\frac{1}{\sqrt{2\pi}} e^{-a_n^2}\boldsymbol{\phi}_n. \tag{111}$$

In order to find the expression for the Hessian, it is is convenient to first determine

$$\begin{aligned}\frac{\partial}{\partial y_n}\frac{y_n - t_n}{y_n(1-y_n)} &= \frac{y_n(1-y_n)}{y_n^2(1-y_n)^2} - \frac{(y_n - t_n)(1-2y_n)}{y_n^2(1-y_n)^2} \\ &= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1-y_n)^2}.\end{aligned} \tag{112}$$

Then using (109)–(112) we have

$$\begin{aligned}\nabla\nabla E &= \sum_{n=1}^{N}\left\{\frac{\partial}{\partial y_n}\left[\frac{y_n - t_n}{y_n(1-y_n)}\right]\frac{1}{\sqrt{2\pi}}e^{-a_n^2}\boldsymbol{\phi}_n\nabla y_n \right. \\ &\qquad \left. + \frac{y_n - t_n}{y_n(1-y_n)}\frac{1}{\sqrt{2\pi}}e^{-a_n^2}(-2a_n)\boldsymbol{\phi}_n\nabla a_n\right\} \\ &= \sum_{n=1}^{N}\left(\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)}\frac{1}{\sqrt{2\pi}}e^{-a_n^2} - 2a_n(y_n - t_n)\right)\frac{e^{-2a_n^2}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}}{\sqrt{2\pi}y_n(1-y_n)}.\end{aligned}$$

4.23 **NOTE**: In the 1st printing of PRML, the text of the exercise contains a typographical error. Following the equation, it should say that $\mathbf{H}$ is the matrix of second derivatives of the *negative* log likelihood.

The BIC approximation can be viewed as a large N approximation to the log model evidence. From (4.138), we have

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) \\ &= \mathbf{H} - \nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}})\end{aligned}$$

and if $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0)$, this becomes

$$\mathbf{A} = \mathbf{H} + \mathbf{V}_0^{-1}.$$

If we assume that the prior is broad, or equivalently that the number of data points is large, we can neglect the term $\mathbf{V}_0^{-1}$ compared to $\mathbf{H}$. Using this result, (4.137) can be rewritten in the form

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \tag{113}$$

as required. Note that the phrasing of the question is misleading, since the assumption of a broad prior, or of large N, is required in order to derive this form, as well as in the subsequent simplification.

We now again invoke the broad prior assumption, allowing us to neglect the second term on the right hand side of (113) relative to the first term.

Since we assume i.i.d. data, $\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})$ consists of a sum of terms, one term for each datum, and we can consider the following approximation:

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{H}_n = N\widehat{\mathbf{H}}$$

where $\mathbf{H}_n$ is the contribution from the n^{th} data point and

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}_n.$$

Combining this with the properties of the determinant, we have

$$\ln|\mathbf{H}| = \ln|N\widehat{\mathbf{H}}| = \ln\left(N^M|\widehat{\mathbf{H}}|\right) = M\ln N + \ln|\widehat{\mathbf{H}}|$$

where M is the dimensionality of $\boldsymbol{\theta}$. Note that we are assuming that $\widehat{\mathbf{H}}$ has full rank M. Finally, using this result together (113), we obtain (4.139) by dropping the $\ln|\widehat{\mathbf{H}}|$ since this $O(1)$ compared to $\ln N$.

Chapter 5 Neural Networks

5.2 The likelihood function for an i.i.d. data set, $\{(\mathbf{x}_1, \mathbf{t}_1), \ldots, (\mathbf{x}_N, \mathbf{t}_N)\}$, under the conditional distribution (5.16) is given by

$$\prod_{n=1}^{N} \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}\right).$$

If we take the logarithm of this, using (2.43), we get

$$\begin{aligned}\sum_{n=1}^{N} \ln \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n,\mathbf{w}),\beta^{-1}\mathbf{I}\right)\\ = -\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{t}_n-\mathbf{y}(\mathbf{x}_n,\mathbf{w})\right)^{\mathrm{T}}(\beta\mathbf{I})\left(\mathbf{t}_n-\mathbf{y}(\mathbf{x}_n,\mathbf{w})\right)+\text{const}\\ = -\frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{t}_n-\mathbf{y}(\mathbf{x}_n,\mathbf{w})\|^2+\text{const},\end{aligned}$$

where 'const' comprises terms which are independent of $\mathbf{w}$. The first term on the right hand side is proportional to the negative of (5.11) and hence maximizing the log-likelihood is equivalent to minimizing the sum-of-squares error.

5.5 For the given interpretation of $y_k(\mathbf{x},\mathbf{w})$, the conditional distribution of the target vector for a multiclass neural network is

$$p(\mathbf{t}|\mathbf{w}_1,\ldots,\mathbf{w}_K)=\prod_{k=1}^{K}y_k^{t_k}.$$

Thus, for a data set of N points, the likelihood function will be

$$p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K}y_{nk}^{t_{nk}}.$$

Taking the negative logarithm in order to derive an error function we obtain (5.24) as required. Note that this is the same result as for the multiclass logistic regression model, given by (4.108) .

5.6 Differentiating (5.21) with respect to the activation a_n corresponding to a particular data point n, we obtain

$$\frac{\partial E}{\partial a_n}=-t_n\frac{1}{y_n}\frac{\partial y_n}{\partial a_n}+(1-t_n)\frac{1}{1-y_n}\frac{\partial y_n}{\partial a_n}. \tag{114}$$

From (4.88), we have

$$\frac{\partial y_n}{\partial a_n}=y_n(1-y_n). \tag{115}$$

Substituting (115) into (114), we get

$$\begin{aligned}\frac{\partial E}{\partial a_n} &= -t_n\frac{y_n(1-y_n)}{y_n}+(1-t_n)\frac{y_n(1-y_n)}{(1-y_n)}\\ &= y_n-t_n\end{aligned}$$

as required.

5.9 This simply corresponds to a scaling and shifting of the binary outputs, which directly gives the activation function, using the notation from (5.19), in the form

$$y = 2\sigma(a) - 1.$$

The corresponding error function can be constructed from (5.21) by applying the inverse transform to y_n and t_n, yielding

$$\begin{aligned} E(\mathbf{w}) &= -\sum_{n=1}^{N} \frac{1+t_n}{2} \ln \frac{1+y_n}{2} + \left(1 - \frac{1+t_n}{2}\right) \ln \left(1 - \frac{1+y_n}{2}\right) \\ &= -\frac{1}{2} \sum_{n=1}^{N} \{(1+t_n)\ln(1+y_n) + (1-t_n)\ln(1-y_n)\} + N \ln 2 \end{aligned}$$

where the last term can be dropped, since it is independent of $\mathbf{w}$.

To find the corresponding activation function we simply apply the linear transformation to the logistic sigmoid given by (5.19), which gives

$$\begin{aligned} y(a) &= 2\sigma(a) - 1 = \frac{2}{1+e^{-a}} - 1 \\ &= \frac{1-e^{-a}}{1+e^{-a}} = \frac{e^{a/2} - e^{-a/2}}{e^{a/2} + e^{-a/2}} \\ &= \tanh(a/2). \end{aligned}$$

5.10 From (5.33) and (5.35) we have

$$\mathbf{u}_i^{\mathrm{T}} \mathbf{H} \mathbf{u}_i = \mathbf{u}_i^{\mathrm{T}} \lambda_i \mathbf{u}_i = \lambda_i.$$

Assume that $\mathbf{H}$ is positive definite, so that (5.37) holds. Then by setting $\mathbf{v} = \mathbf{u}_i$ it follows that

$$\lambda_i = \mathbf{u}_i^{\mathrm{T}} \mathbf{H} \mathbf{u}_i > 0 \tag{116}$$

for all values of i. Thus, if $\mathbf{H}$ is positive definite, all of its eigenvalues will be positive.

Conversely, assume that (116) holds. Then, for any vector, $\mathbf{v}$, we can make use of (5.38) to give

$$\begin{aligned} \mathbf{v}^{\mathrm{T}} \mathbf{H} \mathbf{v} &= \left(\sum_i c_i \mathbf{u}_i\right)^{\mathrm{T}} \mathbf{H} \left(\sum_j c_j \mathbf{u}_j\right) \\ &= \left(\sum_i c_i \mathbf{u}_i\right)^{\mathrm{T}} \left(\sum_j \lambda_j c_j \mathbf{u}_j\right) \\ &= \sum_i \lambda_i c_i^2 > 0 \end{aligned}$$

where we have used (5.33) and (5.34) along with (116). Thus, if all of the eigenvalues are positive, the Hessian matrix will be positive definite.

5.11 **NOTE**: In PRML, Equation (5.32) contains a typographical error: $=$ should be $\simeq$.

We start by making the change of variable given by (5.35) which allows the error function to be written in the form (5.36). Setting the value of the error function $E(\mathbf{w})$ to a constant value C we obtain

$$E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2 = C.$$

Re-arranging gives

$$\sum_i \lambda_i \alpha_i^2 = 2C - 2E(\mathbf{w}^\star) = \widetilde{C}$$

where $\widetilde{C}$ is also a constant. This is the equation for an ellipse whose axes are aligned with the coordinates described by the variables $\{\alpha_i\}$. The length of axis j is found by setting $\alpha_i = 0$ for all $i \neq j$, and solving for α_j giving

$$\alpha_j = \left(\frac{\widetilde{C}}{\lambda_j}\right)^{1/2}$$

which is inversely proportional to the square root of the corresponding eigenvalue.

5.12 **NOTE**: See note in Solution 5.11.

From (5.37) we see that, if $\mathbf{H}$ is positive definite, then the second term in (5.32) will be positive whenever $(\mathbf{w} - \mathbf{w}^\star)$ is non-zero. Thus the smallest value which $E(\mathbf{w})$ can take is $E(\mathbf{w}^\star)$, and so $\mathbf{w}^\star$ is the minimum of $E(\mathbf{w})$.

Conversely, if $\mathbf{w}^\star$ is the minimum of $E(\mathbf{w})$, then, for any vector $\mathbf{w} \neq \mathbf{w}^\star$, $E(\mathbf{w}) > E(\mathbf{w}^\star)$. This will only be the case if the second term of (5.32) is positive for all values of $\mathbf{w} \neq \mathbf{w}^\star$ (since the first term is independent of $\mathbf{w}$). Since $\mathbf{w} - \mathbf{w}^\star$ can be set to any vector of real numbers, it follows from the definition (5.37) that $\mathbf{H}$ must be positive definite.

5.19 If we take the gradient of (5.21) with respect to $\mathbf{w}$, we obtain

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \frac{\partial E}{\partial a_n}\nabla a_n = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n,$$

where we have used the result proved earlier in the solution to Exercise 5.6. Taking the second derivatives we have

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\left\{\frac{\partial y_n}{\partial a_n}\nabla a_n \nabla a_n + (y_n - t_n)\nabla\nabla a_n\right\}.$$

Dropping the last term and using the result (4.88) for the derivative of the logistic sigmoid function, proved in the solution to Exercise 4.12, we finally get

$$\nabla\nabla E(\mathbf{w}) \simeq \sum_{n=1}^{N} y_n(1 - y_n)\nabla a_n \nabla a_n = \sum_{n=1}^{N} y_n(1 - y_n)\mathbf{b}_n \mathbf{b}_n^{\mathrm{T}}$$

where $\mathbf{b}_n \equiv \nabla a_n$.

5.25 The gradient of (5.195) is given

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star)$$

and hence update formula (5.196) becomes

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star).$$

Pre-multiplying both sides with $\mathbf{u}_j^{\mathrm{T}}$ we get

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau)} && (117)\\ &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_j^{\mathrm{T}}\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j\mathbf{u}_j^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j(w_j^{(\tau-1)} - w_j^\star), && (118) \end{aligned}$$

where we have used (5.198). To show that

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^\tau\}\, w_j^\star$$

for $\tau = 1, 2, \ldots$, we can use proof by induction. For $\tau = 1$, we recall that $\mathbf{w}^{(0)} = \mathbf{0}$ and insert this into (118), giving

$$\begin{aligned} w_j^{(1)} &= w_j^{(0)} - \rho\eta_j(w_j^{(0)} - w_j^\star) \\ &= \rho\eta_j w_j^\star \\ &= \{1 - (1 - \rho\eta_j)\}\, w_j^\star. \end{aligned}$$

Now we assume that the result holds for $\tau = N - 1$ and then make use of (118)

$$\begin{aligned} w_j^{(N)} &= w_j^{(N-1)} - \rho\eta_j(w_j^{(N-1)} - w_j^\star) \\ &= w_j^{(N-1)}(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \left\{1 - (1 - \rho\eta_j)^{N-1}\right\} w_j^\star(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \left\{(1 - \rho\eta_j) - (1 - \rho\eta_j)^N\right\} w_j^\star + \rho\eta_j w_j^\star \\ &= \left\{1 - (1 - \rho\eta_j)^N\right\} w_j^\star \end{aligned}$$

as required.

Provided that $|1 - \rho\eta_j| < 1$ then we have $(1 - \rho\eta_j)^\tau \to 0$ as $\tau \to \infty$, and hence $\left\{1 - (1 - \rho\eta_j)^N\right\} \to 1$ and $\mathbf{w}^{(\tau)} \to \mathbf{w}^\star$.

If τ is finite but $\eta_j \gg (\rho\tau)^{-1}$, τ must still be large, since $\eta_j\rho\tau \gg 1$, even though $|1 - \rho\eta_j| < 1$. If τ is large, it follows from the argument above that $w_j^{(\tau)} \simeq w_j^\star$.

If, on the other hand, $\eta_j \ll (\rho\tau)^{-1}$, this means that $\rho\eta_j$ must be small, since $\rho\eta_j\tau \ll 1$ and τ is an integer greater than or equal to one. If we expand,

$$(1-\rho\eta_j)^\tau = 1 - \tau\rho\eta_j + O(\rho\eta_j^2)$$

and insert this into (5.197), we get

$$\begin{aligned} |w_j^{(\tau)}| &= |\left\{1-(1-\rho\eta_j)^\tau\right\} w_j^\star| \\ &= |\left\{1-(1-\tau\rho\eta_j+O(\rho\eta_j^2))\right\} w_j^\star| \\ &\simeq \tau\rho\eta_j|w_j^\star| \ll |w_j^\star| \end{aligned}$$

Recall that in Section 3.5.3 we showed that when the regularization parameter (called α in that section) is much larger than one of the eigenvalues (called λ_j in that section) then the corresponding parameter value w_i will be close to zero. Conversely, when α is much smaller than λ_i then w_i will be close to its maximum likelihood value. Thus α is playing an analogous role to $\rho\tau$.

5.27 If $\mathbf{s}(\mathbf{x},\boldsymbol{\xi}) = \mathbf{x}+\boldsymbol{\xi}$, then

$$\frac{\partial s_k}{\partial \xi_i} = I_{ki}, \text{ i.e., } \frac{\partial \mathbf{s}}{\partial \boldsymbol{\xi}} = \mathbf{I},$$

and since the first order derivative is constant, there are no higher order derivatives. We now make use of this result to obtain the derivatives of y w.r.t. ξ_i:

$$\frac{\partial y}{\partial \xi_i} = \sum_k \frac{\partial y}{\partial s_k}\frac{\partial s_k}{\partial \xi_i} = \frac{\partial y}{\partial s_i} = b_i$$

$$\frac{\partial y}{\partial \xi_i \partial \xi_j} = \frac{\partial b_i}{\partial \xi_j} = \sum_k \frac{\partial b_i}{\partial s_k}\frac{\partial s_k}{\partial \xi_j} = \frac{\partial b_i}{\partial s_j} = B_{ij}$$

Using these results, we can write the expansion of $\widetilde{E}$ as follows:

$$\begin{aligned} \widetilde{E} &= \frac{1}{2}\iiint \{y(\mathbf{x})-t\}^2 p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \iiint \{y(\mathbf{x})-t\}\mathbf{b}^{\mathrm{T}}\boldsymbol{\xi} p(\boldsymbol{\xi})p(t|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &+ \frac{1}{2}\iiint \boldsymbol{\xi}^{\mathrm{T}}\left(\{y(\mathbf{x})-t\}\mathbf{B}+\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\boldsymbol{\xi} p(\boldsymbol{\xi})p(t|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \end{aligned}$$

The middle term will again disappear, since $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$ and thus we can write $\widetilde{E}$ on the form of (5.131) with

$$\Omega = \frac{1}{2}\iiint \boldsymbol{\xi}^{\mathrm{T}}\left(\{y(\mathbf{x})-t\}\mathbf{B}+\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\boldsymbol{\xi} p(\boldsymbol{\xi})p(t|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

Again the first term within the parenthesis vanishes to leading order in $\boldsymbol{\xi}$ and we are left with

$$\begin{aligned}\Omega &\simeq \frac{1}{2}\iint \boldsymbol{\xi}^{\mathrm{T}}\left(\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\boldsymbol{\xi} p(\boldsymbol{\xi})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\iint \mathrm{Trace}\left[\left(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\right)\left(\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\right] p(\boldsymbol{\xi})p(\mathbf{x})\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int \mathrm{Trace}\left[\mathbf{I}\left(\mathbf{b}\mathbf{b}^{\mathrm{T}}\right)\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int \mathbf{b}^{\mathrm{T}}\mathbf{b}p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \frac{1}{2}\int \|\nabla y(\mathbf{x})\|^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x},\end{aligned}$$

where we used the fact that $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}] = \mathbf{I}$.

5.28 The modifications only affect derivatives with respect to weights in the convolutional layer. The units within a feature map (indexed m) have different inputs, but all share a common weight vector, $\mathbf{w}^{(m)}$. Thus, errors $\delta^{(m)}$ from all units within a feature map will contribute to the derivatives of the corresponding weight vector. In this situation, (5.50) becomes

$$\frac{\partial E_n}{\partial w_i^{(m)}} = \sum_j \frac{\partial E_n}{\partial a_j^{(m)}}\frac{\partial a_j^{(m)}}{\partial w_i^{(m)}} = \sum_j \delta_j^{(m)} z_{ji}^{(m)}.$$

Here $a_j^{(m)}$ denotes the activation of the j^{th} unit in the m^{th} feature map, whereas $w_i^{(m)}$ denotes the i^{th} element of the corresponding feature vector and, finally, $z_{ji}^{(m)}$ denotes the i^{th} input for the j^{th} unit in the m^{th} feature map; the latter may be an actual input or the output of a preceding layer.

Note that $\delta_j^{(m)} = \partial E_n/\partial a_j^{(m)}$ will typically be computed recursively from the δs of the units in the following layer, using (5.55). If there are layer(s) preceding the convolutional layer, the standard backward propagation equations will apply; the weights in the convolutional layer can be treated as if they were independent parameters, for the purpose of computing the δs for the preceding layer's units.

5.29 This is easily verified by taking the derivative of (5.138), using (1.46) and standard derivatives, yielding

$$\frac{\partial \Omega}{\partial w_i} = \frac{1}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)}\sum_j \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\frac{(w_i-\mu_j)}{\sigma^2}.$$

Combining this with (5.139) and (5.140), we immediately obtain the second term of (5.141).

5.34 **NOTE**: In the 1^{st} printing of PRML, the l.h.s. of (5.154) should be replaced with $\gamma_{nk} = \gamma_k(\mathbf{t}_n|\mathbf{x}_n)$. Accordingly, in (5.155) and (5.156), γ_k should be replaced by γ_{nk} and in (5.156), t_l should be t_{nl}.

We start by using the chain rule to write

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{119}$$

Note that because of the coupling between outputs caused by the softmax activation function, the dependence on the activation of a single output unit involves all the output units.

For the first factor inside the sum on the r.h.s. of (119), standard derivatives applied to the n^{th} term of (5.153) gives

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{120}$$

For the for the second factor, we have from (4.106) that

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j (I_{jk} - \pi_k). \tag{121}$$

Combining (119), (120) and (121), we get

$$\begin{aligned} \frac{\partial E_n}{\partial a_k^\pi} &= -\sum_{j=1}^{K} \frac{\gamma_{nj}}{\pi_j} \pi_j (I_{jk} - \pi_k) \\ &= -\sum_{j=1}^{K} \gamma_{nj} (I_{jk} - \pi_k) = -\gamma_{nk} + \sum_{j=1}^{K} \gamma_{nj} \pi_k = \pi_k - \gamma_{nk}, \end{aligned}$$

where we have used the fact that, by (5.154), $\sum_{j=1}^{K} \gamma_{nj} = 1$ for all n.

5.39 Using (4.135), we can approximate (5.174) as

$$p(\mathcal{D}|\alpha, \beta) \simeq p(\mathcal{D}|\mathbf{w}_{\text{MAP}}, \beta) p(\mathbf{w}_{\text{MAP}}|\alpha) \int \exp\left\{ -\frac{1}{2} (\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\text{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_{\text{MAP}}) \right\} \mathrm{d}\mathbf{w},$$

where $\mathbf{A}$ is given by (5.166), as $p(\mathcal{D}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha)$ is proportional to $p(\mathbf{w}|\mathcal{D}, \alpha, \beta)$. Using (4.135), (5.162) and (5.163), we can rewrite this as

$$p(\mathcal{D}|\alpha, \beta) \simeq \prod_{n}^{N} \mathcal{N}(t_n | y(\mathbf{x}_n, \mathbf{w}_{\text{MAP}}), \beta^{-1}) \mathcal{N}(\mathbf{w}_{\text{MAP}} | \mathbf{0}, \alpha^{-1}\mathbf{I}) \frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}}.$$

Taking the logarithm of both sides and then using (2.42) and (2.43), we obtain the desired result.

5.40 For a K-class neural network, the likelihood function is given by

$$\prod_n \prod_k y_k(\mathbf{x}_n, \mathbf{w})^{t_{nk}}$$

and the corresponding error function is given by (5.24).

Again we would use a Laplace approximation for the posterior distribution over the weights, but the corresponding Hessian matrix, $\mathbf{H}$, in (5.166), would now be derived from (5.24). Similarly, (5.24), would replace the binary cross entropy error term in the regularized error function (5.184).

The predictive distribution for a new pattern would again have to be approximated, since the resulting marginalization cannot be done analytically. However, in contrast to the two-class problem, there is no obvious candidate for this approximation, although Gibbs (1997) discusses various alternatives.

Chapter 6 Kernel Methods

6.1 We first of all note that $J(\mathbf{a})$ depends on $\mathbf{a}$ only through the form $\mathbf{Ka}$. Since typically the number N of data points is greater than the number M of basis functions, the matrix $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ will be rank deficient. There will then be M eigenvectors of $\mathbf{K}$ having non-zero eigenvalues, and $N - M$ eigenvectors with eigenvalue zero. We can then decompose $\mathbf{a} = \mathbf{a}_{\|} + \mathbf{a}_{\perp}$ where $\mathbf{a}_{\|}^{\mathrm{T}}\mathbf{a}_{\perp} = 0$ and $\mathbf{K}\mathbf{a}_{\perp} = \mathbf{0}$. Thus the value of $\mathbf{a}_{\perp}$ is not determined by $J(\mathbf{a})$. We can remove the ambiguity by setting $\mathbf{a}_{\perp} = \mathbf{0}$, or equivalently by adding a regularizer term

$$\frac{\epsilon}{2}\mathbf{a}_{\perp}^{\mathrm{T}}\mathbf{a}_{\perp}$$

to $J(\mathbf{a})$ where ϵ is a small positive constant. Then $\mathbf{a} = \mathbf{a}_{\|}$ where $\mathbf{a}_{\|}$ lies in the span of $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ and hence can be written as a linear combination of the columns of $\mathbf{\Phi}$, so that in component notation

$$a_n = \sum_{i=1}^{M} u_i \phi_i(\mathbf{x}_n)$$

or equivalently in vector notation

$$\mathbf{a} = \mathbf{\Phi}\mathbf{u}. \tag{122}$$

Substituting (122) into (6.7) we obtain

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{K}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}\mathbf{u} \\ &= \frac{1}{2}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{u} \end{aligned} \tag{123}$$

We now make use of the matrix identity (C.7) to give

$$\alpha^{-1}\mathbf{I}_M - \alpha^{-1}\mathbf{I}_M\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}(\alpha^{-1}\mathbf{I}_M)\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\alpha^{-1}\mathbf{I}_M = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} = \mathbf{S}_N,$$

where we have also used (3.54). Substituting this in (126), we obtain

$$\sigma_N^2(\mathbf{x}_{N+1}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})$$

as derived for the linear regression model in Section 3.3.2.

6.23 **NOTE**: In the 1st printing of PRML, a typographical mistake appears in the text of the exercise at line three, where it should say "... a training set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$".

If we assume that the target variables, $t_1, \ldots, t_D$, are independent given the input vector, $\mathbf{x}$, this extension is straightforward.

Using analogous notation to the univariate case,

$$p(\mathbf{t}_{N+1}|\mathbf{T}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{m}(\mathbf{x}_{N+1}), \sigma(\mathbf{x}_{N+1})\mathbf{I}),$$

where $\mathbf{T}$ is a $N \times D$ matrix with the vectors $\mathbf{t}_1^{\mathrm{T}}, \ldots, \mathbf{t}_N^{\mathrm{T}}$ as its rows,

$$\mathbf{m}(\mathbf{x}_{N+1})^{\mathrm{T}} = \mathbf{k}^{\mathrm{T}}\mathbf{C}_N\mathbf{T}$$

and $\sigma(\mathbf{x}_{N+1})$ is given by (6.67). Note that $\mathbf{C}_N$, which only depend on the input vectors, is the same in the uni- and multivariate models.

6.25 Substituting the gradient and the Hessian into the Newton-Raphson formula we obtain

$$\begin{aligned}\mathbf{a}_N^{\text{new}} &= \mathbf{a}_N + (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1}\mathbf{a}_N\right] \\ &= (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right] \\ &= \mathbf{C}_N(\mathbf{I} + \mathbf{W}_N\mathbf{C}_N)^{-1}\left[\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\right]\end{aligned}$$

Chapter 7 Sparse Kernel Machines

7.1 From Bayes' theorem we have

$$p(t|\mathbf{x}) \propto p(\mathbf{x}|t)p(t)$$

where, from (2.249),

$$p(\mathbf{x}|t) = \frac{1}{N_t}\sum_{n=1}^{N}\frac{1}{Z_k}k(\mathbf{x}, \mathbf{x}_n)\delta(t, t_n).$$

Here N_t is the number of input vectors with label t ($+1$ or -1) and $N = N_{+1} + N_{-1}$. $\delta(t, t_n)$ equals 1 if $t = t_n$ and 0 otherwise. Z_k is the normalisation constant for the kernel. The minimum misclassification-rate is achieved if, for each new input vector, $\tilde{\mathbf{x}}$, we chose $\tilde{t}$ to maximise $p(\tilde{t}|\tilde{\mathbf{x}})$. With equal class priors, this is equivalent to maximizing $p(\tilde{\mathbf{x}}|\tilde{t})$ and thus

$$\tilde{t} = \begin{cases} +1 & \text{iff } \dfrac{1}{N_{+1}} \displaystyle\sum_{i:t_i=+1} k(\tilde{\mathbf{x}}, \mathbf{x}_i) \geqslant \dfrac{1}{N_{-1}} \sum_{j:t_j=-1} k(\tilde{\mathbf{x}}, \mathbf{x}_j) \\ -1 & \text{otherwise.} \end{cases}$$

Here we have dropped the factor $1/Z_k$ since it only acts as a common scaling factor. Using the encoding scheme for the label, this classification rule can be written in the more compact form

$$\tilde{t} = \text{sign}\left(\sum_{n=1}^{N} \frac{t_n}{N_{t_n}} k(\tilde{\mathbf{x}}, \mathbf{x}_n) \right).$$

Now we take $k(\mathbf{x}, \mathbf{x}_n) = \mathbf{x}^{\mathrm{T}}\mathbf{x}_n$, which results in the kernel density

$$p(\mathbf{x}|t = +1) = \frac{1}{N_{+1}} \sum_{n:t_n=+1} \mathbf{x}^{\mathrm{T}}\mathbf{x}_n = \mathbf{x}^{\mathrm{T}}\bar{\mathbf{x}}^{+}.$$

Here, the sum in the middle experssion runs over all vectors $\mathbf{x}_n$ for which $t_n = +1$ and $\bar{\mathbf{x}}^{+}$ denotes the mean of these vectors, with the corresponding definition for the negative class. Note that this density is improper, since it cannot be normalized. However, we can still compare likelihoods under this density, resulting in the classification rule

$$\tilde{t} = \begin{cases} +1 & \text{if } \tilde{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{x}}^{+} \geqslant \tilde{\mathbf{x}}^{\mathrm{T}}\bar{\mathbf{x}}^{-}, \\ -1 & \text{otherwise.} \end{cases}$$

The same argument would of course also apply in the feature space $\phi(\mathbf{x})$.

7.4 From Figure 4.1 and (7.4), we see that the value of the margin

$$\rho = \frac{1}{\|\mathbf{w}\|} \quad \text{and so} \quad \frac{1}{\rho^2} = \|\mathbf{w}\|^2.$$

From (7.16) we see that, for the maximum margin solution, the second term of (7.7) vanishes and so we have

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2.$$

Using this together with (7.8), the dual (7.10) can be written as

$$\frac{1}{2}\|\mathbf{w}\|^2 = \sum_{n}^{N} a_n - \frac{1}{2}\|\mathbf{w}\|^2,$$

from which the desired result follows.

7.8 This follows from (7.67) and (7.68), which in turn follow from the KKT conditions, (E.9)–(E.11), for μ_n, ξ_n, $\widehat{\mu}_n$ and $\widehat{\xi}_n$, and the results obtained in (7.59) and (7.60).

For example, for μ_n and ξ_n, the KKT conditions are

$$\begin{aligned} \xi_n &\geqslant 0 \\ \mu_n &\geqslant 0 \\ \mu_n \xi_n &= 0 \end{aligned} \tag{127}$$

and from (7.59) we have that

$$\mu_n = C - a_n. \tag{128}$$

Combining (127) and (128), we get (7.67); similar reasoning for $\widehat{\mu}_n$ and $\widehat{\xi}_n$ lead to (7.68).

7.10 We first note that this result is given immediately from (2.113)–(2.115), but the task set in the exercise was to practice the technique of completing the square. In this solution and that of Exercise 7.12, we broadly follow the presentation in Section 3.5.1. Using (7.79) and (7.80), we can write (7.84) in a form similar to (3.78)

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{1}{(2\pi)^{N/2}} \prod_{i=1}^{M} \alpha_i \int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} \tag{129}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w}$$

and $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$.

Completing the square over $\mathbf{w}$, we get

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \mathbf{m}) + E(\mathbf{t}) \tag{130}$$

where $\mathbf{m}$ and $\boldsymbol{\Sigma}$ are given by (7.82) and (7.83), respectively, and

$$E(\mathbf{t}) = \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m}\right). \tag{131}$$

Using (130), we can evaluate the integral in (129) to obtain

$$\int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{t})\}\, (2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}. \tag{132}$$

Considering this as a function of $\mathbf{t}$ we see from (7.83), that we only need to deal with the factor $\exp\{-E(\mathbf{t})\}$. Using (7.82), (7.83), (C.7) and (7.86), we can re-write

(131) as follows

$$\begin{aligned} E(\mathbf{t}) &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m}\right) \\ &= \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \beta\mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\beta\right) \\ &= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta\mathbf{I} - \beta\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\beta\right)\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta\mathbf{I} - \beta\boldsymbol{\Phi}(\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\beta\right)\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\right)^{-1}\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}. \end{aligned}$$

This gives us the last term on the r.h.s. of (7.85); the two preceding terms are given implicitly, as they form the normalization constant for the posterior Gaussian distribution $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)$.

7.12 Using the results (129)–(132) from Solution 7.10, we can write (7.85) in the form of (3.86):

$$\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{N}\ln\alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi). \quad (133)$$

By making use of (131) and (7.83) together with (C.22), we can take the derivatives of this w.r.t α_i, yielding

$$\frac{\partial}{\partial\alpha_i}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2. \quad (134)$$

Setting this to zero and re-arranging, we obtain

$$\alpha_i = \frac{1 - \alpha_i\Sigma_{ii}}{m_i^2} = \frac{\gamma_i}{m_i^2},$$

where we have used (7.89). Similarly, for β we see that

$$\frac{\partial}{\partial\beta}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - \mathrm{Tr}\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right]\right). \quad (135)$$

Using (7.83), we can rewrite the argument of the trace operator as

$$\begin{aligned} \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= \boldsymbol{\Sigma}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= (\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= (\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})\beta^{-1}. \end{aligned} \quad (136)$$

Here the first factor on the r.h.s. of the last line equals (7.89) written in matrix form. We can use this to set (135) equal to zero and then re-arrange to obtain (7.88).

7.15 Using (7.94), (7.95) and (7.97)–(7.99), we can rewrite (7.85) as follows

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) &= -\frac{1}{2}\Big\{ N\ln(2\pi) + \ln|\mathbf{C}_{-i}||1+\alpha_i^{-1}\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| \\
&\quad + \mathbf{t}^{\mathrm{T}}\left(\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\right)\mathbf{t}\Big\} \\
&= -\frac{1}{2}\left\{N\ln(2\pi) + \ln|\mathbf{C}_{-i}| + \mathbf{t}^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\mathbf{t}\right\} \\
&\quad + \frac{1}{2}\left[-\ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| + \mathbf{t}^{\mathrm{T}}\frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^{\mathrm{T}}\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t}\right] \\
&= L(\alpha_{-i}) + \frac{1}{2}\left[\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right] \\
&= L(\alpha_{-i}) + \lambda(\alpha_i)
\end{aligned}$$

7.18 As the RVM can be regarded as a regularized logistic regression model, we can follow the sequence of steps used to derive (4.91) in Exercise 4.13 to derive the first term of the r.h.s. of (7.110), whereas the second term follows from standard matrix derivatives (see Appendix C). Note however, that in Exercise 4.13 we are dealing with the *negative* log-likelihood.

To derive (7.111), we make use of (106) and (107) from Exercise 4.13. If we write the first term of the r.h.s. of (7.110) in component form we get

$$\begin{aligned}
\frac{\partial}{\partial w_j}\sum_{n=1}^{N}(t_n - y_n)\phi_{ni} &= -\sum_{n=1}^{N}\frac{\partial y_n}{\partial a_n}\frac{\partial a_n}{\partial w_j}\phi_{ni} \\
&= -\sum_{n=1}^{N} y_n(1-y_n)\phi_{nj}\phi_{ni},
\end{aligned}$$

which, written in matrix form, equals the first term inside the parenthesis on the r.h.s. of (7.111). The second term again follows from standard matrix derivatives.

Chapter 8 Graphical Models

8.1 We want to show that, for (8.5),

$$\sum_{x_1}\cdots\sum_{x_K} p(\mathbf{x}) = \sum_{x_1}\cdots\sum_{x_K}\prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) = 1.$$

We assume that the nodes in the graph has been numbered such that x_1 is the root node and no arrows lead from a higher numbered node to a lower numbered node.

8.23 This follows from the fact that the message that a node, x_i, will send to a factor f_s, consists of the product of all other messages received by x_i. From (8.63) and (8.69), we have

$$\begin{aligned} p(x_i) &= \prod_{s \in \mathrm{ne}(x_i)} \mu_{f_s \to x_i}(x_i) \\ &= \mu_{f_s \to x_i}(x_i) \prod_{t \in \mathrm{ne}(x_i) \setminus f_s} \mu_{f_t \to x_i}(x_i) \\ &= \mu_{f_s \to x_i}(x_i)\, \mu_{x_i \to f_s}(x_i). \end{aligned}$$

8.28 If a graph has one or more cycles, there exists at least one set of nodes and edges such that, starting from an arbitrary node in the set, we can visit all the nodes in the set and return to the starting node, without traversing any edge more than once.

Consider one particular such cycle. When one of the nodes n_1 in the cycle sends a message to one of its neighbours n_2 in the cycle, this causes a pending messages on the edge to the next node n_3 in that cycle. Thus sending a pending message along an edge in the cycle always generates a pending message on the next edge in that cycle. Since this is true for every node in the cycle it follows that there will always exist at least one pending message in the graph.

8.29 We show this by induction over the number of nodes in the tree-structured factor graph.

First consider a graph with two nodes, in which case only two messages will be sent across the single edge, one in each direction. None of these messages will induce any pending messages and so the algorithm terminates.

We then assume that for a factor graph with N nodes, there will be no pending messages after a finite number of messages have been sent. Given such a graph, we can construct a new graph with $N + 1$ nodes by adding a new node. This new node will have a single edge to the original graph (since the graph must remain a tree) and so if this new node receives a message on this edge, it will induce no pending messages. A message sent from the new node will trigger propagation of messages in the original graph with N nodes, but by assumption, after a finite number of messages have been sent, there will be no pending messages and the algorithm will terminate.

Chapter 9 Mixture Models and EM

9.1 Since both the E- and the M-step minimise the distortion measure (9.1), the algorithm will never change from a particular assignment of data points to prototypes, unless the new assignment has a lower value for (9.1).

Since there is a finite number of possible assignments, each with a corresponding unique minimum of (9.1) w.r.t. the prototypes, $\{\boldsymbol{\mu}_k\}$, the K-means algorithm will

converge after a finite number of steps, when no re-assignment of data points to prototypes will result in a decrease of (9.1). When no-reassignment takes place, there also will not be any change in $\{\boldsymbol{\mu}_k\}$.

9.3 From (9.10) and (9.11), we have

$$p\left(\mathbf{x}\right) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_k}.$$

Exploiting the 1-of-K representation for $\mathbf{z}$, we can re-write the r.h.s. as

$$\sum_{j=1}^{K} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{I_{kj}} = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where $I_{kj} = 1$ if $k = j$ and 0 otherwise.

9.7 Consider first the optimization with respect to the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. For this we can ignore the terms in (9.36) which depend on $\ln \pi_k$. We note that, for each data point n, the quantities z_{nk} are all zero except for a particular element which equals one. We can therefore partition the data set into K groups, denoted $\mathbf{X}_k$, such that all the data points $\mathbf{x}_n$ assigned to component k are in group $\mathbf{X}_k$. The complete-data log likelihood function can then be written

$$\ln p\left(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}\right) = \sum_{k=1}^{K} \left\{ \sum_{n \in \mathbf{X}_k} \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

This represents the sum of K independent terms, one for each component in the mixture. When we maximize this term with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ we will simply be fitting the k^{th} component to the data set $\mathbf{X}_k$, for which we will obtain the usual maximum likelihood results for a single Gaussian, as discussed in Chapter 2.

For the mixing coefficients we need only consider the terms in $\ln \pi_k$ in (9.36), but we must introduce a Lagrange multiplier to handle the constraint $\sum_k \pi_k = 1$. Thus we maximize

$$\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

which gives

$$0 = \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda.$$

Multiplying through by π_k and summing over k we obtain $\lambda = -N$, from which we have

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk} = \frac{N_k}{N}$$

where N_k is the number of data points in group $\mathbf{X}_k$.

9.8 Using (2.43), we can write the r.h.s. of (9.40) as

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{K}\gamma(z_{nj})(\mathbf{x}_n-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_j)+\text{const.},$$

where 'const.' summarizes terms independent of $\boldsymbol{\mu}_j$ (for all j). Taking the derivative of this w.r.t. $\boldsymbol{\mu}_k$, we get

$$-\sum_{n=1}^{N}\gamma(z_{nk})\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k-\boldsymbol{\Sigma}^{-1}\mathbf{x}_n\right),$$

and setting this to zero and rearranging, we obtain (9.17).

9.12 Since the expectation of a sum is the sum of the expectations we have

$$\mathbb{E}[\mathbf{x}]=\sum_{k=1}^{K}\pi_k\mathbb{E}_k[\mathbf{x}]=\sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k$$

where $\mathbb{E}_k[\mathbf{x}]$ denotes the expectation of $\mathbf{x}$ under the distribution $p(\mathbf{x}|k)$. To find the covariance we use the general relation

$$\mathrm{cov}[\mathbf{x}]=\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}$$

to give

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}\\ &= \sum_{k=1}^{K}\pi_k\mathbb{E}_k[\mathbf{x}\mathbf{x}^{\mathrm{T}}]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}\\ &= \sum_{k=1}^{K}\pi_k\left\{\boldsymbol{\Sigma}_k+\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}\right\}-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}.\end{aligned}$$

9.15 This is easily shown by calculating the derivatives of (9.55), setting them to zero and solve for μ_{ki}. Using standard derivatives, we get

$$\begin{aligned}\frac{\partial}{\partial\mu_{ki}}\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] &= \sum_{n=1}^{N}\gamma(z_{nk})\left(\frac{x_{ni}}{\mu_{ki}}-\frac{1-x_{ni}}{1-\mu_{ki}}\right)\\ &= \frac{\sum_n\gamma(z_{nk})x_{ni}-\sum_n\gamma(z_{nk})\mu_{ki}}{\mu_{ki}(1-\mu_{ki})}.\end{aligned}$$

Setting this to zero and solving for μ_{ki}, we get

$$\mu_{ki}=\frac{\sum_n\gamma(z_{nk})x_{ni}}{\sum_n\gamma(z_{nk})},$$

which equals (9.59) when written in vector form.

9.17 This follows directly from the equation for the incomplete log-likelihood, (9.51). The largest value that the argument to the logarithm on the r.h.s. of (9.51) can have is 1, since $\forall n, k : 0 \leqslant p(\mathbf{x}_n|\boldsymbol{\mu}_k) \leqslant 1$, $0 \leqslant \pi_k \leqslant 1$ and $\sum_k^K \pi_k = 1$. Therefore, the maximum value for $\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi})$ equals 0.

9.20 If we take the derivatives of (9.62) w.r.t. α, we get

$$\frac{\partial}{\partial \alpha}\mathbb{E}\left[\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)\right] = \frac{M}{2}\frac{1}{\alpha} - \frac{1}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right].$$

Setting this equal to zero and re-arranging, we obtain (9.63).

9.23 **NOTE**: In the 1st printing of PRML, the task set in this exercise is to show that the two sets of re-estimation equations are formally equivalent, without any restriction. However, it really should be restricted to stationary points of the objective function.

Considering the case when the optimization has converged, we can start with α_i, as defined by (7.87), and use (7.89) to re-write this as

$$\alpha_i^\star = \frac{1 - \alpha_i^\star \Sigma_{ii}}{m_N^2},$$

where $\alpha_i^\star = \alpha_i^{\mathrm{new}} = \alpha_i$ is the value reached at convergence. We can re-write this as

$$\alpha_i^\star(m_i^2 + \Sigma_{ii}) = 1$$

which is easily re-written as (9.67).

For β, we start from (9.68), which we re-write as

$$\frac{1}{\beta^\star} = \frac{\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2}{N} + \frac{\sum_i \gamma_i}{\beta^\star N}.$$

As in the α-case, $\beta^\star = \beta^{\mathrm{new}} = \beta$ is the value reached at convergence. We can re-write this as

$$\frac{1}{\beta^\star}\left(N - \sum_i \gamma_i\right) = \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2,$$

which can easily be re-written as (7.88).

9.25 This follows from the fact that the Kullback-Leibler divergence, $\mathrm{KL}(q\|p)$, is at its minimum, 0, when q and p are identical. This means that

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathrm{KL}(q\|p) = \mathbf{0},$$

since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. Therefore, if we compute the gradient of both sides of (9.70) w.r.t. $\boldsymbol{\theta}$, the contribution from the second term on the r.h.s. will be $\mathbf{0}$, and so the gradient of the first term must equal that of the l.h.s.

9.26 From (9.18) we get

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk}). \tag{137}$$

We get N_k^{new} by recomputing the responsibilities, $\gamma(z_{mk})$, for a specific data point, $\mathbf{x}_m$, yielding

$$N_k^{\text{new}} = \sum_{n\neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}). \tag{138}$$

Combining this with (137), we get (9.79).

Similarly, from (9.17) we have

$$\boldsymbol{\mu}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_n \gamma^{\text{old}}(z_{nk})\mathbf{x}_n$$

and recomputing the responsibilities, $\gamma(z_{mk})$, we get

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}}\left(\sum_{n\neq m} \gamma^{\text{old}}(z_{nk})\mathbf{x}_n + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\right) \\
&= \frac{1}{N_k^{\text{new}}}\left(N_k^{\text{old}}\boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk})\mathbf{x}_m + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\right) \\
&= \frac{1}{N_k^{\text{new}}}\Big(\left(N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk})\right)\boldsymbol{\mu}_k^{\text{old}} \\
&\qquad -\gamma^{\text{old}}(z_{mk})\mathbf{x}_m + \gamma^{\text{new}}(z_{mk})\mathbf{x}_m\Big) \\
&= \boldsymbol{\mu}_k^{\text{old}} + \left(\frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\right)(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}),
\end{aligned}$$

where we have used (9.79).

Chapter 10 Approximate Inference

10.1 Starting from (10.3), we use the product rule together with (10.4) to get

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right\} \mathrm{d}\mathbf{Z} \\
&= \int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X}\mid\mathbf{Z})\,p(\mathbf{X})}{q(\mathbf{Z})}\right\} \mathrm{d}\mathbf{Z} \\
&= \int q(\mathbf{Z}) \left(\ln\left\{\frac{p(\mathbf{X}\mid\mathbf{Z})}{q(\mathbf{Z})}\right\} + \ln p(\mathbf{X})\right) \mathrm{d}\mathbf{Z} \\
&= -\mathrm{KL}(q\parallel p) + \ln p(\mathbf{X}).
\end{aligned}$$

Rearranging this, we immediately get (10.2).