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

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

or equivalently in vector notation

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

Substituting (199) into (6.7) we obtain

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

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

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

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 semi-definite 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

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

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

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

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

$$\frac{1}{2}\|\mathbf{w}\|^2 = \sum_n a_n - \frac{1}{2}\|\mathbf{w}\|^2,$$

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

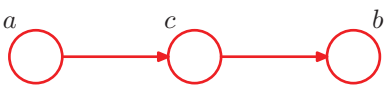


Figure 4 Directed graph representing the joint distribution given in Table 8.2.

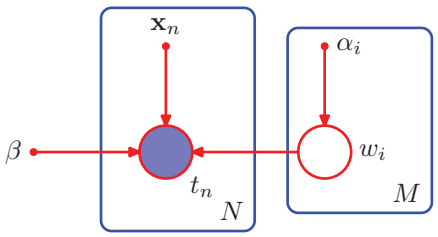


Figure 5 The graphical representation of the relevance vector machine (RVM); Solution 8.5.

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.

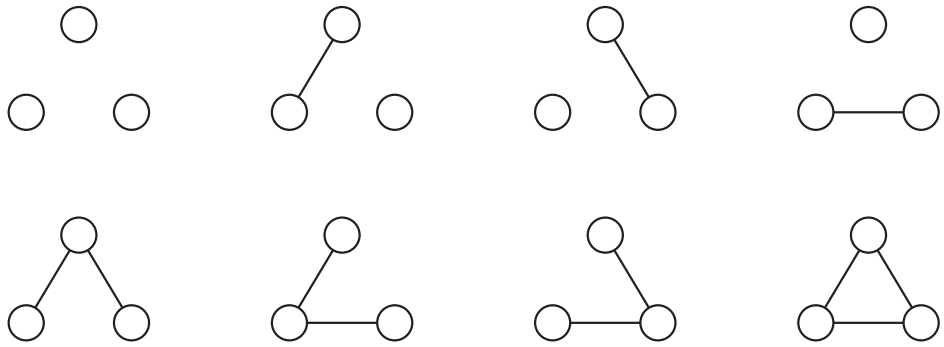


Figure 6 The set of 8 distinct undirected graphs which can be constructed over $M = 3$ nodes.

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.

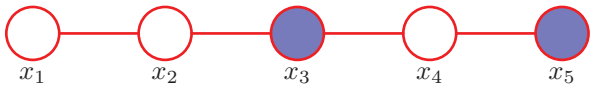
8.16 Observing $\mathbf{x}_N = \widehat{\mathbf{x}}_N$ will only change the initial expression (message) for the β-recursion, which now becomes

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

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.

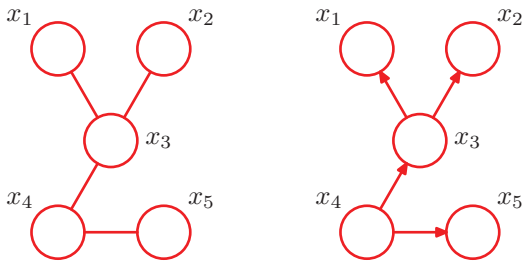


Figure 8 The graph on the left is an undirected tree. If we pick x_4 to be the root node and direct all the edges in the graph to point from the root to the leaf nodes (x_1, x_2 and x_5), we obtain the directed tree shown on the right.

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

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

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

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

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

9.4 From Bayes' theorem we have

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X}}.$$

To maximize this w.r.t. $\boldsymbol{\theta}$, we only need to consider the numerator on the r.h.s. and we shall find it more convenient to operate with the logarithm of this expression,

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \tag{243}$$

where we recognize the first term as the l.h.s. of (9.29). Thus we follow the steps in Section 9.3 in dealing with the latent variables, $\mathbf{Z}$. Note that the second term in (243) does not involve $\mathbf{Z}$ and will not affect the corresponding E-step, which hence gives (9.30). In the M-step, however, we are maximizing over $\boldsymbol{\theta}$ and so we need to include the second term of (243), yielding

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta}).$$

9.5 Consider any two of the latent variable nodes, which we denote $\mathbf{z}_l$ and $\mathbf{z}_m$. We wish to determine whether these variables are independent, conditioned on the observed data $\mathbf{x}_1,\ldots,\mathbf{x}_N$ and on the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$. To do this we consider every possible path from $\mathbf{z}_l$ to $\mathbf{z}_m$. The plate denotes that there are N separate copies of the notes $\mathbf{z}_n$ and $\mathbf{x}_n$. Thus the only paths which connect $\mathbf{z}_l$ and $\mathbf{z}_m$ are those which go via one of the parameter nodes $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ or $\boldsymbol{\pi}$. Since we are conditioning on these parameters they represent observed nodes. Furthermore, any path through one of these parameter nodes must be tail-to-tail at the parameter node, and hence all such paths are blocked. Thus $\mathbf{z}_l$ and $\mathbf{z}_m$ are independent, and since this is true for any pair of such nodes it follows that the posterior distribution factorizes over the data set.

9.6 In this case, the expected complete-data log likelihood function becomes

$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln\pi_k + \ln\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}\right)\right\}$$

where $\gamma(z_{nk})$ is defined in (9.16). Differentiating this w.r.t. $\boldsymbol{\Sigma}^{-1}$, using (C.24) and (C.28), we get

$$\frac{N}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^{\mathrm{T}}$$

where we have also used that $\sum_{k=1}^{K} \gamma(z_{nk}) = 1$ for all n. Setting this equal to zero and rearranging, we obtain

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}.$$

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

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

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

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

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

which gives

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

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

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

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

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

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

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

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

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

9.9 If we differentiate (9.40) w.r.t. $\boldsymbol{\Sigma}_k^{-1}$, while keeping the $\gamma(z_{nk})$ fixed, we get

$$\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})\right]=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\frac{1}{2}\left(\boldsymbol{\Sigma}_k-(x_n-\boldsymbol{\mu}_k)(x_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\right)$$

where we have used (C.28). Setting this equal to zero and rearranging, we obtain (9.19).

Appendix E For π_k, we add a Lagrange multiplier term to (9.40) to enforce the constraint

$$\sum_{k=1}^{K}\pi_k=1$$

yielding

$$\mathbb{E}_{\mathbf{Z}}\left[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})\right]+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right).$$

Differentiating this w.r.t. π_k, we get

$$\sum_{n=1}^{N}\gamma(z_{nk})\frac{1}{\pi_k}+\lambda=\frac{N_k}{\pi_k}+\lambda$$

where we have used (9.18). Setting this equal to zero and rearranging, we get

$$N_k=-\pi_k\lambda.$$

Summing both sides over k, making use of (9.9), we see that $-\lambda=N$ and thus

$$\pi_k=\frac{N_k}{N}.$$

9.10 For the mixture model the joint distribution can be written

$$p(\mathbf{x}_a,\mathbf{x}_b)=\sum_{k=1}^{K}\pi_k p(\mathbf{x}_a,\mathbf{x}_b|k).$$

We can find the conditional density $p(\mathbf{x}_b|\mathbf{x}_a)$ by making use of the relation

$$p(\mathbf{x}_b|\mathbf{x}_a)=\frac{p(\mathbf{x}_a,\mathbf{x}_b)}{p(\mathbf{x}_a)}.$$

For mixture model the marginal density of $\mathbf{x}_a$ is given by

$$p(\mathbf{x}_a) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_a|k)$$

where

$$p(\mathbf{x}_a|k) = \int p(\mathbf{x}_a, \mathbf{x}_b|k)\, \mathrm{d}\mathbf{x}_b.$$

Thus we can write the conditional density in the form

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{\sum_{k=1}^{K} \pi_k p(\mathbf{x}_a, \mathbf{x}_b|k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_a|j)}.$$

Now we decompose the numerator using

$$p(\mathbf{x}_a, \mathbf{x}_b|k) = p(\mathbf{x}_b|\mathbf{x}_a, k)p(\mathbf{x}_a|k)$$

which allows us finally to write the conditional density as a mixture model of the form

$$p(\mathbf{x}_b|\mathbf{x}_a) = \sum_{k=1}^{K} \lambda_k p(\mathbf{x}_b|\mathbf{x}_a, k) \tag{244}$$

where the mixture coefficients are given by

$$\lambda_k \equiv p(k|\mathbf{x}_a) = \frac{\pi_k p(\mathbf{x}_a|k)}{\sum_j \pi_j p(\mathbf{x}_a|j)} \tag{245}$$

and $p(\mathbf{x}_b|\mathbf{x}_a, k)$ is the conditional for component k.

9.11 As discussed in Section 9.3.2, $\gamma(z_{nk}) \to r_{nk}$ as $\epsilon \to 0$. $\boldsymbol{\Sigma}_k = \epsilon\mathbf{I}$ for all k and are no longer free parameters. π_k will equal the proportion of data points assigned to cluster k and assuming reasonable initialization of $\boldsymbol{\pi}$ and $\{\boldsymbol{\mu}_k\}$, π_k will remain strictly positive. In this situation, we can maximize (9.40) w.r.t. $\{\boldsymbol{\mu}_k\}$ independently of $\boldsymbol{\pi}$, leaving us with

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \epsilon\mathbf{I}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(-\frac{1}{2\epsilon}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right) + \text{const.}$$

which equal the negative of (9.1) upto a scaling factor (which is independent of $\{\boldsymbol{\mu}_k\}$).

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

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

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

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

to give

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

9.13 The expectation of $\mathbf{x}$ under the mixture distribution is given by

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

Now we make use of (9.58) and (9.59) to give

$$\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
&= \sum_{n=1}^{N} \mathbf{x}_n \frac{1}{N} \sum_{k=1}^{K} \gamma(z_{nk}) \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\
&= \overline{\mathbf{x}}
\end{aligned}$$

where we have used $\pi_k = N_k/N$, and the fact that $\gamma(z_{nk})$ are posterior probabilities and hence $\sum_k \gamma(z_{nk}) = 1$.

Now suppose we initialize a mixture of Bernoulli distributions by setting the means to a common value $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}}$ for $k = 1, \ldots, K$ and then run the EM algorithm. In the E-step we first compute the responsibilities which will be given by

$$\gamma(z_{nk}) = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)} = \frac{\pi_k}{\sum_{j=1}^{K} \pi_j} = \pi_k$$

and are therefore independent of n. In the subsequent M-step the revised means are given by

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \\ &= \frac{1}{N_k}\pi_k\sum_{n=1}^{N}\mathbf{x}_n \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\ &= \overline{\mathbf{x}}\end{aligned}$$

where again we have made use of $\pi_k = N_k/N$. Note that since these are again the same for all k it follows from the previous discussion that the responsibilities on the next E-step will again be given by $\gamma(z_{nk}) = \pi_k$ and hence will be unchanged. The revised mixing coefficients are given by

$$\frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk}) = \pi_k$$

and so are also unchanged. Thus the EM algorithm has converged and no further changes will take place with subsequent E and M steps. Note that this is a degenerate solution in which all of the components of the mixture are identical, and so this distribution is equivalent to a single multivariate Bernoulli distribution.

9.14 Forming the product of (9.52) and (9.53), we get

$$\prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k}\prod_{j=1}^{K}\pi_j^{z_j} = \prod_{k=1}^{K}\left(p(\mathbf{x}|\boldsymbol{\mu}_k)\pi_k\right)^{z_k}.$$

If we marginalize this over $\mathbf{z}$, we get

$$\begin{aligned}\sum_{\mathbf{z}}\prod_{k=1}^{K}\left(p(\mathbf{x}|\boldsymbol{\mu}_k)\pi_k\right)^{z_k} &= \sum_{j=1}^{K}\prod_{k=1}^{K}\left(p(\mathbf{x}|\boldsymbol{\mu}_k)\pi_k\right)^{I_{jk}} \\ &= \sum_{j=1}^{K}\pi_j p(\mathbf{x}|\boldsymbol{\mu}_j)\end{aligned}$$

where we have exploited the 1-of-K coding scheme used for $\mathbf{z}$.

9.15 This is easily shown by calculating the derivatives of (9.55), setting them to zero and

solve for μ_{ki}. Using standard derivatives, we get

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

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

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

which equals (9.59) when written in vector form.

9.16 This is identical with the maximization w.r.t. π_k in the Gaussian mixture model, detailed in the second half of Solution 9.9.

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

9.18 From Solution 9.4, which dealt with MAP estimation for a general mixture model, we know that the E-step will remain unchanged. In the M-step we maximize

$$\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)+\ln p(\boldsymbol{\theta})$$

which in the case of the given model becomes,

$$\begin{aligned}&\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln\pi_k+\sum_{i=1}^{D}\left[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right]\right\}\\ &+\sum_{j=1}^{K}\sum_{i'=1}^{D}\left\{(a_j-1)\ln\mu_{ji'}+(b_j-1)\ln(1-\mu_{ji'})\right\}+\sum_{l=1}^{K}(\alpha_l-1)\ln\pi_l \qquad (246)\end{aligned}$$

where we have used (9.55), (2.13) and (2.38), and we have dropped terms independent of $\{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\pi}$. Note that we have assumed that each parameter μ_{ki} has the same prior for each i, but this can differ for different components k.

Differentiating (246) w.r.t. μ_{ki} yields

$$\begin{aligned}&\sum_{n=1}^{N}\gamma(z_{nk})\left\{\frac{x_{ni}}{\mu_{ki}}-\frac{1-x_{ni}}{1-\mu_{ki}}\right\}+\frac{a_k}{\mu_{ki}}-\frac{1-b_k}{1-\mu_{ki}}\\ &\qquad=\frac{N_k\overline{x}_{ki}+a-1}{\mu_{ki}}-\frac{N_k-N_k\overline{x}_{ki}+b-1}{1-\mu_{ki}}\end{aligned}$$

where N_k is given by (9.57) and $\overline{x}_{ki}$ is the i^{th} element of $\overline{\mathbf{x}}$ defined in (9.58). Setting this equal to zero and rearranging, we get

$$\mu_{ki} = \frac{N_k \overline{x}_{ki} + a - 1}{N_k + a - 1 + b - 1}. \tag{247}$$

Note that if $a_k = b_k = 1$ for all k, this reduces to the standard maximum likelihood result. Also, as N becomes large, (247) will approach the maximum likelihood result.

Appendix E

When maximizing w.r.t. π_k, we need to enforce the constraint $\sum_k \pi_k = 1$, which we do by adding a Lagrange multiplier term to (246). Dropping terms independent of $\boldsymbol{\pi}$ we are left with

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}) \ln \pi_k + \sum_{l=1}^{K} (\alpha_l - 1) \ln \pi_l + \lambda \left(\sum_{j=1}^{K} \pi_j - 1 \right).$$

Differentiating this w.r.t. π_k, we get

$$\frac{N_k + \alpha_k - 1}{\pi_k} + \lambda$$

and setting this equal to zero and rearranging, we have

$$N_k + \alpha_k - 1 = -\lambda \pi_k.$$

Summing both sides over k, using $\sum_k \pi_k = 1$, we see that $-\lambda = N + \alpha_0 - K$, where α_0 is given by (2.39), and thus

$$\pi_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}. \tag{248}$$

Also in this case, if $\alpha_k = 1$ for all k, we recover the maximum likelihood result exactly. Similarly, as N gets large, (248) will approach the maximum likelihood result.

9.19 As usual we introduce a latent variable $\mathbf{z}_n$ corresponding to each observation. The conditional distribution of the observed data set, given the latent variables, is then

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}.$$

Similarly, the distribution of the latent variables is given by

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \pi_k^{z_{nk}}.$$

The expected value of the complete-data log likelihood function is given by

$$\sum_{n=1}^{N}\sum_{k=1}^{k}\gamma(z_{nk})\left\{\ln\pi_k+\sum_{i=1}^{D}\sum_{j=1}^{M}x_{nij}\ln\mu_{kij}\right\}$$

where as usual we have defined responsibilities given by

$$\gamma(z_{nk})=\mathbb{E}[z_{nk}]=\frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{M}\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}.$$

These represent the E-step equations.

To derive the M-step equations we add to the expected complete-data log likelihood function a set of Lagrange multiplier terms given by

$$\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)+\sum_{k=1}^{K}\sum_{i=1}^{D}\eta_{ki}\left(\sum_{j=1}^{M}\mu_{kij}-1\right)$$

to enforce the constraint $\sum_k \pi_k = 1$ as well as the set of constraints

$$\sum_{j=1}^{M}\mu_{kij}=1$$

for all values of i and k. Maximizing with respect to the mixing coefficients π_k, and eliminating the Lagrange multiplier λ in the usual way, we obtain

$$\pi_k=\frac{N_k}{N}$$

where we have defined

$$N_k=\sum_{n=1}^{N}\gamma(z_{nk}).$$

Similarly maximizing with respect to the parameters μ_{kij}, and again eliminating the Lagrange multipliers, we obtain

$$\mu_{kij}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_{nij}.$$

This is an intuitively reasonable result which says that the value of μ_{kij} for component k is given by the fraction of those counts assigned to component k which have non-zero values of the corresponding elements i and j.

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

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

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

9.21 Taking the derivative of (9.62) w.r.t. β, we obtain

$$\frac{\partial}{\partial\beta}\mathbb{E}\left[\ln p(\mathbf{t},\mathbf{w}|\alpha,\beta)\right] = \frac{N}{2}\frac{1}{\beta} - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left[(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right]. \tag{249}$$

From (3.49)–(3.51), we see that

$$\begin{aligned}\mathbb{E}\left[(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right] &= \mathbb{E}\left[t_n^2 - 2t_n\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}]\right]\\ &= t_n^2 - 2t_n\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}_n + \mathrm{Tr}\left[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}(\mathbf{m}_N\mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N)\right]\\ &= (t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}_n)^2 + \mathrm{Tr}\left[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\mathbf{S}_N\right].\end{aligned}$$

Substituting this into (249) and rearranging, we obtain

$$\frac{1}{\beta} = \frac{1}{N}\left(\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \mathrm{Tr}\left[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{S}_N\right]\right).$$

9.22 **NOTE**: In PRML, a pair of braces is missing from (9.66), which should read

$$\mathbb{E}_{\mathbf{w}}\left[\ln\left\{p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\alpha})\right\}\right].$$

Moreover $\mathbf{m}_N$ should be $\mathbf{m}$ in the numerator on the r.h.s. of (9.68).
Using (7.76)–(7.83) and associated definitions, we can rewrite (9.66) as

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}}\left[\ln\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}\right) + \ln\mathcal{N}\left(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1}\right)\right]\\ &= \frac{1}{2}\mathbb{E}_{\mathbf{w}}\left[N\ln\beta - \beta\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \sum_{i=1}^{M}\ln\alpha_i - \mathrm{Tr}\left[\mathbf{A}\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\right] + \mathrm{const}\\ &= \frac{1}{2}\Bigg(N\ln\beta - \beta\left(\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 + \mathrm{Tr}[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}]\right)\\ &\qquad + \sum_{i=1}^{M}\ln\alpha_i - \mathrm{Tr}[\mathbf{A}(\mathbf{m}\mathbf{m}^{\mathrm{T}} + \boldsymbol{\Sigma})]\Bigg) + \mathrm{const}.\end{aligned} \tag{250}$$

Differentiating this w.r.t. α_i, using (C.23), and setting the result equal to zero, we get

$$\frac{1}{2}\frac{1}{\alpha_i} - \frac{1}{2}\left(m_i^2 + \Sigma_{ii}\right) = 0$$

which we can rearrange to obtain (9.67).

Differentiating (250) w.r.t. β and setting the result equal to zero we get

$$\frac{N}{2}\frac{1}{\beta} - \frac{1}{2}\left(\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}\|^2 + \mathrm{Tr}\left[\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Sigma}\right]\right) = 0. \tag{251}$$

Using (7.83), (C.6) and (C.7) together with the fact that $\mathbf{A}$ is diagonal, we can rewrite $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Sigma}$ as follows:

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

Using this together with (7.89), we obtain (9.68) from (251).

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

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

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

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

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

which is easily re-written as (9.67).

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

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

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

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

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

9.24 This is analogous to Solution 10.1, with the integrals replaced by sums.

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

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

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

9.26 From (9.18) we get

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk}). \tag{252}$$

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

$$N_k^{\text{new}} = \sum_{n\neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}). \tag{253}$$

Combining this with (252), we get (9.79).

Similarly, from (9.17) we have

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

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

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

where we have used (9.79).

9.27 Following the treatment of $\boldsymbol{\mu}_k$ in Solution 9.26, (9.19) gives

$$\boldsymbol{\Sigma}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})^{\mathrm{T}}$$

where N_k^{old} is given by (252). Recomputing the responsibilities $\gamma(z_{mk})$, and using (253), we get

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

where we have also used (9.79).

For π_k, (9.22) gives

$$\pi_k^{\text{old}}=\frac{N_k^{\text{old}}}{N}=\frac{1}{N}\sum_{n=1}^{N}\gamma^{\text{old}}(z_{nk})$$

and thus recomputing $\gamma(z_{nk})$ we get

$$\begin{aligned}
\pi_k^{\text{new}} &= \frac{1}{N}\left(\sum_{n\neq m}^{N}\gamma^{\text{old}}(z_{nk})+\gamma^{\text{new}}(z_{mk})\right)\\
&= \frac{1}{N}\left(N\pi_k^{\text{old}}-\gamma^{\text{old}}(z_{mk})+\gamma^{\text{new}}(z_{mk})\right)\\
&= \pi_k^{\text{old}}-\frac{\gamma^{\text{old}}(z_{mk})}{N}+\frac{\gamma^{\text{new}}(z_{mk})}{N}.
\end{aligned}$$

Chapter 10 Variational Inference and EM

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

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

Rearranging this, we immediately get (10.2).

10.2 By substituting $\mathbb{E}[z_1] = m_1 = \mu_1$ and $\mathbb{E}[z_2] = m_2 = \mu_2$ in (10.13) and (10.15), respectively, we see that both equations are satisfied and so this is a solution.

To show that it is indeed the only solution when $p(\mathbf{z})$ is non-singular, we first substitute $\mathbb{E}[z_1] = m_1$ and $\mathbb{E}[z_2] = m_2$ in (10.13) and (10.15), respectively. Next, we substitute the r.h.s. of (10.13) for m_1 in (10.15), yielding

$$\begin{aligned}m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\,(m_2 - \mu_2) - \mu_1\right) \\ &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2)\end{aligned}$$

which we can rewrite as

$$m_2\left(1 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\right) = \mu_2\left(1 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\right).$$

Thus, unless $\Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} = 1$, the solution $\mu_2 = m_2$ is unique. If $p(\mathbf{z})$ is non-singular,

$$|\mathbf{\Lambda}| = \Lambda_{11}\Lambda_{22} - \Lambda_{21}\Lambda_{12} \neq 0$$

which we can rewrite as

$$\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{21}\Lambda_{12} \neq 1$$

as desired. Since $\mu_2 = m_2$ is the unique solution to (10.15), $\mu_1 = m_1$ is the unique solution to (10.13).

10.3 Starting from (10.16) and optimizing w.r.t. $q_j(\mathbf{Z}_j)$, we get

$$\begin{aligned}
\mathrm{KL}(p \,\|\, q) &= -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i)\right] \mathrm{d}\mathbf{Z} + \text{const.} \\
&= -\int \left(p(\mathbf{Z}) \ln q_j(\mathbf{Z}_j) + p(\mathbf{Z}) \sum_{i \neq j} \ln q_i(\mathbf{Z}_i)\right) \mathrm{d}\mathbf{Z} + \text{const.} \\
&= -\int p(\mathbf{Z}) \ln q_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z} + \text{const.} \\
&= -\int \ln q_j(\mathbf{Z}_j) \left[\int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i\right] \mathrm{d}\mathbf{Z}_j + \text{const.} \\
&= -\int F_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j + \text{const.},
\end{aligned}$$

where terms independent of $q_j(\mathbf{Z}_j)$ have been absorbed into the constant term and we have defined

$$F_j(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i.$$

We use a Lagrange multiplier to ensure that $q_j(\mathbf{Z}_j)$ integrates to one, yielding

$$-\int F_j(\mathbf{Z}_j) \ln q_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j + \lambda \left(\int q_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j - 1\right).$$

Using the results from Appendix D, we then take the functional derivative of this w.r.t. q_j and set this to zero, to obtain

$$-\frac{F_j(\mathbf{Z}_j)}{q_j(\mathbf{Z}_j)} + \lambda = 0.$$

From this, we see that

$$\lambda q_j(\mathbf{Z}_j) = F_j(\mathbf{Z}_j).$$

Integrating both sides over $\mathbf{Z}_j$, we see that, since $q_j(\mathbf{Z}_j)$ must intgrate to one,

$$\lambda = \int F_j(\mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j = \int \left[\int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i\right] \mathrm{d}\mathbf{Z}_j = 1,$$

and thus

$$q_j(\mathbf{Z}_j) = F_j(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i.$$

10.4 The Kullback-Leibler divergence takes the form

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Substituting the Gaussian for $q(\mathbf{x})$we obtain

$$\begin{aligned}
\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \left\{ -\frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \mathrm{d}\mathbf{x} + \text{const.} \\
&= \frac{1}{2} \left\{ \ln |\boldsymbol{\Sigma}| + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \right] \right) \right\} + \text{const.} \\
&= \frac{1}{2} \left\{ \ln |\boldsymbol{\Sigma}| + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - 2 \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbb{E}[\mathbf{x}] + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}} \right] \right) \right\} \\
&\quad + \text{const.} \qquad (254)
\end{aligned}$$

Differentiating this w.r.t. $\boldsymbol{\mu}$, using results from Appendix C, and setting the result to zero, we see that

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]. \qquad (255)$$

Similarly, differentiating (254) w.r.t. $\boldsymbol{\Sigma}^{-1}$, again using results from Appendix C and also making use of (255) and (1.42), we see that

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}} \right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} = \mathrm{cov}[\mathbf{x}].$$

10.5 We assume that $q(\mathbf{Z}) = q(\mathbf{z})q(\boldsymbol{\theta})$ and so we can optimize w.r.t. $q(\mathbf{z})$ and $q(\boldsymbol{\theta})$ independently.

For $q(\mathbf{z})$, this is equivalent to minimizing the Kullback-Leibler divergence, (10.4), which here becomes

$$\mathrm{KL}(\, q \parallel p\,) = -\iint q\,(\boldsymbol{\theta})\, q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{X})}{q\,(\mathbf{z})\, q\,(\boldsymbol{\theta})}\, \mathrm{d}\mathbf{z}\, \mathrm{d}\boldsymbol{\theta}.$$

For the particular chosen form of $q(\boldsymbol{\theta})$, this is equivalent to

$$\begin{aligned}
\mathrm{KL}(\, q \parallel p\,) &= -\int q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z}, \boldsymbol{\theta}_0 \mid \mathbf{X})}{q\,(\mathbf{z})}\, \mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X})\, p\,(\boldsymbol{\theta}_0 \mid \mathbf{X})}{q\,(\mathbf{z})}\, \mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\,(\mathbf{z}) \ln \frac{p\,(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X})}{q\,(\mathbf{z})}\, \mathrm{d}\mathbf{z} + \text{const.},
\end{aligned}$$

where const accumulates all terms independent of $q(\mathbf{z})$. This KL divergence is minimized when $q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{\theta}_0, \mathbf{X})$, which corresponds exactly to the E-step of the EM algorithm.

To determine $q(\boldsymbol{\theta})$, we consider

$$\begin{aligned}&\int q(\boldsymbol{\theta}) \int q(\mathbf{z}) \ln \frac{p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}) q(\mathbf{z})} \, \mathrm{d}\mathbf{z} \, \mathrm{d}\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \, \mathbb{E}_{q(\mathbf{z})} [\ln p(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z})] \, \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} + \text{const.}\end{aligned}$$

where the last term summarizes terms independent of $q(\boldsymbol{\theta})$. Since $q(\boldsymbol{\theta})$ is constrained to be a point density, the contribution from the entropy term (which formally diverges) will be constant and independent of $\boldsymbol{\theta}_0$. Thus, the optimization problem is reduced to maximizing expected complete log posterior distribution

$$\mathbb{E}_{q(\mathbf{z})} [\ln p(\mathbf{X}, \boldsymbol{\theta}_0, \mathbf{z})],$$

w.r.t. $\boldsymbol{\theta}_0$, which is equivalent to the M-step of the EM algorithm.

10.6 We start by rewriting (10.19) as

$$\mathrm{D}(p\|q) = \frac{4}{1-\alpha^2} \left(1 - \int p(x)^{\gamma_p} q(x)^{\gamma_q} \, \mathrm{d}x\right) \tag{256}$$

so that

$$\gamma_p = \frac{1+\alpha}{2} \quad \text{and} \quad \gamma_q = \frac{1-\alpha}{2}. \tag{257}$$

We note that

$$\lim_{\alpha \to 1} \gamma_q = 0 \tag{258}$$

$$\lim_{\alpha \to 1} \gamma_p = 1 \tag{259}$$

$$1 - \gamma_p = \gamma_q. \tag{260}$$

Based on observation (258), we make a Maclaurin expansion of $q(x)^{\gamma_q}$ in γ_q as follows

$$q^{\gamma_q} = \exp(\gamma_q \ln q) = 1 + \gamma_q \ln q + O\left(\gamma_q^2\right) \tag{261}$$

where q is a shorthand notation for $q(x)$. Similarly, based on (259), we make a Taylor expansion of $p(x)^{\gamma_p}$ in γ_p around 1,

$$\begin{aligned}p^{\gamma_p} &= \exp(\gamma_p \ln p) \\ &= p - (1-\gamma_p) p \ln p + O\left((\gamma_p - 1)^2\right) \\ &= p - \gamma_q p \ln p + O\left(\gamma_q^2\right)\end{aligned} \tag{262}$$

where we have used (260) and we have adopted corresponding shorthand notation for $p(x)$.

Using (261) and (262), we can rewrite (256) as

$$\begin{aligned}
&\mathrm{D}(p\|q) \\
&= \frac{4}{1-\alpha^2}\left(1-\int\left[p-\gamma_q p\ln p+O\left(\gamma_q^2\right)\right]\left[1+\gamma_q\ln q+O\left(\gamma_q^2\right)\right]\,\mathrm{d}x\right) \\
&= \frac{4}{1-\alpha^2}\left(1-\int p+\gamma_q\left(p\ln q-p\ln p\right)\,\mathrm{d}x+O\left(\gamma_q^2\right)\right) \qquad (263)
\end{aligned}$$

where $O\left(\gamma_q^2\right)$ account for all higher order terms. From (257) we have

$$\begin{aligned}
\frac{4}{1-\alpha^2}\gamma_q &= \frac{2(1-\alpha)}{1-\alpha^2}=\frac{2}{(1+\alpha)} \\
\frac{4}{1-\alpha^2}\gamma_q^2 &= \frac{(1-\alpha)^2}{1-\alpha^2}=\frac{1-\alpha}{(1+\alpha)}
\end{aligned}$$

and thus

$$\begin{aligned}
\lim_{\alpha\to 1}\frac{4}{1-\alpha^2}\gamma_q &= 1 \\
\lim_{\alpha\to 1}\frac{4}{1-\alpha^2}\gamma_q^2 &= 0.
\end{aligned}$$

Using these results together with (259), and (263), we see that

$$\lim_{\alpha\to 1}\mathrm{D}(p\|q)=-\int p(\ln q-\ln p)\,\mathrm{d}x=\mathrm{KL}(p\|q).$$

The proof that $\alpha\to -1$ yields $\mathrm{KL}(q\|p)$ is analogous.

10.7 **NOTE**: See note in Solution 10.9.

We take the μ-dependent term from the last line of (10.25) as our starting point. We can rewrite this as follows

$$\begin{aligned}
&-\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0\left(\mu-\mu_0\right)^2+\sum_{n=1}^{N}\left(x_n-\mu\right)^2\right\} \\
&= -\frac{\mathbb{E}[\tau]}{2}\left\{\left(\lambda_0+N\right)\mu^2+\sum_{n=1}^{N}x_n^2-2\mu\left(\lambda_0\mu_0+N\overline{x}\right)\right\} \\
&= -\frac{\mathbb{E}[\tau]}{2}\left\{\left(\lambda_0+N\right)\left(\mu-\frac{\lambda_0\mu_0+N\overline{x}}{\lambda_0+N}\right)^2+\sum_{n=1}^{N}x_n^2-\frac{\left(\lambda_0 m u_0+N\overline{x}\right)^2}{\lambda_0+N}\right\}
\end{aligned}$$

where in the last step we have completed the square over μ. The last two terms are indepenedent of μ and hence can be dropped. Taking the exponential of the remainder, we obtain an unnormalized Gaussian with mean and precision given by (10.26) and (10.27), respectively.

For the posterior over τ, we take the last two lines of (10.28) as our starting point. Gathering terms that multiply τ and $\ln \tau$ into two groups, we can rewrite this as

$$\left(a_0 + \frac{N+1}{2} - 1\right) \ln \tau - \left(b_0 + \frac{1}{2}\mathbb{E}\left[\sum_{n=1}^{N}(x-\mu)^2 + \lambda_0\left(\mu - \mu_0\right)\right]\right)\tau + \text{const.}$$

Taking the exponential of this we get an unnormalized Gamma distribution with shape and inverse scale parameters given by (10.29) and (10.30), respectively.

10.8 **NOTE**: See note in Solution 10.9.

If we substitute the r.h.s. of (10.29) and (10.30) for a and b, respectively, in (B.27) and (B.28), we get

$$\begin{aligned}
\mathbb{E}[\tau] &= \frac{2a_0 + N + 1}{2b_0 + \mathbb{E}\left[\lambda_0(\mu - \mu_0) + \sum_{n=1}^{N}(x_n - \mu)^2\right]} \\
\text{var}[\tau] &= \frac{2a_0 + N + 1}{2\left(b_0 + \frac{1}{2}\mathbb{E}\left[\lambda_0(\mu - \mu_0) + \sum_{n=1}^{N}(x_n - \mu)^2\right]\right)^2} \\
&= \frac{\mathbb{E}[\tau]}{b_0 + \frac{1}{2}\mathbb{E}\left[\lambda_0(\mu - \mu_0) + \sum_{n=1}^{N}(x_n - \mu)^2\right]}
\end{aligned}$$

From this we see directly that

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}[\tau] &= \frac{N}{\mathbb{E}\left[\sum_{n=1}^{N}(x_n - \mu)^2\right]} \\
\lim_{N\to\infty} \text{var}[\tau] &= 0
\end{aligned}$$

as long as the data set is not singular.

10.9 **NOTE**: In PRML, an extra term of $1/2$ should be added to the r.h.s. of (10.29), with consequential changes to (10.31) and (10.33), which should read

$$\frac{1}{\mathbb{E}[\tau]} = \mathbb{E}\left[\frac{1}{N+1}\sum_{n=1}^{N}(x_n - \mu)^2\right] = \frac{N}{N+1}\left(\overline{x^2} - 2\overline{x}\mathbb{E}[\mu] + \mathbb{E}[\mu^2]\right)$$

and

$$\frac{1}{\mathbb{E}[\tau]} = (\overline{x^2} - \overline{x}^2) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})^2$$

respectively.

Assuming $a_0 = b_0 = \lambda_0 = 0$, (10.29), (10.30) and (B.27) give

$$\begin{aligned}\frac{1}{\mathbb{E}[\tau]} &= \frac{1}{N+1}\sum_{n=1}^{N}(x_n-\mu)^2 \\ &= \frac{1}{N+1}\sum_{n=1}^{N}\left(x_n^2 - 2x_n\mu + \mu^2\right)\end{aligned}$$

Taking the expectation of this under $q(\mu)$, making use of (10.32), we get

$$\begin{aligned}\frac{1}{\mathbb{E}[\tau]} &= \frac{1}{N+1}\sum_{n=1}^{N}\left(x_n^2 - 2x_n\overline{x} + \overline{x}^2 + \frac{1}{N\mathbb{E}[\tau]}\right) \\ &= \frac{N}{N+1}\left(\frac{1}{N\mathbb{E}[\tau]} - \overline{x}^2 + \frac{1}{N}\sum_{n=1}^{N}x_n^2\right)\end{aligned}$$

which we can rearrange to obtain (10.33).

10.10 **NOTE**: In PRML, there are errors that affect this exercise. $\mathcal{L}_m$ used in (10.34) and (10.35) should really be $\mathcal{L}$, whereas $\mathcal{L}_m$ used in (10.36) is given in Solution 10.11 below.

This completely analogous to Solution 10.1. Starting from (10.35), we can use the product rule to get,

$$\begin{aligned}\mathcal{L} &= \sum_m\sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m)\ln\left\{\frac{p(\mathbf{Z},\mathbf{X},m)}{q(\mathbf{Z}|m)\,q(m)}\right\} \\ &= \sum_m\sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m)\ln\left\{\frac{p(\mathbf{Z},m|\mathbf{X})\,p(\mathbf{X})}{q(\mathbf{Z}|m)\,q(m)}\right\} \\ &= \sum_m\sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m)\ln\left\{\frac{p(\mathbf{Z},m|\mathbf{X})}{q(\mathbf{Z}|m)\,q(m)}\right\} + \ln p(\mathbf{X}).\end{aligned}$$

Rearranging this, we obtain (10.34).

10.11 **NOTE**: Consult note preceding Solution 10.10 for some relevant corrections.

We start by rewriting the lower bound as follows

$$\begin{aligned}
\mathcal{L} &= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}, m)}{q(\mathbf{Z}|m)q(m)} \right\} \\
&= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \left\{ \ln p(\mathbf{Z}, \mathbf{X}|m) + \ln p(m) - \ln q(\mathbf{Z}|m) - \ln q(m) \right\} \\
&= \sum_m q(m) \Bigg(\ln p(m) - \ln q(m) \\
&\qquad + \sum_{\mathbf{Z}} q(\mathbf{Z}|m) \{ \ln p(\mathbf{Z}, \mathbf{X}|m) - \ln q(\mathbf{Z}|m) \} \Bigg) \\
&= \sum_m q(m) \left\{ \ln \left(p(m) \exp\{\mathcal{L}_m\} \right) - \ln q(m) \right\}, \qquad (264)
\end{aligned}$$

where

$$\mathcal{L}_m = \sum_{\mathbf{Z}} q(\mathbf{Z}|m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}|m)}{q(\mathbf{Z}|m)} \right\}.$$

We recognize (264) as the negative KL divergence between $q(m)$ and the (not necessarily normalized) distribution $p(m) \exp\{\mathcal{L}_m\}$. This will be maximized when the KL divergence is minimized, which will be the case when

$$q(m) \propto p(m) \exp\{\mathcal{L}_m\}.$$

10.12 This derivation is given in detail in Section 10.2.1, starting with the paragraph containing (10.43) (page 476) and ending with (10.49).

10.13 In order to derive the optimal solution for $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ we start with the result (10.54) and keep only those term which depend on $\boldsymbol{\mu}_k$ or $\boldsymbol{\Lambda}_k$ to give

$$\begin{aligned}
\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= \ln \mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1} \right) + \ln \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0) \\
&\qquad + \sum_{n=1}^N \mathbb{E}[z_{nk}] \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1} \right) + \text{const.} \\
&= -\frac{\beta_0}{2} (\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{1}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Lambda}_k \mathbf{W}_0^{-1} \right) \\
&\quad + \frac{(\nu_0 - D - 1)}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{1}{2} \sum_{n=1}^N \mathbb{E}[z_{nk}] (\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \\
&\quad + \frac{1}{2} \left(\sum_{n=1}^N \mathbb{E}[z_{nk}] \right) \ln |\boldsymbol{\Lambda}_k| + \text{const.} \qquad (265)
\end{aligned}$$

Using the product rule of probability, we can express $\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ as $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ $+ \ln q^\star(\boldsymbol{\Lambda}_k)$. Let us first of all identify the distribution for $\boldsymbol{\mu}_k$. To do this we need only consider terms on the right hand side of (265) which depend on $\boldsymbol{\mu}_k$, giving

$$\begin{aligned}
&\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) \\
&\quad = -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\left[\beta_0 + \sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right]\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\left[\beta_0\mathbf{m}_0 + \sum_{n=1}^{N}\mathbb{E}[z_{nk}]\mathbf{x}_n\right] \\
&\qquad +\text{const.} \\
&\quad = -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\left[\beta_0 + N_k\right]\boldsymbol{\Lambda}_k\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Lambda}_k\left[\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k\right] + \text{const.}
\end{aligned}$$

where we have made use of (10.51) and (10.52). Thus we see that $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ depends quadratically on $\boldsymbol{\mu}_k$ and hence $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ is a Gaussian distribution. Completing the square in the usual way allows us to determine the mean and precision of this Gaussian, giving

$$q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k) \tag{266}$$

where

$$\begin{aligned}
\beta_k &= \beta_0 + N_k \\
\mathbf{m}_k &= \frac{1}{\beta_k}\left(\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k\right).
\end{aligned}$$

Next we determine the form of $q^\star(\boldsymbol{\Lambda}_k)$ by making use of the relation

$$\ln q^\star(\boldsymbol{\Lambda}_k) = \ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) - \ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k).$$

On the right hand side of this relation we substitute for $\ln q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ using (265), and we substitute for $\ln q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ using the result (266). Keeping only those terms which depend on $\boldsymbol{\Lambda}_k$ we obtain

$$\begin{aligned}
\ln q^\star(\boldsymbol{\Lambda}_k) = &-\frac{\beta_0}{2}(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0) + \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_0^{-1}\right) \\
&+ \frac{(\nu_0 - D - 1)}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \\
&+ \frac{1}{2}\left(\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\right)\ln|\boldsymbol{\Lambda}_k| + \frac{\beta_k}{2}(\boldsymbol{\mu}_k - \mathbf{m}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_k) \\
&- \frac{1}{2}\ln|\boldsymbol{\Lambda}_k| + \text{const.} \\
= &\frac{(\nu_k - D - 1)}{2}\ln|\boldsymbol{\Lambda}_k| - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k\mathbf{W}_k^{-1}\right) + \text{const.}
\end{aligned}$$

Here we have defined

$$\begin{aligned}\mathbf{W}_k^{-1} &= \mathbf{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + \sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \\ &\quad -\beta_k(\boldsymbol{\mu}_k - \mathbf{m}_k)(\boldsymbol{\mu}_k - \mathbf{m}_k)^{\mathrm{T}} \\ &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}} \qquad (267) \\ \nu_k &= \nu_0 + \sum_{n=1}^{N}\mathbb{E}[z_{nk}] \\ &= \nu_0 + N_k,\end{aligned}$$

where we have made use of the result

$$\begin{aligned}\sum_{n=1}^{N}\mathbb{E}[z_{nk}]\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} &= \sum_{n=1}^{N}\mathbb{E}[z_{nk}](\mathbf{x}_n - \overline{\mathbf{x}}_k)(\mathbf{x}_n - \overline{\mathbf{x}}_k)^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} \\ &= N_k\mathbf{S}_k + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} \qquad (268)\end{aligned}$$

and we have made use of (10.53). Note that the terms involving $\boldsymbol{\mu}_k$ have cancelled out in (267) as we expect since $q^\star(\boldsymbol{\Lambda}_k)$ is independent of $\boldsymbol{\mu}_k$.

Thus we see that $q^\star(\boldsymbol{\Lambda}_k)$ is a Wishart distribution of the form

$$q^\star(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k).$$

10.14 We can express the required expectation as an integration with respect to the variational posterior distribution $q^\star(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q^\star(\boldsymbol{\Lambda}_k)$. Thus we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\ &= \iint \mathrm{Tr}\left\{\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right\} q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)q^\star(\boldsymbol{\Lambda}_k)\,\mathrm{d}\boldsymbol{\mu}_k\,\mathrm{d}\boldsymbol{\Lambda}_k.\end{aligned}$$

Next we use the result $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k)$ to perform the integration over $\boldsymbol{\mu}_k$ using the standard expressions for expectations under a Gaussian distribution, giving

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_k] &= \mathbf{m}_k \\ \mathbb{E}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}] &= \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}\end{aligned}$$

from which we obtain the expectation with respect to $\boldsymbol{\mu}_k$ in the form

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right] \\ &= \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \mathbf{x}_n\mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_k\mathbf{x}_n^{\mathrm{T}} + \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1} \\ &= (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}.\end{aligned}$$

Finally, taking the expectation with respect to $\boldsymbol{\Lambda}_k$ we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\right]\\
&\quad=\int \mathrm{Tr}\left\{\boldsymbol{\Lambda}_k\left[(\mathbf{x}_n-\mathbf{m}_k)(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}+\beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}\right]\right\}q^\star(\boldsymbol{\Lambda}_k)\,\mathrm{d}\boldsymbol{\Lambda}_k\\
&\quad=\int\left\{(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n-\mathbf{m}_k)+D\beta_k^{-1}\right\}q^\star(\boldsymbol{\Lambda}_k)\,\mathrm{d}\boldsymbol{\Lambda}_k\\
&\quad=D\beta_k^{-1}+\nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k)
\end{aligned}$$

as required. Here we have used $q^\star(\boldsymbol{\Lambda}_k)=\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k,\nu_k)$, together with the standard result for the expectation under a Wishart distribution to give $\mathbb{E}[\boldsymbol{\Lambda}_k]=\nu_k\mathbf{W}_k$.

10.15 By substituting (10.58) into (B.17) and then using (B.24) together with the fact that $\sum_k N_k=N$, we obtain (10.69).

10.16 To derive (10.71) we make use of (10.38) to give

$$\begin{aligned}
&\mathbb{E}[\ln p(D|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})]\\
&\quad=\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left\{\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|]-\mathbb{E}[(\mathbf{x}_n-\boldsymbol{\mu}_k)\boldsymbol{\Lambda}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)]-D\ln(2\pi)\right\}.
\end{aligned}$$

We now use $\mathbb{E}[z_{nk}]=r_{nk}$ together with (10.64) and the definition of $\widetilde{\Lambda}_k$ given by (10.65) to give

$$\begin{aligned}
\mathbb{E}[\ln p(D|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})]=\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\Big\{&\ln\widetilde{\Lambda}_k\\
&-D\beta_k^{-1}-\nu_k(\mathbf{x}_n-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n-\mathbf{m}_k)-D\ln(2\pi)\Big\}.
\end{aligned}$$

Now we use the definitions (10.51) to (10.53) together with the result (268) to give (10.71).

We can derive (10.72) simply by taking the logarithm of $p(\mathbf{z}|\boldsymbol{\pi})$ given by (10.37)

$$\mathbb{E}[\ln p(\mathbf{z}|\boldsymbol{\pi})]=\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\mathbb{E}[\ln\pi_k]$$

and then making use of $\mathbb{E}[z_{nk}]=r_{nk}$ together with the definition of $\widetilde{\pi}_k$ given by (10.65).

10.17 The result (10.73) is obtained by using the definition of $p(\boldsymbol{\pi})$ given by (10.39) together with the definition of $\widetilde{\pi}_k$ given by (10.66).

For the result (10.74) we start with the definition of the prior $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ given by (10.40) to give

$$\mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{k=1}^{K}\left\{D\ln\beta_0 - D\ln(2\pi) + \mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \beta_0\mathbb{E}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)]\right\} + K\ln B(\mathbf{W}_0, \nu_0) + \sum_{k=1}^{K}\left\{\frac{(\nu_0 - D - 1)}{2}\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \frac{1}{2}\mathrm{Tr}(\mathbf{W}_0^{-1}\mathbb{E}[\boldsymbol{\Lambda}_k])\right\}.$$

Now consider the term $\mathbb{E}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)]$. To evaluate this expression we first perform the expectation with respect to $q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)$ then the subsequently perform the expectation with respect to $q^\star(\boldsymbol{\Lambda}_k)$. Using the standard results for moments under a Gaussian we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_k] &= \mathbf{m}_k \\ \mathbb{E}[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}] &= \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k^{-1}\boldsymbol{\Lambda}_k^{-1}\end{aligned}$$

and hence

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}_k\boldsymbol{\Lambda}_k}[(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)] &= \mathrm{Tr}\left(\mathbb{E}_{\boldsymbol{\mu}_k\boldsymbol{\Lambda}_k}\left[\boldsymbol{\Lambda}_k(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}}\right]\right) \\ &= \mathrm{Tr}\left(\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\boldsymbol{\Lambda}_k(\beta_k^{-1}\boldsymbol{\Lambda}_k^{-1} + \mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} - \mathbf{m}_k\mathbf{m}_0^{\mathrm{T}} + \mathbf{m}_0\mathbf{m}_0^{\mathrm{T}})\right]\right) \\ &= K\beta_k^{-1} + (\mathbf{m}_k - \mathbf{m}_0)^{\mathrm{T}}\mathbb{E}[\boldsymbol{\Lambda}_k](\mathbf{m}_k - \mathbf{m}_0).\end{aligned}$$

Now we use (B.80) to give $\mathbb{E}[\boldsymbol{\Lambda}_k] = \nu_k\mathbf{W}_k$ and $\mathbb{E}[\ln\boldsymbol{\Lambda}_k] = \ln\widetilde{\Lambda}_k$ from (10.65) to give (10.74).

For (10.75) we take use the result (10.48) for $q^\star(\mathbf{z})$ to give

$$\mathbb{E}[\ln q(\mathbf{z})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\ln r_{nk}$$

and using $\mathbb{E}[z_{nk}] = r_{nk}$ we obtain (10.75).

The solution (10.76) for $\mathbb{E}[\ln q(\boldsymbol{\pi})]$ is simply the negative entropy of the corresponding Dirichlet distribution (10.57) and is obtained from (B.22).

Finally, we need the entropy of the Gaussian-Wishart distribution $q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. First of all we note that this distribution factorizes into a product of factors $q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ and the entropy of the product is the sum of the entropies of the individual terms, as is easily verified. Next we write

$$\ln q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \ln q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) + \ln q(\boldsymbol{\Lambda}_k).$$

Consider first the quantity $\mathbb{E}[\ln q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)]$. Taking the expectation first with respect to $\boldsymbol{\mu}_k$ we can make use of the standard result (B.41) for the entropy of a Gaussian to

give

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}_k\boldsymbol{\Lambda}_k}[\ln q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)] &= \mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\frac{1}{2}\ln|\boldsymbol{\Lambda}_k| + \frac{D}{2}\left(\ln\beta_k - 1 - \ln(2\pi)\right)\right] \\ &= \frac{1}{2}\ln\widetilde{\Lambda}_k + \frac{D}{2}\left(\ln\beta_k - 1 - \ln(2\pi)\right).\end{aligned}$$

The term $\mathbb{E}[\ln q(\boldsymbol{\Lambda}_k)]$ is simply the negative entropy of a Wishart distribution, which we write as $-\mathrm{H}[q(\boldsymbol{\Lambda}_k)]$.

10.18 We start with β_k, which appears in (10.71), (10.74) and (10.77). Using these, we can differentiate (10.70) w.r.t. β_k^{-1}, to get

$$\frac{\partial\mathcal{L}}{\partial\beta_k^{-1}} = \frac{D}{2}\left(-N_k - \beta_0 + \beta_k\right).$$

Setting this equal to zero and rearranging the terms, we obtain (10.60). We then consider $\mathbf{m}_k$, which appears in the quadratic terms of (10.71) and (10.74). Thus differentiation of (10.70) w.r.t. $\mathbf{m}_k$ gives

$$\frac{\partial\mathcal{L}}{\partial\mathbf{m}_k} = -N_k\nu_k\left(\mathbf{W}_k\mathbf{m}_k - \mathbf{W}_k\overline{\mathbf{x}}_k\right) - \beta_0\nu_k\left(\mathbf{W}_k\mathbf{m}_k - \mathbf{W}_k\mathbf{m}_0\right).$$

Setting this equal to zero, using (10.60) and rearranging the terms, we obtain (10.61).

Next we tackle $\{\mathbf{W}_k, \nu_k\}$. Here we need to perform a joint optimization w.r.t. $\mathbf{W}_k$ and ν_k for each $k = 1, \ldots, K$. Like β_k, $\mathbf{W}_k$ and ν_k appear in (10.71), (10.74) and (10.77). Using these, we can rewrite the r.h.s. of (10.70) as

$$\begin{aligned}\frac{1}{2}\sum_k^K\Bigg(&N_k\ln\widetilde{\Lambda}_k - N_k\nu_k\left\{\mathrm{Tr}\left(\mathbf{S}_k\mathbf{W}_k\right) + \mathrm{Tr}\left(\mathbf{W}_k\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^{\mathrm{T}}\right)\right\} \\ &+ \ln\widetilde{\Lambda}_k - \beta_0\nu_k\left(\mathbf{m}_k - \mathbf{m}_0\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{m}_k - \mathbf{m}_0\right) + \left(\nu_0 - D - 1\right)\ln\widetilde{\Lambda}_k \\ &- \nu_k\mathrm{Tr}\left(\mathbf{W}_0^{-1}\mathbf{W}_k\right) - \ln\widetilde{\Lambda}_k + 2\mathrm{H}\left[q(\boldsymbol{\Lambda}_k\right]\Bigg) \qquad (269)\end{aligned}$$

where we have dropped terms independent of $\{\mathbf{W}_k, \nu_k\}$, $\ln\widetilde{\Lambda}_k$ is given by (10.65),

$$\mathrm{H}\left[q(\boldsymbol{\Lambda}_k\right] = -\ln B(\mathbf{W}_k, \nu_k) - \frac{\nu_k - D - 1}{2}\ln\widetilde{\Lambda}_k\frac{\nu_k D}{2} \qquad (270)$$

where we have used (10.65) and (B.81), and, from (B.79),

$$\ln B(\mathbf{W}_k, \nu_k) = \frac{\nu_k}{2}\ln|\mathbf{W}_k| - \frac{\nu_k D}{2} - \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k + 1 - i}{2}\right). \qquad (271)$$

Restricting attention to a single component, k, and making use of (270), (269) gives

$$\frac{1}{2}\left(N_k + \nu_0 - \nu_k\right)\ln\widetilde{\Lambda}_k - \frac{\nu_k}{2}\mathrm{Tr}\left(\mathbf{W}_k\mathbf{F}_k\right) - \ln B(\mathbf{W}_k, \nu_k) + \frac{\nu_k D}{2} \qquad (272)$$

where

$$\begin{aligned}\mathbf{F}_k &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + N_k\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^{\mathrm{T}} \\ &\quad + \beta_0\left(\mathbf{m}_k - \mathbf{m}_0\right)\left(\mathbf{m}_k - \mathbf{m}_0\right)^{\mathrm{T}} \\ &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^{\mathrm{T}} \qquad (273)\end{aligned}$$

as we shall show below. Differentiating (272) w.r.t. ν_k, making use of (271) and (10.65), and setting the result to zero, we get

$$\begin{aligned}0 &= \frac{1}{2}\Bigg(\left(N_k+\nu_0-\nu_k\right)\frac{\mathrm{d}\ln\widetilde{\Lambda}_k}{\mathrm{d}\nu_k} - \ln\widetilde{\Lambda}_k - \mathrm{Tr}\left(\mathbf{W}_k\mathbf{F}_k\right) \\ &\quad + \ln|\mathbf{W}_k| + D\ln 2 + \sum_{i=1}^{D}\ln\Gamma\left(\frac{\nu_k+1-i}{2}\right) + D\Bigg) \\ &= \frac{1}{2}\left(\left(N_k+\nu_0-\nu_k\right)\frac{\mathrm{d}\ln\widetilde{\Lambda}_k}{\mathrm{d}\nu_k} - \mathrm{Tr}\left(\mathbf{W}_k\mathbf{F}_k\right) + D\right). \qquad (274)\end{aligned}$$

Similarly, differentiating (272) w.r.t. $\mathbf{W}_k$, making use of (271), (273) and (10.65), and setting the result to zero, we get

$$\begin{aligned}0 &= \frac{1}{2}\left(\left(N_k+\nu_0-\nu_k\right)\mathbf{W}_k^{-1} - \mathbf{F}_k + \mathbf{W}_k^{-1}\right) \\ &= \frac{1}{2}\Bigg(\left(N_k+\nu_0-\nu_k\right)\mathbf{W}_k^{-1} - \mathbf{W}_0^{-1} - N_k\mathbf{S}_k \\ &\quad - \frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^{\mathrm{T}} + \mathbf{W}_k^{-1}\Bigg) \qquad (275)\end{aligned}$$

We see that the simultaneous equations (274) and (275) are satisfied if and only if

$$\begin{aligned}0 &= N_k + \nu_0 - \nu_k \\ 0 &= \mathbf{W}_0^{-1} + N_k\mathbf{S}_k + \frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_0\right)^{\mathrm{T}} - \mathbf{W}_k^{-1}\end{aligned}$$

from which (10.63) and (10.62) follow, respectively. However, we still have to derive (273). From (10.60) and (10.61), derived above, we have

$$\mathbf{m}_k = \frac{\beta_0\mathbf{m}_0 + N_k\overline{\mathbf{x}}_k}{\beta_0 + N_k}$$

and using this, we get

$$
\begin{aligned}
&N_k\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)^{\mathrm{T}}+\beta_0\left(\mathbf{m}_k-\mathbf{m}_0\right)\left(\mathbf{m}_k-\mathbf{m}_0\right)^{\mathrm{T}}= \\
&\qquad N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}}-N_k\overline{\mathbf{x}}_k\frac{\left(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k\right)^{\mathrm{T}}}{\beta_0+N_k}-\frac{\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k}{\beta_0+N_k}N_k\overline{\mathbf{x}}_k^{\mathrm{T}} \\
&+\frac{N_k(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)^{\mathrm{T}}}{(\beta_0+N_k)^2}+\frac{\beta_0(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k)^{\mathrm{T}}}{(\beta_0+N_k)^2} \\
&\qquad -\beta_0\mathbf{m}_0\frac{\left(\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k\right)^{\mathrm{T}}}{\beta_0+N_k}-\frac{\beta_0\mathbf{m}_0+N_k\overline{\mathbf{x}}_k}{\beta_0+N_k}\beta_0\mathbf{m}_0^{\mathrm{T}}+\beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}}.
\end{aligned}
$$

We now gather the coefficients of the terms on the r.h.s. as follows.

Coefficients of $\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}}$:

$$
\begin{aligned}
&N_k-\frac{N_k^2}{\beta_0+N_k}-\frac{N_k^2}{\beta_0+N_k}+\frac{N_k^3}{(\beta_0+N_k)^2}+\frac{\beta_0N_k^2}{(\beta_0+N_k)^2} \\
&\quad=\quad N_k-\frac{N_k^2}{\beta_0+N_k}-\frac{N_k^2}{\beta_0+N_k}+\frac{N_k^3}{(\beta_0+N_k)^2}+\frac{\beta_0N_k^2}{(\beta_0+N_k)^2} \\
&\quad=\quad \frac{N_k\beta_0^2+N_k^2\beta_0+N_k^2\beta_0+N_k^3-2(N_k^2\beta_0+N_k^3)+N_k^3+\beta_0N_k^2}{(\beta_0+N_k)^2} \\
&\quad=\quad \frac{N_k\beta_0^2+\beta_0N_k^2}{(\beta_0+N_k)^2}=\frac{N_k\beta_0(\beta_0+N_k)}{(\beta_0+N_k)^2}=\frac{N_k\beta_0}{\beta_0+N_k}
\end{aligned}
$$

Coefficients of $\overline{\mathbf{x}}_k\mathbf{m}_0^{\mathrm{T}}$ and $\mathbf{m}_0\overline{\mathbf{x}}_k^{\mathrm{T}}$ (these are identical):

$$
\begin{aligned}
&-\frac{N_k\beta_0}{\beta_0+N_k}+\frac{\beta_0N_k^2}{(\beta_0+N_k)^2}+\frac{\beta_0^2N_k}{(\beta_0+N_k)^2}-\frac{N_k\beta_0}{\beta_0+N_k} \\
&\qquad\qquad=\frac{N_k\beta_0}{\beta_0+N_k}-\frac{2N_k\beta_0}{\beta_0+N_k}=-\frac{N_k\beta_0}{\beta_0+N_k}
\end{aligned}
$$

Coefficients of $\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}}$:

$$
\begin{aligned}
&\frac{N_k\beta_0^2}{(\beta_0+N_k)^2}+\frac{\beta_0^3}{(\beta_0+N_k)^2}-\frac{2\beta_0^2}{\beta_0+N_k}+\beta_0 \\
&\quad=\quad \frac{\beta_0^2(N_k+\beta_0)}{(\beta_0+N_k)^2}-\frac{2\beta_0^2}{\beta_0+N_k}+\beta_0 \\
&\quad=\quad \frac{\beta_0^2-2\beta_0^2+\beta_0^2+N_k\beta_0}{\beta_0+N_k}=\frac{N_k\beta_0}{\beta_0+N_k}
\end{aligned}
$$

Thus

$$
\begin{aligned}
&N_k\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_k\right)^{\mathrm{T}}+\beta_0\left(\mathbf{m}_k-\mathbf{m}_0\right)\left(\mathbf{m}_k-\mathbf{m}_0\right)^{\mathrm{T}} \\
&\qquad\qquad=\frac{N_k\beta_0}{N_k+\beta_0}\left(\overline{\mathbf{x}}_k-\mathbf{m}_0\right)\left(\overline{\mathbf{x}}_k-\mathbf{m}_0\right)^{\mathrm{T}} \qquad (276)
\end{aligned}
$$

as desired.

Now we turn our attention to $\boldsymbol{\alpha}$, which appear in (10.72) and (10.73), through (10.66), and (10.76). Using these together with (B.23) and (B.24), we can differentiate (10.70) w.r.t. α_k and set it equal to zero, yielding

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha_k} &= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\frac{\partial \ln \widehat{\pi}_k}{\partial \alpha_k} - \ln \widehat{\pi}_k - \frac{\partial \ln C(\boldsymbol{\alpha})}{\partial \alpha_k} \\ &= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\left\{\psi_1(\alpha_k) - \psi_1(\widehat{\alpha})\frac{\partial \widehat{\alpha}}{\partial \alpha_k}\right\} \\ &\quad + \psi(\widehat{\alpha}) - \psi(\alpha_k) - \psi(\widehat{\alpha})\frac{\partial \widehat{\alpha}}{\partial \alpha_k} + \psi(\alpha_k) \\ &= [N_k + (\alpha_0 - 1) - (\alpha_k - 1)]\left\{\psi_1(\alpha_k) - \psi_1(\widehat{\alpha})\right\} = 0 \end{aligned} \tag{277}$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ are di- and trigamma functions, respectively. If we assume that $\alpha_0 > 0$, (10.58) must hold for (277) to hold, since the trigamma function is strictly positive and monotoncally decreasing for arguments greater than zero.

Finally, we maximize (10.70) w.r.t. r_{nk}, subject to the constraints $\sum_k r_{nk} = 1$ for all $n = 1, \ldots, N$. Note that r_{nk} not only appears in (10.72) and (10.75), but also in (10.71) through N_k, $\overline{\mathbf{x}}_k$ and $\mathbf{S}_k$, and so we must substitute using (10.51), (10.52) and (10.53), respectively. To simplify subsequent calculations, we start by considering the last two terms inside the braces of (10.71), which we write together as

$$\frac{1}{2}\sum_{k=1}^{K} \nu_k \mathrm{Tr}\left(\mathbf{W}_k \mathbf{Q}_k\right) \tag{278}$$

where, using (10.51), (10.52) and (10.53),

$$\begin{aligned}\mathbf{Q}_k &= \sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \overline{\mathbf{x}}_k\right)\left(\mathbf{x}_n - \overline{\mathbf{x}}_k\right)^{\mathrm{T}} + N_k\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)\left(\overline{\mathbf{x}}_k - \mathbf{m}_k\right)^{\mathrm{T}} \\ &= \sum_{n=1}^{N} r_{nk}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - 2N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} \\ &\quad + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\mathbf{m}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} + N_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\ &= \sum_{n=1}^{N} r_{nk}\left(\mathbf{x}_n - \mathbf{m}_k\right)\left(\mathbf{x}_n - \mathbf{m}_k\right)^{\mathrm{T}}. \end{aligned} \tag{279}$$

Using (10.51), (278) and (279), we can now consider all terms in (10.71) which

depend on r_{kn} and add the appropriate Lagrange multiplier terms, yielding

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\left(\ln\widetilde{\Lambda}_k - D\beta_k^{-1}\right) - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right) + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln\widetilde{\pi}_k - \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}\ln r_{nk} + \sum_{n=1}^{N}\lambda_n\left(1-\sum_{k=1}^{K} r_{nk}\right).$$

Taking the derivative of this w.r.t. r_{kn} and setting it equal to zero we obtain

$$0 = \frac{1}{2}\ln\widetilde{\Lambda}_k - \frac{D}{2\beta_k} - \frac{1}{2}\nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right) + \ln\widetilde{\pi}_k - \ln r_{nk} - 1 - \lambda_n.$$

Moving $\ln r_{nk}$ to the l.h.s. and exponentiating both sides, we see that for each n,

$$r_{nk} \propto \widetilde{\pi}_k\widetilde{\Lambda}_k^{1/2}\exp\left\{-\frac{D}{2\beta_k} - \frac{1}{2}\nu_k\left(\mathbf{x}_n-\mathbf{m}_k\right)^{\mathrm{T}}\mathbf{W}_k\left(\mathbf{x}_n-\mathbf{m}_k\right)\right\}$$

which is in agreement with (10.67); the normalized form is then given by (10.49).

10.19 We start by performing the integration over $\boldsymbol{\pi}$ in (10.80), making use of the result

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\widehat{\alpha}}$$

to give

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K}\frac{\alpha_k}{\widehat{\alpha}}\iint \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k)\,\mathrm{d}\boldsymbol{\mu}_k\,\mathrm{d}\boldsymbol{\Lambda}_k.$$

The variational posterior distribution over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ is given from (10.59) by

$$q(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k,\nu_k).$$

Using this result we next perform the integration over $\boldsymbol{\mu}_k$. This can be done explicitly by completing the square in the exponential in the usual way, or we can simply appeal to the general result (2.109) and (2.110) for the linear-Gaussian model from Chapter 2 to give

$$\int \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1})\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_k,(\beta_k\boldsymbol{\Lambda}_k)^{-1}\right)\,\mathrm{d}\boldsymbol{\mu}_k = \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k,\left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right).$$

Thus we have

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K}\frac{\alpha_k}{\widehat{\alpha}}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k,\left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k,\nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k.$$

The final integration over $\boldsymbol{\Lambda}_k$ is the convolution of a Wishart with a Gaussian. Omitting multiplicative constants which are independent of $\widehat{\mathbf{x}}$ we have

$$\begin{aligned}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k, \left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\, \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k \\ \propto \int |\boldsymbol{\Lambda}_k|^{1/2+(\nu_k-D-1)/2} \exp\left\{-\frac{1}{2\left(1+\beta_k^{-1}\right)}\mathrm{Tr}\left[\boldsymbol{\Lambda}_k(\widehat{\mathbf{x}}-\mathbf{m}_k)(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\right]\right. \\ \left. -\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Lambda}_k\mathbf{W}_k^{-1}\right]\right\}\mathrm{d}\boldsymbol{\Lambda}_k.\end{aligned}$$

We can now perform this integration by observing that the argument of the integral is an un-normalized Wishart distribution (unsurprisingly since the Wishart is the conjugate prior for the precision of a Gaussian) and so we can write down the result of this integration, up to an overall constant, by using the known normalization coefficient for the Wishart, given by (B.79). Thus we have

$$\begin{aligned}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k, \left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\, \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k \\ \propto \left|\mathbf{W}_k^{-1} + \frac{1}{\left(1+\beta_k^{-1}\right)}(\widehat{\mathbf{x}}-\mathbf{m}_k)(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\right|^{-(\nu_k+1)/2} \\ \propto \left|\mathbf{I} + \frac{1}{\left(1+\beta_k^{-1}\right)}\mathbf{W}_k(\widehat{\mathbf{x}}-\mathbf{m}_k)(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\right|^{-(\nu_k+1)/2}\end{aligned}$$

where we have omitted factors independent of $\widehat{\mathbf{x}}$ since we are only interested in the functional dependence on $\widehat{\mathbf{x}}$, and we have made use of the result $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ and omitted an overall factor of $|\mathbf{W}_k^{-1}|$. Next we use the identity

$$\left|\mathbf{I} + \mathbf{a}\mathbf{b}^{\mathrm{T}}\right| = (1 + \mathbf{a}^{\mathrm{T}}\mathbf{b})$$

where $\mathbf{a}$ and $\mathbf{b}$ are D-dimensional vectors and $\mathbf{I}$ is the $D \times D$ unit matrix, to give

$$\begin{aligned}\int \mathcal{N}\left(\widehat{\mathbf{x}}|\mathbf{m}_k, \left(1+\beta_k^{-1}\right)\boldsymbol{\Lambda}_k^{-1}\right)\, \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k)\,\mathrm{d}\boldsymbol{\Lambda}_k \\ \propto \left\{1 + \frac{1}{\left(1+\beta_k^{-1}\right)}(\widehat{\mathbf{x}}-\mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\widehat{\mathbf{x}}-\mathbf{m}_k)\right\}^{-(\nu_k+1)/2}.\end{aligned}$$

We recognize this result as being a Student distribution, and by comparison with the standard form (B.68) for the Student we see that it has mean $\mathbf{m}_k$, precision given by (10.82) and degrees of freedom parameter $\nu_k + 1 - D$. We can re-instate the normalization coefficient using the standard form for the Student distribution given in Appendix B.

10.20 Consider first the posterior distribution over the precision of component k given by

$$q^\star(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_k, \nu_k).$$

From (10.63) we see that for large N we have $\nu_k \to N_k$, and similarly from (10.62) we see that $\mathbf{W}_k \to N_k^{-1}\mathbf{S}_k^{-1}$. Thus the mean of the distribution over $\boldsymbol{\Lambda}_k$, given by $\mathrm{E}[\boldsymbol{\Lambda}_k] = \nu_k\mathbf{W}_k \to \mathbf{S}_k^{-1}$ which is the maximum likelihood value (this assumes that the quantities r_{nk} reduce to the corresponding EM values, which is indeed the case as we shall show shortly). In order to show that this posterior is also sharply peaked, we consider the differential entropy, $\mathrm{H}[\boldsymbol{\Lambda}_k]$ given by (B.82), and show that, as $N_k \to \infty$, $\mathrm{H}[\boldsymbol{\Lambda}_k] \to 0$, corresponding to the density collapsing to a spike. First consider the normalizing constant $B(\mathbf{W}_k, \nu_k)$ given by (B.79). Since $\mathbf{W}_k \to N_k^{-1}\mathbf{S}_k^{-1}$ and $\nu_k \to N_k$,

$$-\ln B(\mathbf{W}_k, \nu_k) \to -\frac{N_k}{2}\left(D\ln N_k + \ln|\mathbf{S}_k| - D\ln 2\right) + \sum_{i=1}^{D}\ln\Gamma\left(\frac{N_k+1-i}{2}\right).$$

We then make use of Stirling's approximation (1.146) to obtain

$$\ln\Gamma\left(\frac{N_k+1-i}{2}\right) \simeq \frac{N_k}{2}\left(\ln N_k - \ln 2 - 1\right)$$

which leads to the approximate limit

$$\begin{aligned} -\ln B(\mathbf{W}_k, \nu_k) &\to -\frac{N_k D}{2}\left(\ln N_k - \ln 2 - \ln N_k + \ln 2 + 1\right) - \frac{N_k}{2}\ln|\mathbf{S}_k| \\ &= -\frac{N_k}{2}\left(\ln|\mathbf{S}_k| + D\right). \end{aligned} \tag{280}$$

Next, we use (10.241) and (B.81) in combination with $\mathbf{W}_k \to N_k^{-1}\mathbf{S}_k^{-1}$ and $\nu_k \to N_k$ to obtain the limit

$$\begin{aligned} \mathbb{E}\left[\ln|\boldsymbol{\Lambda}|\right] &\to D\ln\frac{N_k}{2} + D\ln 2 - D\ln N_k - \ln|\mathbf{S}_k| \\ &= -\ln|\mathbf{S}_k|, \end{aligned}$$

where we approximated the argument to the digamma function by $N_k/2$. Substituting this and (280) into (B.82), we get

$$\mathrm{H}[\boldsymbol{\Lambda}] \to 0$$

when $N_k \to \infty$.

Next consider the posterior distribution over the mean $\boldsymbol{\mu}_k$ of the k^{th} component given by

$$q^\star(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k\boldsymbol{\Lambda}_k).$$

From (10.61) we see that for large N the mean $\mathbf{m}_k$ of this distribution reduces to $\overline{\mathbf{x}}_k$ which is the corresponding maximum likelihood value. From (10.60) we see that

$\beta_k \to N_k$ and Thus the precision $\beta_k \mathbf{\Lambda}_k \to \beta_k \nu_k \mathbf{W}_k \to N_k \mathbf{S}_k^{-1}$ which is large for large N and hence this distribution is sharply peaked around its mean.

Now consider the posterior distribution $q^\star(\boldsymbol{\pi})$ given by (10.57). For large N we have $\alpha_k \to N_k$ and so from (B.17) and (B.19) we see that the posterior distribution becomes sharply peaked around its mean $\mathrm{E}[\pi_k] = \alpha_k/\overline{\alpha} \to N_k/N$ which is the maximum likelihood solution.

For the distribution $q^\star(\mathbf{z})$ we consider the responsibilities given by (10.67). Using (10.65) and (10.66), together with the asymptotic result for the digamma function, we again obtain the maximum likelihood expression for the responsibilities for large N.

Finally, for the predictive distribution we first perform the integration over $\boldsymbol{\pi}$, as in the solution to Exercise 10.19, to give

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K} \frac{\alpha_k}{\overline{\alpha}} \iint \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) q(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k)\, \mathrm{d}\boldsymbol{\mu}_k\, \mathrm{d}\mathbf{\Lambda}_k.$$

The integrations over $\boldsymbol{\mu}_k$ and $\mathbf{\Lambda}_k$ are then trivial for large N since these are sharply peaked and hence approximate delta functions. We therefore obtain

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K} \frac{N_k}{N} \mathcal{N}(\widehat{\mathbf{x}}|\overline{\mathbf{x}}_k, \mathbf{W}_k)$$

which is a mixture of Gaussians, with mixing coefficients given by N_k/N.

10.21 The number of equivalent parameter settings equals the number of possible assignments of K parameter sets to K mixture components: K for the first component, times $K-1$ for the second component, times $K-2$ for the third and so on, giving the result $K!$.

10.22 The mixture distribution over the parameter space takes the form

$$q(\mathbf{\Theta}) = \frac{1}{K!} \sum_{\kappa=1}^{K!} q_\kappa(\boldsymbol{\theta}_\kappa)$$

where $\boldsymbol{\theta}_\kappa = \{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \boldsymbol{\pi}\}$, κ indexes the components of this mixture and $\mathbf{\Theta} = \{\boldsymbol{\theta}_\kappa\}$. With this model, (10.3) becomes

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{\Theta}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{\Theta})}{q(\mathbf{\Theta})} \right\} \mathrm{d}\mathbf{\Theta} \\
&= \frac{1}{K!} \sum_{\kappa=1}^{K!} \int q_\kappa(\boldsymbol{\theta}_\kappa) \ln p(\mathbf{X}, \boldsymbol{\theta}_\kappa)\, \mathrm{d}\boldsymbol{\theta}_\kappa \\
&\qquad - \frac{1}{K!} \sum_{\kappa=1}^{K!} \int q_\kappa(\boldsymbol{\theta}_\kappa) \ln \left(\frac{1}{K!} \sum_{\kappa'=1}^{K!} q_{\kappa'}(\boldsymbol{\theta}_{\kappa'}) \right) \mathrm{d}\boldsymbol{\theta}_\kappa \\
&= \int q(\boldsymbol{\theta}) \ln p(\mathbf{X}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \ln q(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} + \ln K!
\end{aligned}$$

where $q(\boldsymbol{\theta})$ corresponds to any one of the $K!$ equivalent $q_\kappa(\boldsymbol{\theta}_\kappa)$ distributions. Note that in the last step, we use the assumption that the overlap between these distributions is negligible and hence

$$\int q_\kappa(\boldsymbol{\theta}) \ln q_{\kappa'}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \simeq 0$$

when $\kappa \neq \kappa'$.

10.23 When we are treating $\boldsymbol{\pi}$ as a parameter, there is neither a prior, nor a variational posterior distribution, over $\boldsymbol{\pi}$. Therefore, the only term remaining from the lower bound, (10.70), that involves $\boldsymbol{\pi}$ is the second term, (10.72). Note however, that (10.72) involves the *expectations* of $\ln \pi_k$ under $q(\boldsymbol{\pi})$, whereas here, we operate directly with π_k, yielding

$$\mathbb{E}_{q(\mathbf{Z})}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \ln \pi_k.$$

Adding a Langrange term, as in (9.20), taking the derivative w.r.t. π_k and setting the result to zero we get

$$\frac{N_k}{\pi_k} + \lambda = 0, \tag{281}$$

where we have used (10.51). By re-arranging this to

$$N_k = -\lambda \pi_k$$

and summing both sides over k, we see that $-\lambda = \sum_k N_k = N$, which we can use to eliminate λ from (281) to get (10.83).

10.24 The singularities that may arise in maximum likelihood estimation are caused by a mixture component, k, collapsing on a data point, $\mathbf{x}_n$, i.e., $r_{kn} = 1$, $\boldsymbol{\mu}_k = \mathbf{x}_n$ and $|\boldsymbol{\Lambda}_k| \to \infty$.

However, the prior distribution $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ defined in (10.40) will prevent this from happening, also in the case of MAP estimation. Consider the product of the expected complete log-likelihood and $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ as a function of $\boldsymbol{\Lambda}_k$:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{Z})}\left[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] \\
&= \frac{1}{2} \sum_{n=1}^{N} r_{kn} \left(\ln |\boldsymbol{\Lambda}_k| - (\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \right) \\
&\quad + \ln |\boldsymbol{\Lambda}_k| - \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\boldsymbol{\mu}_k - \mathbf{m}_0) \\
&\quad + (\nu_0 - D - 1) \ln |\boldsymbol{\Lambda}_k| - \mathrm{Tr}\left[\mathbf{W}_0^{-1} \boldsymbol{\Lambda}_k\right] + \text{const.}
\end{aligned}$$

where we have used (10.38), (10.40) and (10.50), together with the definitions for the Gaussian and Wishart distributions; the last term summarizes terms independent of $\boldsymbol{\Lambda}_k$. Using (10.51)–(10.53), we can rewrite this as

$$(\nu_0 + N_k - D) \ln |\boldsymbol{\Lambda}_k| - \mathrm{Tr}\left[(\mathbf{W}_0^{-1} + \beta_0 (\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k \mathbf{S}_k) \boldsymbol{\Lambda}_k\right],$$

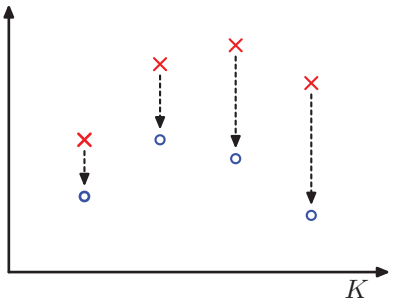


Figure 9 Illustration of the true log marginal likelihood for a Gaussian mixture model (×) and the corresponding variational bound obtained from a factorized approximation (○) as functions of the number of mixture components, K. The dashed arrows emphasize the typical increase in the difference between the true log marginal likelihood and the bound. As a consequence, the bound tends to have its peak at a lower value of K than the true log marginal likelihood.

where we have dropped the constant term. Using (C.24) and (C.28), we can compute the derivative of this w.r.t. $\mathbf{\Lambda}_k$ and setting the result equal to zero, we find the MAP estimate for $\mathbf{\Lambda}_k$ to be

$$\mathbf{\Lambda}_k^{-1} = \frac{1}{\nu_0 + N_k - D}(\mathbf{W}_0^{-1} + \beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k\mathbf{S}_k).$$

From this we see that $|\mathbf{\Lambda}_k^{-1}|$ can never become 0, because of the presence of $\mathbf{W}_0^{-1}$ (which we must chose to be positive definite) in the expression on the r.h.s.

10.25 As the number of mixture components grows, so does the number of variables that may be correlated, but which are treated as independent under a variational approximation, as illustrated in Figure 10.2. As a result of this, the proportion of probability mass under the true distribution, $p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}|\mathbf{X})$, that the variational approximation $q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma})$ does not capture, will grow. The consequence will be that the second term in (10.2), the KL divergence between $q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma})$ and $p(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}|\mathbf{X})$, will increase. Since this KL divergence is the difference between the true log marginal and the corresponding the lower bound, the latter must decrease compared to the former. Thus, as illustrated in Figure 9, chosing the number of components based on the lower bound will tend to underestimate the optimal number of components.

10.26 Extending the variational treatment of Section 10.3 to also include β, we specify the prior for β

$$p(\beta) = \mathrm{Gam}\,(\beta|c_0, d_0) \tag{282}$$

and modify (10.90) as

$$p(\mathbf{t}, \mathbf{w}, \alpha, \beta) = p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)p(\alpha)p(\beta) \tag{283}$$

where the first factor on the r.h.s. correspond to (10.87) with the dependence on β made explicit.

The formulae for $q^\star(\alpha)$, (10.93)–(10.95), remain unchanged. For $q(\mathbf{w})$, we follow the path mapped out in Section 10.3, incorporating the modifications required by the

changed treatment of β; (10.96)–(10.98) now become

$$\begin{aligned}\ln q^\star(\mathbf{w}) &= \mathbb{E}_\beta\left[\ln p(\mathbf{t}|\mathbf{w},\beta)\right] + \mathbb{E}_\alpha\left[\ln p(\mathbf{w}|\alpha)\right] + \text{const} \\ &= -\frac{\mathbb{E}[\beta]}{2}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right\}^2 - \frac{\mathbb{E}[\alpha]}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbb{E}[\alpha]\mathbf{I} + \mathbb{E}[\beta]\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)\mathbf{w} + \mathbb{E}[\beta]\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} + \text{const}.\end{aligned}$$

Accordingly, (10.100) and (10.101) become

$$\begin{aligned}\mathbf{m}_N &= \mathbb{E}[\beta]\mathbf{S}_N\mathbf{\Phi}\mathbf{t} \\ \mathbf{S}_N &= \left(\mathbb{E}[\alpha]\mathbf{I} + \mathbb{E}[\beta]\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}.\end{aligned}$$

For $q(\beta)$, we use (10.9), (282) and (283) to obtain

$$\begin{aligned}\ln q^\star(\beta) &= \mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{t}|\mathbf{w},\beta)\right] + \ln p(\beta) + \text{const} \\ &= \frac{N}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}_{\mathbf{w}}\left[\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right\}^2\right] + (c_0 - 1)\ln\beta - d_0\beta\end{aligned}$$

which we recognize as the logarithm of a Gamma distribution with parameters

$$\begin{aligned}c_N &= c_0 + \frac{N}{2} \\ d_N &= d_0 + \frac{1}{2}\mathbb{E}\left[\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\phi_n - t_n\right)^2\right] \\ &= d_0 + \frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\right) + \mathbf{t}^{\mathrm{T}}\mathbf{t}\right) - \mathbf{t}^{\mathrm{T}}\mathbf{\Phi}\mathbb{E}[\mathbf{w}] \\ &= d_0 + \frac{1}{2}\left(\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \mathrm{Tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{S}_N\right)\right)\end{aligned}$$

where we have used (10.103) and, from (B.38),

$$\mathbb{E}[\mathbf{w}] = \mathbf{m}_N. \tag{284}$$

Thus, from (B.27),

$$\mathbb{E}[\beta] = \frac{c_N}{d_N}. \tag{285}$$

In the lower bound, (10.107), the first term will be modified and two new terms added on the r.h.s.We start with the modified log likelihood term:

$$\begin{aligned}\mathbb{E}_\beta\left[\mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{t}|\mathbf{w},\beta)\right]\right] &= \frac{N}{2}\left(\mathbb{E}[\beta] - \ln(2\pi)\right) - \frac{\mathbb{E}[\beta]}{2}\mathbb{E}\left[\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2\right] \\ &= \frac{N}{2}\left(\psi(c_N) - \ln d_N - \ln(2\pi)\right) \\ &\quad -\frac{c_N}{2d_N}\left(|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \mathrm{Tr}\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{S}_N\right)\right)\end{aligned}$$

where we have used (284), (285), (10.103) and (B.30). Next we consider the term corresponding log prior over β:

$$\begin{aligned}\mathbb{E}\left[\ln p(\beta)\right] &= (c_0-1)\mathbb{E}[\ln\beta] - d_0\mathbb{E}[\beta] + c_0\ln d_0 - \ln\Gamma(c_0) \\ &= (c_0-1)(\psi(c_N)-\ln d_N) - \frac{d_0 c_N}{d_N} + c_0\ln d_0 - \ln\Gamma(c_0)\end{aligned}$$

where we have used (285) and (B.30). Finally, from (B.31), we get the last term in the form of the negative entropy of the posterior over β:

$$-\mathbb{E}\left[\ln q^\star(\beta)\right] = (c_N-1)\psi(c_N) + \ln d_N - c_N - \ln\Gamma(c_N).$$

Finally, the predictive distribution is given by (10.105) and (10.106), with $1/\beta$ replaced by $1/\mathbb{E}[\beta]$.

10.27 Consider each of the five terms in the lower bound (10.107) in turn. For the terms arising from the likelihood function we have

$$\begin{aligned}\mathbb{E}[\ln p(\mathbf{t}|\mathbf{w})] &= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}\left[\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right] \\ &= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta \\ &\quad -\frac{\beta}{2}\left\{\mathbf{t}^{\mathrm{T}}\mathbf{t} - 2\mathbb{E}[\mathbf{w}^{\mathrm{T}}]\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \mathrm{Tr}\left(\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\right\}.\end{aligned}$$

The prior over $\mathbf{w}$ gives rise to

$$\mathbb{E}[\ln p(\mathbf{w}|\alpha)] = -\frac{M}{2}\ln(2\pi) + \frac{M}{2}\mathbb{E}[\ln\alpha] - \frac{\mathbb{E}[\alpha]}{2}\mathbb{E}[\mathbf{w}^{\mathrm{T}}\mathbf{w}].$$

Similarly, the prior over α gives

$$\mathbb{E}[\ln p(\alpha)] = a_0\ln b_0 + (a_0-1)\mathbb{E}[\ln\alpha] - b_0\mathbb{E}[\alpha] - \ln\Gamma(a_0).$$

The final two terms in $\mathcal{L}$ represent the negative entropies of the Gaussian and gamma distributions, and are given by (B.41) and (B.31) respectively, so that

$$-\mathbb{E}[\ln q(\mathbf{w})] = \frac{1}{2}\ln|\mathbf{S}_N| + \frac{M}{2}(1+\ln(2\pi)).$$

Similarly we have

$$-\mathbb{E}[\ln q(\alpha)] = -(a_n-1)\psi(a_N) + a_N + \ln\Gamma(a_N) + \ln b_N.$$

Now we substitute the following expressions for the moments

$$\begin{aligned}\mathbb{E}[\mathbf{w}] &= \mathbf{m}_N \\ \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] &= \mathbf{m}_N\mathbf{m}_N^{\mathrm{T}} + \mathbf{S}_N \\ \mathbb{E}[\alpha] &= \frac{a_N}{b_N} \\ \mathbb{E}[\ln\alpha] &= \psi(a_N) - \ln b_N.\end{aligned}$$

and combine the various terms together to obtain (10.107).

10.28 **NOTE**: In PRML, Equations (10.119)–(10.121) contain errors; please consult the PRML Errata for relevant corrections.

We start by writing the complete-data likelihood, given by (10.37) and (10.38) in a form corresponding to (10.113). From (10.37) and (10.38), we have

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\mathbf{Z}|\boldsymbol{\pi}) \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right)^{z_{nk}} \end{aligned}$$

which is a product over data points, just like (10.113). Focussing on the individual factors of this product, we have

$$p(\mathbf{x}_n, \mathbf{z}_n|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right)^{z_{nk}} = \exp\left\{\sum_{k=1}^{K} z_{nk} \left(\ln \pi_k + \frac{1}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{D}{2} \ln(2\pi) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right)\right\}.$$

Drawing on results from Solution 2.57, we can rewrite this in the form of (10.113), with

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k \\ \overrightarrow{\boldsymbol{\Lambda}_k} \\ \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k \\ \ln |\boldsymbol{\Lambda}_k| \\ \ln \pi_k \end{bmatrix}_{k=1,\ldots,K} \tag{286}$$

$$\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) = \left[z_{nk} \begin{bmatrix} \mathbf{x}_n \\ \overrightarrow{\frac{1}{2}\mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{bmatrix} \right]_{k=1,\ldots,K} \tag{287}$$

$$h(\mathbf{x}_n, \mathbf{z}_n) = \prod_{k=1}^{K} \left((2\pi)^{-D/2}\right)^{z_{nk}} \tag{288}$$

$$g(\boldsymbol{\eta}) = 1$$

where we have introduce the notation

$$[\mathbf{v}_k]_{k=1,\ldots,K} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_K \end{bmatrix}$$

and the operator $\overrightarrow{\mathbf{M}}$ which returns a vector formed by stacking the columns of the argument matrix on top of each other.

Next we seek to rewrite the prior over the parameters, given by (10.39) and (10.40), in a form corresponding to (10.114) and which also matches (286). From, (10.39), (10.40) and Appendix B, we have

$$\begin{aligned} p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \mathrm{Dir}\,(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}\left(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0\right) \\ &= C(\boldsymbol{\alpha}_0) \left(\frac{\beta_0}{2\pi}\right)^{KD/2} B(\mathbf{W}_0, \nu_0)^K \exp\left\{ \sum_{k=1}^{K} (\alpha_0 - 1) \ln \pi_k \right. \\ &\left. + \frac{\nu_0 - D}{2} \ln |\boldsymbol{\Lambda}_k| - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Lambda}_k \left[\beta_0(\boldsymbol{\mu}_k - \mathbf{m}_0)(\boldsymbol{\mu}_k - \mathbf{m}_0)^{\mathrm{T}} + \mathbf{W}_0\right]\right) \right\}. \end{aligned}$$

we can rewrite this in the form of (10.114) with $\boldsymbol{\eta}$ given by (286),

$$\boldsymbol{\chi}_0 = \begin{bmatrix} \beta_0 \mathbf{m}_0 \\ -\frac{1}{2}\left(\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}}\right) \\ -\beta_0/2 \\ (\nu_0 - D)/2 \\ \alpha_0 - 1 \end{bmatrix}_{k=1,\ldots,K} \tag{289}$$

$$g(\boldsymbol{\eta}) = 1$$

$$f(\upsilon_0, \boldsymbol{\chi}_0) = C(\boldsymbol{\alpha}_0) \left(\frac{\beta_0}{2\pi}\right)^{KD/2} B(\mathbf{W}_0, \nu_0)^K$$

and υ_0 replaces ν_0 in (10.114) to avoid confusion with ν_0 in (10.40).

Having rewritten the Bayesian mixture of Gaussians as a conjugate model from the exponential family, we now proceed to rederive (10.48), (10.57) and (10.59). By exponentiating both sides of (10.115) and making use of (286)–(288), we obtain (10.47), with ρ_{nk} given by (10.46), from which (10.48) follows.

Next we can use (10.50) to take the expectation w.r.t. $\mathbf{Z}$ in (10.121), substituting r_{nk} for $\mathbb{E}[z_{nk}]$ in (287). Combining this with (289), (10.120) and (10.121) become

$$\upsilon_N = \upsilon_0 + N = 1 + N$$

and

$$\upsilon_N \boldsymbol{\chi}_N = \left[\begin{array}{c} \beta_0 \mathbf{m}_0 \\ -\frac{1}{2}\left(\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}}\right) \\ -\beta_0/2 \\ (\nu_0 - D)/2 \\ \alpha_0 - 1 \end{array}\right]_{k=1,\ldots,K} + \sum_{n=1}^{N} \left[r_{nk} \left[\begin{array}{c} \mathbf{x}_n \\ \frac{1}{2}\overrightarrow{\mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}} \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{array}\right]\right]_{k=1,\ldots,K}$$

$$= \left[\begin{array}{c} \beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k \\ -\frac{1}{2}\left(\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}} + N_k \overrightarrow{\left(\mathbf{S}_k + \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}}\right)}\right) \\ -(\beta_0 + N_k)/2 \\ (\nu_0 - D + N_k)/2 \\ \alpha_0 - 1 + N_k \end{array}\right]_{k=1,\ldots,K} \tag{290}$$

where we use υ_N instead of ν_N in (10.119)–(10.121), to avoid confusion with ν_k, which appears in (10.59) and (10.63). From the bottom row of (287) and (290), we see that the inner product of $\boldsymbol{\eta}$ and $\upsilon_N \boldsymbol{\chi}_N$ gives us the r.h.s. of (10.56), from which (10.57) follows. The remaining terms of this inner product are

$$\sum_{k=1}^{K} \Big\{ \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \left(\beta_0 \mathbf{m}_0 + N_k \overline{\mathbf{x}}_k\right) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k \left[\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}} + N_k \overrightarrow{\left(\mathbf{S}_k + \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}}\right)}\right]\right) - \frac{1}{2}(\beta_0 + N_k) \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \frac{1}{2}(\nu_0 + N_k - D) \ln |\boldsymbol{\Lambda}| \Big\}.$$

Restricting our attention to parameters corresponding to a single mixture component and making use of (10.60), (10.61) and (10.63), we can rewrite this as

$$-\frac{1}{2}\beta_k \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k + \beta_k \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \mathbf{m}_k - \frac{1}{2}\beta_k \mathbf{m}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \mathbf{m}_k + \frac{1}{2}\ln |\boldsymbol{\Lambda}| + \frac{1}{2}\beta_k \mathbf{m}_k^{\mathrm{T}} \boldsymbol{\Lambda}_k \mathbf{m}_k - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}_k \left[\beta_0 \overrightarrow{\mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}}} + \overrightarrow{\mathbf{W}_0^{-1}} + N_k \overrightarrow{\left(\mathbf{S}_k + \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}}\right)}\right]\right) + \frac{1}{2}(\nu_k - D - 1) \ln |\boldsymbol{\Lambda}|.$$

The first four terms match the logarithm of $\mathcal{N}\left(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}\right)$ from the r.h.s. of (10.59); the missing $D/2[\ln \beta_k - \ln(2\pi)]$ can be accounted for in $f(\upsilon_N, \boldsymbol{\chi}_N)$. To make the remaining terms match the logarithm of $\mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0)$ from the r.h.s. of (10.59), we need to show that

$$\beta_0 \mathbf{m}_0 \mathbf{m}_0^{\mathrm{T}} + N_k \overline{\mathbf{x}}_k \overline{\mathbf{x}}_k^{\mathrm{T}} - \beta_k \mathbf{m}_k \mathbf{m}_k^{\mathrm{T}}$$

equals the last term on the r.h.s. of (10.62). Using (10.60), (10.61) and (276), we get

$$\begin{aligned}
&\beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - \beta_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - \beta_0\mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} - \beta_0\mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} + \beta_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} - \beta_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0\mathbf{m}_0\mathbf{m}_0^{\mathrm{T}} - \beta_0\mathbf{m}_0\mathbf{m}_k^{\mathrm{T}} - \beta_0\mathbf{m}_k\mathbf{m}_0^{\mathrm{T}} + \beta_0\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&\quad + N_k\overline{\mathbf{x}}_k\overline{\mathbf{x}}_k^{\mathrm{T}} - N_k\overline{\mathbf{x}}_k\mathbf{m}_k^{\mathrm{T}} - N_k\mathbf{m}_k\overline{\mathbf{x}}_k^{\mathrm{T}} + N_k\mathbf{m}_k\mathbf{m}_k^{\mathrm{T}} \\
&= \beta_0(\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^{\mathrm{T}} + N_k(\overline{\mathbf{x}}_k - \mathbf{m}_k)(\overline{\mathbf{x}}_k - \mathbf{m}_k)^{\mathrm{T}} \\
&= \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{\mathbf{x}}_k - \mathbf{m}_0)(\overline{\mathbf{x}}_k - \mathbf{m}_0)^{\mathrm{T}}.
\end{aligned}$$

Thus we have recovered $\ln \mathcal{W}(\mathbf{\Lambda}_k|\mathbf{W}_0, \nu_0)$ (missing terms are again accounted for by $f(\upsilon_N, \boldsymbol{\chi}_N)$) and thereby (10.59).

10.29 **NOTE**: In PRML, the use of λ to denote the varitional parameter leads to inconsistencies w.r.t. exisiting literature. To remedy this λ should be replaced by η from the beginning of Section 10.5 up to and including the last line before equation (10.141). For further details, please consult the PRML Errata.

Standard rules of differentiation give

$$\frac{d\ln(x)}{dx} = \frac{1}{x}$$

$$\frac{d^2\ln(x)}{dx^2} = -\frac{1}{x^2}.$$

Since its second derivative is negative for all value of x, $\ln(x)$ is concave for $0 < x < \infty$.

From (10.133) we have

$$\begin{aligned}
g(\eta) &= \min_x \{\eta x - f(x)\} \\
&= \min_x \{\eta x - \ln(x)\}.
\end{aligned}$$

We can minimize this w.r.t. x by setting the corresponding derivative to zero and solving for x:

$$\frac{dg}{dx} = \eta - \frac{1}{x} = 0 \quad \Longrightarrow \quad x = \frac{1}{\eta}.$$

Substituting this in (10.133), we see that

$$g(\eta) = 1 - \ln\left(\frac{1}{\eta}\right).$$

If we substitute this into (10.132), we get

$$f(x) = \min_\eta \left\{\eta x - 1 + \ln\left(\frac{1}{\eta}\right)\right\}.$$

Again, we can minimize this w.r.t. η by setting the corresponding derivative to zero and solving for η:

$$\frac{df}{d\eta} = x - \frac{1}{\eta} = 0 \quad \Longrightarrow \quad \eta = \frac{1}{x},$$

and substituting this into (10.132), we find that

$$f(x) = \frac{1}{x}x - 1 + \ln\left(\frac{1}{1/x}\right) = \ln(x).$$

10.30 **NOTE**: Please consult note preceding Solution 10.29 for relevant corrections.

Differentiating the log logistic function, we get

$$\frac{d}{dx}\ln\sigma = \left(1 + e^{-x}\right)^{-1} e^{-x} = \sigma(x)e^{-x} \tag{291}$$

and, using (4.88),

$$\frac{d^2}{dx^2}\ln\sigma = \sigma(x)(1-\sigma(x))e^{-x} - \sigma(x)e^{-x} = -\sigma(x)^2 e^{-x}$$

which will always be negative and hence $\ln\sigma(x)$ is concave.

From (291), we see that the first order Taylor expansion of $\ln\sigma(x)$ around ξ becomes

$$\ln\sigma(x) = \ln\sigma(\xi) + (x-\xi)\sigma(\xi)e^{-\xi} + O\left((x-xi)^2\right).$$

Since $\ln\sigma(x)$ is concave, its tangent line will be an upper bound and hence

$$\ln\sigma(x) \leqslant \ln\sigma(\xi) + (x-\xi)\sigma(\xi)e^{-\xi}. \tag{292}$$

Following the presentation in Section 10.5, we define

$$\eta = \sigma(\xi)e^{-\xi}. \tag{293}$$

Using (4.60), we have

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

and hence

$$\sigma(\xi) = 1 - \eta.$$

From this and (293)

$$\begin{aligned}\xi &= \ln\sigma(\xi) - \ln\eta \\ &= \ln(1-\eta) - \ln\eta.\end{aligned}$$

Using these results in (292), we have

$$\ln\sigma(x) \leqslant \ln(1-\eta) + x\eta - \eta\left[\ln(1-\eta) - \ln\eta\right].$$

By exponentiating both sides and making use of (10.135), we obtain (10.137).

10.31 **NOTE**: Please consult note preceding Solution 10.29 for relevant corrections.

Taking the derivative of $f(x)$ w.r.t. x we get

$$\frac{\mathrm{d}f}{\mathrm{d}x} = -\frac{1}{e^{x/2}+e^{-x/2}}\frac{1}{2}\left(e^{x/2}-e^{-x/2}\right) = -\frac{1}{2}\tanh\left(\frac{x}{2}\right)$$

where we have used (5.59). From (5.60), we get

$$f''(x) = \frac{d^2f}{dx^2} = -\frac{1}{4}\left(1-\tanh\left(\frac{x}{2}\right)^2\right).$$

Since $\tanh(x/2)^2 < 1$ for finite values of x, $f''(x)$ will always be negative and so $f(x)$ is concave.

Next we define $y = x^2$, noting that y will always be non-negative, and express f as a function of y:

$$f(y) = -\ln\left\{\exp\left(\frac{\sqrt{y}}{2}\right) + \exp\left(-\frac{\sqrt{y}}{2}\right)\right\}.$$

We then differentiate f w.r.t. y, yielding

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}y} &= -\left\{\exp\left(\frac{\sqrt{y}}{2}\right) + \exp\left(-\frac{\sqrt{y}}{2}\right)\right\}^{-1} \\ &\qquad \frac{1}{4\sqrt{y}}\left\{\exp\left(\frac{\sqrt{y}}{2}\right) - \exp\left(-\frac{\sqrt{y}}{2}\right)\right\} \qquad (294)\\ &= -\frac{1}{4\sqrt{y}}\tanh\left(\frac{\sqrt{y}}{2}\right). \qquad (295)\end{aligned}$$

and, using (5.60),

$$\begin{aligned}\frac{\mathrm{d}^2f}{\mathrm{d}y^2} &= \frac{1}{8y^{3/2}}\tanh\left(\frac{\sqrt{y}}{2}\right) - \frac{1}{16y}\left\{1-\tanh\left(\frac{\sqrt{y}}{2}\right)^2\right\} \\ &= \frac{1}{8y}\left(\tanh\left(\frac{\sqrt{y}}{2}\right)\left\{\frac{1}{\sqrt{y}} + \frac{1}{2}\tanh\left(\frac{\sqrt{y}}{2}\right)\right\} - \frac{1}{2}\right). \qquad (296)\end{aligned}$$

We see that this will be positive if the factor inside the outermost parenthesis is positive, which is equivalent to

$$\frac{1}{\sqrt{y}}\tanh\left(\frac{\sqrt{y}}{2}\right) > \frac{1}{2}\left\{1-\tanh^2\left(\frac{\sqrt{y}}{2}\right)\right\}.$$

If we divide both sides by $\tanh\left(\sqrt{y}/2\right)$, substitute a for $\sqrt{y}/2$ and then make use of (5.59), we can write this as

$$\begin{aligned}\frac{1}{a} &> \frac{e^a+e^{-a}}{e^a-e^{-a}} - \frac{e^a-e^{-a}}{e^a+e^{-a}} \\ &= \frac{\left(e^a+e^{-a}\right)^2-\left(e^a-e^{-a}\right)^2}{\left(e^a-e^{-a}\right)\left(e^a+e^{-a}\right)} \\ &= \frac{4}{e^{2a}-e^{-2a}}.\end{aligned}$$

Taking the inverse of both sides of this inequality we get

$$a < \frac{1}{4}\left(e^{2a}-e^{-2a}\right).$$

If differentiate both sides w.r.t. a we see that the derivatives are equal at $a=0$ and for $a>0$, the derivative of the r.h.s. will be greater than that of the l.h.s. Thus, the r.h.s. will grow faster and the inequality will hold for $a>0$. Consequently (296) will be positive for $y>0$ and approach $+\infty$ as y approaches 0.

Now we use (295) to make a Taylor expansion of $f(x^2)$ around ξ^2, which gives

$$\begin{aligned}f(x^2) &= f(\xi^2)+(x^2-\xi^2)f'(\xi^2)+O\left((x^2-\xi^2)^2\right) \\ &\geqslant -\ln\left\{\exp\left(\frac{\xi}{2}\right)+\exp\left(-\frac{\xi}{2}\right)\right\}-(x^2-\xi^2)\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right).\end{aligned}$$

where we have used the fact that f is convex function of x^2 and hence its tangent will be a lower bound. Defining

$$\lambda(\xi)=\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right)$$

we recover (10.143), from which (10.144) follows.

10.32 We can see this from the lower bound (10.154), which is simply a sum of the prior and indepedent contributions from the data points, all of which are quadratic in $\mathbf{w}$. A new data point would simply add another term to this sum and we can regard terms from the previously arrived data points and the original prior collectively as a revised prior, which should be combined with the contributions from the new data point.

The corresponding sufficient statistics, (10.157) and (10.158), can be rewritten di-

rectly in the corresponding sequential form,

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N}(t_n - 1/2)\boldsymbol{\phi}_n\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N-1}(t_n - 1/2)\boldsymbol{\phi}_n + (t_N - 1/2)\boldsymbol{\phi}_N\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_{N-1}^{-1}\mathbf{S}_{N-1}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N-1}(t_n - 1/2)\boldsymbol{\phi}_n\right) + (t_N - 1/2)\boldsymbol{\phi}_N\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_{N-1}^{-1}\mathbf{m}_{N-1} + (t_N - 1/2)\boldsymbol{\phi}_N\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} \\
&= \mathbf{S}_0^{-1} + 2\sum_{n=1}^{N-1}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} + 2\lambda(\xi_N)\boldsymbol{\phi}_N\boldsymbol{\phi}_N^{\mathrm{T}} \\
&= \mathbf{S}_{N-1}^{-1} + 2\lambda(\xi_N)\boldsymbol{\phi}_N\boldsymbol{\phi}_N^{\mathrm{T}}.
\end{aligned}$$

The update formula for the variational parameters, (10.163), remain the same, but each parameter is updated only once, although this update will be part of an iterative scheme, alternating between updating $\mathbf{m}_N$ and $\mathbf{S}_N$ with ξ_N kept fixed, and updating ξ_N with $\mathbf{m}_N$ and $\mathbf{S}_N$ kept fixed. Note that updating ξ_N will not affect $\mathbf{m}_{N-1}$ and $\mathbf{S}_{N-1}$. Note also that this updating policy differs from that of the batch learning scheme, where all variational parameters are updated using statistics based on all data points.

10.33 Taking the derivative of (10.161) w.r.t. ξ_n, we get

$$\begin{aligned}
\frac{\partial Q}{\partial \xi_n} &= \frac{1}{\sigma(\xi_n)}\sigma'(\xi_n) - \frac{1}{2} - \lambda'(\xi_n)\left(\phi_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\phi - \xi_n^2\right) + \lambda(\xi_n)2\xi_n \\
&= \frac{1}{\sigma(\xi_n)}\sigma(\xi_n)(1-\sigma(x)) - \frac{1}{2} - \lambda'(\xi_n)\left(\phi_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\phi - \xi_n^2\right) \\
&\qquad + \frac{1}{2\xi_n}\left[\sigma(x) - \frac{1}{2}\right]2\xi_n \\
&= -\lambda'(\xi_n)\left(\phi_n^{\mathrm{T}}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\phi - \xi_n^2\right)
\end{aligned}$$

where we have used (4.88) and (10.141). Setting this equal to zero, we obtain (10.162), from which (10.163) follows.

10.34 **NOTE**: In PRML, there are a number of sign errors in Equation (10.164); the correct form is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\xi}) &= \frac{1}{2}\ln\frac{|\mathbf{S}_N|}{|\mathbf{S}_0|}+\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N-\frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\\ &\quad+\sum_{n=1}^{N}\left\{\ln\sigma(\xi_n)-\frac{1}{2}\xi_n+\lambda(\xi_n)\xi_n^2\right\}.\end{aligned}$$

We can differentiate $\mathcal{L}$ w.r.t. ξ_n using (3.117) and results from Solution 10.33, to obtain

$$\frac{\partial\mathcal{L}}{\partial\xi_n}=\frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_N^{-1}\frac{\partial\mathbf{S}_N}{\partial\xi_n}\right)+\frac{1}{2}\mathrm{Tr}\left(\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\frac{\partial\mathbf{S}_N}{\partial\xi_n}\right)+\lambda'(\xi_n)\xi_n^2 \tag{297}$$

where we have defined

$$\mathbf{a}_N=\mathbf{S}_N^{-1}\mathbf{m}_N. \tag{298}$$

From (10.158) and (C.21), we get

$$\begin{aligned}\frac{\partial\mathbf{S}_N}{\partial\xi_n} &= \frac{\partial\left(\mathbf{S}_N^{-1}\right)^{-1}}{\partial\xi_n}=-\mathbf{S}_N\frac{\partial\mathbf{S}_N^{-1}}{\partial\xi_n}\mathbf{S}_N\\ &= -\mathbf{S}_N 2\lambda'(\xi_n)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N.\end{aligned}$$

Substituting this into (297) and setting the result equal to zero, we get

$$-\frac{1}{2}\mathrm{Tr}\left(\left(\mathbf{S}_N^{-1}+\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\right)\mathbf{S}_N 2\lambda'(\xi_n)\phi_n\phi_n^{\mathrm{T}}\mathbf{S}_N\right)+\lambda'(\xi_n)\xi_n^2=0.$$

Rearranging this and making use of (298) we get

$$\begin{aligned}\xi_n^2 &= \phi_n^{\mathrm{T}}\mathbf{S}_N\left(\mathbf{S}_N^{-1}+\mathbf{a}_N\mathbf{a}_N^{\mathrm{T}}\right)\mathbf{S}_N\phi_n\\ &= \phi_n^{\mathrm{T}}\left(\mathbf{S}_N+\mathbf{m}_N\mathbf{m}_N^{\mathrm{T}}\right)\phi_n\end{aligned}$$

where we have also used the symmetry of $\mathbf{S}_N$.

10.35 **NOTE**: See note in Solution 10.34.

From (2.43), (4.140) and (10.153), we see that

$$\begin{aligned}p(\mathbf{w})h(\mathbf{w},\boldsymbol{\xi}) &= (2\pi)^{-W/2}|\mathbf{S}_0|^{-1/2}\\ &\quad\exp\left\{-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}+2\sum_{n=1}^{N}\lambda(\xi_n)\phi_n\phi_n^{\mathrm{T}}\right)\mathbf{w}\right.\\ &\qquad\left.+\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\sum_{n=1}^{N}\phi_n\left[t_n-\frac{1}{2}\right]\right)\right\}\\ &\quad\exp\left\{-\frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0+\sum_{n=1}^{N}\frac{\xi_n}{2}+\lambda(\xi_n)\xi_n^2\right\}\prod_{n=1}^{N}\sigma(\xi_n).\end{aligned}$$

Using (10.157) and (10.158), we can complete the square over $\mathbf{w}$, yielding

$$\begin{aligned} p(\mathbf{w})h(\mathbf{w},\boldsymbol{\xi}) &= (2\pi)^{-W/2}\,|\mathbf{S}_0|^{-1/2}\prod_{n=1}^{N}\sigma(\xi_n) \\ &\quad \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\} \\ &\quad \exp\left\{\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 \sum_{n=1}^{N}\frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\right\}. \end{aligned}$$

Now we can do the integral over $\mathbf{w}$ in (10.159), in effect replacing the first exponential factor with $(2\pi)^{W/2}\,|\mathbf{S}_N|^{1/2}$. Taking logarithm, we then obtain (10.164).

10.36 If we denote the joint distribution corresponding to the first j factors by $p_j(\boldsymbol{\theta},\mathcal{D})$, with corresponding evidence $p_j(\mathcal{D})$, then we have

$$\begin{aligned} p_j(\mathcal{D}) &= \int p_j(\boldsymbol{\theta},\mathcal{D})\,\mathrm{d}\boldsymbol{\theta} = \int p_{j-1}(\boldsymbol{\theta},\mathcal{D})f_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= p_{j-1}(\mathcal{D})\int p_{j-1}(\boldsymbol{\theta}|\mathcal{D})f_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &\simeq p_{j-1}(\mathcal{D})\int q_{j-1}(\boldsymbol{\theta})f_j(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = p_{j-1}(\mathcal{D})Z_j. \end{aligned}$$

By applying this result recursively we see that the evidence is given by the product of the normalization constants

$$p(\mathcal{D}) = \prod_j Z_j.$$

10.37 Here we use the general expectation-propagation equations (10.204)–(10.207). The initial $q(\boldsymbol{\theta})$ takes the form

$$q_{\text{init}}(\boldsymbol{\theta}) = \widetilde{f}_0(\boldsymbol{\theta})\prod_{i\neq 0}\widetilde{f}_i(\boldsymbol{\theta})$$

where $\widetilde{f}_0(\boldsymbol{\theta}) = f_0(\boldsymbol{\theta})$. Thus

$$q^{\backslash 0}(\boldsymbol{\theta}) \propto \prod_{i\neq 0}\widetilde{f}_i(\boldsymbol{\theta})$$

and $q^{\text{new}}(\boldsymbol{\theta})$ is determined by matching moments (sufficient statistics) against

$$q^{\backslash 0}(\boldsymbol{\theta})f_0(\boldsymbol{\theta}) = q_{\text{init}}(\boldsymbol{\theta}).$$

Since by definition this belongs to the same exponential family form as $q^{\text{new}}(\boldsymbol{\theta})$ it follows that

$$q^{\text{new}}(\boldsymbol{\theta}) = q_{\text{init}}(\boldsymbol{\theta}) = q^{\backslash 0}(\boldsymbol{\theta})f_0(\boldsymbol{\theta}).$$

Thus

$$\widetilde{f}_0(\boldsymbol{\theta}) = \frac{Z_0\, q^{\text{new}}(\boldsymbol{\theta})}{q^{\backslash 0}(\boldsymbol{\theta})} = Z_0 f_0(\boldsymbol{\theta})$$

where

$$Z_0 = \int q^{\backslash 0}(\boldsymbol{\theta}) f_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \int q^{\text{new}}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = 1.$$

10.38 The ratio is given by

$$\begin{aligned} q^{\backslash n}(\boldsymbol{\theta}) &\propto \exp\left\{-\frac{1}{2v}\|\boldsymbol{\theta}-\mathbf{m}\|^2 + \frac{1}{2v_n}\|\boldsymbol{\theta}-\mathbf{m}_n\|^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}\left(\frac{1}{v}-\frac{1}{v_n}\right) + \boldsymbol{\theta}^{\mathrm{T}}\left(\frac{1}{v}\mathbf{m}-\frac{1}{v_n}\mathbf{m}_n\right)\right\} \end{aligned}$$

from which we obtain the variance given by (10.215). The mean is then obtained by completing the square and is therefore given by

$$\begin{aligned} \mathbf{m}^{\backslash n} &= v^{\backslash n}\left(v^{-1}\mathbf{m} - v_n^{-1}\mathbf{m}_n\right) \\ &= v^{\backslash n}\left(v^{-1}\mathbf{m} - v_n^{-1}\mathbf{m}_n\right) + v^{\backslash n}v_n^{-1}\mathbf{m} - v^{\backslash n}v_n^{-1}\mathbf{m} \\ &= v^{\backslash n}\left(v^{-1} - v_n^{-1}\right)\mathbf{m} + v^{\backslash n}v_n^{-1}\left(\mathbf{m}-\mathbf{m}_n\right) \end{aligned}$$

Hence we obtain (10.214).

The normalization constant is given by

$$Z_n = \int \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}^{\backslash n}, v^{\backslash n}\mathbf{I})\left\{(1-w)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\theta},\mathbf{I}) + w\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I})\right\}\,\mathrm{d}\boldsymbol{\theta}.$$

The first term can be integrated by using the result (2.115) while the second term is trivial since the background distribution does not depend on $\boldsymbol{\theta}$ and hence can be taken outside the integration. We therefore obtain (10.216).

10.39 **NOTE**: In PRML, a term $v^{\backslash n}D$ should be added to the r.h.s. of (10.245).

We derive (10.244) by noting

$$\begin{aligned} \nabla_{\mathbf{m}^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}\nabla_{\mathbf{m}^{\backslash n}} \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z_n}\int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\left\{-\frac{1}{v^{\backslash n}}(\mathbf{m}^{\backslash n}-\boldsymbol{\theta})\right\}\,\mathrm{d}\boldsymbol{\theta} \\ &= -\frac{\mathbf{m}^{\backslash n}}{v^{\backslash n}} + \frac{\mathbb{E}[\boldsymbol{\theta}]}{v^{\backslash n}}. \end{aligned}$$

We now use this to derive (10.217) by substituting for Z_n using (10.216) to give

$$\begin{aligned} \nabla_{\mathbf{m}^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}(1-w)\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n}, (v^{\backslash n}+1)\mathbf{I})\frac{1}{v^{\backslash n}+1}(\mathbf{x}_n - \mathbf{m}^{\backslash n}) \\ &= \rho_n \frac{1}{v^{\backslash n}+1}(\mathbf{x}_n - \mathbf{m}^{\backslash n}) \end{aligned}$$

where we have defined

$$\rho_n = (1-w)\frac{1}{Z_n}\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n},(v^{\backslash n}+1)\mathbf{I}) = 1 - \frac{w}{Z_n}\mathcal{N}(\mathbf{x}_n|\mathbf{0},a\mathbf{I}).$$

Similarly for (10.245) we have

$$\begin{aligned}\nabla_{v^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}\nabla_{v^{\backslash n}} \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z_n}\int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\left\{\frac{1}{2(v^{\backslash n})^2}\left(\mathbf{m}^{\backslash n}-\boldsymbol{\theta}\right)^{\mathrm{T}}\left(\mathbf{m}^{\backslash n}-\boldsymbol{\theta}\right) - \frac{D}{2v^{\backslash n}}\right\}\mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{2(v^{\backslash n})^2}\left\{\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}] - 2\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}]\mathbf{m}^{\backslash n} + \|\mathbf{m}^{\backslash n}\|^2\right\} - \frac{D}{2v^{\backslash n}}.\end{aligned}$$

Re-arranging we obtain (10.245). Now we substitute for Z_n using (10.216) to give

$$\begin{aligned}\nabla_{v^{\backslash n}} \ln Z_n &= \frac{1}{Z_n}(1-w)\mathcal{N}(\mathbf{x}_n|\mathbf{m}^{\backslash n},(v^{\backslash n}+1)\mathbf{I}) \\ &\qquad \left[\frac{1}{2(v^{\backslash n}+1)^2}\|\mathbf{x}_n-\mathbf{m}^{\backslash n}\|^2 - \frac{D}{2(v^{\backslash n}+1)}\right].\end{aligned}$$

Next we note that the variance is given by

$$v\mathbf{I} = \mathbb{E}[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{\theta}]\mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}]$$

and so, taking the trace, we have

$$Dv = \mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}] - \mathbb{E}[\boldsymbol{\theta}^{\mathrm{T}}]\mathbb{E}[\boldsymbol{\theta}]$$

where D is the dimensionality of $\boldsymbol{\theta}$. Combining the above results we obtain (10.218).

Chapter 11 Sampling Methods

11.1 Since the samples are independent, for the mean, we have

$$\mathbb{E}\left[\widehat{f}\right] = \frac{1}{L}\sum_{l=1}^{L}\int f(z^{(l)})p(z^{(l)})\,\mathrm{d}z^{(l)} = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[f\right] = \mathbb{E}\left[f\right].$$

Using this together with (1.38) and (1.39), for the variance, we have

$$\begin{aligned}\mathrm{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\left(\widehat{f}-\mathbb{E}\left[\widehat{f}\right]\right)^2\right] \\ &= \mathbb{E}\left[\widehat{f}^2\right] - \mathbb{E}\left[f\right]^2.\end{aligned}$$

Now note

$$\begin{aligned}\mathbb{E}\left[f(z^{(k)}), f(z^{(m)})\right] &= \begin{cases} \mathrm{var}[f] + \mathbb{E}[f^2] & \text{if } n = k, \\ \mathbb{E}[f^2] & \text{otherwise,} \end{cases} \\ &= \mathbb{E}[f^2] + \delta_{mk}\mathrm{var}[f],\end{aligned}$$

where we again exploited the fact that the samples are independent.

Hence

$$\begin{aligned}\mathrm{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\frac{1}{L}\sum_{m=1}^{L} f(z^{(m)})\frac{1}{L}\sum_{k=1}^{L} f(z^{(k)})\right] - \mathbb{E}[f]^2 \\ &= \frac{1}{L^2}\sum_{m=1}^{L}\sum_{k=1}^{L}\left\{\mathbb{E}[f^2] + \delta_{mk}\mathrm{var}[f]\right\} - \mathbb{E}[f]^2 \\ &= \frac{1}{L}\mathrm{var}[f] \\ &= \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right].\end{aligned}$$

11.2 From (1.27) we have,

$$p_y(y) = p_z(h(y))\,|h'(y)|\,.$$

Differentiating (11.6) w.r.t. y and using the fact that $p_z(h(y)) = 1$, we see that

$$p_y(y) = p(y).$$

11.3 Using the standard integral

$$\int \frac{1}{a^2 + u^2}\,\mathrm{d}u = \frac{1}{a}\tan^{-1}\left(\frac{u}{a}\right) + C$$

where C is a constant, we can integrate the r.h.s. of (11.8) to obtain

$$z = h(y) = \int_{-\infty}^{y} p(\widehat{y})\,\mathrm{d}\widehat{y} = \frac{1}{\pi}\tan^{-1}(y) + \frac{1}{2}$$

where we have chosen the constant $C = 1/2$ to ensure that the range of the cumulative distribution function is $[0, 1]$.

Thus the required transformation function becomes

$$y = h^{-1}(z) = \tan\left(\pi\left(z - \frac{1}{2}\right)\right).$$

11.4 We need to calculate the determinant of the Jacobian

$$\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}.$$

In doing so, we will find it helpful to make use of intermediary variables in polar coordinates

$$\theta = \tan^{-1}\frac{z_2}{z_1} \tag{299}$$

$$r^2 = z_1^2 + z_2^2 \tag{300}$$

from which it follows that

$$z_1 = r\cos\theta \tag{301}$$

$$z_2 = r\sin\theta. \tag{302}$$

From (301) and (302) we have

$$\frac{\partial(z_1, z_2)}{\partial(r, \theta)} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix}$$

and thus

$$\left|\frac{\partial(z_1, z_2)}{\partial(r, \theta)}\right| = r(\cos^2\theta + \sin^2\theta) = r. \tag{303}$$

From (11.10), (11.11) and (300)–(302) we have

$$y_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\cos\theta \tag{304}$$

$$y_2 = z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\sin\theta \tag{305}$$

which give

$$\frac{\partial(y_1, y_2)}{\partial(r, \theta)} = \begin{pmatrix} -2\cos\theta\left(-2\ln r^2\right)^{-1/2} r^{-1} & -2\sin\theta\left(-2\ln r^2\right)^{-1/2} r^{-1} \\ -\sin\theta\left(-2\ln r^2\right)^{1/2} & \cos\theta\left(-2\ln r^2\right)^{1/2} \end{pmatrix}$$

and thus

$$\left|\frac{\partial(r, \theta)}{\partial(y_1, y_2)}\right| = \left|\frac{\partial(y_1, y_2)}{\partial(r, \theta)}\right|^{-1} = \left(-2r^{-1}(\cos^2\theta + \sin^2\theta)\right)^{-1} = -\frac{r}{2}.$$

Combining this with (303), we get

$$\begin{aligned} \left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right| &= \left|\frac{\partial(z_1, z_2)}{\partial(r, \theta)}\frac{\partial(r, \theta)}{\partial(y_1, y_2)}\right| \\ &= \left|\frac{\partial(z_1, z_2)}{\partial(r, \theta)}\right|\left|\frac{\partial(r, \theta)}{\partial(y_1, y_2)}\right| = -\frac{r^2}{2} \end{aligned} \tag{306}$$

However, we only retain the absolute value of this, since both sides of (11.12) must be non-negative. Combining this with

$$p(z_1, z_2) = \frac{1}{\pi}$$

which follows from the fact that z_1 and z_2 are uniformly distributed on the unit circle, we can rewrite (11.12) as

$$p(y_1, y_2) = \frac{1}{2\pi} r^2. \tag{307}$$

By squaring the left- and rightmost sides of (304) and (305), adding up the results and rearranging, we see that

$$r^2 = \exp\left(-\frac{1}{2}\left(y_1^2 + y_2^2\right)\right)$$

which toghether with (307) give (11.12).

11.5 Since $\mathbb{E}[\mathbf{z}] = \mathbf{0}$,

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\boldsymbol{\mu} + \mathbf{L}\mathbf{z}] = \boldsymbol{\mu}.$$

Similarly, since $\mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] = \mathbf{I}$,

$$\begin{aligned}
\operatorname{cov}[\mathbf{y}] &= \mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] - \mathbb{E}[\mathbf{y}]\,\mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\mu} + \mathbf{L}\mathbf{z})(\boldsymbol{\mu} + \mathbf{L}\mathbf{z})^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \mathbf{L}\mathbf{L}^{\mathrm{T}} \\
&= \boldsymbol{\Sigma}.
\end{aligned}$$

11.6 The probability of acceptance follows trivially from the mechanism used to accept or reject the sample. The probability of a sample u drawn uniformly from the interval $[0, kq(\mathbf{z})]$ being less than or equal to a value $\widetilde{p}(\mathbf{z}) \leqslant kq(\mathbf{z})$ is simply

$$p(\text{acceptance}|\mathbf{z}) = \int_0^{\widetilde{p}(\mathbf{z})} \frac{1}{kq(\mathbf{z})}\,\mathrm{d}u = \frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})}.$$

Therefore, the probability density for drawing a sample, $\mathbf{z}$, is

$$q(\mathbf{z})p(\text{acceptance}|\mathbf{z}) = q(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})} = \frac{\widetilde{p}(\mathbf{z})}{k}. \tag{308}$$

Since $\widetilde{p}(\mathbf{z})$ is proportional to $p(\mathbf{x})$,

$$p(\mathbf{z}) = \frac{1}{Z_{\widetilde{p}}}\widetilde{p}(\mathbf{z}),$$

where

$$Z_{\widetilde{p}} = \int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

As the l.h.s. of (308) is a probability density that integrates to 1, it follows that

$$\int \frac{\widetilde{p}(\mathbf{z})}{k}\,\mathrm{d}\mathbf{z} = 1$$

and so $k = Z_{\widetilde{p}}$, and

$$\frac{\widetilde{p}(\mathbf{z})}{k} = p(\mathbf{z}),$$

as required.

11.7 **NOTE**: In PRML, the roles of y and z in the text of the exercise should be swapped in order to be consistent with the notation used in Section 11.1.2, including (11.16); this is opposite to the notation used in Section 11.1.1.

We will suppose that y has a uniform distribution on $[0, 1]$ and we seek the distribution of z, which is derived from

$$z = b \tan y + c.$$

From (11.5), we have

$$q(z) = p(y) \left| \frac{\mathrm{d}y}{\mathrm{d}z} \right| \tag{309}$$

From the inverse transformation

$$y = \tan^{-1}(u(z)) = \tan^{-1}\left(\frac{z-c}{b}\right)$$

where we have implicitly defined $u(z)$, we see that

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}z} &= \frac{\mathrm{d}}{\mathrm{d}u} \tan^{-1}(u) \frac{\mathrm{d}u}{\mathrm{d}z} \\ &= \frac{1}{1+u^2} \frac{\mathrm{d}u}{\mathrm{d}z} \\ &= \frac{1}{1+(z-c)^2/b^2} \frac{1}{b}. \end{aligned}$$

Substituting this into (309), using the fact that $p(y) = 1$ and finally absorbing the factor $1/b$ into k, we obtain (11.16). .

11.8 **NOTE**: In PRML, equation (11.17) and the following end of the sentence need to modified as follows:

$$q(z) = k_i \lambda_i \exp\left\{-\lambda_i \left(z - z_i\right)\right\} \qquad \widehat{z}_{i-1,i} < z \leqslant \widehat{z}_{i,i+1}$$

where $\widehat{z}_{i-1,i}$ is the point of intersection of the tangent lines at z_{i-1} and z_i, λ_i is the slope of the tangent at z_i and k_i accounts for the corresponding offset.

We start by determining $\widetilde{q}(z)$ with coefficients $\widetilde{k}_i$, such that $\widetilde{q}(z) \geqslant p(z)$ everywhere. From Figure 11.6, we see that

$$\widetilde{q}(z_i) = p(z_i)$$

and thus, from (11.17),

$$\begin{aligned} \widetilde{q}(z_i) &= \widetilde{k}_i \lambda_i \exp\left(-\lambda_i (z_i - z_i)\right) \\ &= \widetilde{k}_i \lambda_i = p(z_i). \end{aligned} \tag{310}$$

Next we compute the normalization constant for q in terms of k_i,

$$\begin{aligned} Z_q &= \int \widetilde{q}(z)\,\mathrm{d}z \\ &= \sum_{i=1}^{K} \widetilde{k}_i \lambda_i \int_{\widehat{z}_{i-1,i}}^{\widehat{z}_{i,i+1}} \exp\left(-\lambda_i(z - z_i)\right)\,\mathrm{d}z \\ &= \sum_{i=1}^{K} k_i \end{aligned} \tag{311}$$

where K denotes the number of grid points and

$$\begin{aligned} k_i &= \widetilde{k}_i \lambda_i \int_{\widehat{z}_{i-1,i}}^{\widehat{z}_{i,i+1}} \exp\left(-\lambda_i(z - z_i)\right)\,\mathrm{d}z \\ &= \widetilde{k}_i \left(\exp\left\{-\lambda_i\left(\widehat{z}_{i-1,i} - z_i\right)\right\} - \exp\left\{-\lambda_i\left(\widehat{z}_{i,i+1} - z_i\right)\right\}\right). \end{aligned} \tag{312}$$

Note that $\widehat{z}_{0,1}$ and $\widehat{z}_{K,K+1}$ equal the lower and upper limits on z, respectively, or $-\infty/+\infty$ where no such limits exist.

11.9 **NOTE**: See correction detailed in Solution 11.8

To generate a sample from $q(z)$, we first determine the segment of the envelope function from which the sample will be generated. The probability mass in segment i is given from (311) as k_i/Z_q. Hence we draw v from $U(v|0,1)$ and obtain

$$i = \begin{cases} 1 & \text{if } v \leqslant k_1/Z_q \\ m & \text{if } \sum_{j=1}^{m-1} k_j/Z_q < v \leqslant \sum_{j=1}^{m} k_j/Z_q, \quad 1 < m < K \\ K & \text{otherwise.} \end{cases}$$

Next, we can use the techniques of Section 11.1.1 to sample from the exponential distribution corresponding to the chosen segment i. We must, however, take into account that we now only want to sample from a finite interval of the exponential distribution and so the lower limit in (11.6) will be $\widehat{z}_{i-1,i}$. If we consider the uniformly distributed variable w, $U(w|0,1)$, (11.6), (310) and (311) give

$$\begin{aligned} w &= h(z) = \int_{\widehat{z}_{i-1,i}}^{z} q(\widetilde{z})\,\mathrm{d}\widetilde{z} \\ &= \frac{\widetilde{k}_i}{k_i} \lambda_i \exp\left(\lambda_i z_i\right) \int_{\widehat{z}_{i-1,i}}^{z} \exp\left(-\lambda_i \widetilde{z}\right)\,\mathrm{d}\widetilde{z} \\ &= \frac{\widetilde{k}_i}{k_i} \exp\left(\lambda_i z_i\right) \left[\exp\left(-\lambda_i \widehat{z}_{i-1,i}\right) - \exp\left(-\lambda_i z\right)\right]. \end{aligned}$$

Thus, by drawing a sample $w^\star$ and transforming it according to

$$\begin{aligned} z^\star &= \frac{1}{\lambda_i}\ln\left[w^\star\frac{k_i}{\widetilde{k}_i\exp\left(\lambda_i z_i\right)} - \exp\left(-\lambda_i\widehat{z}_{i-1,i}\right)\right] \\ &= \frac{1}{\lambda_i}\ln\left[w^\star\left(\exp\left\{-\lambda_i\widehat{z}_{i-1,i}\right\} - \exp\left\{-\lambda_i\widehat{z}_{i,i+1}\right\}\right) - \exp\left(-\lambda_i\widehat{z}_{i-1,i}\right)\right] \end{aligned}$$

where we have used (312), we obtain a sample from $q(z)$.

11.10 **NOTE**: In PRML, "$z^{(1)} = 0$" should be "$z^{(0)} = 0$" on the first line following Equation (11.36)

From (11.34)–(11.36) and the fact that $\mathbb{E}\left[z^{(\tau)}\right] = 0$ for all values of τ, we have

$$\begin{aligned} \mathbb{E}_{z^{(\tau)}}\left[\left(z^{(\tau)}\right)^2\right] &= 0.5\,\mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}\right)^2\right] + 0.25\,\mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}+1\right)^2\right] \\ &\quad +0.25\,\mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}-1\right)^2\right] \\ &= \mathbb{E}_{z^{(\tau-1)}}\left[\left(z^{(\tau-1)}\right)^2\right] + \frac{1}{2}. \end{aligned} \tag{313}$$

With $z^{(0)} = 0$ specified in the text following (11.36), (313) gives

$$\mathbb{E}_{z^{(1)}}\left[\left(z^{(1)}\right)^2\right] = \mathbb{E}_{z^{(0)}}\left[\left(z^{(0)}\right)^2\right] + \frac{1}{2} = \frac{1}{2}.$$

Assuming that

$$\mathbb{E}_{z^{(k)}}\left[\left(z^{(k)}\right)^2\right] = \frac{k}{2}$$

(313) immediatly gives

$$\mathbb{E}_{z^{(k+1)}}\left[\left(z^{(k+1)}\right)^2\right] = \frac{k}{2} + \frac{1}{2} = \frac{k+1}{2}$$

and thus

$$\mathbb{E}_{z^{(\tau)}}\left[\left(z^{(\tau)}\right)^2\right] = \frac{\tau}{2}.$$

11.11 This follows from the fact that in Gibbs sampling, we sample a single variable, z_k, at the time, while all other variables, $\{z_i\}_{i\neq k}$, remain unchanged. Thus, $\{z_i'\}_{i\neq k} = \{z_i\}_{i\neq k}$ and we get

$$\begin{aligned} p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') &= p^\star(z_k,\{z_i\}_{i\neq k})p^\star(z_k'|\{z_i\}_{i\neq k}) \\ &= p^\star(z_k|\{z_i\}_{i\neq k})p^\star(\{z_i\}_{i\neq k})p^\star(z_k'|\{z_i\}_{i\neq k}) \\ &= p^\star(z_k|\{z_i'\}_{i\neq k})p^\star(\{z_i'\}_{i\neq k})p^\star(z_k'|\{z_i'\}_{i\neq k}) \\ &= p^\star(z_k|\{z_i'\}_{i\neq k})p^\star(z_k',\{z_i'\}_{i\neq k}) \\ &= p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z}), \end{aligned}$$

where we have used the product rule together with $T(\mathbf{z},\mathbf{z}') = p^\star(z_k'|\{z_i\}_{i\neq k})$.

11.12 Gibbs sampling is *not* ergodic w.r.t. the distribution shown in Figure 11.15, since the two regions of non-zero probability do not overlap when projected onto either the z_1-ot the z_2-axis. Thus, as the initial sample will fall into one and only one of the two regions, all subsequent samples will also come from that region. However, had the initial sample fallen into the other region, the Gibbs sampler would have remained in that region. Thus, the corresponding Markov chain would have two stationary distributions, which is counter to the definition of the equilibrium distribution.

11.13 The joint distribution over x, μ and τ can be written

$$p(x, \mu, \tau|\mu_0, s_0, a, b) = \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathcal{N}\left(\mu|\mu_0, s_0\right) \text{Gam}\left(\tau|a, b\right).$$

From Bayes' theorem we see that

$$p(\mu|x, \tau, \mu_0, s_0) \propto \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathcal{N}\left(\mu|\mu_0, s_0\right)$$

which, from (2.113)–(2.117), is also a Gaussian,

$$\mathcal{N}\left(\mu|\widehat{\mu}, \widehat{s}\right)$$

with parameters

$$\begin{aligned} \widehat{s}^{-1} &= s_0^{-1} + \tau \\ \widehat{\mu} &= \widehat{s}\left(\tau x + s_0^{-1}\mu_0\right). \end{aligned}$$

Similarly,

$$p(\tau|x, \mu, a, b) \propto \mathcal{N}\left(x|\mu, \tau^{-1}\right) \text{Gam}\left(\tau|a, b\right)$$

which, we know from Section 2.3.6, is a gamma distribution

$$\text{Gam}\left(\tau|\widehat{a}, \widehat{b}\right)$$

with parameters

$$\begin{aligned} \widehat{a} &= a + \frac{1}{2} \\ \widehat{b} &= b + \frac{1}{2}\left(x - \mu\right)^2. \end{aligned}$$

11.14 **NOTE**: In PRML, α_i^2 should be α^2 in the last term on the r.h.s. of (11.50).

If we take the expectation of (11.50), we obtain

$$\begin{aligned} \mathbb{E}\left[z_i'\right] &= \mathbb{E}\left[\mu_i + \alpha\left(z_i - \mu_i\right) + \sigma_i\left(1 - \alpha^2\right)^{1/2}\nu\right] \\ &= \mu_i + \alpha\left(\mathbb{E}\left[z_i\right] - \mu_i\right) + \sigma_i\left(1 - \alpha^2\right)^{1/2}\mathbb{E}\left[\nu\right] \\ &= \mu_i. \end{aligned}$$

Now we can use this together with (1.40) to compute the variance of z_i',

$$\begin{aligned}\text{var}\left[z_i'\right] &= \mathbb{E}\left[\left(z_i'\right)^2\right] - \mathbb{E}\left[z_i'\right]^2 \\ &= \mathbb{E}\left[\left(\mu_i + \alpha\left(z_i - \mu_i\right) + \sigma_i\left(1-\alpha^2\right)^{1/2}\nu\right)^2\right] - \mu_i^2 \\ &= \alpha^2\mathbb{E}\left[\left(z_i-\mu_i\right)^2\right] + \sigma_i^2\left(1-\alpha^2\right)\mathbb{E}\left[\nu^2\right] \\ &= \sigma_i^2\end{aligned}$$

where we have used the first and second order moments of z_i and ν.

11.15 Using (11.56), we can differentiate (11.57), yielding

$$\frac{\partial H}{\partial r_i} = \frac{\partial K}{\partial r_i} = r_i$$

and thus (11.53) and (11.58) are equivalent.

Similarly, differentiating (11.57) w.r.t. z_i we get

$$\frac{\partial H}{\partial z_i} = \frac{\partial E}{\partial r_i},$$

and from this, it is immediately clear that (11.55) and (11.59) are equivalent.

11.16 From the product rule we know that

$$p(\mathbf{r}|\mathbf{z}) \propto p(\mathbf{r},\mathbf{z}).$$

Using (11.56) and (11.57) to rewrite (11.63) as

$$\begin{aligned}p(\mathbf{z},\mathbf{r}) &= \frac{1}{Z_H}\exp\left(-H(\mathbf{z},\mathbf{r})\right) \\ &= \frac{1}{Z_H}\exp\left(-E(\mathbf{z}) - K(\mathbf{r})\right) \\ &= \frac{1}{Z_H}\exp\left(-\frac{1}{2}\|\mathbf{r}\|^2\right)\exp\left(-E(\mathbf{z})\right).\end{aligned}$$

Thus we see that $p(\mathbf{z},\mathbf{r})$ is Gaussian w.r.t. $\mathbf{r}$ and hence $p(\mathbf{r}|\mathbf{z})$ will be Gaussian too.

11.17 **NOTE**: In PRML, there are sign errors in equations (11.68) and (11.69). In both cases, the sign of the argument to the exponential forming the second argument to the min-function should be changed.

First we note that, if $H(\mathcal{R}) = H(\mathcal{R}')$, then the detailed balance clearly holds, since in this case, (11.68) and (11.69) are identical.

Otherwise, we either have $H(\mathcal{R}) > H(\mathcal{R}')$ or $H(\mathcal{R}) < H(\mathcal{R}')$. We consider the former case, for which (11.68) becomes

$$\frac{1}{Z_H}\exp(-H(\mathcal{R}))\delta V\frac{1}{2},$$