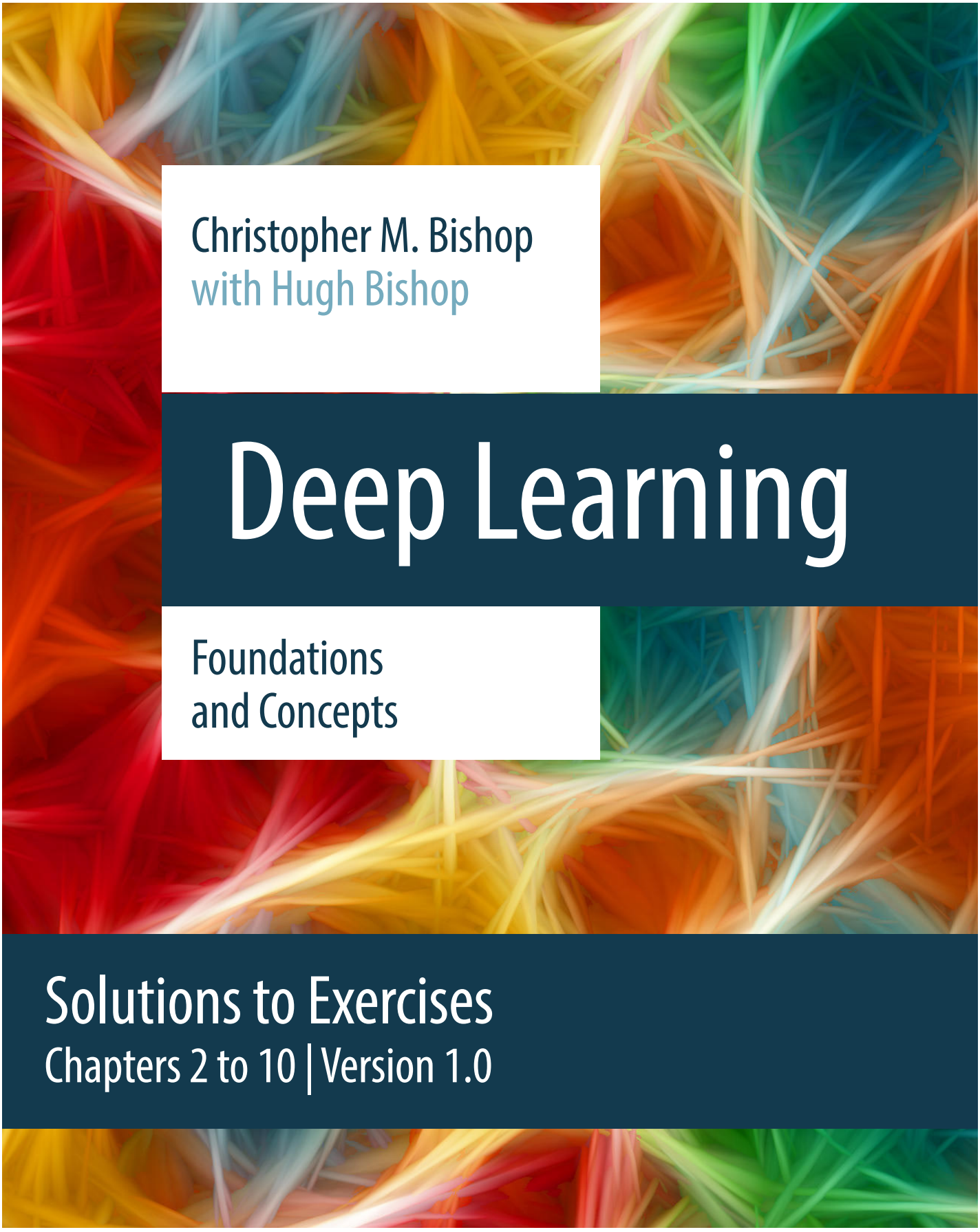
Christopher M. Bishop
with Hugh Bishop
Deep Learning
Foundations
and Concepts
Solutions to Exercises
Chapters 2 to 10 | Version 1.0

This is version 1.0 of the solutions manual for *Deep Learning: Foundations and Concepts* by C. M. Bishop and H. Bishop (Springer, 2024) and contains worked solutions for exercises in Chapters 2 to 10. A full solutions manual including solutions to all exercises in the book will be released soon. The most recent version of the solutions manual, along with a free-to-use digital version of the book as well as downloadable versions of the figures in PDF and JPEG formats, can be found on the book web site:

`https://www.bishopbook.com`

If you have any feedback on the book or associated materials including this solutions manual, please send email to the authors at

`feedback@bishopbook.com`

Contents

Chapter 2 Probabilities

2.1 We first compute $p(T = 1)$ by modifying (2.20)

$$\begin{aligned} p(T=1) &= p(T=1|C=0)p(C=0) + p(T=1|C=1)p(C=1) \\ &= \frac{3}{100} \times \frac{999}{1,000} + \frac{90}{100} \times \frac{1}{1,000} = \frac{3,087}{100,000} = 0.03087. \end{aligned} \tag{1}$$

Then we evaluate $p(C = 1|T = 1)$ by modifying (2.22)

$$p(C=1|T=1) = \frac{p(T=1|C=1)p(C=1)}{p(T=1)} \tag{2}$$

$$= \frac{90}{100} \times \frac{1}{1,000} \times \frac{100,000}{3,087} = \frac{90}{3,087} \simeq 0.029 \tag{3}$$

Hence we see that the probability of having cancer, even after a positive test, remains very small.

2.2 Note that each of the numbers 0, 1, 2, 3, 4, 5, and 6 appears on only one of the dice, which means that when we roll one die against another, there can never be a draw.

Look first at the red die, and notice that it has four copies of the number 2 and two copies of the number 6. Two-thirds of the time, when we roll the red die it will give a 2, and one third of the time it will give a 6. Therefore, if we roll the red die against the yellow die (which always gives a 3), the yellow die will, on average, win two-thirds of the time, and will lose one-third of the time.

Now look at the blue die, and notice that it has four copies of the number 4, and two copies of the number 0. When we roll it against the yellow die, it will therefore give a 4 two thirds of the time, in which case it wins, and a 0 one-third of the time, in which case it loses.

Next consider the green die versus the blue die. The green die has three copies of the number 1 and three copies of the number 5. To work out the probability that the green die will win we first note that there is a probability of 1/2 that the green die will give a 5, in which case it is certain to win against the blue die. Likewise, there is a probability of 1/2 that the green die will give a 1, in which case there is a probability of 1/3 that it will win. The overall probability that the green die will win is then given by multiplying the probabilities:

$$\left(\frac{1}{2} \times 1\right) + \left(\frac{1}{2} \times \frac{1}{3}\right) = \frac{2}{3}. \tag{4}$$

Finally, consider the probability of the red die winning against the green die. There is a probability of 1/3 that the red die will produce a 6, in which case it is certain that the red die will win. There is similarly a probability of 2/3 that the red die will

produce a 2 in which case there is a 1/2 chance that the red die will win. The overall probability of the red die winning is again obtained by multiplying the probabilities:

$$\left(\frac{1}{3}\times 1\right)+\left(\frac{2}{3}\times\frac{1}{2}\right)=\frac{2}{3}. \tag{5}$$

For more information on these dice, see:
`microsoft.com/en-us/research/project/non-transitive-dice/`

2.3 Using the sum and product rules of probability we can write the desired distribution in the form

$$p(\mathbf{y})=\iint p(\mathbf{y}|\mathbf{u},\mathbf{v})p_{\mathbf{u}}(\mathbf{u})p_{\mathbf{v}}(\mathbf{v})\,\mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{v}. \tag{6}$$

Since $\mathbf{y}$ is a deterministic function of $\mathbf{u}$ and $\mathbf{v}$, its conditional distribution is given by a Dirac delta function in the form

$$p(\mathbf{y}|\mathbf{u},\mathbf{v})=\delta(\mathbf{y}-\mathbf{u}-\mathbf{v}). \tag{7}$$

Substituting (7) into (6) allows us to perform the integration over $\mathbf{v}$ to give

$$p(\mathbf{y})=\int p_{\mathbf{u}}(\mathbf{u})p_{\mathbf{v}}(\mathbf{y}-\mathbf{u})\,\mathrm{d}\mathbf{u} \tag{8}$$

as required.

2.4 If we integrate the uniform distribution (2.33) over x we obtain

$$\int_{-\infty}^{\infty}p(x)\,\mathrm{d}x=\int_{c}^{d}\frac{1}{d-c}\,\mathrm{d}x=\frac{d-c}{d-c}=1 \tag{9}$$

and hence this distribution is normalized. 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 (2.46) to give

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

2.5 Integrating the exponential distribution (2.34) over $0 \leqslant x \leqslant \infty$ we obtain

$$\begin{aligned}\int_0^\infty p(x|\lambda)\,\mathrm{d}x &= \int_0^\infty \lambda \exp(-\lambda x)\,\mathrm{d}x \\ &= \int_0^\infty \exp(-y)\,\mathrm{d}y \\ &= [-\exp(-y)]_0^\infty \\ &= 1 \end{aligned} \tag{10}$$

where we have used the change of variables $y = \lambda x$. Hence, the exponential distribution is normalized. Likewise, if we integrate the Laplace distribution (2.35) we obtain

$$\begin{aligned}\int_{-\infty}^\infty p(x|\mu,\lambda)\,\mathrm{d}x &= \int_{-\infty}^\infty \frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)\mathrm{d}x \\ &= \int_{-\infty}^\mu \frac{1}{2\gamma}\exp\left(\frac{x-\mu}{\gamma}\right)\mathrm{d}x + \int_\mu^\infty \frac{1}{2\gamma}\exp\left(-\frac{x-\mu}{\gamma}\right)\mathrm{d}x \\ &= \int_{-\infty}^0 \frac{1}{2}\exp\left(z\right)\,\mathrm{d}z + \int_0^\infty \frac{1}{2}\exp\left(-z\right)\,\mathrm{d}z \\ &= \left[\frac{1}{2}\exp(z)\right]_{-\infty}^0 + \left[-\frac{1}{2}\exp(-z)\right]_0^\infty \\ &= \frac{1}{2} + \frac{1}{2} = 1 \end{aligned} \tag{11}$$

where we have made the substitution $z = (x-\mu)/\gamma$ in each of the two integrals. Hence we see that the Laplace distribution is also normalized.

2.6 Integrating the empirical density (2.37) we obtain

$$\begin{aligned}\int_{-\infty}^\infty p(x|\mathcal{D})\,\mathrm{d}x &= \frac{1}{N}\sum_{n=1}^N \int_{-\infty}^\infty \delta(x-x_n)\,\mathrm{d}x \\ &= \frac{1}{N}\sum_{n=1}^N \int_{-\infty}^\infty \delta(y)\,\mathrm{d}y \\ &= \frac{1}{N}\sum_{n=1}^N 1 = 1 \end{aligned} \tag{12}$$

as required. Here we have substituted $y = x - x_n$ into the nth integral in the summation and then used the definition of the Dirac delta function. This result is easily generalized to a multi-dimensional data variable $\mathbf{x}$.

2.7 If we substitute the empirical distribution (2.37) into the definition of the expectation

with respect to a continuous density given by (2.39) we obtain

$$\begin{aligned}\mathbb{E}[f] &= \int_{-\infty}^{\infty} p(x)f(x)\,\mathrm{d}x \\ &\simeq \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\delta(x-x_n)f(x)\,\mathrm{d}x \\ &= \frac{1}{N}\sum_{n=1}^{N}\int_{-\infty}^{\infty}\delta(y_n)f(y_n+x_n)\,\mathrm{d}y_n \\ &= \frac{1}{N}\sum_{n=1}^{N}f(x_n) \qquad (13)\end{aligned}$$

as required. Here we have used the change of variable $y_n = x - x_n$ separately in each of the integrals, along with the property of the Dirac delta function $\delta(y_n)$ integrating to unity with the only non-zero contribution coming from $y_n = 0$.

2.8 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.

2.9 The definition of covariance is given by (2.47) as

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

Using (2.38) 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 (2.38) replaced by (2.39) and the sums replaced by integrals.

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

$$\begin{aligned}\mathbb{E}[x+z] &= \iint (x+z)p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z \qquad (14)\\ &= \int xp(x)\,\mathrm{d}x + \int zp(z)\,\mathrm{d}z \qquad (15)\\ &= \mathbb{E}[x] + \mathbb{E}[z]. \qquad (16)\end{aligned}$$

Similarly for the variances, we first note that

$$(x + z - \mathbb{E}[x + z])^2 = (x - \mathbb{E}[x])^2 + (z - \mathbb{E}[z])^2 + 2(x - \mathbb{E}[x])(z - \mathbb{E}[z]) \quad (17)$$

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

$$\begin{aligned} \mathrm{var}[x + z] &= \iint (x + z - \mathbb{E}[x + z])^2 p(x)p(z)\,\mathrm{d}x\,\mathrm{d}z \\ &= \int (x - \mathbb{E}[x])^2 p(x)\,\mathrm{d}x + \int (z - \mathbb{E}[z])^2 p(z)\,\mathrm{d}z \\ &= \mathrm{var}(x) + \mathrm{var}(z). \end{aligned} \quad (18)$$

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

2.11 Using the definition (2.39) of expectation we have

$$\begin{aligned} \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right] &= \int p(y) \int p(x|y)x\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint p(x, y)x\,\mathrm{d}x\,\mathrm{d}y \\ &= \int p(x)x\,\mathrm{d}x = \mathbb{E}\left[x\right] \end{aligned} \quad (19)$$

where we have used the product rule of probability $p(x|y)p(y) = p(x, y)$. Now we make use of the result (2.46) to write

$$\begin{aligned} &\mathbb{E}_y\left[\mathrm{var}_x[x|y]\right] + \mathrm{var}_y\left[\mathbb{E}_x[x|y]\right] = \\ &\quad \mathbb{E}_y\left[\mathbb{E}_x[x^2|y]\right] - \mathbb{E}_y\left[\mathbb{E}_x[x|y]^2\right] + \mathbb{E}_y\left[\mathbb{E}_x[x|y]^2\right] - \mathbb{E}_y\left[\mathbb{E}_x[x|y]\right]^2. \end{aligned} \quad (20)$$

We now note that the second and third terms on the right-hand side of (20) cancel. The first term on the right-hand side of (20) can be written as

$$\begin{aligned} \mathbb{E}_y\left[\mathbb{E}_x[x^2|y]\right] &= \int p(y) \int p(x|y)x^2\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint p(x, y)x^2\,\mathrm{d}x\,\mathrm{d}y \\ &= \int p(x)x^2\,\mathrm{d}x = \mathbb{E}\left[x^2\right]. \end{aligned} \quad (21)$$

Likewise, we can again make use of the result (2.46) to write the fourth term on the right-hand side of (20) in the form $\mathbb{E}\left[x\right]^2$. Hence we have

$$\mathbb{E}_y\left[\mathrm{var}_x[x|y]\right] + \mathrm{var}_y\left[\mathbb{E}_x[x|y]\right] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 = \mathrm{var}\left[x\right] \quad (22)$$

as required.

2.12 The transformation from Cartesian to polar coordinates is defined by

$$x = r\cos\theta \tag{23}$$
$$y = r\sin\theta \tag{24}$$

and hence we have $x^2 + y^2 = r^2$ where we have used the well-known trigonometric result (3.127). 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 (3.127). Thus the double integral in (2.125) becomes

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2}\right) r\,\mathrm{d}r\,\mathrm{d}\theta \tag{25}$$
$$= 2\pi\int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2}\right)\frac{1}{2}\,\mathrm{d}u \tag{26}$$
$$= \pi\left[\exp\left(-\frac{u}{2\sigma^2}\right)(-2\sigma^2)\right]_0^{\infty} \tag{27}$$
$$= 2\pi\sigma^2 \tag{28}$$

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.

2.13 From the definition (2.49) of the univariate Gaussian distribution, we have

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}x\,\mathrm{d}x. \tag{29}$$

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

$$\mathbb{E}[x] = \int_{-\infty}^{\infty}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}(y+\mu)\,\mathrm{d}y. \tag{30}$$

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 ∞ and then show that these two integrals cancel). In the second term, μ is a constant and pulls outside the integral, leaving a normalized Gaussian distribution which integrates to 1, and so we obtain (2.52).

To derive (2.53) we first substitute the expression (2.49) for the normal distribution into the normalization result (2.51) and re-arrange to obtain

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \mathrm{d}x = \left(2\pi\sigma^2\right)^{1/2}. \tag{31}$$

We now differentiate both sides of (31) with respect to σ^2 and then re-arrange to obtain

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 \,\mathrm{d}x = \sigma^2 \tag{32}$$

which directly shows that

$$\mathbb{E}[(x-\mu)^2] = \mathrm{var}[x] = \sigma^2. \tag{33}$$

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 (2.52) then gives (2.53) as required.

Finally, (2.54) follows directly from (2.52) and (2.53)

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

2.14 For the univariate case, we simply differentiate (2.49) 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 $x = \mu$.

2.15 We use ℓ to denote $\ln p(\mathbf{X}|\mu,\sigma^2)$ from (2.56). 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 μ 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 σ^2/N we get (2.57).

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 μ_{ML} for μ we get (2.58).

2.16 If $m = n$ then $x_n x_m = x_n^2$ and using (2.53) 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 (2.52). Combining these two results we obtain (2.128).

Next we have

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n] = \mu \tag{34}$$

using (2.52).

Finally, consider $\mathbb{E}[\sigma^2_{\mathrm{ML}}]$. From (2.57) and (2.58), and making use of (2.128), we have

$$\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \frac{1}{N}\sum_{m=1}^{N}x_m\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right] \\
&= \left\{\mu^2 + \sigma^2 - 2\left(\mu^2 + \frac{1}{N}\sigma^2\right) + \mu^2 + \frac{1}{N}\sigma^2\right\} \\
&= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned} \tag{35}$$

as required.

2.17 From the definition (2.61), and making use of (2.52) and (2.53), we have

$$\begin{aligned}\mathbb{E}\left[\widehat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - 2x_n\mu + \mu^2\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2 + \sigma^2 - 2\mu\mu + \mu^2\right) \\ &= \sigma^2 \end{aligned} \tag{36}$$

as required.

2.18 Differentiating (2.66) with respect to σ^2 gives

$$\frac{\partial}{\partial\sigma^2}\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\sigma^2) = \frac{1}{2\sigma^4}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2 - \frac{N}{2}\frac{1}{\sigma^2}. \tag{37}$$

Setting the derivative to zero and rearranging then gives

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}_{\mathrm{ML}})-t_n\}^2 \tag{38}$$

as required.

2.19 If we assume that the function $y = f(x)$ is *strictly* monotonic, which is necessary to exclude the possibility for spikes of infinite density in $p(y)$, we are guaranteed that the inverse function $x = f^{-1}(y)$ exists. We can then use (2.71) to write

$$p(y) = q(f^{-1}(y))\left|\frac{\mathrm{d}f^{-1}}{\mathrm{d}y}\right|. \tag{39}$$

Since the only restriction on f is that it is monotonic, it can distribute the probability mass over x arbitrarily over y. This is illustrated in Figure 2.12 on page 44, as a part of Solution **??**. From (39) we see directly that

$$|f'(x)| = \frac{q(x)}{p(f(x))}.$$

2.20 The Jacobian matrix for the transformation from (x_1, x_2) to (y_1, y_2) is defined by

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial y_1}{\partial x_1} & \dfrac{\partial y_1}{\partial x_2} \\ \dfrac{\partial y_2}{\partial x_1} & \dfrac{\partial y_2}{\partial x_2} \end{bmatrix}. \tag{40}$$

For the specific transformation defined by (2.78) and (2.79) we have

$$\frac{\partial y_1}{\partial x_1} = 1 + 5\,\mathrm{sech}^2(x_1) \tag{41}$$

$$\frac{\partial y_1}{\partial x_2} = 0 \tag{42}$$

$$\frac{\partial y_2}{\partial x_1} = x_1^2 \tag{43}$$

$$\frac{\partial y_2}{\partial x_2} = 1 + 5\,\mathrm{sech}^2(x_2). \tag{44}$$

2.21 From the discussion of the introduction of Section 2.5, 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 $h(p) \propto \ln(p)$.

2.22 We wish to maximize the entropy (2.86) subject to the constraint that the probabilities sum to one, so that

$$\sum_i p(x_i) = 1. \tag{45}$$

We introduce a Lagrange multiplier λ to enforce this constraint and hence we maximize

$$-\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right). \tag{46}$$

Setting the derivative with respect to $p(x_i)$ to zero gives

$$-\ln p(x_i) - 1 + \lambda = 0. \tag{47}$$

Solving for $p(x_i)$ we obtain

$$p(x_i) = \exp\left(-1 + \lambda\right). \tag{48}$$

Since the right-hand side does not depend on i, this shows that the probabilities are all equal. From (45) it then follows that $p(x_i) = 1/M$. Substituting this result into (2.86) then shows that the value of the entropy at its maximum is equal to $\ln M$.

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

$$H(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i) = \sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)}. \tag{49}$$

The function $\ln(x)$ is concave$\frown$ and so we can apply Jensen's inequality in the form (2.102) but with the inequality reversed, so that

$$H(x) \leqslant \ln \left(\sum_{i=1}^{M} p(x_i) \frac{1}{p(x_i)} \right) = \ln M. \tag{50}$$

2.24 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 B. 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 (B.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 (2.97).

To eliminate the Lagrange multipliers we substitute (2.97) into each of the three constraints (2.93), (2.94) and (2.95) in turn. The solution is most easily obtained by comparison with the standard form of the Gaussian, and noting that the results

$$\lambda_1 = 1 - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \tag{51}$$

$$\lambda_2 = 0 \tag{52}$$

$$\lambda_3 = \frac{1}{2\sigma^2} \tag{53}$$

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 3.8.

2.25 Substituting the right hand side of (2.98) in the argument of the logarithm on the right hand side of (2.91), 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 (2.95).

2.26 The Kullback-Leibler divergence takes the form

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

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

$$\begin{aligned}
\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\left\{-\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathrm{d}\mathbf{x} + \text{const.} \\
&= \frac{1}{2}\left\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right]\right)\right\} + \text{const.} \\
&= \frac{1}{2}\left\{\ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}] + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]\right)\right\} \\
&\qquad +\text{const.}
\end{aligned} \tag{54}$$

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

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]. \tag{55}$$

Similarly, differentiating (54) w.r.t. $\boldsymbol{\Sigma}^{-1}$, again using results from Appendix A and also making use of (55) and (2.48), we see that

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} = \mathrm{cov}[\mathbf{x}]. \tag{56}$$

2.27 From (2.100) 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 (2.49) and (2.51)– (2.53), 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 (2.91) as the negative differential entropy of a Gaussian. Thus, from (57), (58) and (2.99), we have

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

2.28 First, let us set $\alpha = 1 - \epsilon$. Then

$$p(x)^{(1+\alpha)/2} = p(x)^{1-\epsilon/2} = p(x)\left\{1 - \frac{\epsilon}{2}\ln p(x) + \mathcal{O}(\epsilon^2)\right\}. \tag{59}$$

Likewise, we have

$$q(x)^{(1-\alpha)/2} = q(x)^{\epsilon/2} = 1 + \frac{\epsilon}{2}\ln q(x) + \mathcal{O}(\epsilon^2). \tag{60}$$

We also have

$$1 - \alpha^2 = 1 - 1 + 2\epsilon - \epsilon^2 = 2\epsilon + \mathcal{O}(\epsilon^2). \tag{61}$$

Substituting these expressions into the alpha divergence defined by (2.129) we obtain

$$\begin{aligned}\mathrm{D}_\alpha(p\|q) &= \frac{4}{2\epsilon}\left(1-\int p(x)\left\{1-\frac{\epsilon}{2}\ln p(x)\right\}\left\{1+\frac{\epsilon}{2}\ln q(x)\right\}\,\mathrm{d}x\right)+\mathcal{O}(\epsilon)\\ &= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\,\mathrm{d}x+\mathcal{O}(\epsilon) \qquad (62)\end{aligned}$$

where we have used

$$\int p(x)\,\mathrm{d}x = 1. \qquad (63)$$

Taking the limit $\epsilon \to 0$ gives $\mathrm{D}_1(p\|q) = \mathrm{KL}(p\|q)$ as required. We can similarly consider $\alpha = -1+\epsilon$, giving

$$p(x)^{(1+\alpha)/2} = p(x)^{\epsilon/2} = 1+\frac{\epsilon}{2}\ln p(x)+\mathcal{O}(\epsilon^2) \qquad (64)$$

together with

$$q(x)^{(1-\alpha)/2} = q(x)^{1-\epsilon/2} = q(x)\left\{1-\frac{\epsilon}{2}\ln q(x)+\mathcal{O}(\epsilon^2)\right\}. \qquad (65)$$

We also have

$$1-\alpha^2 = 1-1+2\epsilon-\epsilon^2 = 2\epsilon+\mathcal{O}(\epsilon^2). \qquad (66)$$

Substituting these expressions into the alpha divergence defined by (2.129) we obtain

$$\begin{aligned}\mathrm{D}_\alpha(p\|q) &= \frac{4}{2\epsilon}\left(1-\int q(x)\left\{1-\frac{\epsilon}{2}\ln q(x)\right\}\left\{1+\frac{\epsilon}{2}\ln p(x)\right\}\,\mathrm{d}x\right)+\mathcal{O}(\epsilon)\\ &= -\int q(x)\ln\left\{\frac{p(x)}{q(x)}\right\}\,\mathrm{d}x+\mathcal{O}(\epsilon) \qquad (67)\end{aligned}$$

where we have used

$$\int q(x)\,\mathrm{d}x = 1. \qquad (68)$$

Taking the limit $\epsilon \to 0$ gives $\mathrm{D}_{-1}(p\|q) = \mathrm{KL}(q\|p)$ as required.

2.29 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 (2.110), 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 (2.108) to obtain the desired result (2.130).

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 (2.108) 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 (2.110) it then follows that the mutual information $\mathrm{I}[\mathbf{x}, \mathbf{y}] = 0$. Finally, using (2.109) 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.

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

$$p(\mathbf{x}) = p(\mathbf{y}) \left| \frac{\partial y_i}{\partial x_j} \right| = p(\mathbf{y})|\mathbf{A}| \tag{69}$$

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

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \ln p(\mathbf{y}) \, \mathrm{d}\mathbf{y} = -\int p(\mathbf{x}) \ln \left\{ p(\mathbf{x})|\mathbf{A}|^{-1} \right\} \, \mathrm{d}\mathbf{x} = H(\mathbf{x}) + \ln |\mathbf{A}| \tag{70}$$

as required.

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

$$H(y|x) = -\sum_i \sum_j p(y_i|x_j)p(x_j) \ln p(y_i|x_j) \tag{71}$$

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.

2.32 Consider (2.101) 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 ϵ 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

$$\begin{aligned} f(a) &> f(x_0) + f'(x_0)(a - x_0) \\ &= f(x_0) + f'(x_0)\left((1 - \lambda)(a - b)\right). \end{aligned} \tag{72}$$

Similarly, setting $x = b$ gives

$$f(b) > f(x_0) + f'(x_0)(\lambda(b - a)). \tag{73}$$

Multiplying (72) by λ and (73) 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.

2.33 From (2.101) we know that the result (2.102) 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 (2.102)

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

$$= f\left(\lambda_{M+1}x_{M+1} + (1 - \lambda_{M+1})\sum_{i=1}^{M} \eta_i x_i\right) \tag{75}$$

where we have defined

$$\eta_i = \frac{\lambda_i}{1 - \lambda_{M+1}}. \tag{76}$$

We now apply (2.101) to give

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{i=1}^{M} \eta_i x_i\right). \tag{77}$$

We now note that the quantities λ_i by definition satisfy

$$\sum_{i=1}^{M+1} \lambda_i = 1 \tag{78}$$

and hence we have

$$\sum_{i=1}^{M} \lambda_i = 1 - \lambda_{M+1} \tag{79}$$

Then using (76) we see that the quantities η_i satisfy the property

$$\sum_{i=1}^{M} \eta_i = \frac{1}{1-\lambda_{M+1}} \sum_{i=1}^{M} \lambda_i = 1. \tag{80}$$

Thus we can apply the result (2.102) at order M and so (77) becomes

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leqslant \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1}) \sum_{i=1}^{M} \eta_i f(x_i) = \sum_{i=1}^{M+1} \lambda_i f(x_i) \tag{81}$$

where we have made use of (76).

2.34 For a one-dimensional variable the KL divergence takes the form

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} \mathrm{d}x \\ &= -\int p(x) \ln q(x) \, \mathrm{d}x + \text{const.} \end{aligned} \tag{82}$$

where the constant term is just the negative entropy of the fixed distribution $p(x)$. Substituting for $p(x)$ in the first term using the empirical distribution (2.37), and substituting for $q(x)$ using the model distribution $q(x|\theta)$ gives

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n) \ln q(x|\theta) \, \mathrm{d}x + \text{const.} \\ &= -\frac{1}{N} \sum_{n=1}^{N} \ln q(x_n|\theta) + \text{const.}. \end{aligned} \tag{83}$$

which is the required neagative log likelihood function up to an additive constant.

2.35 From (2.92), making use of (2.107), we have

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

2.36 We first evaluate the marginal and conditional probabilities $p(x)$, $p(y)$, $p(x|y)$, and $p(y|x)$, to give the results shown in the tables below. From these tables, together

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

$p(y|x)$

with the definitions

$$H(x) = -\sum_i p(x_i) \ln p(x_i) \tag{84}$$

$$H(x|y) = -\sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j) \tag{85}$$

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 (2.110) to evaluate the mutual information. The corresponding diagram is shown in Figure 1.

2.37 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 λ_i with $1/K$ in (2.102), taking into account that the logarithm is concave rather than convex and the inequality therefore goes the other way, we obtain the desired result.

2.38 From the product rule of probability we have $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, and so (2.109) can be written as

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{y}) \ln p(\mathbf{y})\, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \\ &= H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{86}$$

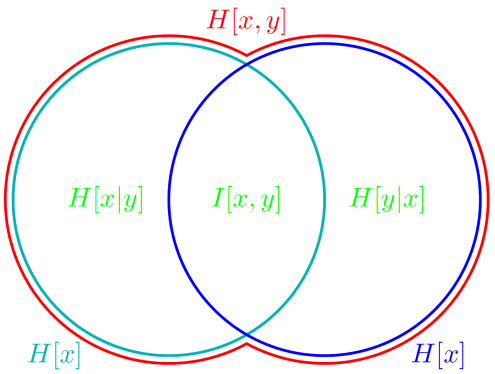

Figure 1 Diagram showing the relationship between marginal, conditional and joint entropy and the mutual information.

2.39 If z_1 and z_2 are independent, then

$$\begin{aligned}\mathrm{cov}[z_1, z_2] &= \iint (z_1 - \bar{z}_1)(z_2 - \bar{z}_2)p(z_1, z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2 \\ &= \iint (z_1 - \bar{z}_1)(z_2 - \bar{z}_2)p(z_1)p(z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2 \\ &= \int (z_1 - \bar{z}_1)p(z_1)\,\mathrm{d}z_1 \int (z_2 - \bar{z}_2)p(z_2)\,\mathrm{d}z_2 \\ &= 0,\end{aligned}$$

where

$$\bar{z}_i = \mathbb{E}[z_i] = \int z_i p(z_i)\,\mathrm{d}z_i.$$

For y_2 we have

$$p(y_2|y_1) = \delta(y_2 - y_1^2),$$

i.e., a spike of probability mass one at y_1^2, which is clearly dependent on y_1. With $\bar{y}_i$ defined analogously to $\bar{z}_i$ above, we get

$$\begin{aligned}\mathrm{cov}[y_1, y_2] &= \iint (y_1 - \bar{y}_1)(y_2 - \bar{y}_2)p(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \iint y_1(y_2 - \bar{y}_2)p(y_2|y_1)p(y_1)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \int (y_1^3 - y_1\bar{y}_2)p(y_1)\,\mathrm{d}y_1 \\ &= 0,\end{aligned}$$

where we have used the fact that all odd moments of y_1 will be zero, since it is symmetric around zero.

2.40 [The original printing of *Deep Learning: Foundations and Concepts* has a typo in this exercise in which the word 'convex' in the first sentence should read 'concave'. Note, however, that the exercise can equally well be solved with 'convex' by following the same reasoning as given here.] We introduce a binary variable C such that $\mathrm{C} = 1$ denotes that the coin lands concave side up, and a binary variable Q for which $\mathrm{Q} = 1$ denotes that the concave side of the coin is heads. From the stated physical properties of the coin, and from the assumed prior probability that the concave side is heads, we have

$$p(\mathrm{C} = 1) = 0.6 \tag{87}$$
$$p(\mathrm{Q} = 1) = 0.1. \tag{88}$$

The data set $\mathcal{D} = \{x_1, \ldots, x_{10}\}$ consists of 10 observations x each of which takes the value H for heads or T for tails. We assume that the data points are independent and identically distributed, so the ordering does not matter. A particular coin flip can land heads up either because it lands concave side up and the concave side is heads

or because it lands convex side up and the concave side is tails. Thus the probability of landing heads is given by

$$\begin{aligned} p(x = H) &= p(\mathrm{C} = 1)p(\mathrm{Q} = 1) + p(\mathrm{C} = 0)p(\mathrm{Q} = 0) \\ &= 0.6 \times 0.1 + 0.4 \times 0.9 \\ &= 0.06 + 0.36 \\ &= 0.42 \end{aligned} \tag{89}$$

from which it follows that $p(x = T) = 0.58$. The probability of observing 8 heads and 2 tails is therefore

$$p(\mathcal{D}) = \left(\begin{array}{c} 10 \\ 8 \end{array} \right) \times (0.42)^8 \times (0.58)^2 \tag{90}$$

where the coefficient in the binomial expansion is given by

$$\left(\begin{array}{c} N \\ K \end{array} \right) = \frac{N!}{K!(N-K)!}. \tag{91}$$

The posterior probability that the concave side is heads is then given by Bayes' theorem

$$p(\mathrm{Q} = 1|\mathcal{D}) = \frac{p(\mathcal{D}|\mathrm{Q} = 1)p(\mathrm{Q} = 1)}{p(\mathcal{D})} \tag{92}$$

For the first term in the numerator on the right-hand side we note that, since this is conditioned on the concave side being heads, we need to use the probabilities of the coin landing concave side up and concave side down, to give

$$p(\mathcal{D}|\mathrm{Q} = 1) = \left(\begin{array}{c} 10 \\ 8 \end{array} \right) \times (0.6)^8 \times (0.4)^2. \tag{93}$$

Substituting into Bayes' theorem, and noting that the binomial coefficients cancel, we obtain

$$p(\mathrm{Q} = 1|\mathcal{D}) = \frac{(0.6)^8 \times (0.4)^2 \times 0.1}{(0.42)^8 \times (0.58)^2} \simeq 0.825 \tag{94}$$

and so we see this is a higher probability than the prior of 0.1, which is intuitively reasonable since we saw a larger number of heads compared to tails in the data set. The probability that the next flip will land heads up is then given by

$$\begin{aligned} p(H|\mathcal{D}) &= p(H|Q = 1, \mathcal{D})p(Q = 1|\mathcal{D}) + p(H|Q = 0, \mathcal{D})p(Q = 0|\mathcal{D}) \\ &= p(H|Q = 1)p(Q = 1|\mathcal{D}) + p(H|Q = 0)p(Q = 0|\mathcal{D}) \\ &\simeq 0.6 \times 0.825 + 0.4 \times (1 - 0.825) \simeq 0.565 \end{aligned} \tag{95}$$

which we see is higher than the probability 0.42 of heads before we observed the data. Again this is intuitively reasonable given the predominance of heads in the data set.

2.41 If we substitute (2.115) into (2.114), we obtain (2.116). We now use (2.66) for the log likelihood of the linear regression model, and note that $\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2)$ corresponds to $\ln p(\mathcal{D}|\mathbf{w})$. We also note that $\sum_i w_i^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w}$. Hence we obtain (2.117) for the regularized error function.

Chapter 3 Standard Distributions

3.1 From the definition (3.2) of the Bernoulli distribution we have

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

The entropy is given by

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

3.2 The normalization of (3.195) 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

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

3.3 Using the definition (3.10) we have

$$\begin{aligned}\binom{N}{n}+\binom{N}{n-1} &= \frac{N!}{n!(N-n)!}+\frac{N!}{(n-1)!(N+1-n)!}\\ &= \frac{(N+1-n)N!+nN!}{n!(N+1-n)!}=\frac{(N+1)!}{n!(N+1-n)!}\\ &= \binom{N+1}{n}. \end{aligned} \tag{96}$$

To prove the binomial theorem (3.197) 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

$$\begin{aligned}(1+x)^{N+1} &= (1+x)\sum_{n=0}^{N}\binom{N}{n}x^n\\ &= \sum_{n=0}^{N}\binom{N}{n}x^n+\sum_{n=1}^{N+1}\binom{N}{n-1}x^n\\ &= \binom{N}{0}x^0+\sum_{n=1}^{N}\left\{\binom{N}{n}+\binom{N}{n-1}\right\}x^n+\binom{N}{N}x^{N+1}\\ &= \binom{N+1}{0}x^0+\sum_{n=1}^{N}\binom{N+1}{n}x^n+\binom{N+1}{N+1}x^{N+1}\\ &= \sum_{n=0}^{N+1}\binom{N+1}{n}x^n \end{aligned} \tag{97}$$

which completes the inductive proof. Finally, using the binomial theorem, the normalization condition (3.198) for the binomial distribution gives

$$\begin{aligned}\sum_{n=0}^{N}\binom{N}{n}\mu^n(1-\mu)^{N-n} &= (1-\mu)^N\sum_{n=0}^{N}\binom{N}{n}\left(\frac{\mu}{1-\mu}\right)^n\\ &= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N=1 \end{aligned} \tag{98}$$

as required.

3.4 Differentiating (3.198) with respect to μ 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 (3.11).

If we differentiate (3.198) twice with respect to μ 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 (3.11) for the mean of the binomial distribution, to obtain

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

Finally, we use (2.46) to obtain the result (3.12) for the variance.

3.5 We differentiate (3.26) 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 (A.19), (A.20), 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 $\mathbf{x} = \boldsymbol{\mu}$.

3.6 First, we define $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, and we assume that $\mathbf{y}$ has the same dimensionality as $\mathbf{x}$, so that $\mathbf{A}$ is a square matrix. We also assume that $\mathbf{A}$ is symmetric and has an inverse. It follows that $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$. From the sum and product rules of probability we have

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \int \delta \left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b} \right) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &\propto \int \delta \left(\mathbf{x} - \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) \right) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \end{aligned} \tag{99}$$

where $\delta(\cdot)$ is the Dirac delta function. Substituting for $p(\mathbf{x})$ using the Gaussian distribution we have

$$\begin{aligned} p(\mathbf{y}) &\propto \int \delta \left(\mathbf{x} - \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) \right) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \mathrm{d}\mathbf{x} \\ &\propto \int \delta \left(\mathbf{x} - \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) \right) \exp \left\{ -\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\} \mathrm{d}\mathbf{x} \\ &\propto \exp \left\{ -\frac{1}{2} (\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}) - \boldsymbol{\mu}) \right\} \\ &\propto \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{b} - \mathbf{A}\boldsymbol{\mu})^{\mathrm{T}} \mathbf{A}^{-1} \boldsymbol{\Sigma}^{-1} \mathbf{A}^{-1} (\mathbf{y} - \mathbf{b} - \mathbf{A}\boldsymbol{\mu}) \right\} \end{aligned} \tag{100}$$

where we have used the property of a symmetric matrix that its inverse is also symmetric. By inspection we see that $p(\mathbf{y})$ is a Gaussian distribution and that its mean is $\mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and that its covariance is

$$\left(\mathbf{A}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}. \tag{101}$$

3.7 From (2.100) we have

$$\mathrm{KL}\left(q(\mathbf{x})\|p(\mathbf{x})\right) = -\int q(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int q(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{102}$$

Using (3.26), (3.40), (3.42) and (3.46), we can rewrite the first integral on the r.h.s. of (102) as

$$\begin{aligned} &-\int q(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \frac{1}{2}\left\{D\ln(2\pi) + \ln|\boldsymbol{\Sigma}_p| + (\mathbf{x}-\boldsymbol{\mu}_p)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}(\mathbf{x}-\boldsymbol{\mu}_p)\right\}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\left\{D\ln(2\pi) + \ln|\boldsymbol{\Sigma}_p| + \mathrm{Tr}[\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q\boldsymbol{\mu}_q^{\mathrm{T}} + \boldsymbol{\Sigma}_q)]\right. \\ &\qquad \left. -\boldsymbol{\mu}_p\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_q - \boldsymbol{\mu}_q^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p + \boldsymbol{\mu}_p^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}_p\right\}. \end{aligned} \tag{103}$$

The second integral on the r.h.s. of (102) we recognize from (2.92) as the negative differential entropy of a multivariate Gaussian. Thus, from (102), (103) and (3.204), we have

$$\begin{aligned} &\mathrm{KL}\left(q(\mathbf{x})\|p(\mathbf{x})\right) \\ &= \frac{1}{2}\left\{\ln\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - D + \mathrm{Tr}\left(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\right) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right\}. \end{aligned} \tag{104}$$

3.8 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 (3.201), a D-dimensional vector $\mathbf{m}$ of Lagrange multipliers for the D constraints given by (3.202), and a $D \times D$ matrix $\mathbf{L}$ of Lagrange multipliers to enforce the D^2 constraints represented by (3.203). Thus we maximize

$$\begin{aligned} \widetilde{\mathrm{H}}[p] \;=\; &-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \lambda\left(\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} - 1\right) \\ &+\mathbf{m}^{\mathrm{T}}\left(\int p(\mathbf{x})\mathbf{x}\,\mathrm{d}\mathbf{x} - \boldsymbol{\mu}\right) \\ &+\mathrm{Tr}\left\{\mathbf{L}\left(\int p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{x} - \boldsymbol{\Sigma}\right)\right\}. \end{aligned} \tag{105}$$

By functional differentiation (Appendix B) 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

$$p(\mathbf{x}) = \exp\left\{\lambda - 1 + \mathbf{m}^{\mathrm{T}}\mathbf{x} + (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{L}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{106}$$

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 (3.202) 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 (3.201) 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 (3.201). 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 (3.203), 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 (3.48) we obtain $\mathbf{L} = -\frac{1}{2}\boldsymbol{\Sigma}$. Finally, the value of λ is simply that value needed to ensure that the Gaussian distribution is correctly normalized, as derived in Section 3.2, and hence is given by

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

3.9 From the definitions of the multivariate differential entropy (2.92) and the multivariate Gaussian distribution (3.26), we get

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

3.10 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 (3.205) which takes the form

$$p(x) = \left(\frac{\tau_1}{2\pi}\right)^{1/2}\left(\frac{\tau_2}{2\pi}\right)^{1/2}\int_{-\infty}^{\infty} \exp\left\{-\frac{\tau_1}{2}(x-\mu_1-x_2)^2 - \frac{\tau_2}{2}(x_2-\mu_2)^2\right\} \mathrm{d}x_2. \tag{107}$$

Since the final result will be a Gaussian distribution for $p(x)$ we need only evaluate its precision, since, from (2.99), 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 (107) 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 (107). 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 (3.99) for the convolution of two linear-Gaussian distributions.

The entropy of x is then given, from (2.99), by

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

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

$$\Lambda^{\mathrm{S}}_{ij} = \frac{\Lambda_{ij}+\Lambda_{ji}}{2}, \qquad \Lambda^{\mathrm{A}}_{ij} = \frac{\Lambda_{ij}-\Lambda_{ji}}{2} \tag{108}$$

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

$$\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)\Lambda_{ij}(x_j-\mu_j) \tag{109}$$

where $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ is the precision matrix. When we substitute $\mathbf{\Lambda} = \mathbf{\Lambda}^{\mathrm{A}} + \mathbf{\Lambda}^{\mathrm{S}}$ into (109) 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.

3.12 We start by pre-multiplying both sides of (3.28) by $\mathbf{u}_i^\dagger$, the conjugate transpose of $\mathbf{u}_i$. This gives us

$$\mathbf{u}_i^\dagger\mathbf{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i^\dagger\mathbf{u}_i. \tag{110}$$

Next consider the conjugate transpose of (3.28) and post-multiply it by $\mathbf{u}_i$, which gives us

$$\mathbf{u}_i^\dagger\mathbf{\Sigma}^\dagger\mathbf{u}_i = \lambda_i^*\mathbf{u}_i^\dagger\mathbf{u}_i. \tag{111}$$

where λ_i^* is the complex conjugate of λ_i. We now subtract (110) from (111) 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 λ_i must be real.

Now consider

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

where we have used (3.28) and the fact that $\mathbf{\Sigma}$ 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 (3.28),

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

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 (3.29).

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 (3.29) 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 (3.29).

3.13 We can write the r.h.s. of (3.31) 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 (3.28)–(3.30), 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 (3.31) holds.

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

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

3.14 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 (3.28) 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, λ_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.

3.15 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

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

3.16 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 (A.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.

3.17 Recall that the transformation (3.34) diagonalizes the coordinate system and that the quadratic form (3.27), corresponding to the square of the Mahalanobis distance, is then given by (3.33). 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 Δ whose volume is $V_D \Delta^D$.

3.18 Multiplying the left hand side of (3.60) by the matrix (3.208) 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

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

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.

3.19 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 3.2.5 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 3.2.4 the distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ is then Gaussian with mean and covariance given by (3.65) and (3.66) respectively.

3.20 Multiplying the left hand side of (3.210) by $(\mathbf{A}+\mathbf{BCD})$ trivially gives the identity matrix $\mathbf{I}$. On the right hand side we obtain

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

3.21 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 $\mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])\right]=\mathbb{E}\left[(\mathbf{z}-\mathbb{E}[\mathbf{z}])\right]=0$, 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 2.10 as a special case.

3.22 For the marginal distribution $p(\mathbf{x})$ we see from (3.76) that the mean is given by the upper partition of (3.92) which is simply $\boldsymbol{\mu}$. Similarly from (3.77) we see that the covariance is given by the top left partition of (3.89) and is therefore given by $\boldsymbol{\Lambda}^{-1}$.

Now consider the conditional distribution $p(\mathbf{y}|\mathbf{x})$. Applying the result (3.65) for the conditional mean we obtain

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

Similarly applying the result (3.66) 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.

3.23 We first define

$$\mathbf{X}=\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A} \tag{112}$$

and

$$\mathbf{W}=-\mathbf{L}\mathbf{A}, \quad \text{and thus} \quad \mathbf{W}^{\mathrm{T}}=-\mathbf{A}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}=-\mathbf{A}^{\mathrm{T}}\mathbf{L}, \tag{113}$$

since $\mathbf{L}$ is symmetric. We can use (112) and (113) to re-write (3.88) as

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

and using (3.60) 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 (112) and (113), respectively, we get

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

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

and

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

as required.

3.24 Substituting the leftmost expression of (3.89) for $\mathbf{R}^{-1}$ in (3.91), we get

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

3.25 Since $\mathbf{y} = \mathbf{x} + \mathbf{z}$ we can write the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ in the form $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{z}} + \mathbf{x}, \mathbf{\Sigma}_{\mathbf{z}})$. 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 $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}})$. This therefore takes the form of (3.83) and (3.84) in which we can identify $\boldsymbol{\mu} \to \boldsymbol{\mu}_{\mathbf{x}}$, $\mathbf{\Lambda}^{-1} \to \mathbf{\Sigma}_{\mathbf{x}}$, $\mathbf{A} \to \mathbf{I}$, $\mathbf{b} \to \boldsymbol{\mu}_{\mathbf{z}}$ and $\mathbf{L}^{-1} \to \mathbf{\Sigma}_{\mathbf{z}}$. We can now obtain the marginal distribution $p(\mathbf{y})$ by making use of the result (3.99) from which we obtain $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}} + \mathbf{\Sigma}_{\mathbf{x}})$. Thus both the means and the covariances are additive, in agreement with the results of Exercise 3.21.

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

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{114}$$

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

$$\begin{aligned} &= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{x}+\mathbf{x}^{\mathrm{T}}\left[\boldsymbol{\Lambda}\boldsymbol{\mu}+\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})\right]+\text{const} \\ &= -\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})(\mathbf{x}-\mathbf{m}) \\ &\quad +\frac{1}{2}\mathbf{m}^{\mathrm{T}}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})\mathbf{m}+\text{const} \end{aligned} \tag{115}$$

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 (115). Then we extract the terms in $\mathbf{y}$ from the final term in (115) and combine these with the remaining terms from the quadratic form (114) which depend on $\mathbf{y}$ to give

$$\begin{aligned} &= -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\left\{\mathbf{L}-\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\right\}\mathbf{y} \\ &\quad +\mathbf{y}^{\mathrm{T}}\left[\left\{\mathbf{L}-\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\right\}\mathbf{b}\right. \\ &\quad \left.+\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\right]. \end{aligned} \tag{116}$$

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 (3.210) to give

$$\left\{\mathbf{L}-\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{L}\right\}^{-1}=\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}} \tag{117}$$

which corresponds to (3.94).

Next we identify the mean $\boldsymbol{\nu}$ of the marginal distribution. To do this we make use of (117) in (116) 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 (3.93).

3.27 To find the conditional distribution $p(\mathbf{x}|\mathbf{y})$ we start from the quadratic form (114) 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

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

where, as in the solution to Exercise 3.26, 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 (3.95) and (3.96).

3.28 Differentiating (3.102) 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 (A.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 (A.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 (A.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.

3.29 The derivation of (3.46) follows directly from the discussion given in the text between (3.42) and (3.46). If $m = n$ then, using (3.46) 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 (3.42). Combining these results we obtain (3.213). From (3.42) and (3.46) we then have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Sigma}_{\mathrm{ML}}\right] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{x}_n - \frac{1}{N}\sum_{m=1}^{N}\mathbf{x}_m\right)\left(\mathbf{x}_n^{\mathrm{T}} - \frac{1}{N}\sum_{l=1}^{N}\mathbf{x}_l^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} - \frac{2}{N}\mathbf{x}_n\sum_{m=1}^{N}\mathbf{x}_m^{\mathrm{T}} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}\mathbf{x}_m\mathbf{x}_l^{\mathrm{T}}\right] \\
&= \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - 2\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \frac{1}{N}\boldsymbol{\Sigma}\right\} \\
&= \left(\frac{N-1}{N}\right)\boldsymbol{\Sigma} \qquad (118)
\end{aligned}$$

as required.

3.30 Using the relation (3.214) 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

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

3.31 Expressed in terms of ξ 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 $p(\theta) \propto \exp\{-m(\theta-\theta_0)^2/2\}$.

3.32 Using (3.133), we can write (3.132) as

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

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. θ_0 to obtain (3.134).

3.33 Differentiating the von Mises distribution (3.129) 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 $\theta = \theta_0 + \pi \,(\mathrm{mod} 2\pi)$. 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.

3.34 From (3.119) and (3.134), we see that $\bar{\theta} = \theta_0^{\mathrm{ML}}$. Using this together with (3.118) and (3.127), we can rewrite (3.137) as follows:

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

3.35 Starting from (3.26), 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 (3.138), with

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

3.36 Taking the first derivative of (3.172) 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 (3.172).

3.37 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 Δ_i, the normalization constraint becomes $\sum_i h_i\Delta_i = 1$. Introducing a Lagrange multiplier λ 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 λ then gives our final result for the maximum likelihood solution for h_k in the form

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

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

3.38 From (3.180) we have

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

where $V(\rho)$ is the volume of a D-dimensional hypersphere with radius ρ, where in turn ρ is the distance from $\mathbf{x}$ to its K^{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

$$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} \propto \int_0^\infty r^{-D} r^{D-1}\,\mathrm{d}r = \int_0^\infty r^{-1}\,\mathrm{d}r$$

which diverges logarithmically.

Chapter 4 Single-layer Networks: Regression

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

$$\sum_{n=1}^{N} \left(\sum_{j=0}^{M} w_j x_n^j - t_n \right) x_n^i = 0. \tag{119}$$

Re-arranging terms then gives the required result.

4.2 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 (4.53), but with A_{ij} replaced by $\widetilde{A}_{ij}$, given by

$$\widetilde{A}_{ij} = A_{ij} + \lambda I_{ij}. \tag{120}$$

4.3 Using (4.6), 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 (4.57) 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 $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

4.4 We first write

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

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 (4.14), 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 $\mathcal{S}$.

4.5 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 (4.14) for the case $\mathbf{R} = \mathbf{I}$.

If we compare (4.60) with (4.9)–(4.11), 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 β.

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 (4.60) with r_n taking positive integer values, although it is valid for any $r_n > 0$.

4.6 Taking the gradient of (4.26) with respect to $\mathbf{w}$ and setting this to zero, we obtain

$$0 = -\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}\boldsymbol{\phi}(\mathbf{x}_n) + \lambda\mathbf{w} \tag{121}$$

which we can rewrite as

$$0 = -\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \lambda\mathbf{w} \tag{122}$$

Rearranging to solve for $\mathbf{w}$ we then obtain

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \tag{123}$$

as required.

4.7 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 (4.14) as required.

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

4.8 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

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

which is equivalent to (4.37).

4.9 We start by expanding the square in (**??**), in a similar fashion to the univariate case in the equation preceding (4.39),

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

Following the treatment of the univariate case, we now substitute this into (4.64) 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 $\mathbb{E}[\mathbf{t}|\mathbf{x}]$.

4.10 This exercise is just a repeat of Exercise 4.9.

4.11 To prove the normalization of the distribution (4.66) 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 (**??**) 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 (4.66) 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.

4.12 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

$$\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\,\mathrm{d}t \tag{124}$$

for each value of $\mathbf{x}$. Setting the derivative of (124) with respect to $y(\mathbf{x})$ to zero gives the stationarity condition

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

which can also be obtained directly by setting the functional derivative of (4.40) with respect to $y(\mathbf{x})$ equal to zero. It follows that $y(\mathbf{x})$ must satisfy

$$\int_{-\infty}^{y(\mathbf{x})}|y(\mathbf{x})-t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty}|y(\mathbf{x})-t|^{q-1}p(t|\mathbf{x})\,\mathrm{d}t. \tag{125}$$

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

$$\int_{-\infty}^{y(\mathbf{x})}p(t|\mathbf{x})\,\mathrm{d}t = \int_{y(\mathbf{x})}^{\infty}p(t|\mathbf{x})\,\mathrm{d}t. \tag{126}$$

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 (124) 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 (124) by choosing the location of the notch to coincide with the largest value of $p(t)$, i.e. with the (conditional) mode.

Chapter 5 Single-layer Networks: Classification

5.1 Consider the component t_k of $\mathbf{t}$. By definition we have

$$p(t_k = 1|\mathbf{x}) = p(\mathcal{C}_k|\mathbf{x}) \tag{127}$$
$$p(t_k = 0|\mathbf{x}) = 1 - p(\mathcal{C}_k|\mathbf{x}). \tag{128}$$

Taking the expectation of t_k then gives

$$\mathbb{E}[t_k|\mathbf{x}] = 1 \times p(\mathcal{C}_k|\mathbf{x}) + 0 \times \{1 - p(\mathcal{C}_k|\mathbf{x})\} = p(\mathcal{C}_k|\mathbf{x}) \tag{129}$$

as required.

5.2 Assume that the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ intersect. Then there exists at least one point $\mathbf{z}$ such that

$$\mathbf{z} = \sum_n \alpha_n \mathbf{x}_n = \sum_m \beta_m \mathbf{y}_m \tag{130}$$

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

$$\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_n \alpha_n \widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 = \sum_n \alpha_n(\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0) > 0, \tag{131}$$

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

$$\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{z} + w_0 = \sum_m \beta_m \widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0 = \sum_m \beta_m(\widehat{\mathbf{w}}^{\mathrm{T}}\mathbf{y}_m + w_0) < 0, \tag{132}$$

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.

5.3 For the purpose of this exercise, we make the contribution of the bias weights explicit in (5.14), giving

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\right\}, \tag{133}$$

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 (133) w.r.t. $\mathbf{w}_0$, giving

$$2N\mathbf{w}_0 + 2(\mathbf{X}\mathbf{W} - \mathbf{T})^{\mathrm{T}}\mathbf{1}. \tag{134}$$

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

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \tag{135}$$

where

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^{\mathrm{T}}\mathbf{1} \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{1}. \tag{136}$$

If we subsitute (135) into (133), we get

$$E_D(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})\right\}, \tag{137}$$

where

$$\overline{\mathbf{T}} = \mathbf{1}\bar{\mathbf{t}}^{\mathrm{T}} \quad \text{and} \quad \overline{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}. \tag{138}$$

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

$$\mathbf{W} = (\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{T}} = \widehat{\mathbf{X}}^{\dagger}\widehat{\mathbf{T}}, \tag{139}$$

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 $\mathbf{x}^\star$,

$$\begin{aligned} \mathbf{y}(\mathbf{x}^\star) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \mathbf{w}_0 \\ &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\ &= \bar{\mathbf{t}} - \widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}). \end{aligned} \tag{140}$$

If we apply (5.97) to $\bar{\mathbf{t}}$, we get

$$\mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{1} = -b. \tag{141}$$

Therefore, applying (5.97) to (140), 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 $\mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = b(\mathbf{1} - \mathbf{1})^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$.

5.4 When we consider several simultaneous constraints, (5.97) becomes

$$\mathbf{A}\mathbf{t}_n + \mathbf{b} = \mathbf{0}, \tag{142}$$

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 (142) to (140), 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 $\mathbf{A}\mathbf{y}(\mathbf{x}^\star) + \mathbf{b} = \mathbf{0}$.

5.5 Using the definitions (5.30) and (5.31) we have

$$\begin{aligned}\text{Precision} \times \text{Recall} &= \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \times \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \\ &= \frac{N_{\text{TP}}^2}{(N_{\text{TP}} + N_{\text{FP}})(N_{\text{TP}} + N_{\text{FN}})}.\end{aligned}$$

Similarly, taking a common denominator we have

$$\begin{aligned}\text{Precision} + \text{Recall} &= \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} + \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \\ &= \frac{N_{\text{TP}}(N_{\text{TP}} + N_{\text{FN}}) + N_{\text{TP}}(N_{\text{TP}} + N_{\text{FP}})}{(N_{\text{TP}} + N_{\text{FP}})(N_{\text{TP}} + N_{\text{FN}})} \\ &= \frac{N_{\text{TP}}(2N_{\text{TP}} + N_{\text{FN}} + N_{\text{FP}})}{(N_{\text{TP}} + N_{\text{FP}})(N_{\text{TP}} + N_{\text{FN}})}.\end{aligned}$$

Substituting these two results into (5.38) we obtain

$$F = \frac{2N_{\text{TP}}}{2N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}}}. \tag{143}$$

as required.

5.6 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 (**??**), by

$$\begin{aligned}p(\text{mistake}) &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_1} p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \qquad (144)\end{aligned}$$

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

$$\begin{aligned}p(\text{mistake}) &\leqslant \int_{\mathcal{R}_1} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\qquad + \int_{\mathcal{R}_2} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \{p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\}^{1/2}\,\mathrm{d}\mathbf{x} \qquad (145)\end{aligned}$$

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

5.7 Substituting $L_{kj} = 1 - \delta_{kj}$ into (5.23), 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.

5.8 From (5.23) 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

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k) p(\mathcal{C}_k) \tag{146}$$

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

5.9 We recognise the sum over data points in (5.100) as the finite-sample approximation to an expectation, as seen in (2.40). Taking the limit $N \to \infty$ we can use (2.39) to write the expectation in the form

$$\mathbb{E}\left[p(\mathcal{C}_k|\mathbf{x})\right] = \int p(\mathcal{C}_k|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \int p(\mathcal{C}_k, \mathbf{x}) \, \mathrm{d}\mathbf{x} = p(\mathcal{C}_k) \tag{147}$$

where we have used the product and sum rules of probability.

5.10 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 λ. Thus, if

$$j = \arg\min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) \tag{148}$$

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

$$\text{choose} \begin{cases} \text{class } j, & \text{if } \min_l \sum_k L_{kl} p(\mathcal{C}_k|\mathbf{x}) < \lambda; \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{149}$$

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 λ, 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 θ. Thus these two criteria for rejection are equivalent provided $\theta = 1 - \lambda$.

5.11 From (5.42) 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

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

5.12 Substituting (5.47) into (5.41), 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 (5.40) we obtain (5.48), with $\mathbf{w}$ and w_0 given by (5.49) and (5.50), respectively.

5.13 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

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln p(\phi_n|\mathcal{C}_k) + \ln \pi_k\right\}. \tag{150}$$

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

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right). \tag{151}$$

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

$$\sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0. \tag{152}$$

Re-arranging then gives

$$-\pi_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k. \tag{153}$$

Summing both sides over k we find that $\lambda = -N$, and using this to eliminate λ we obtain (5.101).

5.14 If we substitute (5.102) into (150) and then use the definition of the multivariate Gaussian, (3.26), we obtain

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \\ -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}, \tag{154}$$

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

Setting the derivative of the r.h.s. of (154) w.r.t. $\boldsymbol{\mu}_k$, obtained by using (A.19), to zero, we get

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) = 0. \tag{155}$$

Making use of (153), we can re-arrange this to obtain (5.103).

Rewriting the r.h.s. of (154) as

$$-\frac{1}{2}b\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\right]\right\}, \tag{156}$$

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

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{k}^{\mathrm{T}} t_{nk}\left\{\boldsymbol{\Sigma} - (\boldsymbol{\phi}_n - \boldsymbol{\mu}_n)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right\} = 0. \tag{157}$$

Again making use of (153), we can re-arrange this to obtain (5.104), with $\mathbf{S}_k$ given by (5.105).

Note that, as in Exercise 3.28, we do not enforce that $\boldsymbol{\Sigma}$ should be symmetric, but simply note that the solution is automatically symmetric.

5.15 We assume that the training set consists of data points $\mathbf{x}_n$ each of which is labelled with the associated class $\mathcal{C}_k$. This allows the parameters $\{\mu_{ki}\}$ to be fitted for each class independently. From (5.64) the log likelihood function for class $\mathcal{C}_k$ is then given by

$$\ln p(\mathcal{D}|\mathcal{C}_k) = \sum_{n=1}^{N}\sum_{i=1}^{D}\{x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\}\,. \tag{158}$$

Setting the derivative with respect to μ_{ki} equal to zero gives

$$0 = \sum_{n=1}^{N}\left\{\frac{x_{ni}}{\mu_{ki}} - \frac{(1-x_{ni})}{(1-\mu_{ki})}\right\}. \tag{159}$$

Rearranging to solve for μ_{ki} we finally obtain

$$\mu_{ki} = \frac{1}{N}\sum_{n=1}^{N} x_{ni} \tag{160}$$

which is the intuitively pleasing result that, for each class k and for each component i, the value of μ_{ki} is given by the average of the values of the corresponding components x_{ni} of those data vectors that belong to class $\mathcal{C}_k$. Since the x_{ni} are binary this is just the fraction of data points for which the corresponding value of i is equal to one.

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

$$p\left(\boldsymbol{\phi} \mid \mathcal{C}_k\right) = \prod_{m=1}^{M} p\left(\boldsymbol{\phi}_m \mid \mathcal{C}_k\right) \tag{161}$$

where

$$p\left(\boldsymbol{\phi}_m \mid \mathcal{C}_k\right) = \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}}, \tag{162}$$

where in turn $\{\mu_{kml}\}$ are the parameters of the multinomial models for $\boldsymbol{\phi}$.

Substituting this into (5.46) we see that

$$\begin{aligned} a_k &= \ln p\left(\boldsymbol{\phi} \mid \mathcal{C}_k\right) p\left(\mathcal{C}_k\right) \\ &= \ln p\left(\mathcal{C}_k\right) + \sum_{m=1}^{M} \ln p\left(\boldsymbol{\phi}_m \mid \mathcal{C}_k\right) \\ &= \ln p\left(\mathcal{C}_k\right) + \sum_{m=1}^{M}\sum_{l=1}^{L} \phi_{ml} \ln \mu_{kml}, \end{aligned} \tag{163}$$

which is linear in ϕ_{ml}.

5.17 We denote the data set by $\mathcal{D} = \{\phi_{nml}\}$ where $n = 1, \ldots, N$. From the naive Bayes assumption we can fit each class $\mathcal{C}_k$ separately to the training data. For class $\mathcal{C}_k$ the log likelihood function takes the form

$$\ln p(\mathcal{D}|\mathcal{C}_k) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{l=1}^{L} \phi_{nml} \ln \mu_{kml}. \tag{164}$$

Note that the parameter μ_{kml} represents the probability that for class $\mathcal{C}_k$ the component m will have its non-zero element in position l. In order to find the maximum likelihood solution we need to take account of the constraint that these probabilities must sum to one, separately for each value of m, so that

$$\sum_{l=1}^{L} \mu_{kml} = 1. \tag{165}$$

We can handle this by introducing Lagrange multipliers, one per component, and then maximize the modified likelihood function given by

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{l=1}^{L} \phi_{nml} \ln \mu_{kml} + \sum_{m=1}^{M} \lambda_m \left(\sum_{l=1}^{L} \mu_{kml} - 1 \right). \tag{166}$$

Setting the derivative with respect to μ_{kml} equal to zero gives

$$0 = \sum_{n=1}^{N} \left\{ \frac{\phi_{nml}}{\mu_{kml}} \right\} + \lambda_m. \tag{167}$$

Rearranging to solve for μ_{kml} we obtain

$$\mu_{kml} = -\frac{1}{\lambda_m} \sum_{n=1}^{N} \phi_{nml}. \tag{168}$$

To find the Lagrange multipliers we substitute this result into the constraint (165) and rearrange to give

$$\lambda_m = -\sum_{n=1}^{N} \sum_{l=1}^{L} \phi_{nml}. \tag{169}$$

We now use this to replace the Lagrange multiplier in (168) to give the final result for the maximum likelihood solution for the parameters in the form

$$\mu_{kml} = \frac{\sum_{n=1}^{N} \phi_{nml}}{\sum_{n=1}^{N} \sum_{l=1}^{L} \phi_{nml}}. \tag{170}$$

5.18 Differentiating (5.42) we obtain

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

5.19 We start by computing the derivative of (5.74) w.r.t. y_n

$$\frac{\partial E}{\partial y_n} = \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \tag{171}$$

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

$$= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \tag{172}$$

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

From (5.72), we see that

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)\left(1-\sigma(a_n)\right) = y_n(1-y_n). \tag{174}$$

Finally, we have

$$\nabla a_n = \boldsymbol{\phi}_n \tag{175}$$

where ∇ denotes the gradient with respect to $\mathbf{w}$. Combining (173), (174) and (175) 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.

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

$$\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n \begin{cases} \geqslant 0 & \text{if } t_n = 1, \\ < 0 & \text{otherwise.} \end{cases} \tag{176}$$

Moreover, from (5.74) 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.

5.21 From (5.76) 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 (5.78).

5.22 From (5.80) we have

$$\frac{\partial E}{\partial y_{nk}} = -\frac{t_{nk}}{y_{nk}}. \tag{177}$$

If we combine this with (5.78) 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 (175), again using the chain rule, we obtain (5.81).

5.23 We consider the two cases where $a \geqslant 0$ and $a < 0$ separately. In the first case, we can use (3.25) to rewrite (5.86) as

$$\begin{aligned}\Phi(a) &= \int_{-\infty}^{0}\mathcal{N}\left(\theta|0,1\right)\,\mathrm{d}\theta + \int_{0}^{a}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\theta^2}{2}\right)\,\mathrm{d}\theta\\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{a/\sqrt{2}}\exp\left(-u^2\right)\sqrt{2}\,\mathrm{d}u\\ &= \frac{1}{2}\left\{1+\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right\},\end{aligned} \tag{178}$$

where, in the last line, we have used (5.87).

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

$$\Phi(a) = 1 - \Phi(-a). \tag{179}$$

Combining this with (178), we get

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

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

5.24 From (5.72) we have that

$$\begin{aligned}\left.\frac{\mathrm{d}\sigma}{\mathrm{d}a}\right|_{a=0} &= \sigma(0)(1-\sigma(0)) \\ &= \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}. \qquad (180)\end{aligned}$$

Since the derivative of a cumulative distribution function is simply the corresponding density function, (5.86) 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 (180), we see that

$$\lambda = \frac{\sqrt{2\pi}}{4} \quad \text{or equivalently} \quad \lambda^2 = \frac{\pi}{8}. \qquad (181)$$

The comparison of the logistic sigmoid function and the scaled probit function is illustrated in Figure 5.12.

Chapter 6 Deep Neural Networks

6.1 On the right-hand side of (6.51) we make the change of variables $u = r^2$ to give

$$\frac{1}{2}S_D \int_0^\infty e^{-u} u^{D/2-1}\,\mathrm{d}u = \frac{1}{2}S_D \Gamma(D/2) \tag{182}$$

where we have used the definition (**??**) of the Gamma function. On the left hand side of (6.51) we can use (2.126) to obtain $\pi^{D/2}$. Equating these we obtain the desired result (6.53).

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

$$V_D = S_D \int_0^1 r^{D-1}\,\mathrm{d}r = \frac{S_D}{D}. \tag{183}$$

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

$$S_2 = 2\pi, \qquad S_3 = 4\pi, \qquad V_2 = \pi a^2, \qquad V_3 = \frac{4}{3}\pi a^3. \tag{184}$$

6.2 The volume of the cube is $(2a)^D$. Combining this with (6.53) and (6.54) we obtain (6.55). Using Stirling's formula (6.56) in (6.55) the ratio becomes, for large D,

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \left(\frac{\pi e}{2D}\right)^{D/2} \frac{1}{D} \tag{185}$$

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 $\sqrt{D}$.

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

$$\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \simeq p(r) S_D r^{D-1} \epsilon \tag{186}$$

from which we obtain (6.58). We can find the stationary points of $p(r)$ by differentiation

$$\frac{\mathrm{d}}{\mathrm{d}r}p(r) \propto \left[(D-1)r^{D-2} + r^{D-1}\left(-\frac{r}{\sigma^2}\right)\right] \exp\left(-\frac{r^2}{2\sigma^2}\right) = 0. \tag{187}$$

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

Next we note that

$$\begin{aligned} p(\widehat{r}+\epsilon) &\propto (\widehat{r}+\epsilon)^{D-1}\exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}+(D-1)\ln(\widehat{r}+\epsilon)\right]. \end{aligned} \tag{188}$$

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 (6.59).

Finally, from (6.57) 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 (6.57) 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 $\exp(D/2)$.

6.4 Using the definition of the $\tanh$ function we have

$$\begin{aligned} \tanh(a) &= \frac{e^a - e^{-a}}{e^a + e^{-a}} \\ &= \frac{1-e^{-2a}}{1+e^{-2a}} \\ &= \frac{2}{1+e^{-2a}} - \frac{1+e^{-2a}}{1+e^{-2a}} \\ &= \frac{2}{1+e^{-2a}} - 1 \\ &= 2\sigma(2a) - 1 \end{aligned} \tag{189}$$

where we have made use of the definition of the sigmoid function in (6.60). Rearranging we obtain

$$\sigma(a) = \frac{1}{2}\left(\tanh(a/2)+1\right). \tag{190}$$

For the case of a logistic sigmoid activation function, the argument of the output-unit activation function in (6.11) is given by

$$\sum_{j=0}^{M} w_{kj}^{(2)}\sigma\left(\sum_{i=0}^{D} w_{ji}^{(1)}x_i\right). \tag{191}$$

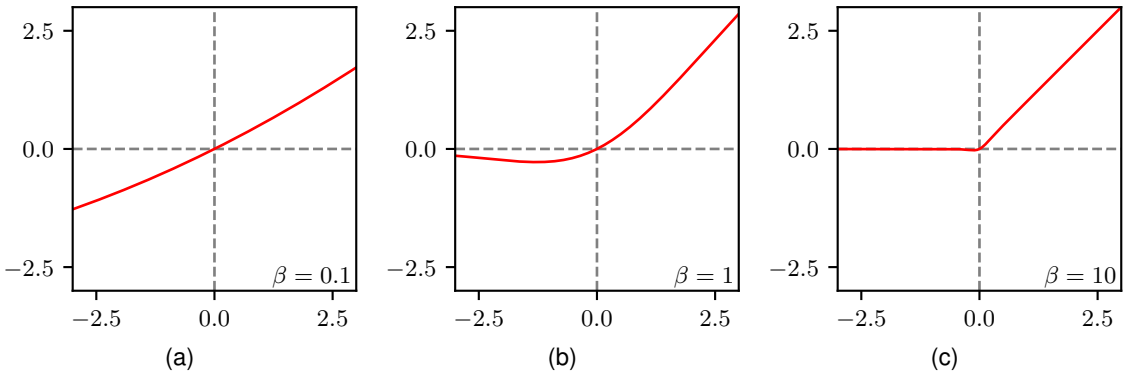

Figure 2 The swish activation function plotted for values of β = 0.1, 1 and 10

Substituting for the sigmoid function using (190) we obtain

$$\sum_{j=0}^{M} \widetilde{w}_{kj}^{(2)} \tanh\left(\sum_{i=0}^{D} \widetilde{w}_{ji}^{(1)} x_i\right) \tag{192}$$

where we have defined

$$\widetilde{w}_{kj}^{(2)} = \frac{1}{2} w_{kj}^{(2)}, \qquad j = 1, \ldots, M \tag{193}$$

$$\widetilde{w}_{k0}^{(2)} = \frac{1}{2} w_{k0}^{(2)} + \frac{1}{2} \tag{194}$$

$$\widetilde{w}_{ji}^{(1)} = \frac{1}{2} w_{ji}^{(1)} \tag{195}$$

which again takes the form (6.11).

6.5 Using the definition of the logistic sigmoid, the swish activation function can be written as

$$\mathrm{sw}(x) = \frac{x}{1 + \exp(-\beta x)}. \tag{196}$$

This is plotted for $\beta = 0.1$, $\beta = 1.0$, and $\beta = 10$ in Figure 2. In the limit $\beta \to \infty$ we can consider the behaviour of the swish function separately for positive and negative values of x. For $x > 0$ the function $\exp(-\beta x) \to 0$ and hence $\mathrm{sw}(x) \to x$, while for $x < 0$ the function $\exp(-\beta x) \to \infty$ and hence $\mathrm{sw}(x) \to 0$. Thus, in this limit the swish function becomes a ReLU function.

6.6 From (6.14), using standard derivatives, we get

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

6.7 The softplus activation function is given by

$$\zeta(a) = \ln\left(1+\exp(a)\right). \tag{197}$$

We can prove the property (6.62) using the following steps

$$\begin{aligned}\zeta(a)-\zeta(-a) &= \ln\left(1+\exp(a)\right) - \ln\left(1+\exp(-a)\right) && (198)\\ &= \ln\left[\exp(a)\left(\exp(-a)+1\right)\right] - \ln\left(1+\exp(-a)\right) && (199)\\ &= a + \ln\left(\exp(-a)+1\right) - \ln\left(1+\exp(-a)\right) && (200)\\ &= a. && (201)\end{aligned}$$

To prove the property (6.63) we have

$$\ln\sigma(a) = -\ln\left(1+\exp(-a)\right) = -\zeta(-a). \tag{202}$$

For the derivative (6.64) of the softplus function we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a}\zeta(a) &= \frac{\mathrm{d}}{\mathrm{d}a}\ln\left(1+\exp(a)\right) \\ &= \frac{\exp(a)}{1+\exp(a)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a).\end{aligned} \tag{203}$$

Finally, to find the inverse (6.65) of the softplus function let $y = \zeta^{-1}(a)$, then

$$a = \zeta(y) = \ln\left(1+\exp(y)\right). \tag{204}$$

Rearranging we obtain

$$\exp(a) = 1+\exp(y) \tag{205}$$

and hence

$$y = \ln(\exp(a)-1). \tag{206}$$

6.8 Differentiating the error (6.25) with respect to σ^2 and setting the derivative to zero gives

$$0 = -\frac{1}{2\sigma^4}\sum_{n=1}^{N}\{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\frac{1}{\sigma^2}. \tag{207}$$

Rearranging to solve for σ^2 we obtain

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\{y(\mathbf{x}_n, \mathbf{w}^\star) - t_n\}^2 \tag{208}$$

as required.

6.9 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 (6.28) is given by

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

If we take the logarithm of this, using (3.26), we get

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

where 'const' comprises terms which are independent of $\mathbf{w}$. The first term on the right hand side is proportional to the negative of (6.29) and hence maximizing the log-likelihood is equivalent to minimizing the sum-of-squares error.

6.10 In this case, the likelihood function becomes

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

with the corresponding log-likelihood function

$$\begin{aligned}
&\ln p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Sigma}) \\
&\qquad = -\frac{N}{2}\left(\ln|\boldsymbol{\Sigma}| + K\ln(2\pi)\right) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}_n),
\end{aligned} \tag{209}$$

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

If we first treat $\boldsymbol{\Sigma}$ as fixed and known, we can drop terms that are independent of $\mathbf{w}$ from (209), and by changing the sign we get the error function

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

If we consider maximizing (209) w.r.t. $\boldsymbol{\Sigma}$, the terms that need to be kept are

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

By rewriting the second term we get

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

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

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

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

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

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

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

The conditional probability of the data label is then

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

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

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

See also Solution **??**.

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

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

The corresponding error function can be constructed from (6.33) by applying the inverse transform to y_n and t_n, yielding

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

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

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

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

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

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

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

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

Taking the negative logarithm in order to derive an error function we obtain (6.36) as required. Note that this is the same result as for the multiclass logistic regression model, given by (5.80) .

6.14 Differentiating (6.33) with respect to the activation a_n corresponding to a particular data point n, we obtain

$$\frac{\partial E}{\partial a_n} = -t_n \frac{1}{y_n} \frac{\partial y_n}{\partial a_n} + (1-t_n) \frac{1}{1-y_n} \frac{\partial y_n}{\partial a_n}. \tag{210}$$

From (5.72), we have

$$\frac{\partial y_n}{\partial a_n} = y_n(1 - y_n). \tag{211}$$

Substituting (211) into (210), we get

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

as required.

6.15 Consider a specific data point n and, to minimize clutter, omit the suffix n on variables such as a_k and y_k. We can use the chain rule of calculus to write

$$\frac{\partial E}{\partial a_k} = \sum_{j=1}^{K} \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial a_k}. \tag{212}$$

From (6.36) we have

$$\frac{\partial E}{\partial y_j} = -\frac{t_j}{y_j}. \tag{213}$$

We can write (6.37) in the form

$$y_j = \frac{\exp(a_j)}{\sum_l \exp(a_l)}. \tag{214}$$

For the derivative $\partial y_j/\partial a_k$ there are two contributions, one from the numerator and one from the denominator, so that

$$\begin{aligned} \frac{\partial y_j}{\partial a_k} &= \frac{\exp(a_j)\delta_{jk}}{\sum_l \exp(a_l)} - \frac{\exp(a_j)\exp(a_k)}{\{\sum_l \exp(a_l)\}^2} \\ &= y_j\delta_{jk} - y_j y_k. \end{aligned} \tag{215}$$

Substituting (213) and (215) into (212) we then have

$$\frac{\partial E}{\partial a_k} = -\sum_{j=1}^{K} \frac{t_j}{y_j}\{y_j\delta_{jk} - y_j y_k\} = y_k - t_k \tag{216}$$

as required. In the final step we have used

$$\sum_{j=1}^{K} t_j = 1 \tag{217}$$

which follows from 1-of-K coding scheme used for the $\{t_j\}$.

6.16 From standard trigonometric rules we get the position of the end of the first arm,

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

Similarly, the position of the end of the second arm relative to the end of the first arm is given by the corresponding equation, with an angle offset of π (see Figure 6.16), which equals a change of sign

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

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

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

6.17 The interpretation of γ_{nk} as a posterior probability follows from Bayes' theorem for the probability of the component indexed by k, given observed data $\mathbf{t}$, in which all quantities are also conditioned on the input variable $\mathbf{x}$. Therefore $\mathbf{x}$ simply appears as a conditioning variable in the right-hand side of all quantities. From Bayes' theorem we have

$$p(k|\mathbf{t}, \mathbf{x}) = \frac{p(\mathbf{t}|k, \mathbf{x})p(k|\mathbf{x})}{p(\mathbf{t}|\mathbf{x})} \tag{218}$$

where, as usual, the denominator can be expressed as a marginalization over the terms in the numerator, so that

$$p(\mathbf{t}|\mathbf{x}) = \sum_l p(\mathbf{t}|l, \mathbf{x})p(l|\mathbf{x}). \tag{219}$$

The quantities $\pi_k(\mathbf{x})$ defined by (6.40) satisfy (6.39) and hence meet the requirements to be viewed as probabilities, and so we equate $p(k|\mathbf{x}) = \pi_k(\mathbf{x})$. Similarly, the class-conditional distribution $p(\mathbf{t}|k, \mathbf{x})$ is given by the Gaussian $\mathcal{N}_{nk} = \mathcal{N}\left(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n), \sigma_k^2(\mathbf{x}_n)\right)$. Substituting into (218) then gives

$$p(k|\mathbf{t}_n, \mathbf{x}_n) = \frac{\pi_k \mathcal{N}_{nk}}{\sum_l \pi_l \mathcal{N}_{nl}} = \gamma_{nk} \tag{220}$$

as required.

6.18 We start by using the chain rule to write

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{221}$$

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

For the first factor inside the sum on the r.h.s. of (221), standard derivatives applied to the n^{th} term of (6.43) gives

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{222}$$

For the for the second factor, we have from (5.78) that

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j(I_{jk} - \pi_k). \tag{223}$$

Combining (221), (222) and (223), we get

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

where we have used the fact that, by (6.44), $\sum_{j=1}^{K} \gamma_{nj} = 1$ for all n.

6.19 Note: see Solution 6.18.

From (6.42) we have

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

and thus

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

From (3.26), (6.43) and (6.44), we get

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

6.20 From (6.41) and (6.43), we see that

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \sigma_k}\frac{\partial \sigma_k}{\partial a_k^\sigma}, \tag{224}$$

where, from (6.41),

$$\frac{\partial \sigma_k}{\partial a_k^\sigma} = \sigma_k. \tag{225}$$

From (3.26), (6.43) and (6.44), we get

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

Combining this with (224) and (225), we get

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

6.21 From (3.42) and (6.38) we have

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

We now introduce the shorthand notation

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

Using this together with (3.42), (3.46), (6.38) and (6.48), we get

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

Chapter 7 Gradient Descent

7.1 Substituting (7.10) into (7.7) we obtain

$$E(\mathbf{w}) = E(\mathbf{w}^\star) + \frac{1}{2}\left(\sum_i \alpha_i \mathbf{u}_i^{\mathrm{T}}\right)\mathbf{H}\left(\sum_j \alpha_j \mathbf{u}_j\right).$$

Making use of (7.8) then gives

$$E(\mathbf{w}) = E(\mathbf{w}^\star) + \frac{1}{2}\left(\sum_i \alpha_i \mathbf{u}_i^{\mathrm{T}}\right)\left(\sum_j \lambda_j \alpha_j \mathbf{u}_j\right).$$

Now making use of (7.9) we have

$$\begin{aligned} E(\mathbf{w}) &= E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \alpha_i \sum_j \lambda_j \alpha_j \delta_{ij} \\ &= E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2 \end{aligned} \tag{226}$$

as required.

7.2 From (7.8) and (7.10) we have

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

Assume that $\mathbf{H}$ is positive definite, so that (7.12) holds. Then by setting $\mathbf{v} = \mathbf{u}_i$ it follows that

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

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

Conversely, assume that () holds. Then, for any vector, $\mathbf{v}$, we can make use of (7.13) to give

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

where we have used (7.8) and (7.9) along with (). Thus, if all of the eigenvalues are positive, the Hessian matrix will be positive definite.

7.3 From (7.12) we see that, if $\mathbf{H}$ is positive definite, then the second term in (7.7) will be positive whenever $(\mathbf{w} - \mathbf{w}^\star)$ is non-zero. Thus the smallest value which $E(\mathbf{w})$ can take is $E(\mathbf{w}^\star)$, and so $\mathbf{w}^\star$ is the minimum of $E(\mathbf{w})$. Conversely, if $\mathbf{w}^\star$ is the minimum of $E(\mathbf{w})$, then, for any vector $\mathbf{w} \neq \mathbf{w}^\star$, $E(\mathbf{w}) > E(\mathbf{w}^\star)$. This will only be the case if the second term of (7.7) is positive for all values of $\mathbf{w} \neq \mathbf{w}^\star$ (since the first term is independent of $\mathbf{w}$). Since $\mathbf{w} - \mathbf{w}^\star$ can be set to any vector of real numbers, it follows from the definition (7.12) that $\mathbf{H}$ must be positive definite.

7.4 The first derivatives of the error function are given by

$$\frac{\partial E}{\partial w} = \sum_n (y_n - t_n) x_n \tag{227}$$

$$\frac{\partial E}{\partial b} = \sum_n (y_n - t_n) \tag{228}$$

where $y_n = y(x_n, w, b)$. The second derivatives are then given by

$$\frac{\partial^2 E}{\partial w^2} = \sum_n x_n^2 \tag{229}$$

$$\frac{\partial^2 E}{\partial w \partial b} = \frac{\partial^2 E}{\partial b \partial w} = \sum_n x_n = N\overline{x} \tag{230}$$

$$\frac{\partial^2 E}{\partial b^2} = \sum_n 1 = N \tag{231}$$

where $\overline{x}$ is the sample mean defined by

$$\overline{x} = \frac{1}{N} \sum_n x_n.$$

The Hessian matrix is given by

$$\mathbf{H} = \begin{pmatrix} \dfrac{\partial^2 E}{\partial w^2} & \dfrac{\partial^2 E}{\partial b \partial w} \\ \dfrac{\partial^2 E}{\partial w \partial b} & \dfrac{\partial^2 E}{\partial b^2} \end{pmatrix} = \begin{pmatrix} \displaystyle\sum_n x_n^2 & N\overline{x} \\ N\overline{x} & N \end{pmatrix}.$$

Note that the Hessian does not depend on the target values for this simple model, nor does it depend on the model parameters. It is a function only of the input data variables. The trace and determinant of the Hessian are given by

$$\mathrm{Tr}\,\mathbf{H} = \sum_n x_n^2 + N \tag{232}$$

$$\det \mathbf{H} = N \sum_n x_n^2 - (N\overline{x})^2 = N^2 \sigma^2 \tag{233}$$

where σ^2 is the sample variance defined by

$$\sigma^2 = \frac{1}{N}\sum_n (x_n - \overline{x})^2.$$

We see that the trace and the determinant are positive. Since the determinant is the product of the two eigenvalues of the Hessian it follows that either both eigenvalues are positive or they are both negative. Since the trace is the sum of the eigenvalues, and the trace is also positive, it follows that both eigenvalues must be positive and so the Hessian must be positive definite. Here we ignore the degenerate case where the number of data points is one or where all the data points are the same.

7.5 Note that the original printing of the book there is a minus sign missing from the right-hand side of (7.64). We first note that the derivative of the logistic sigmoid function is given by (5.18). Using this result, the first derivatives of the error function are given by

$$\begin{aligned}\frac{\partial E}{\partial w} &= -\sum_n \left\{\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right\} y_n(1-y_n)x_n \\ &= \sum_n (y_n - t_n)x_n, \end{aligned} \tag{234}$$

$$\begin{aligned}\frac{\partial E}{\partial b} &= -\sum_n \left\{\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right\} y_n(1-y_n) \\ &= \sum_n (y_n - t_n) \end{aligned} \tag{235}$$

where $y_n = y(x_n, w, b)$. The second derivatives are then given by

$$\frac{\partial^2 E}{\partial w^2} = \sum_n y_n(1-y_n)x_n^2 \tag{236}$$

$$\frac{\partial^2 E}{\partial w \partial b} = \frac{\partial^2 E}{\partial b \partial w} = \sum_n y_n(1-y_n)x_n \tag{237}$$

$$\frac{\partial^2 E}{\partial b^2} = \sum_n y_n(1-y_n). \tag{238}$$

The Hessian matrix is given by

$$\begin{aligned}\mathbf{H} &= \begin{pmatrix} \dfrac{\partial^2 E}{\partial w^2} & \dfrac{\partial^2 E}{\partial b \partial w} \\ \dfrac{\partial^2 E}{\partial w \partial b} & \dfrac{\partial^2 E}{\partial b^2} \end{pmatrix} \\ &= \begin{pmatrix} \displaystyle\sum_n y_n(1-y_n)x_n^2 & \displaystyle\sum_n y_n(1-y_n)x_n \\ \displaystyle\sum_n y_n(1-y_n)x_n & \displaystyle\sum_n y_n(1-y_n) \end{pmatrix}. \end{aligned} \tag{239}$$

Note that the Hessian does not depend on the target values for this simple model, but it is a function of the model parameters w and b, corresponding to the fact that the error function is non-quadratic. Since the logistic sigmoid function satisfies $0 < \sigma(\cdot) < 1$ we see that $y_n(1 - y_n)$ is always a positive quantity. We therefore see that the elements of the leading diagonal of the Hessian are given by the sum of positive terms and are therefore themselves positive. Thus the trace of the Hessian is positive. Note that we ignore the degenerate case where all of the data points are identical, leading to a trace of zero, since this is of no practical interest. For the determinant we first define $c_n = y_n(1-y_n)$ in order to keep the notation uncluttered. We then have

$$\begin{aligned}\det \mathbf{H} &= \left(\sum_n c_n x_n^2\right)\left(\sum_n c_n\right) - \left(\sum_n c_n x_n\right)^2 \\ &= \left(\sum_n c_n\right)\sum_n c_n \left\{ x_n - \frac{\left(\sum_n c_n x_n\right)}{\left(\sum_n c_n\right)} \right\}^2 . \end{aligned} \tag{240}$$

Thus the determinant comprises the sum of terms each of which is positive and hence is itself positive, and hence both the trace and the determinant are positive. Since the determinant is the product of the two eigenvalues of the Hessian it follows that either both eigenvalues are positive or they are both negative. Since the trace is the sum of the eigenvalues, and the trace is also positive, it follows that both eigenvalues must be positive and so the Hessian must be positive definite.

7.6 We start by making the change of variable given by (7.10) which allows the error function to be written in the form (7.11). Setting the value of the error function $E(\mathbf{w})$ to a constant value C we obtain

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

Re-arranging gives

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

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

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

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

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

$$\frac{W^2 - W}{2}$$

as the number of independent of-diagonal elements. If we now add back the W elements on the diagonal we get

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

Finally, we add the W elements of the gradient vector $\mathbf{b}$ to give

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

7.8 From the property (2.52) of the Gaussian distribution we have, for a single data point,

$$\mathbb{E}[x_n] = \mu.$$

Two independent data points will be uncorrelated and hence

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2 \quad \text{if} \quad n \neq m.$$

Therefore, using (2.53) it follows that

$$\mathbb{E}[x_n x_m] = \delta_{nm}\sigma^2 + \mu^2.$$

Using the definition (7.65) of $\overline{x}$ we have

$$\begin{aligned}
\mathbb{E}[\overline{x}] &= \mu \\
\mathbb{E}[\overline{x}^2] &= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[x_n x_m] \\
&= \frac{1}{N}\sigma^2 + \mu^2.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}[(\overline{x} - \mu)^2] &= \mathbb{E}[\overline{x}^2 - 2\overline{x}\mu + \mu^2] \\
&= \frac{1}{N}\sigma^2 + \mu^2 - 2\mu^2 + \mu^2 = \frac{\sigma^2}{N}.
\end{aligned}$$

7.9 Note that in the original printing, the left-hand side of (7.20) should be $\mathrm{var}[a_i^{(l)}]$ instead of $\mathrm{var}[z_j^{(l)}]$. The expectation over $a_i^{(l)}$ involves averaging over both the distribution of w_{ij} and the distribution of $z_j^{(l-1)}$. For a given value of $z_j^{(l-1)}$ the quantity

$w_{ij}z_j^{(l-1)}$ has zero mean since we are assuming that the weights w_{ij} are drawn from a zero-mean Gaussian $\mathcal{N}(0, \epsilon^2)$. Since this is true for every value of j in the summation we have

$$\mathbb{E}[a_i^{(l)}] = 0. \tag{241}$$

To find the variance of $a_i^{(l)}$ we note that the quantity $a_i^{(l)}$ comprises the sum of M terms which are themselves independent random variables, and we know from (2.121) that the total variance of a sum of independent variables is the sum of the variances of the individual variables. Hence we have

$$\begin{aligned}
\text{var}[a_i^{(l)}] &= M\text{var}[w_{ij}z_j^{(l-1)}] \\
&= M\,\mathbb{E}_{w,z}[(w_{ij}z_j^{(l-1)})^2] - M\mathbb{E}_{w,z}[w_{ij}z_j^{(l-1)}]^2 \\
&= M\,\mathbb{E}_{w,z}[(w_{ij}z_j^{(l-1)})^2] \\
&= M\,\mathbb{E}_w[(w_{ij})^2]\,\mathbb{E}_z[(z_j^{(l-1)})^2]
\end{aligned}$$

since $\mathbb{E}_{w,z}[w_{ij}z_j^{(l-1)}] = 0$ as discussed above. Because the weights w_{ij} are drawn from a Gaussian $\mathcal{N}(0, \epsilon^2)$ we have $\mathbb{E}[(w_{ij})^2] = \epsilon^2$. To find $\mathbb{E}[(z_j^{(l-1)})^2]$ we note that

$$z_j^{(l-1)} = \text{ReLU}(a_i^{(l-1)})$$

and therefore

$$(z_j^{(l-1)})^2 = \text{ReLU}(a_i^{(l-1)})^2.$$

When we take the square of the ReLU, the result will be zero if the argument is less zero or negative, and will be the square of the argument if the argument is positive. The quantity $a_i^{(l-1)}$ will have a symmetric distribution about 0 since, for every value of $z_j^{(l-1)}$, the product $w_{ij}z_j^{(l-1)}$ will have a symmetric Gaussian distribution, and so the overall distribution is the sum of symmetric distributions and is therefore itself symmetric. Thus, when we take the expectation of the ReLU of this quantity, half the terms will contribute zero and half the terms will contribute the square of the argument and hence

$$\begin{aligned}
\mathbb{E}[(z_j^{(l-1)})^2] &= \mathbb{E}[(\text{ReLU}(a_i^{(l-1)}))^2] \\
&= \frac{1}{2}\mathbb{E}[(a_i^{(l-1)})^2] \\
&= \frac{1}{2}\text{var}[(a_i^{(l-1)})^2] \\
&= \frac{1}{2}\lambda^2
\end{aligned}$$

where we have used the property that $a_i^{(l-1)}$ has zero mean. Combining these results gives

$$\text{var}[a_i^{(l)}] = \frac{M}{2}\epsilon^2\lambda^2$$

as required. Finally, if the variance of $a_i^{(l)}$ is also to equal λ^2 then

$$\frac{M}{2}\epsilon^2\lambda^2 = \lambda^2$$

from which it follows that ϵ must be chosen to have the value

$$\epsilon = \sqrt{\frac{2}{M}}. \tag{242}$$

7.10 Taking the gradient of (7.7) we obtain

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star). \tag{243}$$

Substituting for $(\mathbf{w} - \mathbf{w}^\star)$ using (7.10) then gives

$$\nabla E = \mathbf{H}\left(\sum_i \alpha_i \mathbf{u}_i\right). \tag{244}$$

Making use of (7.8) we have

$$\nabla E = \sum_i \alpha_i \lambda_i \mathbf{u}_i. \tag{245}$$

If we consider a change $\Delta\alpha_i$ in the coefficients α_i then the corresponding change in $\mathbf{w}$ is obtained by taking finite differences of (7.10) to give

$$\Delta\mathbf{w} = \sum_i \Delta\alpha_i \mathbf{u}_i. \tag{246}$$

Next, using (7.16) we have

$$\Delta\mathbf{w} = -\eta\nabla E. \tag{247}$$

Substituting on both sides then gives

$$\sum_i \Delta\alpha_i \mathbf{u}_i = -\eta \sum_i \alpha_i \lambda_i \mathbf{u}_i. \tag{248}$$

If we multiply both sides by $\mathbf{u}_j^{\mathrm{T}}$ and make use of the orthonormality relation (7.9) we finally obtain

$$\Delta\alpha_j = -\eta\alpha_j\lambda_j \tag{249}$$

as required.

7.11 For small values of μ we can make a Taylor expansion of the first term on the right-hand side of (7.34) in powers of μ to give

$$\begin{aligned}\Delta\mathbf{w}^{(\tau-1)} = &-\eta\nabla E\left(\mathbf{w}^{(\tau-1)}\right) - \eta\mu\nabla\nabla E\left(\mathbf{w}^{(\tau-1)}\right)\Delta\mathbf{w}^{(\tau-2)} + \mathcal{O}(\mu^2) \\ &+ \mu\Delta\mathbf{w}^{(\tau-2)}.\end{aligned}$$

If we now assume that $\eta = \mathcal{O}(\epsilon)$ and $\mu = \mathcal{O}(\epsilon)$, then we can neglect higher-order terms in the Taylor expansion and also omit the term with coefficient $\eta\mu$ since that is $\mathcal{O}(\epsilon^2)$. Here we have assumed that the error surface is slowly varying and hence that the Hessian term $\nabla\nabla E$ is $\mathcal{O}(1)$. We then obtain the standard formula for gradient descent with momentum defined by (7.31).

7.12 If we apply (7.66) recursively we obtain

$$\begin{aligned}
\mu_n &= \beta\mu_{n-1} + (1-\beta)x_n \\
&= \beta\left(\beta\mu_{n-2} + (1-\beta)x_{n-1}\right) + (1-\beta)x_n \\
&= \beta^2\mu_{n-2} + \beta(1-\beta)x_{n-1} + (1-\beta)x_n \\
&= \beta^3\mu_{n-3} + \beta^2(1-\beta)x_{n-2} + \beta(1-\beta)x_{n-1} + (1-\beta)x_n \\
&= \ldots \\
&= \beta^n\mu_0 + \sum_{k=1}^{n}\beta^{k-1}(1-\beta)x_{n-k+1}.
\end{aligned}$$

We now set $\mu_0 = 0$, and then take the expectation of both sides with respect to the distribution of x, noting that the $\{x_n\}$ are independent, identically distributed samples from this distribution. This gives

$$\begin{aligned}
\mathbb{E}[\mu_n] &= \sum_{k=1}^{n}\beta^{k-1}(1-\beta)\mathbb{E}[x_{n-k+1}] \\
&= \sum_{k=1}^{n}\beta^{k-1}(1-\beta)\overline{x}
\end{aligned}$$

where $\overline{x}$ is the true mean of the distribution of x. Making use of the result (7.67) we can write this as

$$\mathbb{E}[\mu_n] = (1-\beta^n)\overline{x}.$$

Thus we see that $\mathbb{E}[\mu_n] \neq \overline{x}$ and hence that the estimate μ_n is biased. This bias is easily corrected by using the estimator

$$\widehat{\mu}_n = \frac{\mu_n}{1-\beta^n}. \tag{250}$$

which has the property $\mathbb{E}[\mu_n] = \overline{x}$ and is therefore unbiased.

7.13 Setting the derivative of (7.69) with respect to λ equal to zero we have

$$\frac{\partial}{\partial\lambda}E(\mathbf{w}^{(\tau)} + \lambda\mathbf{d}) = 0. \tag{251}$$

To evaluate this derivative we can use the chain rule of calculus. Define

$$\mathbf{v} = \mathbf{w}^{(\tau)} + \lambda\mathbf{d}. \tag{252}$$

Then we have

$$\begin{aligned}\frac{\partial}{\partial\lambda}E &= \sum_{i=1}^{M}\frac{\partial v_i}{\partial\lambda}\frac{\partial E}{\partial v_i} \\ &= \sum_{i=1}^{M} d_i\frac{\partial E}{\partial v_i} \\ &= \mathbf{d}^{\mathrm{T}}\nabla E \end{aligned} \tag{253}$$

where $\{d_i\}$ are the components of $\mathbf{d}$. This derivative vanishes for a particular value $\lambda^\star$ which then defines the new location in weight space

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \lambda^\star\mathbf{d}. \tag{254}$$

We therefore have

$$\mathbf{d}^{\mathrm{T}}\nabla E(\mathbf{w}^{(\tau)} + \lambda^\star\mathbf{d}) = \mathbf{d}^{\mathrm{T}}\nabla E(\mathbf{w}^{(\tau+1)}) = 0 \tag{255}$$

and hence the gradient of the error function at the line-search minimum is orthogonal to the search direction.

7.14 Summing both sides of (7.50) over n to compute the sample mean we have

$$\frac{1}{N}\sum_{n=1}^{N}\widetilde{x}_{ni} = \frac{1}{N\sigma_i}\left(\sum_{n=1}^{N}x_{ni} - N\mu_i\right) = 0 \tag{256}$$

where we have used (7.48). Similarly, if we consider the sample variance

$$\frac{1}{N}\sum_{n=1}^{N}(\widetilde{x}_{ni} - 0)^2 = \frac{1}{N\sigma_i^2}\sum_{n=1}^{N}(x_{ni} - \mu_i)^2 = 1 \tag{257}$$

where we have made use of (256) together with (7.49).

Chapter 8 Backpropagation

8.1 From (8.12) we have

$$\delta_j \equiv \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial a_j} \tag{258}$$

which follows from the chain rule of probability. Then (8.8) gives

$$\delta_k = \frac{\partial E_n}{\partial a_k}. \tag{259}$$

Again using the chain rule we have

$$\begin{aligned}\frac{\partial a_k}{\partial a_j} &= \frac{\partial a_k}{\partial z_j}\frac{\partial z_j}{\partial a_j} \\ &= w_{kj}\frac{\partial z_j}{\partial a_j} \\ &= w_{kj}h'(a_j)\end{aligned} \tag{260}$$

where we have used (8.5) and (8.6). Substituting these results into (258) we obtain

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k \tag{261}$$

as required.

8.2 The forward propagation equations in matrix notation are given by (6.19) in the form

$$\mathbf{z}^{(l)} = h^{(l)}\left(\mathbf{W}^{(l)}\mathbf{z}^{(l-1)}\right) \tag{262}$$

where $\mathbf{W}^{(l)}$ is a matrix with elements $\mathbf{w}_{jk}^{(l)}$ comprising the weights in layer l of the network, and the activation function $h^{(l)}(\cdot)$ acts on each element of its vector argument independently. If we define $\boldsymbol{\delta}^{(l)}$ to be the errors vector with elements δ_j, then the backpropagation equations in matrix notation take the form

$$\boldsymbol{\delta}^{(l-1)} = h^{(l-1)\prime}\left(\mathbf{a}^{(l-1)}\right) \odot \left\{\left(\mathbf{W}^{(l-1)}\right)^{\mathrm{T}}\boldsymbol{\delta}^{(l)}\right\} \tag{263}$$

where $\odot$ denotes the Hadamard product which comprises the element-wise multiplication of two vectors. Note that the forward propagation equation (8.5) involves a summation over the second index of w_{ji} whereas the backpropagation equation (8.13) involves a summation over the first index. Hence when we write the backpropagation equation in matrix notation it involves the transpose of the matrix that appears in the forward propagation equation.

8.3 We are interested in determining how the correction term

$$\delta = E'(w_{ij}) - \frac{E(w_{ij} + \epsilon) - E(w_{ij} - \epsilon)}{2\epsilon} \tag{264}$$

depend on ϵ. Using Taylor expansions, we can rewrite the numerator of the first term of (264) as

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

Note that the ϵ^2 terms cancel. Substituting this into (264) we get,

$$\delta = \frac{2\epsilon E'(w_{ij}) + \mathcal{O}(\epsilon^3)}{2\epsilon} - E'(w_{ij}) = \mathcal{O}(\epsilon^2). \tag{265}$$

8.4 If we introduce skip layer weights, $\mathbf{U}$, into the model described in Section 8.1.3, this will only affect the last of the forward propagation equations, (8.20), which becomes

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} u_{ki} x_i. \tag{266}$$

Note that there is no need to include the input bias. The derivative w.r.t. u_{ki} can be expressed using the output $\{\delta_k\}$ of (8.21),

$$\frac{\partial E}{\partial u_{ki}} = \delta_k x_i. \tag{267}$$

8.5 The alternative forward propagation scheme takes the first line of (8.29) as its starting point. However, rather than proceeding with a 'recursive' definition of $\partial y_k / \partial a_j$, we instead make use of a corresponding definition for $\partial a_j / \partial x_i$. More formally

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_j \frac{\partial y_k}{\partial a_j} \frac{\partial a_j}{\partial x_i} \tag{268}$$

where $\partial y_k / \partial a_j$ is defined by (8.33) for logistic sigmoid output units, (8.34) for softmax output units, or simply as δ_{kj}, for the case of linear output units. We define $\partial a_j / \partial x_i = w_{ji}$ if a_j is in the first hidden layer and otherwise

$$\frac{\partial a_j}{\partial x_i} = \sum_l \frac{\partial a_j}{\partial a_l} \frac{\partial a_l}{\partial x_i} \tag{269}$$

where

$$\frac{\partial a_j}{\partial a_l} = w_{jl} h'(a_l). \tag{270}$$

Thus we can evaluate J_{ki} by forward propagating $\partial a_j / \partial x_i$, with initial value w_{ij}, alongside a_j, using (269) and (270).

8.6 Using the chain rule together with (8.5) and (8.77), we have

$$\begin{aligned}\frac{\partial E_n}{\partial w_{kj}^{(2)}} &= \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}^{(2)}} \\ &= \delta_k z_j. \end{aligned} \tag{271}$$

Thus,

$$\frac{\partial^2 E_n}{\partial w_{kj}^{(2)}\,\partial w_{k'j'}^{(2)}} = \frac{\partial \delta_k z_j}{\partial w_{k'j'}^{(2)}} \tag{272}$$

and since z_j is independent of the second layer weights,

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

where we again have used the chain rule together with (8.5) and (8.77). If both weights are in the first layer, we again used the chain rule, this time together with (8.5), (8.12) and (8.13), to get

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

Thus we have

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{j'i'}^{(1)}} = \frac{\partial}{\partial w_{j'i'}}\left(x_i\,h'(a_j)\sum_k w_{kj}^{(2)}\delta_k\right). \tag{273}$$

Now we note that x_i and $w_{kj}^{(2)}$ do not depend on $w_{j'i'}^{(1)}$, while $h'(a_j)$ is only affected in the case where $j = j'$. Using these observations together with (8.5), we get

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{j'i'}^{(1)}} = x_i x_{i'} h''(a_j) I_{jj'}\sum_k w_{kj}^{(2)}\delta_k + x_i h'(a_j)\sum_k w_{kj}^{(2)}\frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}}. \tag{274}$$

From (8.5), (8.12), (8.13), (8.77) and the chain rule, we have

$$\begin{aligned}\frac{\partial \delta_k}{\partial w_{j'i'}^{(1)}} &= \sum_{k'}\frac{\partial^2 E_n}{\partial a_k\,\partial a_{k'}}\frac{\partial a_{k'}}{\partial a_{j'}}\frac{\partial a_{j'}}{\partial w_{j'i'}^{(1)}} \\ &= x_{i'}h'(a_j)\sum_{k'} w_{k'j'}^{(2)} M_{kk'}.\end{aligned} \tag{275}$$

Substituting this back into (274), we obtain (**??**). Finally, from (271) we have

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)}\,\partial w_{kj'}^{(2)}} = \frac{\partial \delta_k z_{j'}}{\partial w_{ji}^{(1)}}. \tag{276}$$

Using (275), we get

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

8.7 If we introduce skip layer weights into the model discussed in Section **??**, three new cases are added to three already covered in Exercise 8.6. The first derivative w.r.t. skip layer weight u_{ki} can be written

$$\frac{\partial E_n}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial u_{ki}} = \frac{\partial E_n}{\partial a_k} x_i. \tag{277}$$

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

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

where we have also used (8.77). When one weight is in the skip layer and the other weight is in the hidden-to-output layer, we can use (277), (8.5) and (8.77) to get

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

Finally, if one weight is a skip layer weight and the other is in the input-to-hidden layer, (277), (8.5), (8.12), (8.13) and (8.77) together give

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

8.8 The multivariate form of (8.38) is

$$E = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}(\mathbf{y}_n - \mathbf{t}_n). \tag{278}$$

The elements of the first and second derivatives then become

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i} \tag{279}$$

and

$$\frac{\partial^2 E}{\partial w_i \partial w_j} = \sum_{n=1}^{N}\left\{\frac{\partial \mathbf{y}_n}{\partial w_j}^{\mathrm{T}}\frac{\partial \mathbf{y}_n}{\partial w_i} + (\mathbf{y}_n - \mathbf{t}_n)^{\mathrm{T}}\frac{\partial^2 \mathbf{y}_n}{\partial w_j\,\partial w_i}\right\}. \tag{280}$$

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

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{B}_n\mathbf{B}_n^{\mathrm{T}}, \tag{281}$$

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

$$(\mathbf{B}_n)_{lk} = \frac{\partial y_{nk}}{\partial w_l}. \tag{282}$$

8.9 Taking the second derivatives of (8.78) with respect to two weights w_r and w_s we obtain

$$\begin{aligned}\frac{\partial^2 E}{\partial w_r \partial w_s} &= \sum_k \int \left\{\frac{\partial y_k}{\partial w_r}\frac{\partial y_k}{\partial w_s}\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &+ \sum_k \int \left\{\frac{\partial^2 y_k}{\partial w_r \partial w_s}(y_k(\mathbf{x}) - \mathbb{E}_{t_k}[t_k|\mathbf{x}])\right\} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{283}$$

Using the result (4.37) that the outputs $y_k(\mathbf{x})$ of the trained network represent the conditional averages of the target data, we see that the second term in (283) vanishes. The Hessian is therefore given by an integral of terms involving only the products of first derivatives. For a finite data set, we can write this result in the form

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \frac{1}{N}\sum_{n=1}^{N}\sum_k \frac{\partial y_k^n}{\partial w_r}\frac{\partial y_k^n}{\partial w_s} \tag{284}$$

which is identical with (8.40) up to a scaling factor.

8.10 If we take the gradient of (6.33) with respect to $\mathbf{w}$, we obtain

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

where we have used the result proved earlier in the solution to Exercise 6.14. Taking the second derivatives we have

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

Dropping the last term and using the result (5.72) for the derivative of the logistic sigmoid function, proved in the solution to Exercise 5.18, we finally get

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

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

8.11 Using the chain rule, we can write the first derivative of (6.36) as

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\partial E}{\partial a_{nk}} \frac{\partial a_{nk}}{\partial w_i}. \tag{285}$$

From Exercise 6.15, we know that

$$\frac{\partial E}{\partial a_{nk}} = y_{nk} - t_{nk}. \tag{286}$$

Using this and (5.78), we can get the derivative of (285) w.r.t. w_j as

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

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

$$(\mathbf{H})_{ij} \simeq \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} y_{nk}(I_{kl} - y_{nl}) \frac{\partial a_{nk}}{\partial w_i} \frac{\partial a_{nl}}{\partial w_j}. \tag{287}$$

8.12 Suppose we have already obtained the inverse Hessian using the first L data points. By separating off the contribution from data point $L + 1$ in (8.40), we obtain

$$\mathbf{H}_{L+1} = \mathbf{H}_L + \nabla a_{L+1} \nabla a_{L+1}^{\mathrm{T}}. \tag{288}$$

We now consider the matrix identity (8.80). If we now identify $\mathbf{H}_L$ with $\mathbf{M}$ and $\mathbf{b}_{L+1}$ with $\mathbf{v}$, we obtain

$$\mathbf{H}_{L+1}^{-1} = \mathbf{H}_L^{-1} - \frac{\mathbf{H}_L^{-1} \nabla a_{L+1} \nabla a_{L+1}^{\mathrm{T}} \mathbf{H}_L^{-1}}{1 + \nabla a_{L+1}^{\mathrm{T}} \mathbf{H}_L^{-1} \nabla a_{L+1}}. \tag{289}$$

In this way, data points are sequentially absorbed until $L+1 = N$ and the whole data set has been processed. This result therefore represents a procedure for evaluating the inverse of the Hessian using a single pass through the data set. The initial matrix $\mathbf{H}_0$ is chosen to be $\alpha\mathbf{I}$, where α is a small quantity, so that the algorithm actually finds the inverse of $\mathbf{H} + \alpha\mathbf{I}$. The results are not particularly sensitive to the precise value of α.

8.13 The function $h(\cdot)$ is given by a soft ReLU

$$h(a) = \ln(1 + \exp(a)) \tag{290}$$

and its derivative is given by

$$h'(a) = \frac{\exp(a)}{1 + \exp(a)}. \tag{291}$$

We can now use the chain rule of calculus in the form

$$\frac{\mathrm{d}y}{\mathrm{d}w_1} = \frac{\mathrm{d}y}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}w_1}. \tag{292}$$

Using (8.44) and (291) we have

$$\frac{\mathrm{d}z}{\mathrm{d}w_1} = \frac{\exp(w_1 x + b_1)}{1 + \exp(w_1 x + b_1)} x. \tag{293}$$

Similarly, using (8.45) and (291) we have

$$\frac{\mathrm{d}y}{\mathrm{d}z} = \frac{\exp(w_2 z + b_2)}{1 + \exp(w_2 z + b_2)} w_2. \tag{294}$$

Finally, combining these derivatives using the chain rule, and then substituting for z using (8.44), we obtain

$$\frac{\partial y}{\partial w_1} = \frac{w_2 x \exp\left(w_1 x + b_1 + b_2 + w_2 \ln[1 + \exp(w_1 x + b_1)]\right)}{(1 + \exp(w_1 x + b_1))\left(1 + \exp(b_2 + w_2 \ln[1 + \exp(w_1 x + b_1)])\right)}. \tag{295}$$

8.14 The evaluation trace equations are given directly from the definition of the logistic map

$$L_1 = x \tag{296}$$
$$L_2 = 4L_1(1 - L_1) \tag{297}$$
$$L_3 = 4L_2(1 - L_2) \tag{298}$$
$$L_4 = 4L_3(1 - L_3). \tag{299}$$
$$\tag{300}$$

The corresponding explicit functions, without simplification, are then given by

$$L_1(x) = x \tag{301}$$
$$L_2(x) = 4x(1-x) \tag{302}$$
$$L_3(x) = 16x(1-x)(1-2x)^2 \tag{303}$$
$$L_4(x) = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2. \tag{304}$$

Finally, taking derivatives, we obtain the following expressions, again without simplification

$$L_1'(x) = 1 \tag{305}$$
$$L_2'(x) = 4(1-x) - 4x \tag{306}$$
$$L_3'(x) = 16(1-x)(1-2x)^2 - 16x(1-2x)^2 - 64x(1-x)(1-2x) \tag{307}$$
$$\begin{aligned} L_4'(x) =& 128x(1-x)(-8+16x)(1-2x)^2(1-8x+8x^2) \\ &+ 64(1-x)(1-2x)^2(1-8x+8x^2)^2 \\ &- 64x(1-2x)^2(1-8x+8x^2)^2 \\ &- 256x(1-x)(1-2x)(1-8x+8x^2)^2. \end{aligned} \tag{308}$$

Note that the complexity of the expressions for the derivatives grows much faster than the complexity of the expressions for the corresponding functions.

8.15 To derive the forward-mode equations we apply (8.57) in the form

$$\dot{v}_i = \sum_{j \in \mathrm{pa}(i)} \dot{v}_j \frac{\partial v_i}{\partial v_j}$$

where $\mathrm{pa}(i)$ denotes parents of the node i in the evaluation trace diagram. Using the evaluation trace diagram in Figure 8.4, together with (8.50) to (8.56), we then have

$$\dot{v}_1 = 1 \tag{309}$$
$$\dot{v}_2 = 0 \tag{310}$$
$$\dot{v}_3 = \dot{v}_1 \frac{\partial v_3}{\partial v_1} + \dot{v}_2 \frac{\partial v_3}{\partial v_2} = \dot{v}_1 v_2 + \dot{v}_2 v_1 \tag{311}$$
$$\dot{v}_4 = \dot{v}_2 \frac{\partial v_4}{\partial v_2} = \dot{v}_2 \cos(v_2) \tag{312}$$
$$\dot{v}_5 = \dot{v}_3 \frac{\partial v_5}{\partial v_3} = \dot{v}_3 \exp(v_3) \tag{313}$$
$$\dot{v}_6 = \dot{v}_3 \frac{\partial v_6}{\partial v_3} + \dot{v}_4 \frac{\partial v_6}{\partial v_4} = \dot{v}_3 - \dot{v}_4 \tag{314}$$
$$\dot{v}_7 = \dot{v}_5 \frac{\partial v_7}{\partial v_5} + \dot{v}_6 \frac{\partial v_7}{\partial v_6} = \dot{v}_5 + \dot{v}_6. \tag{315}$$

8.16 To derive the reverse-mode equations we apply (8.69) in the form

$$\overline{v}_i = \sum_{j \in \mathrm{ch}(i)} \overline{v}_j \frac{\partial v_j}{\partial v_i}. \tag{316}$$

Here $\mathrm{ch}(i)$ denotes the children of node i in the evaluation trace graph. Using the evaluation trace diagram in Figure 8.4, together with (8.50) to (8.56), and starting at the output of the graph and working backwards we then have

$$\overline{v}_7 = 1 \tag{317}$$

$$\overline{v}_6 = \overline{v}_7 \frac{\partial v_7}{\partial v_6} = \overline{v}_7 \tag{318}$$

$$\overline{v}_5 = \overline{v}_7 \frac{\partial v_7}{\partial v_5} = \overline{v}_7 \tag{319}$$

$$\overline{v}_4 = \overline{v}_6 \frac{\partial v_6}{\partial v_4} = -\overline{v}_6 \tag{320}$$

$$\overline{v}_3 = \overline{v}_5 \frac{\partial v_5}{\partial v_3} + \overline{v}_6 \frac{\partial v_6}{\partial v_3} = \overline{v}_5 v_5 + \overline{v}_6 \tag{321}$$

$$\overline{v}_2 = \overline{v}_3 \frac{\partial v_3}{\partial v_2} + \overline{v}_4 \frac{\partial v_4}{\partial v_2} = \overline{v}_3 v_1 + \overline{v}_4 \cos(v_2) \tag{322}$$

$$\overline{v}_1 = \overline{v}_3 \frac{\partial v_3}{\partial v_1} = \overline{v}_3 v_2. \tag{323}$$

8.17 From (8.49) we have the following expression for the partial derivative

$$\frac{\partial f}{\partial x_1} = x_2 + x_2 \exp(x_1 x_2). \tag{324}$$

Evaluating this for $(x_1 x_2) = (1, 2)$ gives

$$\left.\frac{\partial f}{\partial x_1}\right|_{x_1=1, x_2=2} = 2 + 2\exp(2). \tag{325}$$

From the evaluation trace equations (8.50) to (8.56) we have

$$v_1 = 1 \tag{326}$$

$$v_2 = 2 \tag{327}$$

$$v_3 = 2 \tag{328}$$

$$v_4 = \sin(2) \tag{329}$$

$$v_5 = \exp(2) \tag{330}$$

$$v_6 = 2 - \sin(2) \tag{331}$$

$$v_7 = 2 + \exp(2) - \sin(2). \tag{332}$$

For the tangent variables we can then use (8.58) to (8.64) to give

$$\dot{v}_1 = 1 \tag{333}$$
$$\dot{v}_2 = 0 \tag{334}$$
$$\dot{v}_3 = 2 \tag{335}$$
$$\dot{v}_4 = 0 \tag{336}$$
$$\dot{v}_5 = 2\exp(2) \tag{337}$$
$$\dot{v}_6 = 2 \tag{338}$$
$$\dot{v}_7 = 2 + 2\exp(2) \tag{339}$$

and so we see that $\dot{v}_7$ does indeed represent the correct value for the derivative given by (325). Similarly, we can use the evaluation trace equations of reverse-mode automatic differentiation (8.70) to (8.76) to evaluate the adjoint variables as follows

$$\overline{v}_7 = 1 \tag{340}$$
$$\overline{v}_6 = 1 \tag{341}$$
$$\overline{v}_5 = 1 \tag{342}$$
$$\overline{v}_4 = -1 \tag{343}$$
$$\overline{v}_3 = \exp(2) + 1 \tag{344}$$
$$\overline{v}_2 = (\exp(2) + 1) - \cos(2) \tag{345}$$
$$\overline{v}_1 = 2\exp(2) + 2. \tag{346}$$

From (8.68) we have

$$\overline{v}_1 = \frac{\partial f}{\partial x_1} \tag{347}$$

and so again we see that this agrees with the required derivative.

8.18 The vectors $\mathbf{e}_1, \ldots, \mathbf{e}_D$ form a complete orthonormal basis and so we can expand an arbitrary D-dimensional vector $\mathbf{r}$ in the form

$$\mathbf{r} = \sum_{i=1}^{D} \alpha_i \mathbf{e}_i. \tag{348}$$

Taking the product of both sides with $\mathbf{e}_j^{\mathrm{T}}$ we obtain

$$r_j = \mathbf{e}_j^{\mathrm{T}} \mathbf{r} = \alpha_j \tag{349}$$

where r_j is the jth component of $\mathbf{r}$. Hence we can write the expansion in the form

$$\mathbf{r} = \sum_{i=1}^{D} r_i \mathbf{e}_i. \tag{350}$$

Multiplying both sides by the Jacobian then gives

$$\mathbf{J}\mathbf{r} = \sum_{i=1}^{D} r_i \mathbf{J}\mathbf{e}_i = \sum_{i=1}^{D} r_i \frac{\partial \mathbf{f}}{\partial x_i} \tag{351}$$

where $\mathbf{f}(\mathbf{x})$ is the original network function with elements $f_k(\mathbf{x})$. This can be interpreted as a single pass of forward-mode automatic differentiation in which the tangent variables associated with the input variables are given by $\dot{x}_i = r_i$.

One way to see this more clearly is to introduce a function $\mathbf{g}(z)$ where z is a scalar variable and the elements of $\mathbf{g}$ are given by $g_i(z) = r_i z$. From the perspective of a network diagram this can be viewed as introducing an extra layer from a single input z to the original inputs $\{x_i\}$. The overall composite function can be written as $\mathbf{f}(\mathbf{g}(z))$, which is now a function with just one input whose Jacobian is therefore a matrix with a single column which can therefore be evaluated in a single pass of forward-mode automatic differentiation. The elements of this vector are given by

$$\frac{\partial f_k}{\partial z} = \sum_{i=1}^{D} \frac{\partial f_k}{\partial x_i} \frac{\partial x_i}{\partial z} = \sum_{i=1}^{D} J_{ki} r_i = (\mathbf{J}\mathbf{r})_k \tag{352}$$

and are therefore the elements of the Jacobian-vector product as required. The tangent variables at the inputs to the main network are then given by

$$\dot{x}_i = \frac{\partial x_i}{\partial z} = r_i. \tag{353}$$

Thus, we see that if the tangent variable $\dot{x}_i$ for each input i is set to the corresponding element r_i of $\mathbf{r}$, then a single pass of forward-mode automatic differentiation will compute the Jacobian-vector product as required.

Chapter 9 Regularization

9.1 We will start by showing that the group of rotations by multiples of 90° forms a group:

- **Closure:** Suppose $\mathcal{A}$ represents a rotation of a° and $\mathcal{B}$ represents a rotation of b°, and both a and b are multiples of 90, making $\mathcal{A}$ and $\mathcal{B}$ both members of the set. $\mathcal{A} \circ \mathcal{B}$ therefore represents a rotation by $a + b$ which is also a multiple of 90, making this new rotation also a member of the set.
- **Associativity:** Now suppose we have three rotations, $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$, which represent rotations of a, b and c degrees respectively, again each a multiple of 90. Now we can see that both $(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}$ and $\mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})$ correspond to rotations of $(a + b + c)^\circ$, making the set associative under composition of rotations.
- **Identity:** A rotation of 0° is in the set and also leaves other rotations unchanged when composed with them.
- **Inverse:** If we take an element $\mathcal{A}$ in the set which again represents a rotation of a°, then it's inverse $\mathcal{A}^{-1}$ will be a rotation by $360 - a$, meaning $\mathcal{A} \circ \mathcal{A}^{-1}$ will give a rotation of 360° which is the same as a rotation by 0° which is the identity.

By showing that these four axioms are satisfied, we have shown that this is indeed a group. Now we will do the same for the group of translations of an object in a two-dimensional plane.

- **Closure:** Suppose $\mathcal{A}$ represents a translation of a_x in x and a_y in y, and $\mathcal{B}$ represents a translation of b_x in x and b_y in y. $\mathcal{A} \circ \mathcal{B}$ therefore represents a translation by $a_x + b_x$ in x and $a_y + b_y$ in y which is also a translation in a 2D plane.
- **Associativity:** Now suppose we have three translations, $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$, which represent translations of a_x, b_x and c_x in x respectively, and a_y, b_y and c_y in y. Now we can see that both $(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}$ and $\mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})$ correspond to translations of $a_x + b_x + c_x$ in x and $a_y + b_y + c_y$ in y.
- **Identity:** The composition of a translation of 0 in both dimensions with any other translation will leave the latter unchanged and therefore the translation of 0 in both dimensions is the identity for this group.
- **Inverse:** If we take an element $\mathcal{A}$ in the set which again represents a translation of a_x in x and a_y in y, we can see that composing this with a translation of $-a_x$ in x and $-a_y$ in y gives us a translation of 0 in both dimensions which means this negative translation is the inverse of $\mathcal{A}$

9.2 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 (9.52). From (9.53) we then define

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

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

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

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

From this and (9.53) 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.

9.3 We first write the gradient descent formula in terms of continuous time t in the form

$$\mathbf{w}(t+\epsilon) = \mathbf{w}(t) - \epsilon\widetilde{\eta}\nabla\Omega(\mathbf{w}) \tag{354}$$

where ϵ represents some finite time step and we have defined $\widetilde{\eta} = \eta/\epsilon$. We now make a Taylor expansion of the left-hand side in powers of ϵ to give

$$\mathbf{w}(t) + \frac{\mathrm{d}\mathbf{w}(t)}{\mathrm{d}t}\epsilon + \mathcal{O}(\epsilon^2) = \mathbf{w}(t) - \epsilon\widetilde{\eta}\nabla\Omega(\mathbf{w}). \tag{355}$$

We now take the limit $\epsilon \to 0$ to give

$$\frac{\mathrm{d}\mathbf{w}(t)}{\mathrm{d}t} = -\widetilde{\eta}\nabla\Omega(\mathbf{w}). \tag{356}$$

Next we substitute for $\Omega(\mathbf{w})$ using

$$\Omega(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{357}$$

to give

$$\frac{\mathrm{d}\mathbf{w}(t)}{\mathrm{d}t} = -\widetilde{\eta}\mathbf{w}. \tag{358}$$

This has the solution

$$\mathbf{w}(t) = \mathbf{w}(0)\exp\{-\widetilde{\eta}t\} \tag{359}$$

as can easily be verified by substitution. Thus, the elements of $\mathbf{w}$ decay exponentially to zero.

9.4 With the transformed inputs, weights and biases, (9.6) becomes

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

Using (9.8)–(9.10), we can rewrite the argument of $h(\cdot)$ on the r.h.s. as

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

Similarly, with the transformed outputs, weights and biases, (9.7) becomes

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

Using (9.11)–(9.13), we can rewrite this as

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

By subtracting d and subsequently dividing by c on both sides, we recover (9.7) in its original form.

9.5 We can rewrite the constraint (9.20) in the form

$$\sum_{j=1}^{M} |w_j|^q - \eta \leqslant 0. \tag{360}$$

This constraint can be enforced by adding a term to the un-regularized error $E(\mathbf{w})$ using a Lagrange multiplier, which we denote $\lambda/2$, to give

$$E(\mathbf{w}) + \frac{\lambda}{2}\left(\sum_{j=1}^{M} |w_j|^q - \eta\right). \tag{361}$$

Since the term $\lambda\eta/2$ is constant with respect to $\mathbf{w}$, minimizing (361) is equivalent to minimizing

$$E(\mathbf{w}) + \frac{\lambda}{2}\sum_{j=1}^{M} |w_j|^q. \tag{362}$$

Appendix C If the constraint is active then $\lambda \neq 0$ and hence

$$\sum_{j=1}^{M} |w_j|^q = \eta. \tag{363}$$

The strength of the regularization increases as λ increases, driving weights to smaller values. Similarly, the strength of the constraint increases as η decreases, again driving weights to smaller values. However, the precise relationship between λ and η depends on the form of $E(\mathbf{w})$.

9.6 The gradient of (9.56) is given

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

and hence update formula (9.57) becomes

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

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

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau)} && (364)\\ &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_j^{\mathrm{T}}\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j\mathbf{u}_j^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j(w_j^{(\tau-1)} - w_j^\star), && (365) \end{aligned}$$

where we have used (9.59). To show that

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

for $\tau = 1, 2, \ldots$, we can use proof by induction. For $\tau = 1$, we recall that $\mathbf{w}^{(0)} = \mathbf{0}$ and insert this into (365), giving

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

Now we assume that the result holds for $\tau = N - 1$ and then make use of (365)

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

as required. Provided that $|1-\rho\eta_j| < 1$ then we have $(1-\rho\eta_j)^\tau \to 0$ as $\tau \to \infty$, and hence $\left\{1 - (1 - \rho\eta_j)^N\right\} \to 1$ and $\mathbf{w}^{(\tau)} \to \mathbf{w}^\star$. If τ is finite but $\eta_j \gg (\rho\tau)^{-1}$, τ must still be large, since $\eta_j\rho\tau \gg 1$, even though $|1-\rho\eta_j| < 1$. If τ is large, it follows from the argument above that $w_j^{(\tau)} \simeq w_j^\star$. If, on the other hand, $\eta_j \ll (\rho\tau)^{-1}$, this means that $\rho\eta_j$ must be small, since $\rho\eta_j\tau \ll 1$ and τ is an integer greater than or equal to one. If we expand,

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

and insert this into (9.58), we get

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

9.7 Suppose that a set of weights $w_1, \ldots, w_K$ are shared so that $w_1 = w_2 = \ldots = w_K = \lambda$. We can compute the derivative of the error function with respect to λ using the chain rule of calculus

$$\frac{\partial E}{\partial \lambda} = \sum_{i=1}^{K} \frac{\partial E}{\partial w_i}\frac{\partial w_i}{\partial \lambda} = \sum_{i=1}^{K} \frac{\partial E}{\partial w_i} \tag{366}$$

where we have used

$$\frac{\partial w_i}{\partial \lambda} = 1. \tag{367}$$

Hence, first run the standard backpropagation algorithm (or automatic differentiation) to evaluate the individual gradients $\partial E/\partial w_i$ for all weights. Then, for every

group of weights in a network that are shared, sum up the gradients over all of those weights and the use this combined gradient to update the weights in that group. Note that, as long as the weights in each group are initialized to the same value, this will ensure that they remain equal after the update.

9.8 From the formula

$$p(w) = \sum_{j=1}^{M} \pi_j \mathcal{N}(w|\mu_j, \sigma_j^2) \tag{368}$$

we can identify the following probabilities

$$p(j) = \pi_j \tag{369}$$

$$p(w|j) = \mathcal{N}(w|\mu_j, \sigma_j^2). \tag{370}$$

Hence from Bayes' theorem we have

$$p(j|w) = \frac{p(j)p(w|j)}{p(w)} = \frac{\pi_j \mathcal{N}(w|\mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w|\mu_k, \sigma_k^2)}. \tag{371}$$

9.9 This is easily verified by taking the derivative of (9.22), using (2.49) and standard derivatives, yielding

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

Combining this with (9.23) and (9.24), we immediately obtain the second term of (9.25).

9.10 Since the μ_js only appear in the regularization term, $\Omega(\mathbf{w})$, from (9.23) we have

$$\frac{\partial \widetilde{E}}{\partial \mu_j} = \lambda \frac{\partial \Omega}{\partial \mu_j}. \tag{372}$$

Using (3.25), (9.22) and (9.24) and standard rules for differentiation, we can calculate the derivative of $\Omega(\mathbf{w})$ as follows:

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

Combining this with (372), we get (9.26).

9.11 Following the same line of argument as in Solution 9.10, we need the derivative of $\Omega(\mathbf{w})$ w.r.t. σ_j. Again using (3.25), (9.22) and (9.24) and standard rules for differentiation, we find this to be

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

Combining this with (372), we get (9.28).

9.12 From the definition (9.30) we have

$$\pi_k = \frac{\exp(\eta_k)}{\sum_l \exp(\eta_l)}. \tag{373}$$

Taking the derivative then gives

$$\begin{aligned}\frac{\partial\pi_k}{\partial\eta_j} &= \frac{\exp(\eta_k)}{\sum_l\exp(\eta_l)}\delta_{jk} - \frac{\exp(\eta_k)}{\left(\sum_l\exp(\eta_l)\right)^2}\exp(\eta_j)\\ &= \delta_{jk}\pi_k - \pi_j\pi_k.\end{aligned} \tag{374}$$

From (9.22) and (**??**) we then have

$$\begin{aligned}\frac{\partial\widetilde{E}}{\partial\eta_j} &= -\lambda\sum_i\sum_k \gamma_k(w_i)\frac{1}{\pi_k}\frac{\partial\pi_k}{\partial\eta_j}\\ &= -\lambda\sum_i\sum_k \gamma_k(w_i)\left\{\delta_{jk}-\pi_j\right\}\\ &= \lambda\sum_i\left\{\pi_j-\gamma_j(w_i)\right\}\end{aligned} \tag{375}$$

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

9.13 The result is easily proved by substituting (9.36) into (9.37), and then substituting (9.35) into the resulting expression, giving

$$\begin{aligned}\mathbf{y} =& \mathbf{F}_3(\mathbf{z}_2) + \mathbf{z}_2\\ =& \mathbf{F}_3(\mathbf{F}_2(\mathbf{z}_1)+\mathbf{z}_1) + \mathbf{F}_2(\mathbf{z}_1) + \mathbf{z}_1\\ =& \mathbf{F}_3(\mathbf{F}_2(\mathbf{F}_1(\mathbf{x})+\mathbf{x}) + \mathbf{F}_1(\mathbf{x}) + \mathbf{x})\\ &+ \mathbf{F}_2(\mathbf{F}_1(\mathbf{x})+\mathbf{x}))\\ &+ \mathbf{F}_1(\mathbf{x}) + \mathbf{x}.\end{aligned} \tag{376}$$

9.14 Using (9.49), we can rewrite (9.47) as

$$\begin{aligned} E_{\mathrm{COM}} &= \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right] \\ &= \frac{1}{M^2}\mathbb{E}_{\mathbf{x}}\left[\left\{\sum_{m=1}^{M}\epsilon_m(\mathbf{x})\right\}^2\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}\left[\epsilon_m(\mathbf{x})^2\right] = \frac{1}{M}E_{\mathrm{AV}} \end{aligned}$$

where we have used (9.46) in the last step.

9.15 We start by rearranging the r.h.s. of (9.46), by moving the factor $1/M$ inside the sum and the expectation operator outside the sum, yielding

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2\right].$$

If we then identify $\epsilon_m(\mathbf{x})$ and $1/M$ with x_i and λ_i in (2.102), respectively, and take $f(x) = x^2$, we see from (2.102) that

$$\left(\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})\right)^2 \leqslant \sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2.$$

Since this holds for all values of $\mathbf{x}$, it must also hold for the expectation over $\mathbf{x}$, proving (9.64).

9.16 If $E(y(\mathbf{x}))$ is convex, we can apply (2.102) as follows:

$$\begin{aligned} E_{\mathrm{AV}} &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathbf{x}}\left[E(y(\mathbf{x}))\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}E(y(\mathbf{x}))\right] \\ &\geqslant \mathbb{E}_{\mathbf{x}}\left[E\left(\sum_{m=1}^{M}\frac{1}{M}y(\mathbf{x})\right)\right] \\ &= E_{\mathrm{COM}} \end{aligned}$$

where $\lambda_i = 1/M$ for $i = 1, \ldots, M$ in (2.102) and we have implicitly defined versions of E_{AV} and E_{COM} corresponding to $E(y(\mathbf{x}))$.

9.17 To prove that (9.67) is a sufficient condition for (9.66) we have to show that (9.66) follows from (9.67). To do this, consider a fixed set of $y_m(\mathbf{x})$ and imagine varying the α_m over all possible values allowed by (9.67) and consider the values taken by $y_{\text{COM}}(\mathbf{x})$ as a result. The maximum value of $y_{\text{COM}}(\mathbf{x})$ occurs when $\alpha_k = 1$ where $y_k(\mathbf{x}) \geqslant y_m(\mathbf{x})$ for $m \neq k$, and hence all $\alpha_m = 0$ for $m \neq k$. An analogous result holds for the minimum value. For other settings of $\boldsymbol{\alpha}$,

$$y_{\min}(\mathbf{x}) < y_{\text{COM}}(\mathbf{x}) < y_{\max}(\mathbf{x}),$$

since $y_{\text{COM}}(\mathbf{x})$ is a convex combination of points, $y_m(\mathbf{x})$, such that

$$\forall m : y_{\min}(\mathbf{x}) \leqslant y_m(\mathbf{x}) \leqslant y_{\max}(\mathbf{x}).$$

Thus, (9.67) is a sufficient condition for (9.66).

Showing that (9.67) is a necessary condition for (9.66) is equivalent to showing that (9.66) is a sufficient condition for (9.67). The implication here is that if (9.66) holds for any choice of values of the committee members $\{y_m(\mathbf{x})\}$ then (9.67) will be satisfied. Suppose, without loss of generality, that α_k is the smallest of the α values, i.e. $\alpha_k \leqslant \alpha_m$ for $k \neq m$. Then consider $y_k(\mathbf{x}) = 1$, together with $y_m(\mathbf{x}) = 0$ for all $m \neq k$. Then $y_{\min}(\mathbf{x}) = 0$ while $y_{\text{COM}}(\mathbf{x}) = \alpha_k$ and hence from (9.66) we obtain $\alpha_k \geqslant 0$. Since α_k is the smallest of the α values it follows that all of the coefficients must satisfy $\alpha_k \geqslant 0$. Similarly, consider the case in which $y_m(\mathbf{x}) = 1$ for all m. Then $y_{\min}(\mathbf{x}) = y_{\max}(\mathbf{x}) = 1$, while $y_{\text{COM}}(\mathbf{x}) = \sum_m \alpha_m$. From (9.66) it then follows that $\sum_m \alpha_m = 1$, as required.

9.18 From (3.2) the Bernoulli distribution for the elements of the dropout matrix can be written as

$$\text{Bern}(R_{ni}|\rho) = \rho^{R_{ni}}(1-\rho)^{1-R_{ni}}. \tag{377}$$

Hence we have

$$\begin{aligned}\mathbb{E}[R_{ni}] &= \sum_{R_{ni}\in\{0,1\}} \text{Bern}(R_{ni}|\rho) R_{ni} \\ &= \rho. \end{aligned} \tag{378}$$

Two elements R_{ni} and R_{nj} will be independent unless $j = i$. Hence, for $j \neq i$ we have

$$\begin{aligned}\mathbb{E}[R_{ni}R_{nj}] &= \sum_{R_{ni}\in\{0,1\}} \sum_{R_{nj}\in\{0,1\}} \text{Bern}(R_{ni}|\rho)\text{Bern}(R_{nj}|\rho) R_{ni}R_{nj} \\ &= \left(\sum_{R_{ni}\in\{0,1\}} \text{Bern}(R_{ni}|\rho) R_{ni}\right)\left(\sum_{R_{nj}\in\{0,1\}} \text{Bern}(R_{nj}|\rho) R_{nj}\right) \\ &= \rho^2\end{aligned}$$

whereas if $j = i$ we have

$$R_{ni}R_{nj} = R_{ni}^2 = R_{ni}$$

and therefore

$$\mathbb{E}[R_{ni}R_{nj}] = \mathbb{E}[R_{ni}] = \rho.$$

Combining these we obtain

$$\mathbb{E}[R_{ni}R_{nj}] = \delta_{ij}\rho + (1 - \delta_{ij})\rho^2. \tag{379}$$

To find the expected value of the error function (9.69) we first expand out the square to give

$$E(\mathbf{W}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk}^2 - 2\sum_{i=1}^{D} w_{ki}R_{ni}x_{ni} + \left(\sum_{i=1}^{D} w_{ki}R_{ni}x_{ni}\right)\left(\sum_{j=1}^{D} w_{kj}R_{nj}x_{nj}\right)\right\}.$$

Next we take the expectation of the error and substitute for the expectations of the dropout matrix elements using (378) and (379) to give

$$\begin{aligned}\mathbb{E}\left[E(\mathbf{W})\right] &= \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk}^2 - 2\rho\sum_{i=1}^{D} w_{ki}x_{ni} + \rho^2\left(\sum_{i=1}^{D} w_{ki}x_{ni}\right)^2\right\} \\ &\quad + (\rho - \rho^2)\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D} w_{ki}^2 x_{ni}^2 \qquad (380) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{t_{nk} - \rho\sum_{i=1}^{D} w_{ki}x_{ni}\right\}^2 \\ &\quad + \rho(1 - \rho)\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D} w_{ki}^2 x_{ni}^2. \qquad (381)\end{aligned}$$

Finally, we can find a solution for the weights that minimize this expected error function by setting the derivatives with respect to w_{ki} equal to zero. For this it is

more convenient to work with the expression (380) giving

$$\begin{aligned}\frac{\partial}{\partial w_{ki}}\mathbb{E}\left[E(\mathbf{W})\right] =& -2\rho\sum_{n=1}^{N}x_{ni} + 2\rho^2\sum_{j=1}^{D}w_{kj}\left(\sum_{n=1}^{N}x_{nj}x_{ni}\right) \\ &+ 2\rho(1-\rho)w_{ki}\left(\sum_{n=1}^{N}x_{ni}^2\right) \\ =& -2\rho\sum_{n=1}^{N}x_{ni} \\ &+ 2\sum_{j=1}^{D}w_{kj}\left\{\rho^2\left(\sum_{n=1}^{N}x_{nj}x_{ni}\right) + \delta_{ji}\rho(1-\rho)\left(\sum_{n=1}^{N}x_{ni}^2\right)\right\}.\end{aligned}$$

This can be written in matrix form as

$$0 = -\mathbf{B} + \mathbf{W}\mathbf{M} \tag{382}$$

where $\mathbf{W}$ has elements w_{ij}, $\mathbf{B}$ is a diagonal matrix, and the elements of $\mathbf{B}$ and $\mathbf{M}$ are given by

$$B_{ii} = -2\rho\sum_{n=1}^{N}x_{ni} \tag{383}$$

$$M_{ji} = 2\left\{\rho^2\left(\sum_{n=1}^{N}x_{nj}x_{ni}\right) + \delta_{ji}\rho(1-\rho)\left(\sum_{n=1}^{N}x_{ni}^2\right)\right\}. \tag{384}$$

Hence the minimizing weights are given by

$$\mathbf{W}^\star = \mathbf{B}\mathbf{M}^{-1}. \tag{385}$$

Chapter 10 Convolutional Networks

10.1 We can impose the constraint $\|\mathbf{x}\|^2 = K$ by using a Lagrange multiplier λ and maximizing

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + \lambda\left(\|\mathbf{x}\|^2 - K\right). \tag{386}$$

Taking the gradient with respect to $\mathbf{x}$ and setting this gradient to zero gives

$$\mathbf{w} + 2\lambda\mathbf{x} = 0 \tag{387}$$

which shows that $\mathbf{x} = \alpha\mathbf{w}$ where $\alpha = -1/(2\lambda)$.

10.2 Let us represent the input array as a vector $\mathbf{x} = (x_1, x_2, .., x_5)^{\mathrm{T}}$. We will start with the case where there is only one convolutional filter, of width 3, the weights of which we will denote using a vector $\mathbf{k} = (k_1, k_2, k_3)^{\mathrm{T}}$. If we look at Figure 3, we can see that the three outputs, which we represent with the vector $\mathbf{y} = (y_1, y_2, y_3)^{\mathrm{T}}$ are given by:

$$\mathbf{y} = \begin{bmatrix} x_1k_1 + x_2k_2 + x_3k_3 \\ x_2k_1 + x_3k_2 + x_4k_3 \\ x_3k_1 + x_4k_2 + x_5k_3 \end{bmatrix}. \tag{388}$$

Now we wish to find a matrix $\mathbf{K}$ such that $\mathbf{y} = \mathbf{K}\mathbf{x}$. We can see that it must be a 3×5 matrix, and each entry K_{ij} is given by the contribution that x_j makes to y_i. Therefore K is given by:

$$K = \begin{bmatrix} k_1 & k_2 & k_3 & 0 & 0 \\ 0 & k_1 & k_2 & k_3 & 0 \\ 0 & 0 & k_1 & k_2 & k_3 \end{bmatrix}. \tag{389}$$

This is an example of a Toeplitz matrix, where each descending diagonal from left to right is constant. Convolution operations in 1D can always be represented as a multiplication of the input array by a Toeplitz matrix.

10.3 Simple matrix multiplication shows that the convolution is given by

32	14	−18
−22	−24	12
22	6	18

10.4 There are many possibilities for indexing the elements of I, K, and C. One choice is simply to choose the indices in $K(l, m)$ to have the ranges $1 \leqslant l \leqslant L$ and $1 \leqslant m \leqslant M$ giving

$$C(j, k) = \sum_{l=1}^{L}\sum_{m=1}^{M} I(j + l, k + m)K(l, m). \tag{390}$$

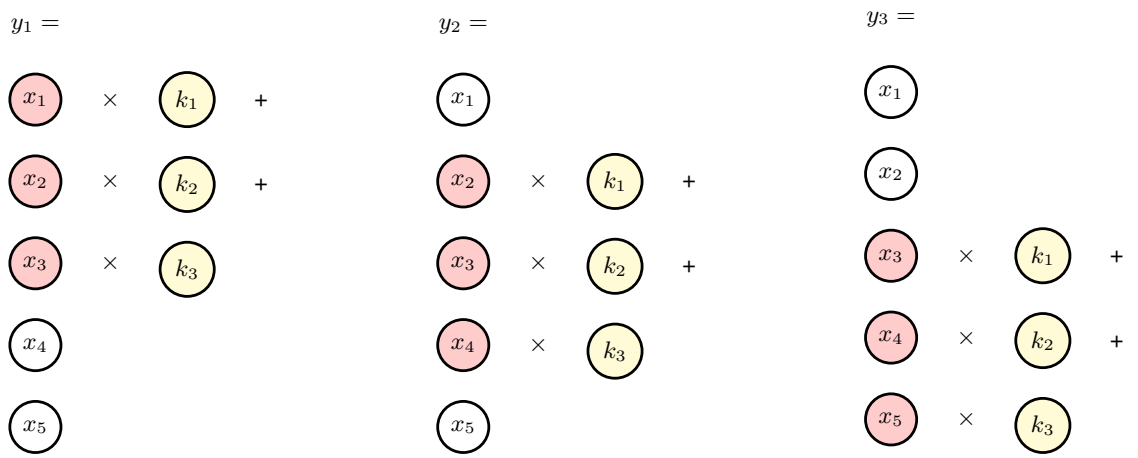

Figure 3 Figure showing a convolution operation of a filter $\mathbf{k}$ over an input array $\mathbf{x}$ with output $\mathbf{y}$.

If we similarly choose the two indices of $I(\cdot, \cdot)$ to run from $1, \ldots, J$ and $1, \ldots, K$ respectively, it follows that $0 \leqslant j \leqslant J - L$ and $0 \leqslant k \leqslant K - M$. For the convolutional form we likewise have

$$C(j,k) = \sum_{l=1}^{L} \sum_{m=1}^{M} I(j-l, k-m) K(l,m) \tag{391}$$

where now $L+1 \leqslant j \leqslant J+1$ and $M+1 \leqslant k \leqslant K+1$ as is easily verified. Finally, we can define $\lambda = L - l + 1$ and $\mu = M - m + 1$ which allows us to rewrite (391) in the form

$$C(j,k) = \sum_{\lambda=1}^{L} \sum_{\mu=1}^{M} \widetilde{I}(j+\lambda, k+\mu) \widetilde{K}(\lambda, \mu) \tag{392}$$

where we have defined

$$\widetilde{I}(j+\lambda, k+\mu) = I(j-L+\lambda-1, k-M+\mu-1) \tag{393}$$

$$\widetilde{K}(\lambda, \mu) = K(L-\lambda+1, M-\mu+1). \tag{394}$$

The convolution and cross-correlation representations differ in whether the index variables l and m that label the kernel elements run from low to high values or or vice versa. In a machine learning application it is usually irrelevant which of these forms is used, since the algorithm will learn the same value for the kernel in the corresponding locations, so that learning with the inverted representation will lead to the same kernel but with its values in reverse order.

10.5 If we substitute $z = x - y$ into (10.21) we obtain

$$F(x) = \int_{-\infty}^{\infty} G(x-z) k(z) \, \mathrm{d}z. \tag{395}$$

If we now discretize the x and z variables into bins of width Δ we can approximate this integral using

$$F(j\Delta) \simeq \sum_{l=-\infty}^{\infty} G(j\Delta - l\Delta)k(l\Delta)\Delta. \tag{396}$$

This can now be written as a one-dimensional version of the convolutional layer defined by (10.19) in the form

$$C(j) \simeq \sum_{l=1}^{L} I(j-l)K(l) \tag{397}$$

where we have defined $C(j) = F(j\Delta)$, $I(l) = G(l\Delta)$, and $K(l) = k(l\Delta)\Delta$.

10.6 We saw in Section 10.2.3 that convolving an image of dimensions $J \times K$ and additional padding P, with a filter of dimensions $M \times M$ will yield a feature map of dimension $(J + 2P - M + 1) \times (K + 2P - M + 1)$. Now if we substitute $P = (M-1)/2$, we see that the dimensions of the feature map are given by $(J+(2((M-1)/2)-M+1) \times (K+(2((M-1)/2)-M+1)$ which simplifies to $J \times K$.

10.7 In the case with no padding and a stride of 1, an $M \times M$ kernel would convolve $J - M + 1$ times horizontally and $K - M + 1$ times vertically, giving us $(J - M + 1) \times (K - M + 1)$ features. After applying padding P to each of the edges of the image, we have a new image of size $(J + 2P) \times (K + 2P)$, and hence the dimensionality of the feature layer would be $(J+2P-M+1) \times (K+2P-M+1)$. When we apply a stride, we divide the number of convolutions in each dimension by the stride and use the floor operator to account for the case where there is some remainder of the image in a given direction that is less than the stride. The initial 1 is not divided by the stride as it represents the first operation and is therefore unaffected by the stride. This gives us

$$\left\lfloor \frac{J + 2P - M}{S} + 1 \right\rfloor \times \left\lfloor \frac{K + 2P - M}{S} + 1 \right\rfloor \tag{398}$$

features.

10.8 We assume connections between padding inputs and features count as connections for simplicity. We also haven't included max pooling or activation function connections. As every convolutional layer in the VGG-16 network uses a 3×3 filter, a given node in a convolutional layer takes a number of inputs equal to 9 times the number of channels in the previous layer plus 1 for the bias. The first convolutional layer therefore has $224 \times 224 \times (3 \times 9 + 1) \times 64 = 96,337,920$ connections to the previous layer. The number of connections for a fully connected layer is equal to the number of input features plus 1 for the bias, multiplied by the number of nodes, which for the first fully connected layer is equal to $(7 \times 7 \times 512 + 1) \times 4096 = 102,764,544$. The number of the connections for the rest of the layers are shown in Table 1.

The number of learnable parameters in a convolutional layer is independent of the height and width dimensions of that layer. For a layer with a 3×3 filter, the number of parameters for a given kernel is equal to 9 multiplied by the number of channels in the previous layer, plus 1 for the bias. For a given layer, the number of such kernels is then the same as the number of channels. So for example the first convolutional layer has $64 \times (3 \times 3 \times 3 + 1) = 1,792$ learnable parameters. For a fully connected layer, the number of parameters is just equal to the number of connections as there are no shared weights. The number of learnable parameters for each layer are also shown in Table 1.

Layer	Connections	Learnable Parameters
Convolution 1	$96,337,920$	$1,792$
Convolution 2	$1,852,899,328$	$36,928$
Convolution 3	$926,449,664$	$73,856$
Convolution 4	$1,852,899,328$	$147,584$
Convolution 5	$926,449,664$	$295,168$
Convolution 6	$1,852,899,328$	$590,080$
Convolution 7	$1,852,899,328$	$590,080$
Convolution 8	$926,449,664$	$1,180,160$
Convolution 9	$1,852,899,328$	$2,359,808$
Convolution 10	$1,852,899,328$	$2,359,808$
Convolution 11	$462,522,368$	$2,359,808$
Convolution 12	$462,522,368$	$2,359,808$
Convolution 13	$462,522,368$	$2,359,808$
Fully Connected 1	$102,764,544$	$102,764,544$
Fully Connected 2	$16,781,312$	$16,781,312$
Fully Connected 3	$4,097,000$	$4,097,000$
Total	$15,504,292,840$	$138,357,544$

Table 1 Table showing the number of connections and learnable parameters in each layer of the VGG-16 network.

10.9 The convolution operation can be written as

$$C(j,k) = \sum_{l=1}^{L} \sum_{m=1}^{M} I(j-l, k-m) K(l,m) \tag{399}$$

where $j = 1, \ldots, J$ and $k = 1, \ldots, K$. The kernel K is swept across the image I giving a total number of positions of $(J - L + 1) \times (K - M + 1)$ and for each position, the number of operations is $L \times M$. Thus the total number of operations is

$$(J - L + 1)(K - M + 1)LM. \tag{400}$$

Now suppose that the kernel is separable, in other words that it factorizes in the form

$$K(l,m) = F(l)G(m). \tag{401}$$

Substituting (401) into (399) we obtain

$$C(j,k) = \sum_{l=1}^{L} F(l) \sum_{m=1}^{M} I(j-l, k-m)G(m). \tag{402}$$

Consider first the summation over m. This involves a one-dimensional kernel $G(m)$ which must be swept over the image for a total number of $J \times (K-M+1)$ positions, and in each position there are M operations to perform giving a total number of operations equal to $(K-M+1)JM$. This gives rise to an intermediate array of dimension $J \times (K-M+1)$. Now the summation over l is performed which is also a convolution involving a one-dimensional kernel $F(l)$. The number of positions for this kernel is given by $(K-M+1) \times (J-L+1)$ and in each position we have to perform L operations giving a total of $(K-M+1)(J-L+1)L$. Overall, the total number of operations is therefore given by

$$(K-M+1)JM + (K-M+1)(J-L+1)L. \tag{403}$$

To see that this represents a saving in computation consider the case where the image is large compared to the kernel size so that $J \gg L$ and $K \gg M$. Then (400) is approximately given by $JKLM$ whereas (403) is approximately given by $JK(L+M)$. Note that as well as saving on compute, a separable kernel uses less storage. However, since it restricts the form of the kernel it can lead to a significant reduction in generalization accuracy.

10.10 The derivatives of a cost function with respect to an activation value can be evaluated using backpropagation, which corresponds to an application of the chain rule of calculus. This backpropagation starts with the derivatives of the cost function with respect to the local activations, which for the cost function defined by (10.12) are given by

$$\delta_{ijk} = \frac{\partial F(\mathbf{I})}{\partial a_{ijk}} = 2a_{ijk} \tag{404}$$

and hence, up to a factor of 2, are given by the activation values themselves. These values are then back-propagated through the network using (8.13) until the input layer is reached. This input layer represents the image and the associated δ values correspond to derivatives of the cost function (10.12) with respect to the pixel values, as required.

10.11 Consider a 1-hot encoding scheme for the C object classes using binary variables $y_i \in \{0,1\}$ where $i = 1, \ldots, C$ and $y_i = 1$ represents the presence of an object from class i. We then introduce an additional class with a binary variable $y_{C+1} \in \{0,1\}$ where $y_{C+1} = 1$ means there is no object from any of the given classes in the image. Since we assume that a given image either contains an object from one of the classes or no objects, the variables $y_1, \ldots, y_{C+1}$ form a 1-hot encoding where all variables have the value 0 except for a single variable taking the value 1. We can train a model $f(\cdot)$ to take an image as input and to return a probability distribution over these

variables. These probabilities must sum to one

$$\sum_{i=1}^{C+1} p_f(y_i = 1) = 1. \tag{405}$$

Now instead suppose we introduce a binary variable $b \in \{0, 1\}$ such that $b = 1$ means an object (of any class) is present in the image and $b = 0$ means that no object is present. We can train a model $g(\cdot)$ to output a probability distribution over this variable, such that

$$p_g(b = 1) + p_g(b = 0) = 1. \tag{406}$$

We also introduce binary variables $z_1, \ldots, z_C \in \{0, 1\}$ to predict the class of the object, conditional on their being an object present. These variables have a 1-hot encoding. We then train an associated model $h(\cdot)$ to output a probability distribution satisfying

$$\sum_{i=1}^{C} p_h(z_i = 1) = 1. \tag{407}$$

To relate these sets of probabilities we can use the product rule in the form

$$p(\text{object present and class } i) = p(\text{class } i|\text{object present})p(\text{object present}) \tag{408}$$

which gives the following results

$$p_f(y_i = 1) = p_h(z_i = 1)p_g(b = 1) \quad i = 1, \ldots, C \tag{409}$$

$$p_f(y_{C+1} = 1) = p_g(b = 0). \tag{410}$$

10.12 We first note that evaluating the scalar product between two N-dimensional vectors requires N multiplies and $N - 1$ additions giving a total of $2N - 1$ computational steps.

For the network in Figure 10.22, the number of computational steps required to calculate the first convolution operation is given by the number of features in the second layer, which is $4 \times 4 = 16$, multiplied by the number of steps needed to evaluate the output of the filter. Since the filter size is $3 \times 3 = 9$ each filter evaluation requires $9 \times 8 = 72$ steps. Hence the total number of computational steps for the first convolutional layer is $16 \times 72 = 1,152$. For one evaluation of a 2×2 max pooling filter, there are 3 computational steps required. This is multiplied by the number of such operations required for the max pooling layer which is 4, giving 12 total steps. The fully connected layer is equivalent to a scalar product between two vectors of dimensionality 4 and hence requires $4 + 3 = 7$ operations. Therefore a single evaluation of this network requires a total of $1,152 + 12 + 7 = 1,171$ computational steps.

For the network in Figure 10.23, there are $6 \times 6 = 36$ features in the second layer each of which requires $9 \times 8 = 72$ computations in the convolutional layer, giving a total of $36 \times 72 = 2,592$ computational steps. The max pooling layer has 9 nodes and hence requires $9 \times 3 = 27$ computations. Finally, the fully connected layer

requires $4 \times 7 = 28$ computations, giving a total of $2,592 + 27 + 28 = 2,647$ operations for one pass through the whole network.

Therefore the improvement in efficiency of using one pass through the second network compared to 8 passes through the first network is equal to $1,171 \times 8/2,647 = 3.54$.

10.13 The padded input vector is given by

$$\mathbf{x} = \begin{pmatrix} 0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ 0 \end{pmatrix}. \tag{411}$$

For a filter with elements (w_1, w_2, w_3) and a stride of 2, the output vector will be two-dimensional and can be written as $\mathbf{y} = (y_1, y_2)^{\mathrm{T}}$, in which the elements are given by

$$y_1 = w_2 x_1 + w_3 x_2 \tag{412}$$
$$y_2 = w_1 x_2 + w_2 x_3 + w_3 x_4. \tag{413}$$

We can write this convolution operation using matrix notation in the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{414}$$

where the matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} w_1 & w_2 & w_3 & 0 & 0 & 0 \\ 0 & 0 & w_1 & w_2 & w_3 & 0 \end{pmatrix}. \tag{415}$$

Now consider the up-sampling operation using a filter (w_1, w_2, w_3) operating on a vector $\mathbf{z} = (z_1, z_2)^{\mathrm{T}}$ with a stride of 2. The resulting six-dimensional vector $\mathbf{h} = (h_1, h_2, h_3, h_4, h_5, h_6)$ has elements given by

$$h_1 = w_1 z_1 \tag{416}$$
$$h_2 = w_2 z_1 \tag{417}$$
$$h_3 = w_3 z_1 + w_1 z_2 \tag{418}$$
$$h_4 = w_2 z_2 \tag{419}$$
$$h_5 = w_3 z_2 \tag{420}$$
$$h_6 = 0. \tag{421}$$
$$\tag{422}$$

This can be written using matrix notation in the form

$$\mathbf{h} = \mathbf{B}\mathbf{z} \tag{423}$$

where the matrix $\mathbf{B}$ is given by

$$\mathbf{B} = \begin{pmatrix} w_1 & 0 \\ w_2 & 0 \\ w_3 & w_1 \\ 0 & w_2 \\ 0 & w_3 \\ 0 & 0 \end{pmatrix}. \tag{424}$$

By inspection we see that $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$ and hence up-sampling can be seen as the transpose of convolution.