# Chapter 1 Pattern Recognition

**1.1** Substituting (1.1) into (1.2) and then differentiating with respect to $w_i$ we obtain

$$\sum_{n=1}^{N}\left(\sum_{j=0}^{M} w_j x_n^j - t_n\right) x_n^i = 0. \tag{1}$$

Re-arranging terms then gives the required result.

**1.2** For the regularized sum-of-squares error function given by (1.4) the corresponding linear equations are again obtained by differentiation, and take the same form as (1.122), but with $A_{ij}$ replaced by $\widetilde{A}_{ij}$, given by

$$\widetilde{A}_{ij} = A_{ij} + \lambda I_{ij}. \tag{2}$$

**1.3** Let us denote apples, oranges and limes by $a$, $o$ and $l$ respectively. The marginal probability of selecting an apple is given by

$$\begin{aligned} p(a) &= p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g) \\ &= \frac{3}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.34 \end{aligned} \tag{3}$$

where the conditional probabilities are obtained from the proportions of apples in each box.

To find the probability that the box was green, given that the fruit we selected was an orange, we can use Bayes' theorem

$$p(g|o) = \frac{p(o|g)p(g)}{p(o)}. \tag{4}$$

The denominator in (4) is given by

$$\begin{aligned} p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g) \\ &= \frac{4}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.36 \end{aligned} \tag{5}$$

from which we obtain

$$p(g|o) = \frac{3}{10} \times \frac{0.6}{0.36} = \frac{1}{2}. \tag{6}$$

**1.4** We are often interested in finding the most probable value for some quantity. In the case of probability distributions over discrete variables this poses little problem. However, for continuous variables there is a subtlety arising from the nature of probability densities and the way they transform under non-linear changes of variable.

Consider first the way a function $f(x)$ behaves when we change to a new variable $y$ where the two variables are related by $x = g(y)$. This defines a new function of $y$ given by

$$\widetilde{f}(y) = f(g(y)). \tag{7}$$

Suppose $f(x)$ has a mode (i.e. a maximum) at $\widehat{x}$ so that $f'(\widehat{x}) = 0$. The corresponding mode of $\widetilde{f}(y)$ will occur for a value $\widehat{y}$ obtained by differentiating both sides of (7) with respect to $y$

$$\widetilde{f}'(\widehat{y}) = f'(g(\widehat{y}))g'(\widehat{y}) = 0. \tag{8}$$

Assuming $g'(\widehat{y}) \neq 0$ at the mode, then $f'(g(\widehat{y})) = 0$. However, we know that $f'(\widehat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables $x$ and $y$ are related by $\widehat{x} = g(\widehat{y})$, as one would expect. Thus, finding a mode with respect to the variable $x$ is completely equivalent to first transforming to the variable $y$, then finding a mode with respect to $y$, and then transforming back to $x$.

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by ((1.27)). Let us write $g'(y) = s|g'(y)|$ where $s \in \{-1, +1\}$. Then ((1.27)) can be written

$$p_y(y) = p_x(g(y))sg'(y).$$

Differentiating both sides with respect to $y$ then gives

$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y). \tag{9}$$

Due to the presence of the second term on the right hand side of (9) the relationship $\widehat{x} = g(\widehat{y})$ no longer holds. Thus the value of $x$ obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to $y$ and then transforming back to $x$. This causes modes of densities to be dependent on the choice of variables. In the case of linear transformation, the second term on the right hand side of (9) vanishes, and so the location of the maximum transforms according to $\widehat{x} = g(\widehat{y})$.

This effect can be illustrated with a simple example, as shown in Figure 1. We begin by considering a Gaussian distribution $p_x(x)$ over $x$ with mean $\mu = 6$ and standard deviation $\sigma = 1$, shown by the red curve in Figure 1. Next we draw a sample of $N = 50,000$ points from this distribution and plot a histogram of their values, which as expected agrees with the distribution $p_x(x)$.

Now consider a non-linear change of variables from $x$ to $y$ given by

$$x = g(y) = \ln(y) - \ln(1 - y) + 5. \tag{10}$$

The inverse of this function is given by

$$y = g^{-1}(x) = \frac{1}{1 + \exp(-x + 5)} \tag{11}$$

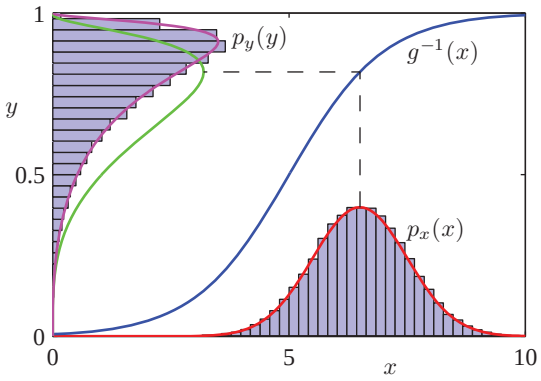


**Figure 1** Example of the transformation of the mode of a density under a non-linear change of variables, illustrating the different behaviour compared to a simple function. See the text for details.

which is a *logistic sigmoid* function, and is shown in Figure 1 by the blue curve.

If we simply transform $p_x(x)$ as a function of $x$ we obtain the green curve $p_x(g(y))$ shown in Figure 1, and we see that the mode of the density $p_x(x)$ is transformed via the sigmoid function to the mode of this curve. However, the density over $y$ transforms instead according to (1.27) and is shown by the magenta curve on the left side of the diagram. Note that this has its mode shifted relative to the mode of the green curve.

To confirm this result we take our sample of $50,000$ values of $x$, evaluate the corresponding values of $y$ using (11), and then plot a histogram of their values. We see that this histogram matches the magenta curve in Figure 1 and not the green curve!

**1.5** Expanding the square we have

$$\begin{aligned}\mathbb{E}[(f(x)-\mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\end{aligned}$$

as required.

**1.6** The definition of covariance is given by (1.41) as

$$\mathrm{cov}[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

Using (1.33) and the fact that $p(x,y) = p(x)p(y)$ when $x$ and $y$ are independent, we obtain

$$\begin{aligned}\mathbb{E}[xy] &= \sum_x \sum_y p(x,y)xy \\ &= \sum_x p(x)x \sum_y p(y)y \\ &= \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

and hence $\mathrm{cov}[x, y] = 0$. The case where $x$ and $y$ are continuous variables is analogous, with (1.33) replaced by (1.34) and the sums replaced by integrals.

**1.7** The transformation from Cartesian to polar coordinates is defined by

$$x = r\cos\theta \tag{12}$$
$$y = r\sin\theta \tag{13}$$

and hence we have $x^2 + y^2 = r^2$ where we have used the well-known trigonometric result (2.177). Also the Jacobian of the change of variables is easily seen to be

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

where again we have used (2.177). Thus the double integral in (1.125) becomes

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r\,\mathrm{d}r\,\mathrm{d}\theta \tag{14}$$
$$= 2\pi\int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}\,\mathrm{d}u \tag{15}$$
$$= \pi\left[\exp\left(-\frac{u}{2\sigma^2}\right)\left(-2\sigma^2\right)\right]_0^{\infty} \tag{16}$$
$$= 2\pi\sigma^2 \tag{17}$$

where we have used the change of variables $r^2 = u$. Thus

$$I = \left(2\pi\sigma^2\right)^{1/2}.$$

Finally, using the transformation $y = x - \mu$, the integral of the Gaussian distribution becomes

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{2\sigma^2}\right)\,\mathrm{d}y$$
$$= \frac{I}{\left(2\pi\sigma^2\right)^{1/2}} = 1$$

as required.

**1.8** From the definition (1.46) of the univariate Gaussian distribution, we have

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x\,\mathrm{d}x. \tag{18}$$

Now change variables using $y = x - \mu$ to give

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} (y + \mu)\, \mathrm{d}y. \tag{19}$$

We now note that in the factor $(y + \mu)$ the first term in $y$ corresponds to an odd integrand and so this integral must vanish (to show this explicitly, write the integral as the sum of two integrals, one from $-\infty$ to 0 and the other from 0 to $\infty$ and then show that these two integrals cancel). In the second term, $\mu$ is a constant and pulls outside the integral, leaving a normalized Gaussian distribution which integrates to 1, and so we obtain (1.49).

To derive (1.50) we first substitute the expression (1.46) for the normal distribution into the normalization result (1.48) and re-arrange to obtain

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \mathrm{d}x = \left(2\pi\sigma^2\right)^{1/2}. \tag{20}$$

We now differentiate both sides of (20) with respect to $\sigma^2$ and then re-arrange to obtain

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} (x - \mu)^2\, \mathrm{d}x = \sigma^2 \tag{21}$$

which directly shows that

$$\mathbb{E}[(x - \mu)^2] = \mathrm{var}[x] = \sigma^2. \tag{22}$$

Now we expand the square on the left-hand side giving

$$\mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \sigma^2.$$

Making use of (1.49) then gives (1.50) as required.

Finally, (1.51) follows directly from (1.49) and (1.50)

$$\mathbb{E}[x^2] - \mathbb{E}[x]^2 = \left(\mu^2 + \sigma^2\right) - \mu^2 = \sigma^2.$$

**1.9** For the univariate case, we simply differentiate (1.46) with respect to $x$ to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathcal{N}\left(x|\mu, \sigma^2\right) = -\mathcal{N}\left(x|\mu, \sigma^2\right) \frac{x - \mu}{\sigma^2}.$$

Setting this to zero we obtain $x = \mu$.

Similarly, for the multivariate case we differentiate (1.52) with respect to $\mathbf{x}$ to obtain

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\nabla_{\mathbf{x}}\left\{(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \\ &= -\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), \end{aligned}$$

where we have used (C.19), (C.20)[1] and the fact that $\boldsymbol{\Sigma}^{-1}$ is symmetric. Setting this derivative equal to $\mathbf{0}$, and left-multiplying by $\boldsymbol{\Sigma}$, leads to the solution $\mathbf{x} = \boldsymbol{\mu}$.

[1] **NOTE**: In PRML, there are mistakes in (C.20); all instances of $\mathbf{x}$ (vector) in the denominators should be $x$ (scalar).

**1.10** Since $x$ and $z$ are independent, their joint distribution factorizes $p(x, z) = p(x)p(z)$, and so

$$\begin{aligned}\mathbb{E}[x+z] &= \iint (x+z)p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z &(23)\\ &= \int xp(x)\,\mathrm{d}x + \int zp(z)\,\mathrm{d}z &(24)\\ &= \mathbb{E}[x] + \mathbb{E}[z]. &(25)\end{aligned}$$

Similarly for the variances, we first note that

$$(x+z-\mathbb{E}[x+z])^2 = (x-\mathbb{E}[x])^2 + (z-\mathbb{E}[z])^2 + 2(x-\mathbb{E}[x])(z-\mathbb{E}[z]) \quad (26)$$

where the final term will integrate to zero with respect to the factorized distribution $p(x)p(z)$. Hence

$$\begin{aligned}\mathrm{var}[x+z] &= \iint (x+z-\mathbb{E}[x+z])^2 p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z\\ &= \int (x-\mathbb{E}[x])^2 p(x)\,\mathrm{d}x + \int (z-\mathbb{E}[z])^2 p(z)\,\mathrm{d}z\\ &= \mathrm{var}(x) + \mathrm{var}(z). &(27)\end{aligned}$$

For discrete variables the integrals are replaced by summations, and the same results are again obtained.

**1.11** We use $\ell$ to denote $\ln p(\mathbf{X}|\mu, \sigma^2)$ from (1.54). By standard rules of differentiation we obtain

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu).$$

Setting this equal to zero and moving the terms involving $\mu$ to the other side of the equation we get

$$\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n = \frac{1}{\sigma^2} N\mu$$

and by multiplying ing both sides by $\sigma^2/N$ we get (1.55).

Similarly we have

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\frac{1}{\sigma^2}$$

and setting this to zero we obtain

$$\frac{N}{2}\frac{1}{\sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n - \mu)^2.$$

Multiplying both sides by $2(\sigma^2)^2/N$ and substituting $\mu_{\mathrm{ML}}$ for $\mu$ we get (1.56).

**1.12** If $m = n$ then $x_n x_m = x_n^2$ and using (1.50) we obtain $\mathbb{E}[x_n^2] = \mu^2 + \sigma^2$, whereas if $n \neq m$ then the two data points $x_n$ and $x_m$ are independent and hence $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2$ where we have used (1.49). Combining these two results we obtain (1.130).

Next we have

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \mu \tag{28}$$

using (1.49).

Finally, consider $\mathbb{E}[\sigma^2_{\mathrm{ML}}]$. From (1.55) and (1.56), and making use of (1.130), we have

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \frac{1}{N}\sum_{m=1}^{N}x_m\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right] \\
&= \left\{\mu^2 + \sigma^2 - 2\left(\mu^2 + \frac{1}{N}\sigma^2\right) + \mu^2 + \frac{1}{N}\sigma^2\right\} \\
&= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned} \tag{29}$$

as required.

**1.13** In a similar fashion to solution 1.12, substituting $\mu$ for $\mu_{\mathrm{ML}}$ in (1.56) and using (1.49) and (1.50) we have

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

**1.14** Define

$$w_{ij}^{\mathrm{S}} = \frac{1}{2}(w_{ij} + w_{ji}) \qquad w_{ij}^{\mathrm{A}} = \frac{1}{2}(w_{ij} - w_{ji}). \tag{30}$$

from which the (anti)symmetry properties follow directly, as does the relation $w_{ij} = w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}$. We now note that

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_i x_j = \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}x_i x_j - \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ji}x_i x_j = 0 \tag{31}$$

from which we obtain (1.132). The number of independent components in $w_{ij}^{\mathrm{S}}$ can be found by noting that there are $D^2$ parameters in total in this matrix, and that entries off the leading diagonal occur in constrained pairs $w_{ij} = w_{ji}$ for $j \neq i$. Thus we start with $D^2$ parameters in the matrix $w_{ij}^{\mathrm{S}}$, subtract $D$ for the number of parameters on the leading diagonal, divide by two, and then add back $D$ for the leading diagonal and we obtain $(D^2 - D)/2 + D = D(D+1)/2$.

**1.15** The redundancy in the coefficients in (1.133) arises from interchange symmetries between the indices $i_k$. Such symmetries can therefore be removed by enforcing an ordering on the indices, as in (1.134), so that only one member in each group of equivalent configurations occurs in the summation.

To derive (1.135) we note that the number of independent parameters $n(D, M)$ which appear at order $M$ can be written as

$$n(D, M) = \sum_{i_1=1}^{D} \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \tag{32}$$

which has $M$ terms. This can clearly also be written as

$$n(D, M) = \sum_{i_1=1}^{D} \left\{ \sum_{i_2=1}^{i_1} \cdots \sum_{i_M=1}^{i_{M-1}} 1 \right\} \tag{33}$$

where the term in braces has $M-1$ terms which, from (32), must equal $n(i_1, M-1)$. Thus we can write

$$n(D, M) = \sum_{i_1=1}^{D} n(i_1, M-1) \tag{34}$$

which is equivalent to (1.135).

To prove (1.136) we first set $D = 1$ on both sides of the equation, and make use of $0! = 1$, which gives the value 1 on both sides, thus showing the equation is valid for $D = 1$. Now we assume that it is true for a specific value of dimensionality $D$ and then show that it must be true for dimensionality $D+1$. Thus consider the left-hand side of (1.136) evaluated for $D+1$ which gives

$$\begin{aligned} \sum_{i=1}^{D+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} &= \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!} \\ &= \frac{(D+M-1)!D + (D+M-1)!M}{D!M!} \\ &= \frac{(D+M)!}{D!M!} \end{aligned} \tag{35}$$

which equals the right hand side of (1.136) for dimensionality $D+1$. Thus, by induction, (1.136) must hold true for all values of $D$.

Finally we use induction to prove (1.137). For $M = 2$ we find obtain the standard result $n(D, 2) = \frac{1}{2}D(D + 1)$, which is also proved in Exercise 1.14. Now assume that (1.137) is correct for a specific order $M - 1$ so that

$$n(D, M - 1) = \frac{(D + M - 2)!}{(D - 1)!\,(M - 1)!}. \tag{36}$$

Substituting this into the right hand side of (1.135) we obtain

$$n(D, M) = \sum_{i=1}^{D} \frac{(i + M - 2)!}{(i - 1)!\,(M - 1)!} \tag{37}$$

which, making use of (1.136), gives

$$n(D, M) = \frac{(D + M - 1)!}{(D - 1)!\,M!} \tag{38}$$

and hence shows that (1.137) is true for polynomials of order $M$. Thus by induction (1.137) must be true for all values of $M$.

**1.16** **NOTE**: In PRML, this exercise contains two typographical errors. On line 4, $M6th$ should be $M^{\text{th}}$ and on the l.h.s. of (1.139), $N(d, M)$ should be $N(D, M)$.

The result (1.138) follows simply from summing up the coefficients at all order up to and including order $M$. To prove (1.139), we first note that when $M = 0$ the right hand side of (1.139) equals 1, which we know to be correct since this is the number of parameters at zeroth order which is just the constant offset in the polynomial. Assuming that (1.139) is correct at order $M$, we obtain the following result at order $M + 1$

$$\begin{aligned} N(D, M + 1) &= \sum_{m=0}^{M+1} n(D, m) \\ &= \sum_{m=0}^{M} n(D, m) + n(D, M + 1) \\ &= \frac{(D + M)!}{D!M!} + \frac{(D + M)!}{(D - 1)!(M + 1)!} \\ &= \frac{(D + M)!(M + 1) + (D + M)!D}{D!(M + 1)!} \\ &= \frac{(D + M + 1)!}{D!(M + 1)!} \end{aligned}$$

which is the required result at order $M + 1$.

Now assume $M \gg D$. Using Stirling's formula we have

$$\begin{aligned} n(D,M) &\simeq \frac{(D+M)^{D+M}e^{-D-M}}{D!\,M^M e^{-M}} \\ &= \frac{M^{D+M}e^{-D}}{D!\,M^M}\left(1+\frac{D}{M}\right)^{D+M} \\ &\simeq \frac{M^D e^{-D}}{D!}\left(1+\frac{D(D+M)}{M}\right) \\ &\simeq \frac{(1+D)e^{-D}}{D!}M^D \end{aligned}$$

which grows like $M^D$ with $M$. The case where $D \gg M$ is identical, with the roles of $D$ and $M$ exchanged. By numerical evaluation we obtain $N(10,3) = 286$ and $N(100,3) = 176{,}851$.

**1.17** Using integration by parts we have

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty u^x e^{-u}\,\mathrm{d}u \\ &= \left[-e^{-u}u^x\right]_0^\infty + \int_0^\infty xu^{x-1}e^{-u}\,\mathrm{d}u = 0 + x\Gamma(x). \end{aligned} \tag{39}$$

For $x = 1$ we have

$$\Gamma(1) = \int_0^\infty e^{-u}\,\mathrm{d}u = \left[-e^{-u}\right]_0^\infty = 1. \tag{40}$$

If $x$ is an integer we can apply proof by induction to relate the gamma function to the factorial function. Suppose that $\Gamma(x+1) = x!$ holds. Then from the result (39) we have $\Gamma(x+2) = (x+1)\Gamma(x+1) = (x+1)!$. Finally, $\Gamma(1) = 1 = 0!$, which completes the proof by induction.

**1.18** On the right-hand side of (1.142) we make the change of variables $u = r^2$ to give

$$\frac{1}{2}S_D\int_0^\infty e^{-u}u^{D/2-1}\,\mathrm{d}u = \frac{1}{2}S_D\Gamma(D/2) \tag{41}$$

where we have used the definition (1.141) of the Gamma function. On the left hand side of (1.142) we can use (1.126) to obtain $\pi^{D/2}$. Equating these we obtain the desired result (1.143).

The volume of a sphere of radius 1 in $D$-dimensions is obtained by integration

$$V_D = S_D\int_0^1 r^{D-1}\,\mathrm{d}r = \frac{S_D}{D}. \tag{42}$$

For $D = 2$ and $D = 3$ we obtain the following results

$$S_2 = 2\pi, \qquad S_3 = 4\pi, \qquad V_2 = \pi a^2, \qquad V_3 = \frac{4}{3}\pi a^3. \tag{43}$$

**1.19** The volume of the cube is $(2a)^D$. Combining this with (1.143) and (1.144) we obtain (1.145). Using Stirling's formula (1.146) in (1.145) the ratio becomes, for large $D$,

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \left(\frac{\pi e}{2D}\right)^{D/2} \frac{1}{D} \tag{44}$$

which goes to 0 as $D \to \infty$. The distance from the center of the cube to the mid point of one of the sides is $a$, since this is where it makes contact with the sphere. Similarly the distance to one of the corners is $a\sqrt{D}$ from Pythagoras' theorem. Thus the ratio is $\sqrt{D}$.

**1.20** Since $p(\mathbf{x})$ is radially symmetric it will be roughly constant over the shell of radius $r$ and thickness $\epsilon$. This shell has volume $S_D r^{D-1}\epsilon$ and since $\|\mathbf{x}\|^2 = r^2$ we have

$$\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \simeq p(r) S_D r^{D-1} \epsilon \tag{45}$$

from which we obtain (1.148). We can find the stationary points of $p(r)$ by differentiation

$$\frac{\mathrm{d}}{\mathrm{d}r} p(r) \propto \left[(D-1)r^{D-2} + r^{D-1}\left(-\frac{r}{\sigma^2}\right)\right] \exp\left(-\frac{r^2}{2\sigma^2}\right) = 0. \tag{46}$$

Solving for $r$, and using $D \gg 1$, we obtain $\widehat{r} \simeq \sqrt{D}\sigma$.

Next we note that

$$\begin{aligned} p(\widehat{r}+\epsilon) &\propto (\widehat{r}+\epsilon)^{D-1} \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2} + (D-1)\ln(\widehat{r}+\epsilon)\right]. \end{aligned} \tag{47}$$

We now expand $p(r)$ around the point $\widehat{r}$. Since this is a stationary point of $p(r)$ we must keep terms up to second order. Making use of the expansion $\ln(1+x) = x - x^2/2 + O(x^3)$, together with $D \gg 1$, we obtain (1.149).

Finally, from (1.147) we see that the probability density at the origin is given by

$$p(\mathbf{x} = \mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{1/2}}$$

while the density at $\|\mathbf{x}\| = \widehat{r}$ is given from (1.147) by

$$p(\|\mathbf{x}\| = \widehat{r}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\widehat{r}^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{D}{2}\right)$$

where we have used $\widehat{r} \simeq \sqrt{D}\sigma$. Thus the ratio of densities is given by $\exp(D/2)$.

**1.21** Since the square root function is monotonic for non-negative numbers, we can take the square root of the relation $a \leqslant b$ to obtain $a^{1/2} \leqslant b^{1/2}$. Then we multiply both sides by the non-negative quantity $a^{1/2}$ to obtain $a \leqslant (ab)^{1/2}$.

The probability of a misclassification is given, from (1.78), by

$$\begin{aligned} p(\text{mistake}) &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_1} p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{48}$$

Since we have chosen the decision regions to minimize the probability of misclassification we must have $p(\mathcal{C}_2|\mathbf{x}) \leqslant p(\mathcal{C}_1|\mathbf{x})$ in region $\mathcal{R}_1$, and $p(\mathcal{C}_1|\mathbf{x}) \leqslant p(\mathcal{C}_2|\mathbf{x})$ in region $\mathcal{R}_2$. We now apply the result $a \leqslant b \Rightarrow a^{1/2} \leqslant b^{1/2}$ to give

$$\begin{aligned} p(\text{mistake}) &\leqslant \int_{\mathcal{R}_1} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad + \int_{\mathcal{R}_2} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \{p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\}^{1/2}\,\mathrm{d}\mathbf{x} \end{aligned} \tag{49}$$

since the two integrals have the same integrand. The final integral is taken over the whole of the domain of $\mathbf{x}$.

**1.22** Substituting $L_{kj} = 1 - \delta_{kj}$ into (1.81), and using the fact that the posterior probabilities sum to one, we find that, for each $\mathbf{x}$ we should choose the class $j$ for which $1 - p(\mathcal{C}_j|\mathbf{x})$ is a minimum, which is equivalent to choosing the $j$ for which the posterior probability $p(\mathcal{C}_j|\mathbf{x})$ is a maximum. This loss matrix assigns a loss of one if the example is misclassified, and a loss of zero if it is correctly classified, and hence minimizing the expected loss will minimize the misclassification rate.

**1.23** From (1.81) we see that for a general loss matrix and arbitrary class priors, the expected loss is minimized by assigning an input $\mathbf{x}$ to class the $j$ which minimizes

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k) p(\mathcal{C}_k)$$

and so there is a direct trade-off between the priors $p(\mathcal{C}_k)$ and the loss matrix $L_{kj}$.

**1.24** A vector $\mathbf{x}$ belongs to class $\mathcal{C}_k$ with probability $p(\mathcal{C}_k|\mathbf{x})$. If we decide to assign $\mathbf{x}$ to class $\mathcal{C}_j$ we will incur an expected loss of $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$, whereas if we select the reject option we will incur a loss of $\lambda$. Thus, if

$$j = \arg\min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) \tag{50}$$

then we minimize the expected loss if we take the following action

$$\text{choose} \begin{cases} \text{class } j, & \text{if } \min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) < \lambda; \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{51}$$

For a loss matrix $L_{kj} = 1 - I_{kj}$ we have $\sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_l|\mathbf{x})$ and so we reject unless the smallest value of $1 - p(\mathcal{C}_l|\mathbf{x})$ is less than $\lambda$, or equivalently if the largest value of $p(\mathcal{C}_l|\mathbf{x})$ is less than $1 - \lambda$. In the standard reject criterion we reject if the largest posterior probability is less than $\theta$. Thus these two criteria for rejection are equivalent provided $\theta = 1 - \lambda$.

**1.25** The expected squared loss for a vectorial target variable is given by

$$\mathbb{E}[L] = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{t}.$$

Our goal is to choose $\mathbf{y}(\mathbf{x})$ so as to minimize $\mathbb{E}[L]$. We can do this formally using the calculus of variations to give

$$\frac{\delta \mathbb{E}[L]}{\delta \mathbf{y}(\mathbf{x})} = \int 2(\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{t} = 0.$$

Solving for $\mathbf{y}(\mathbf{x})$, and using the sum and product rules of probability, we obtain

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t} p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{t}}{\int p(\mathbf{t}, \mathbf{x}) \, \mathrm{d}\mathbf{t}} = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) \, \mathrm{d}\mathbf{t}$$

which is the conditional average of $\mathbf{t}$ conditioned on $\mathbf{x}$. For the case of a scalar target variable we have

$$y(\mathbf{x}) = \int t p(t|\mathbf{x}) \, \mathrm{d}t$$

which is equivalent to (1.89).

**1.26** **NOTE**: In PRML, there is an error in equation (1.90); the integrand of the second integral should be replaced by $\mathrm{var}[t|\mathbf{x}]p(\mathbf{x})$.

We start by expanding the square in (1.151), in a similar fashion to the univariate case in the equation preceding (1.90),

$$\begin{aligned} \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}\|^2 \\ &= \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 + (\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}])^{\mathrm{T}}(\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}) \\ &\quad + (\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t})^{\mathrm{T}}(\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]) + \|\mathbb{E}[\mathbf{t}|\mathbf{x}] - \mathbf{t}\|^2. \end{aligned}$$

Following the treatment of the univariate case, we now substitute this into (1.151) and perform the integral over $\mathbf{t}$. Again the cross-term vanishes and we are left with

$$\mathbb{E}[L] = \int \|\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]\|^2 p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \mathrm{var}[\mathbf{t}|\mathbf{x}] p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

from which we see directly that the function $\mathbf{y}(\mathbf{x})$ that minimizes $\mathbb{E}[L]$ is given by $\mathbb{E}[\mathbf{t}|\mathbf{x}]$.

**1.27** Since we can choose $y(\mathbf{x})$ independently for each value of $\mathbf{x}$, the minimum of the expected $L_q$ loss can be found by minimizing the integrand given by

$$\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\,\mathrm{d}t \tag{52}$$

for each value of $\mathbf{x}$. Setting the derivative of (52) with respect to $y(\mathbf{x})$ to zero gives the stationarity condition

$$\begin{aligned}&\int q|y(\mathbf{x}) - t|^{q-1}\mathrm{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})\,\mathrm{d}t \\ &\quad = \quad q\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t - q\int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t = 0\end{aligned}$$

which can also be obtained directly by setting the functional derivative of (1.91) with respect to $y(\mathbf{x})$ equal to zero. It follows that $y(\mathbf{x})$ must satisfy

$$\int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t. \tag{53}$$

For the case of $q = 1$ this reduces to

$$\int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})\,\mathrm{d}t. \tag{54}$$

which says that $y(\mathbf{x})$ must be the conditional median of $t$.

For $q \to 0$ we note that, as a function of $t$, the quantity $|y(\mathbf{x}) - t|^q$ is close to 1 everywhere except in a small neighbourhood around $t = y(\mathbf{x})$ where it falls to zero. The value of (52) will therefore be close to 1, since the density $p(t)$ is normalized, but reduced slightly by the 'notch' close to $t = y(\mathbf{x})$. We obtain the biggest reduction in (52) by choosing the location of the notch to coincide with the largest value of $p(t)$, i.e. with the (conditional) mode.

**1.28** From the discussion of the introduction of Section 1.6, we have

$$h(p^2) = h(p) + h(p) = 2\,h(p).$$

We then assume that for all $k \leqslant K$, $h(p^k) = k\,h(p)$. For $k = K + 1$ we have

$$h(p^{K+1}) = h(p^K p) = h(p^K) + h(p) = K\,h(p) + h(p) = (K + 1)\,h(p).$$

Moreover,

$$h(p^{n/m}) = n\,h(p^{1/m}) = \frac{n}{m}m\,h(p^{1/m}) = \frac{n}{m}\,h(p^{m/m}) = \frac{n}{m}\,h(p)$$

and so, by continuity, we have that $h(p^x) = x\,h(p)$ for any real number $x$.

Now consider the positive real numbers $p$ and $q$ and the real number $x$ such that $p = q^x$. From the above discussion, we see that

$$\frac{h(p)}{\ln(p)} = \frac{h(q^x)}{\ln(q^x)} = \frac{x\,h(q)}{x\,\ln(q)} = \frac{h(q)}{\ln(q)}$$

and hence $h(p) \propto \ln(p)$.

**1.29** The entropy of an $M$-state discrete variable $x$ can be written in the form

$$H(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i) = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)}. \tag{55}$$

The function $\ln(x)$ is concave$\frown$ and so we can apply Jensen's inequality in the form (1.115) but with the inequality reversed, so that

$$H(x) \leqslant \ln\left(\sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)}\right) = \ln M. \tag{56}$$

**1.30** **NOTE**: In PRML, there is a minus sign ('$-$') missing on the l.h.s. of (1.103).

From (1.113) we have

$$\mathrm{KL}(p\|q) = -\int p(x) \ln q(x)\,\mathrm{d}x + \int p(x) \ln p(x)\,\mathrm{d}x. \tag{57}$$

Using (1.46) and (1.48)– (1.50), we can rewrite the first integral on the r.h.s. of (57) as

$$\begin{aligned} -\int p(x) \ln q(x)\,\mathrm{d}x &= \int \mathcal{N}(x|\mu,\sigma^2) \frac{1}{2}\left(\ln(2\pi s^2) + \frac{(x-m)^2}{s^2}\right) \mathrm{d}x \\ &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{1}{s^2}\int \mathcal{N}(x|\mu,\sigma^2)(x^2 - 2xm + m^2)\,\mathrm{d}x\right) \\ &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2}\right). \end{aligned} \tag{58}$$

The second integral on the r.h.s. of (57) we recognize from (1.103) as the negative differential entropy of a Gaussian. Thus, from (57), (58) and (1.110), we have

$$\begin{aligned} \mathrm{KL}(p\|q) &= \frac{1}{2}\left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1 - \ln(2\pi\sigma^2)\right) \\ &= \frac{1}{2}\left(\ln\left(\frac{s^2}{\sigma^2}\right) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1\right). \end{aligned}$$

**1.31** We first make use of the relation $\mathrm{I}(\mathbf{x};\mathbf{y}) = \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}|\mathbf{x})$ which we obtained in (1.121), and note that the mutual information satisfies $\mathrm{I}(\mathbf{x};\mathbf{y}) \geqslant 0$ since it is a form of Kullback-Leibler divergence. Finally we make use of the relation (1.112) to obtain the desired result (1.152).

To show that statistical independence is a sufficient condition for the equality to be satisfied, we substitute $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ into the definition of the entropy, giving

$$\begin{aligned}\mathrm{H}(\mathbf{x},\mathbf{y}) &= \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= \iint p(\mathbf{x})p(\mathbf{y}) \left\{\ln p(\mathbf{x}) + \ln p(\mathbf{y})\right\}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= \int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int p(\mathbf{y}) \ln p(\mathbf{y})\,\mathrm{d}\mathbf{y} \\ &= \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}).\end{aligned}$$

To show that statistical independence is a necessary condition, we combine the equality condition

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y})$$

with the result (1.112) to give

$$\mathrm{H}(\mathbf{y}|\mathbf{x}) = \mathrm{H}(\mathbf{y}).$$

We now note that the right-hand side is independent of $\mathbf{x}$ and hence the left-hand side must also be constant with respect to $\mathbf{x}$. Using (1.121) it then follows that the mutual information $\mathrm{I}[\mathbf{x},\mathbf{y}] = 0$. Finally, using (1.120) we see that the mutual information is a form of KL divergence, and this vanishes only if the two distributions are equal, so that $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ as required.

**1.32** When we make a change of variables, the probability density is transformed by the Jacobian of the change of variables. Thus we have

$$p(\mathbf{x}) = p(\mathbf{y}) \left|\frac{\partial y_i}{\partial x_j}\right| = p(\mathbf{y})|\mathbf{A}| \tag{59}$$

where $|\cdot|$ denotes the determinant. Then the entropy of $\mathbf{y}$ can be written

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \ln p(\mathbf{y})\,\mathrm{d}\mathbf{y} = -\int p(\mathbf{x}) \ln \left\{p(\mathbf{x})|\mathbf{A}|^{-1}\right\}\,\mathrm{d}\mathbf{x} = H(\mathbf{x}) + \ln|\mathbf{A}| \tag{60}$$

as required.

**1.33** The conditional entropy $H(y|x)$ can be written

$$H(y|x) = -\sum_i \sum_j p(y_i|x_j)p(x_j) \ln p(y_i|x_j) \tag{61}$$

which equals 0 by assumption. Since the quantity $-p(y_i|x_j)\ln p(y_i|x_j)$ is non-negative each of these terms must vanish for any value $x_j$ such that $p(x_j) \neq 0$. However, the quantity $p \ln p$ only vanishes for $p = 0$ or $p = 1$. Thus the quantities $p(y_i|x_j)$ are all either 0 or 1. However, they must also sum to 1, since this is a normalized probability distribution, and so precisely one of the $p(y_i|x_j)$ is 1, and the rest are 0. Thus, for each value $x_j$ there is a unique value $y_i$ with non-zero probability.

**1.34** Obtaining the required functional derivative can be done simply by inspection. However, if a more formal approach is required we can proceed as follows using the techniques set out in Appendix D. Consider first the functional

$$I[p(x)] = \int p(x)f(x)\,\mathrm{d}x.$$

Under a small variation $p(x) \to p(x) + \epsilon\eta(x)$ we have

$$I[p(x) + \epsilon\eta(x)] = \int p(x)f(x)\,\mathrm{d}x + \epsilon \int \eta(x)f(x)\,\mathrm{d}x$$

and hence from (D.3) we deduce that the functional derivative is given by

$$\frac{\delta I}{\delta p(x)} = f(x).$$

Similarly, if we define

$$J[p(x)] = \int p(x)\ln p(x)\,\mathrm{d}x$$

then under a small variation $p(x) \to p(x) + \epsilon\eta(x)$ we have

$$\begin{aligned} J[p(x) + \epsilon\eta(x)] &= \int p(x)\ln p(x)\,\mathrm{d}x \\ &\quad + \epsilon\left\{\int \eta(x)\ln p(x)\,\mathrm{d}x + \int p(x)\frac{1}{p(x)}\eta(x)\,\mathrm{d}x\right\} + O(\epsilon^2) \end{aligned}$$

and hence

$$\frac{\delta J}{\delta p(x)} = p(x) + 1.$$

Using these two results we obtain the following result for the functional derivative

$$-\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2.$$

Re-arranging then gives (1.108).

To eliminate the Lagrange multipliers we substitute (1.108) into each of the three constraints (1.105), (1.106) and (1.107) in turn. The solution is most easily obtained

by comparison with the standard form of the Gaussian, and noting that the results

$$\lambda_1 = 1 - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \tag{62}$$

$$\lambda_2 = 0 \tag{63}$$

$$\lambda_3 = \frac{1}{2\sigma^2} \tag{64}$$

do indeed satisfy the three constraints.

Note that there is a typographical error in the question, which should read "Use calculus of variations to show that the stationary point of the functional shown just before (1.108) is given by (1.108)".

For the multivariate version of this derivation, see Exercise 2.14.

**1.35** **NOTE**: In PRML, there is a minus sign ('$-$') missing on the l.h.s. of (1.103).

Substituting the right hand side of (1.109) in the argument of the logarithm on the right hand side of (1.103), we obtain

$$\begin{aligned}
\mathrm{H}[x] &= -\int p(x)\ln p(x)\,\mathrm{d}x \\
&= -\int p(x)\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\
&= \frac{1}{2}\left(\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\int p(x)(x-\mu)^2\,\mathrm{d}x\right) \\
&= \frac{1}{2}\left(\ln(2\pi\sigma^2) + 1\right),
\end{aligned}$$

where in the last step we used (1.107).

**1.36** Consider (1.114) with $\lambda = 0.5$ and $b = a + 2\epsilon$ (and hence $a = b - 2\epsilon$),

$$\begin{aligned}
0.5f(a) + 0.5f(b) &> f(0.5a + 0.5b) \\
&= 0.5f(0.5a + 0.5(a + 2\epsilon)) + 0.5f(0.5(b - 2\epsilon) + 0.5b) \\
&= 0.5f(a + \epsilon) + 0.5f(b - \epsilon)
\end{aligned}$$

We can rewrite this as

$$f(b) - f(b - \epsilon) > f(a + \epsilon) - f(a)$$

We then divide both sides by $\epsilon$ and let $\epsilon \to 0$, giving

$$f'(b) > f'(a).$$

Since this holds at all points, it follows that $f''(x) \geqslant 0$ everywhere.

To show the implication in the other direction, we make use of Taylor's theorem (with the remainder in Lagrange form), according to which there exist an $x^\star$ such that

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x^\star)(x - x_0)^2.$$

Since we assume that $f''(x) > 0$ everywhere, the third term on the r.h.s. will always be positive and therefore

$$f(x) > f(x_0) + f'(x_0)(x - x_0)$$

Now let $x_0 = \lambda a + (1 - \lambda)b$ and consider setting $x = a$, which gives

$$\begin{aligned} f(a) &> f(x_0) + f'(x_0)(a - x_0) \\ &= f(x_0) + f'(x_0)\left((1 - \lambda)(a - b)\right). \end{aligned} \tag{65}$$

Similarly, setting $x = b$ gives

$$f(b) > f(x_0) + f'(x_0)(\lambda(b - a)). \tag{66}$$

Multiplying (65) by $\lambda$ and (66) by $1 - \lambda$ and adding up the results on both sides, we obtain

$$\lambda f(a) + (1 - \lambda)f(b) > f(x_0) = f(\lambda a + (1 - \lambda)b)$$

as required.

**1.37** From (1.104), making use of (1.111), we have

$$\begin{aligned} \mathrm{H}[\mathbf{x}, \mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\right)\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\right)\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} - \int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &= \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]. \end{aligned}$$

**1.38** From (1.114) we know that the result (1.115) holds for $M = 1$. We now suppose that it holds for some general value $M$ and show that it must therefore hold for $M + 1$. Consider the left hand side of (1.115)

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) = f\left(\lambda_{M+1} x_{M+1} + \sum_{i=1}^{M} \lambda_i x_i\right) \tag{67}$$

$$= f\left(\lambda_{M+1} x_{M+1} + (1 - \lambda_{M+1}) \sum_{i=1}^{M} \eta_i x_i\right) \tag{68}$$

where we have defined

$$\eta_i = \frac{\lambda_i}{1-\lambda_{M+1}}. \tag{69}$$

We now apply (1.114) to give

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1}) f\left(\sum_{i=1}^{M}\eta_i x_i\right). \tag{70}$$

We now note that the quantities $\lambda_i$ by definition satisfy

$$\sum_{i=1}^{M+1}\lambda_i = 1 \tag{71}$$

and hence we have

$$\sum_{i=1}^{M}\lambda_i = 1-\lambda_{M+1} \tag{72}$$

Then using (69) we see that the quantities $\eta_i$ satisfy the property

$$\sum_{i=1}^{M}\eta_i = \frac{1}{1-\lambda_{M+1}}\sum_{i=1}^{M}\lambda_i = 1. \tag{73}$$

Thus we can apply the result (1.115) at order $M$ and so (70) becomes

$$f\left(\sum_{i=1}^{M+1}\lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1})\sum_{i=1}^{M}\eta_i f(x_i) = \sum_{i=1}^{M+1}\lambda_i f(x_i) \tag{74}$$

where we have made use of (69).

**1.39** From Table 1.3 we obtain the marginal probabilities by summation and the conditional probabilities by normalization, to give

| $x$ | |
|---|---|
| 0 | 2/3 |
| 1 | 1/3 |

$p(x)$

| $y$ | |
|---|---|
| 0 | 1 |
| 1/3 | 2/3 |

$p(y)$

| $x$ \ $y$ | 0 | 1 |
|---|---|---|
| 0 | 1 | 1/2 |
| 1 | 0 | 1/2 |

$p(x|y)$

| $x$ \ $y$ | 0 | 1 |
|---|---|---|
| 0 | 1/2 | 1/2 |
| 1 | 0 | 1 |

$p(y|x)$

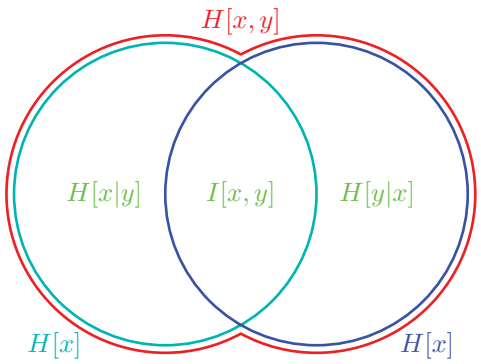


**Figure 2** Diagram showing the relationship between marginal, conditional and joint entropies and the mutual information.

From these tables, together with the definitions

$$H(x) = -\sum_i p(x_i) \ln p(x_i) \tag{75}$$

$$H(x|y) = -\sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j) \tag{76}$$

and similar definitions for $H(y)$ and $H(y|x)$, we obtain the following results

**(a)** $H(x) = \ln 3 - \frac{2}{3} \ln 2$

**(b)** $H(y) = \ln 3 - \frac{2}{3} \ln 2$

**(c)** $H(y|x) = \frac{2}{3} \ln 2$

**(d)** $H(x|y) = \frac{2}{3} \ln 2$

**(e)** $H(x, y) = \ln 3$

**(f)** $I(x; y) = \ln 3 - \frac{4}{3} \ln 2$

where we have used (1.121) to evaluate the mutual information. The corresponding diagram is shown in Figure 2.

**1.40** The arithmetic and geometric means are defined as

$$\bar{x}_{\mathrm{A}} = \frac{1}{K} \sum_k^K x_k \quad \text{and} \quad \bar{x}_{\mathrm{G}} = \left( \prod_k^K x_k \right)^{1/K},$$

respectively. Taking the logarithm of $\bar{x}_{\mathrm{A}}$ and $\bar{x}_{\mathrm{G}}$, we see that

$$\ln \bar{x}_{\mathrm{A}} = \ln \left( \frac{1}{K} \sum_k^K x_k \right) \quad \text{and} \quad \ln \bar{x}_{\mathrm{G}} = \frac{1}{K} \sum_k^K \ln x_k.$$

By matching $f$ with $\ln$ and $\lambda_i$ with $1/K$ in (1.115), taking into account that the logarithm is concave rather than convex and the inequality therefore goes the other way, we obtain the desired result.

**1.41** From the product rule we have $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, and so (1.120) can be written as

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y}) \ln p(\mathbf{y})\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{77}$$

# Chapter 2 Density Estimation

**2.1** From the definition (2.2) of the Bernoulli distribution we have

$$\begin{aligned} \sum_{x\in\{0,1\}} p(x|\mu) &= p(x=0|\mu) + p(x=1|\mu) \\ &= (1-\mu) + \mu = 1 \\ \sum_{x\in\{0,1\}} x p(x|\mu) &= 0.p(x=0|\mu) + 1.p(x=1|\mu) = \mu \\ \sum_{x\in\{0,1\}} (x-\mu)^2 p(x|\mu) &= \mu^2 p(x=0|\mu) + (1-\mu)^2 p(x=1|\mu) \\ &= \mu^2(1-\mu) + (1-\mu)^2\mu = \mu(1-\mu). \end{aligned}$$

The entropy is given by

$$\begin{aligned} \mathrm{H}[x] &= -\sum_{x\in\{0,1\}} p(x|\mu) \ln p(x|\mu) \\ &= -\sum_{x\in\{0,1\}} \mu^x (1-\mu)^{1-x} \{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= -(1-\mu)\ln(1-\mu) - \mu\ln\mu. \end{aligned}$$

**2.2** The normalization of (2.261) follows from

$$p(x=+1|\mu) + p(x=-1|\mu) = \left(\frac{1+\mu}{2}\right) + \left(\frac{1-\mu}{2}\right) = 1.$$

The mean is given by

$$\mathbb{E}[x] = \left(\frac{1+\mu}{2}\right) - \left(\frac{1-\mu}{2}\right) = \mu.$$

To evaluate the variance we use

$$\mathbb{E}[x^2] = \left(\frac{1-\mu}{2}\right) + \left(\frac{1+\mu}{2}\right) = 1$$

from which we have

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1 - \mu^2.$$

Finally the entropy is given by

$$\begin{aligned} \mathrm{H}[x] &= -\sum_{x=-1}^{x=+1} p(x|\mu) \ln p(x|\mu) \\ &= -\left(\frac{1-\mu}{2}\right) \ln\left(\frac{1-\mu}{2}\right) - \left(\frac{1+\mu}{2}\right) \ln\left(\frac{1+\mu}{2}\right). \end{aligned}$$

**2.3** Using the definition (2.10) we have

$$\begin{aligned} \binom{N}{n} + \binom{N}{n-1} &= \frac{N!}{n!(N-n)!} + \frac{N!}{(n-1)!(N+1-n)!} \\ &= \frac{(N+1-n)N! + nN!}{n!(N+1-n)!} = \frac{(N+1)!}{n!(N+1-n)!} \\ &= \binom{N+1}{n}. \end{aligned} \tag{78}$$

To prove the binomial theorem (2.263) we note that the theorem is trivially true for $N = 0$. We now assume that it holds for some general value $N$ and prove its correctness for $N + 1$, which can be done as follows

$$\begin{aligned} (1+x)^{N+1} &= (1+x) \sum_{n=0}^{N} \binom{N}{n} x^n \\ &= \sum_{n=0}^{N} \binom{N}{n} x^n + \sum_{n=1}^{N+1} \binom{N}{n-1} x^n \\ &= \binom{N}{0} x^0 + \sum_{n=1}^{N} \left\{ \binom{N}{n} + \binom{N}{n-1} \right\} x^n + \binom{N}{N} x^{N+1} \\ &= \binom{N+1}{0} x^0 + \sum_{n=1}^{N} \binom{N+1}{n} x^n + \binom{N+1}{N+1} x^{N+1} \\ &= \sum_{n=0}^{N+1} \binom{N+1}{n} x^n \end{aligned} \tag{79}$$

which completes the inductive proof. Finally, using the binomial theorem, the normalization condition (2.264) for the binomial distribution gives

$$\begin{aligned}\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n} &= (1-\mu)^N\sum_{n=0}^{N}\binom{N}{n}\left(\frac{\mu}{1-\mu}\right)^n \\ &= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N = 1 \qquad (80)\end{aligned}$$

as required.

**2.4** Differentiating (2.264) with respect to $\mu$ we obtain

$$\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left[\frac{n}{\mu}-\frac{(N-n)}{(1-\mu)}\right]=0.$$

Multiplying through by $\mu(1-\mu)$ and re-arranging we obtain (2.11).

If we differentiate (2.264) twice with respect to $\mu$ we obtain

$$\sum_{n=1}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n}\left\{\left[\frac{n}{\mu}-\frac{(N-n)}{(1-\mu)}\right]^2-\frac{n}{\mu^2}-\frac{(N-n)}{(1-\mu)^2}\right\}=0.$$

We now multiply through by $\mu^2(1-\mu)^2$ and re-arrange, making use of the result (2.11) for the mean of the binomial distribution, to obtain

$$\mathbb{E}[n^2]=N\mu(1-\mu)+N^2\mu^2.$$

Finally, we use (1.40) to obtain the result (2.12) for the variance.

**2.5** Making the change of variable $t=y+x$ in (2.266) we obtain

$$\Gamma(a)\Gamma(b)=\int_0^\infty x^{a-1}\left\{\int_x^\infty \exp(-t)(t-x)^{b-1}\,\mathrm{d}t\right\}\mathrm{d}x. \qquad (81)$$

We now exchange the order of integration, taking care over the limits of integration

$$\Gamma(a)\Gamma(b)=\int_0^\infty\int_0^t x^{a-1}\exp(-t)(t-x)^{b-1}\,\mathrm{d}x\,\mathrm{d}t. \qquad (82)$$

The change in the limits of integration in going from (81) to (82) can be understood by reference to Figure 3. Finally we change variables in the $x$ integral using $x=t\mu$ to give

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty \exp(-t)t^{a-1}t^{b-1}t\,\mathrm{d}t\int_0^1\mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu \\ &= \Gamma(a+b)\int_0^1\mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu. \qquad (83)\end{aligned}$$

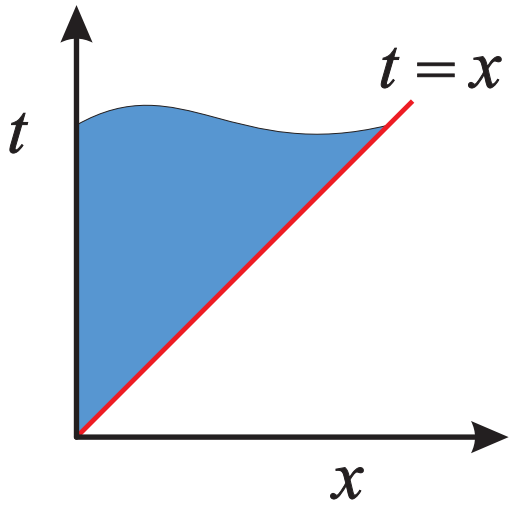


**Figure 3** Plot of the region of integration of (81) in $(x, t)$ space.

**2.6** From (2.13) the mean of the beta distribution is given by

$$\mathbb{E}[\mu] = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{(a+1)-1}(1-\mu)^{b-1}\,\mathrm{d}\mu.$$

Using the result (2.265), which follows directly from the normalization condition for the Beta distribution, we have

$$\mathbb{E}[\mu] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)} = \frac{a}{a+b}$$

where we have used the property $\Gamma(x+1) = x\Gamma(x)$. We can find the variance in the same way, by first showing that

$$\begin{aligned} \mathbb{E}[\mu^2] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)} \mu^{(a+2)-1}(1-\mu)^{b-1}\,\mathrm{d}\mu \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)} = \frac{a}{(a+b)} \frac{a+1}{(a+1+b)}. \end{aligned} \tag{84}$$

Now we use the result (1.40), together with the result (2.15) to derive the result (2.16) for $\mathrm{var}[\mu]$. Finally, we obtain the result (2.269) for the mode of the beta distribution simply by setting the derivative of the right hand side of (2.13) with respect to $\mu$ to zero and re-arranging.

**2.7** **NOTE**: In PRML, the exercise text contains a typographical error. On the third line, "mean value of $x$" should be "mean value of $\mu$".

Using the result (2.15) for the mean of a Beta distribution we see that the prior mean is $a/(a+b)$ while the posterior mean is $(a+n)/(a+b+n+m)$. The maximum likelihood estimate for $\mu$ is given by the relative frequency $n/(n+m)$ of observations

of $x = 1$. Thus the posterior mean will lie between the prior mean and the maximum likelihood solution provided the following equation is satisfied for $\lambda$ in the interval $(0, 1)$

$$\lambda \frac{a}{a+b} + (1-\lambda)\frac{n}{n+m} = \frac{a+n}{a+b+n+m}.$$

which represents a convex combination of the prior mean and the maximum likelihood estimator. This is a linear equation for $\lambda$ which is easily solved by re-arranging terms to give

$$\lambda = \frac{1}{1 + (n+m)/(a+b)}.$$

Since $a > 0$, $b > 0$, $n > 0$, and $m > 0$, it follows that the term $(n+m)/(a+b)$ lies in the range $(0, \infty)$ and hence $\lambda$ must lie in the range $(0, 1)$.

**2.8** To prove the result (2.270) we use the product rule of probability

$$\begin{aligned} \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right] &= \int \left\{ \int x p(x|y)\,\mathrm{d}x \right\} p(y)\,\mathrm{d}y \\ &= \iint x p(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int x p(x)\,\mathrm{d}x = \mathbb{E}_x[x]. \end{aligned} \tag{85}$$

For the result (2.271) for the conditional variance we make use of the result (1.40), as well as the relation (85), to give

$$\begin{aligned} \mathbb{E}_y\left[\mathrm{var}_x[x|y]\right] + \mathrm{var}_y\left[\mathbb{E}_x[x|y]\right] &= \mathbb{E}_y\left[\mathbb{E}_x[x^2|y] - \mathbb{E}_x[x|y]^2\right] \\ &\quad + \mathbb{E}_y\left[\mathbb{E}_x[x|y]^2\right] - \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right]^2 \\ &= \mathbb{E}_x[x^2] - \mathbb{E}_x[x]^2 = \mathrm{var}_x[x] \end{aligned}$$

where we have made use of $\mathbb{E}_y\left[\mathbb{E}_x[x^2|y]\right] = \mathbb{E}_x[x^2]$ which can be proved by analogy with (85).

**2.9** When we integrate over $\mu_{M-1}$ the lower limit of integration is 0, while the upper limit is $1 - \sum_{j=1}^{M-2} \mu_j$ since the remaining probabilities must sum to one (see Figure 2.4). Thus we have

$$\begin{aligned} p_{M-1}(\mu_1, \ldots, \mu_{M-2}) &= \int_0^{1-\sum_{j=1}^{M-2}\mu_j} p_M(\mu_1, \ldots, \mu_{M-1})\,\mathrm{d}\mu_{M-1} \\ &= C_M \left[ \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \right] \int_0^{1-\sum_{j=1}^{M-2}\mu_j} \mu_{M-1}^{\alpha_{M-1}-1} \left( 1 - \sum_{j=1}^{M-1} \mu_j \right)^{\alpha_M - 1} \mathrm{d}\mu_{M-1}. \end{aligned}$$

In order to make the limits of integration equal to 0 and 1 we change integration variable from $\mu_{M-1}$ to $t$ using

$$\mu_{M-1} = t\left(1 - \sum_{j=1}^{M-2} \mu_j\right)$$

which gives

$$\begin{aligned} & p_{M-1}(\mu_1,\ldots,\mu_{M-2}) \\ &= C_M\left[\prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\right]\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1}\int_0^1 t^{\alpha_{M-1}-1}(1-t)^{\alpha_M-1}\,\mathrm{d}t \\ &= C_M\left[\prod_{k=1}^{M-2}\mu_k^{\alpha_k-1}\right]\left(1-\sum_{j=1}^{M-2}\mu_j\right)^{\alpha_{M-1}+\alpha_M-1}\frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} \end{aligned} \tag{86}$$

where we have used (2.265). The right hand side of (86) is seen to be a normalized Dirichlet distribution over $M-1$ variables, with coefficients $\alpha_1,\ldots,\alpha_{M-2},\alpha_{M-1}+\alpha_M$, (note that we have effectively combined the final two categories) and we can identify its normalization coefficient using (2.38). Thus

$$\begin{aligned} C_M &= \frac{\Gamma(\alpha_1+\ldots+\alpha_M)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1}+\alpha_M)}\cdot\frac{\Gamma(\alpha_{M-1}+\alpha_M)}{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)} \\ &= \frac{\Gamma(\alpha_1+\ldots+\alpha_M)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_M)} \end{aligned} \tag{87}$$

as required.

**2.10** Using the fact that the Dirichlet distribution (2.38) is normalized we have

$$\int\prod_{k=1}^{M}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu}=\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}{\Gamma(\alpha_0)} \tag{88}$$

where $\int \mathrm{d}\boldsymbol{\mu}$ denotes the integral over the $(M-1)$-dimensional simplex defined by $0\leqslant\mu_k\leqslant 1$ and $\sum_k\mu_k=1$. Now consider the expectation of $\mu_j$ which can be written

$$\begin{aligned} \mathbb{E}[\mu_j] &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}\int\mu_j\prod_{k=1}^{M}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}\cdot\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_j+1)\cdots\Gamma(\alpha_M)}{\Gamma(\alpha_0+1)}=\frac{\alpha_j}{\alpha_0} \end{aligned}$$

where we have made use of (88), noting that the effect of the extra factor of $\mu_j$ is to increase the coefficient $\alpha_j$ by 1, and then made use of $\Gamma(x+1)=x\Gamma(x)$. By similar reasoning we have

$$\begin{aligned} \mathrm{var}[\mu_j] &= \mathbb{E}[\mu_j^2]-\mathbb{E}[\mu_j]^2=\frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)}-\frac{\alpha_j^2}{\alpha_0^2} \\ &= \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}. \end{aligned}$$

Likewise, for $j \neq l$ we have

$$\begin{aligned}\mathrm{cov}[\mu_j \mu_l] &= \mathbb{E}[\mu_j \mu_l] - \mathbb{E}[\mu_j]\mathbb{E}[\mu_l] = \frac{\alpha_j \alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j}{\alpha_0}\frac{\alpha_l}{\alpha_0} \\ &= -\frac{\alpha_j \alpha_l}{\alpha_0^2(\alpha_0+1)}.\end{aligned}$$

**2.11** We first of all write the Dirichlet distribution (2.38) in the form

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = K(\boldsymbol{\alpha}) \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}$$

where

$$K(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_M)}.$$

Next we note the following relation

$$\begin{aligned}\frac{\partial}{\partial \alpha_j}\prod_{k=1}^{M} \mu_k^{\alpha_k - 1} &= \frac{\partial}{\partial \alpha_j}\prod_{k=1}^{M} \exp\left((\alpha_k - 1)\ln \mu_k\right) \\ &= \prod_{k=1}^{M} \ln \mu_j \exp\left\{(\alpha_k - 1)\ln \mu_k\right\} \\ &= \ln \mu_j \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}\end{aligned}$$

from which we obtain

$$\begin{aligned}\mathbb{E}[\ln \mu_j] &= K(\boldsymbol{\alpha}) \int_0^1 \cdots \int_0^1 \ln \mu_j \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}\, \mathrm{d}\mu_1 \ldots \mathrm{d}\mu_M \\ &= K(\boldsymbol{\alpha}) \frac{\partial}{\partial \alpha_j} \int_0^1 \cdots \int_0^1 \prod_{k=1}^{M} \mu_k^{\alpha_k - 1}\, \mathrm{d}\mu_1 \ldots \mathrm{d}\mu_M \\ &= K(\boldsymbol{\alpha}) \frac{\partial}{\partial \mu_j} \frac{1}{K(\boldsymbol{\alpha})} \\ &= -\frac{\partial}{\partial \mu_j} \ln K(\boldsymbol{\alpha}).\end{aligned}$$

Finally, using the expression for $K(\boldsymbol{\alpha})$, together with the definition of the digamma function $\psi(\cdot)$, we have

$$\mathbb{E}[\ln \mu_j] = \psi(\alpha_j) - \psi(\alpha_0).$$

**2.12** The normalization of the uniform distribution is proved trivially

$$\int_a^b \frac{1}{b-a}\,\mathrm{d}x = \frac{b-a}{b-a} = 1.$$

For the mean of the distribution we have

$$\mathbb{E}[x] = \int_a^b \frac{1}{b-a}x\,\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

The variance can be found by first evaluating

$$\mathbb{E}[x^2] = \int_a^b \frac{1}{b-a}x^2\,\mathrm{d}x = \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

and then using (1.40) to give

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

**2.13** Note that this solution is the multivariate version of Solution 1.30.

From (1.113) we have

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

Using (2.43), (2.57), (2.59) and (2.62), we can rewrite the first integral on the r.h.s. of () as

$$\begin{aligned}
&-\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\quad= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}^2)\frac{1}{2}\left(D\ln(2\pi) + \ln|\mathbf{L}| + (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right)\,\mathrm{d}\mathbf{x} \\
&\quad= \frac{1}{2}\left(D\ln(2\pi) + \ln|\mathbf{L}| + \mathrm{Tr}[\mathbf{L}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma})]\right. \\
&\qquad\qquad \left. -\boldsymbol{\mu}\mathbf{L}^{-1}\mathbf{m} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} + \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m}\right). \qquad (89)
\end{aligned}$$

The second integral on the r.h.s. of () we recognize from (1.104) as the negative differential entropy of a multivariate Gaussian. Thus, from (), (89) and (B.41), we have

$$\begin{aligned}
\mathrm{KL}(p\|q) = \frac{1}{2}\Bigg(&\ln\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|} + \mathrm{Tr}[\mathbf{L}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma})] \\
&- \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} - \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\mu} + \mathbf{m}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{m} - D\Bigg)
\end{aligned}$$

**2.14** As for the univariate Gaussian considered in Section 1.6, we can make use of Lagrange multipliers to enforce the constraints on the maximum entropy solution. Note that we need a single Lagrange multiplier for the normalization constraint (2.280), a $D$-dimensional vector $\mathbf{m}$ of Lagrange multipliers for the $D$ constraints given by (2.281), and a $D \times D$ matrix $\mathbf{L}$ of Lagrange multipliers to enforce the $D^2$ constraints represented by (2.282). Thus we maximize

$$\begin{aligned}\widetilde{\mathrm{H}}[p] = & -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \lambda \left( \int p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - 1 \right) \\ & + \mathbf{m}^{\mathrm{T}} \left( \int p(\mathbf{x}) \mathbf{x} \, \mathrm{d}\mathbf{x} - \boldsymbol{\mu} \right) \\ & + \mathrm{Tr} \left\{ \mathbf{L} \left( \int p(\mathbf{x}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \, \mathrm{d}\mathbf{x} - \boldsymbol{\Sigma} \right) \right\}. \end{aligned} \tag{90}$$

By functional differentiation (Appendix D) the maximum of this functional with respect to $p(\mathbf{x})$ occurs when

$$0 = -1 - \ln p(\mathbf{x}) + \lambda + \mathbf{m}^{\mathrm{T}} \mathbf{x} + \mathrm{Tr}\{\mathbf{L}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\}.$$

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

$$p(\mathbf{x}) = \exp \left\{ \lambda - 1 + \mathbf{m}^{\mathrm{T}} \mathbf{x} + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{L} (\mathbf{x} - \boldsymbol{\mu}) \right\}. \tag{91}$$

We now find the values of the Lagrange multipliers by applying the constraints. First we complete the square inside the exponential, which becomes

$$\lambda - 1 + \left( \mathbf{x} - \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^{-1} \mathbf{m} \right)^{\mathrm{T}} \mathbf{L} \left( \mathbf{x} - \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^{-1} \mathbf{m} \right) + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{m} - \frac{1}{4} \mathbf{m}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{m}.$$

We now make the change of variable

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} + \frac{1}{2} \mathbf{L}^{-1} \mathbf{m}.$$

The constraint (2.281) then becomes

$$\int \exp \left\{ \lambda - 1 + \mathbf{y}^{\mathrm{T}} \mathbf{L} \mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{m} - \frac{1}{4} \mathbf{m}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{m} \right\} \left( \mathbf{y} + \boldsymbol{\mu} - \frac{1}{2} \mathbf{L}^{-1} \mathbf{m} \right) \mathrm{d}\mathbf{y} = \boldsymbol{\mu}.$$

In the final parentheses, the term in $\mathbf{y}$ vanishes by symmetry, while the term in $\boldsymbol{\mu}$ simply integrates to $\boldsymbol{\mu}$ by virtue of the normalization constraint (2.280) which now takes the form

$$\int \exp \left\{ \lambda - 1 + \mathbf{y}^{\mathrm{T}} \mathbf{L} \mathbf{y} + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{m} - \frac{1}{4} \mathbf{m}^{\mathrm{T}} \mathbf{L}^{-1} \mathbf{m} \right\} \mathrm{d}\mathbf{y} = 1.$$

and hence we have

$$-\frac{1}{2} \mathbf{L}^{-1} \mathbf{m} = \mathbf{0}$$

where again we have made use of the constraint (2.280). Thus $\mathbf{m} = \mathbf{0}$ and so the density becomes

$$p(\mathbf{x}) = \exp\left\{\lambda - 1 + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

Substituting this into the final constraint (2.282), and making the change of variable $\mathbf{x} - \boldsymbol{\mu} = \mathbf{z}$ we obtain

$$\int \exp\left\{\lambda - 1 + \mathbf{z}^{\mathrm{T}}\mathbf{L}\mathbf{z}\right\}\mathbf{z}\mathbf{z}^{\mathrm{T}}\,\mathrm{d}\mathbf{x} = \boldsymbol{\Sigma}.$$

Applying an analogous argument to that used to derive (2.64) we obtain $\mathbf{L} = -\frac{1}{2}\boldsymbol{\Sigma}$. Finally, the value of $\lambda$ is simply that value needed to ensure that the Gaussian distribution is correctly normalized, as derived in Section 2.3, and hence is given by

$$\lambda - 1 = \ln\left\{\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\right\}.$$

**2.15** From the definitions of the multivariate differential entropy (1.104) and the multivariate Gaussian distribution (2.43), we get

$$\begin{aligned}
\mathrm{H}[\mathbf{x}] &= -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\,\mathrm{d}\mathbf{x} \\
&= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\frac{1}{2}\left(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right]\right) \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + D\right)
\end{aligned}$$

**2.16** We have $p(x_1) = \mathcal{N}(x_1|\mu_1, \tau_1^{-1})$ and $p(x_2) = \mathcal{N}(x_2|\mu_2, \tau_2^{-1})$. Since $x = x_1 + x_2$ we also have $p(x|x_2) = \mathcal{N}(x|\mu_1 + x_2, \tau_1^{-1})$. We now evaluate the convolution integral given by (2.284) which takes the form

$$p(x) = \left(\frac{\tau_1}{2\pi}\right)^{1/2}\left(\frac{\tau_2}{2\pi}\right)^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{\tau_1}{2}(x - \mu_1 - x_2)^2 - \frac{\tau_2}{2}(x_2 - \mu_2)^2\right\}\,\mathrm{d}x_2. \tag{92}$$

Since the final result will be a Gaussian distribution for $p(x)$ we need only evaluate its precision, since, from (1.110), the entropy is determined by the variance or equivalently the precision, and is independent of the mean. This allows us to simplify the calculation by ignoring such things as normalization constants.

We begin by considering the terms in the exponent of (92) which depend on $x_2$ which are given by

$$\begin{aligned}
&-\frac{1}{2}x_2^2(\tau_1 + \tau_2) + x_2\left\{\tau_1(x - \mu_1) + \tau_2\mu_2\right\} \\
&\quad = -\frac{1}{2}(\tau_1 + \tau_2)\left\{x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right\}^2 + \frac{\left\{\tau_1(x - \mu_1) + \tau_2\mu_2\right\}^2}{2(\tau_1 + \tau_2)}
\end{aligned}$$

where we have completed the square over $x_2$. When we integrate out $x_2$, the first term on the right hand side will simply give rise to a constant factor independent of $x$. The second term, when expanded out, will involve a term in $x^2$. Since the precision of $x$ is given directly in terms of the coefficient of $x^2$ in the exponent, it is only such terms that we need to consider. There is one other term in $x^2$ arising from the original exponent in (92). Combining these we have

$$-\frac{\tau_1}{2}x^2 + \frac{\tau_1^2}{2(\tau_1+\tau_2)}x^2 = -\frac{1}{2}\frac{\tau_1\tau_2}{\tau_1+\tau_2}x^2$$

from which we see that $x$ has precision $\tau_1\tau_2/(\tau_1+\tau_2)$.

We can also obtain this result for the precision directly by appealing to the general result (2.115) for the convolution of two linear-Gaussian distributions.

The entropy of $x$ is then given, from (1.110), by

$$\mathrm{H}[x] = \frac{1}{2}\ln\left\{\frac{2\pi(\tau_1+\tau_2)}{\tau_1\tau_2}\right\}.$$

**2.17** We can use an analogous argument to that used in the solution of Exercise 1.14. Consider a general square matrix $\mathbf{\Lambda}$ with elements $\Lambda_{ij}$. Then we can always write $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$ where

$$\Lambda_{ij}^{\mathrm{S}} = \frac{\Lambda_{ij}+\Lambda_{ji}}{2}, \qquad \Lambda_{ij}^{\mathrm{A}} = \frac{\Lambda_{ij}-\Lambda_{ji}}{2} \tag{93}$$

and it is easily verified that $\mathbf{\Lambda}^{\mathrm{S}}$ is symmetric so that $\Lambda_{ij}^{\mathrm{S}} = \Lambda_{ji}^{\mathrm{S}}$, and $\mathbf{\Lambda}^{\mathrm{A}}$ is antisymmetric so that $\Lambda_{ij}^{\mathrm{A}} = -\Lambda_{ji}^{\mathrm{S}}$. The quadratic form in the exponent of a $D$-dimensional multivariate Gaussian distribution can be written

$$\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)\Lambda_{ij}(x_j-\mu_j) \tag{94}$$

where $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ is the precision matrix. When we substitute $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$ into (94) we see that the term involving $\mathbf{\Lambda}^{\mathrm{A}}$ vanishes since for every positive term there is an equal and opposite negative term. Thus we can always take $\mathbf{\Lambda}$ to be symmetric.

**2.18** We start by pre-multiplying both sides of (2.45) by $\mathbf{u}_i^\dagger$, the conjugate transpose of $\mathbf{u}_i$. This gives us

$$\mathbf{u}_i^\dagger\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i^\dagger\mathbf{u}_i. \tag{95}$$

Next consider the conjugate transpose of (2.45) and post-multiply it by $\mathbf{u}_i$, which gives us

$$\mathbf{u}_i^\dagger\mathbf{\Sigma}^\dagger\mathbf{u}_i = \lambda_i^*\mathbf{u}_i^\dagger\mathbf{u}_i. \tag{96}$$

where $\lambda_i^*$ is the complex conjugate of $\lambda_i$. We now subtract (95) from (96) and use the fact the $\mathbf{\Sigma}$ is real and symmetric and hence $\mathbf{\Sigma} = \mathbf{\Sigma}^\dagger$, to get

$$0 = (\lambda_i^* - \lambda_i)\mathbf{u}_i^\dagger\mathbf{u}_i.$$

Hence $\lambda_i^* = \lambda_i$ and so $\lambda_i$ must be real.

Now consider

$$\begin{aligned}\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j\lambda_j &= \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u}_j \\ &= \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{u}_j \\ &= (\boldsymbol{\Sigma}\mathbf{u}_i)^{\mathrm{T}}\mathbf{u}_j \\ &= \lambda_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j,\end{aligned}$$

where we have used (2.45) and the fact that $\boldsymbol{\Sigma}$ is symmetric. If we assume that $0 \neq \lambda_i \neq \lambda_j \neq 0$, the only solution to this equation is that $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 0$, i.e., that $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal.

If $0 \neq \lambda_i = \lambda_j \neq 0$, any linear combination of $\mathbf{u}_i$ and $\mathbf{u}_j$ will be an eigenvector with eigenvalue $\lambda = \lambda_i = \lambda_j$, since, from (2.45),

$$\begin{aligned}\boldsymbol{\Sigma}(a\mathbf{u}_i + b\mathbf{u}_j) &= a\lambda_i\mathbf{u}_i + b\lambda_j\mathbf{u}_j \\ &= \lambda(a\mathbf{u}_i + b\mathbf{u}_j).\end{aligned}$$

Assuming that $\mathbf{u}_i \neq \mathbf{u}_j$, we can construct

$$\begin{aligned}\mathbf{u}_\alpha &= a\mathbf{u}_i + b\mathbf{u}_j \\ \mathbf{u}_\beta &= c\mathbf{u}_i + d\mathbf{u}_j\end{aligned}$$

such that $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$ are mutually orthogonal and of unit length. Since $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal to $\mathbf{u}_k$ ($k \neq i$, $k \neq j$), so are $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$. Thus, $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$ satisfy (2.46).

Finally, if $\lambda_i = 0$, $\boldsymbol{\Sigma}$ must be singular, with $\mathbf{u}_i$ lying in the nullspace of $\boldsymbol{\Sigma}$. In this case, $\mathbf{u}_i$ will be orthogonal to the eigenvectors projecting onto the rowspace of $\boldsymbol{\Sigma}$ and we can chose $\|\mathbf{u}_i\| = 1$, so that (2.46) is satisfied. If more than one eigenvalue equals zero, we can chose the corresponding eigenvectors arbitrily, as long as they remain in the nullspace of $\boldsymbol{\Sigma}$, and so we can chose them to satisfy (2.46).

**2.19** We can write the r.h.s. of (2.48) in matrix form as

$$\sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}} = \mathbf{M},$$

where $\mathbf{U}$ is a $D \times D$ matrix with the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_D$ as its columns and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_D$ along its diagonal.

Thus we have

$$\mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\mathbf{U} = \boldsymbol{\Lambda}.$$

However, from (2.45)–(2.47), we also have that

$$\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{U} = \mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda} = \boldsymbol{\Lambda},$$

and so $\mathbf{M} = \boldsymbol{\Sigma}$ and (2.48) holds.

Moreover, since $\mathbf{U}$ is orthonormal, $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$ and so

$$\boldsymbol{\Sigma}^{-1} = \left(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}} = \sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}.$$

**2.20** Since $\mathbf{u}_1, \ldots, \mathbf{u}_D$ constitute a basis for $\mathbb{R}^D$, we can write

$$\mathbf{a} = \hat{a}_1\mathbf{u}_1 + \hat{a}_2\mathbf{u}_2 + \ldots + \hat{a}_D\mathbf{u}_D,$$

where $\hat{a}_1, \ldots, \hat{a}_D$ are coefficients obtained by projecting $\mathbf{a}$ on $\mathbf{u}_1, \ldots, \mathbf{u}_D$. Note that they typically do *not* equal the elements of $\mathbf{a}$.

Using this we can write

$$\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{a} = \left(\hat{a}_1\mathbf{u}_1^{\mathrm{T}} + \ldots + \hat{a}_D\mathbf{u}_D^{\mathrm{T}}\right)\boldsymbol{\Sigma}\left(\hat{a}_1\mathbf{u}_1 + \ldots + \hat{a}_D\mathbf{u}_D\right)$$

and combining this result with (2.45) we get

$$\left(\hat{a}_1\mathbf{u}_1^{\mathrm{T}} + \ldots + \hat{a}_D\mathbf{u}_D^{\mathrm{T}}\right)\left(\hat{a}_1\lambda_1\mathbf{u}_1 + \ldots + \hat{a}_D\lambda_D\mathbf{u}_D\right).$$

Now, since $\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = 1$ only if $i = j$, and 0 otherwise, this becomes

$$\hat{a}_1^2\lambda_1 + \ldots + \hat{a}_D^2\lambda_D$$

and since $\mathbf{a}$ is real, we see that this expression will be strictly positive for any non-zero $\mathbf{a}$, if all eigenvalues are strictly positive. It is also clear that if an eigenvalue, $\lambda_i$, is zero or negative, there exist a vector $\mathbf{a}$ (e.g. $\mathbf{a} = \mathbf{u}_i$), for which this expression will be less than or equal to zero. Thus, that a matrix has eigenvectors which are all strictly positive is a sufficient and necessary condition for the matrix to be positive definite.

**2.21** A $D \times D$ matrix has $D^2$ elements. If it is symmetric then the elements not on the leading diagonal form pairs of equal value. There are $D$ elements on the diagonal so the number of elements not on the diagonal is $D^2 - D$ and only half of these are independent giving

$$\frac{D^2 - D}{2}.$$

If we now add back the $D$ elements on the diagonal we get

$$\frac{D^2 - D}{2} + D = \frac{D(D+1)}{2}.$$

**2.22** Consider a matrix $\mathbf{M}$ which is symmetric, so that $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$. The inverse matrix $\mathbf{M}^{-1}$ satisfies

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}.$$

Taking the transpose of both sides of this equation, and using the relation (C.1), we obtain

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}}\mathbf{M}^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I}$$

since the identity matrix is symmetric. Making use of the symmetry condition for $\mathbf{M}$ we then have

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}}\mathbf{M} = \mathbf{I}$$

and hence, from the definition of the matrix inverse,

$$\left(\mathbf{M}^{-1}\right)^{\mathrm{T}} = \mathbf{M}^{-1}$$

and so $\mathbf{M}^{-1}$ is also a symmetric matrix.

**2.23** Recall that the transformation (2.51) diagonalizes the coordinate system and that the quadratic form (2.44), corresponding to the square of the Mahalanobis distance, is then given by (2.50). This corresponds to a shift in the origin of the coordinate system and a rotation so that the hyper-ellipsoidal contours along which the Mahalanobis distance is constant become axis aligned. The volume contained within any one such contour is unchanged by shifts and rotations. We now make the further transformation $z_i = \lambda_i^{1/2} y_i$ for $i = 1, \ldots, D$. The volume within the hyper-ellipsoid then becomes

$$\int \prod_{i=1}^{D} \mathrm{d}y_i = \prod_{i=1}^{D} \lambda_i^{1/2} \int \prod_{i=1}^{D} \mathrm{d}z_i = |\boldsymbol{\Sigma}|^{1/2} V_D \Delta^D$$

where we have used the property that the determinant of $\boldsymbol{\Sigma}$ is given by the product of its eigenvalues, together with the fact that in the $z$ coordinates the volume has become a sphere of radius $\Delta$ whose volume is $V_D \Delta^D$.

**2.24** Multiplying the left hand side of (2.76) by the matrix (2.287) trivially gives the identity matrix. On the right hand side consider the four blocks of the resulting partitioned matrix:

**upper left**

$$\mathbf{AM} - \mathbf{BD}^{-1}\mathbf{CM} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} = \mathbf{I}$$

**upper right**

$$\begin{aligned} -\mathbf{AMBD}^{-1} + \mathbf{BD}^{-1} + \mathbf{BD}^{-1}\mathbf{CMBD}^{-1} \\ = \quad -(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} + \mathbf{BD}^{-1} \\ = \quad -\mathbf{BD}^{-1} + \mathbf{BD}^{-1} = \mathbf{0} \end{aligned}$$

**lower left**

$$\mathbf{CM} - \mathbf{DD}^{-1}\mathbf{CM} = \mathbf{CM} - \mathbf{CM} = \mathbf{0}$$

**lower right**

$$-\mathbf{CMBD}^{-1} + \mathbf{DD}^{-1} + \mathbf{DD}^{-1}\mathbf{CMBD}^{-1} = \mathbf{DD}^{-1} = \mathbf{I}.$$

Thus the right hand side also equals the identity matrix.

**2.25** We first of all take the joint distribution $p(\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c)$ and marginalize to obtain the distribution $p(\mathbf{x}_a, \mathbf{x}_b)$. Using the results of Section 2.3.2 this is again a Gaussian distribution with mean and covariance given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

From Section 2.3.1 the distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is then Gaussian with mean and covariance given by (2.81) and (2.82) respectively.

**2.26** Multiplying the left hand side of (2.289) by $(\mathbf{A} + \mathbf{BCD})$ trivially gives the identity matrix $\mathbf{I}$. On the right hand side we obtain

$$\begin{aligned}
&(\mathbf{A} + \mathbf{BCD})(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}) \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\qquad -\mathbf{BCDA}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BC}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1} \\
&\quad = \mathbf{I} + \mathbf{BCDA}^{-1} - \mathbf{BCDA}^{-1} = \mathbf{I}
\end{aligned}$$

**2.27** From $\mathbf{y} = \mathbf{x} + \mathbf{z}$ we have trivially that $\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{z}]$. For the covariance we have

$$\begin{aligned}
\mathrm{cov}[\mathbf{y}] &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}] + \mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x}] + \mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] + \mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}\right] \\
&\quad + \underbrace{\mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^{\mathrm{T}}\right]}_{=0} + \underbrace{\mathbb{E}\left[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right]}_{=0} \\
&= \mathrm{cov}[\mathbf{x}] + \mathrm{cov}[\mathbf{z}]
\end{aligned}$$

where we have used the independence of $\mathbf{x}$ and $\mathbf{z}$, together with $\mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])] = \mathbb{E}[(\mathbf{z} - \mathbb{E}[\mathbf{z}])] = 0$, to set the third and fourth terms in the expansion to zero. For 1-dimensional variables the covariances become variances and we obtain the result of Exercise 1.10 as a special case.

**2.28** For the marginal distribution $p(\mathbf{x})$ we see from (2.92) that the mean is given by the upper partition of (2.108) which is simply $\boldsymbol{\mu}$. Similarly from (2.93) we see that the covariance is given by the top left partition of (2.105) and is therefore given by $\boldsymbol{\Lambda}^{-1}$.

Now consider the conditional distribution $p(\mathbf{y}|\mathbf{x})$. Applying the result (2.81) for the conditional mean we obtain

$$\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{Ax} + \mathbf{b}.$$

Similarly applying the result (2.82) for the covariance of the conditional distribution we have

$$\mathrm{cov}[\mathbf{y}|\mathbf{x}] = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} = \mathbf{L}^{-1}$$

as required.

**2.29** We first define

$$\mathbf{X} = \boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} \tag{97}$$

and

$$\mathbf{W} = -\mathbf{L}\mathbf{A}, \quad \text{and thus} \quad \mathbf{W}^{\mathrm{T}} = -\mathbf{A}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}} = -\mathbf{A}^{\mathrm{T}}\mathbf{L}, \tag{98}$$

since $\mathbf{L}$ is symmetric. We can use (97) and (98) to re-write (2.104) as

$$\mathbf{R} = \begin{pmatrix} \mathbf{X} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{W} & \mathbf{L} \end{pmatrix}$$

and using (2.76) we get

$$\begin{pmatrix} \mathbf{X} & \mathbf{W}^{\mathrm{T}} \\ \mathbf{W} & \mathbf{L} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} \\ -\mathbf{L}^{-1}\mathbf{W}\mathbf{M} & \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} \end{pmatrix}$$

where now

$$\mathbf{M} = \left(\mathbf{X} - \mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{W}\right)^{-1}.$$

Substituting $\mathbf{X}$ and $\mathbf{W}$ using (97) and (98), respectively, we get

$$\mathbf{M} = \left(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} - \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A}\right)^{-1} = \boldsymbol{\Lambda}^{-1},$$

$$-\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} = \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} = \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}$$

and

$$\begin{aligned} \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^{\mathrm{T}}\mathbf{L}^{-1} &= \mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{L}^{-1} \\ &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}, \end{aligned}$$

as required.

**2.30** Substituting the leftmost expression of (2.105) for $\mathbf{R}^{-1}$ in (2.107), we get

$$\begin{aligned} &\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{S}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{S}\mathbf{b} \end{pmatrix} \\ &\qquad = \begin{pmatrix} \boldsymbol{\Lambda}^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b}\right) + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b}\right) + \left(\mathbf{S}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)\mathbf{S}\mathbf{b} \end{pmatrix} \\ &\qquad = \begin{pmatrix} \boldsymbol{\mu} - \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{b} \end{pmatrix} \\ &\qquad = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{b} \end{pmatrix} \end{aligned}$$

**2.31** Since $\mathbf{y} = \mathbf{x} + \mathbf{z}$ we can write the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ in the form $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_\mathbf{z} + \mathbf{x}, \boldsymbol{\Sigma}_\mathbf{z})$. This gives a decomposition of the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ in the form $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ where $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})$. This therefore takes the form of (2.99) and (2.100) in which we can identify $\boldsymbol{\mu} \to \boldsymbol{\mu}_\mathbf{x}$, $\boldsymbol{\Lambda}^{-1} \to \boldsymbol{\Sigma}_\mathbf{x}$, $\mathbf{A} \to \mathbf{I}$, $\mathbf{b} \to \boldsymbol{\mu}_\mathbf{z}$ and $\mathbf{L}^{-1} \to \boldsymbol{\Sigma}_\mathbf{z}$. We can now obtain the marginal distribution $p(\mathbf{y})$ by making use of the result (2.115) from which we obtain $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_\mathbf{x} + \boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z} + \boldsymbol{\Sigma}_\mathbf{x})$. Thus both the means and the covariances are additive, in agreement with the results of Exercise 2.27.

**2.32** The quadratic form in the exponential of the joint distribution is given by

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathrm{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\mathrm{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{99}$$

We now extract all of those terms involving $\mathbf{x}$ and assemble them into a standard Gaussian quadratic form by completing the square

$$\begin{aligned} &= -\frac{1}{2}\mathbf{x}^\mathrm{T}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^\mathrm{T}\left[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathrm{T}\mathbf{L}(\mathbf{y} - \mathbf{b})\right] + \text{const} \\ &= -\frac{1}{2}(\mathbf{x} - \mathbf{m})^\mathrm{T}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})(\mathbf{x} - \mathbf{m}) \\ &\quad + \frac{1}{2}\mathbf{m}^\mathrm{T}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})\mathbf{m} + \text{const} \end{aligned} \tag{100}$$

where

$$\mathbf{m} = (\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\left[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathrm{T}\mathbf{L}(\mathbf{y} - \mathbf{b})\right].$$

We can now perform the integration over $\mathbf{x}$ which eliminates the first term in (100). Then we extract the terms in $\mathbf{y}$ from the final term in (100) and combine these with the remaining terms from the quadratic form (99) which depend on $\mathbf{y}$ to give

$$\begin{aligned} = &-\frac{1}{2}\mathbf{y}^\mathrm{T}\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{L}\right\}\mathbf{y} \\ &+\mathbf{y}^\mathrm{T}\left[\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{L}\right\}\mathbf{b}\right. \\ &\left.+\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\right]. \end{aligned} \tag{101}$$

We can identify the precision of the marginal distribution $p(\mathbf{y})$ from the second order term in $\mathbf{y}$. To find the corresponding covariance, we take the inverse of the precision and apply the Woodbury inversion formula (2.289) to give

$$\left\{\mathbf{L} - \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^\mathrm{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{L}\right\}^{-1} = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathrm{T} \tag{102}$$

which corresponds to (2.110).

Next we identify the mean $\boldsymbol{\nu}$ of the marginal distribution. To do this we make use of (102) in (101) and then complete the square to give

$$-\frac{1}{2}(\mathbf{y} - \boldsymbol{\nu})^\mathrm{T}\left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathrm{T}\right)^{-1}(\mathbf{y} - \boldsymbol{\nu}) + \text{const}$$

where

$$\boldsymbol{\nu} = \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)\left[(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{b} + \mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\right].$$

Now consider the two terms in the square brackets, the first one involving $\mathbf{b}$ and the second involving $\boldsymbol{\mu}$. The first of these contribution simply gives $\mathbf{b}$, while the term in $\boldsymbol{\mu}$ can be written

$$\begin{aligned} &= \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\right)\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} \\ &= \mathbf{A}(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{I} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\mu} \end{aligned}$$

where we have used the general result $(\mathbf{B}\mathbf{C})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}$. Hence we obtain (2.109).

**2.33** To find the conditional distribution $p(\mathbf{x}|\mathbf{y})$ we start from the quadratic form (99) corresponding to the joint distribution $p(\mathbf{x}, \mathbf{y})$. Now, however, we treat $\mathbf{y}$ as a constant and simply complete the square over $\mathbf{x}$ to give

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) \\ = & -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x} + \mathbf{x}^{\mathrm{T}}\left\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}\mathbf{L}(\mathbf{y} - \mathbf{b})\right\} + \mathrm{const} \\ = & -\frac{1}{2}(\mathbf{x} - \mathbf{m})^{\mathrm{T}}(\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{x} - \mathbf{m}) \end{aligned}$$

where, as in the solution to Exercise 2.32, we have defined

$$\mathbf{m} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\left\{\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b})\right\}$$

from which we obtain directly the mean and covariance of the conditional distribution in the form (2.111) and (2.112).

**2.34** Differentiating (2.118) with respect to $\boldsymbol{\Sigma}$ we obtain two terms:

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

For the first term, we can apply (C.28) directly to get

$$-\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}}\ln|\boldsymbol{\Sigma}| = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathrm{T}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1}.$$

For the second term, we first re-write the sum

$$\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = N\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right],$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}.$$

Using this together with (C.21), in which $x = \Sigma_{ij}$ (element $(i, j)$ in $\boldsymbol{\Sigma}$), and properties of the trace we get

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{ij}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) &= N\frac{\partial}{\partial \Sigma_{ij}}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= N\mathrm{Tr}\left[\frac{\partial}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\right] \\
&= -N\mathrm{Tr}\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right] \\
&= -N\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\right)_{ij}
\end{aligned}$$

where we have used (C.26). Note that in the last step we have ignored the fact that $\Sigma_{ij} = \Sigma_{ji}$, so that $\partial\boldsymbol{\Sigma}/\partial\Sigma_{ij}$ has a 1 in position $(i, j)$ only and 0 everywhere else. Treating this result as valid nevertheless, we get

$$-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}.$$

Combining the derivatives of the two terms and setting the result to zero, we obtain

$$\frac{N}{2}\boldsymbol{\Sigma}^{-1} = \frac{N}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}.$$

Re-arrangement then yields

$$\boldsymbol{\Sigma} = \mathbf{S}$$

as required.

**2.35** **NOTE**: In PRML, this exercise contains a typographical error; $\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_m\right]$ should be $\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_m^{\mathrm{T}}\right]$ on the l.h.s. of (2.291).

The derivation of (2.62) is detailed in the text between (2.59) (page 82) and (2.62) (page 83).

If $m = n$ then, using (2.62) we have $\mathbb{E}[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$, whereas if $n \neq m$ then the two data points $\mathbf{x}_n$ and $\mathbf{x}_m$ are independent and hence $\mathbb{E}[\mathbf{x}_n\mathbf{x}_m] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ where we have used (2.59). Combining these results we obtain (2.291). From (2.59) and

(2.62) we then have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Sigma}_{\mathrm{ML}}\right] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{x}_n - \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\left(\mathbf{x}_n^{\mathrm{T}} - \frac{1}{N}\sum_{l=1}^{N}\mathbf{x}_l^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \frac{2}{N}\mathbf{x}_n\sum_{m=1}^{N}\mathbf{x}_m^{\mathrm{T}} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbf{x}_m\mathbf{x}_l^{\mathrm{T}}\right] \\
&= \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - 2\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right\} \\
&= \left(\frac{N-1}{N}\right)\boldsymbol{\Sigma} \qquad (103)
\end{aligned}$$

as required.

**2.36** **NOTE**: In PRML, there are mistakes that affect this solution. The sign in (2.129) is incorrect, and this equation should read

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1}z(\theta^{(N-1)}).$$

Then, in order to be consistent with the assumption that $f(\theta) > 0$ for $\theta > \theta^\star$ and $f(\theta) < 0$ for $\theta < \theta^\star$ in Figure 2.10, we should find the root of the expected *negative* log likelihood. This lead to sign changes in (2.133) and (2.134), but in (2.135), these are cancelled against the change of sign in (2.129), so in effect, (2.135) remains unchanged. Also, $\mathbf{x}_n$ should be $x_n$ on the l.h.s. of (2.133). Finally, the labels $\mu$ and $\mu_{\mathrm{ML}}$ in Figure 2.11 should be interchanged and there are corresponding changes to the caption (see errata on the PRML web site for details).

Consider the expression for $\sigma^2_{(N)}$ and separate out the contribution from observation $x_N$ to give

$$\begin{aligned}
\sigma^2_{(N)} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N-1}(x_n - \mu)^2 + \frac{(x_N - \mu)^2}{N} \\
&= \frac{N-1}{N}\sigma^2_{(N-1)} + \frac{(x_N - \mu)^2}{N} \\
&= \sigma^2_{(N-1)} - \frac{1}{N}\sigma^2_{(N-1)} + \frac{(x_N - \mu)^2}{N} \\
&= \sigma^2_{(N-1)} + \frac{1}{N}\left\{(x_N - \mu)^2 - \sigma^2_{(N-1)}\right\}. \qquad (104)
\end{aligned}$$

If we substitute the expression for a Gaussian distribution into the result (2.135) for

the Robbins-Monro procedure applied to maximizing likelihood, we obtain

$$\begin{aligned}
\sigma^2_{(N)} &= \sigma^2_{(N-1)} + a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\left\{-\frac{1}{2}\ln\sigma^2_{(N-1)} - \frac{(x_N-\mu)^2}{2\sigma^2_{(N-1)}}\right\} \\
&= \sigma^2_{(N-1)} + a_{N-1}\left\{-\frac{1}{2\sigma^2_{(N-1)}} + \frac{(x_N-\mu)^2}{2\sigma^4_{(N-1)}}\right\} \\
&= \sigma^2_{(N-1)} + \frac{a_{N-1}}{2\sigma^4_{(N-1)}}\left\{(x_N-\mu)^2 - \sigma^2_{(N-1)}\right\}. \qquad (105)
\end{aligned}$$

Comparison of (105) with (104) allows us to identify

$$a_{N-1} = \frac{2\sigma^4_{(N-1)}}{N}.$$

**2.37** **NOTE**: In PRML, this exercise requires the additional assumption that we can use the known true mean, $\boldsymbol{\mu}$, in (2.122). Furthermore, for the derivation of the Robbins-Monro sequential estimation formula, we assume that the covariance matrix is restricted to be diagonal. Starting from (2.122), we have

$$\begin{aligned}
\boldsymbol{\Sigma}^{(N)}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \\
&\quad + \frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \frac{N-1}{N}\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}} + \frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} \\
&= \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}} + \frac{1}{N}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right). \qquad (106)
\end{aligned}$$

From Solution 2.34, we know that

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}}\ln p(\mathbf{x}_N|\boldsymbol{\mu},\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}) \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-1}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-1} \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-2}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\mathrm{T}} - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)
\end{aligned}$$

where we have used the assumption that $\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}$, and hence $\left(\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)^{-1}$, is diag-

onal. If we substitute this into the multivariate form of (2.135), we get

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} + \mathbf{A}_{N-1}\frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right)^{-2}\left(\left(\mathbf{x}_N - \boldsymbol{\mu}\right)\left(\mathbf{x}_N - \boldsymbol{\mu}\right)^{\mathrm{T}} - \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right) \tag{107}$$

where $\mathbf{A}_{N-1}$ is a matrix of coefficients corresponding to $a_{N-1}$ in (2.135). By comparing (106) with (107), we see that if we choose

$$\mathbf{A}_{N-1} = \frac{2}{N}\left(\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right)^{2}.$$

we recover (106). Note that if the covariance matrix was restricted further, to the form $\sigma^2\mathbf{I}$, i.e. a spherical Gaussian, the coefficient in (107) would again become a scalar.

**2.38** The exponent in the posterior distribution of (2.140) takes the form

$$-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 = -\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n\right) + \text{const.}$$

where 'const.' denotes terms independent of $\mu$. Following the discussion of (2.71) we see that the variance of the posterior distribution is given by

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

Similarly the mean is given by

$$\begin{aligned}\mu_N &= \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n\right) \\ &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}. \qquad (108)\end{aligned}$$

(109)

**2.39** From (2.142), we see directly that

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{N-1}{\sigma^2} + \frac{1}{\sigma^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \tag{110}$$

We also note for later use, that

$$\frac{1}{\sigma_N^2} = \frac{\sigma^2+N\sigma_0^2}{\sigma_0^2\sigma^2} = \frac{\sigma^2+\sigma_{N-1}^2}{\sigma_{N-1}^2\sigma^2} \tag{111}$$

and similarly

$$\frac{1}{\sigma_{N-1}^2} = \frac{\sigma^2 + (N-1)\sigma_0^2}{\sigma_0^2\sigma^2}. \tag{112}$$

Using (2.143), we can rewrite (2.141) as

$$\begin{aligned}\mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{\sigma_0^2 \sum_{n=1}^{N} x_n}{N\sigma_0^2 + \sigma^2} \\ &= \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{n=1}^{N-1} x_n}{N\sigma_0^2 + \sigma^2} + \frac{\sigma_0^2 x_N}{N\sigma_0^2 + \sigma^2}.\end{aligned}$$

Using (2.141), (111) and (112), we can rewrite the first term of this expression as

$$\frac{\sigma_N^2}{\sigma_{N-1}^2}\frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{n=1}^{N-1} x_n}{(N-1)\sigma_0^2 + \sigma^2} = \frac{\sigma_N^2}{\sigma_{N-1}^2}\mu_{N-1}.$$

Similarly, using (111), the second term can be rewritten as

$$\frac{\sigma_N^2}{\sigma^2}x_N$$

and so

$$\mu_N = \frac{\sigma_N^2}{\sigma_{N-1}^2}\mu_{N-1} + \frac{\sigma_N^2}{\sigma^2}x_N. \tag{113}$$

Now consider

$$\begin{aligned}p(\mu|\mu_N, \sigma_N^2) &= p(\mu|\mu_{N-1}, \sigma_{N-1}^2)p(x_N|\mu, \sigma^2) \\ &= \mathcal{N}(\mu|\mu_{N-1}, \sigma_{N-1}^2)\mathcal{N}(x_N|\mu, \sigma^2) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu_{N-1}^2 - 2\mu\mu_{N-1} + \mu^2}{\sigma_{N-1}^2} + \frac{x_N^2 - 2x_N\mu + \mu^2}{\sigma^2}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{\sigma^2(\mu_{N-1}^2 - 2\mu\mu_{N-1} + \mu^2)}{\sigma_{N-1}^2\sigma^2}\right.\right. \\ &\qquad \left.\left. + \frac{\sigma_{N-1}^2(x_N^2 - 2x_N\mu + \mu^2)}{\sigma_{N-1}^2\sigma^2}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\frac{(\sigma_{N-1}^2 + \sigma^2)\mu^2 - 2(\sigma^2\mu_{N-1} + \sigma_{N-1}^2 x_N)\mu}{\sigma_{N-1}^2\sigma^2}\right\} + C,\end{aligned}$$

where $C$ accounts for all the remaining terms that are independent of $\mu$. From this, we can directly read off

$$\frac{1}{\sigma_N^2} = \frac{\sigma^2 + \sigma_{N-1}^2}{\sigma_{N-1}^2\sigma^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}$$

and

$$\begin{aligned}\mu_N &= \frac{\sigma^2\mu_{N-1} + \sigma^2_{N-1}x_N}{\sigma^2_{N-1} + \sigma^2} \\ &= \frac{\sigma^2}{\sigma^2_{N-1} + \sigma^2}\mu_{N-1} + \frac{\sigma^2_{N-1}}{\sigma^2_{N-1} + \sigma^2}x_N \\ &= \frac{\sigma^2_N}{\sigma^2_{N-1}}\mu_{N-1} + \frac{\sigma^2_N}{\sigma^2}x_N\end{aligned}$$

and so we have recovered (110) and (113).

**2.40** The posterior distribution is proportional to the product of the prior and the likelihood function

$$p(\boldsymbol{\mu}|\mathbf{X}) \propto p(\boldsymbol{\mu}) \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Thus the posterior is proportional to an exponential of a quadratic form in $\boldsymbol{\mu}$ given by

$$\begin{aligned}&-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) \\ &= -\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\mathbf{x}_n\right) + \text{const}\end{aligned}$$

where 'const.' denotes terms independent of $\boldsymbol{\mu}$. Using the discussion following (2.71) we see that the mean and covariance of the posterior distribution are given by

$$\boldsymbol{\mu}_N = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}N\boldsymbol{\mu}_{\mathrm{ML}}\right) \tag{114}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \tag{115}$$

where $\boldsymbol{\mu}_{\mathrm{ML}}$ is the maximum likelihood solution for the mean given by

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n.$$

**2.41** If we consider the integral of the Gamma distribution over $\tau$ and make the change of variable $b\tau = u$ we have

$$\begin{aligned}\int_0^\infty \mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau &= \frac{1}{\Gamma(a)}\int_0^\infty b^a\tau^{a-1}\exp(-b\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\Gamma(a)}\int_0^\infty b^a u^{a-1}\exp(-u)b^{1-a}b^{-1}\,\mathrm{d}u \\ &= 1\end{aligned}$$

where we have used the definition (1.141) of the Gamma function.

**2.42** We can use the same change of variable as in the previous exercise to evaluate the mean of the Gamma distribution

$$\begin{aligned}\mathbb{E}[\tau] &= \frac{1}{\Gamma(a)}\int_0^\infty b^a \tau^{a-1}\tau \exp(-b\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\Gamma(a)}\int_0^\infty b^a u^a \exp(-u) b^{-a} b^{-1}\,\mathrm{d}u \\ &= \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}\end{aligned}$$

where we have used the recurrence relation $\Gamma(a+1) = a\Gamma(a)$ for the Gamma function. Similarly we can find the variance by first evaluating

$$\begin{aligned}\mathbb{E}[\tau^2] &= \frac{1}{\Gamma(a)}\int_0^\infty b^a \tau^{a-1}\tau^2 \exp(-b\tau)\,\mathrm{d}\tau \\ &= \frac{1}{\Gamma(a)}\int_0^\infty b^a u^{a+1} \exp(-u) b^{-a-1} b^{-1}\,\mathrm{d}u \\ &= \frac{\Gamma(a+2)}{b^2\Gamma(a)} = \frac{(a+1)\Gamma(a+1)}{b^2\Gamma(a)} = \frac{a(a+1)}{b^2}\end{aligned}$$

and then using

$$\mathrm{var}[\tau] = \mathbb{E}[\tau^2] - \mathbb{E}[\tau]^2 = \frac{a(a+1)}{b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}.$$

Finally, the mode of the Gamma distribution is obtained simply by differentiation

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left\{\tau^{a-1}\exp(-b\tau)\right\} = \left[\frac{a-1}{\tau} - b\right]\tau^{a-1}\exp(-b\tau) = 0$$

from which we obtain

$$\mathrm{mode}[\tau] = \frac{a-1}{b}.$$

Notice that the mode only exists if $a \geqslant 1$, since $\tau$ must be a non-negative quantity. This is also apparent in the plot of Figure 2.13.

**2.43** To prove the normalization of the distribution (2.293) consider the integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{|x|^q}{2\sigma^2}\right)\mathrm{d}x = 2\int_0^\infty \exp\left(-\frac{x^q}{2\sigma^2}\right)\mathrm{d}x$$

and make the change of variable

$$u = \frac{x^q}{2\sigma^2}.$$

Using the definition (1.141) of the Gamma function, this gives

$$I = 2\int_0^\infty \frac{2\sigma^2}{q}(2\sigma^2 u)^{(1-q)/q}\exp(-u)\,\mathrm{d}u = \frac{2(2\sigma^2)^{1/q}\Gamma(1/q)}{q}$$

from which the normalization of (2.293) follows.

For the given noise distribution, the conditional distribution of the target variable given the input variable is

$$p(t|\mathbf{x},\mathbf{w},\sigma^2) = \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}\exp\left(-\frac{|t-y(\mathbf{x},\mathbf{w})|^q}{2\sigma^2}\right).$$

The likelihood function is obtained by taking products of factors of this form, over all pairs $\{\mathbf{x}_n, t_n\}$. Taking the logarithm, and discarding additive constants, we obtain the desired result.

**2.44** From Bayes' theorem we have

$$p(\mu,\lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu,\lambda)p(\mu,\lambda),$$

where the factors on the r.h.s. are given by (2.152) and (2.154), respectively. Writing this out in full, we get

$$\begin{aligned} p(\mu,\lambda) \quad \propto \quad & \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^N x_n - \frac{\lambda}{2}\sum_{n=1}^N x_n^2\right\} \\ & (\beta\lambda)^{1/2}\exp\left[-\frac{\beta\lambda}{2}\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right)\right]\lambda^{a-1}\exp\left(-b\lambda\right), \end{aligned}$$

where we have used the defintions of the Gaussian and Gamma distributions and we have ommitted terms independent of $\mu$ and $\lambda$. We can rearrange this to obtain

$$\begin{aligned} & \lambda^{N/2}\lambda^{a-1}\exp\left\{-\left(b + \frac{1}{2}\sum_{n=1}^N x_n^2 + \frac{\beta}{2}\mu_0^2\right)\lambda\right\} \\ & \qquad (\lambda(N+\beta))^{1/2}\exp\left[-\frac{\lambda(N+\beta)}{2}\left(\mu^2 - \frac{2}{N+\beta}\left\{\beta\mu_0 + \sum_{n=1}^N x_n\right\}\mu\right)\right] \end{aligned}$$

and by completing the square in the argument of the second exponential,

$$\begin{aligned} & \lambda^{N/2}\lambda^{a-1}\exp\left\{-\left(b + \frac{1}{2}\sum_{n=1}^N x_n^2 + \frac{\beta}{2}\mu_0^2 - \frac{\left(\beta\mu_0 + \sum_{n=1}^N x_n\right)^2}{2(N+\beta)}\right)\lambda\right\} \\ & \qquad (\lambda(N+\beta))^{1/2}\exp\left[-\frac{\lambda(N+\beta)}{2}\left(\mu - \frac{\beta\mu_0 + \sum_{n=1}^N x_n}{N+\beta}\right)\right] \end{aligned}$$

we arrive at an (unnormalised) Gaussian-Gamma distribution,

$$\mathcal{N}\left(\mu|\mu_N,((N+\beta)\lambda)^{-1}\right)\mathrm{Gam}\left(\lambda|a_N,b_N\right),$$

with parameters

$$\begin{aligned}\mu_N &= \frac{\beta\mu_0+\sum_{n=1}^N x_n}{N+\beta}\\ a_N &= a+\frac{N}{2}\\ b_N &= b+\frac{1}{2}\sum_{n=1}^N x_n^2+\frac{\beta}{2}\mu_0^2-\frac{N+\beta}{2}\mu_N^2.\end{aligned}$$

**2.45** We do this, as in the univariate case, by considering the likelihood function of $\mathbf{\Lambda}$ for a given data set $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$:

$$\begin{aligned}\prod_{n=1}^N\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu},\mathbf{\Lambda}^{-1}) &\propto |\mathbf{\Lambda}|^{N/2}\exp\left(-\frac{1}{2}\sum_{n=1}^N(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Lambda}(\mathbf{x}_n-\boldsymbol{\mu})\right)\\ &= |\mathbf{\Lambda}|^{N/2}\exp\left(-\frac{1}{2}\mathrm{Tr}\left[\mathbf{\Lambda}\mathbf{S}\right]\right),\end{aligned}$$

where $\mathbf{S}=\sum_n(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}$. By simply comparing with (2.155), we see that the functional dependence on $\mathbf{\Lambda}$ is indeed the same and thus a product of this likelihood and a Wishart prior will result in a Wishart posterior.

**2.46** From (2.158), we have

$$\begin{aligned}&\int_0^\infty\frac{b^a e^{(-b\tau)}\tau^{a-1}}{\Gamma(a)}\left(\frac{\tau}{2\pi}\right)^{1/2}\exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}\mathrm{d}\tau\\ &\qquad=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\int_0^\infty\tau^{a-1/2}\exp\left\{-\tau\left(b+\frac{(x-\mu)^2}{2}\right)\right\}\mathrm{d}\tau.\end{aligned}$$

We now make the proposed change of variable $z=\tau\Delta$, where $\Delta=b+(x-\mu)^2/2$, yielding

$$\begin{aligned}&\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\Delta^{-a-1/2}\int_0^\infty z^{a-1/2}\exp(-z)\,\mathrm{d}z\\ &\qquad=\frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\Delta^{-a-1/2}\Gamma(a+1/2)\end{aligned}$$

where we have used the definition of the Gamma function (1.141). Finally, we substitute $b + (x-\mu)^2/2$ for $\Delta$, $\nu/2$ for $a$ and $\nu/2\lambda$ for $b$:

$$\begin{aligned}
&\frac{\Gamma(-a+1/2)}{\Gamma(a)} b^a \left(\frac{1}{2\pi}\right)^{1/2} \Delta^{a-1/2} \\
&= \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)} \left(\frac{\nu}{2\lambda}\right)^{\nu/2} \left(\frac{1}{2\pi}\right)^{1/2} \left(\frac{\nu}{2\lambda} + \frac{(x-\mu)^2}{2}\right)^{-(\nu+1)/2} \\
&= \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)} \left(\frac{\nu}{2\lambda}\right)^{\nu/2} \left(\frac{1}{2\pi}\right)^{1/2} \left(\frac{\nu}{2\lambda}\right)^{-(\nu+1)/2} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2} \\
&= \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\nu\pi}\right)^{1/2} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2}
\end{aligned}$$

**2.47** Ignoring the normalization constant, we write (2.159) as

$$\begin{aligned}
\mathrm{St}(x|\mu,\lambda,\nu) &\propto \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-(\nu-1)/2} \\
&= \exp\left(-\frac{\nu-1}{2} \ln\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]\right). \qquad (116)
\end{aligned}$$

For large $\nu$, we make use of the Taylor expansion for the logarithm in the form

$$\ln(1+\epsilon) = \epsilon + O(\epsilon^2) \qquad (117)$$

to re-write (116) as

$$\begin{aligned}
&\exp\left(-\frac{\nu-1}{2} \ln\left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]\right) \\
&= \exp\left(-\frac{\nu-1}{2}\left[\frac{\lambda(x-\mu)^2}{\nu} + O(\nu^{-2})\right]\right) \\
&= \exp\left(-\frac{\lambda(x-\mu)^2}{2} + O(\nu^{-1})\right).
\end{aligned}$$

We see that in the limit $\nu \to \infty$ this becomes, up to an overall constant, the same as a Gaussian distribution with mean $\mu$ and precision $\lambda$. Since the Student distribution is normalized to unity for all values of $\nu$ it follows that it must remain normalized in this limit. The normalization coefficient is given by the standard expression (2.42) for a univariate Gaussian.

**2.48** Substituting expressions for the Gaussian and Gamma distributions into (2.161), we

have

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^\infty \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\int_0^\infty \eta^{D/2}\eta^{\nu/2-1}e^{-\nu\eta/2}e^{-\eta\Delta^2/2}\,\mathrm{d}\eta.\end{aligned}$$

Now we make the change of variable

$$\tau = \eta\left[\frac{\nu}{2}+\frac{1}{2}\Delta^2\right]^{-1}$$

which gives

$$\begin{aligned}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\left[\frac{\nu}{2}+\frac{1}{2}\Delta^2\right]^{-D/2-\nu/2} \\ &\qquad \int_0^\infty \tau^{D/2+\nu/2-1}e^{-\tau}\,\mathrm{d}\tau \\ &= \frac{\Gamma(\nu/2+d/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2}\end{aligned}$$

as required.

The correct normalization of the multivariate Student's t-distribution follows directly from the fact that the Gaussian and Gamma distributions are normalized. From (2.161) we have

$$\begin{aligned}\int \mathrm{St}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu\right)\,\mathrm{d}\mathbf{x} &= \iint \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\mathrm{Gam}\left(\eta|\nu/2,\nu/2\right)\,\mathrm{d}\eta\,\mathrm{d}\mathbf{x} \\ &= \iint \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1}\right)\,\mathrm{d}\mathbf{x}\,\mathrm{Gam}\left(\eta|\nu/2,\nu/2\right)\,\mathrm{d}\eta \\ &= \int \mathrm{Gam}\left(\eta|\nu/2,\nu/2\right)\,\mathrm{d}\eta = 1.\end{aligned}$$

**2.49** If we make the change of variable $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$, we can write

$$\mathbb{E}[\mathbf{x}] = \int \mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)\mathbf{x}\,\mathrm{d}\mathbf{x} = \int \mathrm{St}(\mathbf{z}|\mathbf{0},\boldsymbol{\Lambda},\nu)(\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}.$$

In the factor $(\mathbf{z}+\boldsymbol{\mu})$ the first term vanishes as a consequence of the fact that the zero-mean Student distribution is an even function of $\mathbf{z}$ that is $\mathrm{St}(-\mathbf{z}|\mathbf{0},\boldsymbol{\Lambda},\nu) = \mathrm{St}(-\mathbf{z}|\mathbf{0},\boldsymbol{\Lambda},\nu)$. This leaves the second term, which equals $\boldsymbol{\mu}$ since the Student distribution is normalized.

The covariance of the multivariate Student can be re-expressed by using the expression for the multivariate Student distribution as a convolution of a Gaussian with a

Gamma distribution given by (2.161) which gives

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \int \mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{x} \\ &= \int_0^\infty \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\eta\boldsymbol{\Lambda})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{x}\,\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\ &= \int_0^\infty \eta^{-1}\boldsymbol{\Lambda}^{-1}\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta\end{aligned}$$

where we have used the standard result for the covariance of a multivariate Gaussian. We now substitute for the Gamma distribution using (2.146) to give

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \frac{1}{\Gamma(\nu/2)}\left(\frac{\nu}{2}\right)^{\nu/2}\int_0^\infty e^{-\nu\eta/2}\eta^{\nu/2-2}\,\mathrm{d}\eta\boldsymbol{\Lambda}^{-1} \\ &= \frac{\nu}{2}\frac{\Gamma(\nu/2-2)}{\Gamma(\nu/2)}\boldsymbol{\Lambda}^{-1} \\ &= \frac{\nu}{\nu-2}\boldsymbol{\Lambda}^{-1}\end{aligned}$$

where we have used the integral representation for the Gamma function, together with the standard result $\Gamma(1+x) = x\Gamma(x)$.

The mode of the Student distribution is obtained by differentiation

$$\nabla_{\mathbf{x}}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{D/2}}\left[1+\frac{\Delta^2}{\nu}\right]^{-D/2-\nu/2-1}\frac{1}{\nu}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}).$$

Provided $\boldsymbol{\Lambda}$ is non-singular we therefore obtain

$$\mathrm{mode}[\mathbf{x}] = \boldsymbol{\mu}.$$

**2.50** Just like in univariate case (Exercise 2.47), we ignore the normalization coefficient, which leaves us with

$$\left[1+\frac{\Delta^2}{\nu}\right]^{-\nu/2-D/2} = \exp\left\{-\left(\frac{\nu}{2}+\frac{D}{2}\right)\ln\left[1+\frac{\Delta^2}{\nu}\right]\right\}$$

where $\Delta^2$ is the squared Mahalanobis distance given by

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}).$$

Again we make use of (117) to give

$$\exp\left\{-\left(\frac{\nu}{2}+\frac{D}{2}\right)\ln\left[1+\frac{\Delta^2}{\nu}\right]\right\} = \exp\left\{-\frac{\Delta^2}{2}+O(1/\nu)\right\}.$$

As in the univariate case, in the limit $\nu\to\infty$ this becomes, up to an overall constant, the same as a Gaussian distribution, here with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$; the univariate normalization argument also applies in the multivariate case.

**2.51** Using the relation (2.296) we have

$$1 = \exp(iA)\exp(-iA) = (\cos A + i\sin A)(\cos A - i\sin A) = \cos^2 A + \sin^2 A.$$

Similarly, we have

$$\begin{aligned}\cos(A-B) &= \Re\exp\{i(A-B)\} \\ &= \Re\exp(iA)\exp(-iB) \\ &= \Re(\cos A + i\sin A)(\cos B - i\sin B) \\ &= \cos A\cos B + \sin A\sin B.\end{aligned}$$

Finally

$$\begin{aligned}\sin(A-B) &= \Im\exp\{i(A-B)\} \\ &= \Im\exp(iA)\exp(-iB) \\ &= \Im(\cos A + i\sin A)(\cos B - i\sin B) \\ &= \sin A\cos B - \cos A\sin B.\end{aligned}$$

**2.52** Expressed in terms of $\xi$ the von Mises distribution becomes

$$p(\xi) \propto \exp\left\{m\cos(m^{-1/2}\xi)\right\}.$$

For large $m$ we have $\cos(m^{-1/2}\xi) = 1 - m^{-1}\xi^2/2 + O(m^{-2})$ and so

$$p(\xi) \propto \exp\left\{-\xi^2/2\right\}$$

and hence $p(\theta) \propto \exp\{-m(\theta-\theta_0)^2/2\}$.

**2.53** Using (2.183), we can write (2.182) as

$$\sum_{n=1}^{N}(\cos\theta_0\sin\theta_n - \cos\theta_n\sin\theta_0) = \cos\theta_0\sum_{n=1}^{N}\sin\theta_n - \sin\sum_{n=1}^{N}\cos\theta_n = 0.$$

Rearranging this, we get

$$\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n} = \frac{\sin\theta_0}{\cos\theta_0} = \tan\theta_0,$$

which we can solve w.r.t. $\theta_0$ to obtain (2.184).

**2.54** Differentiating the von Mises distribution (2.179) we have

$$p'(\theta) = -\frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta-\theta_0)\}\sin(\theta-\theta_0)$$

which vanishes when $\theta = \theta_0$ or when $\theta = \theta_0 + \pi \pmod{2\pi}$. Differentiating again we have

$$p''(\theta) = -\frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta-\theta_0)\}\left[\sin^2(\theta-\theta_0) + \cos(\theta-\theta_0)\right].$$

Since $I_0(m) > 0$ we see that $p''(\theta) < 0$ when $\theta = \theta_0$, which therefore represents a maximum of the density, while $p''(\theta) > 0$ when $\theta = \theta_0 + \pi \, (\mathrm{mod} 2\pi)$, which is therefore a minimum.

**2.55** **NOTE**: In PRML, equation (2.187), which will be the starting point for this solution, contains a typo. The "$-$" on the r.h.s. should be a "$+$", as is easily seen from (2.178) and (2.185).

From (2.169) and (2.184), we see that $\bar{\theta} = \theta_0^{\mathrm{ML}}$. Using this together with (2.168) and (2.177), we can rewrite (2.187) as follows:

$$\begin{aligned} A(m_{\mathrm{ML}}) &= \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{\mathrm{ML}} + \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{\mathrm{ML}} \\ &= \bar{r}\cos\bar{\theta}\cos\theta_0^{\mathrm{ML}} + \bar{r}\sin\bar{\theta}\sin\theta_0^{\mathrm{ML}} \\ &= \bar{r}\left(\cos^2\theta_0^{\mathrm{ML}} + \sin^2\theta_0^{\mathrm{ML}}\right) \\ &= \bar{r}. \end{aligned}$$

**2.56** We can most conveniently cast distributions into standard exponential family form by taking the exponential of the logarithm of the distribution. For the Beta distribution (2.13) we have

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left\{(a-1)\ln\mu + (b-1)\ln(1-\mu)\right\}$$

which we can identify as being in standard exponential form (2.194) with

$$h(\mu) = 1 \tag{118}$$

$$g(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{119}$$

$$\mathbf{u}(\mu) = \begin{pmatrix} \ln\mu \\ \ln(1-\mu) \end{pmatrix} \tag{120}$$

$$\boldsymbol{\eta}(a,b) = \begin{pmatrix} a-1 \\ b-1 \end{pmatrix}. \tag{121}$$

Applying the same approach to the gamma distribution (2.146) we obtain

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\exp\left\{(a-1)\ln\lambda - b\lambda\right\}.$$

from which it follows that

$$h(\lambda) = 1 \tag{122}$$
$$g(a,b) = \frac{b^a}{\Gamma(a)} \tag{123}$$
$$\mathbf{u}(\lambda) = \begin{pmatrix} \lambda \\ \ln\lambda \end{pmatrix} \tag{124}$$
$$\boldsymbol{\eta}(a,b) = \begin{pmatrix} -b \\ a-1 \end{pmatrix}. \tag{125}$$

Finally, for the von Mises distribution (2.179) we make use of the identity (2.178) to give

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m\cos\theta\cos\theta_0 + m\sin\theta\sin\theta_0\}$$

from which we find

$$h(\theta) = 1 \tag{126}$$
$$g(\theta_0, m) = \frac{1}{2\pi I_0(m)} \tag{127}$$
$$\mathbf{u}(\theta) = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{128}$$
$$\boldsymbol{\eta}(\theta_0, m) = \begin{pmatrix} m\cos\theta_0 \\ m\sin\theta_0 \end{pmatrix}. \tag{129}$$

**2.57** Starting from (2.43), we can rewrite the argument of the exponential as

$$-\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

The last term is indepedent of $\mathbf{x}$ but depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and so should go into $g(\boldsymbol{\eta})$. The second term is already an inner product and can be kept as is. To deal with the first term, we define the $D^2$-dimensional vectors $\mathbf{z}$ and $\boldsymbol{\lambda}$, which consist of the columns of $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ and $\boldsymbol{\Sigma}^{-1}$, respectively, stacked on top of each other. Now we can write the multivariate Gaussian distribution on the form (2.194), with

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\lambda} \end{bmatrix}$$
$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}$$
$$h(\mathbf{x}) = (2\pi)^{-D/2}$$
$$g(\boldsymbol{\eta}) = |\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right).$$

**2.58** Taking the first derivative of (2.226) we obtain, as in the text,

$$-\nabla \ln g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

Taking the gradient again gives

$$\begin{aligned} -\nabla\nabla \ln g(\boldsymbol{\eta}) &= g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}\, \mathrm{d}\mathbf{x} \\ &\quad + \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &= \mathbb{E}[\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathrm{T}}] - \mathbb{E}[\mathbf{u}(\mathbf{x})]\mathbb{E}[\mathbf{u}(\mathbf{x})^{\mathrm{T}}] \\ &= \mathrm{cov}[\mathbf{u}(\mathbf{x})] \end{aligned}$$

where we have used the result (2.226).

**2.59**

$$\begin{aligned} \int \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) \mathrm{d}x &= \frac{1}{\sigma} \int f(y) \frac{\mathrm{d}x}{\mathrm{d}y}\, \mathrm{d}y \\ &= \frac{1}{\sigma} \int f(y)\sigma\, \mathrm{d}y \\ &= \frac{\sigma}{\sigma} \int f(y)\, \mathrm{d}y = 1, \end{aligned}$$

since $f(x)$ integrates to 1.

**2.60** The value of the density $p(\mathbf{x})$ at a point $\mathbf{x}_n$ is given by $h_{j(n)}$, where the notation $j(n)$ denotes that data point $\mathbf{x}_n$ falls within region $j$. Thus the log likelihood function takes the form

$$\sum_{n=1}^{N} \ln p(\mathbf{x}_n) = \sum_{n=1}^{N} \ln h_{j(n)}.$$

We now need to take account of the constraint that $p(\mathbf{x})$ must integrate to unity. Since $p(\mathbf{x})$ has the constant value $h_i$ over region $i$, which has volume $\Delta_i$, the normalization constraint becomes $\sum_i h_i \Delta_i = 1$. Introducing a Lagrange multiplier $\lambda$ we then minimize the function

$$\sum_{n=1}^{N} \ln h_{j(n)} + \lambda \left( \sum_i h_i \Delta_i - 1 \right)$$

with respect to $h_k$ to give

$$0 = \frac{n_k}{h_k} + \lambda \Delta_k$$

where $n_k$ denotes the total number of data points falling within region $k$. Multiplying both sides by $h_k$, summing over $k$ and making use of the normalization constraint,

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

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

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

**2.61** From (2.246) we have

$$p(\mathbf{x}) = \frac{K}{NV(\rho)}$$

where $V(\rho)$ is the volume of a $D$-dimensional hypersphere with radius $\rho$, where in turn $\rho$ is the distance from $\mathbf{x}$ to its $K^{\text{th}}$ nearest neighbour in the data set. Thus, in polar coordinates, if we consider sufficiently large values for the radial coordinate $r$, we have

$$p(\mathbf{x}) \propto r^{-D}.$$

If we consider the integral of $p(\mathbf{x})$ and note that the volume element $\mathrm{d}\mathbf{x}$ can be written as $r^{D-1}\,\mathrm{d}r$, we get

$$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} \propto \int r^{-D} r^{D-1}\,\mathrm{d}r = \int r^{-1}\,\mathrm{d}r$$

which diverges logarithmically.

# Chapter 3 Linear Models for Regression

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

Using (3.6), we have

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

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

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

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

**3.2** We first write

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v} &= \boldsymbol{\Phi}\widetilde{\mathbf{v}} \\ &= \boldsymbol{\varphi}_1 \tilde{v}^{(1)} + \boldsymbol{\varphi}_2 \tilde{v}^{(2)} + \ldots + \boldsymbol{\varphi}_M \tilde{v}^{(M)} \end{aligned}$$

where $\boldsymbol{\varphi}_m$ is the $m$-th column of $\boldsymbol{\Phi}$ and $\widetilde{\mathbf{v}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{v}$. By comparing this with the least squares solution in (3.15), we see that

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

corresponds to a projection of $\mathbf{t}$ onto the space spanned by the columns of $\boldsymbol{\Phi}$. To see that this is indeed an orthogonal projection, we first note that for any column of $\boldsymbol{\Phi}$, $\boldsymbol{\varphi}_j$,

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\varphi}_j = \left[\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right]_j = \boldsymbol{\varphi}_j$$

and therefore

$$(\mathbf{y} - \mathbf{t})^{\mathrm{T}} \boldsymbol{\varphi}_j = (\boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} - \mathbf{t})^{\mathrm{T}} \boldsymbol{\varphi}_j = \mathbf{t}^{\mathrm{T}} \left(\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}} - \mathbf{I}\right)^{\mathrm{T}} \boldsymbol{\varphi}_j = 0$$

and thus $(\mathbf{y} - \mathbf{t})$ is ortogonal to every column of $\boldsymbol{\Phi}$ and hence is orthogonal to $\mathcal{S}$.

**3.3** If we define $\mathbf{R} = \mathrm{diag}(r_1, \ldots, r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

$$E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}\mathbf{R}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}).$$

Setting the derivative with respect to $\mathbf{w}$ to zero, and re-arranging, then gives

$$\mathbf{w}^{\star} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{t}$$

which reduces to the standard solution (3.15) for the case $\mathbf{R} = \mathbf{I}$.

If we compare (3.104) with (3.10)–(3.12), we see that $r_n$ can be regarded as a precision (inverse variance) parameter, particular to the data point $(\mathbf{x}_n, t_n)$, that either replaces or scales $\beta$.

Alternatively, $r_n$ can be regarded as an *effective* number of replicated observations of data point $(\mathbf{x}_n, t_n)$; this becomes particularly clear if we consider (3.104) with $r_n$ taking positive integer values, although it is valid for any $r_n > 0$.

**3.4** Let

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

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

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

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

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

since the $\epsilon_{ni}$ are all independent with variance $\sigma^2$.

From this and (3.106) we see that

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

as required.

**3.5** We can rewrite (3.30) as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**3.7** From Bayes' theorem we have

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}),$$

where the factors on the r.h.s. are given by (3.10) and (3.48), respectively. Writing this out in full, we get

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}) &\propto \left[\prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right)\right] \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right) \\
&\propto \exp\left(-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})\right) \\
&\quad \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right) \\
&= \exp\Big(-\frac{1}{2}\big(\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} - \beta\mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \beta\mathbf{t}^{\mathrm{T}}\mathbf{t} \\
&\quad \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big(\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} - \left(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right)^{\mathrm{T}}\mathbf{w} \\
&\quad -\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) + \beta\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\big)\Big) \\
&= \exp\left(-\frac{1}{2}\left(\mathbf{w}-\mathbf{m}_N\right)^{\mathrm{T}}\mathbf{S}_N^{-1}\left(\mathbf{w}-\mathbf{m}_N\right)\right) \\
&\quad \exp\left(-\frac{1}{2}\left(\beta\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N\right)\right)
\end{aligned}$$

where we have used (3.50) and (3.51) when completing the square in the last step. The first exponential corrsponds to the posterior, unnormalized Gaussian distribution over $\mathbf{w}$, while the second exponential is independent of $\mathbf{w}$ and hence can be absorbed into the normalization factor.

**3.8** Combining the prior

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

and the likelihood

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left(-\frac{\beta}{2}(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right) \tag{130}$$

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

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

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

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

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

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

with

$$\mathbf{S}_{N+1}^{-1}=\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}. \tag{131}$$

and

$$\mathbf{m}_{N+1}=\mathbf{S}_{N+1}(\mathbf{S}_N^{-1}\mathbf{m}_N+\beta\boldsymbol{\phi}_{N+1}t_{N+1}). \tag{132}$$

**3.9** Identifying (2.113) with (3.49) and (2.114) with (130), such that

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

$$\mathbf{y}\Rightarrow t_{N+1}\quad\mathbf{A}\Rightarrow\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathrm{T}}=\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\quad\mathbf{b}\Rightarrow\mathbf{0}\quad\mathbf{L}^{-1}\Rightarrow\beta\mathbf{I},$$

(2.116) and (2.117) directly give

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

where $\mathbf{S}_{N+1}$ and $\mathbf{m}_{N+1}$ are given by (131) and (132), respectively.

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

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

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

**3.11** From (3.59) we have

$$\sigma_{N+1}^2(\mathbf{x})=\frac{1}{\beta}+\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x}) \tag{133}$$

where $\mathbf{S}_{N+1}$ is given by (131). From (131) and (3.110) we get

$$\begin{aligned}
\mathbf{S}_{N+1}&=\left(\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\right)^{-1}\\
&=\mathbf{S}_N-\frac{\left(\mathbf{S}_N\boldsymbol{\phi}_{N+1}\beta^{1/2}\right)\left(\beta^{1/2}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\right)}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}\\
&=\mathbf{S}_N-\frac{\beta\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}.
\end{aligned}$$

Using this and (3.59), we can rewrite (133) as

$$\begin{aligned}\sigma^2_{N+1}(\mathbf{x}) &= \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{S}_N - \frac{\beta\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}\right)\boldsymbol{\phi}(\mathbf{x}) \\ &= \sigma^2_N(\mathbf{x}) - \frac{\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}{1+\beta\boldsymbol{\phi}^{\mathrm{T}}_{N+1}\mathbf{S}_N\boldsymbol{\phi}_{N+1}}. \qquad (134)\end{aligned}$$

Since $\mathbf{S}_N$ is positive definite, the numerator and denominator of the second term in (134) will be non-negative and positive, respectively, and hence $\sigma^2_{N+1}(\mathbf{x}) \leqslant \sigma^2_N(\mathbf{x})$.

**3.12** It is easiest to work in log space. The log of the posterior distribution is given by

$$\begin{aligned}\ln p(\mathbf{w},\beta|\mathbf{t}) &= \ln p(\mathbf{w},\beta) + \sum_{n=1}^{N}\ln p(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{M}{2}\ln\beta - \frac{1}{2}\ln|\mathbf{S}_0| - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ &\quad -b_0\beta + (a_0-1)\ln\beta \\ &\quad +\frac{N}{2}\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - t_n\}^2 + \text{const.}\end{aligned}$$

Using the product rule, the posterior distribution can be written as $p(\mathbf{w},\beta|\mathbf{t}) = p(\mathbf{w}|\beta,\mathbf{t})p(\beta|\mathbf{t})$. Consider first the dependence on $\mathbf{w}$. We have

$$\ln p(\mathbf{w}|\beta,\mathbf{t}) = -\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\left[\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \mathbf{S}_0^{-1}\right]\mathbf{w} + \mathbf{w}^{\mathrm{T}}\left[\beta\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right] + \text{const.}$$

Thus we see that $p(\mathbf{w}|\beta,\mathbf{t})$ is a Gaussian distribution with mean and covariance given by

$$\mathbf{m}_N = \mathbf{S}_N\left[\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right] \qquad (135)$$

$$\beta\mathbf{S}_N^{-1} = \beta\left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right). \qquad (136)$$

To find $p(\beta|\mathbf{t})$ we first need to complete the square over $\mathbf{w}$ to ensure that we pick up all terms involving $\beta$ (any terms independent of $\beta$ may be discarded since these will be absorbed into the normalization coefficient which itself will be found by inspection at the end). We also need to remember that a factor of $(M/2)\ln\beta$ will be absorbed by the normalisation factor of $p(\mathbf{w}|\beta,\mathbf{t})$. Thus

$$\begin{aligned}\ln p(\beta|\mathbf{t}) = &-\frac{\beta}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{\beta}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N \\ &+ \frac{N}{2}\ln\beta - b_0\beta + (a_0-1)\ln\beta - \frac{\beta}{2}\sum_{n=1}^{N}t_n^2 + \text{const.}\end{aligned}$$

We recognize this as the log of a Gamma distribution. Reading off the coefficients of $\beta$ and $\ln \beta$ we then have

$$a_N = a_0 + \frac{N}{2} \tag{137}$$

$$b_N = b_0 + \frac{1}{2}\left(\mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \sum_{n=1}^{N} t_n^2\right). \tag{138}$$

**3.13** Following the line of presentation from Section 3.3.2, the predictive distribution is now given by

$$p(t|\mathbf{x},\mathbf{t}) = \iint \mathcal{N}\left(t|\phi(\mathbf{x})^{\mathrm{T}}\mathbf{w},\beta^{-1}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N\right)\,\mathrm{d}\mathbf{w}$$
$$\mathrm{Gam}\left(\beta|a_N,b_N\right)\,\mathrm{d}\beta \tag{139}$$

We begin by performing the integral over $\mathbf{w}$. Identifying (2.113) with (3.49) and (2.114) with (3.8), using (3.3), such that

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{m}_N \quad \boldsymbol{\Lambda}^{-1} \Rightarrow \mathbf{S}_N$$
$$\mathbf{y} \Rightarrow t \quad \mathbf{A} \Rightarrow \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} = \boldsymbol{\phi}^{\mathrm{T}} \quad \mathbf{b} \Rightarrow 0 \quad \mathbf{L}^{-1} \Rightarrow \beta^{-1},$$

(2.115) and (136) give

$$\begin{aligned} p(t|\beta) &= \mathcal{N}\left(t|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N,\beta^{-1}+\boldsymbol{\phi}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}\right) \\ &= \mathcal{N}\left(t|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N,\beta^{-1}\left(1+\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{S}_0+\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}\right)\right). \end{aligned}$$

Substituting this back into (139) we get

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) = \int \mathcal{N}\left(t|\boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N,\beta^{-1}s\right)\mathrm{Gam}\left(\beta|a_N,b_N\right)\,\mathrm{d}\beta,$$

where we have defined

$$s = 1 + \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{S}_0+\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi})^{-1}\boldsymbol{\phi}.$$

We can now use (2.158)– (2.160) to obtain the final result:

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) = \mathrm{St}\left(t|\mu,\lambda,\nu\right)$$

where

$$\mu = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{m}_N \quad \lambda = \frac{a_N}{b_N}s^{-1} \quad \nu = 2a_N.$$

**3.14** For $\alpha = 0$ the covariance matrix $\mathbf{S}_N$ becomes

$$\mathbf{S}_N = (\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}. \tag{140}$$

Let us define a new set of orthonormal basis functions given by linear combinations of the original basis functions so that

$$\boldsymbol{\psi}(\mathbf{x}) = \mathbf{V}\boldsymbol{\phi}(\mathbf{x}) \tag{141}$$

where $\mathbf{V}$ is an $M \times M$ matrix. Since both the original and the new basis functions are linearly independent and span the same space, this matrix must be invertible and hence

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{V}^{-1}\boldsymbol{\psi}(\mathbf{x}).$$

For the data set $\{\mathbf{x}_n\}$, (141) and (3.16) give

$$\boldsymbol{\Psi} = \boldsymbol{\Phi}\mathbf{V}^{\mathrm{T}}$$

and consequently

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}\mathbf{V}^{-\mathrm{T}}$$

where $\mathbf{V}^{-\mathrm{T}}$ denotes $(\mathbf{V}^{-1})^{\mathrm{T}}$. Orthonormality implies

$$\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} = \mathbf{I}.$$

Note that $(\mathbf{V}^{-1})^{\mathrm{T}} = (\mathbf{V}^{\mathrm{T}})^{-1}$ as is easily verified. From (140), the covariance matrix then becomes

$$\mathbf{S}_N = \beta^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1} = \beta^{-1}(\mathbf{V}^{-\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\mathbf{V}^{-1})^{-1} = \beta^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}.$$

Here we have used the orthonormality of the $\psi_i(\mathbf{x})$. Hence the equivalent kernel becomes

$$k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}')$$

as required. From the orthonormality condition, and setting $j = 1$, it follows that

$$\sum_{n=1}^{N} \psi_i(\mathbf{x}_n)\psi_1(\mathbf{x}_n) = \sum_{n=1}^{N} \psi_i(\mathbf{x}_n) = \delta_{i1}$$

where we have used $\psi_1(\mathbf{x}) = 1$. Now consider the sum

$$\begin{aligned}\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) &= \sum_{n=1}^{N} \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_n) = \sum_{n=1}^{N}\sum_{i=1}^{M} \psi_i(\mathbf{x})\psi_i(\mathbf{x}_n) \\ &= \sum_{i=1}^{M} \psi_i(\mathbf{x})\delta_{i1} = \psi_1(\mathbf{x}) = 1\end{aligned}$$

which proves the summation constraint as required.

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

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

**3.16** The likelihood function is a product of independent univariate Gaussians and so can be written as a joint Gaussian distribution over $\mathbf{t}$ with diagonal covariance matrix in the form

$$p(\mathbf{t}|\mathbf{w},\beta) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}_N). \tag{142}$$

Identifying (2.113) with the prior distribution $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$ and (2.114) with (142), such that

$$\mathbf{x} \Rightarrow \mathbf{w} \quad \boldsymbol{\mu} \Rightarrow \mathbf{0} \quad \mathbf{\Lambda}^{-1} \Rightarrow \alpha^{-1}\mathbf{I}_M$$

$$\mathbf{y} \Rightarrow \mathbf{t} \quad \mathbf{A} \Rightarrow \mathbf{\Phi} \quad \mathbf{b} \Rightarrow \mathbf{0} \quad \mathbf{L}^{-1} \Rightarrow \beta^{-1}\mathbf{I}_N,$$

(2.115) gives

$$p(\mathbf{t}|\alpha,\beta) = \mathcal{N}(\mathbf{t}|\mathbf{0},\beta^{-1}\mathbf{I}_N + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}).$$

Taking the log we obtain

$$\begin{aligned} \ln p(\mathbf{t}|\alpha,\beta) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\beta^{-1}\mathbf{I}_N + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\right| \\ - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta^{-1}\mathbf{I}_N + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\right)\mathbf{t}. \end{aligned} \tag{143}$$

Using the result (C.14) for the determinant we have

$$\begin{aligned} \left|\beta^{-1}\mathbf{I}_N + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\right| &= \beta^{-N}\left|\mathbf{I}_N + \beta\alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\right| \\ &= \beta^{-N}\left|\mathbf{I}_M + \beta\alpha^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right| \\ &= \beta^{-N}\alpha^{-M}\left|\alpha\mathbf{I}_M + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right| \\ &= \beta^{-N}\alpha^{-M}\left|\mathbf{A}\right| \end{aligned}$$

where we have used (3.81). Next consider the quadratic term in $\mathbf{t}$ and make use of the identity (C.7) together with (3.81) and (3.84) to give

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

where in the last step, we have exploited results from Solution 3.18. Substituting for the determinant and the quadratic term in (143) we obtain (3.86).

**3.17** Using (3.11), (3.12) and (3.52) together with the definition for the Gaussian, (2.43), we can rewrite (3.77) as follows:

$$\begin{aligned} p(\mathbf{t}|\alpha,\beta) &= \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w} \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\beta E_D(\mathbf{w})\right)\exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)\,\mathrm{d}\mathbf{w} \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-E(\mathbf{w})\right)\,\mathrm{d}\mathbf{w}, \end{aligned}$$

where $E(\mathbf{w})$ is defined by (3.79).

**3.18** We can rewrite (3.79)

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

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

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

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

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

as required.

**3.19** From (3.80) we see that the integrand of (3.85) is an unnormalized Gaussian and hence integrates to the inverse of the corresponding normalizing constant, which can be read off from the r.h.s. of (2.43) as

$$(2\pi)^{M/2}\,|\mathbf{A}^{-1}|^{1/2}.$$

Using (3.78), (3.85) and the properties of the logarithm, we get

$$\begin{aligned}\ln p(\mathbf{t}|\alpha,\beta) &= \frac{M}{2}(\ln\alpha - \ln(2\pi)) + \frac{N}{2}(\ln\beta - \ln(2\pi)) + \ln\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \\ &= \frac{M}{2}(\ln\alpha - \ln(2\pi)) + \frac{N}{2}(\ln\beta - \ln(2\pi)) - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| + \frac{M}{2}\ln(2\pi)\end{aligned}$$

which equals (3.86).

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

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

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

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

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

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

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

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

for $M$ and re-arranging.

**3.21** The eigenvector equation for the $M \times M$ real, symmetric matrix $\mathbf{A}$ can be written as

$$\mathbf{A}\mathbf{u}_i = \eta_i\mathbf{u}_i$$

where $\{\mathbf{u}_i\}$ are a set of $M$ orthonormal vectors, and the $M$ eigenvalues $\{\eta_i\}$ are all real. We first express the left hand side of (3.117) in terms of the eigenvalues of $\mathbf{A}$. The log of the determinant of $\mathbf{A}$ can be written as

$$\ln|\mathbf{A}| = \ln\prod_{i=1}^{M}\eta_i = \sum_{i=1}^{M}\ln\eta_i.$$

Taking the derivative with respect to some scalar $\alpha$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}| = \sum_{i=1}^{M}\frac{1}{\eta_i}\frac{\mathrm{d}}{\mathrm{d}\alpha}\eta_i.$$

We now express the right hand side of (3.117) in terms of the eigenvector expansion and show that it takes the same form. First we note that $\mathbf{A}$ can be expanded in terms of its own eigenvectors to give

$$\mathbf{A} = \sum_{i=1}^{M}\eta_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$$

and similarly the inverse can be written as

$$\mathbf{A}^{-1} = \sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}.$$

Thus we have

$$\begin{aligned}
\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right) &= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{j=1}^{M}\eta_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right) \\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\left\{\sum_{j=1}^{M}\frac{\mathrm{d}\eta_j}{\mathrm{d}\alpha}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} + \eta_j\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}} + \mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right\}\right) \\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\frac{\mathrm{d}\eta_j}{\mathrm{d}\alpha}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\right) \\
&\quad + \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\eta_j\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}} + \mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right) \qquad (144)
\end{aligned}$$

where $\mathbf{b}_j = \mathrm{d}\mathbf{u}_j/\mathrm{d}\alpha$. Using the properties of the trace and the orthognality of

eigenvectors, we can rewrite the second term as

$$\begin{aligned}
&\mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\eta_j\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}}+\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right)\\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\frac{1}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}2\eta_j\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{2\eta_j}{\eta_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\left(\mathbf{b}_j\mathbf{u}_j^{\mathrm{T}}+\mathbf{u}_j\mathbf{b}_j^{\mathrm{T}}\right)\right)\\
&= \mathrm{Tr}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{i}^{M}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right).
\end{aligned}$$

However,

$$\sum_{i}^{M}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}=\mathbf{I}$$

which is constant and thus its derivative w.r.t. $\alpha$ will be zero and the second term in (144) vanishes.

For the first term in (144), we again use the properties of the trace and the orthognality of eigenvectors to obtain

$$\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)=\sum_{i=1}^{M}\frac{1}{\eta_i}\frac{\mathrm{d}\eta_i}{\mathrm{d}\alpha}.$$

We have now shown that both the left and right hand sides of (3.117) take the same form when expressed in terms of the eigenvector expansion. Next, we use (3.117) to differentiate (3.86) w.r.t. $\alpha$, yielding

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln p(\mathbf{t}|\alpha\beta) &= \frac{M}{2}\frac{1}{\alpha}-\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\frac{1}{2}\mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)\\
&= \frac{1}{2}\left(\frac{M}{\alpha}-\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\mathrm{Tr}\left(\mathbf{A}^{-1}\right)\right)\\
&= \frac{1}{2}\left(\frac{M}{\alpha}-\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N-\sum_{i}\frac{1}{\lambda_i+\alpha}\right)
\end{aligned}$$

which we recognize as the r.h.s. of (3.89), from which (3.92) can be derived as detailed in Section 3.5.2, immediately following (3.89).

**3.22** Using (3.82) and (3.93)—the derivation of latter is detailed in Section 3.5.2—we get the derivative of (3.86) w.r.t. $\beta$ as the r.h.s. of (3.94). Rearranging this, collecting the $\beta$-dependent terms on one side of the equation and the remaining term on the other, we obtain (3.95).

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

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

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

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

Now we are ready to do the integration, first over $\mathbf{w}$ and then $\beta$, and re-arrange the

terms to obtain the desired result

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

**3.24** Substituting the r.h.s. of (3.10), (3.112) and (3.113) into (3.119), we get

$$p(\mathbf{t}) = \frac{\mathcal{N}\left(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_0,\beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}\left(\beta|a_0,b_0\right)}{\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}\left(\beta|a_N,b_N\right)}. \tag{145}$$

Using the definitions of the Gaussian and Gamma distributions, we can write this as

$$\begin{aligned} &\left(\frac{\beta}{2\pi}\right)^{N/2}\exp\left(-\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{w}\|^2\right) \\ &\qquad\left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_0|^{1/2}\exp\left(-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right) \\ &\qquad\qquad\Gamma(a_0)^{-1}b_0^{a_0}\beta^{a_0-1}\exp(-b_0\beta) \\ &\left\{\left(\frac{\beta}{2\pi}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\left(-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right)\right. \\ &\qquad\qquad\left.\Gamma(a_N)^{-1}b_N^{a_N}\beta^{a_N-1}\exp(-b_N\beta)\right\}^{-1}. \end{aligned} \tag{146}$$

Concentrating on the factors corresponding to the denominator in (145), i.e. the fac-

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

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

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

# Chapter 4 Linear Models for Classification

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

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

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

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

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

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

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

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

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

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

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

We can take the derivative of (147) w.r.t. $\mathbf{w}_0$, giving

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

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

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

where

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

If we subsitute (148) into (147), we get

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

where

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

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

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

where we have defined $\widehat{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$ and $\widehat{\mathbf{T}} = \mathbf{T} - \overline{\mathbf{T}}$.

Now consider the prediction for a new input vector $\mathbf{x}^\star$,

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

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

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

Therefore, applying (4.157) to (149), we obtain

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

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

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

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

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

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

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

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

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

From (4.22) we can construct the Lagrangian function

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

Taking the gradient of $L$ we obtain

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

and setting this gradient to zero gives

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

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

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

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

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

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

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

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

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

We then use the identity

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

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

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

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

**4.7** From (4.59) we have

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

The inverse of the logistic sigmoid is easily found as follows

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

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

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

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

**4.9** The likelihood function is given by

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

and taking the logarithm, we obtain

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

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

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

Setting the derivative with respect to $\pi_k$ equal to zero, we obtain

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

Re-arranging then gives

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

which is linear in $\phi_{ml}$.

**4.12** Differentiating (4.59) we obtain

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

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

$$\frac{\partial E}{\partial y_n} = \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \tag{158}$$

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

$$= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \tag{159}$$

$$= \frac{y_n - t_n}{y_n(1-y_n)}. \tag{160}$$

From (4.88), we see that

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

Finally, we have

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

where $\nabla$ denotes the gradient with respect to $\mathbf{w}$. Combining (160), (161) and (162) using the chain rule, we obtain

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

as required.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**4.17** From (4.104) we have

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

Combining these results we obtain (4.106).

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

From (4.108) we have

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

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

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

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

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

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

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

Also

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

From (4.115) and (4.116) we have

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

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

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

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

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

Then using (165)–(168) we have

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

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

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

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

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

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

where

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

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

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

and hence

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

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

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

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

can use (2.42) to rewrite (4.114) as

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

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

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

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

Combining this with (169), we get

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

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

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

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

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

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

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

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

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

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

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

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

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

We now again invoke the broad prior assumption, allowing us to neglect the second term on the right hand side of (169) relative to the first term.

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

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

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

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

Combining this with the properties of the determinant, we have

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

where $M$ is the dimensionality of $\boldsymbol{\theta}$. Note that we are assuming that $\widehat{\mathbf{H}}$ has full rank $M$. Finally, using this result together (169), we obtain (4.139) by dropping the $\ln|\widehat{\mathbf{H}}|$ since this $O(1)$ compared to $\ln N$.

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

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

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

Section 2.3.2. Denoting the unit vector

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

we have

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

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

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

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

**4.25** From (4.88) we have that

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

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

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

Setting this equal to (170), we see that

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

This is illustrated in Figure 4.9.

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

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

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

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

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

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

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

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

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

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

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

# Chapter 5 Neural Networks

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

See Solution 3.1.

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

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