tors inside $\{\ldots\})^{-1}$ in (146), we can use (135)–(138) to get

$$\begin{aligned}
&\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}\left(\beta|a_N, b_N\right) \\
&= \left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\Bigg(-\frac{\beta}{2}\big(\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} \\
&\quad + \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\big)\Bigg)\Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_N-1}\exp(-b_N\beta) \\
&= \left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\Bigg(-\frac{\beta}{2}\big(\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 \\
&\quad - \mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} - \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} - \mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\big)\Bigg) \\
&\qquad \Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_0+N/2-1} \\
&\qquad \exp\left(-\left(b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \mathbf{t}^{\mathrm{T}}\mathbf{t}\right)\right)\beta\right) \\
&= \left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\left(-\frac{\beta}{2}\left((\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0(\mathbf{w}-\mathbf{m}_0) + \|\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\|^2\right)\right) \\
&\quad \Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_N+N/2-1}\exp(-b_0\beta).
\end{aligned}$$

Substituting this into (146), the exponential factors along with $\beta^{a_0+N/2-1}(\beta/2\pi)^{M/2}$ cancel and we are left with (3.118).

Chapter 4 Linear Models for Classification

4.1 Assume that the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ intersect. Then there exist a point $\mathbf{z}$ such that

$$\mathbf{z} = \sum_n \alpha_n\mathbf{x}_n = \sum_m \beta_m\mathbf{y}_m$$

where $\beta_m \geqslant 0$ for all m and $\sum_m \beta_m = 1$. If $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ also were to be linearly separable, we would have that

$$\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_n \alpha_n\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 = \sum_n \alpha_n($$

since $\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 > 0$ and the $\{\alpha_n\}$ are all non-negative and sum to 1, but by the corresponding argument

$$\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_m \beta_m\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0 = \sum_m \beta_m(\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0) < 0,$$

which is a contradiction and hence $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ cannot be linearly separable if their convex hulls intersect.

If we instead assume that $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ are linearly separable and consider a point $\mathbf{z}$ in the intersection of their convex hulls, the same contradiction arise. Thus no such point can exist and the intersection of the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ must be empty.

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

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\right\}, \tag{147}$$

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 (147) 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

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \tag{148}$$

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 (148) into (147), 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 $\mathbf{x}^\star$,

$$\begin{aligned}\mathbf{y}(\mathbf{x}^\star) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \mathbf{w}_0 \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\ &= \bar{\mathbf{t}} - \widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}).\end{aligned} \tag{149}$$

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

4.3 When we consider several simultaneous constraints, (4.157) becomes

$$\mathbf{A}\mathbf{t}_n + \mathbf{b} = \mathbf{0}, \tag{150}$$

where $\mathbf{A}$ is a matrix and $\mathbf{b}$ is a column vector such that each row of $\mathbf{A}$ and element of $\mathbf{b}$ correspond to one linear constraint.

If we apply (150) to (149), we obtain

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

since $\mathbf{A}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{A}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = \mathbf{b}\mathbf{1}^{\mathrm{T}} - \mathbf{b}\mathbf{1}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$. Thus $\mathbf{A}\mathbf{y}(\mathbf{x}^{\star}) + \mathbf{b} = \mathbf{0}$.

4.4 **NOTE**: In 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

$$\nabla L = \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} \tag{151}$$

and setting this gradient to zero gives

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

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

4.5 Starting with the numerator on the r.h.s. of (4.25), we can use (4.23) and (4.27) to rewrite it as follows:

$$\begin{aligned}(m_2 - m_1)^2 &= \left(\mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)\right)^2 \\ &= \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}.\end{aligned} \tag{152}$$

Similarly, we can use (4.20), (4.23), (4.24), and (4.28) to rewrite the denominator of the r.h.s. of (4.25):

$$\begin{aligned} s_1^2 + s_2^2 &= \sum_{n\in\mathcal{C}_1} (y_n - m_1)^2 + \sum_{k\in\mathcal{C}_2} (y_k - m_2)^2 \\ &= \sum_{n\in\mathcal{C}_1} \left(\mathbf{w}^{\mathrm{T}}(\mathbf{x}_n - \mathbf{m}_1)\right)^2 + \sum_{k\in\mathcal{C}_2} \left(\mathbf{w}^{\mathrm{T}}(\mathbf{x}_k - \mathbf{m}_2)\right)^2 \\ &= \sum_{n\in\mathcal{C}_1} \mathbf{w}^{\mathrm{T}}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w} \\ &\qquad + \sum_{k\in\mathcal{C}_2} \mathbf{w}^{\mathrm{T}}(\mathbf{x}_k - \mathbf{m}_2)(\mathbf{x}_k - \mathbf{m}_2)^{\mathrm{T}}\mathbf{w} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}. \end{aligned} \tag{153}$$

Substituting (152) and (153) in (4.25) we obtain (4.26).

4.6 Using (4.21) and (4.34) along with the chosen target coding scheme, we can re-write the l.h.s. of (4.33) as follows:

$$\begin{aligned} \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - w_0 - t_n\right)\mathbf{x}_n &= \sum_{n=1}^{N} \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - \mathbf{w}^{\mathrm{T}}\mathbf{m} - t_n\right)\mathbf{x}_n \\ &= \sum_{n=1}^{N} \left\{\left(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - \mathbf{x}_n t_n\right\} \\ &= \sum_{n\in\mathcal{C}_1} \left\{\left(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - \mathbf{x}_n t_n\right\} \\ &\qquad \sum_{m\in\mathcal{C}_2} \left\{\left(\mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} - \mathbf{x}_m\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - \mathbf{x}_m t_m\right\} \\ &= \left(\sum_{n\in\mathcal{C}_1} \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - N_1\mathbf{m}_1\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} - N_1\mathbf{m}_1\frac{N}{N_1} \\ &\qquad \left(\sum_{m\in\mathcal{C}_2} \mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} - N_2\mathbf{m}_2\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} + N_2\mathbf{m}_2\frac{N}{N_2} \\ &= \left(\sum_{n\in\mathcal{C}_1} \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} + \sum_{m\in\mathcal{C}_2} \mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} - (N_1\mathbf{m}_1 + N_2\mathbf{m}_2)\mathbf{m}^{\mathrm{T}}\right)\mathbf{w} \\ &\qquad -N(\mathbf{m}_1 - \mathbf{m}_2). \end{aligned} \tag{154}$$

We then use the identity

$$\begin{aligned}\sum_{i\in\mathcal{C}_k}(\mathbf{x}_i-\mathbf{m}_k)(\mathbf{x}_i-\mathbf{m}_k)^{\mathrm{T}} &= \sum_{i\in\mathcal{C}_k}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}-\mathbf{x}_i\mathbf{m}_k^{\mathrm{T}}-\mathbf{m}_k\mathbf{x}_i^{\mathrm{T}}+\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}}\right)\\ &= \sum_{i\in\mathcal{C}_k}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}-N_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}}\end{aligned}$$

together with (4.28) and (4.36) to rewrite (154) as

$$\begin{aligned}&\Big(\mathbf{S}_{\mathrm{W}}+N_1\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}+N_2\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\\ &\qquad -(N_1\mathbf{m}_1+N_2\mathbf{m}_2)\frac{1}{N}(N_1\mathbf{m}_1+N_2\mathbf{m}_2)\Big)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\left(N_1-\frac{N_1^2}{N}\right)\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}-\frac{N_1N_2}{N}(\mathbf{m}_1\mathbf{m}_2^{\mathrm{T}}+\mathbf{m}_2\mathbf{m}_1)\right.\\ &\qquad \left.+\left(N_2-\frac{N_2^2}{N}\right)\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\right)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\frac{(N_1+N_2)N_1-N_1^2}{N}\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}-\frac{N_1N_2}{N}(\mathbf{m}_1\mathbf{m}_2^{\mathrm{T}}+\mathbf{m}_2\mathbf{m}_1)\right.\\ &\qquad \left.+\frac{(N_1+N_2)N_2-N_2^2}{N}\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\right)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\frac{N_2N_1}{N}\left(\mathbf{m}_1\mathbf{m}_1^{\mathrm{T}}-\mathbf{m}_1\mathbf{m}_2^{\mathrm{T}}-\mathbf{m}_2\mathbf{m}_1+\mathbf{m}_2\mathbf{m}_2^{\mathrm{T}}\right)\right)\mathbf{w}\\ &\qquad -N(\mathbf{m}_1-\mathbf{m}_2)\\ =\;&\left(\mathbf{S}_{\mathrm{W}}+\frac{N_2N_1}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w}-N(\mathbf{m}_1-\mathbf{m}_2),\end{aligned}$$

where in the last line we also made use of (4.27). From (4.33), this must equal zero, and hence we obtain (4.37).

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

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

4.8 Substituting (4.64) into (4.58), we see that the normalizing constants cancel and we are left with

$$\begin{aligned} a &= \ln \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_1\right)\right)p(\mathcal{C}_1)}{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_2\right)\right)p(\mathcal{C}_2)} \\ &= -\frac{1}{2}\left(\mathbf{x}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{x} - \mathbf{x}\boldsymbol{\Sigma}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} + \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_1\right. \\ &\qquad \left. -\mathbf{x}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{x} + \mathbf{x}\boldsymbol{\Sigma}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_2\right) + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= \left(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\mu}_2\right) + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}. \end{aligned}$$

Substituting this into the rightmost form of (4.57) we obtain (4.65), with $\mathbf{w}$ and w_0 given by (4.66) and (4.67), respectively.

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

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln p(\phi_n|\mathcal{C}_k) + \ln\pi_k\right\}. \tag{155}$$

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

$$-\pi_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k. \tag{156}$$

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

4.10 If we substitute (4.160) into (155) and then use the definition of the multivariate Gaussian, (2.43), we obtain

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \\ -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}, \tag{157}$$

where we have dropped terms independent of $\{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Sigma}$.

Setting the derivative of the r.h.s. of (157) w.r.t. $\boldsymbol{\mu}_k$, obtained by using (C.19), to zero, we get

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) = 0.$$

Making use of (156), we can re-arrange this to obtain (4.161).

Rewriting the r.h.s. of (157) as

$$-\frac{1}{2}b\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\right]\right\},$$

we can use (C.24) and (C.28) to calculate the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$. Setting this to zero we obtain

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{k}^{\mathrm{T}} t_{nk}\left\{\boldsymbol{\Sigma} - (\boldsymbol{\phi}_n - \boldsymbol{\mu}_n)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right\} = 0.$$

Again making use of (156), we can re-arrange this to obtain (4.162), with $\mathbf{S}_k$ given by (4.163).

Note that, as in Exercise 2.34, we do not enforce that $\boldsymbol{\Sigma}$ should be symmetric, but simply note that the solution is automatically symmetric.

4.11 The generative model for $\boldsymbol{\phi}$ corresponding to the chosen coding scheme is given by

$$p(\boldsymbol{\phi} \mid \mathcal{C}_k) = \prod_{m=1}^{M} p(\boldsymbol{\phi}_m \mid \mathcal{C}_k)$$

where

$$p(\boldsymbol{\phi}_m \mid \mathcal{C}_k) = \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}},$$

where in turn $\{\mu_{kml}\}$ are the parameters of the multinomial models for $\boldsymbol{\phi}$. Substituting this into (4.63) we see that

$$\begin{aligned} a_k &= \ln p(\boldsymbol{\phi} \mid \mathcal{C}_k)\, p(\mathcal{C}_k) \\ &= \ln p(\mathcal{C}_k) + \sum_{m=1}^{M} \ln p(\boldsymbol{\phi}_m \mid \mathcal{C}_k) \\ &= \ln p(\mathcal{C}_k) + \sum_{m=1}^{M} \sum_{l=1}^{L} \phi_{ml} \ln \mu_{kml}, \end{aligned}$$

which is linear in ϕ_{ml}.

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

$$\begin{aligned} \frac{\partial E}{\partial y_n} &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} && (158) \\ &= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\ &= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} && (159) \\ &= \frac{y_n - t_n}{y_n(1-y_n)}. && (160) \end{aligned}$$

From (4.88), we see that

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)\left(1 - \sigma(a_n)\right) = y_n(1 - y_n). \tag{161}$$

Finally, we have

$$\nabla a_n = \boldsymbol{\phi}_n \tag{162}$$

where ∇ denotes the gradient with respect to $\mathbf{w}$. Combining (160), (161) and (162) 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.

4.14 If the data set is linearly separable, any decision boundary separating the two classes will have the property

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n \begin{cases} \geqslant 0 & \text{if } t_n = 1, \\ < 0 & \text{otherwise.} \end{cases}$$

Moreover, from (4.90) we see that the negative log-likelihood will be minimized (i.e., the likelihood maximized) when $y_n = \sigma\left(\mathbf{w}_{\mathrm{T}}\boldsymbol{\phi}_n\right) = t_n$ for all n. This will be the case when the sigmoid function is saturated, which occurs when its argument, $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$, goes to $\pm\infty$, i.e., when the magnitude of $\mathbf{w}$ goes to infinity.

4.15 **NOTE**: In PRML, "concave" should be "convex" on the last line of the exercise.

Assuming that the argument to the sigmoid function (4.87) is finite, the diagonal elements of $\mathbf{R}$ will be strictly positive. Then

$$\mathbf{v}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{v} = \left(\mathbf{v}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}^{1/2}\right)\left(\mathbf{R}^{1/2}\boldsymbol{\Phi}\mathbf{v}\right) = \left\|\mathbf{R}^{1/2}\boldsymbol{\Phi}\mathbf{v}\right\|^2 > 0$$

where $\mathbf{R}^{1/2}$ is a diagonal matrix with elements $(y_n(1 - y_n))^{1/2}$, and thus $\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}$ is positive definite.

Now consider a Taylor expansion of $E(\mathbf{w})$ around a minima, $\mathbf{w}^\star$,

$$E(\mathbf{w}) = E(\mathbf{w}^\star) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^\star\right)^{\mathrm{T}}\mathbf{H}\left(\mathbf{w} - \mathbf{w}^\star\right)$$

where the linear term has vanished since $\mathbf{w}^\star$ is a minimum. Now let

$$\mathbf{w} = \mathbf{w}^\star + \lambda\mathbf{v}$$

where $\mathbf{v}$ is an arbitrary, non-zero vector in the weight space and consider

$$\frac{\partial^2 E}{\partial \lambda^2} = \mathbf{v}^{\mathrm{T}} \mathbf{H} \mathbf{v} > 0.$$

This shows that $E(\mathbf{w})$ is convex. Moreover, at the minimum of $E(\mathbf{w})$,

$$\mathbf{H}\left(\mathbf{w} - \mathbf{w}^{\star}\right) = 0$$

and since $\mathbf{H}$ is positive definite, $\mathbf{H}^{-1}$ exists and $\mathbf{w} = \mathbf{w}^{\star}$ must be the unique minimum.

4.16 If the values of the $\{t_n\}$ were known then each data point for which $t_n = 1$ would contribute $p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n))$ to the log likelihood, and each point for which $t_n = 0$ would contribute $1 - p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n))$ to the log likelihood. A data point whose probability of having $t_n = 1$ is given by π_n will therefore contribute

$$\pi_n p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n)) + (1 - \pi_n)(1 - p(t_n = 1|\boldsymbol{\phi}(\mathbf{x}_n)))$$

and so the overall log likelihood for the data set is given by

$$\sum_{n=1}^{N} \pi_n \ln p\left(t_n = 1 \mid \boldsymbol{\phi}(\mathbf{x}_n)\right) + (1 - \pi_n) \ln\left(1 - p\left(t_n = 1 \mid \boldsymbol{\phi}(\mathbf{x}_n)\right)\right). \tag{163}$$

This can also be viewed from a sampling perspective by imagining sampling the value of each t_n some number M times, with probability of $t_n = 1$ given by π_n, and then constructing the likelihood function for this expanded data set, and dividing by M. In the limit $M \to \infty$ we recover (163).

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).

4.18 **NOTE**: In PRML, the text of the exercise refers equation (4.91) where it should refer to (4.106).

From (4.108) we have

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}.$$

If we combine this with (4.106) using the chain rule, we get

$$\begin{aligned}\frac{\partial E}{\partial a_{nj}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} \\ &= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk} \left(I_{kj} - y_{nj}\right) \\ &= y_{nj} - t_{nj},\end{aligned}$$

where we have used that $\forall n : \sum_k t_{nk} = 1$.

If we combine this with (162), again using the chain rule, we obtain (4.109).

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

$$\frac{\partial E}{\partial y_n} = \frac{y_n - t_n}{y_n(1 - y_n)}. \tag{164}$$

Also

$$\nabla a_n = \boldsymbol{\phi}_n. \tag{165}$$

From (4.115) and (4.116) we have

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \Phi(a_n)}{\partial a_n} = \frac{1}{\sqrt{2\pi}} e^{-a_n^2}. \tag{166}$$

Combining (164), (165) and (166), we get

$$\nabla E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla a_n = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1 - y_n)} \frac{1}{\sqrt{2\pi}} e^{-a_n^2} \boldsymbol{\phi}_n. \tag{167}$$

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

$$\begin{aligned}\frac{\partial}{\partial y_n} \frac{y_n - t_n}{y_n(1 - y_n)} &= \frac{y_n(1 - y_n)}{y_n^2(1 - y_n)^2} - \frac{(y_n - t_n)(1 - 2y_n)}{y_n^2(1 - y_n)^2} \\ &= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1 - y_n)^2}.\end{aligned} \tag{168}$$

Then using (165)–(168) we have

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

4.20 **NOTE**: In PRML, equation (4.110) contains an incorrect leading minus sign ('$-$') on the right hand side.

We first write out the components of the $MK \times MK$ Hessian matrix in the form

$$\frac{\partial^2 E}{\partial w_{ki} \partial w_{jl}} = \sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\phi_{ni}\phi_{nl}.$$

To keep the notation uncluttered, consider just one term in the summation over n and show that this is positive semi-definite. The sum over n will then also be positive semi-definite. Consider an arbitrary vector of dimension MK with elements u_{ki}. Then

$$\begin{aligned}
\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} &= \sum_{i,j,k,l} u_{ki} y_k (I_{kj} - y_j)\phi_i \phi_l u_{jl} \\
&= \sum_{j,k} b_j y_k (I_{kj} - y_j) b_k \\
&= \sum_k y_k b_k^2 - \left(\sum_k b_k y_k \right)^2
\end{aligned}$$

where

$$b_k = \sum_i u_{ki}\phi_{ni}.$$

We now note that the quantities y_k satisfy $0 \leqslant y_k \leqslant 1$ and $\sum_k y_k = 1$. Furthermore, the function $f(b) = b^2$ is a concave function. We can therefore apply Jensen's inequality to give

$$\sum_k y_k b_k^2 = \sum_k y_k f(b_k) \geqslant f\left(\sum_k y_k b_k \right) = \left(\sum_k y_k b_k \right)^2$$

and hence

$$\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} \geqslant 0.$$

Note that the equality will never arise for finite values of a_k where a_k is the set of arguments to the softmax function. However, the Hessian can be positive *semi*-definite since the basis vectors ϕ_{ni} could be such as to have zero dot product for a linear subspace of vectors u_{ki}. In this case the minimum of the error function would comprise a continuum of solutions all having the same value of the error function.

4.21 **NOTE**: In PRML, Equation (4.116) contains a minor typographical error. On the l.h.s., $\mathbf{\Phi}$ should be Φ (i.e. not bold).

We consider the two cases where $a \geqslant 0$ and $a < 0$ separately. In the first case, we

can use (2.42) to rewrite (4.114) as

$$\begin{aligned}\Phi(a) &= \int_{-\infty}^{0} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta + \int_{0}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) \mathrm{d}\theta \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{a} \exp\left(-\frac{\theta^2}{2}\right) \mathrm{d}\theta \\ &= \frac{1}{2}\left(1 + \frac{1}{\sqrt{2}}\mathrm{erf}(a)\right),\end{aligned}$$

where, in the last line, we have used (4.115).

When $a < 0$, the symmetry of the Gaussian distribution gives

$$\Phi(a) = 1 - \Phi(-a).$$

Combining this with (169), we get

$$\begin{aligned}\Phi(a) &= 1 - \frac{1}{2}\left(1 + \frac{1}{\sqrt{2}}\mathrm{erf}(-a)\right) \\ &= \frac{1}{2}\left(1 + \frac{1}{\sqrt{2}}\mathrm{erf}(a)\right),\end{aligned}$$

where we have used the fact that the erf function is is anti-symmetric, i.e., $\mathrm{erf}(-a) = -\mathrm{erf}(a)$.

4.22 Starting from (4.136), using (4.135), we have

$$\begin{aligned}p(\mathcal{D}) &= \int p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &\simeq p(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}})\, p(\boldsymbol{\theta}_{\mathrm{MAP}}) \\ &\qquad \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})\mathbf{A}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})\right) \mathrm{d}\boldsymbol{\theta} \\ &= p(\mathcal{D} \mid \boldsymbol{\theta}_{\mathrm{MAP}})\, p(\boldsymbol{\theta}_{\mathrm{MAP}})\, \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}},\end{aligned}$$

where $\mathbf{A}$ is given by (4.138). Taking the logarithm of this yields (4.137).

4.23 **NOTE**: In 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

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \tag{169}$$

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 (169) 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 (169), we obtain (4.139) by dropping the $\ln|\widehat{\mathbf{H}}|$ since this $O(1)$ compared to $\ln N$.

4.24 Consider a rotation of the coordinate axes of the M-dimensional vector $\mathbf{w}$ such that $\mathbf{w} = (w_\parallel, \mathbf{w}_\perp)$ where $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi} = w_\parallel\|\boldsymbol{\phi}\|$, and $\mathbf{w}_\perp$ is a vector of length $M-1$. We then have

$$\begin{aligned}\int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} &= \iint \sigma\left(w_\parallel\|\boldsymbol{\phi}\|\right) q(\mathbf{w}_\perp|w_\parallel)q(w_\parallel)\,\mathrm{d}w_\parallel\,\mathrm{d}\mathbf{w}_\perp \\ &= \int \sigma(w_\parallel\|\boldsymbol{\phi}\|)q(w_\parallel)\,\mathrm{d}w_\parallel.\end{aligned}$$

Note that the joint distribution $q(\mathbf{w}_\perp, w_\parallel)$ is Gaussian. Hence the marginal distribution $q(w_\parallel)$ is also Gaussian and can be found using the standard results presented in

Section 2.3.2. Denoting the unit vector

$$\mathbf{e} = \frac{1}{\|\boldsymbol{\phi}\|}\boldsymbol{\phi}$$

we have

$$q(w_{\|}) = \mathcal{N}(w_{\|}|\mathbf{e}^{\mathrm{T}}\mathbf{m}_N, \mathbf{e}^{\mathrm{T}}\mathbf{S}_N\mathbf{e}).$$

Defining $a = w_{\|}\|\boldsymbol{\phi}\|$ we see that the distribution of a is given by a simple re-scaling of the Gaussian, so that

$$q(a) = \mathcal{N}(a|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N, \boldsymbol{\phi}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi})$$

where we have used $\|\boldsymbol{\phi}\|\mathbf{e} = \boldsymbol{\phi}$. Thus we obtain (4.151) with μ_a given by (4.149) and σ_a^2 given by (4.150).

4.25 From (4.88) we have that

$$\begin{aligned}\left.\frac{\mathrm{d}\sigma}{\mathrm{d}a}\right|_{a=0} &= \sigma(0)(1-\sigma(0)) \\ &= \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}. \end{aligned} \tag{170}$$

Since the derivative of a cumulative distribution function is simply the corresponding density function, (4.114) gives

$$\begin{aligned}\left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} &= \lambda\mathcal{N}\left(0|0,1\right) \\ &= \lambda\frac{1}{\sqrt{2\pi}}.\end{aligned}$$

Setting this equal to (170), we see that

$$\lambda = \frac{\sqrt{2\pi}}{4} \quad \text{or equivalently} \quad \lambda^2 = \frac{\pi}{8}.$$

This is illustrated in Figure 4.9.

4.26 First of all consider the derivative of the right hand side with respect to μ, making use of the definition of the probit function, giving

$$\left(\frac{1}{2\pi}\right)^{1/2}\exp\left\{-\frac{\mu^2}{2(\lambda^{-2}+\sigma^2)}\right\}\frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}}.$$

Now make the change of variable $a = \mu + \sigma z$, so that the left hand side of (4.152) becomes

$$\int_{-\infty}^{\infty}\Phi(\lambda\mu+\lambda\sigma z)\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}z^2\right\}\sigma\,\mathrm{d}z$$

where we have substituted for the Gaussian distribution. Now differentiate with respect to μ, making use of the definition of the probit function, giving

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}z^2-\frac{\lambda^2}{2}(\mu+\sigma z)^2\right\}\sigma\,\mathrm{d}z.$$

The integral over z takes the standard Gaussian form and can be evaluated analytically by making use of the standard result for the normalization coefficient of a Gaussian distribution. To do this we first complete the square in the exponent

$$\begin{aligned}
&-\frac{1}{2}z^2-\frac{\lambda^2}{2}(\mu+\sigma z)^2\\
&\quad=\quad -\frac{1}{2}z^2(1+\lambda^2\sigma^2)-z\lambda^2\mu\sigma-\frac{1}{2}\lambda^2\mu^2\\
&\quad=\quad -\frac{1}{2}\left[z+\lambda^2\mu\sigma(1+\lambda^2\sigma^2)^{-1}\right]^2(1+\lambda^2\sigma^2)+\frac{1}{2}\frac{\lambda^4\mu^2\sigma^2}{(1+\lambda^2\sigma^2)}-\frac{1}{2}\lambda^2\mu^2.
\end{aligned}$$

Integrating over z then gives the following result for the derivative of the left hand side

$$\begin{aligned}
&\frac{1}{(2\pi)^{1/2}}\frac{1}{(1+\lambda^2\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}\lambda^2\mu^2+\frac{1}{2}\frac{\lambda^4\mu^2\sigma^2}{(1+\lambda^2\sigma^2)}\right\}\\
&\quad=\quad \frac{1}{(2\pi)^{1/2}}\frac{1}{(1+\lambda^2\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2}\frac{\lambda^2\mu^2}{(1+\lambda^2\sigma^2)}\right\}.
\end{aligned}$$

Thus the derivatives of the left and right hand sides of (4.152) with respect to μ are equal. It follows that the left and right hand sides are equal up to a function of σ^2 and λ. Taking the limit $\mu\to-\infty$ the left and right hand sides both go to zero, showing that the constant of integration must also be zero.

Chapter 5 Neural Networks

5.1 **NOTE**: In PRML, the text of this exercise contains a typographical error. On line 2, $g(\cdot)$ should be replaced by $h(\cdot)$.

See Solution 3.1.

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.

5.3 In this case, the likelihood function becomes

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \boldsymbol{\Sigma}\right),$$

with the corresponding log-likelihood function

$$\begin{aligned}
&\ln p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma}) \\
&\quad = -\frac{N}{2}\left(\ln|\boldsymbol{\Sigma}| + K\ln(2\pi)\right) - \frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{y}_n), \quad (171)
\end{aligned}$$

where $\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n, \mathbf{w})$ and K is the dimensionality of $\mathbf{y}$ and $\mathbf{t}$.

If we first treat $\boldsymbol{\Sigma}$ as fixed and known, we can drop terms that are independent of $\mathbf{w}$ from (171), and by changing the sign we get the error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{y}_n).$$

If we consider maximizing (171) w.r.t. $\boldsymbol{\Sigma}$, the terms that need to be kept are

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

By rewriting the second term we get

$$-\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\right].$$

Using results from Appendix C, we can maximize this by setting the derivative w.r.t. $\mathbf{\Sigma}^{-1}$ to zero, yielding

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}.$$

Thus the optimal value for $\mathbf{\Sigma}$ depends on $\mathbf{w}$ through $\mathbf{y}_n$.

A possible way to address this mutual dependency between $\mathbf{w}$ and $\mathbf{\Sigma}$ when it comes to optimization, is to adopt an iterative scheme, alternating between updates of $\mathbf{w}$ and $\mathbf{\Sigma}$ until some convergence criterion is reached.

5.4 Let $t \in \{0, 1\}$ denote the data set label and let $k \in \{0, 1\}$ denote the true class label. We want the network output to have the interpretation $y(\mathbf{x}, \mathbf{w}) = p(k = 1|\mathbf{x})$. From the rules of probability we have

$$p(t = 1|\mathbf{x}) = \sum_{k=0}^{1} p(t = 1|k)p(k|\mathbf{x}) = (1 - \epsilon)y(\mathbf{x}, \mathbf{w}) + \epsilon(1 - y(\mathbf{x}, \mathbf{w})).$$

The conditional probability of the data label is then

$$p(t|\mathbf{x}) = p(t = 1|\mathbf{x})^t(1 - p(t = 1|\mathbf{x})^{1-t}.$$

Forming the likelihood and taking the negative logarithm we then obtain the error function in the form

$$\begin{aligned} E(\mathbf{w}) \quad = \quad & -\sum_{n=1}^{N}\{t_n \ln[(1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) + \epsilon(1 - y(\mathbf{x}_n, \mathbf{w}))] \\ & +(1 - t_n)\ln[1 - (1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) - \epsilon(1 - y(\mathbf{x}_n, \mathbf{w}))]\}. \end{aligned}$$

See also Solution 4.16.

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

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

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

$$\frac{\partial E}{\partial a_n} = -t_n \frac{1}{y_n}\frac{\partial y_n}{\partial a_n} + (1-t_n)\frac{1}{1-y_n}\frac{\partial y_n}{\partial a_n}. \tag{172}$$

From (4.88), we have

$$\frac{\partial y_n}{\partial a_n} = y_n(1-y_n). \tag{173}$$

Substituting (173) into (172), 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.

5.7 See Solution 4.17.

5.8 From (5.59), using standard derivatives, we get

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

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

$$\lambda_i = \mathbf{u}_i^{\mathrm{T}} \mathbf{H} \mathbf{u}_i > 0 \tag{174}$$

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

Conversely, assume that (174) 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 (174). 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.

5.13 From exercise 2.21 we know that a $W \times W$ matrix has $W(W+1)/2$ independent elements. Add to that the W elements of the gradient vector $\mathbf{b}$ and we get

$$\frac{W(W+1)}{2} + W = \frac{W(W+1) + 2W}{2} = \frac{W^2 + 3W}{2} = \frac{W(W+3)}{2}.$$

5.14 We are interested in determining how the correction term

$$\delta = E'(w_{ij}) - \frac{E(w_{ij} + \epsilon) - E(w_{ij} - \epsilon)}{2\epsilon} \tag{175}$$

depend on ϵ.

Using Taylor expansions, we can rewrite the numerator of the first term of (175) as

$$\begin{aligned} &E(w_{ij}) + \epsilon E'(w_{ij}) + \frac{\epsilon^2}{2} E''(w_{ij}) + O(\epsilon^3) \\ &\quad - E(w_{ij}) + \epsilon E'(w_{ij}) - \frac{\epsilon^2}{2} E''(w_{ij}) + O(\epsilon^3) = 2\epsilon E'(w_{ij}) + O(\epsilon^3). \end{aligned}$$

Note that the ϵ^2-terms cancel. Substituting this into (175) we get,

$$\delta = \frac{2\epsilon E'(w_{ij}) + O(\epsilon^3)}{2\epsilon} - E'(w_{ij}) = O(\epsilon^2).$$

5.15 The alternative forward propagation scheme takes the first line of (5.73) as its starting point. However, rather than proceeding with a 'recursive' definition of $\partial y_k / \partial a_j$, we instead make use of a corresponding definition for $\partial a_j / \partial x_i$. More formally

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_j \frac{\partial y_k}{\partial a_j} \frac{\partial a_j}{\partial x_i}$$

where $\partial y_k/\partial a_j$ is defined by (5.75), (5.76) or simply as δ_{kj}, for the case of linear output units. We define $\partial a_j/\partial x_i = w_{ji}$ if a_j is in the first hidden layer and otherwise

$$\frac{\partial a_j}{\partial x_i} = \sum_l \frac{\partial a_j}{\partial a_l}\frac{\partial a_l}{\partial x_i} \tag{176}$$

where

$$\frac{\partial a_j}{\partial a_l} = w_{jl}h'(a_l). \tag{177}$$

Thus we can evaluate J_{ki} by forward propagating $\partial a_j/\partial x_i$, with initial value w_{ij}, alongside a_j, using (176) and (177).

5.16 The multivariate form of (5.82) is

$$E = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}(\mathbf{y}_n - \mathbf{t}_n).$$

The elements of the first and second derivatives then become

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i}$$

and

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \sum_{n=1}^{N}\left\{\frac{\partial \mathbf{y}_n}{\partial w_j}^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i} + (\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial^2 \mathbf{y}_n}{\partial w_j\,\partial w_i}\right\}.$$

As for the univariate case, we again assume that the second term of the second derivative vanishes and we are left with

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{B}_n\mathbf{B}_n^{\mathrm{T}},$$

where $\mathbf{B}_n$ is a $W \times K$ matrix, K being the dimensionality of $\mathbf{y}_n$, with elements

$$(\mathbf{B}_n)_{lk} = \frac{\partial y_{nk}}{\partial w_l}.$$

5.17 Taking the second derivatives of (5.193) with respect to two weights w_r and w_s we obtain

$$\begin{aligned}\frac{\partial^2 E}{\partial w_r \partial w_s} &= \sum_k \int \left\{\frac{\partial y_k}{\partial w_r}\frac{\partial y_k}{\partial w_s}\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad + \sum_k \int \left\{\frac{\partial^2 y_k}{\partial w_r \partial w_s}(y_k(\mathbf{x}) - \mathbb{E}_{t_k}[t_k|\mathbf{x}])\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{178}$$

Using the result (1.89) that the outputs $y_k(\mathbf{x})$ of the trained network represent the conditional averages of the target data, we see that the second term in (178) vanishes. The Hessian is therefore given by an integral of terms involving only the products of first derivatives. For a finite data set, we can write this result in the form

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \frac{1}{N} \sum_{n=1}^{N} \sum_k \frac{\partial y_k^n}{\partial w_r} \frac{\partial y_k^n}{\partial w_s}$$

which is identical with (5.84) up to a scaling factor.

5.18 If we introduce skip layer weights, $\mathbf{U}$, into the model described in Section 5.3.2, this will only affect the last of the forward propagation equations, (5.64), which becomes

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} u_{ki} x_i.$$

Note that there is no need to include the input bias. The derivative w.r.t. u_{ki} can be expressed using the output $\{\delta_k\}$ of (5.65),

$$\frac{\partial E}{\partial u_{ki}} = \delta_k x_i.$$

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

5.20 Using the chain rule, we can write the first derivative of (5.24) as

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\partial E}{\partial a_{nk}} \frac{\partial a_{nk}}{\partial w_i}. \tag{179}$$

From Exercise 5.7, we know that

$$\frac{\partial E}{\partial a_{nk}} = y_{nk} - t_{nk}.$$

Using this and (4.106), we can get the derivative of (179) w.r.t. w_j as

$$\frac{\partial^2 E}{\partial w_i \, \partial w_j} = \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl}) \frac{\partial a_{nk}}{\partial w_i} \frac{\partial a_{nl}}{\partial w_j} + (y_{nk} - t_{nk}) \frac{\partial^2 a_{nk}}{\partial w_i \, \partial w_j} \right).$$

For a trained model, the network outputs will approximate the conditional class probabilities and so the last term inside the parenthesis will vanish in the limit of a large data set, leaving us with

$$(\mathbf{H})_{ij} \simeq \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl}) \frac{\partial a_{nk}}{\partial w_i} \frac{\partial a_{nl}}{\partial w_j}.$$

5.21 **NOTE**: In PRML, the text in the exercise could be misunderstood; a clearer formulation is: "Extend the expression (5.86) for the outer product approximation of the Hessian matrix to the case of $K > 1$ output units. Hence, derive a form that allows (5.87) to be used to incorporate sequentially contributions from individual outputs as well as individual patterns. This, together with the identity (5.88), will allow the use of (5.89) for finding the inverse of the Hessian by sequentially incorporating contributions from individual outputs and patterns."

From (5.44) and (5.46), we see that the multivariate form of (5.82) is

$$E = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} (y_{nk} - t_{nk})^2 .$$

Consequently, the multivariate form of (5.86) is given by

$$\mathbf{H}_{NK} = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{b}_{nk} \mathbf{b}_{nk}^{\mathrm{T}} \tag{180}$$

where $\mathbf{b}_{nk} \equiv \nabla a_{nk} = \nabla y_{nk}$. The double index indicate that we will now iterate over outputs as well as patterns in the sequential build-up of the Hessian. However, in terms of the end result, there is no real need to attribute terms in this sum to specific outputs or specific patterns. Thus, by changing the indexation in (180), we can write it

$$\mathbf{H}_J = \sum_{j=1}^{J} \mathbf{c}_j \mathbf{c}_j^{\mathrm{T}} \tag{181}$$

where $J = NK$ and

$$\begin{aligned} \mathbf{c}_j &= \mathbf{b}_{n(j)k(j)} \\ n(j) &= (j-1) \oslash K + 1 \\ k(j) &= (j-1) \odot K + 1 \end{aligned}$$

with $\oslash$ and $\odot$ denoting integer division and remainder, respectively. The advantage of the indextion in (181) is that now we have a single indexed sum and so we can use (5.87)–(5.89) as they stand, just replacing $\mathbf{b}_L$ with $\mathbf{c}_L$, letting L run from 0 to $J-1$.

5.22 **NOTE**: The first printing of PRML contained typographical errors in equation (5.95). On the r.h.s., $H_{kk'}$ should be $M_{kk'}$. Moreover, the indices j and j' should be swapped on the r.h.s.

Using the chain rule together with (5.48) and (5.92), we have

$$\begin{aligned}\frac{\partial E_n}{\partial w_{kj}^{(2)}} &= \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}^{(2)}} \\ &= \delta_k z_j \end{aligned} \tag{182}$$

Thus,

$$\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\,\partial w_{k'j'}^{(2)}} = \frac{\partial \delta_k z_j}{\partial w_{k'j'}^{(2)}}$$

and since z_j is independent of the second layer weights,

$$\begin{aligned}\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\,\partial w_{k'j'}^{(2)}} &= z_j \frac{\partial \delta_k}{\partial w_{k'j'}^{(2)}} \\ &= z_j \frac{\partial^2 E_n}{\partial a_k\,\partial a_{k'}}\frac{\partial a_k}{\partial w_{k'j'}^{(2)}} \\ &= z_j z_{j'} M_{kk'},\end{aligned}$$

where we again have used the chain rule together with (5.48) and (5.92).

If both weights are in the first layer, we again used the chain rule, this time together with (5.48), (5.55) and (5.56), to get

$$\begin{aligned}\frac{\partial E_n}{\partial w_{ji}^{(1)}} &= \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}^{(1)}} \\ &= x_i \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \\ &= x_i\, h'(a_j) \sum_k w_{kj}^{(2)} \delta_k.\end{aligned}$$

Thus we have

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{j'i'}^{(1)}} = \frac{\partial}{\partial w_{j'i'}}\left(x_i\, h'(a_j) \sum_k w_{kj}^{(2)} \delta_k \right).$$

Now we note that x_i and $w_{kj}^{(2)}$ do not depend on $w_{j'i'}^{(1)}$, while $h'(a_j)$ is only affected in the case where $j = j'$. Using these observations together with (5.48), we get

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{j'i'}^{(1)}} = x_i x_{i'} h''(a_j) I_{jj'} \sum_k w_{kj}^{(2)} \delta_k + x_i h'(a_j) \sum_k w_{kj}^{(2)} \frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}}. \tag{183}$$

From (5.48), (5.55), (5.56), (5.92) and the chain rule, we have

$$\begin{aligned}\frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}} &= \sum_{k'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial a_{j'}} \frac{\partial a_{j'}}{\partial w_{j'i'}^{(1)}} \\ &= x_{i'} h'(a_j) \sum_{k'} w_{k'j'}^{(2)} M_{kk'}. \end{aligned} \tag{184}$$

Substituting this back into (183), we obtain (5.94).

Finally, from (182) we have

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} = \frac{\partial \delta_k z_{j'}}{\partial w_{ji}^{(1)}}.$$

Using (184), we get

$$\begin{aligned}\frac{\partial^2 E_n}{\partial w_{ij}^{(1)} \partial w_{kj'}^{(2)}} &= z_{j'} x_i h'(a_j) \sum_{k'} w_{k'j}^{(2)} M_{kk'} + \delta_k I_{jj'} h'(a_j) x_i \\ &= x_i h'(a_j) \left(\delta_k I_{jj'} + \sum_{k'} w_{k'j}^{(2)} M_{kk'} \right).\end{aligned}$$

5.23 If we introduce skip layer weights into the model discussed in Section 5.4.5, three new cases are added to three already covered in Exercise 5.22.

The first derivative w.r.t. skip layer weight u_{ki} can be written

$$\frac{\partial E_n}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} x_i. \tag{185}$$

Using this, we can consider the first new case, where both weights are in the skip layer,

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki} \partial u_{k'i'}} &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial u_{k'i'}} x_i \\ &= M_{kk'} x_i x_{i'},\end{aligned}$$

where we have also used (5.92).

When one weight is in the skip layer and the other weight is in the hidden-to-output layer, we can use (185), (5.48) and (5.92) to get

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki} \partial w_{k'j}^{(2)}} &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k'j}^{(2)}} x_i \\ &= M_{kk'} z_j x_i.\end{aligned}$$

Finally, if one weight is a skip layer weight and the other is in the input-to-hidden layer, (185), (5.48), (5.55), (5.56) and (5.92) together give

$$\begin{aligned}\frac{\partial^2 E_n}{\partial u_{ki}\,\partial w^{(1)}_{ji'}} &= \frac{\partial}{\partial w^{(1)}_{ji'}}\left(\frac{\partial E_n}{\partial a_k}x_i\right)\\ &= \sum_{k'}\frac{\partial^2 E_n}{\partial a_k\,\partial a_{k'}}\frac{\partial a_{k'}}{\partial w^{(1)}_{ji'}}x_i\\ &= x_i x_{i'} h'(a_j)\sum_{k'} M_{kk'} w^{(2)}_{k'j}.\end{aligned}$$

5.24 With the transformed inputs, weights and biases, (5.113) becomes

$$z_j = h\left(\sum_i \widetilde{w}_{ji}\widetilde{x}_i + \widetilde{w}_{j0}\right).$$

Using (5.115)–(5.117), we can rewrite the argument of $h(\cdot)$ on the r.h.s. as

$$\begin{aligned}&\sum_i \frac{1}{a}w_{ji}(ax_i + b) + w_{j0} - \frac{b}{a}\sum_i w_{ji}\\ &= \sum_i w_{ji}x_i + \frac{b}{a}\sum_i w_{ji} + w_{j0} - \frac{b}{a}\sum_i w_{ji}\\ &= \sum_i w_{ji}x_i + w_{j0}.\end{aligned}$$

Similarly, with the transformed outputs, weights and biases, (5.114) becomes

$$\widetilde{y}_k = \sum_i \widetilde{w}_{kj}z_j + \widetilde{w}_{k0}.$$

Using (5.118)–(5.120), we can rewrite this as

$$\begin{aligned}cy_k + d &= \sum_k cw_{kj}z_j + cw_{k0} + d\\ &= c\left(\sum_i w_{kj}z_j + w_{k0}\right) + d.\end{aligned}$$

By subtracting d and subsequently dividing by c on both sides, we recover (5.114) in its original form.

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

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau)} && (186)\\ &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_j^{\mathrm{T}}\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^{\star}) \\ &= w_j^{(\tau-1)} - \rho\eta_j\mathbf{u}_j^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^{\star}) \\ &= w_j^{(\tau-1)} - \rho\eta_j(w_j^{(\tau-1)} - w_j^{\star}), && (187) \end{aligned}$$

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 (187), 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 (187)

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

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 $\rho\tau$.

5.26 **NOTE**: In PRML, equation (5.201) should read

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_k)^2\Big|_{\mathbf{x}_n}.$$

In this solution, we will indicate dependency on $\mathbf{x}_n$ with a subscript n on relevant symbols.

Substituting the r.h.s. of (5.202) into (5.201) and then using (5.70), we get

$$\Omega_n = \frac{1}{2}\sum_k \left(\sum_i \tau_{ni}\frac{\partial y_{nk}}{\partial x_{ni}}\right)^2 \tag{188}$$

$$= \frac{1}{2}\sum_k \left(\sum_i \tau_{ni} J_{nki}\right)^2 \tag{189}$$

where J_{nki} denoted J_{ki} evaluated at $\mathbf{x}_n$. Summing (189) over n, we get (5.128).

By applying $\mathcal{G}$ from (5.202) to the equations in (5.203) and making use of (5.205) we obtain (5.204). From this, we see that β_{nl} can be written in terms of α_{ni}, which in turn can be written as functions of β_{ni} from the previous layer. For the input layer, using (5.204) and (5.205), we get

$$\begin{aligned} \beta_{nj} &= \sum_i w_{ji}\alpha_{ni} \\ &= \sum_i w_{ji}\mathcal{G}x_{ni} \\ &= \sum_i w_{ji}\sum_{i'}\tau_{ni'}\frac{\partial x_{ni}}{\partial x_{ni'}} \\ &= \sum_i w_{ji}\tau_{ni}. \end{aligned} \tag{190}$$

Thus we see that, starting from (190), $\boldsymbol{\tau}_n$ is propagated forward by subsequent application of the equations in (5.204), yielding the β_{nl} for the output layer, from which Ω_n can be computed using (5.201),

$$\Omega_n = \frac{1}{2}\sum_k \left(\mathcal{G}y_{nk}\right)^2 = \frac{1}{2}\sum_k \alpha_{nk}^2.$$

Considering $\partial\Omega_n/\partial w_{rs}$, we start from (5.201) and make use of the chain rule, together with (5.52), (5.205) and (5.207), to obtain

$$\begin{aligned}\frac{\partial \Omega_n}{\partial w_{rs}} &= \sum_k \left(\mathcal{G}y_{nk}\right)\mathcal{G}\left(\delta_{nkr}z_{ns}\right) \\ &= \sum_k \alpha_{nk}\left(\phi_{nkr}z_{ns} + \delta_{nkr}\alpha_{ns}\right).\end{aligned}$$

The backpropagation formula for computing δ_{nkr} follows from (5.74), which is used in computing the Jacobian matrix, and is given by

$$\delta_{nkr} = h'(a_{nr})\sum_l w_{lr}\delta_{nkl}.$$

Using this together with (5.205) and (5.207), we can obtain backpropagation equations for ϕ_{nkr},

$$\begin{aligned}\phi_{nkr} &= \mathcal{G}\delta_{nkr} \\ &= \mathcal{G}\left(h'(a_{nr})\sum_l w_{lr}\delta_{nkl}\right) \\ &= h''(a_{nr})\beta_{nr}\sum_l w_{lr}\delta_{nkl} + h'(a_{nr})\sum_l w_{lr}\phi_{nkl}.\end{aligned}$$

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).

5.30 Since the μ_js only appear in the regularization term, $\Omega(\mathbf{w})$, from (5.139) we have

$$\frac{\partial \widetilde{E}}{\partial \mu_j} = \lambda \frac{\partial \Omega}{\partial \mu_j}. \tag{191}$$

Using (2.42), (5.138) and (5.140) and standard rules for differentiation, we can calculate the derivative of $\Omega(\mathbf{w})$ as follows:

$$\begin{aligned} \frac{\partial \Omega}{\partial \mu_j} &= -\sum_i \frac{1}{\sum_{j'} \pi_{j'} \mathcal{N}\left(w_i|\mu_{j'}, \sigma_{j'}^2\right)} \pi_j \mathcal{N}\left(w_i|\mu_j, \sigma_j^2\right) \frac{w_i - \mu_j}{\sigma_j^2} \\ &= -\sum_i \gamma_j(\mathbf{w}_i) \frac{w_i - \mu_j}{\sigma_j^2}. \end{aligned}$$

Combining this with (191), we get (5.142).

5.31 Following the same line of argument as in Solution 5.30, we need the derivative of $\Omega(\mathbf{w})$ w.r.t. σ_j. Again using (2.42), (5.138) and (5.140) and standard rules for differentiation, we find this to be

$$\begin{aligned} \frac{\partial \Omega}{\partial \sigma_j} &= -\sum_i \frac{1}{\sum_{j'} \pi_{j'} \mathcal{N}\left(w_i|\mu_{j'}, \sigma_{j'}^2\right)} \pi_j \frac{1}{(2\pi)^{1/2}} \left\{ -\frac{1}{\sigma_j^2} \exp\left(-\frac{(w_i - \mu_j)^2}{2\sigma_j^2}\right) \right. \\ &\qquad \left. + \frac{1}{\sigma_j} \exp\left(-\frac{(w_i - \mu_j)^2}{2\sigma_j^2}\right) \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\} \\ &= \sum_i \gamma_j(w_i) \left\{ \frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right\}. \end{aligned}$$

Combining this with (191), we get (5.143).

5.32 **NOTE**: In the first printing of PRML, there is a leading λ missing on the r.h.s. of equation (5.147). Moreover, in the text of the exercise (last line), the equation of the constraint to be used should read "$\sum_k \gamma_k(w_i) = 1$ for all i".

Equation (5.208) follows from (5.146) in exactly the same way that (4.106) follows from (4.104) in Solution 4.17.

Just as in Solutions 5.30 and 5.31, η_j only affect $\widetilde{E}$ through $\Omega(\mathbf{w})$. However, η_j will affect π_k for all values of k (not just $j = k$). Thus we have

$$\frac{\partial \Omega}{\partial \eta_j} = \sum_k \frac{\partial \Omega}{\partial \pi_k} \frac{\partial \pi_k}{\partial \eta_j}. \tag{192}$$

From (5.138) and (5.140), we get

$$\frac{\partial \Omega}{\partial \pi_k} = -\sum_i \frac{\gamma_k(w_i)}{\pi_k}.$$

Substituting this and (5.208) into (192) yields

$$\begin{aligned} \frac{\partial \Omega}{\partial \eta_j} &= \frac{\partial \widetilde{E}}{\partial \eta_j} = -\sum_k \sum_i \frac{\gamma_k(w_i)}{\pi_k} \{\delta_{jk}\pi_j - \pi_j \pi_k\} \\ &= \sum_i \{\pi_j - \gamma_j(w_i)\}, \end{aligned}$$

where we have used the fact that $\sum_k \gamma_k(w_i) = 1$ for all i.

5.33 From standard trigometric rules we get the position of the end of the first arm,

$$\left(x_1^{(1)}, x_2^{(1)}\right) = (L_1 \cos(\theta_1), L_1 \sin(\theta_1)).$$

Similarly, the position of the end of the second arm relative to the end of the first arm is given by the corresponding equation, with an angle offset of π (see Figure 5.18), which equals a change of sign

$$\begin{aligned} \left(x_1^{(2)}, x_2^{(2)}\right) &= (L_2 \cos(\theta_1 + \theta_2 - \pi), L_1 \sin(\theta_1 + \theta_2 - \pi)) \\ &= -(L_2 \cos(\theta_1 + \theta_2), L_2 \sin(\theta_1 + \theta_2)). \end{aligned}$$

Putting this together, we must also taken into account that θ_2 is measured relative to the first arm and so we get the position of the end of the second arm relative to the attachment point of the first arm as

$$(x_1, x_2) = (L_1 \cos(\theta_1) - L_2 \cos(\theta_1 + \theta_2), L_1 \sin(\theta_1) - L_2 \sin(\theta_1 + \theta_2)).$$

5.34 **NOTE**: In 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

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{193}$$

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 (193), standard derivatives applied to the n^{th} term of (5.153) gives

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{194}$$

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

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j (I_{jk} - \pi_k). \tag{195}$$

Combining (193), (194) and (195), 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.

5.35 **NOTE**: See Solution 5.34.

From (5.152) we have

$$a_{kl}^\mu = \mu_{kl}$$

and thus

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = \frac{\partial E_n}{\partial \mu_{kl}}.$$

From (2.43), (5.153) and (5.154), we get

$$\begin{aligned} \frac{\partial E_n}{\partial \mu_{kl}} &= -\frac{\pi_k \mathcal{N}_{nk}}{\sum_{k'} \pi_{k'} \mathcal{N}_{nk'}} \frac{t_{nl} - \mu_{kl}}{\sigma_k^2(\mathbf{x}_n)} \\ &= \gamma_{nk}(\mathbf{t}_n|\mathbf{x}_n) \frac{\mu_{kl} - t_{nl}}{\sigma_k^2(\mathbf{x}_n)}. \end{aligned}$$

5.36 **NOTE**: In PRML, equation (5.157) is incorrect and the correct equation appears at the end of this solution ; see also Solution 5.34.

From (5.151) and (5.153), we see that

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \sigma_k} \frac{\partial \sigma_k}{\partial a_k^\sigma}, \tag{196}$$

where, from (5.151),

$$\frac{\partial \sigma_k}{\partial a_k^\sigma} = \sigma_k. \tag{197}$$

From (2.43), (5.153) and (5.154), we get

$$\begin{aligned}
\frac{\partial E_n}{\partial \sigma_k} &= -\frac{1}{\sum_{k'} \mathcal{N}_{nk'}} \left(\frac{L}{2\pi}\right)^{L/2} \left\{ -\frac{L}{\sigma^{L+1}} \exp\left(-\frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right) \right. \\
&\qquad \left. + \frac{1}{\sigma^L} \exp\left(-\frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right) \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3} \right\} \\
&= \gamma_{nk} \left(\frac{L}{\sigma_k} - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3}\right).
\end{aligned}$$

Combining this with (196) and (197), we get

$$\frac{\partial E_n}{\partial a_k^\sigma} = \gamma_{nk} \left(L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right).$$

5.37 From (2.59) and (5.148) we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{t}|\mathbf{x}\right] &= \int \mathbf{t} p\left(\mathbf{t}|\mathbf{x}\right) \, \mathrm{d}\mathbf{t} \\
&= \int \mathbf{t} \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right) \, \mathrm{d}\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k(\mathbf{x}) \int \mathbf{t} \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right) \, \mathrm{d}\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}).
\end{aligned}$$

We now introduce the shorthand notation

$$\bar{\mathbf{t}}_k = \boldsymbol{\mu}_k(\mathbf{x}) \quad \text{and} \quad \bar{\mathbf{t}} = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \bar{\mathbf{t}}_k.$$

Using this together with (2.59), (2.62), (5.148) and (5.158), we get

$$\begin{aligned} s^2(\mathbf{x}) &= \mathbb{E}\left[\|\mathbf{t}-\mathbb{E}\left[\mathbf{t}|\mathbf{x}\right]\|^2|\mathbf{x}\right] = \int \|\mathbf{t}-\overline{\mathbf{t}}\|^2 p\left(\mathbf{t}|\mathbf{x}\right)\,\mathrm{d}\mathbf{t} \\ &= \int \left(\mathbf{t}^{\mathrm{T}}\mathbf{t}-\mathbf{t}^{\mathrm{T}}\overline{\mathbf{t}}-\overline{\mathbf{t}}^{\mathrm{T}}\mathbf{t}+\overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}\right)\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}),\sigma_k^2(\mathbf{x})\right)\,\mathrm{d}\mathbf{t} \\ &= \sum_{k=1}^{K}\pi_k(\mathbf{x})\left\{\sigma_k^2+\overline{\mathbf{t}}_k^{\mathrm{T}}\overline{\mathbf{t}}_k-\overline{\mathbf{t}}_k^{\mathrm{T}}\overline{\mathbf{t}}-\overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}_k+\overline{\mathbf{t}}^{\mathrm{T}}\overline{\mathbf{t}}\right\} \\ &= \sum_{k=1}^{K}\pi_k(\mathbf{x})\left\{\sigma_k^2+\|\overline{\mathbf{t}}_k-\overline{\mathbf{t}}\|^2\right\} \\ &= \sum_{k=1}^{K}\pi_k(\mathbf{x})\left\{\sigma_k^2+\left\|\boldsymbol{\mu}_k(\mathbf{x})-\sum_{l}^{K}\pi_l\boldsymbol{\mu}_l(\mathbf{x})\right\|^2\right\}. \end{aligned}$$

5.38 Making the following substitions from the r.h.s. of (5.167) and (5.171),

$$\mathbf{x}\Rightarrow\mathbf{w}\quad\boldsymbol{\mu}\Rightarrow\mathbf{w}_{\mathrm{MAP}}\quad\boldsymbol{\Lambda}^{-1}\Rightarrow\mathbf{A}^{-1}$$

$$\mathbf{y}\Rightarrow t\quad\mathbf{A}\Rightarrow\mathbf{g}^{\mathrm{T}}\quad\mathbf{b}\Rightarrow y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}})-\mathbf{g}^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}}\quad\mathbf{L}^{-1}\Rightarrow\beta^{-1},$$

in (2.113) and (2.114), (2.115) becomes

$$\begin{aligned} p(t) &= \mathcal{N}\left(t|\mathbf{g}^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}}+y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}})-\mathbf{g}^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}},\beta^{-1}+\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}\right) \\ &= \mathcal{N}\left(t|y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}}),\sigma^2\right), \end{aligned}$$

where σ^2 is defined by (5.173).

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

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

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

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

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.

5.41 **NOTE**: In PRML, the final "const" term in Equation (5.183) should be ommitted.

This solutions is similar to Solution 5.39, with the difference that the log-likelihood term is now given by (5.181). Again using (4.135), the corresponding approximation of the marginal likelihood becomes

$$p(\mathcal{D}|\alpha) \simeq p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}})p(\mathbf{w}_{\mathrm{MAP}}|\alpha) \int \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})\right) \mathrm{d}\mathbf{w}, \quad (198)$$

where now

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\mathbf{w}) = \mathbf{H} + \alpha\mathbf{I}.$$

Performing the integral in (198) using (4.135) and then taking the logarithm on, we get (5.183).

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 $\mathbf{a} = \mathbf{a}_\parallel$ where $\mathbf{a}_\parallel$ lies in the span of $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$ and hence can be written as a linear combination of the columns of $\boldsymbol{\Phi}$, so that in component notation

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

or equivalently in vector notation

$$\mathbf{a} = \boldsymbol{\Phi}\mathbf{u}. \tag{199}$$

Substituting (199) into (6.7) we obtain

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2}\left(\mathbf{K}\boldsymbol{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\mathbf{K}\boldsymbol{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{K}\boldsymbol{\Phi}\mathbf{u} \\ &= \frac{1}{2}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{u} - \mathbf{t}\right)^{\mathrm{T}}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{u} - \mathbf{t}\right) + \frac{\lambda}{2}\mathbf{u}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{u} \end{aligned} \tag{200}$$

Since the matrix $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$ has full rank we can define an equivalent parametrization given by

$$\mathbf{w} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{u}$$

and substituting this into (200) we recover the original regularized error function (6.2).

6.2 Starting with an initial weight vector $\mathbf{w} = \mathbf{0}$ the Perceptron learning algorithm increments $\mathbf{w}$ with vectors $t_n\boldsymbol{\phi}(\mathbf{x}_n)$ where n indexes a pattern which is misclassified by the current model. The resulting weight vector therefore comprises a linear combination of vectors of the form $t_n\boldsymbol{\phi}(\mathbf{x}_n)$ which we can represent in the form

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(\mathbf{x}_n) \tag{201}$$

where α_n is an integer specifying the number of times that pattern n was used to update $\mathbf{w}$ during training. The corresponding predictions made by the trained Perceptron are therefore given by

$$\begin{aligned} y(\mathbf{x}) &= \mathrm{sign}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right) \\ &= \mathrm{sign}\left(\sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right) \\ &= \mathrm{sign}\left(\sum_{n=1}^{N} \alpha_n t_n k(\mathbf{x}_n, \mathbf{x})\right). \end{aligned}$$

Thus the predictive function of the Perceptron has been expressed purely in terms of the kernel function. The learning algorithm of the Perceptron can similarly be written as

$$\alpha_n \to \alpha_n + 1$$

for patterns which are misclassified, in other words patterns which satisfy

$$t_n \left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right) \geqslant 0.$$

Using (201) together with $\alpha_n \geqslant 0$, this can be written in terms of the kernel function in the form

$$t_n \left(\sum_{m=1}^{N} k(\mathbf{x}_m, \mathbf{x}_n) \right) \geqslant 0$$

and so the learning algorithm depends only on the elements of the Gram matrix.

6.3 The distance criterion for the nearest neighbour classifier can be expressed in terms of the kernel as follows

$$\begin{aligned} D(\mathbf{x}, \mathbf{x}_n) &= \|\mathbf{x} - \mathbf{x}_n\|^2 \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n - 2\mathbf{x}^{\mathrm{T}}\mathbf{x}_n \\ &= k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}_n, \mathbf{x}_n) - 2k(\mathbf{x}, \mathbf{x}_n) \end{aligned}$$

where $k(\mathbf{x}, \mathbf{x}_n) = \mathbf{x}^{\mathrm{T}}\mathbf{x}_n$. We then obtain a non-linear kernel classifier by replacing the linear kernel with some other choice of kernel function.

6.4 An example of such a matrix is

$$\begin{pmatrix} 2 & -2 \\ -3 & 4 \end{pmatrix}.$$

We can verify this by calculating the determinant of

$$\begin{pmatrix} 2-\lambda & -2 \\ -3 & 4-\lambda \end{pmatrix},$$

setting the resulting expression equal to zero and solve for the eigenvalues λ, yielding

$$\lambda_1 \simeq 5.65 \quad \text{and} \quad \lambda_2 \simeq 0.35,$$

which are both positive.

6.5 The results (6.13) and (6.14) are easily proved by using (6.1) which defines the kernel in terms of the scalar product between the feature vectors for two input vectors. If $k_1(\mathbf{x}, \mathbf{x}')$ is a valid kernel then there must exist a feature vector $\boldsymbol{\phi}(\mathbf{x})$ such that

$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}').$$

It follows that

$$ck_1(\mathbf{x}, \mathbf{x}') = \mathbf{u}(\mathbf{x})^{\mathrm{T}} \mathbf{u}(\mathbf{x}')$$

where

$$\mathbf{u}(\mathbf{x}) = c^{1/2} \boldsymbol{\phi}(\mathbf{x})$$

and so $ck_1(\mathbf{x}, \mathbf{x}')$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

Similarly, for (6.14) we can write

$$f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') = \mathbf{v}(\mathbf{x})^{\mathrm{T}}\mathbf{v}(\mathbf{x}')$$

where we have defined

$$\mathbf{v}(\mathbf{x}) = f(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}).$$

Again, we see that $f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

Alternatively, these results can be proved be appealing to the general result that the Gram matrix, $\mathbf{K}$, whose elements are given by $k(\mathbf{x}_n, \mathbf{x}_m)$, should be positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$, by following a similar argument to Solution 6.7 below.

6.6 Equation (6.15) follows from (6.13), (6.17) and (6.18).

For (6.16), we express the exponential as a power series, yielding

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \exp(k_1(\mathbf{x}, \mathbf{x}')) \\ &= \sum_{m=0}^{\infty} \frac{(k_1(\mathbf{x}, \mathbf{x}'))^m}{m!}. \end{aligned}$$

Since this is a polynomial in $k_1(\mathbf{x}, \mathbf{x}')$ with positive coefficients, (6.16) follows from (6.15).

6.7 (6.17) is most easily proved by making use of the result, discussed on page 295, that a necessary and sufficient condition for a function $k(\mathbf{x}, \mathbf{x}')$ to be a valid kernel is that the Gram matrix $\mathbf{K}$, whose elements are given by $k(\mathbf{x}_n, \mathbf{x}_m)$, should be positive semidefinite for all possible choices of the set $\{\mathbf{x}_n\}$. A matrix $\mathbf{K}$ is positive semidefinite if, and only if,

$$\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \geqslant 0$$

for any choice of the vector $\mathbf{a}$. Let $\mathbf{K}_1$ be the Gram matrix for $k_1(\mathbf{x}, \mathbf{x}')$ and let $\mathbf{K}_2$ be the Gram matrix for $k_2(\mathbf{x}, \mathbf{x}')$. Then

$$\mathbf{a}^{\mathrm{T}}(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{K}_1\mathbf{a} + \mathbf{a}^{\mathrm{T}}\mathbf{K}_2\mathbf{a} \geqslant 0$$

where we have used the fact that $\mathbf{K}_1$ and $\mathbf{K}_2$ are positive semi-definite matrices, together with the fact that the sum of two non-negative numbers will itself be non-negative. Thus, (6.17) defines a valid kernel.

To prove (6.18), we take the approach adopted in Solution 6.5. Since we know that $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ are valid kernels, we know that there exist mappings $\boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}(\mathbf{x})$ such that

$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}') \qquad \text{and} \qquad k_2(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}').$$

Hence

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \\ &= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')\,\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}') \\ &= \sum_{m=1}^{M}\phi_m(\mathbf{x})\phi_m(\mathbf{x}')\sum_{n=1}^{N}\psi_n(\mathbf{x})\psi_n(\mathbf{x}') \\ &= \sum_{m=1}^{M}\sum_{n=1}^{N}\phi_m(\mathbf{x})\phi_m(\mathbf{x}')\psi_n(\mathbf{x})\psi_n(\mathbf{x}') \\ &= \sum_{k=1}^{K}\varphi_k(\mathbf{x})\varphi_k(\mathbf{x}') \\ &= \boldsymbol{\varphi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}'), \end{aligned}$$

where $K = MN$ and

$$\varphi_k(\mathbf{x}) = \phi_{((k-1)\oslash N)+1}(\mathbf{x})\psi_{((k-1)\odot N)+1}(\mathbf{x}),$$

where in turn $\oslash$ and $\odot$ denote integer division and remainder, respectively.

6.8 If we consider the Gram matrix, $\mathbf{K}$, corresponding to the l.h.s. of (6.19), we have

$$(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = k_3\left(\boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j)\right) = (\mathbf{K}_3)_{ij}$$

where $\mathbf{K}_3$ is the Gram matrix corresponding to $k_3(\cdot,\cdot)$. Since $k_3(\cdot,\cdot)$ is a valid kernel,

$$\mathbf{u}^{\mathrm{T}}\mathbf{K}\mathbf{u} = \mathbf{u}^{\mathrm{T}}\mathbf{K}_3\mathbf{u} \geqslant 0.$$

For (6.20), let $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}$, so that $(\mathbf{K})_{ij} = \mathbf{x}_i^{\mathrm{T}}\mathbf{A}\mathbf{x}_j$, and consider

$$\begin{aligned} \mathbf{u}^{\mathrm{T}}\mathbf{K}\mathbf{u} &= \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}\mathbf{u} \\ &= \mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v} \geqslant 0 \end{aligned}$$

where , $\mathbf{v} = \mathbf{X}\mathbf{u}$ and we have used that $\mathbf{A}$ is positive semidefinite.

6.9 Equations (6.21) and (6.22) are special cases of (6.17) and (6.18), respectively, where $k_a(\cdot,\cdot)$ and $k_b(\cdot,\cdot)$ only depend on particular elements in their argument vectors. Thus (6.21) and (6.22) follow from the more general results.

6.10 Any solution of a linear learning machine based on this kernel must take the form

$$y(\mathbf{x}) = \sum_{n=1}^{N}\alpha_n k(\mathbf{x}_n, \mathbf{x}) = \left(\sum_{n=1}^{N}\alpha_n f(\mathbf{x}_n)\right) f(\mathbf{x}) = Cf(\mathbf{x}).$$

6.11 As discussed in Solution 6.6, the exponential kernel (6.16) can be written as an infinite sum of terms, each of which can itself be written as an inner product of feature vectors, according to (6.15). Thus, by concatenating the feature vectors of the indvidual terms in that sum, we can write this as an inner product of infinite dimension feature vectors. More formally,

$$\begin{aligned}\exp\left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'/\sigma^2\right) &= \sum_{m=0}^{\infty}\boldsymbol{\phi}_m(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}_0(\mathbf{x}') \\ &= \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}')\end{aligned}$$

where $\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}} = \left[\boldsymbol{\phi}_0(\mathbf{x})^{\mathrm{T}}, \boldsymbol{\phi}_1(\mathbf{x})^{\mathrm{T}}, \ldots\right]$. Hence, we can write (6.23) as

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\varphi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\varphi}(\mathbf{x}')$$

where

$$\boldsymbol{\varphi}(\mathbf{x}) = \exp\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}}{\sigma^2}\right)\boldsymbol{\psi}(\mathbf{x}).$$

6.12 **NOTE**: In PRML, there is an error in the text relating to this exercise. Immediately following (6.27), it says: $|A|$ denotes the number of *subsets* in A; it should have said: $|A|$ denotes the number of *elements* in A.

Since A may be equal to D (the subset relation was not defined to be strict), $\boldsymbol{\phi}(D)$ must be defined. This will map to a vector of $2^{|D|}$ 1s, one for each possible subset of D, including D itself as well as the empty set. For $A \subset D$, $\boldsymbol{\phi}(A)$ will have 1s in all positions that correspond to subsets of A and 0s in all other positions. Therefore, $\boldsymbol{\phi}(A_1)^{\mathrm{T}}\boldsymbol{\phi}(A_2)$ will count the number of subsets shared by A_1 and A_2. However, this can just as well be obtained by counting the number of elements in the intersection of A_1 and A_2, and then raising 2 to this number, which is exactly what (6.27) does.

6.13 In the case of the transformed parameter $\boldsymbol{\psi}(\boldsymbol{\theta})$, we have

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{M}\mathbf{g}_{\boldsymbol{\psi}} \tag{202}$$

where $\mathbf{M}$ is a matrix with elements

$$M_{ij} = \frac{\partial\psi_i}{\partial\theta_j}$$

(recall that $\boldsymbol{\psi}(\boldsymbol{\theta})$ is assumed to be differentiable) and

$$\mathbf{g}_{\boldsymbol{\psi}} = \nabla_{\boldsymbol{\psi}} \ln p\left(\mathbf{x}|\boldsymbol{\psi}(\boldsymbol{\theta})\right).$$

The Fisher information matrix then becomes

$$\begin{aligned}\mathbf{F} &= \mathbb{E}_{\mathbf{x}}\left[\mathbf{M}\mathbf{g}_{\boldsymbol{\psi}}\mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\right] \\ &= \mathbf{M}\mathbb{E}_{\mathbf{x}}\left[\mathbf{g}_{\boldsymbol{\psi}}\mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}}\right]\mathbf{M}^{\mathrm{T}}.\end{aligned} \tag{203}$$

Substituting (202) and (203) into (6.33), we get

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \mathbf{M}^{\mathrm{T}} \left(\mathbf{M} \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}_{\boldsymbol{\psi}} \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \right] \mathbf{M}^{\mathrm{T}} \right)^{-1} \mathbf{M} \mathbf{g}_{\boldsymbol{\psi}} \\ &= \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \mathbf{M}^{\mathrm{T}} \left(\mathbf{M}^{\mathrm{T}} \right)^{-1} \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}_{\boldsymbol{\psi}} \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \right]^{-1} \mathbf{M}^{-1} \mathbf{M} \mathbf{g}_{\boldsymbol{\psi}} \\ &= \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}_{\boldsymbol{\psi}} \mathbf{g}_{\boldsymbol{\psi}}^{\mathrm{T}} \right]^{-1} \mathbf{g}_{\boldsymbol{\psi}}, \end{aligned} \tag{204}$$

where we have used (C.3) and the fact that $\boldsymbol{\psi}(\boldsymbol{\theta})$ is assumed to be invertible. Since $\boldsymbol{\theta}$ was simply replaced by $\boldsymbol{\psi}(\boldsymbol{\theta})$, (204) corresponds to the original form of (6.33).

6.14 In order to evaluate the Fisher kernel for the Gaussian we first note that the covariance is assumed to be fixed, and hence the parameters comprise only the elements of the mean $\boldsymbol{\mu}$. The first step is to evaluate the Fisher score defined by (6.32). From the definition (2.43) of the Gaussian we have

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = \nabla_{\boldsymbol{\mu}} \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) = \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}).$$

Next we evaluate the Fisher information matrix using the definition (6.34), giving

$$\mathbf{F} = \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) \mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^{\mathrm{T}} \right] = \mathbf{S}^{-1} \mathbb{E}_{\mathbf{x}} \left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \right] \mathbf{S}^{-1}.$$

Here the expectation is with respect to the original Gaussian distribution, and so we can use the standard result

$$\mathbb{E}_{\mathbf{x}} \left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \right] = \mathbf{S}$$

from which we obtain

$$\mathbf{F} = \mathbf{S}^{-1}.$$

Thus the Fisher kernel is given by

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}' - \boldsymbol{\mu}),$$

which we note is just the squared Mahalanobis distance.

6.15 The determinant for the 2×2 Gram matrix

$$\begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_2, x_1) & k(x_2, x_2) \end{pmatrix}$$

equals

$$k(x_1, x_1) k(x_2, x_2) - k(x_1, x_2)^2,$$

where we have used the fact that $k(x_1, x_2) = k(x_2, x_1)$. Then (6.96) follows directly from the fact that this determinant must be non-negative for a positive semidefinite matrix.

6.16 **NOTE**: In PRML, a detail is missing in this exercise; the text "where $\mathbf{w}_{\perp}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) = 0$ for all n," should be inserted at the beginning of the line immediately following equation (6.98).

We start by rewriting (6.98) as

$$\mathbf{w} = \mathbf{w}_{\|} + \mathbf{w}_{\perp} \tag{205}$$

where

$$\mathbf{w}_{\|} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\mathbf{x}_n).$$

Note that since $\mathbf{w}_{\perp}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) = 0$ for all n,

$$\mathbf{w}_{\perp}^{\mathrm{T}} \mathbf{w}_{\|} = 0. \tag{206}$$

Using (205) and (206) together with the fact that $\mathbf{w}_{\perp}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) = 0$ for all n, we can rewrite (6.97) as

$$\begin{aligned} J(\mathbf{w}) &= f\left((\mathbf{w}_{\|} + \mathbf{w}_{\perp})^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_1), \ldots, (\mathbf{w}_{\|} + \mathbf{w}_{\perp})^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_N)\right) \\ &\quad + g\left((\mathbf{w}_{\|} + \mathbf{w}_{\perp})^{\mathrm{T}} (\mathbf{w}_{\|} + \mathbf{w}_{\perp})\right) \\ &= f\left(\mathbf{w}_{\|}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_1), \ldots, \mathbf{w}_{\|}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_N)\right) + g\left(\mathbf{w}_{\|}^{\mathrm{T}} \mathbf{w}_{\|} + \mathbf{w}_{\perp}^{\mathrm{T}} \mathbf{w}_{\perp}\right). \end{aligned}$$

Since $g(\cdot)$ is monotonically increasing, it will have its minimum w.r.t. $\mathbf{w}_{\perp}$ at $\mathbf{w}_{\perp} = \mathbf{0}$, in which case

$$\mathbf{w} = \mathbf{w}_{\|} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\mathbf{x}_n)$$

as desired.

6.17 **NOTE**: In PRML, there are typographical errors in the text relating to this exercise. In the sentence following immediately after (6.39), $f(\mathbf{x})$ should be replaced by $y(\mathbf{x})$. Also, on the l.h.s. of (6.40), $y(\mathbf{x}_n)$ should be replaced by $y(\mathbf{x})$. There were also errors in Appendix D, which might cause confusion; please consult the errata on the PRML website.

Following the discussion in Appendix D we give a first-principles derivation of the solution. First consider a variation in the function $y(\mathbf{x})$ of the form

$$y(\mathbf{x}) \to y(\mathbf{x}) + \epsilon \eta(\mathbf{x}).$$

Substituting into (6.39) we obtain

$$E[y + \epsilon\eta] = \frac{1}{2} \sum_{n=1}^{N} \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) + \epsilon\eta(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}^2 \nu(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}.$$

Now we expand in powers of ϵ and set the coefficient of ϵ, which corresponds to the functional first derivative, equal to zero, giving

$$\sum_{n=1}^{N} \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\} \eta(\mathbf{x}_n + \boldsymbol{\xi}) \nu(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} = 0. \tag{207}$$

This must hold for every choice of the variation function $\eta(\mathbf{x})$. Thus we can choose

$$\eta(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{z})$$

where $\delta(\,\cdot\,)$ is the Dirac delta function. This allows us to evaluate the integral over $\boldsymbol{\xi}$ giving

$$\sum_{n=1}^{N} \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}\, \delta(\mathbf{x}_n + \boldsymbol{\xi} - \mathbf{z}) \nu(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} = \sum_{n=1}^{N} \{y(\mathbf{z}) - t_n\}\, \nu(\mathbf{z} - \mathbf{x}_n).$$

Substituting this back into (207) and rearranging we then obtain the required result (6.40).

6.18 From the product rule we have

$$p(t|x) = \frac{p(t, x)}{p(x)}.$$

With $p(t, x)$ given by (6.42) and

$$f(x - x_n, t - t_n) = \mathcal{N}\left([x - x_n, t - t_n]^{\mathrm{T}} | \mathbf{0}, \sigma^2 \mathbf{I}\right)$$

this becomes

$$\begin{aligned} p(t|x) &= \frac{\sum_{n=1}^{N} \mathcal{N}\left([x - x_n, t - t_n]^{\mathrm{T}} | \mathbf{0}, \sigma^2 \mathbf{I}\right)}{\int \sum_{m=1}^{N} \mathcal{N}\left([x - x_m, t - t_m]^{\mathrm{T}} | \mathbf{0}, \sigma^2 \mathbf{I}\right)\, \mathrm{d}t} \\ &= \frac{\sum_{n=1}^{N} \mathcal{N}\left(x - x_n | 0, \sigma^2\right) \mathcal{N}\left(t - t_n | 0, \sigma^2\right)}{\sum_{m=1}^{N} \mathcal{N}\left(x - x_m | 0, \sigma^2\right)}. \end{aligned}$$

From (6.46), (6.47), the definition of $f(x, t)$ and the properties of the Gaussian distribution, we can rewrite this as

$$\begin{aligned} p(t|x) &= \sum_{n=1}^{N} k(x, x_n) \mathcal{N}\left(t - t_n | 0, \sigma^2\right) \\ &= \sum_{n=1}^{N} k(x, x_n) \mathcal{N}\left(t | t_n, \sigma^2\right) \end{aligned} \tag{208}$$

where

$$k(x, x_n) = \frac{\mathcal{N}\left(x - x_n | 0, \sigma^2\right)}{\sum_{m=1}^{N} \mathcal{N}\left(x - x_m | 0, \sigma^2\right)}.$$

We see that this a Gaussian mixture model where $k(x, x_n)$ play the role of input dependent mixing coefficients.

Using (208) it is straightforward to calculate various expectations:

$$\begin{aligned}\mathbb{E}[t|x] &= \int t\, p(t|x)\, \mathrm{d}t \\ &= \int t \sum_{n=1}^{N} k(x, x_n) \mathcal{N}\left(t|t_n, \sigma^2\right)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n) \int t\, \mathcal{N}\left(t|t_n, \sigma^2\right)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n)\, t_n\end{aligned}$$

and

$$\begin{aligned}\mathrm{var}[t|x] &= \mathbb{E}\left[(t - \mathbb{E}[t|x])^2\right] \\ &= \int (t - \mathbb{E}[t|x])^2\, p(t|x)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n) \int (t - \mathbb{E}[t|x])^2\, \mathcal{N}\left(t|t_n, \sigma^2\right)\, \mathrm{d}t \\ &= \sum_{n=1}^{N} k(x, x_n) \left(\sigma^2 + t_n^2 - 2t_n \mathbb{E}[t|x] + \mathbb{E}[t|x]^2\right) \\ &= \sigma^2 - \mathbb{E}[t|x]^2 + \sum_{n=1}^{N} k(x, x_n)\, t_n^2.\end{aligned}$$

6.19 Changing variables to $\mathbf{z}_n = \mathbf{x}_n - \boldsymbol{\xi}_n$ we obtain

$$E = \frac{1}{2} \sum_{n=1}^{N} \int [y(\mathbf{z}_n) - t_n]^2\, g(\mathbf{x}_n - \mathbf{z}_n)\, \mathrm{d}\mathbf{z}_n.$$

If we set the functional derivative of E with respect to the function $y(\mathbf{x})$, for some general value of $\mathbf{x}$, to zero using the calculus of variations (see Appendix D) we have

$$\begin{aligned}\frac{\delta E}{\delta \mathbf{y}(\mathbf{x})} &= \sum_{n=1}^{N} \int [y(\mathbf{z}_n) - t_n]\, g(\mathbf{x}_n - \mathbf{z}_n) \delta(\mathbf{x} - \mathbf{z}_n)\, \mathrm{d}\mathbf{z}_n \\ &= \sum_{n=1}^{N} [y(\mathbf{x}) - t_n]\, g(\mathbf{x}_n - \mathbf{x}) = 0.\end{aligned}$$

Solving for $y(\mathbf{x})$ we obtain

$$y(\mathbf{x}) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n \tag{209}$$

where we have defined

$$k(\mathbf{x}, \mathbf{x}_n) = \frac{g(\mathbf{x}_n - \mathbf{x})}{\sum_n g(\mathbf{x}_n - \mathbf{x})}.$$

This an expansion in kernel functions, where the kernels satisfy the summation constraint $\sum_n k(\mathbf{x}, \mathbf{x}_n) = 1$.

6.20 Given the joint distribution (6.64), we can identify t_{N+1} with $\mathbf{x}_a$ and $\mathbf{t}$ with $\mathbf{x}_b$ in (2.65). Note that this means that we are prepending rather than appending t_{N+1} to $\mathbf{t}$ and $\mathbf{C}_{N+1}$ therefore gets redefined as

$$\mathbf{C}_{N+1} = \begin{pmatrix} c & \mathbf{k}^{\mathrm{T}} \\ \mathbf{k} & \mathbf{C}_N \end{pmatrix}.$$

It then follows that

$$\begin{array}{lll} \boldsymbol{\mu}_a = 0 & \boldsymbol{\mu}_b = \mathbf{0} & \mathbf{x}_b = \mathbf{t} \\ \boldsymbol{\Sigma}_{aa} = c & \boldsymbol{\Sigma}_{bb} = \mathbf{C}_N & \boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^{\mathrm{T}} = \mathbf{k}^{\mathrm{T}} \end{array}$$

in (2.81) and (2.82), from which (6.66) and (6.67) follows directly.

6.21 Both the Gaussian process and the linear regression model give rise to Gaussian predictive distributions $p(t_{N+1}|\mathbf{x}_{N+1})$ so we simply need to show that these have the same mean and variance. To do this we make use of the expression (6.54) for the kernel function defined in terms of the basis functions. Using (6.62) the covariance matrix $\mathbf{C}_N$ then takes the form

$$\mathbf{C}_N = \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N \tag{210}$$

where $\boldsymbol{\Phi}$ is the design matrix with elements $\Phi_{nk} = \phi_k(\mathbf{x}_n)$, and $\mathbf{I}_N$ denotes the $N \times N$ unit matrix. Consider first the mean of the Gaussian process predictive distribution, which from (210), (6.54), (6.66) and the definitions in the text preceding (6.66) is given by

$$m_{N+1} = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\mathbf{t}.$$

We now make use of the matrix identity (C.6) to give

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

Thus the mean becomes

$$m_{N+1} = \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

which we recognize as the mean of the predictive distribution for the linear regression model given by (3.58) with $\mathbf{m}_N$ defined by (3.53) and $\mathbf{S}_N$ defined by (3.54).

For the variance we similarly substitute the expression (210) for the kernel function into the Gaussian process variance given by (6.67) and then use (6.54) and the definitions in the text preceding (6.66) to obtain

$$\begin{aligned}\sigma^2_{N+1}(\mathbf{x}_{N+1}) &= \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{N+1}) + \beta^{-1} \\ &\quad -\alpha^{-2}\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}_{N+1}) \\ &= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}\big(\alpha^{-1}\mathbf{I}_M \\ &\quad -\alpha^{-2}\boldsymbol{\Phi}^{\mathrm{T}}\left(\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \beta^{-1}\mathbf{I}_N\right)^{-1}\boldsymbol{\Phi}\big)\boldsymbol{\phi}(\mathbf{x}_{N+1}). \end{aligned} \tag{211}$$

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

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

where we have also used (3.54). Substituting this in (211), we obtain

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

as derived for the linear regression model in Section 3.3.2.

6.22 From (6.61) we have

$$p\left(\begin{bmatrix}\mathbf{t}_{1\ldots N} \\ \mathbf{t}_{N+1\ldots N+L}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{t}_{1\ldots N} \\ \mathbf{t}_{N+1\ldots N+L}\end{bmatrix}\middle|\,\mathbf{0}, \mathbf{C}\right)$$

with $\mathbf{C}$ specified by (6.62).

For our purposes, it is useful to consider the following partition[2] of $\mathbf{C}$:

$$\mathbf{C} = \begin{pmatrix}\mathbf{C}_{bb} & \mathbf{C}_{ba} \\ \mathbf{C}_{ab} & \mathbf{C}_{aa}\end{pmatrix},$$

where $\mathbf{C}_{aa}$ corresponds to $\mathbf{t}_{N+1\ldots N+L}$ and $\mathbf{C}_{bb}$ corresponds to $\mathbf{t}_{1\ldots N}$. We can use this together with (2.94)–(2.97) and (6.61) to obtain the conditional distribution

$$p(\mathbf{t}_{N+1\ldots N+L}|\mathbf{t}_{1\ldots N}) = \mathcal{N}\left(\mathbf{t}_{N+1\ldots N+L}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}^{-1}\right) \tag{212}$$

where, from (2.78)–(2.80),

$$\begin{aligned}\boldsymbol{\Lambda}^{-1}_{aa} &= \mathbf{C}_{aa} - \mathbf{C}_{ab}\mathbf{C}^{-1}_{bb}\mathbf{C}_{ba} \\ \boldsymbol{\Lambda}_{ab} &= -\boldsymbol{\Lambda}_{aa}\mathbf{C}_{ab}\mathbf{C}^{-1}_{bb}\end{aligned} \tag{213}$$

[2]The indexing and ordering of this partition have been chosen to match the indexing used in (2.94)–(2.97) as well as the ordering of elements used in the single variate case, as seen in (6.64)–(6.65).

and

$$\boldsymbol{\mu}_{a|b} = -\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\mathbf{t}_{1\ldots N} = \mathbf{C}_{ab}\mathbf{C}_{bb}^{-1}\mathbf{t}_{1\ldots N}. \tag{214}$$

Restricting (212) to a single test target, we obtain the corresponding marginal distribution, where $\mathbf{C}_{aa}$, $\mathbf{C}_{ba}$ and $\mathbf{C}_{bb}$ correspond to c, $\mathbf{k}$ and $\mathbf{C}_N$ in (6.65), respectively. Making the matching substitutions in (213) and (214), we see that they equal (6.67) and (6.66), respectively.

6.23 **NOTE**: In 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.

6.24 Since the diagonal elements of a diagonal matrix are also the eigenvalues of the matrix, $\mathbf{W}$ is positive definite (see Appendix C). Alternatively, for an arbitrary, non-zero vector $\mathbf{x}$,

$$\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} = \sum_i x_i^2 W_{ii} > 0.$$

If $\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} > 0$ and $\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x} > 0$ for an arbitrary, non-zero vector $\mathbf{x}$, then

$$\mathbf{x}^{\mathrm{T}}(\mathbf{W} + \mathbf{V})\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x} > 0.$$

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

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

6.26 Using (2.115) the mean of the posterior distribution $p(a_{N+1}|\mathbf{t}_N)$ is given by

$$\mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N^*.$$

Combining this with the condition

$$\mathbf{C}_N^{-1}\mathbf{a}_N^* = \mathbf{t}_N - \boldsymbol{\sigma}_N$$

satisfied by $\mathbf{a}_N^*$ we obtain (6.87).

Similarly, from (2.115) the variance of the posterior distribution $p(a_{N+1}|\mathbf{t}_N)$ is given by

$$\begin{aligned}\text{var}[a_{N+1}|\mathbf{t}_N] &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k} + \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{C}_N(\mathbf{I} + \mathbf{W}_N\mathbf{C}_N)^{-1}\mathbf{C}_N^{-1}\mathbf{k} \\ &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\left[\mathbf{I} - (\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\mathbf{C}_N^{-1}\right]\mathbf{k} \\ &= c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}(\mathbf{C}_N^{-1} + \mathbf{W}_N)^{-1}\mathbf{W}_N\mathbf{k} \\ &= c - \mathbf{k}^{\mathrm{T}}(\mathbf{W}_N^{-1} + \mathbf{C}_N)^{-1}\mathbf{k}\end{aligned}$$

as required.

6.27 Using (4.135), (6.80) and (6.85), we can approximate (6.89) as follows:

$$\begin{aligned}p(\mathbf{t}_N|\boldsymbol{\theta}) &= \int p(\mathbf{t}_N|\mathbf{a}_N)p(\mathbf{a}_N|\boldsymbol{\theta})\,\mathrm{d}\mathbf{a}_N \\ &\simeq p(\mathbf{t}_N|\mathbf{a}_N^\star)p(\mathbf{a}_N^\star|\boldsymbol{\theta}) \\ &\qquad \int \exp\left\{-\frac{1}{2}\left(\mathbf{a}_N - \mathbf{a}_N^\star\right)^{\mathrm{T}}\mathbf{H}\left(\mathbf{a}_N - \mathbf{a}_N^\star\right)\right\}\mathrm{d}\mathbf{a}_N \\ &= \exp\left(\Psi(\mathbf{a}_N^\star)\right)\frac{(2\pi)^{N/2}}{|\mathbf{H}|^{1/2}}.\end{aligned}$$

Taking the logarithm, we obtain (6.90).

To derive (6.91), we gather the terms from (6.90) that involve $\mathbf{C}_N$, yielding

$$\begin{aligned}-\frac{1}{2}\left(\mathbf{a}_N^{\star\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N^\star + \ln|\mathbf{C}_N| + \ln|\mathbf{W}_N + \mathbf{C}_N^{-1}|\right) \\ = -\frac{1}{2}\mathbf{a}_N^{\star\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N^\star - \frac{1}{2}\ln|\mathbf{C}_N\mathbf{W}_N + \mathbf{I}|.\end{aligned}$$

Applying (C.21) and (C.22) to the first and second terms, respectively, we get (6.91).

Applying (C.22) to the l.h.s. of (6.92), we get

$$\begin{aligned}-\frac{1}{2}\sum_{n=1}^{N}\frac{\partial\ln|\mathbf{W}_N + \mathbf{C}_N^{-1}|}{\partial a_n^\star}\frac{\partial a_n^\star}{\partial\theta_j} &= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left(\left(\mathbf{W}_N + \mathbf{C}_N^{-1}\right)^{-1}\frac{\partial\mathbf{W}}{\partial a_n^\star}\right)\frac{\partial a_n^\star}{\partial\theta_j} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\mathrm{Tr}\left(\left(\mathbf{C}_N\mathbf{W}_N + \mathbf{I}\right)^{-1}\mathbf{C}_N\frac{\partial\mathbf{W}}{\partial a_n^\star}\right)\frac{\partial a_n^\star}{\partial\theta_j}. \quad (215)\end{aligned}$$

Using the definition of $\mathbf{W}$ together with (4.88), we have

$$\begin{aligned}\frac{\mathrm{d}W_{nn}}{\mathrm{d}a_n^\star} &= \frac{\mathrm{d}\sigma_n^\star(1-\sigma_n^\star)}{\mathrm{d}a_n^\star} \\ &= \sigma_n^\star(1-\sigma_n^\star)^2 - \sigma_n^{\star 2}(1-\sigma_n^\star) \\ &= \sigma_n^\star(1-\sigma_n^\star)(1-2\sigma_n^\star)\end{aligned}$$

and substituting this into (215) we the the r.h.s. of (6.92).

Gathering all the terms in (6.93) involving $\partial a_n^\star / \partial \theta_j$ on one side, we get

$$(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N) \frac{\partial a_n^\star}{\partial \theta_j} = \frac{\partial \mathbf{C}_N}{\partial \theta_j} (\mathbf{t}_N - \boldsymbol{\sigma}_N).$$

Left-multiplying both sides with $(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1}$, we obtain (6.94).

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

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

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 $\phi(\mathbf{x})$.

7.2 Consider multiplying both sides of (7.5) by $\gamma > 0$. Accordingly, we would then replace all occurences of $\mathbf{w}$ and b in (7.3) with $\gamma\mathbf{w}$ and γb, respectively. However, as discussed in the text following (7.3), its solution w.r.t. $\mathbf{w}$ and b is invariant to a common scaling factor and hence would remain unchanged.

7.3 Given a data set of two data points, $\mathbf{x}_1 \in \mathcal{C}_+$ ($t_1 = +1$) and $\mathbf{x}_2 \in \mathcal{C}_-$ ($t_2 = -1$), the maximum margin hyperplane is determined by solving (7.6) subject to the constraints

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 + b = +1 \tag{216}$$
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 + b = -1. \tag{217}$$

We do this by introducing Lagrange multipliers λ and η, and solving

$$\arg\min_{\mathbf{w},b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \lambda\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 + b - 1\right) + \eta\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_2 + b + 1\right) \right\}.$$

Taking the derivative of this w.r.t. $\mathbf{w}$ and b and setting the results to zero, we obtain

$$0 = \mathbf{w} + \lambda\mathbf{x}_1 + \eta\mathbf{x}_2 \tag{218}$$
$$0 = \lambda + \eta. \tag{219}$$

Equation (219) immediately gives $\lambda = -\eta$, which together with (218) give

$$\mathbf{w} = \lambda\left(\mathbf{x}_1 - \mathbf{x}_2\right). \tag{220}$$

For b, we first rearrange and sum (216) and (217) to obtain

$$2b = -\mathbf{w}^{\mathrm{T}}\left(\mathbf{x}_1 + \mathbf{x}_2\right).$$

Using (220), we can rewrite this as

$$\begin{aligned} b &= -\frac{\lambda}{2}\left(\mathbf{x}_1 - \mathbf{x}_2\right)^{\mathrm{T}}\left(\mathbf{x}_1 + \mathbf{x}_2\right) \\ &= -\frac{\lambda}{2}\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1 - \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2\right). \end{aligned}$$

Note that the Lagrange multiplier λ remains undetermined, which reflects the inherent indeterminacy in the magnitude of $\mathbf{w}$ and b.

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.

7.5 These properties follow directly from the results obtained in the solution to the previous exercise, 7.4.

7.6 If $p(t = 1|y) = \sigma(y)$, then

$$p(t = -1|y) \;=\; 1 - p(t = 1|y) \;=\; 1 - \sigma(y) \;=\; \sigma(-y),$$

where we have used (4.60). Thus, given i.i.d. data $\mathcal{D} = \{(t_1, \mathbf{x}_n), \ldots, (t_N, \mathbf{x}_N)\}$, we can write the corresponding likelihood as

$$p(\mathcal{D}) = \prod_{t_n=1} \sigma(y_n) \prod_{t_{n'}=-1} \sigma(-y_{n'}) = \prod_{n=1}^{N} \sigma(t_n y_n),$$

where $y_n = y(\mathbf{x}_n)$, as given by (7.1). Taking the negative logarithm of this, we get

$$\begin{aligned} -\ln p(\mathcal{D}) &= -\ln \prod_{n=1}^{N} \sigma(t_n y_n) \\ &= \sum_{n=1}^{N} \ln \sigma(t_n y_n) \\ &= \sum_{n=1}^{N} \ln(1 + \exp(-t_n y_n)), \end{aligned}$$

where we have used (4.59). Combining this with the regularization term $\lambda\|\mathbf{w}\|^2$, we obtain (7.47).

7.7 We start by rewriting (7.56) as

$$\begin{aligned} L &= \sum_{n=1}^{N} C\xi_n + \sum_{n=1}^{N} C\widehat{\xi}_n + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n) \\ &\quad - \sum_{n=1}^{N} a_n(\epsilon + \xi_n + \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b - t_n) \\ &\quad - \sum_{n=1}^{N} \widehat{a}_n(\epsilon + \widehat{\xi}_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - b + t_n), \end{aligned}$$

where we have used (7.1). We now use (7.1), (7.57), (7.59) and (7.60) to rewrite this as

$$\begin{aligned} L &= \sum_{n=1}^{N}(a_n + \mu_n)\xi_n + \sum_{n=1}^{N}(\widehat{a}_n + \widehat{\mu}_n)\widehat{\xi}_n \\ &\quad + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m) - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n) \\ &\quad - \sum_{n=1}^{N}(a_n\xi_n + \widehat{a}_n\widehat{\xi}_n) - \epsilon\sum_{n=1}^{N}(a_n + \widehat{a}_n) + \sum_{n=1}^{N}(a_n - \widehat{a}_n)t_n \\ &\quad - \sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m) - b\sum_{n=1}^{N}(a_n - \widehat{a}_n). \end{aligned}$$

If we now eliminate terms that cancel out and use (7.58) to eliminate the last term, what we are left with equals the r.h.s. of (7.61).

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

$$\begin{aligned} \xi_n &\geqslant 0 \\ \mu_n &\geqslant 0 \\ \mu_n\xi_n &= 0 \end{aligned} \tag{221}$$

and from (7.59) we have that

$$\mu_n = C - a_n. \tag{222}$$

Combining (221) and (222), we get (7.67); similar reasoning for $\widehat{\mu}_n$ and $\widehat{\xi}_n$ lead to (7.68).

7.9 From (7.76), (7.79) and (7.80), we make the substitutions

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{0} \quad \boldsymbol{\Lambda} \Rightarrow \mathrm{diag}(\boldsymbol{\alpha})$$

$$\mathbf{y} \Rightarrow \mathbf{t} \quad \mathbf{A} \Rightarrow \mathbf{\Phi} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L} \Rightarrow \beta\mathbf{I},$$

in (2.113) and (2.114), upon which the desired result follows from (2.116) and (2.117).

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)

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{1}{(2\pi)^{N/2}} \prod_{i=1}^{M} \alpha_i \int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} \tag{223}$$

where

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

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

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

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{w} - \mathbf{m}) + E(\mathbf{t}) \tag{224}$$

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

$$E(\mathbf{t}) = \frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} - \mathbf{m}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{m}\right). \tag{225}$$

Using (224), we can evaluate the integral in (223) to obtain

$$\int \exp\{-E(\mathbf{w})\}\, \mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{t})\}\,(2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2}. \tag{226}$$

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 (225) as follows

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

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 $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)$.

7.11 If we make the same substitutions as in Exercise 7.9, the desired result follows from (2.115).

7.12 Using the results (223)–(226) from Solution 7.10, we can write (7.85) in the form of (3.86):

$$\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{N}\ln\alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi). \quad (227)$$

By making use of (225) and (7.83) together with (C.22), we can take the derivatives of this w.r.t α_i, yielding

$$\frac{\partial}{\partial\alpha_i}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2. \quad (228)$$

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

$$\frac{\partial}{\partial\beta}\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 - \mathrm{Tr}\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right]\right). \quad (229)$$

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

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= \boldsymbol{\Sigma}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= (\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= (\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})\beta^{-1}. \quad (230)
\end{aligned}$$

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 (229) equal to zero and then re-arrange to obtain (7.88).

7.13 We start by introducing prior distributions over $\boldsymbol{\alpha}$ and β,

$$\begin{aligned}
p(\alpha_i) &= \mathrm{Gam}\left(\alpha_i|a_{\alpha 0}, b_{\beta 0}\right), i = 1, \ldots, N, \\
p(\beta) &= \mathrm{Gam}\left(\beta|a_{\beta 0}, b_{\beta 0}\right).
\end{aligned}$$

Note that we use an independent, common prior for all α_i. We can then combine this with (7.84) to obtain

$$p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta)p(\boldsymbol{\alpha})p(\beta).$$

Rather than maximizing the r.h.s. directly, we first take the logarithm, which enables us to use results from Solution 7.12. Using (227) and (B.26), we get

$$\begin{aligned}\ln p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = &\frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{N}\ln\alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\ln(2\pi) \\ &-N\ln\Gamma(a_{\alpha 0})^{-1} + Na_{\alpha 0}\ln b_{\alpha 0} + \sum_{i=1}^{N}\left((a_{\alpha 0}-1)\ln\alpha_i - b_{\alpha 0}\alpha_i\right) \\ &-\ln\Gamma(a_{\beta 0})^{-1} + a_{\beta 0}\ln b_{\beta 0} + (a_{\beta 0}-1)\ln\beta - b_{\beta 0}\beta.\end{aligned}$$

Using (228), we obtain the derivative of this w.r.t. α_i as

$$\frac{\partial}{\partial\alpha_i}\ln p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2 + \frac{a_{\alpha 0}-1}{\alpha_i} - b_{\alpha 0}.$$

Setting this to zero and rearranging (cf. Solution 7.12) we obtain

$$\alpha_i^{\text{new}} = \frac{\gamma_i + 2a_{\alpha 0} - 2}{m_i^2 - 2b_{\alpha 0}},$$

where we have used (7.89).

For β, we can use (229) together with (B.26) to get

$$\frac{\partial}{\partial\beta}\ln p(\boldsymbol{\alpha}, \beta, \mathbf{t}|\mathbf{X}) = \frac{1}{2}\left(\frac{N}{\beta} - \|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}\|^2 - \text{Tr}\left[\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}\right]\right) + \frac{a_{\beta 0}-1}{\beta} - b_{\beta 0}.$$

Setting this equal to zero and using (7.89) and (230), we get

$$\frac{1}{\beta^{\text{new}}} = \frac{\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}\|^2 + 2b_{\beta 0}}{a_{\beta 0} + 2 + N - \sum_i \gamma_i}.$$

7.14 If we make the following substitions from (7.81) into (2.113),

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{m} \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \boldsymbol{\Sigma},$$

and from (7.76) and (7.77) into (2.114)

$$\mathbf{y} \Rightarrow t \quad \mathbf{A} \Rightarrow \boldsymbol{\phi}(\mathbf{x})^{\text{T}} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L} \Rightarrow \beta^{\star}\mathbf{I},$$

(7.90) and (7.91) can be read off directly from (2.115).

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

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

7.16 If we differentiate (7.97) twice w.r.t. α_i, we get

$$\frac{\mathrm{d}^2\lambda}{\mathrm{d}\alpha_i^2} = -\frac{1}{2}\left(\frac{1}{\alpha_i^2} + \frac{1}{(\alpha_i + s_i)^2}\right).$$

This second derivative must be negative and thus the solution given by (7.101) corresponds to a maximum.

7.17 Using (7.83), (7.86) and (C.7), we have

$$\mathbf{C}^{-1} = \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\left(\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}} = \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}}.$$

Substituting this into (7.102) and (7.103), we immediately obtain (7.106) and (7.107), respectively.

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 (161) and (162) 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.

7.19 **NOTE**: In PRML, on line 1 of the text of this exercise, "approximate log marginal" should be "approximate marginal".

We start by taking the logarithm of (7.114), which, omitting terms that do not depend on $\boldsymbol{\alpha}$, leaves us with

$$\ln p(\mathbf{w}^\star|\boldsymbol{\alpha}) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| = -\frac{1}{2}\left(\ln|\boldsymbol{\Sigma}^{-1}| + \sum_i (w_i^\star)^2\,\alpha_i - \ln\alpha_i\right),$$

where we have used (7.80). Making use of (7.113) and (C.22), we can differentiate this to obtain (7.115), from which we get (7.116) by using $\gamma_i = 1 - \alpha_i\Sigma_{ii}$.

Chapter 8 Probabilistic 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. We can then marginalize over the nodes in reverse order, starting with x_K

$$\begin{aligned}\sum_{x_1}\cdots\sum_{x_K} p(\mathbf{x}) &= \sum_{x_1}\cdots\sum_{x_K} p(x_K|\mathrm{pa}_K)\prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k)\\ &= \sum_{x_1}\cdots\sum_{x_{K-1}}\prod_{k=1}^{K-1} p(x_k|\mathrm{pa}_k),\end{aligned}$$

since each of the conditional distributions is assumed to be correctly normalized and none of the other variables depend on x_K. Repeating this process $K-2$ times we are left with

$$\sum_{x_1} p(x_1|\emptyset) = 1.$$

8.2 Consider a directed graph in which the nodes of the graph are numbered such that are no edges going from a node to a lower numbered node. If there exists a directed cycle in the graph then the subset of nodes belonging to this directed cycle must also satisfy the same numbering property. If we traverse the cycle in the direction of the edges the node numbers cannot be monotonically increasing since we must end up back at the starting node. It follows that the cycle cannot be a directed cycle.

Table 1 Comparison of the distribution $p(a, b)$ with the product of marginals $p(a)p(b)$ showing that these are not equal for the given joint distribution $p(a, b, c)$.

a	b	$p(a, b)$
0	0	0.336
0	1	0.264
1	0	0.256
1	1	0.144

a	b	$p(a)p(b)$
0	0	0.355
0	1	0.245
1	0	0.237
1	1	0.163

8.3 The distribution $p(a, b)$ is found by summing the complete joint distribution $p(a, b, c)$ over the states of c so that

$$p(a, b) = \sum_{c\in\{0,1\}} p(a, b, c)$$

and similarly the marginal distributions $p(a)$ and $p(b)$ are given by

$$p(a) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c) \quad \text{and} \quad p(b) = \sum_{a\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c). \tag{231}$$

Table 1 shows the joint distribution $p(a, b)$ as well as the product of marginals $p(a)p(b)$, demonstrating that these are not equal for the specified distribution.

The conditional distribution $p(a, b|c)$ is obtained by conditioning on the value of c and normalizing

$$p(a, b|c) = \frac{p(a, b, c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a, b, c)}.$$

Similarly for the conditionals $p(a|c)$ and $p(b|c)$ we have

$$p(a|c) = \frac{\sum_{b\in\{0,1\}} p(a, b, c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a, b, c)}$$

and

$$p(b|c) = \frac{\sum_{a\in\{0,1\}} p(a, b, c)}{\sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} p(a, b, c)}. \tag{232}$$

Table 2 compares the conditional distribution $p(a, b|c)$ with the product of marginals $p(a|c)p(b|c)$, showing that these are equal for the given joint distribution $p(a, b, c)$ for both $c = 0$ and $c = 1$.

8.4 In the previous exercise we have already computed $p(a)$ in (231) and $p(b|c)$ in (232). There remains to compute $p(c|a)$ which is done using

$$p(c|a) = \frac{\sum_{b\in\{0,1\}} p(a, b, c)}{\sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} p(a, b, c)}.$$

The required distributions are given in Table 3.

Table 2 Comparison of the conditional distribution $p(a, b|c)$ with the product of marginals $p(a|c)p(b|c)$ showing that these are equal for the given distribution.

a	b	c	$p(a, b\|c)$
0	0	0	0.400
0	1	0	0.100
1	0	0	0.400
1	1	0	0.100
0	0	1	0.277
0	1	1	0.415
1	0	1	0.123
1	1	1	0.185

a	b	c	$p(a\|c)p(b\|c)$
0	0	0	0.400
0	1	0	0.100
1	0	0	0.400
1	1	0	0.100
0	0	1	0.277
0	1	1	0.415
1	0	1	0.123
1	1	1	0.185

Table 3 Tables of $p(a)$, $p(c|a)$ and $p(b|c)$ evaluated by marginalizing and conditioning the joint distribution of Table 8.2.

a	$p(a)$
0	0.600
1	0.400

c	a	$p(c\|a)$
0	0	0.400
1	0	0.600
0	1	0.600
1	1	0.400

b	c	$p(b\|c)$
0	0	0.800
1	0	0.200
0	1	0.400
1	1	0.600

Multiplying the three distributions together we recover the joint distribution $p(a, b, c)$ given in Table 8.2, thereby allowing us to verify the validity of the decomposition $p(a, b, c) = p(a)p(c|a)p(b|c)$ for this particular joint distribution. We can express this decomposition using the graph shown in Figure 4.

8.5 **NOTE**: In PRML, Equation (7.79) contains a typographical error: $p(t_n|\mathbf{x}_n, \mathbf{w}, \beta^{-1})$ should be $p(t_n|\mathbf{x}_n, \mathbf{w}, \beta)$. This correction is provided for completeness only; it does not affect this solution.

The solution is given in Figure 5.

8.6 **NOTE**: In PRML, the text of the exercise should be slightly altered; please consult the PRML errata.

In order to interpret (8.104) suppose initially that $\mu_0 = 0$ and that $\mu_i = 1 - \epsilon$ where $\epsilon \ll 1$ for $i = 1, \ldots, K$. We see that, if all of the $x_i = 0$ then $p(y = 1|x_1, \ldots, x_K) = 0$ while if L of the $x_i = 1$ then $p(y = 1|x_1, \ldots, x_K) = 1 - \epsilon^L$ which is close to 1. For $\epsilon \to 0$ this represents the logical OR function in which $y = 1$ if one or more of the $x_i = 1$, and $y = 0$ otherwise. More generally, if just one of the $x_i = 1$ with all remaining $x_{j \neq i} = 0$ then $p(y = 1|x_1, \ldots, x_K) = \mu_i$ and so we can interpret μ_i as the probability of $y = 1$ given that only this one $x_i = 1$. We can similarly interpret μ_0 as the probability of $y = 1$ when all of the $x_i = 0$. An example of the application of this model would be in medical diagnosis in which y represents the presence or absence of a symptom, and each of the x_i represents the presence or absence of some disease. For the i^{th} disease there is a probability μ_i that it will give rise to the symptom. There is also a background probability μ_0 that

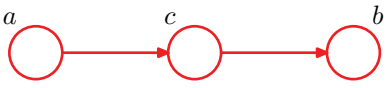


Figure 4 Directed graph representing the joint distribution given in Table 8.2.

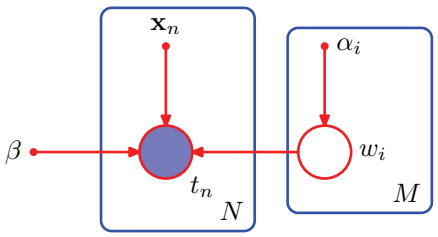


Figure 5 The graphical representation of the relevance vector machine (RVM); Solution 8.5.

the symptom will be observed even in the absence of disease. In practice we might observe that the symptom is indeed present (so that $y = 1$) and we wish to infer the posterior probability for each disease. We can do this using Bayes' theorem once we have defined prior probabilities $p(x_i)$ for the diseases.

8.7 Starting with $\boldsymbol{\mu}$, (8.11) and (8.15) directly gives

$$\mu_1 = \sum_{j \in \emptyset} w_{1j} \mathbb{E}[x_j] + b_1 = b_1,$$

$$\mu_2 = \sum_{j \in \{x_1\}} w_{2j} \mathbb{E}[x_j] + b_2 = w_{21} b_1 + b_2$$

and

$$\mu_3 = \sum_{j \in \{x_2\}} w_{3j} \mathbb{E}[x_j] + b_3 = w_{32}(w_{21} b_1 + b_2) + b_3.$$

Similarly for $\boldsymbol{\Sigma}$, using (8.11) and (8.16), we get

$$\mathrm{cov}[x_1, x_1] = \sum_{k \in \emptyset} w_{1j} \mathrm{cov}[x_1, x_k] + I_{11} v_1 = v_1,$$

$$\mathrm{cov}[x_1, x_2] = \sum_{k \in \{x_1\}} w_{2j} \mathrm{cov}[x_1, x_k] + I_{12} v_2 = w_{21} v_1,$$

$$\mathrm{cov}[x_1, x_3] = \sum_{k \in \{x_2\}} w_{3j} \mathrm{cov}[x_1, x_k] + I_{13} v_3 = w_{32} w_{21} v_1,$$

$$\mathrm{cov}[x_2, x_2] = \sum_{k \in \{x_1\}} w_{2j} \mathrm{cov}[x_2, x_k] + I_{22} v_2 = w_{21}^2 v_1 + v_2,$$

$$\mathrm{cov}[x_2, x_3] = \sum_{k \in \{x_2\}} w_{3j} \mathrm{cov}[x_2, x_k] + I_{23} v_3 = w_{32}(w_{21}^2 v_1 + v_2)$$

and

$$\mathrm{cov}[x_3, x_3] = \sum_{k \in \{x_2\}} w_{3j} \mathrm{cov}[x_3, x_k] + I_{33} v_3 = w_{32}^2(w_{21}^2 v_1 + v_2) + v_3,$$

where the symmetry of $\boldsymbol{\Sigma}$ gives the below diagonal elements.

8.8 $a \perp\!\!\!\perp b, c \mid d$ can be written as

$$p(a, b, c|d) = p(a|d)p(b, c|d).$$

Summing (or integrating) both sides with respect to c, we obtain

$$p(a, b|d) = p(a|d)p(b|d) \qquad \text{or} \qquad a \perp\!\!\!\perp b \mid d,$$

as desired.

8.9 Consider Figure 8.26. In order to apply the d-separation criterion we need to consider all possible paths from the central node $\mathbf{x}_i$ to all possible nodes external to the Markov blanket. There are three possible categories of such paths. First, consider paths via the parent nodes. Since the link from the parent node to the node $\mathbf{x}_i$ has its tail connected to the parent node, it follows that for any such path the parent node must be either tail-to-tail or head-to-tail with respect to the path. Thus the observation of the parent node will block any such path. Second consider paths via one of the child nodes of node $\mathbf{x}_i$ which do not pass directly through any of the co-parents. By definition such paths must pass to a child of the child node and hence will be head-to-tail with respect to the child node and so will be blocked. The third and final category of path passes via a child node of $\mathbf{x}_i$ and then a co-parent node. This path will be head-to-head with respect to the observed child node and hence will not be blocked by the observed child node. However, this path will either tail-to-tail or head-to-tail with respect to the co-parent node and hence observation of the co-parent will block this path. We therefore see that all possible paths leaving node $\mathbf{x}_i$ will be blocked and so the distribution of $\mathbf{x}_i$, conditioned on the variables in the Markov blanket, will be independent of all of the remaining variables in the graph.

8.10 From Figure 8.54, we see that

$$p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c).$$

Following the examples in Section 8.2.1, we see that

$$\begin{aligned} p(a, b) &= \sum_c \sum_d p(a, b, c, d) \\ &= p(a)p(b) \sum_c p(c|a, b) \sum_d p(d|c) \\ &= p(a)p(b). \end{aligned}$$

Similarly,

$$\begin{aligned} p(a, b|d) &= \frac{\sum_c p(a, b, c, d)}{\sum_a \sum_b \sum_c p(a, b, c, d)} \\ &= \frac{p(d|a, b)p(a)p(b)}{p(d)} \\ &\neq p(a|d)p(b|d) \end{aligned}$$

in general. Note that this result could also be obtained directly from the graph in Figure 8.54 by using d-separation, discussed in Section 8.2.2.

8.11 The described situation correspond to the graph shown in Figure 8.54 with $a = B$, $b = F$, $c = G$ and $d = D$ (cf. Figure 8.21). To evaulate the probability that the tank is empty given the driver's report that the gauge reads zero, we use Bayes' theorem

$$p(F = 0|D = 0) = \frac{p(D = 0|F = 0)p(F = 0)}{p(D = 0)}.$$

To evaluate $p(D = 0|F = 0)$, we marginalize over B and G,

$$p(D = 0|F = 0) = \sum_{B,G} p(D = 0|G)p(G|B, F = 0)p(B) = 0.748 \tag{233}$$

and to evaluate $p(D = 0)$, we marginalize also over F,

$$p(D = 0) = \sum_{B,G,F} p(D = 0|G)p(G|B, F)p(B)p(F) = 0.352. \tag{234}$$

Combining these results with $p(F = 0)$, we get

$$p(F = 0|D = 0) = 0.213.$$

Note that this is slightly lower than the probability obtained in (8.32), reflecting the fact that the driver is not completely reliable.

If we now also observe $B = 0$, we longer marginalize over B in (233) and (234), but instead keep it fixed at its observed value, yielding

$$p(F = 0|D = 0, B = 0) = 0.110$$

which is again lower than what we obtained with a direct observation of the fuel gauge in (8.33). More importantly, in both cases the value is lower than before we observed $B = 0$, since this observation provides an alternative explanation why the gauge should read zero; see also discussion following (8.33).

8.12 In an undirected graph of M nodes there could potentially be a link between each pair of nodes. The number of distinct graphs is then 2 raised to the power of the number of potential links. To evaluate the number of distinct links, note that there are M nodes each of which could have a link to any of the other $M - 1$ nodes, making a total of $M(M - 1)$ links. However, each link is counted twice since, in an undirected graph, a link from node a to node b is equivalent to a link from node b to node a. The number of distinct potential links is therefore $M(M - 1)/2$ and so the number of distinct graphs is $2^{M(M-1)/2}$. The set of 8 possible graphs over three nodes is shown in Figure 6.

8.13 The change in energy is

$$E(x_j = +1) - E(x_j = -1) = 2h - 2\beta \sum_{i \in \mathrm{ne}(j)} x_i - 2\eta y_j$$

where $\mathrm{ne}(j)$ denotes the nodes which are neighbours of x_j.

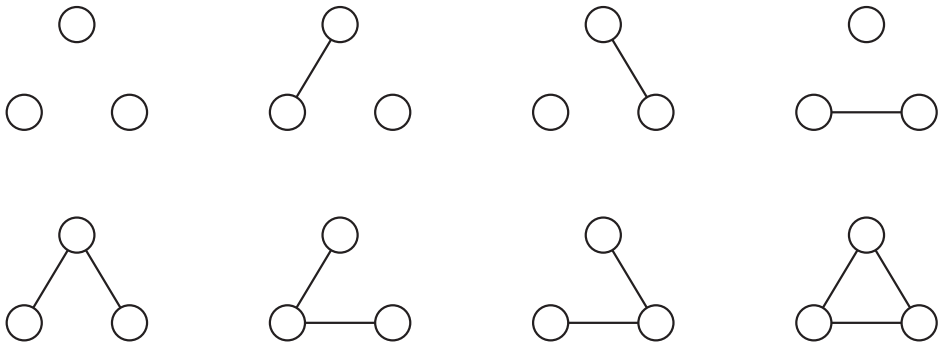


Figure 6 The set of 8 distinct undirected graphs which can be constructed over $M = 3$ nodes.

8.14 The most probable configuration corresponds to the configuration with the lowest energy. Since η is a positive constant (and $h = \beta = 0$) and $x_i, y_i \in \{-1, +1\}$, this will be obtained when $x_i = y_i$ for all $i = 1, \ldots, D$.

8.15 The marginal distribution $p(x_{n-1}, x_n)$ is obtained by marginalizing the joint distribution $p(\mathbf{x})$ over all variables except x_{n-1} and x_n,

$$p(x_{n-1}, x_n) = \sum_{x_1} \cdots \sum_{x_{n-2}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x}).$$

This is analogous to the marginal distribution for a single variable, given by (8.50).

Following the same steps as in the single variable case described in Section 8.4.1, we arrive at a modified form of (8.52),

$$p(x_n) = \frac{1}{Z} \underbrace{\left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \cdots \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right] \cdots\right]}_{\mu_\alpha(x_{n-1})} \psi_{n-1,n}(x_{n-1}, x_n)$$
$$\underbrace{\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]}_{\mu_\beta(x_n)},$$

from which (8.58) immediately follows.

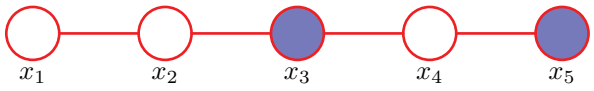
8.16 Observing $\mathbf{x}_N = \widehat{\mathbf{x}}_N$ will only change the initial expression (message) for the β-recursion, which now becomes

$$\mu_\beta(\mathbf{x}_{N-1}) = \psi_{N-1,N}(\mathbf{x}_{N-1}, \widehat{\mathbf{x}}_N).$$

Note that there is no summation over $\mathbf{x}_N$. $p(\mathbf{x}_n)$ can then be evaluated using (8.54)–(8.57) for all $n = 1, \ldots, N-1$.

8.17 With $N = 5$ and x_3 and x_5 observed, the graph from Figure 8.38 will look like in Figure 7. This graph is undirected, but from Figure 8.32 we see that the equivalent

Figure 7 The graph discussed in Solution 8.17.



directed graph can be obtained by simply directing all the edges from left to right. (**NOTE**: In PRML, the labels of the two rightmost nodes in Figure 8.32b should be interchanged to be the same as in Figure 8.32a.) In this directed graph, the edges on the path from x_2 to x_5 meet head-to-tail at x_3 and since x_3 is observed, by d-separation $x_2 \perp\!\!\!\perp x_5 | x_3$; note that we would have obtained the same result if we had chosen to direct the arrows from right to left. Alternatively, we could have obtained this result using graph separation in undirected graphs, illustrated in Figure 8.27.

From (8.54), we have

$$p(x_2) = \frac{1}{Z}\mu_\alpha(x_2)\mu_\beta(x_2). \tag{235}$$

$\mu_\alpha(x_2)$ is given by (8.56), while for $\mu_\beta(x_2)$, (8.57) gives

$$\begin{aligned}\mu_\beta(x_2) &= \sum_{x_3} \psi_{2,3}(x_2, x_3)\mu_\beta(x_3) \\ &= \psi_{2,3}(x_2, \widehat{x}_3)\mu_\beta(\widehat{x}_3)\end{aligned}$$

since x_3 is observed and we denote the observed value $\widehat{x}_3$. Thus, any influence that x_5 might have on $\mu_\beta(\widehat{x}_3)$ will be in terms of a scaling factor that is indepedent of x_2 and which will be absorbed into the normalization constant Z in (235) and so

$$p(x_2|x_3, x_5) = p(x_2|x_3).$$

8.18 The joint probability distribution over the variables in a general directed graphical model is given by (8.5). In the particular case of a tree, each node has a single parent, so pa_k will be a singleton for each node, k, except for the root node for which it will empty. Thus, the joint probability distribution for a tree will be similar to the joint probability distribution over a chain, (8.44), with the difference that the same variable may occur to the right of the conditioning bar in several conditional probability distributions, rather than just one (in other words, although each node can only have one parent, it can have several children). Hence, the argument in Section 8.3.4, by which (8.44) is re-written as (8.45), can also be applied to probability distributions over trees. The result is a Markov random field model where each potential function corresponds to one conditional probability distribution in the directed tree. The prior for the root node, e.g. $p(x_1)$ in (8.44), can again be incorporated in one of the potential functions associated with the root node or, alternatively, can be incorporated as a single node potential.

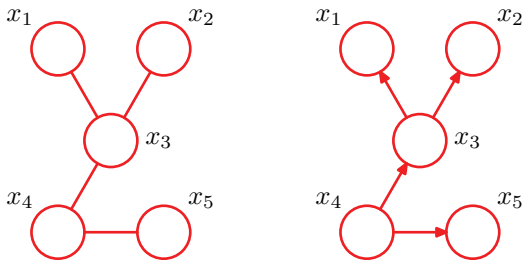


Figure 8 The graph on the left is an undirected tree. If we pick x_4 to be the root node and direct all the edges in the graph to point from the root to the leaf nodes (x_1, x_2 and x_5), we obtain the directed tree shown on the right.

This transformation can also be applied in the other direction. Given an undirected tree, we pick a node arbitrarily as the root. Since the graph is a tree, there is a unique path between every pair of nodes, so, starting at root and working outwards, we can direct all the edges in the graph to point from the root to the leaf nodes. An example is given in Figure 8. Since every edge in the tree correspond to a two-node potential function, by normalizing this appropriately, we obtain a conditional probability distribution for the child given the parent.

Since there is a unique path beween every pair of nodes in an undirected tree, once we have chosen the root node, the remainder of the resulting directed tree is given. Hence, from an undirected tree with N nodes, we can construct N different directed trees, one for each choice of root node.

8.19 If we convert the chain model discussed in Section 8.4.1 into a factor graph, each potential function in (8.49) will become a factor. Under this factor graph model, $p(x_n)$ is given by (8.63) as

$$p(x_n) = \mu_{f_{n-1,n}\to x_n}(x_n)\mu_{f_{n,n+1}\to x_n}(x_n) \tag{236}$$

where we have adopted the indexing of potential functions from (8.49) to index the factors. From (8.64)–(8.66), we see that

$$\mu_{f_{n-1,n}\to x_n}(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_{x_{n-1}\to f_{n-1,n}}(x_{n-1}) \tag{237}$$

and

$$\mu_{f_{n,n+1}\to x_n}(x_n) = \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})\mu_{x_{n+1}\to f_{n,n+1}}(x_{n+1}). \tag{238}$$

From (8.69), we further see that

$$\mu_{x_{n-1}\to f_{n-1,n}}(x_{n-1}) = \mu_{f_{n-2,n-1}\to x_{n-1}}(x_{n-1})$$

and

$$\mu_{x_{n+1}\to f_{n,n+1}}(x_{n+1}) = \mu_{f_{n+1,n+2}\to x_{n+1}}(x_{n+1}).$$

Substituting these into (237) and (238), respectively, we get

$$\mu_{f_{n-1,n}\to x_n}(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)\mu_{f_{n-2,n-1}\to x_{n-1}}(x_{n-1}) \tag{239}$$

and

$$\mu_{f_{n,n+1}\to x_n}(x_n) = \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})\mu_{f_{n+1,n+2}\to x_{n+1}}(x_{n+1}). \tag{240}$$

Since the messages are uniquely identified by the index of their arguments and whether the corresponding factor comes before or after the argument node in the chain, we can rename the messages as

$$\mu_{f_{n-2,n-1}\to x_{n-1}}(x_{n-1}) = \mu_\alpha(x_{n-1})$$

and

$$\mu_{f_{n+1,n+2}\to x_{n+1}}(x_{n+1}) = \mu_\beta(x_{n+1}).$$

Applying these name changes to both sides of (239) and (240), respectively, we recover (8.55) and (8.57), and from these and (236) we obtain (8.54); the normalization constant $1/Z$ can be easily computed by summing the (unnormalized) r.h.s. of (8.54). Note that the end nodes of the chain are variable nodes which send unit messages to their respective neigbouring factors (cf. (8.56)).

8.20 We do the induction over the size of the tree and we grow the tree one node at a time while, at the same time, we update the message passing schedule. Note that we can build up any tree this way.

For a single root node, the required condition holds trivially true, since there are no messages to be passed. We then assume that it holds for a tree with N nodes. In the induction step we add a new leaf node to such a tree. This new leaf node need not to wait for any messages from other nodes in order to send its outgoing message and so it can be scheduled to send it first, before any other messages are sent. Its parent node will receive this message, whereafter the message propagation will follow the schedule for the original tree with N nodes, for which the condition is assumed to hold.

For the propagation of the outward messages from the root back to the leaves, we first follow the propagation schedule for the original tree with N nodes, for which the condition is assumed to hold. When this has completed, the parent of the new leaf node will be ready to send its outgoing message to the new leaf node, thereby completing the propagation for the tree with $N + 1$ nodes.

8.21 **NOTE**: In PRML, this exercise contains a typographical error. On line 2, $f_x(\mathbf{x}_s)$ should be $f_s(\mathbf{x}_s)$.

To compute $p(\mathbf{x}_s)$, we marginalize $p(\mathbf{x})$ over all other variables, analogously to (8.61),

$$p(\mathbf{x}_s) = \sum_{\mathbf{x}\setminus\mathbf{x}_s} p(\mathbf{x}).$$

Using (8.59) and the defintion of $F_s(x, X_s)$ that followed (8.62), we can write this

as

$$\begin{aligned} p(\mathbf{x}_s) &= \sum_{\mathbf{x}\setminus\mathbf{x}_s} f_s(\mathbf{x}_s) \prod_{i\in\mathrm{ne}(f_s)} \prod_{j\in\mathrm{ne}(x_i)\setminus f_s} F_j(x_i, X_{ij}) \\ &= f_s(\mathbf{x}_s) \prod_{i\in\mathrm{ne}(f_s)} \sum_{\mathbf{x}\setminus\mathbf{x}_s} \prod_{j\in\mathrm{ne}(x_i)\setminus f_s} F_j(x_i, X_{ij}) \\ &= f_s(\mathbf{x}_s) \prod_{i\in\mathrm{ne}(f_s)} \mu_{x_i\to f_s}(x_i), \end{aligned}$$

where in the last step, we used (8.67) and (8.68). Note that the marginalization over the different sub-trees rooted in the neighbours of f_s would only run over variables in the respective sub-trees.

8.22 Let X_a denote the set of variable nodes in the connected subgraph of interest and X_b the remaining variable nodes in the full graph. To compute the joint distribution over the variables in X_a, we need to marginalize $p(\mathbf{x})$ over X_b,

$$p(X_a) = \sum_{X_b} p(\mathbf{x}).$$

We can use the sum-product algorithm to perform this marginalization efficiently, in the same way that we used it to marginalize over all variables but x_n when computing $p(x_n)$. Following the same steps as in the single variable case (see Section 8.4.4), we can write can write $p(X_a)$ in a form corresponding to (8.63),

$$\begin{aligned} p(X_a) &= \prod_{s_a} f_{s_a}(X_{s_a}) \prod_{s\in\mathrm{ne}X_a} \sum_{X_s} F_s(x_s, X_s) \\ &= \prod_{s_a} f_{s_a}(X_{s_a}) \prod_{s\in\mathrm{ne}X_a} \mu_{f_s\to x_s}(x_s). \end{aligned} \tag{241}$$

Here, s_a indexes factors that only depend on variables in X_a and so $X_{s_a} \subseteq X_a$ for all values of s_a; s indexes factors that connect X_a and X_b and hence also the corresponding nodes, $x_s \in X_a$. $X_s \subseteq X_b$ denotes the variable nodes connected to x_s via factor f_s. The messages $\mu_{f_s\to x_s}(x_s)$ can be computed using the sum-product algorithm, starting from the leaf nodes in, or connected to nodes in, X_b. Note that the density in (241) may require normalization, which will involve summing the r.h.s. of (241) over all possible combination of values for X_a.

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

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

8.24 **NOTE**: In PRML, this exercise contains a typographical error. On the last line, $f(\mathbf{x}_s)$ should be $f_s(\mathbf{x}_s)$.

See Solution 8.21.

8.25 **NOTE**: In PRML, equation (8.86) contains a typographical error. On the third line, the second summation should sum over x_3, not x_2. Furthermore, in equation (8.79), "$\mu_{x_2 \to f_b}$" (no argument) should be "$\mu_{x_2 \to f_b}(x_2)$".

Starting from (8.63), using (8.73), (8.77) and (8.81)–(8.83), we get

$$\begin{aligned}
\widetilde{p}(x_1) &= \mu_{f_a \to x_1}(x_1) \\
&= \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2) \\
&= \sum_{x_2} f_a(x_1, x_2) \mu_{f_b \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\
&= \sum_{x_2} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) \\
&= \sum_{x_2} \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_2} \sum_{x_3} \sum_{x_4} \widetilde{p}(\mathbf{x}).
\end{aligned}$$

Similarly, starting from (8.63), using (8.73), (8.75) and (8.77)–(8.79), we get

$$\begin{aligned}
\widetilde{p}(x_3) &= \mu_{f_b \to x_3}(x_3) \\
&= \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \to f_b}(x_2) \\
&= \sum_{x_2} f_b(x_2, x_3) \mu_{f_a \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\
&= \sum_{x_2} f_b(x_2, x_3) \sum_{x_1} f_a(x_1, x_2) \sum_{x_4} f_c(x_2, x_4) \\
&= \sum_{x_1} \sum_{x_2} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_1} \sum_{x_2} \sum_{x_4} \widetilde{p}(\mathbf{x}).
\end{aligned}$$

Finally, starting from (8.72), using (8.73), (8.74), (8.77), (8.81) and (8.82), we get

$$\begin{aligned}
\widetilde{p}(x_1, x_2) &= f_a(x_1, x_2)\mu_{x_1 \to f_a}(x_1)\mu_{x_2 \to f_a}(x_2) \\
&= f_a(x_1, x_2)\mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
&= f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_b(x_2, x_4) \\
&= \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_b(x_2, x_4) \\
&= \sum_{x_3} \sum_{x_4} \widetilde{p}(\mathbf{x}).
\end{aligned}$$

8.26 We start by using the product and sum rules to write

$$p(x_a, x_b) = p(x_b|x_a)p(x_a) = \sum_{\mathbf{x}_{\setminus ab}} p(\mathbf{x}) \tag{242}$$

where $\mathbf{x}_{\setminus ab}$ denote the set of all all variables in the graph except x_a and x_b.

We can use the sum-product algorithm from Section 8.4.4 to first evaluate $p(x_a)$, by marginalizing over all other variables (including x_b). Next we successively fix x_a at all its allowed values and for each value, we use the sum-product algorithm to evaluate $p(x_b|x_a)$, by marginalizing over all variables except x_b and x_a, the latter of which will only appear in the formulae at its current, fixed value. Finally, we use (242) to evaluate the joint distribution $p(x_a, x_b)$.

8.27 An example is given by

	$x = 0$	$x = 1$	$x = 2$
$y = 0$	0.0	0.1	0.2
$y = 1$	0.0	0.1	0.2
$y = 2$	0.3	0.1	0.0

for which $\widehat{x} = 2$ and $\widehat{y} = 2$.

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

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 $\{\boldsymbol{\mu}_k\}$.

9.2 Taking the derivative of (9.1), which in this case only involves $\mathbf{x}_n$, w.r.t. $\boldsymbol{\mu}_k$, we get

$$\frac{\partial J}{\partial \mu_k} = -2r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = z(\boldsymbol{\mu}_k).$$

Substituting this into (2.129), with $\boldsymbol{\mu}_k$ replacing θ, we get

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \eta_n(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})$$

where by (9.2), $\boldsymbol{\mu}_k^{\text{old}}$ will be the prototype nearest to $\mathbf{x}_n$ and the factor of 2 has been absorbed into η_n.

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

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