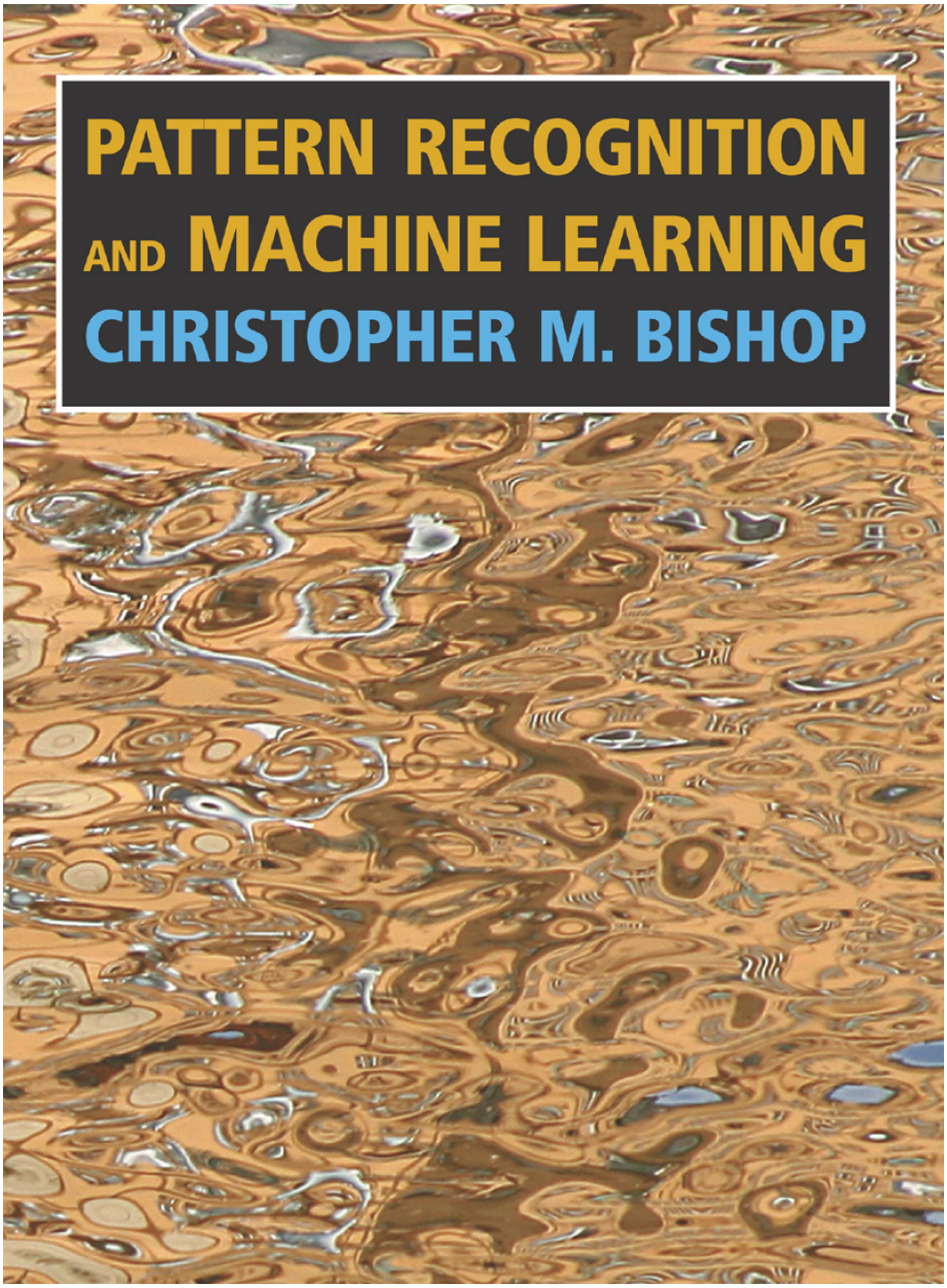
Solutions for



(8TH PRINTING, 2009)

Before You Read

This Book of Solutions is Currently Unfinished

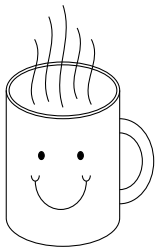
As part of an individual endeavour to do some self-studying, as well as build some personal material, I have begun to develop this document of solutions to Christopher Bishop's "*Pattern Recognition and Machine Learning*" (8^{th} printing, 2009). **These solutions are not official, they are my own, as are all associated mistakes and blemishes, as well as all mathematical and grammatical oversights and inconsistencies** - I am allowed some leeway on this last aspect, as I am not a native English speaker, nor good at math. The degree to which the solutions are detailed varies from question to question, purely by nature of my personal preference and/or disposition. I have likewise struggled to maintain consistency with how I prefer to present certain concepts versus how the author presents them, which may prove to make for unnecessarily cumbersome reading. Nevertheless, I highly recommend the book - for more information on it, follow this link to it's Springer storefront. The book is likewise freely available as a `.pdf` download on this link. I do not invite people to message me in case they find any errors or have any suggestions on fixing certain Exercise solutions, but I do welcome it.

Now, for some commentary on the organization of this particular document: all Exercises are presented on the same order as they were on the book, separated by chapter. Links which reference formulae contained within this document (or lead to an internet web page) are highlighted with the color blue, whilst links which reference formulae contained within the original book "*Pattern Recognition and Machine Learning*" are highlighted with the color red, and lead to a corresponding entry at the end of this document.

Now enjoy this quote for the spuds, and my complimentary "*Good Morning!*" mug:

> "Patterns multiplying
> Redirect our view
> Endless variations
> Make it all seem new
> Can you recognize the patterns that you find
> Stuck in your mind?"
>
> Devo (1982)

"



Bom dia!"

Chapter 1
Introduction

Exercise 1.1

Consider the error function in (1.2), with $\mathbf{y}(x, \mathbf{w})$ as in (1.1). We seek to determine the coefficients $\mathbf{w} = \{w_j\}_{j=0}^{M}$ which minimize (1.2): first, we differentiate (1.2) with respect to w_i, and obtain

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}w_i} = \sum_{n=1}^{N}(x_n)^i\{y(x_n, \mathbf{w}) - t_n\} \quad i \in \{0, \dots, M\}.$$

Solving for ${}^{\mathrm{d}E(\mathbf{w})}/_{\mathrm{d}w_i} = 0$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}w_i} &= 0 \\
\sum_{n=1}^{N}(x_n)^i\{y(x_n, \mathbf{w}) - t_n\} &= 0 \\
\sum_{n=1}^{N}(x_n)^i\left\{\sum_{j=0}^{M} w_j(x_n)^j - t_n\right\} &= 0 \\
\sum_{n=1}^{N}\left\{\sum_{j=0}^{M} w_j(x_n)^j(x_n)^i - t_n\right\} &= 0 && \text{(Apply formula (1.1))} \\
\sum_{n=1}^{N}\sum_{j=0}^{M}(x_n)^i w_j(x_n)^j - \sum_{n=1}^{N}(x_n)^i t_n &= 0 \\
\sum_{j=0}^{M} w_j \overbrace{\sum_{n=1}^{N}(x_n)^{i+j}}^{A_{i,j}} &= \overbrace{\sum_{n=1}^{N}(x_n)^i t_n}^{T_i} \\
\sum_{j=0}^{M} A_{i,j} w_j &= T_i \quad i \in \{0, \dots, M\}. && (1.1)
\end{aligned}$$

It is likewise possible to differentiate (1.2) again with respect to w_k, so that we obtain

$$\frac{\mathrm{d}^2E(\mathbf{w})}{\mathrm{d}w_i\mathrm{d}w_k} = \sum_{n=1}^{N}(x_n)^i(x_n)^k = \sum_{n=1}^{N}(x_n)^{i+k} \quad i, k \in \{0, \dots, M\}.$$

It follows that the solution to (1.1) minimizes (1.2).

Exercise 1.2

Consider the error function in (1.4), where $y(x, \mathbf{w})$ is as in (1.1) and $\lambda \geq 0$. We seek to determine the coefficients $\mathbf{w} = \{w_j\}_{j=0}^{M}$ which minimize (1.4): first, we differentiate (1.4) with respect to w_i, and obtain

$$\frac{\mathrm{d}\widetilde{E}(\mathbf{w})}{\mathrm{d}w_i} = \sum_{n=1}^{N} (x_n)^i \{y(x_n, \mathbf{w}) - t_n\} + \lambda w_i \quad i \in \{0, \dots, M\}.$$

Solving for ${}^{\mathrm{d}\widetilde{E}(\mathbf{w})}/_{\mathrm{d}w_i} = 0$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}\widetilde{E}(\mathbf{w})}{\mathrm{d}w_i} &= 0 \\
\sum_{n=1}^{N} (x_n)^i \{y(x_n, \mathbf{w}) - t_n\} + \lambda w_i &= 0 \\
\sum_{n=1}^{N} \left\{ \sum_{j=0}^{M} (x_n)^i w_j (x_n)^j - t_n \right\} + \lambda w_i &= 0 && \text{(Apply formula (1.1))} \\
\sum_{n=1}^{N} \sum_{j=0}^{M} (x_n)^i w_j (x_n)^j - \sum_{n=1}^{N} (x_n)^i t_n + \lambda w_i &= 0 \\
\sum_{j=0}^{M} w_j \sum_{n=1}^{N} (x_n)^{i+j} + \lambda w_i &= \sum_{n=1}^{N} (x_n)^i t_n \\
\sum_{\substack{j=0 \\ j \neq i}}^{M} w_j \overbrace{\sum_{n=1}^{N} (x_n)^{i+j}}^{A_{i,j}} + w_i \left(\lambda + \overbrace{\sum_{n=1}^{N} (x_n)^{i+i}}^{A_{i,i}} \right) &= \overbrace{\sum_{n=1}^{N} (x_n)^i t_n}^{T_i}
\end{aligned}$$

$$\sum_{\substack{j=0 \\ j \neq i}}^{M} A_{i,j} w_j + (\lambda + A_{i,i}) w_i = T_i \quad i \in \{0, \dots, M\}. \tag{1.2}$$

Differentiating (1.4) again with respect to w_k, we obtain

$$\begin{aligned}
\frac{\mathrm{d}^2 \widetilde{E}(\mathbf{w})}{\mathrm{d}w_i \mathrm{d}w_k} &= \begin{cases} \sum_{n=1}^{N} (x_n)^i (x_n)^k + \lambda & \text{if } i = k, \\ \sum_{n=1}^{N} (x_n)^i (x_n)^k & \text{otherwise.} \end{cases} \\
&= \begin{cases} \sum_{n=1}^{N} (x_n)^{i+k} + \lambda & \text{if } i = k, \\ \sum_{n=1}^{N} (x_n)^{i+k} & \text{otherwise.} \end{cases} \qquad i, k \in \{0, \dots, M\}.
\end{aligned}$$

It follows that the solution to (1.2) minimizes (1.4).

Exercise 1.3

Let B be a random variable which denotes which box is selected (assuming values $r =$ red, $b =$ blue and $g =$ green), and F be a random variable which denotes which fruit is selected (assuming values $a =$ apple, $o =$ orange and $l =$ lime). Let also

$$\begin{array}{llll} p(B=r)=\frac{1}{5} & p(F=a|B=r)=\frac{3}{10} & p(F=a|B=b)=\frac{1}{2} & p(F=a|B=g)=\frac{3}{10} \\ p(B=b)=\frac{1}{5} & p(F=o|B=r)=\frac{2}{5} & p(F=o|B=b)=\frac{1}{2} & p(F=o|B=g)=\frac{3}{10} \\ p(B=g)=\frac{3}{5} & p(F=l|B=r)=\frac{3}{10} & p(F=l|B=b)=0 & p(F=l|B=g)=\frac{2}{5}. \end{array}$$

We choose a box with probability determined as above, and from it sample a fruit, likewise using the above-defined probabilities. In the corresponding experiment, we seek to compute the probability of selecting an apple (i.e.: $p(F=a)$). From the sum probability rule in (1.10), it follows that

$$\begin{aligned} p(F=a) &= p(F=a|B=r)p(B=r)+p(F=a|B=b)p(B=b)+ \\ &\quad +p(F=a|B=g)p(B=g) \\ &= \frac{3}{10}\cdot\frac{1}{5}+\frac{1}{2}\cdot\frac{1}{5}+\frac{3}{10}\cdot\frac{3}{5} \\ p(F=a) &= \frac{17}{50}. \end{aligned}$$

Thereafter, we seek to compute the probability that the green box was picked, given that an orange was selected (i.e.: $p(B=g|F=o)$). It follows from Bayes' Theorem in (1.12) that

$$\begin{aligned} p(B=g|F=o) &= \frac{p(F=o|B=g)p(B=g)}{p(F=o)} \\ &= \frac{p(F=o|B=g)p(B=g)}{\sum_{k\in\{r,b,g\}} p(F=o|B=k)p(B=k)} \\ &= \frac{\frac{3}{10}\frac{3}{5}}{\frac{2}{5}\frac{1}{5}+\frac{1}{2}\frac{1}{5}+\frac{3}{10}\frac{3}{5}} \\ &= \frac{\frac{9}{50}}{\frac{4+5+9}{50}} \\ p(B=g|F=o) &= \frac{1}{2}. \end{aligned}$$

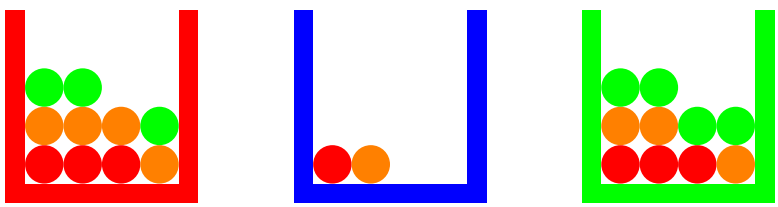


Figure 1.1: Illustration of the boxes described in Exercise 1.3.

Exercise 1.4

Let $p_X(x)$ be a probability density function defined for a continuous random variable X, and suppose that we implicitly define a general (possibly nonlinear) transformation of X by $X = g(Y)$, where $g(y)$ is differentiable. It follows from (1.27) that the probability density function of Y is

$$p_Y(y) = p_X(g(y))\left|\frac{\mathrm{d}g(y)}{\mathrm{d}y}\right| = \begin{cases} p_X(g(y))\frac{\mathrm{d}g(y)}{\mathrm{d}y} & \text{if } \frac{\mathrm{d}g(y)}{\mathrm{d}y} \geq 0, \\ -p_X(g(y))\frac{\mathrm{d}g(y)}{\mathrm{d}y} & \text{if } \frac{\mathrm{d}g(y)}{\mathrm{d}y} < 0. \end{cases}$$

Assume herein that the maximum of $p_Y(y)$ is restricted to locations wherein ${}^{\mathrm{d}p_Y(y)}/_{\mathrm{d}y} = 0$. It follows that a maximum for $p_Y(y)$ may be determined by differentiating it with respect to y, as follows

$$\frac{\mathrm{d}p_Y(y)}{\mathrm{d}y} = \begin{cases} \frac{\mathrm{d}p_X(g(y))}{\mathrm{d}x}\left(\frac{\mathrm{d}g(y)}{\mathrm{d}y}\right)^2 + p_X(g(y))\frac{\mathrm{d}^2g(y)}{\mathrm{d}y^2} & \text{if } \frac{\mathrm{d}g(y)}{\mathrm{d}y} > 0, \\ \frac{\mathrm{d}p_X(g(y))}{\mathrm{d}x}\left(\frac{\mathrm{d}g(y)}{\mathrm{d}y}\right)^2 - p_X(g(y))\frac{\mathrm{d}^2g(y)}{\mathrm{d}y^2} & \text{if } \frac{\mathrm{d}g(y)}{\mathrm{d}y} < 0. \end{cases}$$

The context where ${}^{\mathrm{d}g(y)}/_{\mathrm{d}y} = 0$ is ignored herein. Take $\tilde{y}$ such that the functional relation $g(\tilde{y}) = \hat{x}$ is satisfied, where $\hat{x}$ is the maximum density location over X. Write

$$\begin{aligned} \frac{\mathrm{d}p_Y(\tilde{y})}{\mathrm{d}y} &= \begin{cases} \frac{\mathrm{d}p_X(g(\tilde{y}))}{\mathrm{d}x}\left(\frac{\mathrm{d}g(\tilde{y})}{\mathrm{d}y}\right)^2 + p_X(g(\tilde{y}))\frac{\mathrm{d}^2g(\tilde{y})}{\mathrm{d}y^2} & \text{if } \frac{\mathrm{d}g(\tilde{y})}{\mathrm{d}y} > 0, \\ \frac{\mathrm{d}p_X(g(\tilde{y}))}{\mathrm{d}x}\left(\frac{\mathrm{d}g(\tilde{y})}{\mathrm{d}y}\right)^2 - p_X(g(\tilde{y}))\frac{\mathrm{d}^2g(\tilde{y})}{\mathrm{d}y^2} & \text{if } \frac{\mathrm{d}g(\tilde{y})}{\mathrm{d}y} < 0. \end{cases} \\ &= \begin{cases} p_X(\hat{x})\frac{\mathrm{d}^2g(\tilde{y})}{\mathrm{d}y^2} & \text{if } \frac{\mathrm{d}g(\tilde{y})}{\mathrm{d}y} > 0, \\ -p_X(\hat{x})\frac{\mathrm{d}^2g(\tilde{y})}{\mathrm{d}y^2} & \text{if } \frac{\mathrm{d}g(\tilde{y})}{\mathrm{d}y} < 0. \end{cases} \end{aligned}$$

As $\hat{x}$ is a maximum density location and $p_X(x)$ is a valid probability density function, by assumption, it follows that $p_X(\hat{x}) > 0$. Solving for ${}^{\mathrm{d}p_Y(\tilde{y})}/_{\mathrm{d}y} = 0$, we find that

$$\frac{\mathrm{d}p_Y(\tilde{y})}{\mathrm{d}y} = 0 \quad \Longleftrightarrow \quad \frac{\mathrm{d}^2g(\tilde{y})}{\mathrm{d}y^2} = 0. \tag{1.3}$$

Consequently, that $\hat{x}$ is a maximum density location over X does not provide sufficient conditions to conclude that $\tilde{y}$ such that $\hat{x} = g(\tilde{y})$ is a maximum density location over Y: $\tilde{y}$ (and $g(y)$) must also satisfy the condition in (1.3). Assume now that the transformation is $X = g(Y) = Y + a$, where $a \in \mathbb{R}$ is a constant. It likewise follows from (1.27) that the probability density function of Y is

$$\begin{aligned} p_Y(y) &= p_X(y+a)\left|\frac{\mathrm{d}(y+a)}{\mathrm{d}y}\right| \\ &= p_X(y+a)|1| \\ p_Y(y) &= p_X(y+a). \end{aligned}$$

The density maximum location over Y may be determined by differentiating $p_Y(y)$ with respect to y, as follows

$$\frac{\mathrm{d}p_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}p_X(y+a)}{\mathrm{d}y}.$$

We now solve $\mathrm{d}p_X(y+a)/\mathrm{d}y = 0$. By assumption, we know that $\mathrm{d}p_X(\hat{x})/\mathrm{d}x = 0$. It follows that, for $\hat{y}$ such that $\hat{y} + a = \hat{x}$, we solve $\mathrm{d}p_X(\hat{y})/\mathrm{d}y = 0$. We conclude that the density maximum location over X is $\hat{x} = \hat{y} + a$, that is, the maximum density location is transformed analogously to the random variable (one may also note that linear transformations satisfy the condition in (1.3)).

Exercise 1.5

From the definition (1.38), it follows that

$$\begin{aligned}\mathbb{V}\text{ar}[f(X)] &= \mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2] \\ &= \mathbb{E}[\{f(X)\}^2 - 2f(X)\mathbb{E}[f(X)] + \{\mathbb{E}[f(X)]\}^2] \\ &= \mathbb{E}[\{f(X)\}^2] - 2\mathbb{E}[f(X)]\mathbb{E}[f(X)] + \{\mathbb{E}[f(X)]\}^2 \\ \mathbb{V}\text{ar}[f(X)] &= \mathbb{E}[\{f(X)\}^2] - \{\mathbb{E}[f(X)]\}^2.\end{aligned}$$

Hence we derive (1.39).

Exercise 1.6

Let X and Y be independent random variables with joint density function $p(x,y)$. For this Exercise, the variables are assumed to be continuous. It follows from (1.41) that the covariance between X and Y is computed as

$$\begin{aligned}
\mathbb{C}\text{ov}[X,Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xyp(x,y)\,\mathrm{d}x\mathrm{d}y - \left[\int_{\mathbb{R}} xp(x)\,\mathrm{d}x\right]\left[\int_{\mathbb{R}} yp(y)\,\mathrm{d}y\right] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xyp(x)p(y)\,\mathrm{d}x\mathrm{d}y - \left[\int_{\mathbb{R}} xp(x)\,\mathrm{d}x\right]\left[\int_{\mathbb{R}} yp(y)\,\mathrm{d}y\right] && \text{(Independence)} \\
&= \int_{\mathbb{R}} xp(x)\left[\int_{\mathbb{R}} yp(y)\,\mathrm{d}y\right]\mathrm{d}x - \left[\int_{\mathbb{R}} xp(x)\,\mathrm{d}x\right]\left[\int_{\mathbb{R}} yp(y)\,\mathrm{d}y\right] \\
&= \left[\int_{\mathbb{R}} xp(x)\,\mathrm{d}x\right]\left[\int_{\mathbb{R}} yp(y)\,\mathrm{d}y\right] - \left[\int_{\mathbb{R}} xp(x)\,\mathrm{d}x\right]\left[\int_{\mathbb{R}} yp(y)\,\mathrm{d}y\right] \\
\mathbb{C}\text{ov}[X,Y] &= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] && \text{(Apply (1.34)).}
\end{aligned}$$

We thereby conclude

$$\mathbb{C}\text{ov}[X,Y] = 0. \tag{1.4}$$

We conclude that the covariance between two independent random variables is zero. The demonstration for discrete random variables is analogous.

Exercise 1.7

Let I be defined as the integral in (1.124). We aim to demonstrate that $I = \sqrt{2\pi\sigma^2}$. It is as follows:

$$\begin{aligned}
I &= \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \mathrm{d}x \\
I \cdot I &= \left[\int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \mathrm{d}x\right]\left[\int_{-\infty}^{\infty} \exp\left\{-\frac{y^2}{2\sigma^2}\right\} \mathrm{d}y\right] \quad \text{(Multiply both sides by } I\text{)} \\
I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} \mathrm{d}x\mathrm{d}y.
\end{aligned}$$

Perform a variable transform on x and y to polar coordinates, obtaining

$$\begin{aligned}
I^2 &= \int_{-\pi}^{\pi}\int_{0}^{\infty} \exp\left\{-\frac{(r\cos\theta)^2 + (r\sin\theta)^2}{2\sigma^2}\right\} \left|\begin{matrix}\cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta\end{matrix}\right| \mathrm{d}r\mathrm{d}\theta \\
&= \int_{-\pi}^{\pi}\int_{0}^{\infty} |r(\cos\theta)^2 + r(\sin\theta)^2| \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathrm{d}r\mathrm{d}\theta \\
&= \int_{-\pi}^{\pi}\int_{0}^{\infty} |r| \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \mathrm{d}r\mathrm{d}\theta \\
&= \int_{-\pi}^{\pi}\int_{0}^{\infty} \frac{1}{2}\exp\left\{-\frac{u}{2\sigma^2}\right\} \mathrm{d}u\mathrm{d}\theta \qquad \text{(Set } u = r^2\text{)} \\
&= \frac{1}{2}\int_{-\pi}^{\pi} 2\sigma^2 \, \mathrm{d}\theta \\
I^2 &= 2\pi\sigma^2 \\
I &= \sqrt{2\pi\sigma^2}.
\end{aligned}$$

The normal probability density function with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is as in (1.46). We aim to demonstrate herein that $\int_{\mathbb{R}} p(x|\mu,\sigma^2)\,\mathrm{d}x = 1$. It follows that

$$\begin{aligned}
\int_{\mathbb{R}} p(x|\mu,\sigma^2)\,\mathrm{d}x &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathbf{1}_{\mathbb{R}}(x)\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left\{-\frac{v^2}{2\sigma^2}\right\} \mathbf{1}_{\mathbb{R}}(v+\mu)\,\mathrm{d}v \qquad \text{(Set } v = x-\mu\text{)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{v^2}{2\sigma^2}\right\} \mathrm{d}v \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot I \qquad \text{(Apply (1.124))} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \sqrt{2\pi\sigma^2} \\
\int_{\mathbb{R}} p(x|\mu,\sigma^2)\,\mathrm{d}x &= 1.
\end{aligned}$$

Exercise 1.8

We aim to compute the expected value of a normally distributed random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. From (1.49), write

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x p(x|\mu,\sigma^2)\,\mathrm{d}x && \text{(Apply (1.34))}\\
&= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x && \text{(Apply (1.46))}\\
&= \int_{-\infty}^{\infty} \frac{u+\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\}\mathrm{d}u && \text{(Set } u = x-\mu\text{)}\\
&= \int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\}\mathrm{d}u + \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\}\mathrm{d}u\\
&= \int_{0}^{\infty} \frac{1}{2\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}v\right\}\mathrm{d}v - \int_{-\infty}^{0} \frac{1}{2\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}v\right\}\mathrm{d}v +\\
&+ \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\}\mathrm{d}u && \text{(Set } v = u^2\text{)}\\
&= \frac{2\sigma^2}{2\sqrt{2\pi\sigma^2}} - \frac{2\sigma^2}{2\sqrt{2\pi\sigma^2}} + \frac{\mu}{\sqrt{2\pi\sigma^2}}\cdot\sqrt{2\pi\sigma^2} && \text{(Apply (1.124))}\\
\mathbb{E}[X] &= \mu.
\end{aligned}$$

Differentiating the normalizing condition (1.48) from both sides with respect to σ^2 results in

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x\right]\\
0 &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\left[\overbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}^{f(\sigma^2)} \overbrace{\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}^{g(\sigma^2)}\right]\mathrm{d}x\\
0 &= \int_{-\infty}^{\infty} \left[\overbrace{-\frac{1}{2\sigma^2\sqrt{2\pi\sigma^2}}}^{f'(\sigma^2)} \overbrace{\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}^{g(\sigma^2)} +\right.\\
&\left. + \overbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}^{f(\sigma^2)} \overbrace{\frac{(x-\mu)^2}{2\sigma^4}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}^{g'(\sigma^2)}\right]\mathrm{d}x.
\end{aligned}$$

It follows that

$$\begin{aligned}
0 &= \frac{1}{2\sigma^4}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x + \\
&\quad -\frac{1}{2\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x \\
0 &= \int_{-\infty}^{\infty}\frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x + \\
&\quad -\sigma^2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x \\
0 &= \mathbb{E}[(X-\mu)^2] - \sigma^2\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\sqrt{2\pi\sigma^2} && \text{(Apply (1.124))} \\
\mathbb{E}[(X-\mathbb{E}[X])^2] &= \sigma^2 \\
\mathbb{V}\mathrm{ar}[X] &= \sigma^2 && \text{(Apply (1.38))}.
\end{aligned}$$

As seen in (1.39), the variance of a random variable may be decomposed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \sigma^2 \\
\mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 &= \sigma^2 \\
\mathbb{E}[X^2] - \mu^2 &= \sigma^2 \\
\mathbb{E}[X^2] &= \sigma^2 + \mu^2.
\end{aligned}$$

Exercise 1.9

The mode, or maximum density location, of the normal density function with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, may be determined as the point which maximizes the normal density function, or equivalently the logarithm of the density function, as follows

$$\arg\max_{x\in\mathbb{R}} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\right] = \arg\min_{x\in\mathbb{R}}(x-\mu)^2.$$

We conclude that the mode occurs at $\hat{x} = \mu$. For a D-dimensional normal random vectors, with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$, the maximum may be determined as follows

$$\begin{aligned}&\arg\max_{\mathbf{x}\in\mathbb{R}^D} \log\left[\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right]\\ &= \arg\min_{\mathbf{x}\in\mathbb{R}^D} \operatorname{tr}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}].\end{aligned}$$

We conclude that the mode occurs at $\hat{\mathbf{x}} = \boldsymbol{\mu}$.

Exercise 1.10

Let X and Z be independent random variables with joint density function $p(x, z)$. For this exercise, the variables are assumed to be continuous. It follows that the expected value of $X + Z$ is determined by

$$\begin{aligned}
\mathbb{E}[X+Z] &= \int_{\mathbb{R}}\int_{\mathbb{R}}(x+z)p(x,z)\,\mathrm{d}x\mathrm{d}z && \text{(Apply (1.34))}\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}(x+z)p(x)p(z)\,\mathrm{d}x\mathrm{d}z && \text{(Independence)}\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}xp(x)p(z)\,\mathrm{d}x\mathrm{d}z + \int_{\mathbb{R}}\int_{\mathbb{R}}zp(x)p(z)\,\mathrm{d}x\mathrm{d}z\\
&= \int_{\mathbb{R}}p(z)\left[\int_{\mathbb{R}}xp(x)\,\mathrm{d}x\right]\mathrm{d}z + \int_{\mathbb{R}}p(x)\left[\int_{\mathbb{R}}zp(z)\,\mathrm{d}z\right]\mathrm{d}x\\
&= \int_{\mathbb{R}}p(z)\mathbb{E}[X]\,\mathrm{d}z + \int_{\mathbb{R}}p(x)\mathbb{E}[Z]\,\mathrm{d}x && \text{(Apply (1.34))}\\
\mathbb{E}[X+Z] &= \mathbb{E}[X] + \mathbb{E}[Z] && \text{(Apply (1.30)).}
\end{aligned}$$

Also, the variance of $X + Z$ is determined by

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X+Z] &= \mathbb{E}[(X+Z-\mathbb{E}[X+Z])^2]\\
&= \mathbb{E}[(X+Z-\mathbb{E}[X]-\mathbb{E}[Z])^2]\\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Z-\mathbb{E}[Z])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Z-\mathbb{E}[Z])]\\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Z] + 2\mathbb{C}\mathrm{ov}[X,Z]\\
\mathbb{V}\mathrm{ar}[X+Z] &= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Z].
\end{aligned}$$

The result that, for two independent random variables the covariance is zero, as seen in (1.4), was utilized above. The demonstration in the discrete case is analogous.

Exercise 1.11

To determine the maximum likelihood estimators of the normal density function with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, first we take the logarithm of the likelihood function, yielding (1.54), and differentiate it with respect to μ, obtaining

$$\frac{\mathrm{d} \log p(\mathbf{x}|\mu, \sigma^2)}{\mathrm{d}\mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu).$$

Solving ${}^{\mathrm{d} \log p(\mathbf{x}|\mu,\sigma^2)}/_{\mathrm{d}\mu} = 0$, it follows that

$$\begin{aligned} \frac{\mathrm{d} \log p(x|\mu, \sigma^2)}{\mathrm{d}\mu} &= 0 \\ \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) &= 0 \\ \sum_{n=1}^{N} x_n - N\mu &= 0 \\ \mu &= \frac{\sum_{n=1}^{N} x_n}{N}. \end{aligned}$$

Differentiating (1.54) once more with respect to μ, one obtains

$$\frac{\mathrm{d}^2 \log p(\mathbf{x}|\mu, \sigma^2)}{\mathrm{d}\mu^2} = -\frac{N}{\sigma^2} < 0.$$

Consequently, the maximum likelihood estimate of μ is as in (1.55). Conversely, differentiating (1.54) with respect to σ^2, you obtain

$$\frac{\mathrm{d} \log p(\mathbf{x}|\mu, \sigma^2)}{\mathrm{d}\sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2.$$

Solving ${}^{\mathrm{d} \log p(\mathbf{x}|\mu,\sigma^2)}/_{\mathrm{d}\sigma^2} = 0$, it follows that

$$\begin{aligned} \frac{\mathrm{d} \log p(\mathbf{x}|\mu, \sigma^2)}{\mathrm{d}\sigma^2} &= 0 \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 &= 0 \\ \sigma^2 &= \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2. \end{aligned}$$

Differentiating (1.54) once more with respect to σ^2, one obtains

$$\frac{\mathrm{d}^2 \log p(\mathbf{x}|\mu, \sigma^2)}{\mathrm{d}(\sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{n=1}^{N} (x_n - \mu)^2. \tag{1.5}$$

Evaluating (1.5) at $\sigma^2 = \sum_{n=1}^{N} (x_n - \mu)^2/N$ gives

$$\begin{aligned} \frac{\mathrm{d}^2 \log p(x|\mu, \sum_{n=1}^{N} (x_n - \mu)^2/N)}{\mathrm{d}(\sigma^2)^2} &= \frac{N^3}{2[\sum_{n=1}^{N} (x_n - \mu)^2]^2} - \frac{N^3}{[\sum_{n=1}^{N} (x_n - \mu)^2]^2} \\ &= -\frac{N^3}{[2\sum_{n=1}^{N} (x_n - \mu)^2]^2} < 0. \end{aligned}$$

Note that the maximum likelihood estimator of σ^2 is dependent on μ, which is unknown, whereas the maximum likelihood estimator of μ is not dependent on σ^2. We can therefore plug the maximum likelihood estimator of μ directly onto the maximum likelihood estimator of σ^2. Consequently, the maximum likelihood estimate of σ^2 is as in (1.56).

Exercise 1.12

Let $X_1, \dots, X_N$ be a sample of normally distributed independent random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let the expected value of $X_n X_m$ be written as

$$\begin{aligned}
\mathbb{E}[X_n X_m] &= \mathbb{C}\text{ov}[X_n, X_m] + \mathbb{E}[X_n]\mathbb{E}[X_m] \\
&= \begin{cases} \mu \cdot \mu + \mathbb{V}\text{ar}[X_n] & \text{if } n = m, \\ \mu \cdot \mu & \text{otherwise.} \end{cases} \\
\mathbb{E}[X_n X_m] &= \mu^2 + \sigma^2 I_{n,m}, \qquad (1.6)
\end{aligned}$$

where $I_{n,m}$ is such that $I_{n,m} = 1$ if $n = m$ and $I_{n,m} = 0$ if $n \neq m$. The result that, for two independent random variables the covariance is zero, as seen in (1.4), was utilized above. It follows that the expected value of the maximum likelihood estimators seen previously is as follows

$$\begin{aligned}
\mathbb{E}[\mu_{\text{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[X_n] \\
&= \frac{1}{N}\sum_{n=1}^{N} \mu \\
&= \frac{1}{N} \cdot N \cdot \mu \\
\mathbb{E}[\mu_{\text{ML}}] &= \mu,
\end{aligned}$$

and

$$
\begin{aligned}
\mathbb{E}[\sigma^2_{\text{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(X_n - \mu_{\text{ML}})^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n^2 - 2X_n\mu_{\text{ML}} + \mu^2_{\text{ML}}] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[X_n^2 - 2X_n\frac{1}{N}\sum_{m=1}^{N}X_m + \left(\frac{1}{N}\sum_{m=1}^{N}X_m\right)^2\right] && \text{(Apply (1.55))} \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n^2] - \frac{2}{N^2}\sum_{n=1}^{N}\mathbb{E}\left[X_n\sum_{m=1}^{N}X_m\right] + \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{m=1}^{N}X_m\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n^2] - \frac{2}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[X_nX_m] + \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[X_nX_m] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n^2] - \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[X_nX_m] \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mu^2 + \sigma^2) - \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\mu^2 + \sigma^2 I_{n,m}) && \text{(Apply (1.6))} \\
&= \frac{1}{N}\cdot N\cdot\mu^2 + \frac{1}{N}\cdot N\cdot\sigma^2 - \frac{1}{N^2}\cdot N^2\cdot\mu^2 - \frac{1}{N^2}\cdot N\cdot\sigma^2 \\
&= \sigma^2 - \frac{\sigma^2}{N} \\
\mathbb{E}[\sigma^2_{\text{ML}}] &= \frac{N-1}{N}\sigma^2.
\end{aligned}
$$

Exercise 1.13

Modifying the maximum likelihood estimator of σ^2 by substituting the true value of μ in the place of μ_{ML}, the expected value of the estimator $\tilde{\sigma}^2_{\text{ML}}$ becomes

$$\begin{aligned}
\mathbb{E}[\tilde{\sigma}^2_{\text{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(X_n-\mu)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(X_n-\mu)^2] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(X_n-\mathbb{E}[X_n])^2] \quad \text{(Apply (1.38))} \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{V}\text{ar}[X_n] \\
&= \frac{1}{N}\sum_{n=1}^{N}\sigma^2 \\
\mathbb{E}[\tilde{\sigma}^2_{\text{ML}}] &= \sigma^2.
\end{aligned}$$

Exercise 1.14

Let W be a square matrix of dimension D, composed of elements $\{w_{i,j}\}$. Let W^S be a square matrix of dimension D composed of elements $\{w^S_{i,j}\}$ such that

$$\begin{aligned} w^S_{i,j} &= \frac{w_{i,j} + w_{j,i}}{2} \\ &= \frac{w_{j,i} + w_{i,j}}{2} \\ w^S_{i,j} &= w^S_{j,i}. \end{aligned}$$

Trivially, W^S is symmetric. Moreover, let W^A be a square matrix of dimension D composed of elements $\{w^A_{i,j}\}$ such that

$$\begin{aligned} w^A_{i,j} &= \frac{w_{i,j} - w_{j,i}}{2} \\ &= -\frac{w_{j,i} - w_{i,j}}{2} \\ w^A_{i,j} &= -w^A_{j,i}, \end{aligned} \tag{1.7}$$

note that $w^A_{i,I} = 0, \forall i \in \{1, \dots, D\}$. Trivially, W^A is antisymmetric. Lastly, see that

$$\begin{aligned} w^S_{i,j} + w^A_{i,j} &= \frac{w_{i,j} + w_{j,i}}{2} + \frac{w_{i,j} - w_{j,i}}{2} \\ &= \frac{w_{i,j} + w_{j,i} + w_{i,j} - w_{j,i}}{2} \\ &= \frac{2w_{i,j}}{2} \\ w^S_{i,j} + w^A_{i,j} &= w_{i,j}, \end{aligned} \tag{1.8}$$

and conclude that $W = W^S + W^A$, i.e., it is possible to decompose a square matrix as the sum of a symmetric matrix and an antisymmetric matrix. Returning to the context of polynomial regression, we consider the second-order term as in (1.131). Utilizing the

property that $W = W^S + W^A$, we decompose the second-order term as follows

$$\begin{aligned}
\sum_{i=1}^{D}\sum_{j=1}^{D} w_{i,j}x_ix_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w^S_{i,j}+w^A_{i,j})x_ix_j && \text{(Apply (1.8))}\\
&= \sum_{i=1}^{D}\sum_{j>i}^{D}(w^S_{i,j}+w^A_{i,j})x_ix_j + \sum_{i=1}^{D}(w^S_{i,i}+w^A_{i,i})(x_i)^2 + \\
&+ \sum_{i=1}^{D}\sum_{j<i}^{D}(w^S_{i,j}+w^A_{i,j})x_ix_j \\
&= \sum_{i=1}^{D}\sum_{j>i}^{D}(w^S_{i,j}+w^A_{i,j})x_ix_j + \sum_{i=1}^{D} w^S_{i,i}(x_i)^2 + \\
&+ \sum_{i=1}^{D}\sum_{j>i}^{D}(w^S_{j,i}+w^A_{j,i})x_jx_i && \text{(Apply } w^A_{i,i}=0\text{)}\\
&= \sum_{i=1}^{D}\sum_{j>i}^{D}(w^S_{i,j}+w^A_{i,j})x_ix_j + \sum_{i=1}^{D} w^S_{i,i}(x_i)^2 + \\
&+ \sum_{i=1}^{D}\sum_{j>i}^{D}(w^S_{j,i}-w^A_{i,j})x_jx_i && \text{(Apply (1.7))}\\
&= \sum_{i=1}^{D}\sum_{j>i}^{D} w^S_{i,j}x_ix_j + \sum_{i=1}^{D} w^S_{i,i}(x_i)^2 + \sum_{i=1}^{D}\sum_{j>i}^{D} w^S_{j,i}x_jx_i + \\
&+ \sum_{i=1}^{D}\sum_{j>i}^{D}(w^A_{i,j}-w^A_{i,j})x_ix_j \\
&= \sum_{i=1}^{D}\sum_{j>i}^{D} w^S_{i,j}x_ix_j + \sum_{i=1}^{D} w^S_{i,i}(x_i)^2 + \sum_{i=1}^{D}\sum_{j<i}^{D} w^S_{i,j}x_ix_j \\
\sum_{i=1}^{D}\sum_{j=1}^{D} w_{i,j}x_ix_j &= \sum_{i=1}^{D}\sum_{j=1}^{D} w^S_{i,j}x_ix_j.
\end{aligned}$$

Thereby concluding that the antisymmetric matrix contribution vanishes. We may count the number of independent elements in W^S (consequently W) as follows

$$\begin{aligned}
\sum_{i=1}^{D}\sum_{j=1}^{i} 1 &= \sum_{i=1}^{D}\sum_{j\geq 1}^{D} 1 \\
&= \sum_{i=1}^{D} 1 + \sum_{i=1}^{D}\sum_{j>1}^{D} 1 \\
&= D + \frac{D(D-1)}{2} \\
\sum_{i=1}^{D}\sum_{j=1}^{i} 1 &= \frac{D(D+1)}{2}.
\end{aligned} \tag{1.9}$$

Exercise 1.15

Consider the context of polynomial regression, such that for a model whose input space is of dimension D we hope to study the M^{th} order term, written as in (1.133). It is easy to note that the number of independent terms in the array $w_{i_1,i_2,\ldots,i_M}$ is equal to the number of unique unordered sequences of the form $\{j_1, j_2, \ldots, j_M\}$, where $j_k \in \{1, \ldots, D\}$ $\forall k \in \{1, \ldots, M\}$. We may define a class which contains all such sequences, and only said such sequences, by ordering the indexes so that $j_M \leq j_{M-1} \leq \ldots \leq j_1$, i.e., the following

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1,i_2,\ldots,i_M} x_{i_1}x_{i_2}\ldots x_{i_M} = \sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \tilde{w}_{i_1,i_2,\ldots,i_M} x_{i_1}x_{i_2}\ldots x_{i_M},$$

for some array $\tilde{w}_{i_1,i_2,\ldots,i_M}$. The number of independent elements for the M^{th} order term may be computed similarly to (1.9) as follows

$$\begin{aligned} n(D,M) &= \sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_{M-1}=1}^{i_{M-2}}\sum_{i_M=1}^{i_{M-1}} 1 \\ &= \sum_{i_1=1}^{D}\left[\sum_{i_2=1}^{i_1}\cdots\sum_{i_{M-1}=1}^{i_{M-2}}\sum_{i_M=1}^{i_{M-1}} 1\right] \\ n(D,M) &= \sum_{i_1=1}^{D} n(i_1, M-1). \end{aligned} \tag{1.10}$$

And hence, we conclude that the number of independent parameters satisfies a recurrence relation. We now seek to prove, by induction, (1.136). Firstly, for $D = 1$, it follows that

$$\begin{aligned} \sum_{i=1}^{1}\frac{(i+M-2)!}{(i-1)!(M-1)!} &= \frac{(1+M-1)!}{(1-1)!M!} \\ \frac{(M-1)!}{0!(M-1)!} &= \frac{M!}{0!M!} \\ 1 &= 1. \end{aligned}$$

We conclude that the result holds for $D = 1$. Assume that the result holds for D, we aim now to demonstrate it is valid for $D+1$, as in

$$\begin{aligned} \sum_{i=1}^{D+1}\frac{(i+M-2)!}{(i-1)!(M-1)!} &= \frac{(D+1+M-2)!}{(D+1-1)!(M-1)!} + \sum_{i=1}^{D}\frac{(i+M-2)!}{(i-1)!(M-1)!} \\ &= \frac{(D+M-1)!}{D!(M-1)!} + \frac{(D+M-1)!}{(D-1)!M!} && \text{(By assumption)} \\ &= \frac{(D+M-1)!}{D!M!}(M+D) \\ \sum_{i=1}^{D+1}\frac{(i+M-2)!}{(i-1)!(M-1)!} &= \frac{(D+M)!}{D!M!}. \end{aligned}$$

We conclude that, assuming (1.136) holds for D, it holds for $D+1$. By induction, we have proven (1.136) is true. Lastly, we seek to demonstrate the validity of (1.137), which

will be done by induction. Consider first that, for $D \geq 1$ and $M = 2$, it follows that

$$\begin{aligned} n(D,2) &= \frac{(D+2-1)!}{(D-1)!2!} \\ &= \frac{(D+1)!}{(D-1)!2} \\ &= \frac{(D+1)D(D-1)!}{(D-1)!2} \\ n(D,2) &= \frac{D(D+1)}{2}, \end{aligned}$$

which matches the result demonstrated in (1.9). Subsequently, assuming it holds for $M-1$, we seek to prove it holds for M, utilizing the recurrence relation seen in (1.10), as follows

$$\begin{aligned} n(D,M) &= \sum_{i=1}^{D} n(i, M-1) \\ &= \sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} \quad \text{(By assumption)} \\ n(D,M) &= \frac{(D+M-1)!}{(D-1)!M!}. \end{aligned}$$

Consequently, we have proved that (1.137) holds for $M = 2$ and, assuming it holds for $M-1$, it likewise holds for M. We hence conclude our demonstration by induction.

Exercise 1.16

We seek now to compute the number of independent coefficients in a polynomial regression model with terms up to and including the M^{th} order. It follows that said number is computed as

$$\begin{aligned} N(D,M) &= \sum_{m=0}^{M}\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_m=1}^{i_{m-1}} 1 \\ &= \sum_{m=0}^{M}\left[\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_m=1}^{i_{m-1}} 1\right] \\ N(D,M) &= \sum_{m=0}^{M} n(D,m). \end{aligned}$$

We seek to prove by induction (1.139). First, for $D \geq 1$ we show that (1.139) holds for $M = 0$. See that

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

Secondly, we assume that (1.139) holds for $M-1$, and seek to demonstrate it holds for M:

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

We conclude that (1.139) holds for $M = 0$ and, assuming it holds for $M-1$, it holds for M. We have therefore demonstrated (1.139) by induction. Now, assuming $D \gg M$, we apply Stirling's approximation in the form (1.140) to $N(D,M)$, assuming also D is

sufficiently large, obtaining the following

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

i.e., for sufficiently large D (and $D \gg M$), it follows that $N(D,M)$ grows in a rate approximately proportional to D^M. For $M \gg D$, and M sufficiently large, it follows that

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

i.e., for sufficiently large M (and $M \gg D$), it follows that $N(D,M)$ grows in a rate approximately proportional to M^D. For the cubic polynomial regression model ($M=3$), it follows that $N(10,3)$ and $N(100,3)$ are computed as follows

$$\begin{aligned} N(10,3) &= \frac{(10+3)!}{10!3!} \\ &= \frac{13!}{10!3!} \\ &= \frac{13\cdot 12\cdot 11}{6} \\ N(10,3) &= 286. \end{aligned}$$

and

$$\begin{aligned} N(100,3) &= \frac{(100+3)!}{100!3!} \\ &= \frac{103!}{100!3!} \\ &= \frac{103\cdot 102\cdot 101}{6} \\ N(100,3) &= 176851. \end{aligned}$$

Exercise 1.17

The gamma function is defined as in (1.141). We seek to prove that $\Gamma(x+1) = x\Gamma(x)$. Inspecting $\Gamma(x+1)$, we see that

$$\begin{aligned}
\Gamma(x+1) &= \int_0^\infty u^{x+1-1} e^{-u}\, \mathrm{du} \\
&= \int_0^\infty u^x e^{-u}\, \mathrm{du} \\
&= -\Big[u^x e^{-u}\Big]\Big|_0^\infty + \int_0^\infty x u^{x-1} e^{-u}\, \mathrm{du} \qquad \text{(Integration by parts)} \\
&= 0 - \lim_{u\to\infty} u^x e^{-u} + x\int_0^\infty u^{x-1} e^{-u}\, \mathrm{du} \\
\Gamma(x+1) &= x\Gamma(x) \qquad \text{(Apply (1.141)).} \qquad (1.11)
\end{aligned}$$

Subsequently, we show that $\Gamma(1) = 1$:

$$\begin{aligned}
\Gamma(1) &= \int_0^\infty u^{1-1} e^{-u}\, \mathrm{du} \\
&= \int_0^\infty u^0 e^{-u}\, \mathrm{du} \\
&= \int_0^\infty e^{-u}\, \mathrm{du} \\
&= -\Big[e^{-u}\Big]\Big|_0^\infty \\
&= 1 - 0 \\
\Gamma(1) &= 1. \qquad (1.12)
\end{aligned}$$

We now prove by induction that, for $x \in \mathbb{N}$ it holds that $\Gamma(x+1) = x!$. The case for $x+1 = 1$ in (1.12). Assuming $\Gamma(x+1) = x!$ holds, we now seek to demonstrate it holds for $x+2$.

$$\begin{aligned}
\Gamma(x+2) &= (x+1)\Gamma(x+1) \\
&= (x+1)x! \\
\Gamma(x+2) &= (x+1)!.
\end{aligned}$$

Thereby concluding the proof.

Exercise 1.18

Let $D \geq 1$ and S_D denote the surface area of a D-dimensional unit radius sphere, we may rewrite the result in (1.142) as

$$\begin{aligned}\prod_{i=1}^{D}\left[\int_{-\infty}^{\infty} e^{-x_i^2}\,\mathrm{d}x_i\right] &= S_D \int_0^{\infty} e^{-r^2} r^{D-1}\,\mathrm{d}r \\ &= S_D \int_0^{\infty} e^{-r^2} r^{D-2} r\,\mathrm{d}r \\ &= S_D \int_0^{\infty} e^{-r^2} (r^2)^{D/2-1} r\,\mathrm{d}r.\end{aligned}$$

By applying the transformations $x_i^2 = y_i^2/2$, such that $\mathrm{d}x_i = \mathrm{d}y_i/\sqrt{2}$ and $r^2 = u$, such that $2r\mathrm{d}r = \mathrm{d}u$, we can rewrite the previous result as

$$\begin{aligned}\prod_{i=1}^{D}\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2}} e^{-\frac{1}{2}y_i^2}\,\mathrm{d}y_i\right] &= \frac{1}{2} S_D \int_0^{\infty} e^{-u} u^{D/2-1}\,\mathrm{d}u \\ 2^{-D/2}\prod_{i=1}^{D} I &= \frac{1}{2} S_D \Gamma(D/2) \\ 2^{-D/2}(2\pi)^{D/2} &= \frac{1}{2} S_D \Gamma(D/2) \\ S_D &= \frac{2\pi^{D/2}}{\Gamma(D/2)},\end{aligned}$$

where I is the integral seen in (1.124), hence matching the result (1.143). It follows that the volume of the unit sphere in D dimensions is computed as

$$\begin{aligned}V_D &= \int_0^1 r^{D-1} S_D\,\mathrm{d}r \\ &= \frac{S_D}{D},\end{aligned}$$

hence matching the result (1.144). In particular, for $D = 2$ and $D = 3$, we have that

$$\begin{aligned}S_2 &= \frac{2\pi^{2/2}}{\Gamma(2/2)} \\ &= \frac{2\pi}{\Gamma(1)} \\ S_2 &= 2\pi,\end{aligned}$$

and

$$\begin{aligned}S_3 &= \frac{2\pi^{3/2}}{\Gamma(3/2)} \\ &= \frac{2\pi^{3/2}}{\pi^{1/2}/2} \\ S_3 &= 4\pi.\end{aligned}$$

Exercise 1.19

Let H_D denote the volume of the D dimensional hypercube of side $2a$ ($a > 0$), and V_D denote the volume of the D dimensional sphere of radius a, computed as

$$\begin{aligned} V_D &= \int_0^a r^{D-1} S_D \, \mathrm{d}r \\ &= \frac{a^D S_D}{D} \\ V_D &= \frac{2\pi^{D/2} a^D}{D\Gamma(D/2)} \end{aligned}$$

It follows that the ratio between the volume of the sphere and the volume of the hypercube is

$$\begin{aligned} \frac{V_D}{H_D} &= \frac{\frac{2\pi^{D/2}a^D}{D\Gamma(D/2)}}{(2a)^D} \\ &= \frac{2\pi^{D/2}a^D}{D\Gamma(D/2)2^D a^D} \\ \frac{V_D}{H_D} &= \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}. \end{aligned}$$

Utilizing Stirling's approximation in the form (1.146), assuming D is sufficiently large ($D \gg 1$), we find that this ratio may be written as

$$\begin{aligned} \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} &\approx \frac{\pi^{D/2}}{D2^{D-1}(2\pi)^{1/2}e^{-D/2}(D/2)^{(D+1)/2}} \\ &= \frac{2^{1/2}\pi^{D/2}2^{D/2}e^{D/2}}{D2^{D-1}2^{1/2}\pi^{1/2}D^{D/2}D^{1/2}} \\ &= \frac{2}{\pi^{1/2}}\frac{\pi^{D/2}2^{D/2}e^{D/2}}{D^{D/2+3/2}2^D} \\ &= \frac{2}{\pi^{1/2}}\frac{\pi^{D/2}e^{D/2}}{D^{D/2+3/2}2^{D/2}} \\ &= \frac{2}{\pi^{1/2}D^{3/2}}\frac{\pi^{D/2}e^{D/2}}{D^{D/2}2^{D/2}} \end{aligned}$$

$$\frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} = \frac{2}{\pi^{1/2}D^{3/2}}\left(\frac{(\pi e)/2}{D}\right)^{D/2}. \tag{1.13}$$

We note that, for $D > 1$, the following relation holds

$$\frac{2}{\pi^{1/2}D^{3/2}}\left(\frac{(\pi e)/2}{D}\right)^{D/2} \leq \frac{2}{\pi^{1/2}}\left(\frac{(\pi e)/2}{D}\right)^{D/2}$$

Note that the left-hand side is strictly greater than zero for all $D \gg 1$. Moreover, by applying the limit when $D \to \infty$ to both sides, we obtain

$$\begin{aligned} 0 \leq \lim_{D\to\infty} \frac{2}{\pi^{1/2}D^{3/2}}\left(\frac{(\pi e)/2}{D}\right)^{D/2} &\leq \lim_{D\to\infty} \frac{2}{\pi^{1/2}}\left(\frac{(\pi e)/2}{D}\right)^{D/2} \\ &\leq 0. \end{aligned}$$

We conclude that the ratio in (1.13) approaches zero as $D \to \infty$. We now compute the Euclidean distance of the centre of the hypercube to one of its corners, all of which are equidistant, as follows:

$$\begin{aligned} \ell_D &= \sqrt{\sum_{i=1}^{D} a^2} \\ &= \sqrt{Da^2} \\ \ell_D &= a\sqrt{D}. \end{aligned}$$

the distance from the centre of the hypercube to the centre of one of its sides is

$$c_D = a.$$

Consequently, the ratio between the two distances is

$$\begin{aligned} \frac{\ell_D}{c_D} &= \frac{\sqrt{D}a}{a} \\ &= \sqrt{D}, \end{aligned}$$

which trivially goes to infinity as $D \to \infty$.

Exercise 1.20

Consider the D dimensional normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance $\boldsymbol{\Sigma} = \sigma^2 I$, for $\sigma^2 > 0$, with density function as in (1.147). In order to determine the distribution of the radius of $\mathbf{x}$, we define $r^2 = \sum_{i=1}^{D} x_i^2$ via a spherical coordinate transform, and marginalize with respect to the angular coordinates, yielding the following volume element

$$\mathrm{d}\mathbf{x} = S_D r^{D-1} \mathrm{d}r,$$

where S_D is the surface area of the D-dimensional unit sphere. This results in the following density function for r:

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left\{ -\frac{r^2}{2\sigma^2} \right\} \mathbf{1}_{[0,\infty)}(r).$$

We aim to determine the maximum density location of $p(r)$ by first differentiating it with respect to r as follows

$$\begin{aligned}
\frac{\mathrm{d}p(r)}{\mathrm{d}r} &= \frac{\mathrm{d}}{\mathrm{d}r}\left[\overbrace{\frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}}}^{f(r)} \overbrace{\exp\left\{ -\frac{r^2}{2\sigma^2} \right\}}^{g(r)} \right] \\
&= \overbrace{\frac{(D-1) S_D r^{D-2}}{(2\pi\sigma^2)^{D/2}}}^{f'(r)} \overbrace{\exp\left\{ -\frac{r^2}{2\sigma^2} \right\}}^{g(r)} - \overbrace{\frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}}}^{f(r)} \overbrace{\frac{2r}{2\sigma^2} \exp\left\{ -\frac{r^2}{2\sigma^2} \right\}}^{-g'(r)} \\
\frac{\mathrm{d}p(r)}{\mathrm{d}r} &= \frac{S_D r^{D-2}}{(2\pi\sigma^2)^{D/2}} \exp\left\{ -\frac{r^2}{2\sigma^2} \right\} \left[D - 1 - \frac{r^2}{\sigma^2} \right].
\end{aligned}$$

Solving ${}^{\mathrm{d}p(r)}/_{\mathrm{d}r} = 0$, and assuming $D \gg 1$ (we also discard solutions where $r = 0$, which are points where the density is zero for $D \gg 1$), we find that

$$\begin{aligned}
\frac{\mathrm{d}p(r)}{\mathrm{d}r} = 0 \quad &\iff \quad \frac{S_D r^{D-2}}{(2\pi\sigma^2)^{D/2}} \exp\left\{ -\frac{r^2}{2\sigma^2} \right\} \left[D - 1 - \frac{r^2}{\sigma^2} \right] = 0 \\
\frac{\mathrm{d}p(r)}{\mathrm{d}r} = 0 \quad &\iff \quad D - 1 - \frac{r^2}{\sigma^2} = 0 \\
\frac{\mathrm{d}p(r)}{\mathrm{d}r} = 0 \quad &\iff \quad r = \sqrt{D-1}\sigma \\
\frac{\mathrm{d}p(r)}{\mathrm{d}r} = 0 \quad &\iff \quad r \approx \sqrt{D}\sigma.
\end{aligned}$$

Thereby concluding that, for $D \gg 1$, $\hat{r} = \sqrt{D}\sigma$ is a maximum density location. We now inspect the density at the location $\hat{r} + \varepsilon$. Consider the second order Taylor polynomial

expansion of $p(r)$ around $\hat{r}$ and evaluated at $\hat{r}+\varepsilon$, given as follows

$$
\begin{aligned}
p(\hat{r}+\varepsilon) &\approx p(\hat{r}) + \frac{\mathrm{d}p(\hat{r})}{\mathrm{d}r}\varepsilon + \frac{1}{2}\frac{\mathrm{d}^2p(\hat{r})}{\mathrm{d}r^2}\varepsilon^2 \\
&\approx p(\hat{r}) + \frac{1}{2}\left[\left(\frac{(D-2)S_D\hat{r}^{D-3}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}+\right.\right. \\
&\quad \left.\left. - \frac{S_D\hat{r}^{D-1}}{\sigma^2(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}\right)\left(D-1-\frac{\hat{r}^2}{\sigma^2}\right)+ \right. \\
&\quad \left. - \frac{2S_D\hat{r}^{D-1}}{\sigma^2(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}\right]\varepsilon^2 && \left(\text{Note } \frac{\mathrm{d}p(\hat{r})}{\mathrm{d}r}=0\right) \\
&= p(\hat{r}) + \frac{1}{2}\left[\left(\frac{(D-2)S_D\hat{r}^{D-2}}{\hat{r}(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}+\right.\right. \\
&\quad \left.\left. - \hat{r}\frac{S_D\hat{r}^{D-2}}{\sigma^2(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}\right)\left(D-1-\frac{\hat{r}^2}{\sigma^2}\right)+ \right. \\
&\quad \left. - \frac{2S_D\hat{r}^{D-1}}{\sigma^2(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}\right]\varepsilon^2 \\
&= p(\hat{r}) + \frac{1}{2}\left[\frac{(D-2)}{\hat{r}^2}\frac{\mathrm{d}p(\hat{r})}{\mathrm{d}r} - \frac{\hat{r}}{\sigma^2}\frac{\mathrm{d}p(\hat{r})}{\mathrm{d}r}+\right. \\
&\quad \left. - \frac{2}{\sigma^2}\frac{S_D\hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}\right]\varepsilon^2 \\
&= p(\hat{r}) - \frac{S_D\hat{r}^{D-1}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\hat{r}^2}{2\sigma^2}\right\}\frac{\varepsilon^2}{\sigma^2} && \left(\text{Note } \frac{\mathrm{d}p(\hat{r})}{\mathrm{d}r}=0\right) \\
&= p(\hat{r}) - p(\hat{r})\frac{\varepsilon^2}{\sigma^2} \\
&= p(\hat{r})(1-\varepsilon^2/\sigma^2) \\
&= p(\hat{r})\exp\left\{\log(1-\varepsilon^2/\sigma^2)\right\} \\
p(\hat{r}+\varepsilon) &\approx p(\hat{r})\exp\left\{-\frac{\varepsilon^2}{\sigma^2}\right\} && (\text{Use } \log(1+x)\approx x).
\end{aligned}
$$

We conclude then that the density decays away from $\hat{r}$ at an approximately exponential rate. Lastly, we evaluate the ratio $p(\mathbf{x}=\mathbf{0})/p(\mathbf{x}=\hat{r})$, as follows

$$
\begin{aligned}
\frac{p(\mathbf{x}=\mathbf{0})}{p(\mathbf{x}=\hat{r})} &= \frac{(2\pi\sigma^2)^{D/2}}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{0}{2\sigma^2}\right\}\exp\left\{\frac{\hat{r}^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{D}{2}\right\}.
\end{aligned}
$$

Exercise 1.21

Consider two nonnegative numbers a and b such that $a \leq b$. It follows that $a^{1/2} \leq b^{1/2}$. We may therefore write that

$$\begin{aligned} a &= a^{1/2}a^{1/2} \\ &\leq a^{1/2}b^{1/2} \\ &= (ab)^{1/2}. \end{aligned} \tag{1.14}$$

In a classification problem, in order to minimize the probability of committing a mistake, we must attribute our observations to the class to which it has the highest probability of belonging. Let $\mathcal{R}_1$ denote the region in which $p(\mathbf{x}, \mathcal{C}_1) \geq p(\mathbf{x}, \mathcal{C}_2)$ and $\mathcal{R}_2$ denote the region in which $p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1)$, the probability of committing a mistake is as in (1.78). Note that, as when constrained to $\mathcal{R}_1$ it follows that $p(\mathbf{x}, \mathcal{C}_2) < p(\mathbf{x}, \mathcal{C}_1)$, under that assumption we may write $(p(\mathbf{x}, \mathcal{C}_1), p(\mathbf{x}, \mathcal{C}_2))^{1/2}$ using the result in (1.14) (an analogous result is valid over $\mathcal{R}_2$). We can consequently rewrite the probability of committing a mistake as

$$\begin{aligned} p(\text{mistake}) &\leq \int_{\mathcal{R}_1} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_1 \cup \mathcal{R}_2} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Exercise 1.22

Consider a loss matrix defined as $L_{k,j} = 1 - I_{k,j}$, where I is the identity matrix. We define our classifying criterion by, for each $\mathbf{x}$, minimizing the following:

$$\begin{aligned}\sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) &= \sum_{\substack{k=1 \\ k\neq j}}^{K} p(\mathcal{C}_k|\mathbf{x}) \\ &= \sum_{k=1}^{K} p(\mathcal{C}_k|\mathbf{x}) - p(\mathcal{C}_j)\end{aligned}$$

$$\sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_j|\mathbf{x}) \qquad \text{(Use } \sum_{k=1}^{K} p(\mathcal{C}_k|\mathbf{x}) = 1\text{).} \tag{1.15}$$

That is, in order to minimize (1.15), me must choose the class j that minimizes $1 - p(\mathcal{C}_j|\mathbf{x})$, or equivalently, maximizes $p(\mathcal{C}_j|\mathbf{x})$, that is, the class having the largest posterior probability. The loss matrix $\{L_{k,j}\}$ is often referred to as the $0-1$ loss function, which simply returns whether you were correct or incorrect in your classification.

Exercise 1.23

For a general loss matrix, and general prior probabilities attributed to each class, if we classify an observation $\mathbf{x}$ as belonging to the m-th class, the resulting expected loss is

$$\sum_{j=1}^{K} \mathbf{1}_{\{j\}}(m) \sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}).$$

Therefore, in order to minimize the resulting expected loss, we must attribute an observation $\mathbf{x}$ to the class $\hat{j}$ for which $\sum_{k=1}^{K} L_{k,\hat{j}} p(\mathcal{C}_k|\mathbf{x})$ is minimal amongst the classes $j \in \{1, \dots, K\}$, which we write as

$$\hat{j} = \underset{j \in \{1,\dots,K\}}{\arg\min} \sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}).$$

Exercise 1.24

For a general loss matrix, general prior probabilities attributed to each class, and considering also that if we reject classifying an observation we incur a loss $\lambda \geq 0$. We increase the set of potential classes by adding the class $j = 0$, which denotes the rejection to classify. Therefore, the ideal classification decision is denoted by

$$\hat{j} = \underset{j\in\{0,1,\dots,K\}}{\operatorname{argmin}} \left[\sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) \mathbf{1}_{\{1,\dots,K\}}(j) + \lambda \mathbf{1}_{\{0\}}(j) \right]$$

$$\hat{j} = \begin{cases} \arg\min_{j\in\{1,\dots,K\}} \sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) & \text{if } \min_{j\in\{1,\dots,K\}} \sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) < \lambda, \\ 0 & \text{if } \min_{j\in\{1,\dots,K\}} \sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) \geq \lambda. \end{cases} \tag{1.16}$$

Assume the loss matrix is, as seen previously, such that $L_{k,j} = 1 - I_{k,j}$, where I is the identity matrix. It follows that the expected loss function for a fixed class j may be rewritten as

$$\begin{aligned} \sum_{k=1}^{K} L_{k,j} p(\mathcal{C}_k|\mathbf{x}) \mathbf{1}_{\{1,\dots,K\}}(j) + \lambda \mathbf{1}_{\{0\}}(j) &= \sum_{\substack{k=1 \\ k\neq j}}^{K} p(\mathcal{C}_k|\mathbf{x}) \mathbf{1}_{\{1,\dots,K\}}(j) + \lambda \mathbf{1}_{\{0\}}(j) \\ &= [1 - p(\mathcal{C}_j|\mathbf{x})] \mathbf{1}_{\{1,\dots,K\}}(j) + \lambda \mathbf{1}_{\{0\}}(j). \end{aligned}$$

Utilizing the result in (1.16), we find that

$$\begin{aligned} \hat{j} &= \begin{cases} \arg\min_{j\in\{1,\dots,K\}} \sum_{k=1}^{K} [1 - p(\mathcal{C}_j|\mathbf{x})] & \text{if } \min_{j\in\{1,\dots,K\}} [1 - p(\mathcal{C}_j|\mathbf{x})] < \lambda, \\ 0 & \text{if } \min_{j\in\{1,\dots,K\}} [1 - p(\mathcal{C}_j|\mathbf{x})] \geq \lambda. \end{cases} \\ &= \begin{cases} \arg\max_{j\in\{1,\dots,K\}} \sum_{k=1}^{K} p(\mathcal{C}_j|\mathbf{x}) & \text{if } 1 - \max_{j\in\{1,\dots,K\}} p(\mathcal{C}_j|\mathbf{x}) < \lambda, \\ 0 & \text{if } 1 - \max_{j\in\{1,\dots,K\}} p(\mathcal{C}_j|\mathbf{x}) \geq \lambda. \end{cases} \\ \hat{j} &= \begin{cases} \arg\max_{j\in\{1,\dots,K\}} \sum_{k=1}^{K} p(\mathcal{C}_j|\mathbf{x}) & \text{if } \max_{j\in\{1,\dots,K\}} p(\mathcal{C}_j|\mathbf{x}) > 1 - \lambda, \\ 0 & \text{if } \max_{j\in\{1,\dots,K\}} p(\mathcal{C}_j|\mathbf{x}) \leq 1 - \lambda. \end{cases} \end{aligned}$$

Thereby demonstrating this result relates to the rejection classifier seen previously, wherein the parameter λ is such that $1 - \lambda = \theta$ is equivalent to the discussed "rejection threshold", which rejects classification if the maximum posterior probability across all classes is lower than or equal to $1 - \lambda = \theta$.

Exercise 1.25

Consider that we desire to minimize the expected loss for multivariate input and target spaces, utilizing the expected loss function in (1.151). By differentiating (1.151) with respect to $\mathbf{y}(\mathbf{x})$ we obtain

$$\begin{aligned}
\frac{\mathrm{d}\mathbb{E}[L(\mathbf{T},\mathbf{y}(\mathbf{X}))]}{\mathrm{d}\mathbf{y}(\mathbf{x})} &= 2\int\int (\mathbf{y}(\mathbf{x})-\mathbf{t})p(\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\
&= 2\int\int (\mathbf{y}(\mathbf{x})-\mathbf{t})p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} && \text{(Apply (1.32))} \\
&= 2\int \left[\int \mathbf{y}(\mathbf{x})p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{t} - \int \mathbf{t}p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{t}\right]\mathrm{d}\mathbf{x} \\
&= 2\int \left[\mathbf{y}(\mathbf{x})p(\mathbf{x}) - \int \mathbf{t}p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{t}\right]\mathrm{d}\mathbf{x} && \text{(Apply (1.30))} \\
&= 2\int \left[\mathbf{y}(\mathbf{x})p(\mathbf{x}) - p(\mathbf{x})\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]\right]\mathrm{d}\mathbf{x} && \text{(Apply (1.37))} \\
\frac{\mathrm{d}\mathbb{E}[L(\mathbf{T},\mathbf{y}(\mathbf{X}))]}{\mathrm{d}\mathbf{y}(\mathbf{x})} &= 2\int [\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{X}]]p(\mathbf{x})\,\mathrm{d}\mathbf{x}.
\end{aligned}$$

It is straightforward to conclude that solving ${}^{\mathrm{d}\mathbb{E}[L(\mathbf{T},\mathbf{y}(\mathbf{X}))]}/_{\mathrm{d}\mathbf{y}(\mathbf{x})} = 0$ is equivalent to setting $\mathbf{y}(\mathbf{x}) = \mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]$. Moreover, the case for one-dimensional target space is such that

$$\begin{aligned}
y(\mathbf{x}) &= \mathbb{E}[T|\mathbf{X}=\mathbf{x}] \\
&= \int t p(t|\mathbf{x})\,\mathrm{d}t \quad \text{(Apply (1.37))},
\end{aligned}$$

as previously seen.

Exercise 1.26

We now decompose the expected loss function seen in (1.151) as follows

$$
\begin{aligned}
\mathbb{E}[L(\mathbf{T},\mathbf{y}(\mathbf{X}))] &= \int\int ||\mathbf{y}(\mathbf{x})-\mathbf{t}||^2 p(\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\
&= \int\int ||\mathbf{y}(\mathbf{x})-\mathbf{t}||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} && \text{(Apply (1.32))} \\
&= \int\int ||\mathbf{y}(\mathbf{x})-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]+ \\
&+ \mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]-\mathbf{t}||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\
&= \int\int ||\mathbf{y}(\mathbf{x})-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}+ \\
&+ \int\int ||\mathbf{t}-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t}+ \\
&+ 2\int\int (\mathbf{y}(\mathbf{x})-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}])\times \\
&\times (\mathbf{t}-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}])p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{t} \\
&= \int p(\mathbf{x})\left[\int ||\mathbf{y}(\mathbf{x})-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x})\,\mathrm{d}\mathbf{t}\right]\mathrm{d}\mathbf{x}+ \\
&+ \int p(\mathbf{x})\left[\int ||\mathbf{t}-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x})\,\mathrm{d}\mathbf{t}\right]\mathrm{d}\mathbf{x}+ \\
&+ 2\int \left[(\mathbf{y}(\mathbf{x})-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}])p(\mathbf{x})\times\right. \\
&\left.\times \int (\mathbf{t}-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathrm{x}])p(\mathbf{t}|\mathbf{x})\,\mathrm{d}\mathbf{t}\right]\mathrm{d}\mathbf{x} \\
\mathbb{E}[L(\mathbf{T},\mathbf{y}(\mathbf{X}))] &= \int ||\mathbf{y}(\mathbf{x})-\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]||^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}+ \\
&+ \int \mathbb{V}\mathrm{ar}^\star[\mathbf{T}|\mathbf{X}=\mathbf{x}]p(\mathbf{x})\,\mathrm{d}\mathbf{x} && \text{(Apply (1.37)).}
\end{aligned}
$$

As only the first component is dependent on $\mathbf{y}(\mathbf{x})$, minimizing the above is reduced by minimizing the first component, which is trivially attained when $\mathbf{y}(\mathbf{x})=\mathbb{E}[\mathbf{T}|\mathbf{X}=\mathbf{x}]$.

Exercise 1.27

Consider the L_q loss function in (1.91) for $q > 0$, which we rewrite as

$$\begin{aligned}\mathbb{E}[L_q(y(\mathbf{X}), T)|] &= \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\mathrm{d}t \\ &= \iint |y(\mathbf{x}) - t|^q p(t|\mathbf{x}) p(\mathbf{x})\, \mathrm{d}\mathbf{x}\mathrm{d}t \qquad \text{(Apply (1.32))} \\ \mathbb{E}[L_q(y(\mathbf{X}), T)|] &= \int p(\mathbf{x}) \left[\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\, \mathrm{d}t \right] \mathrm{d}\mathbf{x}.\end{aligned}$$

In order to determine $y(\mathbf{x})$ which minimizes the expected loss function, we must therefore minimize $\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\, \mathrm{d}t$. First we differentiate this term with respect to $y(\mathbf{x})$ (assuming this derivative exists), obtaining the following

$$\frac{\mathrm{d}}{\mathrm{d}y(\mathbf{x})} \left[\int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\, \mathrm{d}t \right] = q \int_{y(\mathbf{x})}^{\infty} |y(\mathbf{x}) - t|^{q-1} p(t|\mathbf{x})\, \mathrm{d}t - q \int_{-\infty}^{y(\mathbf{x})} |y(\mathbf{x}) - t|^{q-1} p(t|\mathbf{x})\, \mathrm{d}t$$

Solving the above for zero, we may determine the solution by solving the following

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

For $q = 1$, the solution is $y(\mathrm{x})$ such that

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

i.e., the median of $T|\mathbf{X} = \mathbf{x}$. On the other hand, for $q \to 0$, one must inspect the previous term directly: assume herein that $0^0 = 0$ and note that

$$\lim_{q \to 0} \int |y(\mathbf{x}) - t|^q p(t|\mathbf{x})\, \mathrm{d}t.$$

That is $\lim_{q \to 0} L_q$ is uniformly 1 across all values in the target space, except at the point in which $y(\mathbf{x}) = t$ (the true value of the target observation). Consequently, a sensible choice for estimator would consequently be the point of highest conditional likelihood for $T|\mathbf{X} = \mathbf{x}$.

Exercise 1.28

Let X be a discrete random variable with probability function $p(x)$, its associated entropy, $h(p)$, is determined as in (1.98). It follows that the entropy of $h(p^2)$ is written as

$$\begin{aligned} h(p^2) &= -\sum_{k\in\mathbb{N}} p(k)\log[p(k)]^2 \\ &= -2\sum_{k\in\mathbb{N}} p(k)\log p(k) \\ h(p^2) &= 2h(p) \qquad\qquad \text{(Apply (1.98)).} \end{aligned}$$

Assume herein that the entropy of $h(p^n) = nh(p)$, we hope to find the form of $h(p^{n+1})$: we obtain

$$\begin{aligned} h(p^{n+1}) &= -\sum_{k\in\mathbb{N}} p(k)\log[p(k)]^{n+1} \\ &= -\sum_{k\in\mathbb{N}} p(k)\log\{[p(k)]^n p(k)\} \\ &= -\sum_{k\in\mathbb{N}} p(k)\{\log[p(k)]^n + \log p(k)\} \\ &= -\sum_{k\in\mathbb{N}} p(k)\log[p(k)]^n - \sum_{k\in\mathbb{N}} p(k)p(k)\log p(k) \\ &= -nh(p) - h(p) \\ h(p^{n+1}) &= -(n+1)h(p). \end{aligned}$$

Thereby concluding by induction that $h(p^n) = nh(n)\ \forall n \geq 1$. For $h(p^{n/m})$, we find that

$$\begin{aligned} h(p^{n/m}) &= -\sum_{k\in\mathbb{N}} p(k)\log[p(k)]^{n/m} \\ &= -\sum_{k\in\mathbb{N}} p(k)\log\{([p(k)]^n)^{1/m}\} \\ &= -\frac{1}{m}\sum_{k\in\mathbb{N}} p(k)\log\{[p(k)]^n\} \\ h(p^{n/m}) &= -\frac{n}{m}\sum_{k\in\mathbb{N}} p(k)\log\{p(k)\}. \end{aligned}$$

By continuity, we can conclude that $h(p^x) = xh(p)$ for all $x > 0$. We therefore obtain

$$\begin{aligned} h(p^x) &= xh(p) \\ h(e^{x\log p}) &= xh(p). \end{aligned}$$

Differentiating both sides with respect to x, we obtain

$$\begin{aligned} h(e^{x\log p})e^{x\log p}\log p &= h(p) \\ \frac{\mathrm{d}h(p^x)}{\mathrm{d}p}p^x\log p &= h(p). \end{aligned}$$

As the above relation is valid for all $x > 0$, we may apply $x \to 0$ on both sides. Assuming the limit $\lim_{x\to 0} {}^{\mathrm{d}h(p^x)}/_{\mathrm{d}p}p^x$ exists, note that the right hand side is constant with respect

to x, yielding the following

$$\begin{aligned}\lim_{x\to 0}\frac{\mathrm{d}h(p^x)}{\mathrm{d}p}p^x\log p &= \lim_{x\to 0}h(p)\\ \log p\lim_{x\to 0}\frac{\mathrm{d}h(p^x)}{\mathrm{d}p}p^x &= h(p)\\ h(p) &\propto \log p.\end{aligned}$$

Exercise 1.29

Let X be a M-state discrete random variable. As $\log(x)$ is a concave function with respect to x, and $\sum_{i=1}^{M} p(x_i) = 1$ with $p(x_i) \geq 0, \forall i \in \{1, \dots, M\}$, we write the following

$$\begin{aligned}
\mathrm{H}[X] &= -\sum_{i=1}^{M} p(x_i) \log p(x_i) \qquad \text{(Apply (1.98))} \\
&= \sum_{i=1}^{M} p(x_i) \log[1/p(x_i)] \\
&\leq \log\left(\sum_{i=1}^{M} \frac{p(x_i)}{p(x_i)}\right) \qquad \text{(Apply (1.115))} \\
\mathrm{H}[X] &\leq \log M.
\end{aligned}$$

Exercise 1.30

The Kullback-Leibler divergence between two normal density functions, with mean $\mu, m \in \mathbb{R}$ and variance $\sigma^2, s^2 > 0$ respectively is computed as follows

$$\begin{aligned}
\mathrm{KL}(p||q) &= -\int_{\mathbb{R}} p(x|\mu,\sigma^2)\log\left\{\frac{p(x|m,s^2)}{p(x|\mu,\sigma^2)}\right\}\mathrm{d}x && \text{(Apply (1.113))}\\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left[-\frac{1}{2}\log s^2+\frac{1}{2}\log\sigma^2+\right.\\
&\left.-\frac{1}{2s^2}(x-m)^2+\frac{1}{2\sigma^2}(x-\mu)^2\right]\mathrm{d}x && \text{(Apply (1.30))}\\
&= \frac{1}{2s^2}\int_{\mathbb{R}}\frac{(x-m)^2}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x+\\
&-\frac{1}{2\sigma^2}\int_{\mathbb{R}}\frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x+\frac{1}{2}\log\frac{\sigma^2}{s^2}\\
&= \frac{1}{2s^2}\int_{\mathbb{R}}\frac{x^2-2xm+m^2}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x+\\
&-\frac{1}{2}+\frac{1}{2}\log\frac{\sigma^2}{s^2}\\
&= \frac{1}{2}\left[\frac{\mu^2+\sigma^2-2\mu m+m^2-s^2}{s^2}+\log\frac{\sigma^2}{s^2}\right]\\
\mathrm{KL}(p||q) &= \frac{1}{2}\left[\frac{(\mu-m)^2+(\sigma-s)(\sigma+s)}{s^2}+\log\frac{\sigma^2}{s^2}\right].
\end{aligned}$$

Exercise 1.31

Let $\mathbf{X}$ and $\mathbf{Y}$ be a pair of continuous random variables with joint density function $p(\mathbf{x}, \mathbf{y})$, the differential entropy associated with this pair is given in (1.112). Since $\mathrm{I}[\mathbf{X}, \mathbf{Y}] = \mathrm{H}[\mathbf{Y}] - \mathrm{H}[\mathbf{Y}|\mathbf{X}] \geq 0$ - from (1.121) -, we can infer that $\mathrm{H}[\mathbf{Y}] \geq \mathrm{H}[\mathbf{Y}|\mathbf{X}]$, which implies

$$\mathrm{H}[\mathbf{X}, \mathbf{Y}] \leq \mathrm{H}[\mathbf{Y}] + \mathrm{H}[\mathbf{X}]. \tag{1.17}$$

Assuming $\mathbf{X}$ and $\mathbf{Y}$ are independent, $\mathrm{I}[\mathbf{X}, \mathbf{Y}] = 0$, and consequently $\mathrm{H}[\mathbf{Y}|\mathbf{X}] = \mathrm{H}[\mathbf{Y}]$. Applying this result in (1.112) we obtain (1.17). On the other hand, assuming the strict equality holds in (1.17), it holds that

$$\begin{aligned} \mathrm{H}[\mathbf{X}] + \mathrm{H}[\mathbf{Y}] &= \mathrm{H}[\mathbf{Y}|\mathbf{X}] + \mathrm{H}[\mathbf{X}] \\ \mathrm{H}[\mathbf{Y}] &= \mathrm{H}[\mathbf{Y}|\mathbf{X}]. \end{aligned}$$

It follows that $\mathrm{I}[\mathbf{X}, \mathbf{Y}] = 0$, and consequently $\mathrm{KL}(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y})) = 0$ - from (1.120) -, which occurs if, and only if, $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$, i.e., $\mathbf{X}$ and $\mathbf{Y}$ must be independent.

Exercise 1.32

Let $\mathbf{X}$ be a continuous random vector, and $\mathbf{A}$ a nonsingular matrix such that we define $\mathbf{Y} = \mathbf{AX}$ (consequently $\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y}$), if the density function associated to $\mathbf{X}$ is $p_{\mathbf{X}}(x)$, it follows from (1.27) that the density associated with $\mathbf{Y}$ is given by

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})\left|\frac{\mathrm{d}\mathbf{A}^{-1}\mathbf{y}}{\mathrm{d}\mathbf{y}}\right| \\ &= p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})|(\mathbf{A}^{-1})^{\top}| && \text{(Apply (C.19))} \\ &= p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})|\mathbf{A}^{-1}| && \text{(Apply } |\mathbf{A}^{\top}| = |\mathbf{A}|\text{)} \\ p_{\mathbf{Y}}(\mathbf{y}) &= \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} && \text{(Apply (C.13)).} \end{aligned} \tag{1.18}$$

It follows from (1.104) that the differential entropy associated with $\mathbf{Y}$ would be

$$\begin{aligned} \mathrm{H}[\mathbf{Y}] &= -\int p_{\mathbf{Y}}(\mathbf{y}) \log p_{\mathbf{Y}}(\mathbf{y})\, \mathrm{d}\mathbf{y} \\ &= -\int \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} \log \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|}\, \mathrm{d}\mathbf{y} && \text{(Apply (1.18))} \\ &= -\int \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} \log p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})\, \mathrm{d}\mathbf{y} + \int \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} \log|\mathbf{A}|\, \mathrm{d}\mathbf{y} \\ &= -\int p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int p_{\mathbf{X}}(\mathbf{x}) \log|\mathbf{A}|\, \mathrm{d}\mathbf{x} \\ \mathrm{H}[\mathbf{Y}] &= \mathrm{H}[\mathbf{X}] + \log|\mathbf{A}| && \text{(Apply (1.104)).} \end{aligned}$$

Exercise 1.33

Let X and Y be discrete random variables whose conditional entropy is $\mathrm{H}[Y|X] = 0$. It follows from (1.111) that

$$\begin{aligned} \mathrm{H}[Y|X] &= 0 \\ -\sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} p(x_i, y_j)\log p(y_j|x_i) &= 0 \\ -\sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} p(y_j|x_i)p(x_i)\log p(y_j|x_i) &= 0 \qquad \text{(Apply (1.32))} \\ \sum_{i\in\mathbb{N}} p(x_i)\left[\sum_{j\in\mathbb{N}} p(y_j|x_i)\log p(y_j|x_i)\right] &= 0. \end{aligned} \tag{1.19}$$

In order for the above equation to equal zero, for all $p(x_i) > 0$ it must follow that $\sum_{j\in\mathbb{N}} p(y_j|x_i)\log p(y_j|x_i) = 0$. Note that all terms within this sum are non-positive, thus this must imply that $p(y_j|x_i)\log p(y_j|x_i) = 0, \forall j \in \mathbb{N}$. This may occur if $p(y_j|x_i) \in \{0, 1\}$. As $p(y|x_i)$ is a probability function, it must be normalized so too that $\sum_{j\in\mathbb{N}} p(y_j|x_i) = 1$, whilst constrained to $p(y_j|x_i) \geq 0, \forall j \in \mathbb{N}$. This implies that only one y_j may yield unit probability. Therefore, in order for the relation (1.19) to be valid, there must be one, and strictly one, y_j such that $p(y_j|x_i) \neq 0$, i.e. that $p(y_j|x_i) = 1$.

Exercise 1.34

We seek the density function p which solves the following optimization problem

$$p = \begin{cases} \max -\int_{-\infty}^{\infty} p(x) \log p(x)\, \mathrm{d}x, \\ \text{contrained to} \begin{cases} \int_{-\infty}^{\infty} p(x)\, \mathrm{d}x = 1, \\ \int_{-\infty}^{\infty} xp(x)\, \mathrm{d}x = \mu, \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, \mathrm{d}x = \sigma^2. \end{cases} \end{cases}$$

Which may be solved by maximizing the related Lagrangian, as defined in (E.4), given as follows

$$\begin{aligned} g(p) = &-\int_{-\infty}^{\infty} p(x) \log p(x)\, \mathrm{d}x + \lambda_1 \left(\int_{-\infty}^{\infty} p(x)\, \mathrm{d}x - 1 \right) + \\ &+ \lambda_2 \left(\int_{-\infty}^{\infty} xp(x)\, \mathrm{d}x - \mu \right) + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, \mathrm{d}x - \sigma^2 \right). \end{aligned}$$

We differentiate $g(p)$ with respect to p, obtaining the following

$$\frac{\mathrm{d}g(p)}{\mathrm{d}p} = -\log p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2.$$

Solving for ${}^{\mathrm{d}g(p)}/_{\mathrm{d}p} = 0$, we find that

$$\begin{aligned} -\log p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 &= 0 \\ \log p(x) &= -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 \\ p(x) &= \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\}. \end{aligned} \tag{1.20}$$

Substituting into the first constraint, we find

$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x \\
1 &= \int_{-\infty}^{\infty} \exp\{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2\}\,\mathrm{d}x \\
\exp\{1\} &= \exp\{\lambda_1\}\int_{-\infty}^{\infty} \exp\{\lambda_2 x+\lambda_3 x^2-2\lambda_3\mu x+\lambda_3\mu^2\}\,\mathrm{d}x \\
\exp\{1-\lambda_3\mu^2\} &= \exp\{\lambda_1\}\int_{-\infty}^{\infty} \exp\left\{\lambda_3 x^2-2\left(\lambda_3\mu-\frac{\lambda_2}{2}\right)x\right\}\mathrm{d}x \\
\exp\{1-\lambda_3\mu^2\} &= \exp\{\lambda_1\}\int_{-\infty}^{\infty} \exp\left\{\lambda_3\left[x^2-2\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)x\right]\right\}\mathrm{d}x \\
\exp\left\{1-\lambda_3\mu^2+\lambda_3\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)^2\right\} &= \exp\{\lambda_1\}\int_{-\infty}^{\infty} \exp\left\{\lambda_3\left(x-\mu+\frac{\lambda_2}{2\lambda_3}\right)^2\right\}\mathrm{d}x \\
\exp\left\{1-\lambda_3\left[\mu^2-\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)^2\right]\right\} &= \exp\{\lambda_1\}\int_{-\infty}^{\infty} \exp\left\{\lambda_3 y^2\right\}\mathrm{d}y \\
\exp\left\{1-\lambda_3\left[\frac{\mu\lambda_2}{\lambda_3}-\frac{\lambda_2^2}{4\lambda_3^2}\right]\right\} &= \exp\{\lambda_1\}\int_{-\infty}^{\infty} \exp\left\{-(-\lambda_3)y^2\right\}\mathrm{d}y \\
\exp\left\{1-\mu\lambda_2+\frac{\lambda_2^2}{4\lambda_3}\right\} &= \exp\{\lambda_1\}\sqrt{-\frac{\pi}{\lambda_3}} \\
\sqrt{-\frac{\lambda_3}{\pi}}\exp\left\{1-\mu\lambda_2+\frac{\lambda_2^2}{4\lambda_3}\right\} &= \exp\{\lambda_1\} \\
\frac{1}{2}\log\left(-\frac{\lambda_3}{\pi}\right)+1-\mu\lambda_2+\frac{\lambda_2^2}{4\lambda_3} &= \lambda_1.
\end{aligned}
$$

Substituting into the second constraint, we find

$$\mu = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x$$
$$\mu = \int_{-\infty}^{\infty} x \exp\{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2\}\,\mathrm{d}x$$
$$\mu = \int_{-\infty}^{\infty} x \exp\left\{-1+\frac{1}{2}\log\left(-\frac{\lambda_3}{\pi}\right)+1-\mu\lambda_2+\frac{\lambda_2^2}{4\lambda_3}+\lambda_2 x+\lambda_3(x-\mu)^2\right\}\mathrm{d}x$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\int_{-\infty}^{\infty} x\sqrt{-\frac{\lambda_3}{\pi}}\exp\left\{x\lambda_2+\lambda_3 x^2-2\lambda_3\mu x+\lambda_3\mu^2\right\}\mathrm{d}x$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\exp\{\lambda_3\mu^2\}\times$$
$$\times\int_{-\infty}^{\infty} x\sqrt{-\frac{\lambda_3}{\pi}}\exp\left\{\lambda_3 x^2-2\left[\lambda_3\mu-\frac{\lambda_2}{2}\right]x\right\}\mathrm{d}x$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\exp\{\lambda_3\mu^2\}\times$$
$$\times\int_{-\infty}^{\infty} x\sqrt{-\frac{\lambda_3}{\pi}}\exp\left\{\lambda_3\left(x^2-2\left[\mu-\frac{\lambda_2}{2\lambda_3}\right]x\right)\right\}\mathrm{d}x$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\exp\{\lambda_3\mu^2\}\exp\left\{-\lambda_3\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)^2\right\}\times$$
$$\times\int_{-\infty}^{\infty} x\sqrt{-\frac{\lambda_3}{\pi}}\exp\left\{-(-\lambda_3)\left(x-\mu+\frac{\lambda_2}{2\lambda_3}\right)^2\right\}\mathrm{d}x$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\exp\left\{\lambda_3\left[\mu^2-\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)^2\right]\right\}\left[\mu-\frac{\lambda_2}{2\lambda_3}\right]$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\exp\left\{\lambda_3\left[2\mu\frac{\lambda_2}{2\lambda_3}-\frac{\lambda_2^2}{4\lambda_3^2}\right]\right\}\left[\mu-\frac{\lambda_2}{2\lambda_3}\right]$$
$$\mu = \exp\left\{\left(\frac{\lambda_2}{4\lambda_3}-\mu\right)\lambda_2\right\}\exp\left\{\lambda_2\left(\mu-\frac{\lambda_2}{4\lambda_3}\right)\right]\right\}\left[\mu-\frac{\lambda_2}{2\lambda_3}\right]$$
$$\mu = \mu-\frac{\lambda_2}{2\lambda_3}$$
$$0 = \lambda_2.$$

Finally, substituting into the third constraint, we find

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x \\
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \exp\{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2\}\,\mathrm{d}x \\
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left\{-1+\frac{1}{2}\log\left(-\frac{\lambda_3}{\pi}\right)+1-\mu\lambda_2+\frac{\lambda_2^2}{4\lambda_3}+\lambda_2 x+\lambda_3(x-\mu)^2\right\}\mathrm{d}x \\
\sigma^2 &= \sqrt{-\frac{\lambda_3}{\pi}}\sqrt{-\frac{2\pi}{2\lambda_3}}\int_{-2\infty}^{\infty}\sqrt{-\frac{2\lambda_3}{2\pi}}(x-\mu)^2\exp\left\{-(-2\lambda_3)\frac{(x-\mu)^2}{2}\right\}\mathrm{d}x \\
\sigma^2 &= -\frac{1}{2\lambda_3} \\
-\frac{1}{2\sigma^2} &= \lambda_3.
\end{aligned}$$

We conclude that the Lagrange multipliers are

$$\begin{aligned}
\lambda_1 &= \frac{1}{2}\log\left(\frac{1}{2\pi\sigma^2}\right)+1, \\
\lambda_2 &= 0, \\
\lambda_3 &= -\frac{1}{2\sigma^2}.
\end{aligned}$$

Substituting these values in (1.20), we find that p must be the normal density function, with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.

Exercise 1.35

Let X denote a normal random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, its differential entropy is computed as

$$\begin{aligned}
\mathrm{H}[X] &= -\int_{\mathbb{R}} p(x|\mu,\sigma^2)\log p(x|\mu,\sigma^2)\,\mathrm{d}x && \text{(Apply (1.104))} \\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2}\right\}\times && (1.21) \\
&\quad\times\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right]\mathrm{d}x && \text{(Apply (1.46))} \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{V}\mathrm{ar}[X] \\
\mathrm{H}[X] &= \frac{1}{2}\left\{1+\log(2\pi\sigma^2)\right\}. && (1.22)
\end{aligned}$$

We conclude that the differential entropy of a normal random variable is as above.

Exercise 1.36

For a strictly convex function $f(x)$, for any $\lambda \in (0,1)$, it follows from (1.114) that, choosing $a = x - h$, $\lambda = 1/2$ and $b = x + h$, for $x \in \mathbb{R}$ and $h > 0$, we write

$$\begin{aligned} f\left(\frac{1}{2}(x-h) + \frac{1}{2}(x+h)\right) &< \frac{1}{2}f(x-h) + \frac{1}{2}f(x+h) \\ f(x) &< \frac{1}{2}f(x-h) + \frac{1}{2}f(x+h) \\ 0 &< \frac{1}{2}f(x-h) - f(x) + \frac{1}{2}f(x+h) \\ 0 &< f(x-h) - 2f(x) + f(x+h) \\ 0 &< \frac{f(x-h) - 2f(x) + f(x+h)}{2h}. \end{aligned}$$

Applying the limit as $h \to 0$ on both sides, we find

$$0 < \frac{\mathrm{d}^2 f(x)}{\mathrm{d}x}.$$

Exercise 1.37

Let $\mathbf{X}$ and $\mathbf{Y}$ be continuous random variables, we desire to demonstrate that the conditional entropy satisfies $\mathrm{H}[\mathbf{X},\mathbf{Y}] = \mathrm{H}[\mathbf{Y}|\mathbf{X}] + \mathrm{H}[\mathbf{X}]$. We therefore write the differential entropy associated with $(\mathbf{X},\mathbf{Y})$ as

$$\begin{aligned}
\mathrm{H}[\mathbf{X},\mathbf{Y}] &= -\int\!\!\int p(\mathbf{x},\mathbf{y}) \log p(\mathbf{y},\mathbf{x})\, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= -\int\!\!\int p(\mathbf{x},\mathbf{y}) \log p(\mathbf{y}|\mathbf{x}) p(\mathbf{x})\, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} && \text{(Apply (1.32))} \\
&= -\int\!\!\int p(\mathbf{x},\mathbf{y}) \log p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} - \int\!\!\int p(\mathbf{x},\mathbf{y}) \log p(\mathbf{x})\, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= -\int\!\!\int p(\mathbf{x},\mathbf{y}) \log p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} - \int \log p(\mathbf{x}) \left[\int p(\mathbf{x},\mathbf{y})\, \mathrm{d}\mathbf{y}\right] \mathrm{d}\mathbf{x} \\
&= \mathrm{H}[\mathbf{Y}|\mathbf{X}] - \int p(\mathbf{x}) \log p(\mathbf{x})\, \mathrm{d}\mathbf{x} && \text{(Apply (1.31))} \\
\mathrm{H}[\mathbf{X},\mathbf{Y}] &= \mathrm{H}[\mathbf{Y}|\mathbf{X}] + \mathrm{H}[\mathbf{X}] && \text{(Apply (1.104)),}
\end{aligned}$$

wherein we utilized also (1.111). Thereby, we conclude our demonstration.

Exercise 1.38

We seek to, utilizing proof by induction, demonstrate that, for a convex function $f(x)$, if (1.114) holds for all $a, b \in \mathbb{R}$, we must therefore be able to extend this to a sequence $\{\lambda_i\}_{i=1}^M$, wherein $\sum_{i=1}^M \lambda_i = 1$, such that (1.115) holds. The result trivially holds to $M = 1$, once it implies $\lambda_1 = 1$ and

$$f(\lambda_1 a) \leq \lambda_1 f(a)$$
$$f(a) \leq f(a).$$

We choose now $\lambda_{M+1} \in [0, 1)$, and take x_0 as the following

$$x_0 = \frac{\sum_{i=1}^M \lambda_i x_i}{1 - \lambda_{M+1}},$$

We also take x_{M+1} as any arbitrary point. It follows, from the property of convexity, that

$$f(\lambda_{M+1} x_{M+1} + (1 - \lambda_M) x_0) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_M) f(x_0).$$

We now assume it holds for M, and verify if this implies it holds for $M + 1$

$$\begin{aligned}
f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) &= f\left(\lambda_{M+1} x_{M+1} + \sum_{i=1}^{M} \lambda_i x_i\right) \\
&= f(\lambda_{M+1} x_{M+1} + (1 - \lambda_{M+1}) x_0) \\
&\leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f(x_0) && \text{(Apply (1.114))} \\
&= \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right) \\
&\leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i) && \text{(By assumption)} \\
&= \lambda_{M+1} f(x_{M+1}) + \sum_{i=1}^{M} \lambda_i f(x_i) \\
f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) &\leq \sum_{i=1}^{M+1} \lambda_i f(x_i).
\end{aligned}$$

We therefore conclude by induction that the extension is valid.

Exercise 1.39

We consider the joint distribution presented in Table 1.1 for the computation of several forms of entropy. First, we determine the distribution of each individual variable via the sum probability rule in (1.10):

$$\begin{aligned}\mathbb{P}(X=0) &= \mathbb{P}(X=0,Y=0)+\mathbb{P}(X=0,Y=1)\\ &= \frac{1}{3}+\frac{1}{3}\\ &= \frac{2}{3}.\end{aligned}$$

Consequently $\mathbb{P}(X=1)=1/3$. For Y

$$\begin{aligned}\mathbb{P}(Y=0) &= \mathbb{P}(X=0,Y=0)+\mathbb{P}(X=1,Y=0)\\ &= \frac{1}{3}+0\\ &= \frac{1}{3}.\end{aligned}$$

Consequently $\mathbb{P}(Y=1)=2/3$. We compute $\mathrm{H}[X]$ as

$$\begin{aligned}\mathrm{H}[X] &= -p_X(0)\log p_X(0)-\log p_X(1)\log p_X(1)\\ &= -\frac{2}{3}\log\frac{2}{3}-\frac{1}{3}\log\frac{1}{3}\\ &= -\frac{2}{3}\log 2+\frac{2}{3}\log 3-\frac{1}{3}\log 1+\frac{1}{3}\log 3\\ &= -\frac{2}{3}\log 2+\frac{2}{3}\log 3-\frac{1}{3}\log 1+\frac{1}{3}\log 3\\ &= \log 3-\frac{2}{3}\log 2.\end{aligned}$$

We compute $\mathrm{H}[Y]$ as

$$\begin{aligned}\mathrm{H}[Y] &= -p_Y(0)\log p_Y(0)-\log p_Y(1)\log p_Y(1)\\ &= -\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3}\\ &= -\frac{1}{3}\log 1+\frac{1}{3}\log 3-\frac{2}{3}\log 2+\frac{2}{3}\log 3\\ &= \log 3-\frac{2}{3}\log 2.\end{aligned}$$

We compute $\mathrm{H}[X,Y]$ as

$$\begin{aligned}\mathrm{H}[X,Y] &= -p(0,0)\log p(0,0)-\log p(0,1)\log p(0,1)\\ &\quad -p(1,0)\log p(1,0)-\log p(1,1)\log p(1,1)\\ &= -\frac{1}{3}\log\frac{1}{3}-\frac{1}{3}\log\frac{1}{3}-0-\frac{1}{3}\log\frac{1}{3}\\ &= \log 3.\end{aligned}$$

We can therefore compute $H[X|Y]$ as

$$\begin{aligned}\mathrm{H}[X|Y] &= \mathrm{H}[X,Y] - \mathrm{H}[Y] \\ &= \log 3 - \log 3 + \frac{2}{3}\log 2 \\ &= \frac{2}{3}\log 2,\end{aligned}$$

and $\mathrm{H}[Y|X]$

$$\begin{aligned}\mathrm{H}[Y|X] &= \mathrm{H}[X,Y] - \mathrm{H}[X] \\ &= \log 3 - \log 3 + \frac{2}{3}\log 2 \\ &= \frac{2}{3}\log 2.\end{aligned}$$

Lastly, the mutual information is

$$\begin{aligned}\mathrm{I}[X,Y] &= \mathrm{H}[Y] - \mathrm{H}[Y|X] \\ &= \log 3 - \frac{2}{3}\log 2 - \frac{2}{3}\log 2 \\ &= \log 3 - \frac{4}{3}\log 2.\end{aligned}$$

Table 1.1: Joint distribution for binary random variables (X, Y) utilized in Exercise 1.39.

		Y	
		0	1
X	0	1/3	1/3
	1	0	1/3

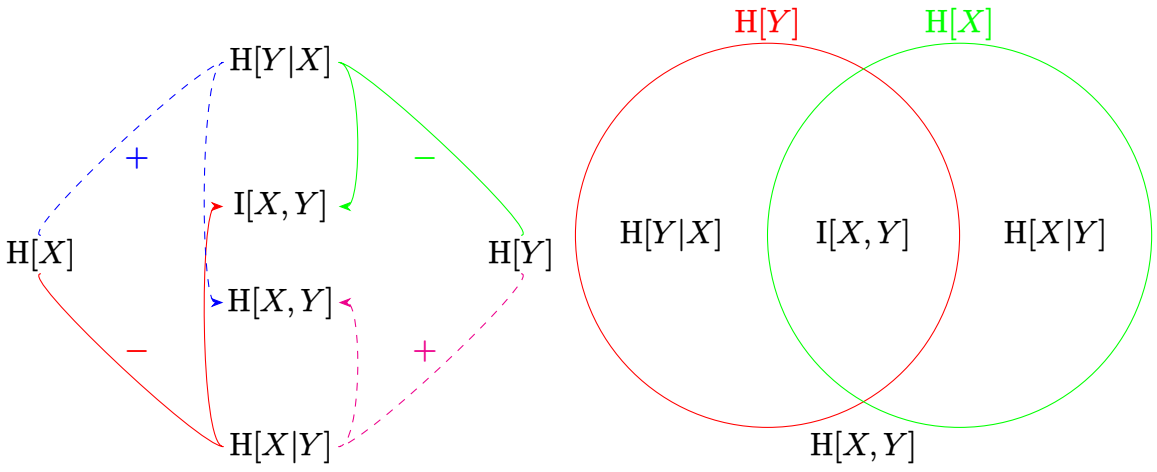


Figure 1.2: Diagrams representing the relationship between different forms of entropy.

Exercise 1.40

Let $\{a_i\}_{i=1}^M$ be a sequence of positive real values, it follows that its geometric mean is computed as

$$\begin{aligned}\left(\prod_{i=1}^M a_i\right)^{1/M} &= e^{\log(\prod_{i=1}^M a_i)^{1/M}} \\ &= e^{\sum_{i=1}^M \frac{1}{M}\log a_i}.\end{aligned}$$

We note herein that, from the concavity of $f(t) = \log t$, it follows by applying (1.115) that

$$\sum_{i=1}^M \frac{1}{M}\log a_i \leq \log\left(\sum_{i=1}^M \frac{1}{M} a_i\right). \tag{1.23}$$

Note that, as $g(t) = e^t$ is a monotonic function, it possesses the property that if $s \leq t$, then $g(s) \leq g(t)$. This, joined with the result in (1.23), implies that

$$\begin{aligned}\left(\prod_{i=1}^M a_i\right)^{1/M} &= e^{\sum_{i=1}^M \frac{1}{M}\log a_i} \\ &\leq e^{\log(\sum_{i=1}^M \frac{1}{M} a_i)} \\ &= \sum_{i=1}^M \frac{1}{M} a_i.\end{aligned}$$

Hence, we conclude that the arithmetic mean is greater than or equal to the geometric mean.

Exercise 1.41

Let $\mathbf{X}$ and $\mathbf{Y}$ be continuous random variables with joint density function denoted by $p(\mathbf{x}, \mathbf{y})$. It follows from (1.120) that the mutual information $\mathrm{I}[\mathbf{X}, \mathbf{Y}]$ is computed as

$$
\begin{aligned}
\mathrm{I}[\mathbf{X}, \mathbf{Y}] &= \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) || p(\mathbf{x})p(\mathbf{y})) \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && \text{(Apply (1.113))} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} && \text{(Apply (1.32))} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})} \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= -\iint p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + \iint p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= -\iint p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} - \mathrm{H}[\mathbf{X}|\mathbf{Y}] && \text{(Apply (1.111))} \\
&= -\int p(\mathbf{y}|\mathbf{x}) \left[\int p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right] \mathrm{d}\mathbf{y} - \mathrm{H}[\mathbf{X}|\mathbf{Y}] \\
&= \mathrm{H}[\mathbf{X}] \int p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} - \mathrm{H}[\mathbf{X}|\mathbf{Y}] && \text{(Apply (1.104))} \\
\mathrm{I}[\mathbf{X}, \mathbf{Y}] &= \mathrm{H}[\mathbf{X}] - \mathrm{H}[\mathbf{X}|\mathbf{Y}] && \text{(Apply (1.30))}.
\end{aligned}
$$

Wherein the demonstration for $\mathrm{H}[\mathbf{Y}] - \mathrm{H}[\mathbf{Y}|\mathbf{X}]$ follows analogously.

Chapter 2
Probability Distributions

Exercise 2.1

Let X be a Bernoulli distributed random variable with parameter $\mu \in [0,1]$, with respective probability function as in (2.2). It follows that

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

I.e., the distribution is normalized. Its expected value is computed as

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^{1} x \cdot p(x|\mu) && \text{(Apply (1.33))} \\ &= 0 \cdot p(0|\mu) + 1 \cdot p(1|\mu) \\ &= 0 \cdot (1-\mu) + 1 \cdot \mu \\ \mathbb{E}[X] &= \mu. \qquad (2.1)\end{aligned}$$

Its variance is computed as

$$\begin{aligned}\mathbb{V}\text{ar}[X] &= \sum_{x=0}^{1} (x - \mathbb{E}[X])^2 \cdot p(x|\mu) && \text{(Apply (1.38))} \\ &= \sum_{x=0}^{1} (x-\mu)^2 \cdot p(x|\mu) && \text{(Apply (2.1))} \\ &= (0-\mu)^2 \cdot p(0|\mu) + (1-\mu)^2 \cdot p(1|\mu) \\ &= \mu^2 \cdot (1-\mu) + (1-\mu)^2 \cdot \mu \\ &= \mu(1-\mu)(\mu + 1 - \mu) \\ \mathbb{V}\text{ar}[X] &= \mu(1-\mu).\end{aligned}$$

Lastly, the entropy associated with X is

$$\begin{aligned}\mathrm{H}[X] &= -\sum_{x=0}^{1} p(x|\mu) \log p(x|\mu) && \text{(Apply (1.98))} \\ &= -p(0|\mu) \log p(0|\mu) - p(1|\mu) \log p(1|\mu) \\ \mathrm{H}[X] &= -(1-\mu)\log(1-\mu) - \mu \log \mu.\end{aligned}$$

Exercise 2.2

Let X be a random variable whose probability function is defined as in (2.261), where $\mu \in [-1, 1]$. We first desire to prove it is normalized:

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

We now compute its expected value as

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\{-1,1\}} x\cdot p(x|\mu) && \text{(Apply (1.33))}\\ &= -1\cdot p(-1|\mu) + 1\cdot p(1|\mu)\\ &= -\frac{1-\mu}{2} + \frac{1+\mu}{2}\\ \mathbb{E}[X] &= \mu. && (2.2)\end{aligned}$$

Its variance is computed as

$$\begin{aligned}\mathbb{V}\text{ar}[X] &= \sum_{x\in\{-1,1\}} (x-\mathbb{E}[X])^2\cdot p(x|\mu) && \text{(Apply (1.38))}\\ &= \sum_{x\in\{-1,1\}} (x-\mu)^2\cdot p(x|\mu) && \text{(Apply (2.2))}\\ &= (-1-\mu)^2\cdot p(-1|\mu) + (1-\mu)^2\cdot p(1|\mu)\\ &= (1+\mu)\cdot\frac{1-\mu}{2} + (1-\mu)\cdot\frac{1+\mu}{2}\\ &= (1+\mu)(1-\mu)\\ \mathbb{V}\text{ar}[X] &= 1-\mu^2.\end{aligned}$$

Lastly, the entropy associated with X is

$$\begin{aligned}\text{H}[X] &= -\sum_{x\in\{-1,1\}} p(x|\mu)\log p(x|\mu) && \text{(Apply (1.98))}\\ &= -p(-1|\mu)\log p(-1|\mu) - p(1|\mu)\log p(1|\mu)\\ &= -\frac{(1-\mu)}{2}\log\frac{(1-\mu)}{2} - \frac{(1+\mu)}{2}\log\frac{(1+\mu)}{2}\\ &= -\frac{(1-\mu)}{2}\log(1-\mu) + \frac{(1-\mu)}{2}\log 2 +\\ &+ \frac{(1+\mu)}{2}\log 2 - \frac{(1+\mu)}{2}\log(1+\mu)\\ \text{H}[X] &= -\frac{1}{2}\Big[(1-\mu)\log(1-\mu) + (1+\mu)\log(1+\mu) - 2\log 2\Big].\end{aligned}$$

Exercise 2.3

First, we desire to prove (2.262). This may be performed as follows

$$\begin{aligned}\binom{N}{m}+\binom{N}{m-1} &= \frac{N!}{m!(N-m)!}+\frac{N!}{(m-1)!(N-m+1)} \\ &= \frac{N!}{m!(N-m+1)!}\Big[N-m+1+m\Big] \\ &= \frac{(N+1)!}{m!(N+1-m)!} \\ \binom{N}{m}+\binom{N}{m-1} &= \binom{N+1}{m}.\end{aligned}$$

We now desire to prove by induction (2.263). Trivially, the result for $N=1$ holds, as

$$\begin{aligned}1+x &= \sum_{m=0}^{1}\binom{1}{0}x^m \\ 1+x &= 1+x.\end{aligned}$$

We now assume the result for an arbitrary N, and desire to show it that implies it holds for $N+1$. See that

$$\begin{aligned}(1+x)^{N+1} &= (1+x)(1+x)^N \\ &= (1+x)\sum_{m=0}^{N}\binom{N}{m}x^m && \text{(By assumption)} \\ &= \sum_{m=0}^{N}\binom{N}{m}x^m+\sum_{m=0}^{N}\binom{N}{m}x^{m+1} \\ &= \binom{N}{0}+\sum_{m=1}^{N}\binom{N}{m}x^m+\sum_{m=1}^{N+1}\binom{N}{m-1}x^m \\ &= \binom{N+1}{0}+\sum_{m=1}^{N}\left[\binom{N}{m}+\binom{N}{m-1}\right]x^m+\binom{N}{N}x^{N+1} \\ &= \binom{N+1}{0}+\sum_{m=1}^{N}\binom{N+1}{m}x^m+\binom{N+1}{N+1}x^{N+1} \\ (1+x)^{N+1} &= \sum_{m=0}^{N+1}\binom{N+1}{m}x^m.\end{aligned}$$

We conclude that the relation in (2.263) holds for all $n \geq 1$. We now seek to demonstrate (2.264): see that

$$\begin{aligned}
\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} &= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \left(\frac{\mu}{1-\mu}\right)^m \quad \text{(Apply (2.263))} \\
&= (1-\mu)^N \left(1 + \frac{\mu}{1-\mu}\right)^N \\
&= (1-\mu)^N \left(\frac{1-\mu+\mu}{1-\mu}\right)^N \\
\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} &= 1.
\end{aligned}$$

Exercise 2.4

Let X be a Binomial random variable with parameters $N \geq 1$ and $\mu \in [0,1]$, its expected value may be determined by differentiating both sides of (2.264) with respect to μ, as follows

$$
\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\right] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}m\mu^{m-1}(1-\mu)^{N-m} + \\
&\quad - \sum_{m=0}^{N}\binom{N}{m}(N-m)\mu^m(1-\mu)^{N-m-1} \\
0 &= \sum_{m=0}^{N}\binom{N}{m}\mu^{m-1}(1-\mu)^{N-m-1}\left[m(1-\mu)-(N-m)\mu\right] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}\mu^{m-1}(1-\mu)^{N-m-1}\left[m-\mu N\right] \\
0 &= \frac{1}{\mu(1-\mu)}\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\left[m-\mu N\right] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\left[m-\mu N\right] \\
0 &= E[X] - \mu N\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} && \text{(Apply (1.33))} \\
E[X] &= \mu N && \text{(Apply (1.26)).}
\end{aligned} \tag{2.3}
$$

In order to determine its second moment we must again differentiate (2.264) with respect to μ, obtaining the following

$$
\begin{aligned}
0 &= \frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\left[\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\right] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}m(m-1)\mu^{m-2}(1-\mu)^{N-m}+ \\
&\quad -2\sum_{m=0}^{N}\binom{N}{m}m(N-m)\mu^{m-1}(1-\mu)^{N-m-1}+ \\
&\quad +\sum_{m=0}^{N}\binom{N}{m}(N-m)(N-m-1)\mu^m(1-\mu)^{N-m-2} \\
0 &= \frac{1}{\mu^2(1-\mu)^2}\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\Big[m(m-1)(1-\mu)^2+ \\
&\quad -2m(N-m)\mu(1-\mu)+(N-m)(N-m-1)\mu^2\Big] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\Big[(m^2-m)(1-2\mu+\mu^2)+ \\
&\quad +(2m^2-2mN)(\mu-\mu^2)+ \\
&\quad +(N^2-Nm-N-mN+m^2+m)\mu^2\Big] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\Big[m^2-m+ \\
&\quad +(-2m^2+2m+2m^2-2mN)\mu+ \\
&\quad +(m^2-m-2m^2+2mN+N^2+ \\
&\quad -Nm-N-mN+m^2+m)\mu^2\Big] \\
0 &= \sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\times \\
&\quad \times\Big[m^2-m-2m(N-1)\mu+N(N-1)\mu^2\Big] \\
0 &= \mathbb{E}[X^2]-\mathbb{E}[X]-2\mathbb{E}[X](N-1)\mu+N(N-1)\mu^2 \qquad \text{(Apply (1.26) and (1.33))}
\end{aligned}
$$

$$
\mathbb{E}[X^2] = N\mu(1-\mu)+N^2\mu^2. \tag{2.4}
$$

Consequently

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2]-\{\mathbb{E}[X]\}^2 && \text{(Apply (1.39))} \\
&= N\mu(1-\mu)+N^2\mu^2-N^2\mu^2 && \text{(Apply (2.3) and (2.4))} \\
&= N\mu(1-\mu).
\end{aligned}
$$

Exercise 2.5

We desire to show that

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2.5}$$

We write

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \left[\int_0^\infty \exp\{-x\}x^{a-1}\,\mathrm{d}x\right]\left[\int_0^\infty \exp\{-y\}y^{b-1}\,\mathrm{d}y\right] && \text{(Apply (1.141))}\\
&= \int_0^\infty\int_0^\infty \exp\{-(x+y)\}x^{a-1}y^{b-1}\,\mathrm{d}y\mathrm{d}x\\
&= \int_0^\infty\int_x^\infty \exp\{-t\}x^{a-1}(t-x)^{b-1}\,\mathrm{d}t\mathrm{d}x && \text{(Set } y=t-x\text{)}\\
&= \int_0^\infty\int_0^t \exp\{-t\}x^{a-1}(t-x)^{b-1}\,\mathrm{d}x\mathrm{d}t\\
&= \int_0^\infty\int_0^1 \exp\{-t\}(t\mu)^{a-1}(t-t\mu)^{b-1}t\,\mathrm{d}\mu\mathrm{d}t && \text{(Set } x=t\mu\text{)}\\
&= \int_0^\infty\int_0^1 \exp\{-t\}t^{a+b-1}\mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu\mathrm{d}t\\
&= \left[\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu\right]\left[\int_0^\infty \exp\{-t\}t^{a+b-1}\,\mathrm{d}t\right]\\
&= \left[\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu\right]\Gamma(a+b) && \text{(Apply (1.141))}\\
\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} &= \int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu.
\end{aligned}$$

Exercise 2.6

Let X be a Beta-distributed random variable, with parameters $a > 0$ and $b > 0$. It follows that the corresponding expected value is computed as

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 x p(x|a,b)\,\mathrm{d}x && \text{(Apply (1.34))}\\
&= \int_0^1 x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\,\mathrm{d}x && \text{(Apply (2.13))}\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+1-1}(1-x)^{b-1}\,\mathrm{d}x\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} && \text{(Apply (2.5))}\\
\mathbb{E}[X] &= \frac{a}{a+b}. && (2.6)
\end{aligned}$$

It follows also that

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^1 x^2 p(x|a,b)\,\mathrm{d}x && \text{(Apply (1.34))}\\
&= \int_0^1 x^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\,\mathrm{d}x && \text{(Apply (2.13))}\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+2-1}(1-x)^{b-1}\,\mathrm{d}x\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} && \text{(Apply (2.5))}\\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} && \text{(Apply (1.11))}\\
\mathbb{E}[X^2] &= \frac{(a+1)a}{(a+b+1)(a+b)},
\end{aligned}$$

Consequently, the variance of X is computed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 && \text{(Apply (1.39))}\\
&= \frac{a^2+a}{(a+b+1)(a+b)} - \frac{a^2}{(a+b)^2}\\
&= \frac{(a^2+a)(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2}\\
&= \frac{a^3+a^2b+a^2+ab-a^3-a^2b-a^2}{(a+b+1)(a+b)^2}\\
\mathbb{V}\mathrm{ar}[X] &= \frac{ab}{(a+b+1)(a+b)^2}.
\end{aligned}$$

Lastly, the mode X is computed by taking the derivative of the logarithm of the probability function with respect to x, i.e.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\log p(x|a,b) &= \frac{\mathrm{d}}{\mathrm{d}x}\Bigg[\log\Gamma(a+b)-\log\Gamma(a)-\log\Gamma(b)+\\ &\qquad +(a-1)\log x+(b-1)\log(1-x)\Bigg]\\ &= \frac{a-1}{x}-\frac{b-1}{1-x}.\end{aligned}$$

Solving ${}^{\mathrm{d}\log p(x|a,b)}/_{\mathrm{d}x}=0$, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\log p(x|a,b) &= 0\\ \frac{a-1}{x}-\frac{b-1}{1-x} &= 0\\ (1-x)(a-1)-x(b-1) &= 0\\ -x(b+a-2)+(a-1) &= 0\\ x &= \frac{a-1}{a+b-2}.\end{aligned}$$

Consequently, the maximum density location (or mode) of X is $x=(a-1)/(a+b-2)$.

Exercise 2.7

Let $X|\Theta = \theta$ be a Binomial distributed random variable with parameters $N \geq 1$ and $\theta \in [0,1]$, and let Θ be a Beta distributed random variable with parameters $a > 0$ and $b > 0$. The distribution of $\Theta|X = x$ is

$$\begin{aligned} p(\theta|a,b,x) &\propto p(x|\theta)p(\theta|a,b) \\ &= \binom{N}{x}\theta^x(1-\theta)^{N-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad \text{(Apply (2.2) and (2.13))} \\ p(\theta|a,b,x) &\propto \theta^{a+x-1}(1-\theta)^{N-x+b-1} \end{aligned}$$

It follows from the result proven in (2.6) that the mean of $\Theta|X = x$ is

$$\begin{aligned} \mathbb{E}[\Theta|X=x] &= \frac{a+x}{a+b+N} \\ &= \frac{a+b}{a+b+N}\frac{a}{a+b} + \frac{N}{a+b+N}\frac{x}{N}. \end{aligned}$$

Note that the mean of Θ is $a/(a+b)$, whilst the maximum likelihood estimator of Θ is X/N. If follows that for $\lambda = (a+b)/(a+b+N)$, consequently $1-\lambda = N/(a+b+N)$, the mean of $\Theta|X = x$ may be written as

$$\mathbb{E}[\Theta|X=x] = \lambda\mathbb{E}[\Theta] + (1-\lambda)\frac{x}{N}.$$

Where, trivially, $\lambda \in [0,1]$.

Exercise 2.8

Let $(X, Y)^\top$ be a pair of random variables with joint distribution $p(x, y)$. It follows that

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathbb{R}} xp(x)\,\mathrm{d}x && \text{(Apply (1.34))}\\
&= \int_{\mathbb{R}} x\left[\int_{\mathbb{R}} p(x,y)\,\mathrm{d}y\right]\mathrm{d}x && \text{(Apply (1.31))}\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xp(x|y)p(y)\,\mathrm{d}y\mathrm{d}x && \text{(Apply (1.32))}\\
&= \int_{\mathbb{R}}\left[\int_{\mathbb{R}} xp(x|y)\,\mathrm{d}x\right]p(y)\,\mathrm{d}y\\
&= \int_{\mathbb{R}} \mathbb{E}[X|Y=y]p(y)\,\mathrm{d}y && \text{(Apply (1.37))}\\
\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X|Y]] && \text{(Apply (1.34)).}
\end{aligned}$$

Moreover, we have that

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \int_{\mathbb{R}} (x-\mathbb{E}[X])^2 p(x)\,\mathrm{d}x && \text{(Apply (1.38))}\\
&= \int_{\mathbb{R}} (x-\mathbb{E}[X])^2\left[\int_{\mathbb{R}} p(x,y)\,\mathrm{d}y\right]\mathrm{d}x && \text{(Apply (1.31))}\\
&= \int_{\mathbb{R}} (x-\mathbb{E}[X])^2\left[\int_{\mathbb{R}} p(x|y)p(y)\,\mathrm{d}y\right]\mathrm{d}x && \text{(Apply (1.32))}\\
&= \int_{\mathbb{R}}\left[\int_{\mathbb{R}} (x-\mathbb{E}[X])^2 p(x|y)\,\mathrm{d}x\right]p(y)\,\mathrm{d}y\\
&= \int_{\mathbb{R}}\Bigg[\int_{\mathbb{R}} (x-\mathbb{E}[X|Y=y]+\\
&\quad + \mathbb{E}[X|Y=y]-\mathbb{E}[X])^2 p(x|y)\,\mathrm{d}x\Bigg]p(y)\,\mathrm{d}y\\
&= \int_{\mathbb{R}}\Bigg[\int_{\mathbb{R}} (x-\mathbb{E}[X|Y=y])^2 p(x|y)\,\mathrm{d}x\\
&\quad + 2\int_{\mathbb{R}} (x-\mathbb{E}[X|Y=y])(\mathbb{E}[X|Y=y]-\mathbb{E}[X])p(x|y)\,\mathrm{d}x\\
&\quad + (\mathbb{E}[X|Y=y]-\mathbb{E}[X])^2 p(x|y)\Bigg]p(y)\,\mathrm{d}y\\
&= \int_{\mathbb{R}} \mathbb{V}\mathrm{ar}[X|Y=y]p(y)\,\mathrm{d}y+\\
&\quad + \int_{\mathbb{R}} (\mathbb{E}[X|Y=y]-\mathbb{E}[X])^2 p(y)\,\mathrm{d}y && \text{(Apply (1.37) and (1.38))}\\
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[\mathbb{V}\mathrm{ar}[X|Y]] + \mathbb{V}\mathrm{ar}[\mathbb{E}[X|Y]] && \text{(Apply (1.34)).}
\end{aligned}$$

Exercise 2.9

We now desire to prove by induction that the Dirichlet distribution is normalized. First, consider that the one-dimensional Dirichlet distribution is the Beta distribution, as such the case for $M = 1$ dimensions has been proven prior in Exercise 2.5. We now assume that it is normalized for $M - 1$ dimensions, and seek to prove it therefore holds for M dimensions. We write the M-dimensional probability density function as in (2.272). In this context, we find that

$$\sum_{k=1}^{M-1} x_k \leq 1$$
$$x_{M-1} \leq 1 - \sum_{k=1}^{M-2} x_k.$$

We seek to integrate the probability density function with respect to x_{M-1}, as follows

$$\begin{aligned}
&\int_0^{1-\sum_{k=1}^{M-2} x_k} p_M(x_1, \ldots, x_{M-1})\, \mathrm{d}x_{M-1} \\
&= \int_0^{1-\sum_{k=1}^{M-2} x_k} C_M x_{M-1}^{\alpha_{M-1}-1} \prod_{k=1}^{M-2} x_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{M-2} x_k - x_{M-1}\right)^{\alpha_M - 1} \mathrm{d}x_{M-1}.
\end{aligned}$$

We make a change of variable $x_{M-1} = t(1-\sum_{k=1}^{M-2} x_k)$, holding $\sum_{k=1}^{M-2} x_k$ as fixed, yielding the following, in which we likewise utilize the result in (2.5):

$$\begin{aligned}
&\int_0^1 C_M t^{\alpha_{M-1}-1} \left(1 - \sum_{k=1}^{M-2} x_k\right)^{\alpha_{M-1}} (1-t)^{\alpha_M - 1} \left(1 - \sum_{k=1}^{M-2} x_k\right)^{\alpha_M - 1} \prod_{k=1}^{M-2} x_k^{\alpha_k - 1}\, \mathrm{d}t \\
&= \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} C_m \left(1 - \sum_{k=1}^{M-2} x_k\right)^{\alpha_M + \alpha_{M-1} - 1} \prod_{k=1}^{M-2} x_k^{\alpha_k - 1}.
\end{aligned}$$

The resulting density function seen above is an $(M-1)$-dimensional Dirichlet distribution, which by assumption is normalized if

$$\begin{aligned}
\frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1}+\alpha_M)} C_m &= C_{M-1} \\
&= \frac{\Gamma(\alpha_1 + \ldots + \alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdot\cdots\cdot\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1}+\alpha_M)} \\
C_M &= \frac{\Gamma(\alpha_1 + \ldots + \alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdot\cdots\cdot\Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}.
\end{aligned}$$

We therefore conclude, by induction, that the Dirichlet distribution is normalized for all dimensions $M \geq 1$.

Exercise 2.10

We now seek to demonstrate certain properties of the Dirichlet distribution. Let X be an M-dimensional Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_M$, it follows that the expected value of the j-th coordinate may be computed as

$$\begin{aligned}
\mathbb{E}[X_j] &= \int_{\mathbb{R}^M} x_j p(\mathbf{x}|\boldsymbol{\alpha}) \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (1.34))} \\
&= \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \dots \cdot \Gamma(\alpha_M)} x_j \prod_{k=1}^{M} x_k^{\alpha_k - 1} \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (2.38))} \\
&= \frac{\alpha_j}{\alpha_0} \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0 + 1)}{\Gamma(\alpha_j + 1) \prod_{\substack{k=1 \\ k \neq j}}^{M} \Gamma(\alpha_k)} x_j^{\alpha_j + 1 - 1} \prod_{\substack{k=1 \\ k \neq j}}^{M} x_k^{\alpha_k - 1} \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (1.11))}
\end{aligned}$$

$$\mathbb{E}[X_j] = \frac{\alpha_j}{\alpha_0} \qquad \text{(Apply (1.26)).} \tag{2.7}$$

The expected value of $X_j X_l$, for $j \neq l$, is given by

$$\begin{aligned}
\mathbb{E}[X_j X_l] &= \int_{\mathbb{R}^M} x_j x_l p(\mathbf{x}|\boldsymbol{\alpha}) \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (1.34))} \\
&= \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \dots \cdot \Gamma(\alpha_M)} x_j x_l \prod_{k=1}^{M} x_k^{\alpha_k - 1} \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (2.38))} \\
&= \frac{\alpha_j \alpha_k}{\alpha_0(\alpha_0 + 1)} \int_{\mathbb{R}^M} \left[\frac{\Gamma(\alpha_0 + 2)}{\Gamma(\alpha_j + 1)\Gamma(\alpha_k + 1) \prod_{\substack{k=1 \\ k \neq j \\ k \neq l}}^{M} \Gamma(\alpha_1) \dots \Gamma(\alpha_M)} \right. \\
&\quad \left. \times x_j^{\alpha_j + 1 - 1} x_l^{\alpha_l + 1 - 1} \prod_{\substack{k=1 \\ k \neq j \\ k \neq l}}^{M} x_k^{\alpha_k - 1} \right] \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (1.11))}
\end{aligned}$$

$$\mathbb{E}[X_j X_l] = \frac{\alpha_j \alpha_l}{\alpha_0(\alpha_0 + 1)} \qquad \text{(Apply (1.26)).} \tag{2.8}$$

The expected value for X_j^2 is given by

$$\begin{aligned}
\mathbb{E}[X_j^2] &= \int_{\mathbb{R}^M} x_j^2 p(\mathbf{x}|\boldsymbol{\alpha}) \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (1.34))} \\
&= \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \dots \cdot \Gamma(\alpha_M)} x_j^2 \prod_{k=1}^{M} x_k^{\alpha_k - 1} \, \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (2.38))} \\
&= \frac{\alpha_j(\alpha_j + 1)}{\alpha_0(\alpha_0 + 1)} \int_{\mathbb{R}^M} \left[\frac{\Gamma(\alpha_0 + 2)}{\Gamma(\alpha_j + 2) \prod_{\substack{k=1 \\ k \neq j}}^{M} \Gamma(\alpha_1) \dots \Gamma(\alpha_M)} \right. \\
&\quad \left. \times x_j^{\alpha_j + 2 - 1} \prod_{\substack{k=1 \\ k \neq j}}^{M} x_k^{\alpha_k - 1} \right] \mathrm{d}x_1 \dots \mathrm{d}x_M && \text{(Apply (1.11))}
\end{aligned}$$

$$\mathbb{E}[X_j^2] = \frac{\alpha_j(\alpha_j + 1)}{\alpha_0(\alpha_0 + 1)} \qquad \text{(Apply (1.26)).} \tag{2.9}$$

It follows that the variance of X_j is computed as

$$\begin{aligned}
\mathbb{V}\text{ar}[X_j] &= \mathbb{E}[X_j^2] - \{\mathbb{E}[X_j]\}^2 && \text{(Apply (1.39))} \\
&= \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha^2}{\alpha_0^2} && \text{(Apply (2.7) and (2.9))} \\
&= \frac{\alpha_0(\alpha_j^2+\alpha_j) - (\alpha_0+1)\alpha_j^2}{\alpha_0^2(\alpha_0+1)} \\
&= \frac{\alpha_0\alpha_j - \alpha_j^2}{\alpha_0^2(\alpha_0+1)} \\
\mathbb{V}\text{ar}[X_j] &= \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}.
\end{aligned}$$

Lastly, the covariance between X_j and X_l is

$$\begin{aligned}
\mathbb{C}\text{ov}[X_j, X_l] &= \mathbb{E}[X_jX_l] - \mathbb{E}[X_j]\mathbb{E}[X_l] && \text{(Apply (1.41))} \\
&= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j\alpha_l}{\alpha_0^2} && \text{(Apply (2.7) and (2.8))} \\
&= \frac{\alpha_0\alpha_j\alpha_l - (\alpha_0+1)\alpha_j\alpha_l}{\alpha_0 2(\alpha_0+1)} \\
\mathbb{C}\text{ov}[X_j, X_l] &= -\frac{\alpha_j\alpha_l}{\alpha_0 2(\alpha_0+1)}.
\end{aligned}$$

Exercise 2.11

Let X be an M-dimensional Dirichlet random variable, we seek do determine the expected value of $\mathbb{E}[\log X_j]$. To do so, we differentiate the corresponding normalizing condition with respect to α_j, obtaining the following

$$
\begin{aligned}
0 &= \int_{\mathbb{R}^M} \frac{\mathrm{d}}{\mathrm{d}\alpha_j}\left[\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdot\cdots\cdot\Gamma(\alpha_M)}\prod_{k=1}^{M} x_k^{\alpha_k-1}\right]\mathrm{d}x_1\ldots\mathrm{d}x_M \\
0 &= \int_{\mathbb{R}^M}\left[\prod_{\substack{k=1\\k\neq j}}^{M}\frac{1}{\Gamma(\alpha_k)}\right]\left[\frac{\Gamma'(\alpha_0)\Gamma(\alpha_j)-\Gamma(\alpha_0)\Gamma'(\alpha_j)}{[\Gamma(\alpha_j)]^2}x_j^{\alpha_j-1}+\right. \\
&\quad \left.+\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)}x_j^{\alpha_j-1}\log x_j\right]\mathrm{d}x_1\ldots\mathrm{d}x_M \\
0 &= \frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)}\int_{\mathbb{R}^M}\frac{\Gamma(\alpha_0)\prod_{j=1}^{M}x_j^{\alpha_j-1}}{\prod_{k=1}^{M}\Gamma(\alpha_k)}\mathrm{d}x_1\ldots\mathrm{d}x_M+ \\
&\quad -\frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)}\int_{\mathbb{R}^M}\frac{\Gamma(\alpha_0)\prod_{j=1}^{M}x_j^{\alpha_j-1}}{\prod_{k=1}^{M}\Gamma(\alpha_k)}\mathrm{d}x_1\ldots\mathrm{d}x_M+ \\
&\quad +\int_{\mathbb{R}^M}\frac{\Gamma(\alpha_0)\prod_{j=1}^{M}x_j^{\alpha_j-1}}{\prod_{k=1}^{M}\Gamma(\alpha_k)}\log x_j\,\mathrm{d}x_1\ldots\mathrm{d}x_M \\
\mathbb{E}[\log X_j] &= \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)}-\frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)} && \text{(Apply (1.26) and (1.34))} \\
\mathbb{E}[\log X_j] &= \psi(\alpha_j)-\psi(\alpha_0) && \text{(Apply (2.277)).}
\end{aligned}
$$

Exercise 2.12

Let U be a continuous random variable uniformly distributed in the interval $[a, b]$, such that its probability density function is as in (2.278). We may prove it is normalized as follows

$$\begin{aligned}\int_a^b p(u|a,b)\,\mathrm{d}u &= \int_a^b \frac{1}{b-a}\,\mathrm{d}u \\ &= \frac{v}{b-a}\bigg|_{v=a}^{v=b} \\ &= \frac{b-a}{b-a} \\ \int_a^b p(u|a,b)\,\mathrm{d}u &= 1.\end{aligned}$$

The expected value of U is computed as

$$\begin{aligned}\mathbb{E}[U] &= \int_a^b up(u|a,b)\,\mathrm{d}u \qquad \text{(Apply (1.30))} \\ &= \int_a^b \frac{u}{b-a}\,\mathrm{d}u \qquad \text{(Apply (2.278))} \\ &= \frac{v^2}{2(b-a)}\bigg|_{v=a}^{v=b} \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} \\ \mathbb{E}[U] &= \frac{b+a}{2}.\end{aligned}$$

Lastly, the variance of U is

$$\begin{aligned}
\mathbb{V}\text{ar}[U] &= \mathbb{E}[U^2] - \{\mathbb{E}[U]\}^2 && \text{(Apply (1.39))} \\
&= \int_a^b \frac{u^2}{b-a}\,\mathrm{d}u - \frac{(b+a)^2}{4} \\
&= \frac{v^3}{3(b-a)}\bigg|_{v=a}^{v=b} - \frac{(b+a)^2}{4} \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{(b+a)^2}{4} \\
&= \frac{4b^3-4a^3-3(b-a)(b+a)^2}{12(b-a)} \\
&= \frac{4b^3-4a^3-3b^3-6ab^2-3a^2b+3ab^2+6a^2b+3a^3}{12(b-a)} \\
&= \frac{b^3-a^3-3ab^2+3a^2b}{12(b-a)} \\
&= \frac{(b-a)^3}{12(b-a)} \\
\mathbb{V}\text{ar}[U] &= \frac{(b-a)^2}{12}.
\end{aligned}$$

Exercise 2.13

Let $p(\mathbf{x})$ be a multivariate normal density function with parameters $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$, and $q(\mathbf{x})$ be a multivariate normal density function with parameters $\mathbf{m} \in \mathbb{R}^D$ and $\mathbf{L} \in \mathbb{R}^{D\times D}$, the corresponding Kullback-Leibler divergence is computed, following (1.113), as

$$
\begin{aligned}
\mathrm{KL}(p(\mathbf{x})||q(\mathbf{x})) &= -\int_{\mathbb{R}^M} p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}\, \mathrm{d}\mathbf{x} \\
&= -\int_{\mathbb{R}^M} \frac{1}{|2\pi|^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \Bigg[-\frac{D}{2}(2\pi) + \\
&- \frac{1}{2}\log|\mathbf{L}| - \frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) + \\
&+ \frac{D}{2}(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \\
&+ \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \Bigg]\, \mathrm{d}\mathbf{x} && \text{(Apply (2.43))} \\
&= -\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} - \mathbf{m}^\top \mathbf{L}^{-1}\mathbf{m}\right] + \\
&- \frac{1}{2}\int_{\mathbb{R}^M} \frac{\mathrm{tr}((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1})}{|2\pi|^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\, \mathrm{d}\mathbf{x} + \\
&+ \frac{1}{2}\int_{\mathbb{R}^M} \frac{\mathrm{tr}(\mathbf{x}\mathbf{x}^\top \mathbf{L}^{-1})}{|2\pi|^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\, \mathrm{d}\mathbf{x} + \\
&- \int_{\mathbb{R}^M} \frac{\mathrm{tr}(\mathbf{x}\mathbf{m}^\top \mathbf{L}^{-1})}{|2\pi|^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\, \mathrm{d}\mathbf{x} && \text{(Apply (1.30))} \\
&= \frac{1}{2}\Bigg[-\mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}) - \log\frac{|\boldsymbol{\Sigma}|}{|\mathbf{L}|} + \mathrm{tr}(\mathbf{m}\mathbf{m}^\top \mathbf{L}^{-1}) \\
&- 2\mathrm{tr}(\boldsymbol{\mu}\mathbf{m}^\top \mathbf{L}^{-1}) + \mathrm{tr}(\boldsymbol{\mu}\boldsymbol{\mu}\mathbf{L}^{-1} + \boldsymbol{\Sigma}\mathbf{L}^{-1}) \Bigg] && \text{(Apply (2.59), (2.62) and (2.64))} \\
\mathrm{KL}(p(\mathbf{x})||q(\mathbf{x})) &= \frac{1}{2}\Bigg[\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{L}^{-1}) - D + \\
&+ (\boldsymbol{\mu}-\mathbf{m})^\top \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) + \log\frac{|\mathbf{L}|}{|\boldsymbol{\Sigma}|}\Bigg].
\end{aligned}
$$

Exercise 2.14

We seek the density function $p(\mathbf{x})$ which solves the following optimization problem

$$p = \begin{cases} \max -\int_{\mathbb{R}^D} p(\mathbf{x}) \log p(\mathbf{x})\, \mathrm{d}\mathbf{x}, \\ \text{contrained to} \begin{cases} \int_{\mathbb{R}^D} p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1, \\ \int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x})\, \mathrm{d}\mathbf{x} = \boldsymbol{\mu}, \\ \int_{\mathbb{R}^D} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top p(\mathbf{x})\, \mathrm{d}x = \boldsymbol{\Sigma}. \end{cases} \end{cases}$$

Which may be solved by maximizing the related Lagrangian, as defined in (E.4), given as follows

$$\begin{aligned} g(p) = &-\int_{\mathbb{R}^D} p(\mathbf{x}) \log p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \lambda_1 \left(\int_{\mathbb{R}^D} p(\mathbf{x})\, \mathrm{d}\mathbf{x} - 1 \right) \\ &+ \boldsymbol{\lambda}_2^\top \left(\int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x})\, \mathrm{d}\mathbf{x} - \boldsymbol{\mu} \right) + \mathrm{tr}\left(\Lambda_3 \left[\int_{\mathbb{R}^D} (\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top p(\mathbf{x})\, \mathrm{d}\mathbf{x} - \boldsymbol{\Sigma} \right] \right). \end{aligned}$$

We differentiate $g(p)$ with respect to p, obtaining the following

$$\frac{\mathrm{d}g(p)}{\mathrm{d}p} = -\log p(\mathbf{x}) - 1 + \lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\Lambda_3 (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top].$$

Solving for ${}^{\mathrm{d}g(p)}/_{\mathrm{d}p} = 0$, we find that

$$p(\mathbf{x}) = \exp\{-1 + \lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\Lambda_3 (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\}. \tag{2.10}$$

Substituting into the first constraint, we obtain

$$\begin{aligned} 1 &= \int_{\mathbb{R}^D} p(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ 1 &= \int_{\mathbb{R}^D} \exp\{-1 + \lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\Lambda_3 (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\}\, \mathrm{d}\mathbf{x} \\ \exp\{1\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\{\mathrm{tr}(\mathbf{x}\boldsymbol{\lambda}_2^\top) + \mathrm{tr}[\Lambda_3 \mathbf{x}\mathbf{x}^\top - 2\Lambda_3 \mathbf{x}\boldsymbol{\mu}^\top + \Lambda_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top]\}\, \mathrm{d}\mathbf{x} \\ \exp\{1 - \mathrm{tr}(\Lambda_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top)\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\{\mathrm{tr}(\Lambda_3 \mathbf{x}\boldsymbol{\lambda}_2^\top \Lambda_3^{-1}) + \mathrm{tr}[\Lambda_3 \mathbf{x}\mathbf{x}^\top - 2\Lambda_3 \mathbf{x}\boldsymbol{\mu}^\top]\}\, \mathrm{d}\mathbf{x} \\ \exp\{1 - \mathrm{tr}(\Lambda_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top)\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\{\mathrm{tr}[\Lambda_3 (\mathbf{x}\mathbf{x}^\top - 2\mathbf{x}(\boldsymbol{\mu}^\top - \boldsymbol{\lambda}_2^\top \Lambda_3^{-1}/2))]\}\, \mathrm{d}\mathbf{x} \\ \exp\{1 - \mathrm{tr}(\Lambda_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top)\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\Big\{ -\frac{1}{2}\mathrm{tr}[(-2\Lambda_3)\times \\ &\qquad \times (\mathbf{x} - \boldsymbol{\mu} + \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\lambda}_2 \Lambda_3^{-1}/2)^\top] + \\ &\qquad - \mathrm{tr}[\Lambda_3 (\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)^\top] \Big\}\, \mathrm{d}\mathbf{x} \\ \exp\{\lambda_1\} 2^{-D/2} |-\Lambda_3|^{-1/2} (2\pi)^{D/2} &= \exp\{1 - \mathrm{tr}(\Lambda_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top) + \mathrm{tr}[\Lambda_3 (\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)^\top]\} \\ \exp\{\lambda_1\} &= |-\Lambda_3|^{1/2} \pi^{-D/2} \exp\{1 - \mathrm{tr}(\Lambda_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top) + \\ &\quad + \mathrm{tr}[\Lambda_3 (\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)^\top]\} \\ \lambda_1 &= \frac{1}{2}\log(|-\Lambda_3|) - \frac{D}{2}\log\pi + 1 - \mathrm{tr}[\Lambda_3 (\boldsymbol{\mu}\boldsymbol{\mu}^\top)] + \\ &\quad + \mathrm{tr}[\Lambda_3 (\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\boldsymbol{\lambda}_2/2)^\top] \\ \lambda_1 &= \frac{1}{2}\log(|-\Lambda_3|) - \frac{D}{2}\log\pi + 1 - \boldsymbol{\mu}^\top \boldsymbol{\lambda}_2 + \frac{\boldsymbol{\lambda}_2^\top \Lambda_3^{-1} \boldsymbol{\lambda}_2}{4}. \end{aligned}$$

Substituting into the second constraint, we obtain

$$
\begin{aligned}
\boldsymbol{\mu} &= \int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \int_{\mathbb{R}^D} \mathbf{x} \exp\{-1 + \lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\boldsymbol{\Lambda}_3 (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\}\, \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \int_{\mathbb{R}^D} \mathbf{x} \exp\left\{ -1 + \frac{1}{2}\log(|-\boldsymbol{\Lambda}_3|) - \frac{D}{2}\pi + 1 - \boldsymbol{\mu}^\top \boldsymbol{\lambda}_2 + \right. \\
&\quad \left. + \frac{\boldsymbol{\lambda}_2^\top \boldsymbol{\Lambda}_3^{-1} \boldsymbol{\lambda}_2}{4} + \boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\boldsymbol{\Lambda}_3 (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \exp\left\{ \frac{\boldsymbol{\lambda}_2^\top \boldsymbol{\Lambda}_3^{-1} \boldsymbol{\lambda}_2}{4} - \boldsymbol{\mu}^\top \boldsymbol{\lambda}_2 \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\boldsymbol{\Lambda}_3|^{1/2}}{\pi^{D/2}} \times \\
&\quad \times \exp\left\{ \mathrm{tr}(\mathbf{x}\boldsymbol{\lambda}_2^\top) + \mathrm{tr}[\boldsymbol{\Lambda}_3 \mathbf{x}\mathbf{x}^\top - 2\boldsymbol{\Lambda}_3 \boldsymbol{\mu}\mathbf{x}^\top + \boldsymbol{\Lambda}_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \exp\left\{ \frac{\boldsymbol{\lambda}_2^\top \boldsymbol{\Lambda}_3^{-1} \boldsymbol{\lambda}_2}{4} - \boldsymbol{\mu}^\top \boldsymbol{\lambda}_2 \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\boldsymbol{\Lambda}_3|^{1/2}}{\pi^{D/2}} \times \\
&\quad \times \exp\left\{ \mathrm{tr}\left[\boldsymbol{\Lambda}_3 (\mathbf{x}\mathbf{x}^\top - 2\boldsymbol{\mu}\mathbf{x}^\top + \boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2 \mathbf{x}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top) \right] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \exp\left\{ \frac{\boldsymbol{\lambda}_2^\top \boldsymbol{\Lambda}_3^{-1} \boldsymbol{\lambda}_2}{4} + \mathrm{tr}[\boldsymbol{\Lambda}_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top] - \boldsymbol{\mu}^\top \boldsymbol{\lambda}_2 \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\boldsymbol{\Lambda}_3|^{1/2}}{\pi^{D/2}} \times \\
&\quad \times \exp\left\{ \mathrm{tr}\left[\boldsymbol{\Lambda}_3 (\mathbf{x}\mathbf{x}^\top - (\boldsymbol{\mu} - \boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2/2)\mathbf{x}^\top) \right] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \exp\left\{ \frac{\boldsymbol{\lambda}_2^\top \boldsymbol{\Lambda}_3^{-1} \boldsymbol{\lambda}_2}{4} + \mathrm{tr}[\boldsymbol{\Lambda}_3 \boldsymbol{\mu}\boldsymbol{\mu}^\top] - \boldsymbol{\mu}^\top \boldsymbol{\lambda}_2 + \right. \\
&\quad \left. - \mathrm{tr}\left[\boldsymbol{\Lambda}_3 \left(\boldsymbol{\mu} - \boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2/2\right)\left(\boldsymbol{\mu} - \boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2/2\right)^\top \right] \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\boldsymbol{\Lambda}_3|^{1/2}}{\pi^{D/2}} \times \\
&\quad \times \exp\left\{ -\frac{1}{2}\mathrm{tr}\left[(-2\boldsymbol{\Lambda}_3)(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2/2)(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2/2)^\top \right] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\mu} &= \boldsymbol{\mu} - \frac{\boldsymbol{\Lambda}_3^{-1}\boldsymbol{\lambda}_2}{2} \\
\boldsymbol{\lambda}_2 &= \mathbf{0}.
\end{aligned}
$$

Finally, substituting into the third constraint, we obtain

$$\begin{aligned}
\boldsymbol{\Sigma} &= \int_{\mathbb{R}^D} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top p(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
\boldsymbol{\Sigma} &= \int_{\mathbb{R}^D} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \exp\{-1+\lambda_1+\boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\boldsymbol{\Lambda}_3(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\}\, \mathrm{d}\mathbf{x} \\
\boldsymbol{\Sigma} &= \int_{\mathbb{R}^D} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \exp\left\{ -1 + \frac{1}{2}\log(|-\boldsymbol{\Lambda}_3|) - \frac{D}{2}\pi + \right. \\
&\quad \left. + 1 + \mathrm{tr}[\boldsymbol{\Lambda}_3(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\Sigma} &= \int_{\mathbb{R}^D} \frac{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top}{\pi^{D/2}|-\boldsymbol{\Lambda}_3|^{-1/2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}[(-2\boldsymbol{\Lambda}_3)(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] \right\} \mathrm{d}\mathbf{x} \\
\boldsymbol{\Sigma} &= (-2\boldsymbol{\Lambda}_3)^{-1} \\
\boldsymbol{\Lambda}_3 &= -\frac{\boldsymbol{\Sigma}^{-1}}{2}.
\end{aligned}$$

Substituting the Lagrangian values into (2.10), we obtain

$$\begin{aligned}
p(\mathbf{x}) &= \exp\{-1+\lambda_1+\boldsymbol{\lambda}_2^\top \mathbf{x} + \mathrm{tr}[\boldsymbol{\Lambda}_3(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]\} \\
&= \exp\left\{ -1 + \frac{1}{2}\log(|\boldsymbol{\Sigma}^{-1}/2|) - \frac{D}{2}\log\pi + 1 - \mathrm{tr}[\boldsymbol{\Sigma}^{-1}/2(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] \right\} \\
p(\mathbf{x}) &= \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}.
\end{aligned}$$

Thereby concluding that a D dimensional multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ maximizes the differential entropy, amongst distributions with fixed and finite mean and variance.

Exercise 2.15

Let $\mathbf{X}$ be a random variable distributed as a D-dimensional multivariate normal with mean $\mu \in \mathbb{R}^D$ and covariance $\Sigma \in \mathbb{R}^{D\times D}$. Its associated differential entropy is computed as

$$
\begin{aligned}
\mathrm{H}[\mathbf{X}] &= -\int_{\mathbb{R}^D} p(\mathbf{x}) \log p(\mathbf{x})\, \mathrm{d}\mathbf{x} && \text{(Apply (1.104))} \\
&= -\int_{\mathbb{R}^D} \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^D/2|\Sigma|^{1/2}} \times \\
&\times \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| + \right. \\
&\left. -\frac{1}{2}\mathrm{tr}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}] \right] \mathrm{d}\mathbf{x} && \text{(Apply (2.43))} \\
&= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{\mathrm{tr}(\Sigma\Sigma^{-1})}{2} && \text{(Apply (1.30) and (2.64))} \\
&= \frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{D}{2} \\
\mathrm{H}[\mathbf{X}] &= \frac{1}{2}\log|\Sigma| + \frac{D}{2}(1+\log(2\pi)).
\end{aligned}
$$

Exercise 2.16

Let X_1 be a random variable distributed as a normal with mean $\mu_1 \in \mathbb{R}$ and precision $\tau_1 > 0$, and X_2 be a random variable distributed as a normal with mean $\mu_2 \in \mathbb{R}$ and precision $\tau_2 > 0$. We define $X = X_1 + X_2$, and moreover note that $X|X_2 = x_2$ is a normal random variable with mean $\mu_1 + x_2 \in \mathbb{R}$ and precision $\tau_1 > 0$. It follows from (1.46) that

$$
\begin{aligned}
p(x) &= \int_{\infty}^{\infty} p(x|x_2)p(x_2)\,\mathrm{d}x_2 \\
&= \int_{\infty}^{\infty} \sqrt{\frac{\tau_1}{2\pi}} \exp\left\{ -\frac{\tau_1}{2}(x-\mu_1-x_2)^2 \right\} \sqrt{\frac{\tau_2}{2\pi}} \exp\left\{ -\frac{\tau_2}{2}(x_2-\mu_2)^2 \right\} \mathrm{d}x_2 \\
&= \int_{\infty}^{\infty} \frac{\sqrt{\tau_1\tau_2}}{2\pi} \exp\left\{ -\frac{\tau_1}{2}(x_2-x+\mu_1)^2 - \frac{\tau_2}{2}(x_2-\mu_2)^2 \right\} \mathrm{d}x_2 \\
&= \exp\left\{ -\frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 \right\} \int_{\infty}^{\infty} \frac{\sqrt{\tau_1\tau_2}}{2\pi} \exp\left\{ -\frac{\tau_1}{2}x_2^2 + \tau_1 x_2(x-\mu_1) \right. \\
&\left. -\frac{\tau_2}{2}x_2^2 + \tau_2 x_2\mu_2 \right\} \mathrm{d}x_2 \\
&= \exp\left\{ -\frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 \right\} \int_{\infty}^{\infty} \frac{\sqrt{\tau_1\tau_2}}{2\pi} \exp\left\{ -\frac{\tau_1+\tau_2}{2}x_2^2 \right. \\
&\left. + x_2[\tau_1 x - \tau_1\mu_1 + \tau_2\mu_2] \right\} \mathrm{d}x_2 \\
&= \exp\left\{ -\frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 \right\} \\
&\times \int_{\infty}^{\infty} \frac{\sqrt{\tau_1\tau_2}}{2\pi} \exp\left\{ -\frac{\tau_1+\tau_2}{2}\left[x_2^2 - 2x_2\frac{\tau_1 x - \tau_1\mu_1 + \tau_2\mu_2}{\tau_1+\tau_2} \right] \right\} \mathrm{d}x_2 \\
&= \exp\left\{ -\frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 + \frac{(\tau_1[x-\mu_1]+\tau_2\mu_2)^2}{2(\tau_1+\tau_2)} \right\} \\
&\sqrt{\frac{\tau_1\tau_2}{\tau_1+\tau_2}} \times \int_{\infty}^{\infty} \frac{\sqrt{\tau_1+\tau_2}}{2\pi} \exp\left\{ -\frac{\tau_1+\tau_2}{2}\left[x_2 - \frac{\tau_1 x - \tau_1\mu_1 + \tau_2\mu_2}{\tau_1+\tau_2} \right]^2 \right\} \mathrm{d}x_2 \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\tau_1\tau_2}{\tau_1+\tau_2}} \exp\left\{ -\frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 + \frac{(\tau_1[x-\mu_1]+\tau_2\mu_2)^2}{2(\tau_1+\tau_2)} \right\} \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\tau_1\tau_2}{\tau_1+\tau_2}} \exp\left\{ -\frac{1}{2}\frac{1}{\tau_1+\tau_2}\left[\tau_1(\tau_1+\tau_2)(x-\mu_1)^2 \right.\right. \\
&\left.\left. + \tau_2(\tau_1+\tau_2)\mu_2^2 - (\tau_1[x-\mu_1]+\tau_2\mu_2)^2 \right] \right\}.
\end{aligned}
$$

Continued:

$$= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_1\tau_2}{\tau_1+\tau_2}}\exp\left\{-\frac{1}{2}\frac{1}{\tau_1+\tau_2}\Big[\tau_1\tau_2(x-\mu_1)^2 + \tau_1\tau_2\mu_2^2 - 2\tau_1(x-\mu_1)\tau_2\mu_2\Big]\right\}$$

$$p(x) = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\tau_1\tau_2}{\tau_1+\tau_2}}\exp\left\{-\frac{1}{2}\frac{\tau_1\tau_2}{\tau_1+\tau_2}(x-\mu_1-\mu_2)^2\right\}.$$

We thereby conclude that X is a normally distributed random variable with mean $\mu_1 + \mu_2 \in \mathbb{R}$ and precision $(\tau_1\tau_2)/(\tau_1+\tau_2) > 0$. It thereafter follows, by applying the result in (1.22), that the differential entropy associated with X is of the form

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

Exercise 2.17

Let $\mathbf{X}$ be a D-dimensional multivariate normal random variable with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and precision matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$. Consider that the precision matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$ may be rewritten as the sum of anti-symmetric and a symmetric matrix, as in $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{\mathrm{A}} + \boldsymbol{\Lambda}^{\mathrm{S}}$. It follows that the density may be rewritten as

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Lambda(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{D/2}|\boldsymbol{\Lambda}|^{1/2}} \qquad \text{(Apply (2.43))} \\ &= \frac{\exp\{-\frac{1}{2}\mathrm{tr}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}]\}}{(2\pi)^{D/2}|\boldsymbol{\Lambda}|^{1/2}} \qquad \text{(Apply (C.9))}. \end{aligned}$$

Write the symmetric and anti-symmetric matrices as follows

$$\Lambda^{\mathrm{A}}_{i,j} = \frac{\Lambda_{i,j} - \Lambda_{j,i}}{2} \quad \text{and} \quad \Lambda^{\mathrm{S}}_{i,j} = \frac{\Lambda_{i,j} + \Lambda_{j,i}}{2}.$$

It is trivial to demonstrate that $\boldsymbol{\Lambda}^{\mathrm{S}}$ is symmetric, $\boldsymbol{\Lambda}^{\mathrm{A}}$ is anti-symmetric, and $\boldsymbol{\Lambda}^{\mathrm{S}} + \boldsymbol{\Lambda}^{\mathrm{A}} = \boldsymbol{\Lambda}$. It follows that the component in the exponent of $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ may be written as follows

$$\begin{aligned} -\frac{1}{2}\mathrm{tr}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}] &= -\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)(x_i-\mu_j)\Lambda_{i,j} \\ &= -\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)(x_i-\mu_j)(\Lambda^{\mathrm{S}}_{i,j} + \Lambda^{\mathrm{A}}_{i,j}) \\ &= -\frac{1}{2}\sum_{i=1}^{D}(x_i-\mu_i)^2\Lambda^{\mathrm{S}}_{i,j} \\ &\quad -\frac{1}{2}\sum_{i=1}^{D}\sum_{j>i}^{D}(x_i-\mu_i)^2(\Lambda^{\mathrm{A}}_{i,j} + \Lambda^{\mathrm{S}}_{i,j}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{D}\sum_{j<i}^{D}(x_i-\mu_i)^2(\Lambda^{\mathrm{A}}_{i,j} + \Lambda^{\mathrm{S}}_{i,j}) \\ &= -\frac{1}{2}\sum_{i=1}^{D}(x_i-\mu_i)^2\Lambda^{\mathrm{S}}_{i,j} \\ &\quad -\frac{1}{2}\sum_{i=1}^{D}\sum_{j>i}^{D}(x_i-\mu_i)^2(\Lambda^{\mathrm{A}}_{i,j} + \Lambda^{\mathrm{S}}_{i,j}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{D}\sum_{j>i}^{D}(x_i-\mu_i)^2(-\Lambda^{\mathrm{A}}_{i,j} + \Lambda^{\mathrm{S}}_{i,j}) \\ &= -\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(x_i-\mu_i)(x_j-\mu_j)\Lambda^{\mathrm{S}}_{i,j} \\ -\frac{1}{2}\mathrm{tr}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}] &= -\frac{1}{2}\mathrm{tr}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}^{\mathrm{S}}]. \end{aligned}$$

We therefore conclude that the anti-symmetric component vanishes for the exponent, and consequently we may, without loss of generality, take $\boldsymbol{\Lambda}$ as symmetric. Note, moreover, that as the inverse of a symmetric matrix is also symmetric (as seen in Exercise 2.22), the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$ may also be chosen as symmetric.

Exercise 2.18

Let $\Sigma \in \mathbb{R}^{D\times D}$ be a symmetric matrix, whose eigenvalue equation is given as in (2.45), we may rewrite it as follows

$$\begin{aligned}
\Sigma\mathbf{u}_i &= \lambda_i\mathbf{u}_i \\
\Sigma[\Re(\mathbf{u}_i) + i\Im(\mathbf{u}_i)] &= [\Re(\lambda_i) + i\Im(\lambda_i)][\Re(\mathbf{u}_i) + i\Im(\mathbf{u}_i)] \\
\Sigma\Re(\mathbf{u}_i) &= [\Re(\lambda_i) + i\Im(\lambda_i)][\Re(\mathbf{u}_i) + i\Im(\mathbf{u}_i)] - i\Sigma\Im(\mathbf{u}_i) \\
&= \Re(\lambda_i)\Re(\mathbf{u}_i) - \Im(\lambda_i)\Im(\mathbf{u}_i) + i[\Re(\lambda_i)\Im(\mathbf{u}_i) + \Im(\lambda_i)\Re(\mathbf{u}_i) - \Sigma\Im(\mathbf{u}_i)].
\end{aligned}$$

We take the complex conjugate of both sides, obtaining the following

$$\Sigma\Re(\mathbf{u}_i) = \Re(\lambda_i)\Re(\mathbf{u}_i) - \Im(\lambda_i)\Im(\mathbf{u}_i) - i[\Re(\lambda_i)\Im(\mathbf{u}_i) + \Im(\lambda_i)\Re(\mathbf{u}_i) - \Sigma\Im(\mathbf{u}_i)].$$

Subtracting the first from the second, we obtain

$$\begin{aligned}
0 &= -2i[\Re(\lambda_i)\Im(\mathbf{u}_i) + \Im(\lambda_i)\Re(\mathbf{u}_i) - \Sigma\Im(\mathbf{u}_i)] \\
\Sigma\Im(\mathbf{u}_i) &= \Re(\lambda_i)\Im(\mathbf{u}_i) + \Im(\lambda_i)\Re(\mathbf{u}_i) \\
\Im(\mathbf{u}_i)^\top\Sigma &= \Re(\lambda_i)\Im(\mathbf{u}_i)^\top + \Im(\lambda_i)\Re(\mathbf{u}_i)^\top
\end{aligned}$$

We thereafter apply the inner product of both sides and $\mathbf{u}_i$ by right-multiplication of $\mathbf{u}_i$

$$\begin{aligned}
\Im(\mathbf{u}_i)^\top\Sigma\mathbf{u}_i &= \Re(\lambda_i)\Im(\mathbf{u}_i)^\top\mathbf{u}_i + \Im(\lambda_i)\Re(\mathbf{u}_i)^\top\mathbf{u}_i \\
\lambda_i\Im(\mathbf{u}_i)^\top\mathbf{u}_i &= \Re(\lambda_i)\Im(\mathbf{u}_i)^\top\mathbf{u}_i + \Im(\lambda_i)\Re(\mathbf{u}_i)^\top\mathbf{u}_i && \text{(Apply (2.48))} \\
i\lambda_i\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \lambda_i\Im(\mathbf{u}_i)^\top\Re(\mathbf{u}_i) &= i\Re(\lambda_i)\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \Re(\lambda_i)\Im(\mathbf{u}_i)^\top\Re(\mathbf{u}_i) + \\
&\quad + i\Im(\lambda_i)\Re(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \Im(\lambda_i)\Re(\mathbf{u}_i)^\top\Re(\mathbf{u}_i).
\end{aligned}$$

It follows that

$$\begin{aligned}
\lambda_i[i\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \Im(\mathbf{u}_i)^\top\Re(\mathbf{u}_i)] &= \Re(\lambda_i)[i\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \Im(\mathbf{u}_i)^\top\Re(\mathbf{u}_i)] + \\
&\quad + \Im(\lambda_i)[i\Re(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \Re(\mathbf{u}_i)^\top\Re(\mathbf{u}_i)] \\
&= i\Re(\lambda_i)\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \Im(\lambda_i)[i\Re(\mathbf{u}_i)^\top\Im(\mathbf{u}_i) + \\
&\quad + \lambda_i\Im(\mathbf{u}_i)^\top\Re(\mathbf{u}_i) \\
i\lambda_i\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i)^\top &= i\Re(\lambda_i)\Im(\mathbf{u}_i)^\top\Im(\mathbf{u}_i)^\top + \Im(\lambda_i)\Re(\mathbf{u}_i)^\top\Re(\mathbf{u}_i).
\end{aligned}$$

Note that the term on the left-hand-side presents the constant $i = \sqrt{-1}$, whilst on the right-hand-side only the term multiplied by $\Re(\lambda_i)$ presents the constant $i = \sqrt{-1}$. This implies that $\Im(\lambda_i) = 0$, that is, that $\lambda_i \in \mathbb{R}$. In order to now prove that $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal, we left-multiply both sides of (2.45) by $\mathbf{u}_j^\top$, obtaining the following

$$\begin{aligned}
\mathbf{u}_j^\top\Sigma\mathbf{u}_i &= \lambda_i\mathbf{u}_j^\top\mathbf{u}_i \\
\mathbf{u}_j^\top\Sigma^\top\mathbf{u}_i &= \lambda_i\mathbf{u}_j^\top\mathbf{u}_i && \text{(Symmetry of } \Sigma) \\
\lambda_j\mathbf{u}_j\mathbf{u}_i &= \lambda_i\mathbf{u}_j^\top\mathbf{u}_i && \text{(Apply (2.48))} \\
(\lambda_j - \lambda_i)\mathbf{u}_j^\top\mathbf{u}_i &= 0.
\end{aligned}$$

Consequently, provided that $\lambda_j \neq \lambda_i$, it must follow that $\mathbf{u}_j^\top\mathbf{u}_i = 0$, i.e., that the eigenvectors are orthogonal. We now seek to demonstrate that the eigenvectors may be chosen to

be orthonormal. Returning to the eigenvalue equation, we find that by left-multiplying both sides of (2.45) by $\mathbf{u}_j^\top$, we obtain the following

$$\begin{aligned}
\mathbf{u}_i^\top \boldsymbol{\Sigma} \mathbf{u}_i &= \lambda_i \mathbf{u}_i^\top \mathbf{u}_i \\
\mathbf{u}_i^\top \boldsymbol{\Sigma}^\top \mathbf{u}_i &= \lambda_i \mathbf{u}_i^\top \mathbf{u}_i \quad \text{(Symmetry of } \boldsymbol{\Sigma}\text{)} \\
\lambda_i \mathbf{u}_i^\top \mathbf{u}_i &= \lambda_i \mathbf{u}_i^\top \mathbf{u}_i \qquad \text{(Apply (2.48))}.
\end{aligned}$$

Note that the above equality holds irrespective of the value of $\mathbf{u}_i^\top \mathbf{u}_i$ (even for $\lambda_i = 0$). Thereby, we may choose $\mathbf{u}_i^\top \mathbf{u}_j = 1$ for $i = j$. Coupled with the fact that $\mathbf{u}_i$ must be orthogonal to $\mathbf{u}_j$ provided $\lambda_i \neq \lambda_j$ (and may be chosen to be orthogonal otherwise), we conclude that we may choose $\mathbf{u}_i^\top \mathbf{u}_j = I_{i,j}$, i.e., we may choose that the eigenvectors form an orthonormal basis.

Exercise 2.19

Let $\mathbf{\Sigma} \in \mathbb{R}^{D\times D}$ be a real symmetric matrix, whose eigenvalue equation is given as in (2.45). We right-multiply both sides by $\mathbf{u}_i^\top$ and subsequently sum with respect to i, obtaining the following

$$\begin{aligned}
\sum_{i=1}^{D} \mathbf{\Sigma}\mathbf{u}_i\mathbf{u}_i^\top &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i\mathbf{u}_i^\top \\
\mathbf{\Sigma}\left[\sum_{i=1}^{D} \mathbf{u}_i\mathbf{u}_i^\top\right] &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i\mathbf{u}_i^\top \\
\mathbf{\Sigma}\cdot \boldsymbol{I} &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i\mathbf{u}_i^\top \quad \text{(Orthonormality of U)} \\
\mathbf{\Sigma} &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i\mathbf{u}_i^\top.
\end{aligned}$$

In order to verify the decomposition of $\mathbf{\Sigma}^{-1}$, right-multiply both sides of (2.45) by $\mathbf{u}_i^\top$, and thereafter left-multiply both sides by $\mathbf{\Sigma}^{-1}$, yielding

$$\begin{aligned}
\mathbf{u}_i\mathbf{u}_i^\top &= \lambda_i \mathbf{\Sigma}^{-1}\mathbf{u}_i\mathbf{u}_i^\top \\
\mathbf{\Sigma}^{-1}\mathbf{u}_i\mathbf{u}_i^\top &= \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top.
\end{aligned}$$

By subsequently summing both sides with respect to i, we obtain

$$\begin{aligned}
\mathbf{\Sigma}^{-1}\left[\sum_{i=1}^{D} \mathbf{u}_i\mathbf{u}_i^\top\right] &= \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top \\
\mathbf{\Sigma}^{-1}\cdot \boldsymbol{I} &= \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top \quad \text{(Orthonormality of U)} \\
\mathbf{\Sigma}^{-1} &= \sum_{i=1}^{D} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top.
\end{aligned}$$

Exercise 2.20

Let $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ be a real symmetric matrix, we say it is a positive-definite matrix if, for all $\mathbf{a} \in \mathbb{R}^D$, it follows that

$$\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} > 0.$$

We seek to demonstrate that a necessary and sufficient condition for such is that all eigenvalues of $\boldsymbol{\Sigma}$ must be positive. First, we seek to prove it is a sufficient condition. We write the eigendecomposition of $\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$ as follows

$$\begin{aligned}
\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} &= \mathbf{a}^\top \left[\sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right] \mathbf{a} \quad \text{(Apply (2.48))} \\
&= \sum_{i=1}^{D} \lambda_i \mathbf{a}^\top \mathbf{u}_i \mathbf{u}_i^\top \mathbf{a} \\
\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} &= \sum_{i=1}^{D} \lambda_i [\mathbf{a}^\top \mathbf{u}_i]^2.
\end{aligned}$$

Trivially, $[\mathbf{a}^\top \mathbf{u}_i]^2 > 0$. It follows that, if we assume all $\lambda_i > 0$, we have that

$$\begin{aligned}
\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} &= \sum_{i=1}^{D} \lambda_i [\mathbf{a}^\top \mathbf{u}_i]^2 \\
&> 0.
\end{aligned}$$

Hence, we conclude that all eigenvalues being positive is a sufficient condition for the matrix $\boldsymbol{\Sigma}$ to be positive-definite. We seek now to prove it is a necessary condition. Consider that, as demonstrated in Exercise 2.18, we may choose $\mathbf{u}_j$ to be orthonormal. Moreover, as $\mathbf{u}_j \in \mathbb{R}^D$, if we assume that $\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} > 0$ holds for all $\mathbf{a} \in \mathbb{R}^D$, it must also hold for $\mathbf{a} = \mathbf{u}_j$, in which case we find that

$$\begin{aligned}
\mathbf{u}_j^\top \boldsymbol{\Sigma} \mathbf{u}_j &> 0 && \text{(By assumption)} \\
\mathbf{u}_j^\top \left[\sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \right] \mathbf{u}_j &> 0 && \text{(Apply (2.48))} \\
\sum_{i=1}^{D} \lambda_i \mathbf{u}_j^\top \mathbf{u}_i \mathbf{u}_i^\top \mathbf{u}_j &> 0 \\
\lambda_j &> 0 \quad \text{(Orthonormality of U).}
\end{aligned}$$

We may repeat this argument for all $j \in \{1, \dots, D\}$, thereby concluding that all eigenvalues must be positive. Hence, we prove it is a necessary condition.

Exercise 2.21

Let $\mathbf{\Sigma} \in \mathbb{R}^{D\times D}$ be a symmetric matrix. It follows that $\mathbf{\Sigma}$ may be decomposed as

$$\begin{aligned}
\mathbf{\Sigma} &= \sum_{i=1}^{D}\sum_{j=1}^{D} \sigma_{i,j}\mathbf{E}_{i,j} \\
&= \sum_{i=1}^{D}\sum_{j>i}^{D} \sigma_{i,j}\mathbf{E}_{i,j} + \sum_{i=1}^{D} \sigma_{i,i}\mathbf{E}_{i,i} + \sum_{i=1}^{D}\sum_{j<i}^{D} \sigma_{i,j}\mathbf{E}_{i,j} \\
&= \sum_{i=1}^{D}\sum_{j>i}^{D} \sigma_{i,j}\mathbf{E}_{i,j} + \sum_{i=1}^{D} \sigma_{i,i}\mathbf{E}_{i,i} + \sum_{i=1}^{D}\sum_{j>i}^{D} \sigma_{j,i}\mathbf{E}_{j,i} \\
\mathbf{\Sigma} &= \sum_{i=1}^{D}\sum_{j>i}^{D} \sigma_{i,j}(\mathbf{E}_{i,j} + \mathbf{E}_{j,i}) + \sum_{i=1}^{D} \sigma_{i,i}\mathbf{E}_{i,i},
\end{aligned}$$

where $\mathbf{E}_{i,j}$ is a matrix composed mostly of zeros, except at the (i,j)-th coordinate, at which it is one. It is therefore easy to see that the number of independent parameters is equal to the number of terms above (equivalently below) or at the diagonal of the D dimensional square matrix, given as

$$\sum_{i=1}^{D}\sum_{j\geq 1}^{D} 1 = \frac{D(D+1)}{2}.$$

Exercise 2.22

Let $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ be a symmetric matrix, it follows that

$$\begin{aligned}
\boldsymbol{\Sigma} &= \boldsymbol{\Sigma}^\top && \text{(Symmetry of } \boldsymbol{\Sigma}\text{)} \\
\boldsymbol{I} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^\top \\
\boldsymbol{I}^\top &= \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1})^\top \\
\boldsymbol{I} &= \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1})^\top && \text{(Symmetry of } \boldsymbol{I}\text{)} \\
\boldsymbol{\Sigma}^{-1} &= (\boldsymbol{\Sigma}^{-1})^\top.
\end{aligned}$$

We thereby conclude that, if $\boldsymbol{\Sigma}$ is symmetric, so too is its inverse.

Exercise 2.23

We seek herein to compute the area contained within an ellipsoid of constant Mahalanobis distance, as seen in (2.44). In order to do so, we diagonalize the corresponding coordinate space via eigendecomposition. The integral is therefore obtained as follows

$$\begin{aligned}
\int_{\{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\leq \Delta^2\}} 1\, \mathrm{d}\mathbf{x} &= \int_{\left\{\sum_{i=1}^D \frac{y_i^2}{\lambda_i}\leq \Delta^2\right\}} 1\, \mathrm{d}\mathbf{y} \qquad \text{(Apply (2.50))} \\
&= \int_{\left\{\sum_{i=1}^D \left(\frac{y_i}{\lambda_i^{1/2}}\right)^2\leq \Delta^2\right\}} 1\, \mathrm{d}\mathbf{y} \\
&= \int_{\left\{\sum_{i=1}^D z_i^2\leq \Delta^2\right\}} \prod_{i=1}^D \lambda_i^{1/2}\, \mathrm{d}\mathbf{z} \quad (\text{Set } z_i = y_i/\sqrt{\lambda_i}) \\
\int_{\{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\leq \Delta^2\}} 1\, \mathrm{d}\mathbf{x} &= |\boldsymbol{\Sigma}|^{1/2} \int_{\left\{\sum_{i=1}^D z_i^2\leq \Delta^2\right\}} 1\, \mathrm{d}\mathbf{z} \qquad \text{(Apply (2.55))}.
\end{aligned}$$

In order to proceed, we define $r^2 = \sum_{i=1}^D z_i^2$ via a spherical coordinate transform, and marginalize with respect to the angular coordinates, yielding the following volume element

$$\mathrm{d}\mathbf{z} = S_D r^{D-1} \mathrm{d}r.$$

Where S_D is the surface area of the D-dimensional unit sphere. We continue as follows

$$\begin{aligned}
\int_{\{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\leq \Delta^2\}} 1\, \mathrm{d}\mathbf{x} &= |\boldsymbol{\Sigma}|^{1/2} \int_0^\Delta S_D r^{D-1}\, \mathrm{d}r \\
&= |\boldsymbol{\Sigma}|^{1/2} \frac{S_D}{D} \Delta^D \\
\int_{\{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\leq \Delta^2\}} 1\, \mathrm{d}\mathbf{x} &= |\boldsymbol{\Sigma}|^{1/2} V_D \Delta^D \qquad \text{(Apply (1.144))}.
\end{aligned}$$

Exercise 2.24

We seek to prove the validity of (2.76). In order to do so, we left-multiply both sides by the inverse of the left-hand term, as follows

$$\begin{aligned}
\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} &= \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} &= \begin{pmatrix} \mathbf{AM}-\mathbf{BD}^{-1}\mathbf{CM} & -\mathbf{AMBD}^{-1}+\mathbf{BD}^{-1}+\mathbf{BD}^{-1}\mathbf{CMBD}^{-1} \\ \mathbf{CM}-\mathbf{DD}^{-1}\mathbf{CM} & -\mathbf{CMBD}^{-1}+\mathbf{DD}^{-1}+\mathbf{DD}^{-1}\mathbf{CMBD}^{-1} \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} &= \begin{pmatrix} [\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C}]\mathbf{M} & \mathbf{BD}^{-1}-[\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C}]\mathbf{MBD}^{-1} \\ \mathbf{CM}-\mathbf{CM} & -\mathbf{CMBD}^{-1}+\boldsymbol{I}+\boldsymbol{I}\cdot\mathbf{CMBD}^{-1} \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} &= \begin{pmatrix} \mathbf{M}^{-1}\mathbf{M} & \mathbf{BD}^{-1}-\mathbf{M}^{-1}\mathbf{MBD}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \\
\begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix}.
\end{aligned}$$

Thereby concluding that the relation is valid.

Exercise 2.25

Let $\mathbf{X}$ be a random variable with multivariate normal distribution, with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$, which may be decomposed as in (2.288). We likewise decompose $\mathbf{X} = (\mathbf{X}_\mathrm{a}, \mathbf{X}_\mathrm{b}, \mathbf{X}_\mathrm{c})^\top$. We seek to determine the distribution of $\mathbf{X}_\mathrm{a}|\mathbf{X}_b = \mathbf{x}_\mathrm{b}$. In order to do so, first we determine the distribution of $(\mathbf{X}_\mathrm{a}, \mathbf{X}_\mathrm{c})|\mathbf{X}_\mathrm{b} = \mathbf{x}_\mathrm{b}$. Utilizing previously established results, we find that it is a multivariate normal distribution, whose mean is

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{a,c|b}} &= \begin{pmatrix}\boldsymbol{\mu}_\mathrm{a}\\ \boldsymbol{\mu}_\mathrm{c}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,b}}\\ \boldsymbol{\Sigma}_{\mathrm{c,b}}\end{pmatrix}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}(\mathbf{x}_\mathrm{b} - \boldsymbol{\mu}_\mathrm{b}) \quad \text{(Apply (2.81))}\\ &= \begin{pmatrix}\boldsymbol{\mu}_\mathrm{a} + \boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\{\mathbf{x}_\mathrm{a} - \boldsymbol{\mu}_\mathrm{b}\}\\ \boldsymbol{\mu}_\mathrm{c} + \boldsymbol{\Sigma}_{\mathrm{c,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\{\mathbf{x}_\mathrm{a} - \boldsymbol{\mu}_\mathrm{b}\}\end{pmatrix}.\end{aligned}$$

Similarly, the associated covariance matrix is

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathrm{a,c|b}} &= \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,a}} & \boldsymbol{\Sigma}_{\mathrm{a,c}}\\ \boldsymbol{\Sigma}_{\mathrm{c,a}} & \boldsymbol{\Sigma}_{\mathrm{c,c}}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,b}}\\ \boldsymbol{\Sigma}_{\mathrm{c,b}}\end{pmatrix}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{b,c}}\end{pmatrix} \qquad \text{(Apply (2.82))}\\ &= \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,a}} & \boldsymbol{\Sigma}_{\mathrm{a,c}}\\ \boldsymbol{\Sigma}_{\mathrm{c,a}} & \boldsymbol{\Sigma}_{\mathrm{c,c}}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\\ \boldsymbol{\Sigma}_{\mathrm{c,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{b,c}}\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,a}} & \boldsymbol{\Sigma}_{\mathrm{a,c}}\\ \boldsymbol{\Sigma}_{\mathrm{c,a}} & \boldsymbol{\Sigma}_{\mathrm{c,c}}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,c}}\\ \boldsymbol{\Sigma}_{\mathrm{c,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{c,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,c}}\end{pmatrix}\\ \boldsymbol{\Sigma}_{\mathrm{a,c|b}} &= \begin{pmatrix}\boldsymbol{\Sigma}_{\mathrm{a,a}} - \boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{a,c}} - \boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,c}}\\ \boldsymbol{\Sigma}_{\mathrm{c,a}} - \boldsymbol{\Sigma}_{\mathrm{c,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{c,c}} - \boldsymbol{\Sigma}_{\mathrm{c,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,c}}\end{pmatrix}.\end{aligned}$$

Subsequently, by marginalizing with respect to $\mathbf{X}_\mathrm{c}$, we obtain, using previous results ,that $\mathbf{X}_\mathrm{a}|\mathbf{X}_\mathrm{b} = \mathbf{x}_\mathrm{b}$ possesses a multivariate normal distribution, with mean

$$\boldsymbol{\mu}_{\mathrm{a|b}} = \boldsymbol{\mu}_\mathrm{a} + \boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}(\mathbf{x}_\mathrm{a} - \boldsymbol{\mu}_\mathrm{b}) \quad \text{(Apply (2.92))}.$$

And covariance matrix

$$\boldsymbol{\Sigma}_{\mathrm{a|b}} = \boldsymbol{\Sigma}_{\mathrm{a,a}} - \boldsymbol{\Sigma}_{\mathrm{a,b}}\boldsymbol{\Sigma}_{\mathrm{b,b}}^{-1}\boldsymbol{\Sigma}_{\mathrm{b,a}} \quad \text{(Apply (2.93))}.$$

Exercise 2.26

We seek to prove the validity of (2.289). In order to do so, we left-multiply both sides by the inverse of its left-hand term, as follows

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

Thereby concluding that the relation is valid.

Exercise 2.27

Let $\mathbf{X}$ and $\mathbf{Z}$ be independent multivariate random variables, respectively of dimensions D. It follows that the expected value of $\mathbf{X}+\mathbf{Z}$ is computed as

$$\begin{aligned}
\mathbb{E}[\mathbf{X}+\mathbf{Z}] &= \int_{\mathbb{R}^D}(\mathbf{x}+\mathbf{z})p(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} && \text{(Apply (1.34))}\\
&= \int_{\mathbb{R}^D}\mathbf{x}p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} + \int_{\mathbb{R}^D}\mathbf{z}p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} && \text{(Independence)}\\
&= \int_{\mathbb{R}^D}p(\mathbf{z})\left[\int_{\mathbb{R}^D}\mathbf{x}p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\mathrm{d}\mathbf{z} + \int_{\mathbb{R}^D}p(\mathbf{x})\left[\int_{\mathbb{R}^D}\mathbf{z}p(\mathbf{z})\,\mathrm{d}\mathbf{z}\right]\mathrm{d}\mathbf{x}\\
&= \int_{\mathbb{R}^D}p(\mathbf{z})\mathbb{E}[\mathbf{X}]\,\mathrm{d}\mathbf{z} + \int_{\mathbb{R}^D}p(\mathbf{x})\mathbb{E}[\mathbf{Z}]\,\mathrm{d}\mathbf{x} && \text{(Apply (1.34))}\\
\mathbb{E}[\mathbf{X}+\mathbf{Z}] &= \mathbb{E}[\mathbf{X}]+\mathbb{E}[\mathbf{Z}] && \text{(Apply (1.30)).}
\end{aligned}$$

Similarly, the covariance of $\mathbf{X}+\mathbf{Z}$ is computed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\mathbf{X}+\mathbf{Z}] &= \int_{\mathbb{R}^D}(\mathbf{x}+\mathbf{z}-\mathbb{E}[\mathbf{X}+\mathbf{Z}])\times\\
&\quad\times(\mathbf{x}+\mathbf{z}-\mathbb{E}[\mathbf{X}+\mathbf{Z}])^\top p(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} && \text{(Apply (1.38))}\\
&= \int_{\mathbb{R}^D}(\mathbf{x}-\mathbb{E}[\mathbf{X}]+\mathbf{z}-\mathbb{E}[\mathbf{Z}])\times\\
&\quad\times(\mathbf{x}-\mathbb{E}[\mathbf{X}]+\mathbf{z}-\mathbb{E}[\mathbf{Z}])^\top p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} && \text{(Independence)}\\
&= \int_{\mathbb{R}^D}(\mathbf{x}-\mathbb{E}[\mathbf{X}])(\mathbf{x}-\mathbb{E}[\mathbf{X}])^\top p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}\\
&\quad+\int_{\mathbb{R}^D}(\mathbf{z}-\mathbb{E}[\mathbf{Z}])(\mathbf{z}-\mathbb{E}[\mathbf{Z}])^\top p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}\\
&\quad+\int_{\mathbb{R}^D}(\mathbf{x}-\mathbb{E}[\mathbf{X}])(\mathbf{z}-\mathbb{E}[\mathbf{Z}])^\top p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}\\
&\quad+\int_{\mathbb{R}^D}(\mathbf{z}-\mathbb{E}[\mathbf{Z}])(\mathbf{x}-\mathbb{E}[\mathbf{X}])^\top p(\mathbf{x})p(\mathbf{z})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}\\
&= \int_{\mathbb{R}^D}p(\mathbf{z})\left[\int_{\mathbb{R}^D}(\mathbf{x}-\mathbb{E}[\mathbf{X}])(\mathbf{x}-\mathbb{E}[\mathbf{X}])^\top p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\mathrm{d}\mathbf{z}\\
&\quad+\int_{\mathbb{R}^D}p(\mathbf{x})\left[\int_{\mathbb{R}^D}(\mathbf{z}-\mathbb{E}[\mathbf{Z}])(\mathbf{z}-\mathbb{E}[\mathbf{Z}])^\top p(\mathbf{z})\,\mathrm{d}\mathbf{z}\right]\mathrm{d}\mathbf{x}\\
&\quad+\left[\int_{\mathbb{R}^D}(\mathbf{x}-\mathbb{E}[\mathbf{X}])p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\left[\int_{\mathbb{R}^D}(\mathbf{z}-\mathbb{E}[\mathbf{Z}])^\top p(\mathbf{z})\,\mathrm{d}\mathbf{z}\right]\\
&\quad+\left[\int_{\mathbb{R}^D}(\mathbf{z}-\mathbb{E}[\mathbf{Z}])p(\mathbf{z})\,\mathrm{d}\mathbf{z}\right]\left[\int_{\mathbb{R}^D}(\mathbf{x}-\mathbb{E}[\mathbf{X}])^\top p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\\
\mathbb{V}\mathrm{ar}[\mathbf{X}+\mathbf{Z}] &= \mathbb{V}\mathrm{ar}[\mathbf{X}]+\mathbb{V}\mathrm{ar}[\mathbf{Z}] && \text{(Apply (1.30) and (1.34)).}
\end{aligned}$$

Trivially, this result agrees with that of Exercise 1.10, which may be verified by fixing the dimensions as $D=1$.

Exercise 2.28

Let $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})^\top$ be a $(D+M)$-dimensional random variable with normal distribution and mean as in (2.108) and covariance matrix as in (2.105). By utilizing (2.92) and (2.93), it trivial to observe that the distribution of $\mathbf{X}$ (i.e., the distribution marginalized over $\mathbf{Y}$) is a D-dimensional multivariate normal with mean

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$$

and covariance

$$\mathbb{V}\text{ar}[\mathbf{X}] = \boldsymbol{\Lambda}^{-1}.$$

Similarly, utilizing (2.81) and (2.82), we find that the conditional distribution $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ is a M-dimensional multivariate normal, with mean

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\
&= \mathbf{A}\boldsymbol{\mu} - \mathbf{A}\boldsymbol{\mu} + \mathbf{A}\mathbf{x} + \mathbf{b} \\
\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] &= \mathbf{A}\mathbf{x}.
\end{aligned}$$

And covariance

$$\begin{aligned}
\mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top - \mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top \\
&= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top \\
\mathbb{V}\text{ar}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] &= \mathbf{L}^{-1}.
\end{aligned}$$

Exercise 2.29

We seek to determine the inverse of the precision matrix in (2.104). To prevent the work from becoming cluttered, we will partition the process into four components: the upper-left, upper-right, lower-right and lower-left. Utilizing the relation proven in Exercise 2.24, we find that the upper-left of the inverse is

$$(\mathbf{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A} - \mathbf{A}^\top \mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1} = \mathbf{\Lambda}^{-1}$$

The upper-right of the inverse is

$$(\mathbf{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A} - \mathbf{A}^\top \mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{L}\mathbf{L}^{-1} = \mathbf{\Lambda}^{-1}\mathbf{A}^\top$$

The lower-left of the inverse is

$$\mathbf{L}^{-1}\mathbf{L}\mathbf{A}(\mathbf{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A} - \mathbf{A}^\top \mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1} = \mathbf{A}\mathbf{\Lambda}^{-1}$$

The lower-right of the inverse is

$$\mathbf{L}^{-1} + \mathbf{L}^{-1}\mathbf{L}\mathbf{A}(\mathbf{\Lambda} + \mathbf{A}^\top \mathbf{L}\mathbf{A} - \mathbf{A}^\top \mathbf{L}\mathbf{L}^{-1}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{L}\mathbf{L}^{-1} = \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\top$$

Hence, we conclude that the inverse of the precision matrix defined in (2.104) equals the covariance matrix in (2.105).

Exercise 2.30

We now seek to prove that, under the conditions imposed in Exercise 2.28, we may, utilizing the precision matrix in (2.104), derive (2.108). It follows that

$$
\begin{aligned}
\mathbb{E}[\mathbf{Z}] &= \mathbf{R}^{-1}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\top\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b}\end{pmatrix} && \text{(Apply (2.107))} \\
&= \begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\top \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\end{pmatrix}\begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\top\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b}\end{pmatrix} && \text{(Apply (2.105))} \\
&= \begin{pmatrix}\boldsymbol{\Lambda}^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\top\mathbf{L}\mathbf{b}] + \boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\mathbf{L}\mathbf{b} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\top\mathbf{L}\mathbf{b}] + [\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top]\mathbf{L}\mathbf{b}\end{pmatrix} \\
&= \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\mathbf{L}\mathbf{b} + \boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\mathbf{L}\mathbf{b} \\ \mathbf{A}\boldsymbol{\mu} - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\mathbf{L}\mathbf{b} + \mathbf{L}^{-1}\mathbf{L}\mathbf{b} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\mathbf{L}\mathbf{b}\end{pmatrix} \\
\mathbb{E}[\mathbf{Z}] &= \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b}\end{pmatrix}.
\end{aligned}
$$

Thereby reaching the desired conclusion.

Exercise 2.31

Let $\mathbf{X}$ and $\mathbf{Z}$ be D-dimensional multivariate normal random variables with means $\boldsymbol{\mu}_{\mathbf{X}} \in \mathbb{R}^D$ and $\boldsymbol{\mu}_{\mathbf{Z}} \in \mathbb{R}^D$ respectively, and covariances matrices $\boldsymbol{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{D\times D}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}} \in \mathbb{R}^{D\times D}$, respectively. We seek to determine the marginal distribution of $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$. To do so, consider that the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, which is determined by the following hierarchical model:

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{x}}) &= \textsc{Multivariate Normal}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{x}}) \\ p(\mathbf{y}|\mathbf{x}) &= \textsc{Multivariate Normal}(\boldsymbol{\mu}_{\mathbf{Z}} + \mathbf{x}, \boldsymbol{\Sigma}_{\mathbf{z}}). \end{aligned}$$

I.e., $\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A}$ is the D-dimensional identity matrix, and $\mathbf{b} = \boldsymbol{\mu}_{\mathrm{Z}}$. From (2.113), (2.114) and (2.115), we obtain that $\mathbf{Y}$ is a D-dimensional multivariate random variable with normal distribution and mean

$$\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\mu}_{\mathbf{Z}}.$$

And covariance matrix

$$\mathbb{V}\mathrm{ar}[\mathbf{Y}] = \boldsymbol{\Sigma}_{\mathrm{X}} + \boldsymbol{\Sigma}_{\mathrm{Z}}.$$

Exercise 2.32

We consider herein the linear-Gaussian model, where $\mathbf{X}$ is a D-dimensional multivariate normal random variable, with mean $\boldsymbol{\mu} \in \mathbb{R}$ and precision matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$, and $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ is a M-dimensional multivariate normal random variable, with mean $\mathbf{Ax} + \mathbf{b} \in \mathbb{R}^M$ and precision matrix $\mathbf{L} \in \mathbb{R}^{M\times M}$, as in (2.113) and (2.113). We aim to determine herein the marginal distribution of $\mathbf{Y}$. For that purpose, first we write the associated joint probability density function

$$\begin{aligned}
p(\mathbf{x},\mathbf{y}) &= p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) \\
&= \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{D/2}|\boldsymbol{\Lambda}^{-1}|^{1/2}}\frac{\exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{Ax}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{Ax}-\mathbf{b})\}}{(2\pi)^{D/2}|\mathbf{L}^{-1}|^{1/2}} \\
p(\mathbf{x},\mathbf{y}) &= \frac{\exp\{-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Lambda}\mathbf{x} + \mathbf{x}^\top\boldsymbol{\Lambda}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Lambda}\boldsymbol{\mu} - \frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b}) + \mathbf{x}^\top\mathbf{A}^\top\mathbf{L}(\mathbf{y}-\mathbf{b})\}}{(2\pi)^{D}|\boldsymbol{\Lambda}^{-1}|^{1/2}|\mathbf{L}^{-1}|^{1/2}}\times \\
&\quad\times\exp\left\{-\frac{1}{2}\mathbf{x}^\top\mathbf{A}^\top\mathbf{LAx}\right\}.
\end{aligned}$$

As we are utilizing the technique of completing the squares, we will restrict our study to the exponent terms dependent on $\mathbf{x}$ or $\mathbf{y}$. We find that

$$\begin{aligned}
\log p(\mathbf{x}) &\propto -\frac{1}{2}\mathbf{x}^\top[\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]\mathbf{x} + \mathbf{x}^\top[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}] - \frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b}) \\
&= -\frac{1}{2}\mathbf{x}^\top[\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]\mathbf{x} + \mathbf{x}^\top[\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}][\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}]+ \\
&\quad - \frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b}).
\end{aligned}$$

Completing the squares with respect to $\mathbf{x}$, we find that the exponent terms may be rewritten as

$$\begin{aligned}
\log p(\mathbf{x}) \propto &-\frac{1}{2}(\mathbf{x} - [\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}])^\top[\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]\times \\
&\times(\mathbf{x} - [\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}])+ \\
&+\frac{1}{2}([\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}])^\top[\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]\times \\
&\times([\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}])+ \\
&-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b}).
\end{aligned}$$

Consider that $\mathbf{x}$ is integrated out of the above written results. Hence, the exponent is of the form

$$\begin{aligned}
\log p(\mathbf{x}) \propto &\frac{1}{2}[\boldsymbol{\mu}^\top\boldsymbol{\Lambda} + \{\mathbf{y}^\top - \mathbf{b}^\top\}\mathbf{LA}][\boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{LA}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}])+ \\
&-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b}).
\end{aligned}$$

We discard terms which are independent of $\mathbf{y}$, resulting in

$$\begin{aligned}
\log p(\mathbf{x}) &\propto \frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\{\mathbf{L}\mathbf{A}[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}]^{-1}\mathbf{A}^\top\mathbf{L}-\mathbf{L}\}(\mathbf{y}-\mathbf{b})+\\
&\quad+(\mathbf{y}-\mathbf{b})^\top\mathbf{L}\mathbf{A}[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}]^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\\
&\quad=-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y}-\mathbf{b})+\\
&\quad-(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)\mathbf{L}\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\\
&\quad=-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y}-\mathbf{b})+\\
&\quad-(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}\times\\
&\quad\times[\mathbf{A}(\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}](\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}+\\
&\quad-\mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}]\boldsymbol{\Lambda}\boldsymbol{\mu}\\
&\quad=-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y}-\mathbf{b})+\\
&\quad-(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\\
&\quad=-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y}-\mathbf{b})+\\
&\quad-(\mathbf{y}-\mathbf{b})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}\mathbf{A}\boldsymbol{\mu}\\
\log p(\mathbf{x}) &\propto -\frac{1}{2}(\mathbf{y}-\mathbf{b}-\mathbf{A}\boldsymbol{\mu})^\top(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y}-\mathbf{b}-\mathbf{A}\boldsymbol{\mu}).
\end{aligned}$$

We thereby conclude, from the form presented in the exponent, that $\mathbf{Y}$ is a M-dimensional multivariate normal random variable with mean $\mathbf{A}\boldsymbol{\mu}+\mathbf{b}\in\mathbb{R}^M$ and covariance matrix $\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top\in\mathbb{R}^{M\times M}$.

Exercise 2.33

We consider now the same setup as Exercise 2.32, and aim to determine the conditional distribution $\mathbf{X}|\mathbf{Y}=\mathbf{y}$. Utilizing similar results, we find that the terms in exponent of the joint probability density function $p(\mathbf{x},\mathbf{y})$ which depend either on $\mathbf{x}$ or $\mathbf{y}$ are

$$\begin{aligned}\log p(\mathbf{x},\mathbf{y}) &\propto -\frac{1}{2}\mathbf{x}^\top[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}]\mathbf{x}+\mathbf{x}^\top[\boldsymbol{\Lambda}\boldsymbol{\mu}+\mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}]-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b})\\ &=-\frac{1}{2}\mathbf{x}^\top[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}]\mathbf{x}+\mathbf{x}^\top[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}][\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}]^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu}+\mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}]\\ &\quad-\frac{1}{2}(\mathbf{y}-\mathbf{b})^\top\mathbf{L}(\mathbf{y}-\mathbf{b}).\end{aligned}$$

Completing the squares with respect to $\mathbf{x}$, we find that the exponent terms may be rewritten as

$$\log p(\mathbf{x}|\mathbf{y}) \propto -\frac{1}{2}(\mathbf{x}-\boldsymbol{\Sigma}[\boldsymbol{\Lambda}\boldsymbol{\mu}+\mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}])^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\Sigma}[\boldsymbol{\Lambda}\boldsymbol{\mu}+\mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}]),$$

where $\boldsymbol{\Sigma}=[\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A}]^{-1}$. As we only seek the distribution $\mathbf{X}|\mathbf{Y}=\mathbf{y}$, we have discarded terms independent on $\mathbf{x}$. We consequently conclude that $\mathbf{X}|\mathbf{Y}=\mathbf{y}$ is a D-variate normal distribution with mean $\boldsymbol{\Sigma}[\boldsymbol{\Lambda}\boldsymbol{\mu}+\mathbf{A}^\top\mathbf{L}\{\mathbf{y}-\mathbf{b}\}]\in\mathbb{R}^D$ and variance $\boldsymbol{\Sigma}\in\mathbb{R}^{D\times D}$.

Exercise 2.34

Let we observe a sample of N multivariate normal random variables with known mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and unknown covariance $\boldsymbol{\Sigma}$. We differentiate the logarithm of the likelihood of the data (2.118) with respect to $\boldsymbol{\Sigma}$, resulting in

$$\begin{aligned}\frac{\partial \log p(\mathbf{x}|\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}}\Bigg[-\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| \\ &\quad - \frac{1}{2}\sum_{n=1}^{N}\operatorname{tr}\{(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\}\Bigg] && \text{(Apply (2.43))} \\ &= \frac{1}{2}\left[\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top\right]\boldsymbol{\Sigma}^{-2} - \frac{N}{2}\boldsymbol{\Sigma}^{-1} && \text{(Apply (C.21), (C.24) and (C.28)).}\end{aligned}$$

Solving for $\partial \log p(\mathbf{x}|\boldsymbol{\Sigma}) / \partial \boldsymbol{\Sigma} = \mathbf{0}$, we find that

$$\begin{aligned}\frac{1}{2}\left[\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top\right]\boldsymbol{\Sigma}^{-2} - \frac{N}{2}\boldsymbol{\Sigma}^{-1} &= 0 \\ N\boldsymbol{\Sigma} &= \sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top \\ \boldsymbol{\Sigma} &= \frac{\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^\top}{N}.\end{aligned}$$

Note that this solution is dependent on the parameter $\boldsymbol{\mu}$. From (2.121), we have that the maximum likelihood estimator of $\boldsymbol{\mu}$ is not dependent on the estimator of the covariance matrix, and we can therefore plug the maximum likelihood estimator of μ directly onto the maximum likelihood estimator of $\boldsymbol{\Sigma}$. Thereafter, we conclude that the maximum likelihood estimator for the covariance matrix is

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_i-\boldsymbol{\mu}_{\text{ML}})^\top}{N}, \tag{2.11}$$

which is the sample covariance.

Exercise 2.35

Let we observe a sample of N multivariate normal random variables with known mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and unknown covariance $\boldsymbol{\Sigma}$. It follows that

$$\begin{aligned}\mathbb{E}[(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_m - \boldsymbol{\mu})^\top] &= \mathbb{E}[\mathbf{x}_n\mathbf{x}_m^\top - \mathbf{x}_n\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbf{x}_m^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top] \\ &= \mathbb{E}[\mathbf{x}_n\mathbf{x}_m^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \qquad \text{(Apply (2.59))} \\ \mathbb{E}[\mathbf{x}_n\mathbf{x}_m^\top] &= \mathbb{E}[(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_m - \boldsymbol{\mu})^\top] + \boldsymbol{\mu}\boldsymbol{\mu}^\top.\end{aligned}$$

We note that, if $n \neq m$, the first term on the right-hand-side is $\mathbf{0}$, whilst if $n = m$, it is $\boldsymbol{\Sigma}$. Hence, we conclude

$$\mathbb{E}[\mathbf{x}_n\mathbf{x}_m^\top] = I_{n,m}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top \quad \text{(Apply (2.64))}, \tag{2.12}$$

where $I_{n,m}$ is (n,m)-th element of the identity matrix. We aim to now verify the expected value of the maximum likelihood estimator of $\boldsymbol{\Sigma}$. It follows that

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Sigma}_{\text{ML}}] &= \mathbb{E}\left[\sum_{n=1}^N \frac{(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top}{N}\right] \qquad \text{(Apply (2.11))} \\ &= \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^\top - \sum_{n=1}^N \mathbf{x}_n\boldsymbol{\mu}_{\text{ML}}^\top - \boldsymbol{\mu}_{\text{ML}}\sum_{n=1}^N \mathbf{x}_n^\top + N\boldsymbol{\mu}_{\text{ML}}\boldsymbol{\mu}_{\text{ML}}^\top\right] \\ &= \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^\top - \sum_{n=1}^N \mathbf{x}_n\left\{\sum_{m=1}^N \frac{\mathbf{x}_m^\top}{N}\right\} - \left\{\sum_{m=1}^N \frac{\mathbf{x}_m}{N}\right\}\sum_{n=1}^N \mathbf{x}_n^\top + \right. \qquad \text{(Apply (2.121))} \\ &\quad \left. + N\left\{\sum_{m=1}^N \frac{\mathbf{x}_m}{N}\right\}\left\{\sum_{m=1}^N \frac{\mathbf{x}_m^\top}{N}\right\}\right] \\ &= \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^\top - \frac{1}{N}\sum_{n=1}^N\sum_{m=1}^N \mathbf{x}_n\mathbf{x}_m^\top\right] \\ &= \frac{1}{N}\left\{\sum_{n=1}^N \mathbb{E}[\mathbf{x}_n\mathbf{x}_n^\top] - \frac{1}{N}\sum_{n=1}^N\sum_{m=1}^N \mathbb{E}[\mathbf{x}_n\mathbf{x}_m^\top]\right\} \\ &= \frac{1}{N}\left\{N\boldsymbol{\Sigma} + N\boldsymbol{\mu}\boldsymbol{\mu}^\top - \frac{1}{N}(N\boldsymbol{\Sigma} + N^2\boldsymbol{\mu}\boldsymbol{\mu}^\top)\right\} \qquad \text{(Apply (2.12))} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\text{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}.\end{aligned}$$

We hence conclude that the maximum likelihood estimator for $\boldsymbol{\Sigma}$ is biased.

Exercise 2.36

We aim to determine a sequential estimation procedure for the variance $\sigma^2 > 0$ of a univariate normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. As previously demonstrated in Exercise 1.11, for a sample of N random variables, the maximum likelihood estimate of σ^2, assuming μ is known is of the form

$$\sigma^{2,(N)}_{\text{ML}} = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}. \tag{2.13}$$

We dissect (2.13) to analyse its dependence on values prior to the N-th observation, as

$$\begin{aligned}
\sigma^{2,(N)}_{\text{ML}} &= \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} \\
&= \frac{(x_N - \mu)^2}{N} + \sum_{i=1}^{N-1} \frac{(x_i - \mu)^2}{N} \\
&= \frac{(x_N - \mu)^2}{N} + \frac{N-1}{N} \sum_{i=1}^{N-1} \frac{(x_i - \mu)^2}{N-1} \\
&= \frac{(x_N - \mu)^2}{N} + \frac{N-1}{N} \sigma^{2,(N-1)}_{\text{ML}} && \text{(Apply (2.13))}
\end{aligned}$$

$$\sigma^{2,(N)}_{\text{ML}} = \sigma^{2,(N-1)}_{\text{ML}} + \frac{1}{N}[(x_N - \mu)^2 - \sigma^{2,(N-1)}_{\text{ML}}]. \tag{2.14}$$

We now compare this approach to that which is obtained by the Robbins-Monro procedure. By substituting the observed values into (2.135), we obtain

$$\begin{aligned}
\sigma^{2,(N)}_{\text{ML}} &= \sigma^{2,(N-1)}_{\text{ML}} - a_{N-1} \frac{\mathrm{d}[-\log p(x_N | \sigma^{2,(N-1)}_{\text{ML}})]}{\mathrm{d}\sigma^2} \\
&= \sigma^{2,(N-1)}_{\text{ML}} - a_{N-1} \frac{\mathrm{d}}{\mathrm{d}\sigma^2} \left[\frac{1}{\sigma^2}(x_n - \mu)^2 + \right. \\
&\quad \left. + \frac{1}{2}\log(2\pi) + \frac{1}{2}\sigma^2 \right] \Bigg|^{\sigma^2 = \sigma^{2,(N-1)}_{\text{ML}}} && \text{(Apply (1.46))} \\
&= \sigma^{2,(N-1)}_{\text{ML}} - a_{N-1} \left[-\frac{1}{2\sigma^{4,(N-1)}_{\text{ML}}}(x_N - \mu)^2 + \frac{1}{2\sigma^{2,(N-1)}_{\text{ML}}} \right]
\end{aligned} \tag{2.15}$$

$$\sigma^{2,(N)}_{\text{ML}} = \sigma^{2,(N-1)}_{\text{ML}} + \frac{a_{N-1}}{2\sigma^{4,(N-1)}_{\text{ML}}} \left[(x_N - \mu)^2 - \sigma^{2,(N-1)}_{\text{ML}} \right]. \tag{2.16}$$

We conclude that, by choosing $a_N = 2(N+1)^{-1}\sigma^{4,(N)}_{\text{ML}}$, the procedure outlined in (2.16) is equivalent to that which is outlined in (2.14).

Exercise 2.37

We desire to determine a sequential estimation procedure for the covariance matrix $\Sigma \in \mathbb{R}^{D\times D}$ from a sample of D-dimensional multivariate normal random variables with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance matrix $\Sigma \in \mathbb{R}^{D\times D}$. As previously demonstrated in Exercise 2.34, the maximum likelihood estimator of Σ, under known $\boldsymbol{\mu}$, is

$$\Sigma_{\text{ML}}^{(N)} = \frac{\sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top}{N}. \tag{2.17}$$

We decompose (2.17) as follows

$$\begin{aligned}
\Sigma_{\text{ML}}^{(N)} &= \frac{\sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top}{N} \\
&= \frac{(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top}{N} + \sum_{i=1}^{N-1} \frac{(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top}{N} \\
&= \frac{(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top}{N} + \frac{N-1}{N}\sum_{i=1}^{N-1} \frac{(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top}{N-1} \\
&= \frac{(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top}{N} + \frac{N-1}{N}\Sigma_{\text{ML}}^{(N-1)} && \text{(Apply (2.17))}
\end{aligned}$$

$$\Sigma_{\text{ML}}^{(N)} = \Sigma_{\text{ML}}^{(N-1)} + \frac{1}{N}\left\{(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top - \Sigma_{\text{ML}}^{(N-1)}\right\}. \tag{2.18}$$

By contrast, the Robbins-Monro procedure approach is determined as follows

$$\begin{aligned}
\Sigma_{\text{ML}}^{(N)} &= \Sigma_{\text{ML}}^{(N-1)} - a_{N-1}\frac{\partial[-\log p(\mathbf{x}_N|\Sigma_{\text{ML}}^{(N-1)})]}{\partial\Sigma} && \text{(Apply (2.135))} \\
&= \Sigma_{\text{ML}}^{(N-1)} - a_{N-1}\frac{\partial}{\partial\Sigma}\left[\frac{\text{tr}[(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top\Sigma^{-1}]}{2} + \right. \\
&\left. + \frac{D}{2}\log 2\pi + \frac{1}{2}\log\Sigma\right]\Bigg|^{\Sigma=\Sigma_{\text{ML}}^{(N-1)}} && \text{(Apply (2.43))} \\
&= \Sigma_{\text{ML}}^{(N-1)} - a_{N-1}\left[\frac{1}{2}\Sigma_{\text{ML}}^{-1,(N-1)} + \right. \\
&\left. - \frac{\Sigma_{\text{ML}}^{-2,(N-1)}(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top}{2}\right] && \text{(Apply (C.21), (C.24) and (C.28))} \\
\Sigma_{\text{ML}}^{(N)} &= \Sigma_{\text{ML}}^{(N-1)} + \frac{a_{N-1}}{2}\times \\
&\times \Sigma_{\text{ML}}^{-2,(N-1)}\left[(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top - \Sigma_{\text{ML}}^{(N-1)}\right].
\end{aligned}$$

By choosing $a_N = 2(N+1)^{-1}\Sigma_{\text{ML}}^{-2,(N)}$, we find that the Robbins-Monro procedure is equivalent to that which is outlined in (2.18).

Exercise 2.38

Let us observe a sample of size N of random variables (denoted by X) which, conditioned by $\Theta = \theta$, are independent and normally distributed with mean $\theta \in \mathbb{R}$ and variance $\sigma^2 > 0$, whilst Θ is a normally distributed random variable with mean $\mu_0 \in \mathbb{R}$ and variance $\sigma_0^2 > 0$. We may determine the distribution of $\Theta|\mathrm{X} = \mathrm{x}$ as follows

$$\begin{aligned}
p(\theta|\mathrm{x}) &\propto p(\mathrm{x}|\theta)p(\theta) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\theta)^2\right\}\exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right\} \quad \text{(Apply (1.46) and (2.137))} \\
&\propto \exp\left\{-\frac{N\sigma_0^2(\theta^2-2\theta\mu_{\mathrm{ML}})+\sigma^2(\theta^2-2\theta\mu_0)}{2\sigma_0^2\sigma^2}\right\} \\
&= \exp\left\{-\frac{(N\sigma_0^2+\sigma^2)\theta^2-2[\sigma^2\mu_0+N\sigma_0^2\mu_{\mathrm{ML}}]\theta}{2\sigma_0^2\sigma^2}\right\} \\
&= \exp\left\{-\frac{N\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left[\theta^2-2\frac{\sigma^2\mu_0+N\sigma_0^2\mu_{\mathrm{ML}}}{N\sigma_0^2+\sigma^2}\theta\right]\right\} \\
p(\theta|\mathrm{x}) &\propto \exp\left\{-\frac{1}{2}\frac{N\sigma_0^2+\sigma^2}{2\sigma^2\sigma_0^2}\left[\theta-\frac{\sigma^2\mu_0+N\sigma_0^2\mu_{\mathrm{ML}}}{N\sigma_0^2+\sigma^2}\right]^2\right\}.
\end{aligned}$$

By method of completing squares, we conclude that $\Theta|\mathrm{X} = \mathrm{x}$ is distributed as univariate normal random variable with expected value

$$\begin{aligned}
\mu_N &= \mathbb{E}[\Theta|\mathrm{X} = \mathrm{x}] \\
\mu_N &= \frac{\sigma^2\mu_0}{N\sigma_0^2+\sigma^2} + \frac{N\sigma_0^2\mu_{\mathrm{ML}}}{N\sigma_0^2+\sigma} \qquad \text{(Apply (1.55))},
\end{aligned} \tag{2.19}$$

and variance

$$\begin{aligned}
\sigma_N^2 &= \mathbb{V}\mathrm{ar}[\Theta|\mathrm{X} = \mathrm{x}] \\
&= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2} \\
&= \left[\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}\right]^{-1} \\
\sigma_N^2 &= \left[\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right]^{-1}.
\end{aligned} \tag{2.20}$$

Exercise 2.39

From the results seen in Exercise 2.38, we can dissect the form of μ_N as

$$\begin{aligned}
\mu_N &= \frac{\sigma^2\mu_0}{N\sigma_0^2+\sigma^2} + \frac{N\sigma_0^2\mu_{\text{ML}}^{(N)}}{N\sigma_0^2+\sigma^2} && \text{(Apply (2.19))}\\
&= \frac{\sigma^2\mu_0}{N\sigma_0^2+\sigma^2} + \frac{\sigma_0^2\sum_{i=1}^N x_i}{N\sigma_0^2+\sigma^2} && \text{(Apply (1.55))}\\
&= \frac{\sigma^2\mu_0}{N\sigma_0^2+\sigma^2} + \frac{\sigma_0^2\sum_{i=1}^{N-1} x_i}{N\sigma_0^2+\sigma^2} + \frac{\sigma_0^2 x_N}{N\sigma_0^2+\sigma^2}\\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\left[\frac{\sigma^2\mu_0}{(N-1)\sigma_0^2+\sigma^2} + \right.\\
&\left. + \frac{(N-1)\sigma_0^2\mu_{\text{ML}}^{(N-1)}}{(N-1)\sigma_0^2+\sigma^2}\right] + \frac{\sigma_0^2 x_N}{N\sigma_0^2+\sigma^2} && \text{(Apply (1.55))}\\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\mu_{N-1} + \frac{\sigma_0^2 x_N}{N\sigma_0^2+\sigma^2} && \text{(Apply (2.19))}\\
\mu_N &= \left[1 - \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\right]\mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N.
\end{aligned}$$

Whereas the form of σ_N^2 is

$$\begin{aligned}
\sigma_N^2 &= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2} && \text{(Apply (2.20))}\\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}\\
\sigma_N^2 &= \left[1 - \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\right]\sigma_{N-1}^2.
\end{aligned}$$

Consider that we sampled $N-1$ observations, in the same context as in Exercise 2.38, resulting in the distribution of $\Theta|\mathrm{X}_{(-N)} = \mathrm{x}_{(-N)}$ with mean $\mu_{N-1} \in \mathbb{R}$ and variance $\sigma_{N-1}^2 > 0$. If we observe an additional variable X_N, we update the prior as follows

$$\begin{aligned}
p(\theta|\mathrm{x}, x_N) &\propto p(x_N|\theta)p(\theta|\mathrm{x}_{(-N)})\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(x_N-\theta)^2\right\}\exp\left\{-\frac{1}{2\sigma_{N-1}^2}(\theta-\mu_{N-1})^2\right\} && \text{(Apply (1.46) and (2.137))}\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\theta^2-2\theta x_N) - \frac{1}{2\sigma_{N-1}^2}(\theta^2-2\theta\mu_{N-1})\right\}\\
&= \exp\left\{-\frac{(\theta^2\sigma_{N-1}^2 - 2\theta x_N\sigma_{N-1}^2 + \sigma^2\theta^2 - 2\theta\sigma^2\mu_{N-1})}{2\sigma^2\sigma_{N-1}^2}\right\}\\
&= \exp\left\{-\frac{\sigma_{N-1}^2+\sigma^2}{2\sigma^2\sigma_{N-1}^2}\left[\theta^2 - 2\frac{\sigma_{N-1}^2 x_N + \sigma^2\mu_{N-1}}{\sigma_{N-1}^2+\sigma^2}\theta\right]\right\}\\
p(\theta|\mathrm{x}, x_N) &\propto \exp\left\{-\frac{\sigma_{N-1}^2+\sigma^2}{2\sigma^2\sigma_{N-1}^2}\left[\theta - \frac{\sigma_{N-1}^2 x_N + \sigma^2\mu_{N-1}}{\sigma_{N-1}^2+\sigma^2}\right]^2\right\}.
\end{aligned}$$

Therefore, in this setting we find that mean is updated as

$$\begin{aligned}
\mu_N &= \frac{\sigma^2}{\sigma_{N-1}^2+\sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2+\sigma^2}x_N \\
&= \frac{\sigma^2}{\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}+\sigma^2}\mu_{N-1} + \frac{\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}}{\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}+\sigma^2}x_N \\
&= \frac{\sigma^2}{\frac{\sigma^2\sigma_0^2+(N-1)\sigma_0^2\sigma^2+\sigma^4}{(N-1)\sigma_0^2+\sigma^2}}\mu_{N-1} + \frac{\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}}{\frac{\sigma^2\sigma_0^2+(N-1)\sigma_0^2\sigma^2+\sigma^4}{(N-1)\sigma_0^2+\sigma^2}}x_N \\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N \\
\mu_N &= \left[1-\frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}\right]\mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2+\sigma^2}x_N.
\end{aligned}$$

And the variance is updated as

$$\begin{aligned}
\sigma_N^2 &= \frac{\sigma^2\sigma_{N-1}^2}{\sigma_{N-1}^2+\sigma^2} \\
&= \frac{\sigma^2\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}}{\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}+\sigma^2} \\
&= \frac{\frac{\sigma^4\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}}{\frac{\sigma^2\sigma_0^2+(N-1)\sigma_0^2\sigma^2+\sigma^4}{(N-1)\sigma_0^2+\sigma^2}} \\
&= \frac{\sigma^4\sigma_0^2}{N\sigma^2\sigma_0^2+\sigma^4} \\
&= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2} \\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2} \\
&= \left[1-\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\right]\frac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2} \\
\sigma_N^2 &= \left[1-\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\right]\sigma_{N-1}^2.
\end{aligned}$$

We thereby conclude that either approach results in the same values for μ_N and σ_N^2 whence the complete N observations are accounted for.

Exercise 2.40

Let us observe a set of N random variables (denoted by X) which, conditioned by $\Theta = \boldsymbol{\theta}$, are independent D-dimensional multivariate normals with mean $\boldsymbol{\theta} \in \mathbb{R}$ and known covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$, whilst Θ is a D-dimensional multivariate normal with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{D\times D}$. We aim herein to determine the distribution of $\Theta|\mathrm{X} = \mathrm{x}$ as follows

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathrm{x}) &\propto p(\mathrm{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&\propto \exp\left\{ -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\theta}) \right\} \times \\
&\times \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0) \right\} && \text{(Apply (2.43))} \\
&\propto \exp\left\{ -\frac{N}{2}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + N\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} - \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \right\} \\
&= \exp\left\{ -\frac{1}{2}\boldsymbol{\theta}^\top(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\theta} + \boldsymbol{\theta}^\top(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}) \times \right. \\
&\left. \times (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}[N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0] \right\} \\
p(\boldsymbol{\theta}|\mathrm{x}) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}[N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0])^\top \times \right. \\
&\left. \times (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\theta} - (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}[N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0]) \right\}.
\end{aligned}
$$

Thereby concluding that $\Theta|\mathrm{X} = \mathrm{x}$ is D-dimensional multivariate normal random variable with expected value

$$
\begin{aligned}
\boldsymbol{\mu}_N &= \mathbb{E}[\Theta|\mathrm{X} = \mathrm{x}] \\
&= (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}[N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0] \\
\boldsymbol{\mu}_N &= (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}/N)^{-1}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0/N].
\end{aligned}
$$

And covariance

$$
\begin{aligned}
\boldsymbol{\Sigma}_N &= (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1} \\
&= \frac{1}{N}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}/N)^{-1}.
\end{aligned}
$$

Exercise 2.41

Let X be a Gamma random variable with parameters $a > 0$ and $b > 0$. We aim to demonstrate herein its corresponding probability density function is normalized, as follows

$$\begin{aligned}
\int_0^\infty p(x|a,b)\,\mathrm{d}x &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{a-1}\exp\{-bx\}\,\mathrm{d}x && \text{(Apply (2.146))}\\
&= \frac{1}{\Gamma(a)}\int_0^\infty u^{a-1}\exp\{-u\}\,\mathrm{d}u && \text{(Set } u = bx\text{)}\\
&= \frac{1}{\Gamma(a)}\Gamma(a) && \text{(Apply (1.141))}\\
\int_0^\infty p(x|a,b)\,\mathrm{d}x &= 1.
\end{aligned}$$

Thereby concluding the distribution is correctly normalized.

Exercise 2.42

Let X be a Gamma random variable with parameters $a > 0$ and $b > 0$. We aim to determine the expected value, variance and mode associated with X. Firstly, we consider the expected value of X:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x p(x|a,b)\,\mathrm{d}x && \text{(Apply (1.34))}\\
&= \int_0^\infty x \frac{b^a}{\Gamma(a)} x^{a-1}\exp\{-bx\}\,\mathrm{d}x && \text{(Apply (2.146))}\\
&= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{a+1-1}\exp\{-bx\}\,\mathrm{d}x\\
&= \frac{\Gamma(a+1)}{b\Gamma(a)}\int_0^\infty \frac{b^{a+1}}{\Gamma(a+1)} x^{a-1}\exp\{-bx\}\,\mathrm{d}x\\
\mathbb{E}[X] &= \frac{a}{b} && \text{(Apply (1.11) and (1.26)).}
\end{aligned} \tag{2.21}$$

Thereafter, the variance is computed as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X] &= \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 && \text{(Apply (1.39))}\\
&= \int_0^\infty x^2 p(x|a,b)\,\mathrm{d}x - \frac{a^2}{b^2} && \text{(Apply (1.34) and (2.21))}\\
&= \int_0^\infty x^2 \frac{b^a}{\Gamma(a)} x^{a-1}\exp\{-bx\}\,\mathrm{d}x\\
&= \frac{\Gamma(a+2)}{b^2\Gamma(a)}\int_0^\infty \frac{b^a}{\Gamma(a+2)} x^{a+2-1}\exp\{-bx\}\,\mathrm{d}x - \frac{a^2}{b^2} && \text{(Apply (1.30))}\\
&= \frac{(a+1)a}{b^2} - \frac{a^2}{b^2} && \text{(Apply (1.11))}\\
\mathbb{V}\mathrm{ar}[X] &= \frac{a}{b^2}.
\end{aligned}$$

We now aim to determine the mode of X. First, we see that, for $a < 1$, it follows that

$$\lim_{x\to 0^+} \frac{b^a}{\Gamma(a)} x^{a-1}\exp\{-x\} = \infty.$$

By which we conclude that, for $a < 1$ the maximum density location (i.e. mode) occurs at $x = 0$. By contrast, for $a \geq 1$, it suffices to take the logarithm of the probability density function (2.146), as follows

$$\log p(x|a,b) = (a-1)\log x - bx + a\log b - \log\Gamma(a),$$

thereafter differentiating it with respect to x and solving for $\mathrm{d}\log p(x|a,b)/\mathrm{d}x = 0$, from which we obtain the following

$$\begin{aligned}
\frac{a-1}{x} - b &= 0\\
x &= \frac{a-1}{b}.
\end{aligned}$$

By which we conclude that the mode of X is such that

$$\hat{x} = \begin{cases} 0 & \text{if } a < 1,\\ \frac{a-1}{b} & \text{if } a \geq 1. \end{cases}$$

Exercise 2.43

Consider the following density in (2.293), where $q > 0$ and $\sigma^2 > 0$. We aim to demonstrate this function is normalized, as follows

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x|\sigma^2, q)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{|x|^q}{2\sigma^2}\right\}\mathrm{d}x \\
&= \int_{-\infty}^{0} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{(-x)^q}{2\sigma^2}\right\}\mathrm{d}x \\
&+ \int_{0}^{\infty} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{x^q}{2\sigma^2}\right\}\mathrm{d}x \\
&= \int_{0}^{\infty} \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{x^q}{2\sigma^2}\right\}\mathrm{d}x \\
&= \int_{0}^{\infty} \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \frac{(2\sigma^2)(2\sigma)^{1/q-1}y^{1/q-1}}{q} \exp\{-y\}\,\mathrm{d}y \quad (\text{Set } y = x^q/(2\sigma^2)) \\
&= \frac{1}{\Gamma(1/y)} \int_{0}^{\infty} y^{1/q-1} \exp\{-y\}\,\mathrm{d}y \\
&= \frac{1}{\Gamma(1/y)}\Gamma(1/y) \qquad (\text{Apply (1.141)}) \\
\int_{-\infty}^{\infty} p(x|\sigma^2, q)\,\mathrm{d}x &= 1.
\end{aligned}$$

Consider now the case that (2.293) is evaluated for $q = 2$, resulting in the following

$$\begin{aligned}
p(x|\sigma^2, 2) &= \frac{2}{2(2\sigma^2)^{1/2}\Gamma(1/2)} \exp\left\{-\frac{|x|^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi\sigma)^{1/2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\},
\end{aligned}$$

wherein the result $\Gamma(1/2) = \sqrt{\pi}$ was utilized. We consider now a set of target variables $T_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n$, wherein ϵ_n are random variables distributed according to (2.293) This implies that the likelihood function associated with said set of N target variables is

$$\begin{aligned}
p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= \prod_{n=1}^{N} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{|t_n - y(\mathbf{x}_n, \mathbf{w})|^q}{2\sigma^2}\right\} \\
&= \frac{q^N}{2^N(2\sigma^2)^{N/q}\{\Gamma(1/q)\}^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}|t_n - y(\mathbf{x}_n, \mathbf{w})|^q\right\}.
\end{aligned}$$

We thereby conclude that the corresponding logarithm likelihood function is

$$\begin{aligned}
\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= N\log(q/2) - \frac{N}{q}\log(2\sigma^2) - N\log\Gamma(1/q) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}|t_n - y(\mathbf{x}_n, \mathbf{w})|^q \\
&\propto -\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q}\log(2\sigma^2).
\end{aligned}$$

Exercise 2.44

We consider a set of random variables which, conditioned on $(\Theta, \Omega)^\top = (\theta, \omega)^\top$, are independent normal random variables with mean $\theta \in \mathbb{R}$ and variance $\omega^{-1} > 0$. Moreover, let Θ, conditioned on $\Omega = \omega$, be distributed as a normal random variable with mean $\mu_0 \in \mathbb{R}$ and variance $(\beta\omega)^{-1} > 0$, and lastly Ω be a gamma distributed random variable with parameters $a > 0$ and $b > 0$. We may obtain the joint distribution of $(\Theta, \Omega)^\top | \mathrm{X} = \mathrm{x}$ as follows

$$
\begin{aligned}
p(\theta, \omega|\mathrm{x}) &\propto p(\mathrm{x}|\theta, \omega)p(\theta|\omega)p(\omega) \\
&\propto \omega^{N/2} \exp\left\{ -\frac{\omega}{2}\sum_{i=1}^{N}(x_i - \theta)^2 \right\} \times \\
&\times \omega^{1/2} \exp\left\{ -\frac{\beta\omega}{2}(\theta - \mu_0)^2 \right\} \omega^{a-1} \exp\{-b\omega\} \qquad \text{(Apply (1.46), (2.137) and (2.146))} \\
&\propto \omega^{(N+1)/2+a-1} \exp\left\{ -\frac{\omega}{2}\left[\sum_{i=1}^{N}(x_i - \mu_{\mathrm{ML}})^2 + 2b \right] \right\} \times \\
&\times \exp\left\{ -\frac{\omega}{2}\Big[N\theta^2 - 2N\theta\mu_{\mathrm{ML}} + N\mu_{\mathrm{ML}}^2 + \beta\theta^2 + \right. \\
&\left. - 2\beta\theta\mu_0 + \beta\mu_0^2 \Big] \right\} \\
&\propto \omega^{(N+1)/2+a-1} \exp\left\{ -\frac{\omega}{2}\left[\sum_{i=1}^{N}(x_i - \mu_{\mathrm{ML}})^2 + 2b \right] \right\} \times \\
&\times \exp\left\{ -\frac{\omega}{2}\Big[\{N + \beta\}\theta^2 + \right. \\
&\left. - 2\theta\{N\mu_{\mathrm{ML}} + \beta\mu_0\} + N\mu_{\mathrm{ML}}^2 + \beta\mu_0^2 \Big] \right\} \\
&= \omega^{(N+1)/2+a-1} \exp\left\{ -\frac{\omega}{2}\left[N\sigma_{\mathrm{ML}}^2 + 2b \right] \right\} \times \\
&\times \exp\left\{ -\frac{\omega}{2}\Big[\{N + \beta\}\left(\theta - \frac{N\mu_{\mathrm{ML}} + \beta\mu_0}{N + \beta} \right)^2 + \right. \\
&- \frac{N^2\mu_{\mathrm{ML}}^2 + 2N\beta\mu_{\mathrm{ML}}\mu_0 + \beta^2\mu_0^2}{N + B} \\
&\left. + \frac{N^2\mu_{\mathrm{ML}}^2 + N\beta\mu_{\mathrm{ML}}^2 + N\beta\mu_0^2 + \beta^2\mu_0^2}{N + \beta} \Big] \right\} \\
p(\theta, \omega|\mathrm{x}) &\propto \omega^{N/2+a-1} \exp\left\{ -\omega\left[\frac{1}{2}N\sigma_{\mathrm{ML}}^2 + \right. \right. \\
&\left. \left. + \frac{1}{2}\frac{N\beta(\mu_{\mathrm{ML}} - \mu_0)^2}{N + B} + b \right] \right\} \\
&\times (\omega\{N + \beta\})^{1/2} \exp\left\{ -\frac{\omega\{N + \beta\}}{2} \times \right. \\
&\left. \times \left(\theta - \frac{N\mu_{\mathrm{ML}} + \beta\mu_0}{N + \beta} \right)^2 \right\}.
\end{aligned}
$$

We thereby conclude that $(\Theta, \Omega)^\top | \mathrm{X} = \mathrm{x}$ is such that, the distribution of $\Theta | \Omega = \omega, \mathrm{X} = \mathrm{x}$ is normal with parameters μ_N given by

$$\mu_N = \frac{N\mu_{\mathrm{ML}} + \beta\mu_0}{N + \beta},$$

and variance given by

$$(\beta_N \omega)^{-1} = (\beta + N)^{-1}\omega^{-1}.$$

Moreover, $\Omega | \mathrm{X} = \mathrm{x}$ is a Gamma random variable with parameters $a_N = N/2 + a$ and b_N given by

$$b_N = \frac{1}{2}N\sigma^2_{\mathrm{ML}} + \frac{1}{2}\frac{N\beta(\mu_{\mathrm{ML}} - \mu_0)^2}{N + \beta} + b.$$

Exercise 2.45

Let us consider that we observe a sample of random variables (denoted by X) which, conditional on $\mathbf{\Lambda} = \mathbf{S}$, are independent D-dimensional multivariate normal random variables with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and precision matrix $\mathbf{S} \in \mathbb{R}^{D\times D}$. Let $\mathbf{\Sigma}$ be a random variable distributed as a Wishart with parameters $\mathbf{W} \in \mathbb{R}^{D\times D}$ and $\nu > 0$. We thereby determine the distribution of $\mathbf{\Lambda}|\mathrm{X} = \mathrm{x}$ as follows

$$\begin{aligned}
p(\mathbf{S}|\mathrm{x}) &\propto p(\mathrm{x}|\mathbf{S})p(\mathbf{S}) \\
&\propto |\mathbf{S}|^{N/2} \exp -\left\{ -\frac{1}{2}\sum_{n=1}^{N} \mathrm{tr}[(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top \mathbf{S}] \right\} \times \\
&\times |\mathbf{S}|^{(\nu-D-1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\mathbf{S}) \right\} \qquad \text{(Apply (2.43) and (2.155))} \\
p(\mathbf{S}|\mathrm{x}) &\propto |\mathbf{S}|^{(\nu+N-D-1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}[(\mathbf{W}^{-1} + N\mathbf{\Sigma}_{\mathrm{ML}})\mathbf{S}] \right\}.
\end{aligned}$$

We note by the functional form of the distribution of $\mathbf{\Lambda}|\mathrm{X} = \mathrm{x}$ that it follows a Wishart distribution, with updated parameters defined by $\nu_N = \nu + N$ and $\mathbf{W}_N = (\mathbf{W}^{-1} + N\mathbf{\Sigma}_{\mathrm{ML}})$. We consequently conclude that this distribution is conjugate with the precision parameter of the multivariate normal distribution.

Exercise 2.46

We aim to demonstrate that, if X is a random variable whose distribution, conditioned on $\Omega = \omega$, is univariate normal with mean $\mu \in \mathbb{R}$ and precision ω, and Ω is a gamma random variable with parameters $a > 0$, $b > 0$, therefore the marginal distribution of X is a Student's t. It is done as follows

$$\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty p(x|\mu,\omega)p(\omega|a,b)\,\mathrm{d}\omega && \text{(Apply (1.32))}\\
&= \int_0^\infty \frac{\omega^{1/2}}{\sqrt{2\pi}}\exp\left\{-\frac{\omega(x-\mu)^2}{2}\right\}\times\\
&\times \frac{b^a}{\Gamma(a)}\omega^{a-1}\exp\{-b\omega\}\,\mathrm{d}\omega && \text{(Apply (1.46) and (2.146))}\\
&= \frac{b^a\Gamma(a+1/2)}{\sqrt{2\pi}\Gamma(a)}\left(b+\frac{(x-\mu)^2}{2}\right)^{-a-1/2}\times\\
&\times \int_0^\infty \frac{(b+\frac{(x-\mu)^2}{2})^{a+1/2}}{\Gamma(a+1/2)}\omega^{a+1/2-1}\times\\
&\times \exp\left\{-\left[b+\frac{(x-\mu)^2}{2}\right]\omega\right\}\mathrm{d}\omega\\
p(x|\mu,a,b) &= \frac{b^a\Gamma(a+1/2)}{\sqrt{2\pi}\Gamma(a)}\left(b+\frac{(x-\mu)^2}{2}\right)^{-a-1/2} && \text{(Apply (1.26)).}
\end{aligned}$$

Defining $a = \nu/2$ and $b = \nu/(2\lambda)$

$$\begin{aligned}
p(x|\mu,\lambda,\nu) &= \frac{(\frac{\nu}{2\lambda})^{\nu/2}\Gamma(\nu/2+1/2)}{\sqrt{2\pi}\Gamma(\nu/2)}\left(\frac{\nu}{2\lambda}+\frac{(x-\mu)^2}{2}\right)^{-\nu/2-1/2}\\
&= \frac{(\frac{\nu}{2\lambda})^{\nu/2}\Gamma(\nu/2+1/2)}{\sqrt{2\pi}\Gamma(\nu/2)}\left(\frac{\nu}{2\lambda}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]\right)^{-\nu/2-1/2}\\
&= \frac{\Gamma(\nu/2+1/2)}{\sqrt{2\pi}\Gamma(\nu/2)}\left(\frac{2\lambda}{\nu}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2}\\
p(x|\mu,\lambda,\nu) &= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2}. \qquad (2.22)
\end{aligned}$$

We thereby conclude that the marginal distribution of X is a Student's t.

Exercise 2.47

We now aim to demonstrate that, for $\nu \to \infty$, the distribution in (2.22) converges to a univariate normal with mean $\mu \in \mathbb{R}$ and precision $\lambda > 0$. We consider herein the following limit, ignoring terms independent of x

$$\begin{aligned}
\lim_{\nu\to\infty}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} &= \lim_{\nu\to\infty}\left(\frac{\nu+\lambda(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2} \\
&= \lim_{\nu\to\infty}\left(\frac{\nu}{\nu+\lambda(x-\mu)^2}\right)^{(\nu+1)/2} \\
&= \lim_{\xi\to\infty}\left(1-\frac{1}{2\xi+\lambda(x-\mu)^2}\lambda(x-\mu)^2\right)^{\xi+1/2} \\
&= \lim_{\xi\to\infty}\left(1-\frac{1}{\xi+\lambda(x-\mu)^2/2}\frac{\lambda(x-\mu)^2}{2}\right)^{\xi+1/2}
\end{aligned}$$

$$\lim_{\nu\to\infty}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\nu/2-1/2} = \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\}. \tag{2.23}$$

We therefore conclude that, applying the limit $\nu \to \infty$ on the components dependent on x we obtain the terms which are dependent on x for the normal distribution. Consequently, once the resulting function presented on the right-hand-side in (2.23) is properly normalized, it is trivial to conclude that it is the normal probability density function.

Exercise 2.48

Consider the context wherein a D-dimensional random variable $\mathbf{X}$, conditioned on $\Omega = \omega$, is distributed as a multivariate normal, with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and precision $\omega\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$, wherein also Ω is a random variable distributed as a Gamma with parameters $a = \nu/2$ and $b = \nu/2$. We aim to determine the marginal distribution of X. It is as follows

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \int_0^\infty p(\mathbf{x}|\boldsymbol{\mu},\omega^{-1}\boldsymbol{\Lambda}^{-1})p(\omega|\nu)\,\mathrm{d}\omega && \text{(Apply (1.32))}\\
&= \int_0^\infty \frac{\exp\{-\frac{\omega}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{D/2}}\times\\
&\times |\omega\boldsymbol{\Lambda}|^{1/2}\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\omega^{\nu/2-1}\exp\{-\nu\omega/2\}\,\mathrm{d}\omega && \text{(Apply (2.43) and (2.146))}\\
&= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{\Gamma([D+\nu]/2)}{(\frac{\nu+\Delta^2}{2})^{(\nu+D)/2}}\\
&\times \int_0^\infty \frac{(\frac{\nu+\Delta^2}{2})^{(\nu+D)/2}}{\Gamma([D+\nu]/2)}\omega^{(D+\nu)/2-1}\exp\{-(\nu+\Delta^2)\omega/2\}\\
&= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}}\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\Gamma([D+\nu]/2)\left(\frac{\nu}{2}\left[1+\frac{\Delta^2}{\nu}\right]\right)^{-(\nu+D)/2} && \text{(Apply (1.26))}\\
p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) &= \frac{|\boldsymbol{\Lambda}|^{1/2}}{(\nu\pi)^{D/2}}\frac{\Gamma([D+\nu]/2)}{\Gamma(\nu/2)}\left(1+\frac{\Delta^2}{\nu}\right)^{-(\nu+D)/2},
\end{aligned}$$

where $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$. In order to demonstrate this distribution is properly normalized, we use the following

$$\begin{aligned}
\int_{\mathbb{R}^D} p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)\,\mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^D}\int_0^\infty p(\mathbf{x}|\boldsymbol{\mu},\omega^{-1}\boldsymbol{\Lambda}^{-1})p(\omega|\nu)\,\mathrm{d}\omega\mathrm{d}\mathbf{x} && \text{(Apply (1.32))}\\
&= \int_0^\infty\int_{\mathbb{R}^D} p(\mathbf{x}|\boldsymbol{\mu},\omega^{-1}\boldsymbol{\Lambda}^{-1})p(\omega|\nu)\,\mathrm{d}\mathbf{x}\mathrm{d}\omega\\
&= \int_0^\infty p(\omega|\nu)\left[\int_{\mathbb{R}} p(\mathbf{x}|\boldsymbol{\mu},\omega^{-1}\boldsymbol{\Lambda}^{-1})\,\mathrm{d}\mathbf{x}\right]\mathrm{d}\omega\\
&= \int_0^\infty p(\omega|\nu)\,\mathrm{d}\omega && \text{(Apply (1.30))}\\
\int_{\mathbb{R}^D} p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)\,\mathrm{d}\mathbf{x} &= 1.
\end{aligned}$$

The above results follow from the fact that the multivariate normal distribution is normalized (which may be determined via eigendecomposition) and from the result that the Gamma distribution is likewise normalized (as seen in Exercise 2.41).

Exercise 2.49

Let $\mathbf{X}$ be a D-dimensional random variable following a Student's t distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^D$, precision $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$ and degrees of freedom $\nu > 0$. We seek to determine the mean, variance and mode of said distribution, which we may do so utilizing the result in Exercise 2.46. It follows that the expected value of $\mathbf{X}$ is

$$\begin{aligned}
\mathbb{E}[\mathbf{X}] &= \int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)\, \mathrm{d}\mathbf{x} && \text{(Apply (1.34))} \\
&= \int_{\mathbb{R}^D} \int_0^\infty \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1}) p(\omega|\nu)\, \mathrm{d}\omega \mathrm{d}\mathbf{x} && \text{(Apply (1.32))} \\
&= \int_0^\infty \left[\int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1})\, \mathrm{d}\mathbf{x} \right] p(\omega|\nu)\, \mathrm{d}\omega \\
&= \int_0^\infty \boldsymbol{\mu} p(\omega|\nu)\, \mathrm{d}\omega && \text{(Apply (2.59))} \\
\mathbb{E}[\mathbf{X}] &= \boldsymbol{\mu} && \text{(Apply (1.26)).}
\end{aligned}$$

The variance of $\mathbf{X}$ is determined as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\mathbf{X}] &= \int_{\mathbb{R}^D} (\mathbf{x} - \mathbb{E}[\mathbf{X}])(\mathbf{x} - \mathbb{E}[\mathbf{X}])^\top p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)\, \mathrm{d}\mathbf{x} && \text{(Apply (1.42))} \\
&= \int_{\mathbb{R}^D} \int_0^\infty (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1}) p(\omega|\nu)\, \mathrm{d}\omega \mathrm{d}\mathbf{x} && \text{(Apply (1.32))} \\
&= \int_0^\infty \left[\int_{\mathbb{R}^D} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1})\, \mathrm{d}\mathbf{x} \right] p(\omega|\nu)\, \mathrm{d}\omega \\
&= \int_0^\infty \omega^{-1}\boldsymbol{\Lambda}^{-1} p(\omega|\nu)\, \mathrm{d}\omega && \text{(Apply (2.64))} \\
&= \int_0^\infty \omega^{-1}\boldsymbol{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\nu/2}}{\Gamma(\nu/2)} \omega^{\nu/2-1} \exp\{-\nu\omega/2\}\, \mathrm{d}\omega && \text{(Apply (2.146))} \\
&= \boldsymbol{\Lambda}^{-1} \frac{\Gamma(\nu/2-1)(\frac{\nu}{2})}{\Gamma(\nu/2)} \int_0^\infty \frac{(\frac{\nu}{2})^{\nu/2-1}}{\Gamma(\nu/2-1)} \omega^{\nu/2-2} \exp\{-\nu\omega/2\}\, \mathrm{d}\omega \\
&= \frac{\frac{\nu}{2}}{\frac{\nu}{2}-1} \boldsymbol{\Lambda}^{-1} && \text{(Apply (1.11) and (1.26))} \\
\mathbb{V}\mathrm{ar}[\mathbf{X}] &= \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1} \quad \nu > 2.
\end{aligned}$$

Lastly, we seek to determine the mode of $\mathbf{X}$. In order to do so, we take the logarithm of the probability density function of $\mathbf{X}$ and differentiate it with respect to $\mathbf{x}$, as follows

$$\begin{aligned}
\frac{\partial \log p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}} \log \left[\int_0^\infty p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1}) p(\omega|\nu)\, \mathrm{d}\omega \right] \\
&= \frac{\int_0^\infty \frac{\partial p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1})}{\partial \mathbf{x}} p(\omega|\nu)\, \mathrm{d}\omega}{\int_0^\infty p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1}) p(\omega|\nu)\, \mathrm{d}\omega}.
\end{aligned} \tag{2.24}$$

From (2.24) it is easy to see that stationary points (i.e. points whose derivative is zero) on $p(\mathbf{x}|\boldsymbol{\mu}, \omega^{-1}\boldsymbol{\Lambda}^{-1})$ are likewise stationary for $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$. Therefore, having previously demonstrated in Exercise 1.9 that the mode of the multivariate normal occurs at $\boldsymbol{\mu}$, so too does the mode of a Student's t distribution occur at $\boldsymbol{\mu}$.

Exercise 2.50

Let $\mathbf{X}$ be a D-dimensional random variable following a Student's t distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^D$, precision $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$ and degrees of freedom $\nu > 0$. We seek to demonstrate that, for $\nu \to \infty$, the distribution of $\mathbf{X}$ is multivariate normal with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and precision $\boldsymbol{\Lambda} \in \mathbb{R}^{D\times D}$. We follow a procedure analogous to that of Exercise 2.47, by applying the limit only to terms dependent on $\mathbf{x}$, as follows

$$\begin{aligned}
\lim_{\nu\to\infty}\left(1+\frac{\Delta^2}{\nu}\right)^{-(\nu+D)/2} &= \lim_{\nu\to\infty}\left(\frac{\nu+\Delta^2}{\nu}\right)^{-(\nu+D)/2} \\
&= \lim_{\nu\to\infty}\left(\frac{\nu}{\nu+\Delta^2}\right)^{\nu/2+D/2} \\
&= \lim_{\nu\to\infty}\left(1-\frac{\Delta^2}{\nu+\Delta^2}\right)^{\nu/2+D/2} \\
&= \lim_{\xi\to\infty}\left(1-\frac{\Delta^2/2}{\xi+\Delta^2/2}\right)^{\xi+D/2} \\
\lim_{\nu\to\infty}\left(1+\frac{\Delta^2}{\nu}\right)^{-(\nu+D)/2} &= \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right\}.
\end{aligned}$$

Observing the final functional form in the right-hand-side, we conclude that for $\nu \to \infty$, $\mathbf{X}$ is multivariate normal with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$.

Exercise 2.51

From the relation (2.296) we may prove that

$$\begin{aligned}
\exp\{iA\}\exp\{-iA\} &= 1 \\
[\cos A + i\sin A][\cos(-A) + i\sin(-A)] &= 1 \\
[\cos A + i\sin A][\cos A - i\sin A] &= 1 \\
[\cos A]^2 - (-1)\cdot[\sin A]^2 &= 1 \\
[\cos A]^2 + [\sin A]^2 &= 1. \qquad (2.25)
\end{aligned}$$

We hence conclude that $[\cos A]^2 + [\sin A]^2 = 1$. Moreover, we have that

$$\begin{aligned}
\cos(A-B) &= \Re[\exp\{i(A-B)\}] \\
&= \Re[\exp\{iA\}\exp\{-iB\}] \\
&= \Re[\{\cos A + i\sin A\}\{\cos(-B) + i\sin(-B)\}] \qquad \text{(Apply (2.296))} \\
&= \Re[\{\cos A + i\sin A\}\{\cos B - i\sin B\}] \\
&= \Re[\cos A\cos B - i\cos A\sin B + i\sin A\cos B + \sin A\sin B]
\end{aligned}$$

$$\cos(A-B) = \cos A\cos B + \sin A\sin B. \qquad (2.26)$$

Lastly, we also have that

$$\begin{aligned}
\sin(A-B) &= \Im[\exp\{i(A-B)\}] \\
&= \Im[\exp\{iA\}\exp\{-iB\}] \\
&= \Im[\{\cos A + i\sin A\}\{\cos(-B) + i\sin(-B)\}] \qquad \text{(Apply (2.296))} \\
&= \Im[\{\cos A + i\sin A\}\{\cos B - i\sin B\}] \\
&= \Im[\cos A\cos B - i\cos A\sin B \\
&\quad + i\sin A\cos B + \sin A\sin B]
\end{aligned}$$

$$\sin(A-B) = \sin A\cos B - \cos A\sin B. \qquad (2.27)$$

Exercise 2.52

Let Θ be a Von Mises random variable with parameters $\theta_0 \in [0, 2\pi)$ and $m > 0$, with probability density as in (2.179). Consider the following Taylor polynomial approximation of the cosine function

$$\cos x = 1 - \frac{x^2}{2} + O(x^4)$$

$$1 - \cos x = \frac{x^2}{2} + O(x^4). \tag{2.28}$$

If we take $\xi = m^{1/2}(\theta - \theta_0)$, we rewrite (2.179) as

$$\begin{aligned} p(\xi|\theta_0, m) &= \frac{1}{2\pi I_0(m)} \exp\{m\cos(m^{-1/2}\xi)\} \\ &= \frac{1}{2\pi I_0(m)} \exp\{m - m[1 - \cos(m^{-1/2}\xi)]\} \\ &= \frac{1}{2\pi I_0(m)} \exp\left\{m - m\left[\frac{m^{-1}\xi^2}{2} + O(m^{-2}\xi^4)\right]\right\} \quad \text{(Apply (2.28))} \\ p(\xi|\theta_0, m) &= \frac{1}{2\pi I_0(m)} \exp\left\{-\frac{\xi^2}{2} + m + O(m^{-1}\xi^4)\right]\right\}. \end{aligned}$$

By inspection of the term in the exponent, it thereby follows that, for $m \to \infty$, $\Xi = m^{1/2}(\Theta - \theta_0)$ is normally distributed with mean 0 and variance 1.

Exercise 2.53

We aim herein to demonstrate that the solution to (2.182) is given by (2.184). Consider the following

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

$$\sum_{n=1}^{N} (\sin\theta_n \cos\theta_0 - \sin\theta_0 \cos\theta_n) = 0 \qquad \text{(Apply (2.27))}$$

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

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

$$\tan\theta_0 = \frac{\sum_{n=1}^{N} \sin\theta_n}{\sum_{n=1}^{N} \cos\theta_n}$$

$$\theta_0 = \arctan\left\{ \frac{\sum_{i=1}^{N} \sin\theta_n}{\sum_{n=1}^{N} \cos\theta_n} \right\}.$$

Hence, we conclude that the solution is as defined in (2.184).

Exercise 2.54

Let Θ be a Von Mises random variable with parameters $\theta_0 \in [0, 2\pi]$ and $m > 0$. We aim herein to determine its mode. For that purpose, we write the logarithm of (2.179) as follows

$$\log p(\theta|\theta_0, m) = -\log(2\pi) - \log I_0(m) + m\cos(\theta - \theta_0), \tag{2.29}$$

wherein we may do so as $I_0(m) > 0$ for all $m > 0$. We thereafter differentiate it with respect to θ and solve for $\mathrm{d}\log p(\theta|\theta_0,m)/\mathrm{d}\theta = 0$, obtaining

$$\begin{aligned} -m\sin(\theta - \theta_0) &= 0 \\ \sin(\theta - \theta_0) &= 0 \\ \theta - \theta_0 &= k\pi \quad k \in \{\dots, -1, 0, 1, \dots\} \\ \theta &= \theta_0 + k\pi \quad k \in \{\dots, -1, 0, 1, \dots\}. \end{aligned}$$

In particular, as $\theta \in [0, 2\pi)$ and $\theta_0 \in [0, 2\pi)$, we will restrict the possible solutions to values of k such that $\theta_0 + k\pi \in [0, 2\pi)$ (in particular, this implies we are considering only $k \in \{-1, 0, 1\}$). It follows therefore that our possible solutions are constrained to

$$\theta \in \{\theta_0 - \pi, \theta_0, \theta_0 + \pi\}. \tag{2.30}$$

We now inspect the values of θ for which the second derivative of (2.29) is negative, as follows

$$\begin{aligned} \frac{\mathrm{d}^2 \log p(\theta|\theta_0, m)}{\mathrm{d}\theta^2} &< 0 \\ -m\cos(\theta - \theta_0) &< 0 \\ \cos(\theta - \theta_0) &> 0 \\ \theta - \theta_0 &\in \left(\frac{(2k-1)\pi}{2}, \frac{(2k+1)\pi}{2}\right) \quad k \in \{\dots, -2, 0, 2, \dots\} \end{aligned}$$

$$\theta \in \left(\theta_0 + \frac{(2k-1)\pi}{2}, \theta_0 + \frac{(2k+1)\pi}{2}\right) \quad k \in \{\dots, -2, 0, 2, \dots\}. \tag{2.31}$$

Similarly to before, we may restrict our candidate points to

$$\theta \in \left(\theta_0 - \frac{5}{2}\pi, \theta_0 - \frac{3}{2}\pi\right) \bigcup \left(\theta_0 - \frac{1}{2}\pi, \theta_0 + \frac{1}{2}\pi\right) \bigcup \left(\theta_0 + \frac{3}{2}\pi, \theta_0 + \frac{5}{2}\pi\right) \tag{2.32}$$

In order for a point to be a maximum, it must belong to the intersection of (2.30) and (2.32), which occurs only at $\theta = \theta_0$. We thereby conclude that $\theta = \theta_0$ is the mode of the Von Mises distribution. In order to determine minimum points for this distribution, we may consider the same inflection points as seen in (2.30) and, in a procedure analogous to (2.31), consider the points at which the second derivative of (2.29) is positive, obtaining the following

$$\begin{aligned} \theta \in &\left(\theta - 2\pi, \theta_0 - \frac{5}{2}\pi\right) \bigcup \left(\theta_0 - \frac{3}{2}\pi, \theta_0 - \frac{1}{2}\pi\right) \bigcup \\ &\bigcup \left(\theta_0 + \frac{1}{2}\pi, \theta_0 + \frac{3}{2}\pi\right) \bigcup \left(\theta_0 + \frac{5}{2}\pi, \theta_0 + 2\pi\right). \end{aligned} \tag{2.33}$$

By taking the intersection of the values in (2.30) and (2.33) we find $\theta \in \{\theta_0 - \pi, \theta_0 + \pi\}$. Applying the restriction that $\theta_0 \in [0, 2\pi)$ and $\theta \in [0, 2\pi)$, we find that the minimum occurs at

$$\begin{aligned}
\theta &= \begin{cases} \theta_0 - \pi & \text{if } \theta_0 + \pi \geq 2\pi, \\ \theta_0 + \pi & \text{if } \theta_0 + \pi < 2\pi. \end{cases} \\
&= \begin{cases} \theta_0 + \pi - 2\pi & \text{if } \theta_0 + \pi \geq 2\pi, \\ \theta_0 + \pi & \text{if } \theta_0 + \pi < 2\pi. \end{cases} \\
\theta &= (\theta + \pi) \bmod 2\pi.
\end{aligned}$$

Exercise 2.55

Consider that a sample of Von Mises random variables with parameters $m > 0$ and $\theta_0 \in [0, 2\pi)$ is observed (and denoted as Θ). The maximum likelihood estimator of m satisfies (2.185). It follows that

$$
\begin{aligned}
A(m_{\text{ML}}) &= \frac{1}{N}\sum_{n=1}^{N}[\cos\theta_n \cos\theta_0^{\text{ML}} + \sin\theta_n \sin\theta_0^{\text{ML}}] \\
&= \cos\theta_0^{\text{ML}} \frac{1}{N}\sum_{n=1}^{N}\cos\theta_n + \sin\theta_0^{\text{ML}} \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n \\
&= \cos\left(\arctan\left\{\frac{\sum_{i=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}\right\}\right)\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n + \\
&+ \sin\left(\arctan\left\{\frac{\sum_{i=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}\right\}\right)\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n \qquad \text{(Apply (2.184))} \\
&= \cos\left(\arctan\left\{\frac{N\bar{r}\sin\bar{\theta}}{N\bar{r}\cos\bar{\theta}}\right\}\right)\bar{r}\cos\bar{\theta} + \\
&+ \sin\left(\arctan\left\{\frac{N\bar{r}\sin\bar{\theta}}{N\bar{r}\cos\bar{\theta}}\right\}\right)\bar{r}\sin\bar{\theta} \qquad \text{(Apply (2.168))} \\
&= \cos(\bar{\theta})\bar{r}\cos\bar{\theta} + \sin(\bar{\theta})\bar{r}\sin\bar{\theta} \qquad \text{(Apply (2.169))} \\
A(m_{\text{ML}}) &= \bar{r} \qquad \text{(Apply (2.25)).}
\end{aligned}
$$

We hence conclude that $A(m_{\text{ML}}) = \bar{r}$, where $\bar{r}$ denotes the mean radius of the observations when viewed as unit vectors in the Euclidean plane.

Exercise 2.56

We aim to demonstrate in this exercise that the Beta, Gamma and Von Mises distributions belong to the class of exponential family distributions, of the form (2.194), where $\boldsymbol{\eta}$ are the natural parameters. First, the Beta distribution, as in (2.13), may be rewritten as

$$\begin{aligned} p(x|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \\ &= \frac{1}{x(1-x)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\{a\log x + b\log(1-x)\}. \end{aligned}$$

By inspection, we conclude that the components of the Beta distribution are

$$\begin{aligned} \boldsymbol{\eta} &= \begin{pmatrix} a \\ b \end{pmatrix} \\ \mathbf{u}(x) &= \begin{pmatrix} \log x \\ \log(1-x) \end{pmatrix} \\ g(\boldsymbol{\eta}) &= \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \\ h(x) &= \frac{1}{x(1-x)}. \end{aligned}$$

We rewrite the Gamma distribution, as in (2.146), as follows

$$\begin{aligned} p(x|a,b) &= \frac{b^a}{\Gamma(a)} x^{a-1}\exp\{-bx\} \\ &= \frac{1}{x}\frac{b^a}{\Gamma(a)} \exp\{a\log x - bx\}. \end{aligned}$$

By inspection, we conclude that the components of the Gamma distribution are

$$\begin{aligned} \boldsymbol{\eta} &= \begin{pmatrix} a \\ b \end{pmatrix} \\ \mathbf{u}(x) &= \begin{pmatrix} \log x \\ -x \end{pmatrix} \\ g(\boldsymbol{\eta}) &= \frac{\eta_2^{\eta_1}}{\Gamma(\eta_1)} \\ h(x) &= \frac{1}{x}. \end{aligned}$$

Lastly, we rewrite the Von Mises distribution, as in (2.179), as follows

$$\begin{aligned} p(\theta|m,\theta_0) &= \frac{1}{2\pi I_0(m)} \exp\{m\cos(\theta-\theta_0)\} \\ &= \frac{1}{2\pi I_0(m)} \exp\{m\cos\theta\cos\theta_0 + m\sin\theta\sin\theta\} \quad \text{(Apply (2.26))}. \end{aligned}$$

By inspection, we conclude that the components of the Von Mises distribution are

$$\begin{aligned}
\boldsymbol{\eta} &= \begin{pmatrix} m\cos\theta_0 \\ m\sin\theta_0 \end{pmatrix} \\
\mathbf{u}(\theta) &= \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \\
g(\boldsymbol{\eta}) &= \frac{1}{2\pi I_0(\sqrt{\eta_1^2+\eta_2^2})} \\
h(\theta) &= 1.
\end{aligned}$$

Exercise 2.57

We aim to demonstrate herein that a D-dimensional multivariate normal random variable with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and variance $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ belongs to the exponential family. In order to do so, we rewrite (2.43) as follows

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{\exp\{-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{ -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \right\} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{\exp\{-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{ -\frac{1}{2}\mathrm{vec}(\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}) + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \right\} \\
p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{D/2}} \frac{\exp\{-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}}{|\boldsymbol{\Sigma}|^{1/2}} \times \\
&\times \exp\left\{ -\frac{1}{2}(\mathbf{x}^\top \otimes \mathbf{x}^\top)\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \right\},
\end{aligned}$$

where $\mathrm{vec}(\mathbf{A})$ denotes the vectorization operator and $\otimes$ the Kronecker product. We thereby conclude, by inspection and comparison to (2.194), that the components are

$$\begin{aligned}
\boldsymbol{\eta} &= \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix} \\
\mathbf{u}(\mathbf{x}) &= \begin{pmatrix} -\frac{1}{2}(\mathbf{x}\otimes\mathbf{x}) \\ \mathbf{x} \end{pmatrix} \\
g(\boldsymbol{\eta}) &= \frac{1}{(2\pi)^{D/2}} \frac{\exp\{-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}}{|\boldsymbol{\Sigma}|^{1/2}} \\
h(\mathbf{x}) &= 1.
\end{aligned}$$

Exercise 2.58

We consider herein the general formula for a probability density function which belongs to the exponential family, such that the normalizing condition (2.195) is satisfied. First, we differentiate both sides of (2.195) with respect to $\boldsymbol{\eta}$, obtaining the following

$$\nabla\left[g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\,\mathrm{d}\mathbf{x}\right]=\mathbf{0}$$

$$\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\,\mathrm{d}\mathbf{x}+g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\mathbf{x}=\mathbf{0},$$

from which we obtain the following

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}=\mathbb{E}[\mathbf{u}(\mathbf{X})] \tag{2.34}$$

$$-\nabla\log g(\boldsymbol{\eta})=\mathbb{E}[\mathbf{u}(\mathbf{X})].$$

Differentiating again both sides of (2.195) with respect to $\boldsymbol{\eta}$, we obtain

$$\begin{aligned}
\mathbf{0}&=\nabla\nabla^\top g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\,\mathrm{d}\mathbf{x}+\\
&+\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\{\mathbf{u}(\mathbf{x})\}^\top\,\mathrm{d}\mathbf{x}+\\
&+\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\nabla^\top g(\boldsymbol{\eta})\,\mathrm{d}\mathbf{x}+\\
&+g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x})\}\mathbf{u}(\mathbf{x})\{\mathbf{u}(\mathbf{x})\}^\top\,\mathrm{d}\mathbf{x}\\
\mathbf{0}&=\frac{\nabla\nabla^\top g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}+\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\mathbb{E}[\{\mathbf{u}(\mathbf{X})\}^\top]+\\
&+\mathbb{E}[\mathbf{u}(\mathbf{X})]\frac{\nabla^\top g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}+\mathbb{E}[\mathbf{u}(\mathbf{X})\{\mathbf{u}(\mathbf{X})\}^\top] && \text{(Apply (1.34))}\\
\mathbf{0}&=-\left(-\frac{\nabla g(\boldsymbol{\eta})\nabla^\top g(\boldsymbol{\eta})}{\{g(\boldsymbol{\beta})\}^2}+\frac{\nabla g(\boldsymbol{\eta})\nabla^\top g(\boldsymbol{\eta})}{\{g(\boldsymbol{\beta})\}^2}-\frac{\nabla\nabla^\top g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\right)+\\
&-2\mathbb{E}[\mathbf{u}(\mathbf{X})]\mathbb{E}[\{\mathbf{u}(\mathbf{X})\}^\top]+\mathbb{E}[\mathbf{u}(\mathbf{X})\{\mathbf{u}(\mathbf{X})\}^\top] && \text{(Apply (2.34))}\\
\mathbf{0}&=\nabla\nabla^\top\log g(\boldsymbol{\eta})+\mathbb{E}[\mathbf{u}(\mathbf{X})]\mathbb{E}[\{\mathbf{u}(\mathbf{X})\}^\top]+\\
&-2\mathbb{E}[\mathbf{u}(\mathbf{X})]\mathbb{E}[\{\mathbf{u}(\mathbf{X})\}^\top]+\mathbb{E}[\mathbf{u}(\mathbf{X})\{\mathbf{u}(\mathbf{X})\}^\top]\\
-\nabla\nabla^\top\log g(\boldsymbol{\eta})&=\mathbb{E}[\mathbf{u}(\mathbf{X})\{\mathbf{u}(\mathbf{X})\}^\top]-\mathbb{E}[\mathbf{u}(\mathbf{X})]\mathbb{E}[\{\mathbf{u}(\mathbf{X})\}^\top]\\
-\nabla\nabla^\top\log g(\boldsymbol{\eta})&=\mathbb{V}\mathrm{ar}(\mathbf{u}(\mathbf{X})) && \text{(Apply (1.42)).}
\end{aligned}$$

Thereby obtaining the desired result.

Exercise 2.59

Consider that $f(x)$ is a properly normalized probability density function. We seek to demonstrate that any probability density function constructed as in (2.236), where $\sigma > 0$, is likewise normalized. It follows that

$$\begin{aligned}\int p(x|\sigma)\,\mathrm{d}x &= \int \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)\mathrm{d}x \quad \text{(Apply (2.236))}\\ &= \int f(y)\,\mathrm{d}y \qquad\qquad \text{(Set } y = x/\sigma\text{)}\\ \int p(x|\sigma)\,\mathrm{d}x &= 1.\end{aligned}$$

Hence concluding that $p(x|\sigma)$ is normalized.

Exercise 2.60

Let us consider a probability density function similar to a histogram, composed of bins of varying and known widths, denoted by Δ_i, and unknown heights, denoted by h_i. For a sample of N random variables drawn from this distribution, we seek to determine the maximum likelihood estimator of h_i. First, we find that the likelihood function associated with the data is such that

$$p(\mathrm{x}) = \prod_{m=1}^{N} \prod_{i\geq 1} h_i^{\mathbf{1}_{m,i}},$$

wherein $\mathbf{1}_{m,i}$ is a function which is 1 if the m-th observation falls in the i-th region, and 0 otherwise. Consequently, the logarithm of the likelihood function is

$$p(\mathrm{x}) = \sum_{m=1}^{N} \prod_{i\geq 1} \mathbf{1}_{m,i} \log h_i$$

Moreover, we note that, as $p(\mathrm{x})$ must be a probability density function, it is constrained such that $\int p(\mathrm{x})\, \mathrm{dx} = 1$ (see (1.30)). More precisely, such that

$$\sum_{i\geq 1} h_i \Delta_i = 1. \tag{2.35}$$

In order to determine the maximum likelihood estimators of h_i, we must therefore utilize a restricted maximization procedure. We must minimize the following, as defined in (E.4)

$$f(\mathbf{h}) = \sum_{m=1}^{N} \prod_{i\geq 1} \mathbf{1}_{m,i} \log h_i + \lambda \left(1 - \sum_{i\geq 1} h_i \Delta_i\right). \tag{2.36}$$

Differentiating (2.36) with respect to h_k and solving for $\partial f(\mathbf{h})/\partial h_k = 0$, we obtain

$$\begin{aligned}
\frac{\partial f(\mathbf{h})}{\partial h_k} &= 0 \\
\sum_{m=1}^{N} \frac{\mathbf{1}_{m,k}}{h_k} - \lambda \Delta_k &= 0 \\
\frac{n_k}{h_k} &= \lambda \Delta_k \\
h_k &= \frac{n_k}{\lambda \Delta_k}.
\end{aligned}$$

Substituting this result in the constraint (2.35), we obtain

$$\begin{aligned}
\sum_{i\geq 1} \frac{n_i}{\lambda \Delta_i} \Delta_i &= 1 \\
\lambda &= \sum_{i\geq 1} n_i \\
\lambda &= N.
\end{aligned}$$

We conclude that the constrained maximum likelihood estimators for the bin heights are

$$h_i = \frac{n_i}{N \Delta_i}.$$

Exercise 2.61

Consider the K-nearest neighbours density model, defined as (2.246) where K is fixed, whilst the volume V is allowed to grow until it encompasses at least K observations. Trivially, V is strictly positive. We note that, integrating over all space is this context is done via

$$\int_0^\infty p(\mathbf{x})\,\mathrm{d}V = \int_0^\infty \frac{K}{NV}\,\mathrm{d}V.$$

Where the above integral is divergent, and consequently the K-nearest neighbours density model constitutes an improper density.

Chapter 3
Linear Models for Regression

Exercise 3.1

We manipulate the $\tanh(a)$, as in (5.59) function as follows

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

$$\tanh(a) = 2\sigma(2a)-1 \qquad \text{(Apply (3.6))}. \tag{3.1}$$

Hence, we conclude that these functions are related. Let $y(x,\mathbf{w})$ and $y(x,\mathbf{u})$ be linear combinations, respectively, of tanh and logistic-sigmoid functions, which are equal for all $x\in\mathbb{R}$. It follows that

$$y(x,\mathbf{w}) = y(x,\mathbf{u})$$

$$w_0+\sum_{j=1}^{M} w_j\sigma\left(\frac{x-\mu_j}{s}\right) = u_0+\sum_{j=1}^{M} u_j\tanh\left(\frac{x-\mu_j}{2s}\right) \qquad \text{(Apply (3.101) and (3.102))}$$

$$w_0+\sum_{j=1}^{M} w_j\sigma\left(\frac{x-\mu_j}{s}\right) = u_0-\sum_{j=1}^{M} u_j+\sum_{j=1}^{M} 2u_j\sigma\left(\frac{x-\mu_j}{s}\right) \qquad \text{(Apply (3.1))}.$$

Therefore, in order for the above to be valid for all x, it must hold that

$$w_0 = u_0-\sum_{j=1}^{M} u_j \quad \text{and} \quad w_j = 2u_j, \quad \forall j\in\{1,\dots,M\}.$$

Exercise 3.2

We define the following product

$$\mathbf{y} = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{v}. \tag{3.2}$$

Assuming $(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}$ exists, we may write its eigendecomposed form as

$$(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1} = \sum_{i=1}^{M}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top \quad \text{(Apply (2.49))}.$$

We denote the column elements of $\mathbf{\Phi}$ as φ_j, such that

$$\mathbf{\Phi} = \begin{pmatrix}\varphi_1 & \dots & \varphi_M\end{pmatrix}. \tag{3.3}$$

We thereby define the following

$$\begin{aligned}
W_i &= \frac{\mathbf{\Phi}\mathbf{u}_i}{\sqrt{\lambda_i}} \\
&= \frac{1}{\sqrt{\lambda_i}}\begin{pmatrix}\varphi_1 & \dots & \varphi_M\end{pmatrix}\mathbf{u}_i && \text{(Apply (3.3))} \\
&= \frac{1}{\sqrt{\lambda_i}}(u_{1,i}\varphi_1 + u_{2,i}\varphi_2 + \ldots + u_{M,i}\varphi_M) \\
W_i &= \frac{1}{\sqrt{\lambda_i}}\sum_{j=1}^{M}u_{j,i}\varphi_j. && (3.4)
\end{aligned}$$

We may finally therefore rewrite (3.2) as follows

$$\begin{aligned}
\mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{v} &= \left[\sum_{i=1}^{M}W_iW_i^\top\right]\mathbf{v} \\
&= \sum_{i=1}^{M}\left[\left(\frac{1}{\sqrt{\lambda_i}}\sum_{j=1}^{M}u_{j,i}\varphi_j\right)\left(\frac{1}{\sqrt{\lambda_i}}\sum_{k=1}^{M}u_{k,i}\varphi_k^\top\right)\right]\mathbf{v} && \text{(Apply (3.4))} \\
&= \sum_{i=1}^{M}\left[\sum_{j=1}^{M}\sum_{k=1}^{M}\frac{1}{\lambda_i}u_{j,i}u_{k,i}(\varphi_k^\top\mathbf{v})\varphi_j\right] \\
&= \sum_{j=1}^{M}\left[\sum_{i=1}^{M}\sum_{k=1}^{M}\frac{1}{\lambda_i}u_{j,i}u_{k,i}(\varphi_k^\top\mathbf{v})\right]\varphi_j \\
\mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{v} &= \sum_{j=1}^{M}\xi_j\varphi_j. && (3.5)
\end{aligned}$$

Where

$$\xi_j = \sum_{i=1}^{M}\sum_{k=1}^{M}\frac{1}{\lambda_i}u_{j,i}u_{k,i}(\varphi_k^\top\mathbf{v}).$$

From (3.5) we conclude that, as the resulting product is a linear combination of elements belonging to the column space of $\mathbf{\Phi}$, operations of the form (3.2) project vectors $\mathbf{v}$ into

the column space of $\boldsymbol{\Phi}$. In order to verify that $\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top$ is an orthogonal projection, first we show it is idempotent, as follows

$$\begin{aligned}\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top &= \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top \\ &= \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top.\end{aligned}$$

Thereby concluding it is idempotent. Moreover, we must prove it is symmetric, which is as follows

$$\begin{aligned}[\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top]^\top &= \boldsymbol{\Phi}[(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}]^\top\boldsymbol{\Phi}^\top \\ &= \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top,\end{aligned}$$

above, we utilized the result that the inverse of a symmetric matrix is itself symmetric, as seen in Exercise 2.22. Having proven that the projection in (3.2) is both idempotent and symmetric, we conclude it is orthogonal. Therefore, we conclude that (3.2) configures an orthogonal projection, which takes any N-dimensional vector and projects it into the column space of the matrix $\boldsymbol{\Phi}$, denoted by $\mathcal{S}$. Figure 3.1 provides a rough sketch on the geometric intuition behind (3.2).

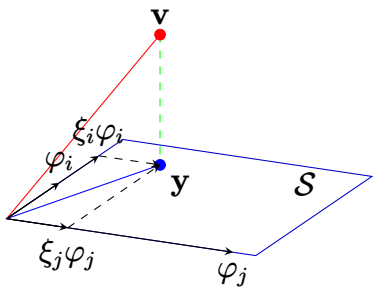


Figure 3.1: Illustration of the least-squares solution provided in (3.2).

Exercise 3.3

Consider that we observe a dataset wherein, for every point t_n, there is a corresponding positive weight factor r_n, such that we seek to minimize the error function (3.104). For that purpose, first we differentiate (3.104) with respect to $\mathbf{w}$, yielding the following

$$\frac{\mathrm{d}E_D(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -\sum_{n=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n)[t_n - \{\boldsymbol{\phi}(\mathbf{x}_n)\}^\top \mathbf{w}].$$

Solving for ${}^{\mathrm{d}E_D(\mathbf{w})}/_{\mathrm{d}\mathbf{w}} = \mathbf{0}$, we find

$$\begin{aligned}
\frac{\mathrm{d}E_D(\mathbf{w})}{\mathrm{d}\mathbf{w}} &= \mathbf{0} \\
\sum_{n=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n)[t_n - \{\boldsymbol{\phi}(\mathbf{x}_n)\}^\top \mathbf{w}] &= \mathbf{0} \\
\sum_{n=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n) t_n - \sum_{n=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n)\{\boldsymbol{\phi}(\mathbf{x}_n)\}^\top \mathbf{w} &= \mathbf{0} \\
\boldsymbol{\Phi}^\top \mathbf{R} \boldsymbol{\Phi} \mathbf{w} &= \boldsymbol{\Phi}^\top \mathbf{R} \mathbf{t} \\
\mathbf{w} &= (\boldsymbol{\Phi}^\top \mathbf{R} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{R} \mathbf{t}, \qquad (3.6)
\end{aligned}$$

where $\mathbf{R}$ is a diagonal matrix, such that $R_{n,n} = r_n$ and $R_{n,m} = 0$ if $n \neq m$. Hence, we conclude that $\mathbf{w}^\star = (\boldsymbol{\Phi}^\top \mathbf{R} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{R} \mathbf{t}$, as determined in (3.6) minimizes (3.104). We may interpret the weighing factor r_n in a number of ways. Of most significant note are: we may consider r_n to be a factor which is inversely proportional to the data variance attributed to the n-th point $\{y_n, \boldsymbol{\phi}(\mathbf{x}_n)\}$ ($r_n \propto \{\mathbb{V}\mathrm{ar}[\epsilon_n]\}^{-1}$), assuming we believe the noise variance is not homoscedastic, thus reweighing the resulting estimates to decrease the effect of noisy data points. Alternatively, we may consider r_n to be a factor which represents the effective number of known or expected data points on a larger dataset of interest which present the same values as the n-th data point $\{y_n, \boldsymbol{\phi}(\mathbf{x}_n)\}$, hence weighing our estimates to increase the effect of data points with several replicas.

Exercise 3.4

Let $\{t_n\}_{n=1}^N$ be a sample of target variables and input variables $\{\mathbf{x}_n\}_{n=1}^N$, and consider that we are adopting the linear model (3.105) for the prediction of our target variables. Consider the addition of normal noise to the input variables x_i, so that we define $\tilde{x}_i = x_i + \epsilon_i$, where $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{V}\text{ar}[\epsilon_i] = \sigma^2$. We desire to minimize the squared-loss error function averaged over the input noise. We write the following

$$\begin{aligned}
\mathbb{E}[\tilde{E}_D(\mathbf{w})] &= \mathbb{E}\left[\frac{1}{2}\sum_{n=1}^N \{y(\tilde{\mathbf{x}}_n, \mathbf{w}) - t_n\}^2\right] \\
&= \frac{1}{2}\sum_{n=1}^N \mathbb{E}\left[\left\{w_0 + \sum_{i=1}^D w_i \tilde{x}_{n,i} - t_n\right\}^2\right] && \text{(Apply (3.105))} \\
&= \frac{1}{2}\sum_{n=1}^N \mathbb{E}\left[\left\{w_0 + \sum_{i=1}^D w_i x_{n,i} - t_n + \sum_{i=1}^D w_i \epsilon_{n,i}\right\}^2\right] \\
&= \frac{1}{2}\sum_{n=1}^N \mathbb{E}\left[\left\{w_0 + \sum_{i=1}^D w_i x_{n,i} - t_n\right\}^2\right] + \\
&+ \sum_{n=1}^N \mathbb{E}\left[\left\{w_0 + \sum_{i=1}^D w_i x_{n,i} - t_n\right\}\left\{\sum_{i=1}^D w_i \epsilon_{n,i}\right\}\right] + \\
&+ \frac{1}{2}\sum_{n=1}^N \mathbb{E}\left[\left\{\sum_{i=1}^D w_i \epsilon_{n,i}\right\}^2\right] \\
\mathbb{E}[\tilde{E}_D(\mathbf{w})] &= E_D(\mathbf{w}) + \frac{N\sigma^2}{2}\sum_{i=1}^D w_i^2 && \text{(Apply (3.105)).} \qquad (3.7)
\end{aligned}$$

We thereby conclude that minimizing (3.7) is equivalent to minimizing the usual squared-loss function, restricted by a square regularization term applied to the parameters $w_1, \ldots, w_D$.

Exercise 3.5

Let us consider, for a sample of target variables $\{t_n\}_{n=1}^N$ and input variables $\{\mathbf{x}_n\}_{n=1}^N$, that we seek a solution to the following problem

$$\begin{cases} \min_{\mathbf{w}\in\mathbb{R}^M} \frac{1}{2}\sum_{n=1}^N \{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 \\ \text{subject to } \sum_{i=1}^M |w_i|^q \leq \eta, \end{cases} \tag{3.8}$$

for $\eta \geq 0$. Note that we may rewrite this constraint as a function $g(\mathbf{w}) \geq 0$ as follows:

$$\sum_{i=1}^D |w_i|^q \leq \eta \quad \Longleftrightarrow \quad 0 \leq \frac{1}{2}\left(\eta - \sum_{i=1}^D |w_i|^q\right) = g(\mathbf{w}).$$

Notably, the solution to (3.8) may likewise be found as that which solve a related Lagrangian, determined as (E.4), and in written as follows

$$\begin{aligned} L(\mathbf{w},\lambda) &= \frac{1}{2}\sum_{n=1}^N \{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 - \lambda g(\mathbf{w}) \\ &= \frac{1}{2}\sum_{n=1}^N \{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 - \frac{\lambda}{2}\left[\eta - \sum_{i=1}^D |w_i|^q\right], \end{aligned} \tag{3.9}$$

for $\lambda \geq 0$. Note that the term $\lambda\eta/2$ in (3.9) is independent of $\mathbf{w}$, and may therefore be ignored when (3.9) is minimized with respect solely to $\mathbf{w}$. We conclude that solving the constrained minimization problem in (3.8) is analogous to minimizing (3.9), which hence is equivalent to minimizing

$$\frac{1}{2}\sum_{n=1}^N \{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{i=1}^D |w_i|^q.$$

We may point to a relation between λ and η as follows: consider that, for a subset of $[0,\infty)$, we find that the solution to (3.8) is such that $g(\mathbf{w}^\star) < 0$, i.e., that the restriction is inactive. This implies that, under this range, $\lambda(\eta) = 0$. By contrast, for any fixed value $\lambda > 0$, we find that $g(\mathbf{w}^\star) = 0$, and moreover that

$$\eta(\lambda) = \frac{1}{2}\sum_{i=1}^M |w_i^\star|^q.$$

Exercise 3.6

Consider that we observe a multivariate target variable $\mathbf{t}_N$ which follows a D-variate normal distribution with mean $\mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n) \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$. We aim herein to determine the maximum likelihood estimator of $\mathbf{W}$ and $\boldsymbol{\Sigma}$. Firstly, we write the associated logarithm of the likelihood function as

$$\begin{aligned} p(\mathbf{T}|\mathbf{W},\boldsymbol{\Sigma}) = &-\frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\boldsymbol{\Sigma}| \\ &- \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{t}_n - \mathbf{W}^\top\boldsymbol{\phi}(\mathbf{t}_n)\}^\top\boldsymbol{\Sigma}^{-1}\{\mathbf{t}_n - \mathbf{W}^\top\boldsymbol{\phi}(\mathbf{x}_n)\}. \end{aligned} \tag{3.10}$$

In order to determine the maximum likelihood estimator of $\mathbf{W}$, we differentiate (3.10) with respect to $\mathbf{W}$, and solve for $\partial p(\mathbf{T}|\mathbf{W},\boldsymbol{\Sigma})/\partial\mathbf{W} = \mathbf{0}$, as follows

$$\begin{aligned} \frac{\partial p(\mathbf{T}|\mathbf{W},\boldsymbol{\Sigma})}{\partial \mathbf{W}} &= \mathbf{0} \\ \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}(\mathbf{x}_n)\{\mathbf{t}_n - \boldsymbol{\phi}(\mathbf{x}_n)\{\boldsymbol{\phi}(\mathbf{x}_n)\}^\top\mathbf{W}\} &= \mathbf{0} \qquad \text{(Apply (C.19))} \\ \sum_{n=1}^{N}\boldsymbol{\phi}(\mathbf{x}_n)\mathbf{t}_n - \sum_{n=1}^{N}\boldsymbol{\phi}(\mathbf{x}_n)\{\boldsymbol{\phi}(\mathbf{x}_n)\}^\top\mathbf{W} &= \mathbf{0} \\ \mathbf{W} &= (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{T}. \end{aligned}$$

We hence conclude that the maximum likelihood estimator of $\mathbf{W}$ is $\mathbf{W}_{\text{ML}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathbf{T}$. By decomposing $\mathbf{T}$ into D columns, we find that

$$\begin{aligned} \mathbf{W}_{\text{ML}} &= (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\begin{pmatrix}\mathrm{T}_1 & \dots & \mathrm{T}_D\end{pmatrix} \\ &= \begin{pmatrix}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathrm{T}_1 & \dots & (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\mathrm{T}_D\end{pmatrix}. \end{aligned}$$

Consequently, every column of $\mathbf{W}_{\text{ML}}$ may be written as a distinct solution to a ordinary least squares problem. In order to determine the maximum likelihood estimator of $\boldsymbol{\Sigma}$, we differentiate (3.10) with respect to $\boldsymbol{\Sigma}$, and solve for $\partial p(\mathbf{T}|\mathbf{W},\boldsymbol{\Sigma})/\partial\boldsymbol{\Sigma} = \mathbf{0}$, as follows

$$\begin{aligned} \mathbf{0} &= \frac{\partial p(\mathbf{T}|\mathbf{W},\boldsymbol{\Sigma})}{\partial\boldsymbol{\Sigma}} \\ \mathbf{0} &= -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-2}(\mathbf{t}_n - \mathbf{W}^\top\boldsymbol{\phi}(\mathbf{x}_n))\times \\ &\quad\times(\mathbf{t}_n - \mathbf{W}^\top\boldsymbol{\phi}(\mathbf{x}_n))^\top \qquad \text{(Apply (C.21), (C.24) and (C.28)).} \end{aligned}$$

We thereby conclude that the maximum likelihood point, with respect to $\boldsymbol{\Sigma}$, occurs at

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^\top\boldsymbol{\phi}(\mathbf{x}_n))(\mathbf{t}_n - \mathbf{W}^\top\boldsymbol{\phi}(\mathbf{x}_n))^\top. \tag{3.11}$$

Note that the expression in (3.11) is dependent on $\mathbf{W}$, whilst the maximum likelihood estimator of $\mathbf{W}$ is independent of $\boldsymbol{\Sigma}$. Hence, we can simply plug $\mathbf{W}_{\text{ML}}$ onto (3.11), such that the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is

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

Exercise 3.7

Consider that we observe a set of target variables $\{t_n\}_{n=1}^N$ which, conditioned on $\mathbf{w}$, are distributed as normal random variables with mean $\mathbf{w}^\top \phi(\mathbf{x}_n) \in \mathbb{R}$ and precision $\beta > 0$. Let also $\mathbf{w}$ be a normally distributed random variable with mean $\mathbf{m}_0 \in \mathbb{R}^D$ and covariance matrix $S_0 \in \mathbb{R}^{D\times D}$. We seek to determine the distribution of $\mathbf{w}|\mathrm{T}$. It is as follows

$$
\begin{aligned}
p(\mathbf{w}|\mathrm{T}) &\propto p(\mathrm{T}|\mathbf{w})p(\mathbf{w}) \\
&\propto \exp\left\{ -\frac{\beta}{2}\sum_{n=1}^N (t_n - \mathbf{w}^\top\phi(\mathbf{x}_n))^2 \right\} \times \\
&\times \exp\left\{ -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \right\} \qquad \text{(Apply (2.43) and (2.137))} \\
&\propto \exp\left\{ \beta\sum_{n=1}^N t_n \mathbf{w}^\top\phi(\mathbf{x}_n) + \right. \\
&- \frac{\beta}{2}\sum_{n=1}^N \mathbf{w}^\top\phi(\mathbf{x}_n)\{\phi(\mathbf{x}_n)\}^\top\mathbf{w} + \\
&\left. - \frac{1}{2}\mathbf{w}^\top\mathbf{S}_0^{-1}\mathbf{w} + \mathbf{w}^\top\mathbf{S}_0^{-1}\mathbf{m}_0 \right\} \\
&= \exp\left\{ \mathbf{w}^\top\left[\beta\sum_{n=1}^N t_n\phi(\mathbf{x}_n) + \mathbf{S}_0^{-1}\mathbf{m}_0\right] + \right. \\
&\left. - \frac{1}{2}\mathbf{w}^\top\left[\beta\sum_{n=1}^N \phi(\mathbf{x}_n)\{\phi(\mathbf{x}_n)\}^\top + \mathbf{S}_0^{-1}\right]\mathbf{w} \right\} \\
&= \exp\left\{ \mathbf{w}^\top\left[\beta\mathbf{\Phi}^\top\mathbf{t} + \mathbf{S}_0^{-1}\mathbf{m}_0\right] + \right. \\
&\left. - \frac{1}{2}\mathbf{w}^\top\left[\beta\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}_0^{-1}\right]\mathbf{w} \right\} \\
p(\mathbf{w}|\mathrm{T}) &\propto \exp\left\{ -\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) \right\},
\end{aligned}
$$

where

$$
\mathbf{m}_N = S_N\left[\beta\mathbf{\Phi}^\top\mathbf{t} + \mathbf{S}_0^{-1}\mathbf{m}_0\right] \tag{3.12}
$$

$$
\mathbf{S}_N = \left[\beta\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}_0^{-1}\right]^{-1}. \tag{3.13}
$$

Hence, via the method of completing the squares, we conclude that $\mathbf{w}|\mathrm{T}$ is a D-dimensional normal random variable with mean $\mathbf{m}_N \in \mathbb{R}^D$ as in (3.12) and covariance matrix $\mathbf{S}_N \in \mathbb{R}^{D\times D}$ as in (3.13).

Exercise 3.8

Consider the same framework as presented in Exercise 3.7, such that, after observing the first sample set, an additional data point, denoted by $\{t_{N+1}, \mathbf{x}_{N+1}\}$, is observed. Utilizing the posterior distribution of $\mathbf{w}$ as defined by the constants in (3.12) and (3.13) as a prior, we seek the posterior distribution after the inclusion of $(N+1)$-th data point. It follows that

$$
\begin{aligned}
p(\mathbf{w}|\mathrm{T}, t_{N+1}) &\propto p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathrm{T}) \\
&\propto \exp\left\{ -\frac{\beta}{2}(t_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_{N+1}))^2 \right\} \times \\
&\times \exp\left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) \right\} \qquad \text{(Apply (1.46) and (2.43))} \\
&\propto \exp\Bigg\{ \beta t_{N+1}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_{N+1}) + \\
&- \frac{\beta}{2}\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top \mathbf{w}^\top + \\
&- \frac{1}{2}\mathbf{w}^\top \mathbf{S}_N^{-1}\mathbf{w} + \mathbf{w}^\top \mathbf{S}_N^{-1}\mathbf{m}_N \Bigg\} \\
&= \exp\Bigg\{ \mathbf{w}^\top [\beta t_{N+1}\boldsymbol{\phi}(\mathbf{x}_{N+1}) + \mathbf{S}_N^{-1}\mathbf{m}_N] + \\
&- \frac{1}{2}\mathbf{w}^\top [\mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top]\mathbf{w}^\top \Bigg\} \\
p(\mathbf{w}|\mathrm{T}, t_{N+1}) &\propto \exp\left\{ -\frac{1}{2}(\mathbf{w} - \tilde{\mathbf{m}})^\top \tilde{\mathbf{S}}^{-1}(\mathbf{w} - \tilde{\mathbf{m}}) \right\}.
\end{aligned}
$$

By method of completing the squares, we find that, with the addition of the $(N+1)$-th data point, the posterior distribution of $\mathbf{w}$ is D-dimensional normal, with mean $\tilde{\mathbf{m}} \in \mathbb{R}^D$ and covariance $\tilde{\mathbf{S}} \in \mathbb{R}^{D\times D}$ determined as

$$
\begin{aligned}
\tilde{\mathbf{m}} &= \tilde{\mathbf{S}}_N[\beta t_{N+1}\boldsymbol{\phi}(\mathbf{x}_{N+1}) + \mathbf{S}_N^{-1}\mathbf{m}_N] \\
\tilde{\mathbf{S}} &= [\mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top]^{-1}.
\end{aligned}
$$

Lastly, we aim herein to demonstrate that $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{S}}$ may be rewritten as $\mathbf{m}_{N+1}$ and $\mathbf{S}_{N+1}$ respectively. It follows that

$$\begin{aligned}
\tilde{\mathbf{S}} &= [\mathbf{S}_N^{-1} + \beta\phi(\mathbf{x}_{N+1})\{\phi(\mathbf{x}_{N+1})\}^\top]^{-1} \\
&= [\beta\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}_0^{-1} + \beta\phi(\mathbf{x}_{N+1})\{\phi(\mathbf{x}_{N+1})\}^\top]^{-1} && \text{(Apply (3.13))} \\
&= \left[\beta\sum_{n=1}^{N}\phi(\mathbf{x}_n)\{\phi(\mathbf{x}_n)\}^\top + \mathbf{S}_0^{-1} + \beta\phi(\mathbf{x}_{N+1})\{\phi(\mathbf{x}_{N+1})\}^\top\right]^{-1} \\
&= \left[\beta\sum_{n=1}^{N+1}\phi(\mathbf{x}_n)\{\phi(\mathbf{x}_n)\}^\top + \mathbf{S}_0^{-1}\right]^{-1} \\
&= \left[\beta\mathbf{\Phi}_{N+1}^\top\mathbf{\Phi}_{N+1} + \mathbf{S}_0^{-1}\right]^{-1}
\end{aligned}$$

$$\tilde{\mathbf{S}} = \mathbf{S}_{N+1} \qquad \text{(Apply (3.13)).} \tag{3.14}$$

And

$$\begin{aligned}
\tilde{\mathbf{m}} &= \tilde{\mathbf{S}}_N[\beta t_{N+1}\phi(\mathbf{x}_{N+1}) + \mathbf{S}_N^{-1}\mathbf{m}_N] \\
&= \mathbf{S}_{N+1}[\beta t_{N+1}\phi(\mathbf{x}_{N+1}) + \beta\mathbf{\Phi}^\top\mathbf{t} + \mathbf{S}_0^{-1}\mathbf{m}_0] && \text{(Apply (3.12))} \\
&= \mathbf{S}_{N+1}\left[\beta t_{N+1}\phi(\mathbf{x}_{N+1}) + \beta\sum_{n=1}^{N}\phi(\mathbf{x}_n)t_n + \mathbf{S}_0^{-1}\mathbf{m}_0\right] \\
&= \mathbf{S}_{N+1}\left[\beta\sum_{n=1}^{N+1}\phi(\mathbf{x}_n)t_n + \mathbf{S}_0^{-1}\mathbf{m}_0\right] \\
&= \mathbf{S}_{N+1}\left[\beta\mathbf{\Phi}_{N+1}^\top\mathbf{t}_{N+1} + \mathbf{S}_0^{-1}\mathbf{m}_0\right] \\
&= \mathbf{S}_{N+1}\left[\beta\mathbf{\Phi}_{N+1}^\top\mathbf{t}_{N+1} + \mathbf{S}_0^{-1}\mathbf{m}_0\right]
\end{aligned}$$

$$\tilde{\mathbf{m}} = \mathbf{m}_{N+1} \qquad \text{(Apply (3.12)).}$$

Hence concluding that the posterior maintains its functional form after the addition of the $(N+1)$-th data point.

Exercise 3.9

We consider now the same framework as that which is studied in Exercise 3.8, yet now we aim to determine the posterior distribution utilizing the tools provided in the linear-Gaussian model. Consider the following hierarchical model

$$p(\mathbf{w}|\mathrm{T}) = \textsc{Multivariate Normal}(\mathbf{m}_N, \mathbf{S}_N)$$
$$p(t_{N+1}|\mathbf{w}) = \textsc{Normal}(\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top \mathbf{w}, \beta^{-1}).$$

Utilizing results seen in (2.113), (2.114) and (2.116) for the linear-Gaussian model, we conclude that the posterior distribution $p(\mathbf{w}|\mathbf{t}, t_{N+1})$ is such that

$$p(\mathbf{w}|\mathrm{T}, t_{N+1}) = \textsc{Multivariate Normal}(\mathbf{m}^\star, \mathbf{S}^\star)$$

where

$$\mathbf{m}^\star = \mathbf{S}^\star[\beta\boldsymbol{\phi}(\mathbf{x}_{N+1})t_{N+1} + \mathbf{S}_N^{-1}\mathbf{m}_N]$$
$$\mathbf{S}^\star = [\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top]^{-1}.$$

Note that the formulae for $\mathbf{m}^\star$ and $\tilde{\mathbf{m}}$ (also $\mathbf{S}^\star$ and $\tilde{\mathbf{S}}$) are equivalent in this Exercise and Exercise 3.8. Hence, we conclude both approaches yield the same result.

Exercise 3.10

We again utilize the linear-Gaussian model framework to analyse the Bayesian linear regression model, under the same framework as seen in Exercise 3.8. Consider the following hierarchical model:

$$\begin{aligned} p(\mathbf{w}|\mathrm{T}) &= \textsc{Multivariate Normal}(\mathbf{m}_N, \mathbf{S}_N) \\ p(t|\mathbf{w}) &= \textsc{Normal}(\{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{w}, \beta^{-1}). \end{aligned}$$

Utilizing results seen in (2.113), (2.114) and (2.115) for the linear-Gaussian model, we conclude that the predictive distribution $p(t)$ is such that

$$p(t) = \textsc{Normal}(\{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{m}_N, \beta^{-1} + \{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})).$$

Note that the variance of t is dependent on the input value, and is defined as

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \tag{3.15}$$

Exercise 3.11

We aim herein to demonstrate that $\sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x})$, where these values denote the input-dependent variance in (3.15). Note, from (3.14), that

$$\mathbf{S}_{N+1} = [\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top]^{-1}$$

$$\mathbf{S}_{N+1} = \mathbf{S}_N - \frac{\beta\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N}{1+\beta\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})} \qquad \text{(Apply (2.289))}. \tag{3.16}$$

It follows that

$$\begin{aligned}
\sigma^2_{N+1}(\mathbf{x}) &= \frac{1}{\beta} + \{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x}) && \text{(Apply (3.15))}\\
&= \frac{1}{\beta} + \{\boldsymbol{\phi}(\mathbf{x})\}^\top\left[\mathbf{S}_N - \frac{\beta\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N}{1+\beta\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})}\right]\boldsymbol{\phi}(\mathbf{x}) && \text{(Apply (3.16))}\\
&= \frac{1}{\beta} + \{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) - \frac{\beta[\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})]^2}{1+\beta\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})}\\
\sigma^2_{N+1}(\mathbf{x}) &= \sigma^2_N(\mathbf{x}) - \frac{\beta[\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})]^2}{1+\beta\{\boldsymbol{\phi}(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})} && \text{(Apply (3.15))}.
\end{aligned}$$

Note that the second term in the right-hand-side is non-positive as, presumably, $\mathbf{S}_N$ is a positive-definite real and symmetric matrix, hence $\mathbf{a}^\top\mathbf{S}_N\mathbf{a} > 0$ for all $\mathbf{a} \in \mathbb{R}^D$ (as was demonstrated in Exercise 2.20). We thereby conclude that

$$\sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x}).$$

Exercise 3.12

Let a sample set of target variables $\{t_n\}_{n=1}^N$ be observed, such that, conditional on $\mathbf{w}$ and β, these random variables are normally distributed, with mean $\mathbf{w}^\top\phi(\mathbf{x}_n) \in \mathbb{R}$ and precision $\beta > 0$, where $\{\mathbf{x}_n\}_{n=1}^N$ denote the corresponding input variables. Let also $\mathbf{w}$, conditioned on β, be a D-dimensional multivariate normal random variable, with mean $\mathbf{m}_0 \in \mathbb{R}^D$ and covariance matrix $\beta^{-1}\mathbf{S}_0 \in \mathbb{R}^{D\times D}$, and β be a Gamma random variable, with parameters $a_0 >$ and $b > 0$. We aim herein to determine the posterior distribution of $\mathbf{w}$ and β. It follows that

$$
\begin{aligned}
p(\mathbf{w},\beta|\mathrm{T}) &\propto p(\mathrm{T}|\mathbf{w},\beta)p(\mathbf{w}|\beta)p(\beta)\\
&\propto \beta^{N/2}\exp\left\{-\frac{\beta}{2}\sum_{n=1}^N(t_n-\mathbf{w}^\top\phi(\mathbf{x}_n))^2\right\}\times\\
&\times \beta^{D/2}\exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^\top\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}\times\\
&\times \beta^{a_0-1}\exp\{-b_0\beta\} \qquad \text{(Apply (2.43), (2.137) and (2.146))}\\
&\propto \beta^{(D+N)/2+a_0-1}\exp\left\{-\frac{\beta}{2}\sum_{n=1}^N t_n^2+\right.\\
&+\beta\sum_{n=1}^N t_n\mathbf{w}^\top\phi(\mathbf{x}_n)+\\
&-\frac{\beta}{2}\sum_{n=1}^N\mathbf{w}^\top\phi(\mathbf{x}_n)\{\phi(\mathbf{x}_n)\}^\top\mathbf{w}+\\
&-\frac{\beta}{2}\mathbf{m}_0^\top\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\mathbf{w}^\top\mathbf{S}_0^{-1}\mathbf{m}_0+\\
&\left.-\frac{\beta}{2}\mathbf{w}^\top\mathbf{S}_0^{-1}\mathbf{w}-b_0\beta\right\}\\
&= \beta^{(D+N)/2+a_0-1}\exp\left\{\beta\mathbf{w}^\top[\mathbf{S}_0^{-1}\mathbf{m}_0+\mathbf{\Phi}^\top\mathbf{t}]+\right.\\
&-\frac{\beta}{2}\mathbf{w}^\top[\mathbf{\Phi}^\top\mathbf{\Phi}+\mathbf{S}_0^{-1}]\mathbf{w}+\\
&\left.-\beta\left(\frac{1}{2}\mathbf{t}^\top\mathbf{t}+\frac{1}{2}\mathbf{m}_0^\top\mathbf{S}_0^{-1}\mathbf{m}_0+b_0\right)\right\}\\
p(\mathbf{w},\beta|\mathrm{T}) &= \beta^{D/2}\exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\}\times\\
&\times \beta^{N/2+a_0-1}\exp\left\{-\beta\left(\frac{1}{2}\mathbf{t}^\top\mathbf{t}+\right.\right.\\
&\left.\left.+\frac{1}{2}\mathbf{m}_0^\top\mathbf{S}_0^{-1}\mathbf{m}_0-\frac{1}{2}\mathbf{m}_N^\top\mathbf{S}_N^{-1}\mathbf{m}_N+b_0\right)\right\},
\end{aligned}
$$

where

$$\mathbf{m}_N = \mathbf{S}_N[\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^\top \mathbf{t}] \tag{3.17}$$

$$\mathbf{S}_N = (\mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{S}_0^{-1})^{-1} \tag{3.18}$$

$$a_N = \frac{N}{2} + a_0 \tag{3.19}$$

$$b_N = \frac{1}{2}\mathbf{t}^\top \mathbf{t} + \frac{1}{2}\mathbf{m}_0^\top \mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{1}{2}\mathbf{m}_N^\top \mathbf{S}_N^{-1}\mathbf{m}_N + b_0. \tag{3.20}$$

We thereby conclude that the posterior distribution of β is Gamma, with parameters $a_N > 0$ and $b_N > 0$ as defined in (3.19) and (3.20), whilst the posterior distribution of $\mathbf{w}$, conditioned on β, is a D-dimensional multivariate normal with mean $\mathbf{m}_N \in \mathbb{R}^D$ and covariance matrix $\mathbf{S}_N \in \mathbb{R}^{D\times D}$, as defined in (3.17) and (3.18).

Exercise 3.13

We consider herein the same sample set framework as in Exercise 3.12, and desire to determine the predictive distribution of a new target variable t, with associated input variable $\mathbf{x}$. In order to marginalize the distribution over $\mathbf{w}|\mathrm{T}$, we utilize the linear-Gaussian model framework to analyse the Bayesian linear regression model. Consider the following hierarchical model:

$$\begin{aligned} p(\mathbf{w}|\mathrm{T},\beta) &= \textsc{Multivariate Normal}(\mathbf{m}_N, \beta^{-1}\mathbf{S}_N) \\ p(t|\mathbf{w},\beta) &= \textsc{Normal}(\{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{w}, \beta^{-1}). \end{aligned}$$

Utilizing results seen in (2.113), (2.114) and (2.115) for the linear-Gaussian model, we conclude that the distribution $p(t|\mathrm{T},\beta)$ is such that

$$p(t|\mathrm{T},\beta) = \textsc{Normal}(\{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{m}_N, \beta^{-1} + \beta^{-1}\{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})).$$

We now marginalize this distribution over $\beta|\mathrm{T}$ as follows

$$\begin{aligned} p(t|\mathrm{T}) &= \int_0^\infty p(t|\mathrm{T},\beta)p(\beta|\mathrm{T})\,\mathrm{d}\beta && \text{(Apply (1.32))} \\ &= \int_0^\infty \frac{\exp\left\{-\frac{(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{2(\beta^{-1}+\beta^{-1}\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right\}}{\sqrt{2\pi}(\beta^{-1}+\beta^{-1}\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^{1/2}} \times \\ &\times \frac{b_N^{a_N}}{\Gamma(a_N)}\beta^{a_N-1}\exp\{-b_N\beta\}\,\mathrm{d}\beta && \text{(Apply (1.46) and (2.146))} \\ &= \int_0^\infty \frac{\exp\left\{-\left[b_N+\frac{(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{2(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right]\beta\right\}}{\sqrt{2\pi}(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^{1/2}} \times \\ &\times \frac{b_N^{a_N}}{\Gamma(a_N)}\beta^{a_N+1/2-1}\,\mathrm{d}\beta \\ &= \frac{b_N^{a_N}}{\Gamma(a_N)}\frac{\Gamma(a_N+1/2)}{\left[b_N+\frac{(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{2(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right]^{a_N+1/2}} \times \\ &\times \int_0^\infty \frac{\exp\left\{-\left[b_N+\frac{(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{2(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right]\beta\right\}}{\sqrt{2\pi}(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^{1/2}} \times \\ &\times \frac{\left[b_N+\frac{(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{2(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right]}{\Gamma(a_N+1/2)}\beta^{a_N+1/2-1}\,\mathrm{d}\beta \\ &= \frac{b_N^{a_N}}{\Gamma(a_N)}\frac{\Gamma(a_N+1/2)}{\sqrt{2\pi}(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^{1/2}} \\ &\times \left[b_N+\frac{(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{2(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right]^{-(a_N+1/2)} && \text{(Apply (1.26))} \\ p(t|\mathrm{T}) &= \frac{\sqrt{2a_N}}{\Gamma(a_N)}\frac{\Gamma(a_N+1/2)}{\sqrt{4a_Nb_N\pi(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}} \times \\ &\times \left[1+\frac{2a_N(t-\mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}))^2}{4a_Nb_N(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right]^{-(a_N+1/2)}. \end{aligned}$$

Hence, we conclude that the predictive distribution of t is a Student's t, with degrees of freedom $\nu = 2a_N$, precision $\lambda = (a_N)/[b_N(1+\{\boldsymbol{\phi}(\mathbf{x})\}^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}))]$ and mean $\mu = \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x})$.

Exercise 3.14

Consider the usual context of Bayesian linear regression with predictors of the form $\phi(\mathbf{x})$. Let a sample set composed of target variables $\{t_n\}_{n=1}^N$ and input variables $\{\mathbf{x}_n\}_{n=1}^N$ be observed, such that we can construct a new basis set $\psi(\mathbf{x})$ which is orthonormal, i.e., such that (3.115) is satisfied. Note moreover that $\psi_0(\mathbf{x}) = 1/\sqrt{N}$. Take $\alpha = 0$ in (3.54), and consider the consequent equivalent kernel as (3.62). We may rewrite it as

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}^\star) &= \beta\{\boldsymbol{\psi}(\mathbf{x})\}^\top \mathbf{S}_N \boldsymbol{\psi}(\mathbf{x}^\star) \\ &= \beta\{\boldsymbol{\psi}(\mathbf{x})\}^\top (\beta\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1} \boldsymbol{\psi}(\mathbf{x}^\star) && \text{(Apply (3.54))} \\ &= \{\boldsymbol{\psi}(\mathbf{x})\}^\top (\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1} \boldsymbol{\psi}(\mathbf{x}^\star) \\ k(\mathbf{x}, \mathbf{x}^\star) &= \{\boldsymbol{\psi}(\mathbf{x})\}^\top \boldsymbol{\psi}(\mathbf{x}^\star) && \text{(Orthornormality of } \boldsymbol{\Psi}\text{).} \end{aligned} \tag{3.21}$$

It follows that

$$\begin{aligned} \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) &= \sum_{n=1}^N \{\boldsymbol{\psi}(\mathbf{x})\}^\top \boldsymbol{\psi}(\mathbf{x}_n) && \text{(Apply (3.21))} \\ &= \sum_{n=1}^N \sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) \\ &= \sum_{n=1}^N \sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) \frac{1}{\sqrt{N}}\sqrt{N} \\ &= \sqrt{N} \sum_{n=1}^N \sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n)\psi_0(\mathbf{x}_n) \\ &= \sqrt{N} \sum_{j=0}^{M-1} \psi_j(\mathbf{x}) \left[\sum_{n=1}^N \psi_j(\mathbf{x}_n)\psi_0(\mathbf{x}_n) \right] \\ &= \sqrt{N} \sum_{j=0}^{M-1} \psi_j(\mathbf{x}) I_{j,0} && \text{(Apply (3.115))} \\ &= \sqrt{N} \sum_{j=1}^{M-1} \psi_j(\mathbf{x}) I_{j,0} + \sqrt{N}\psi_0(\mathbf{x}) I_{0,0} \\ &= \sqrt{N} \frac{1}{\sqrt{N}} \\ \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) &= 1. \end{aligned}$$

Exercise 3.15

Consider that we are utilizing the empirical Bayes' method do determine the values of the hyperparameters $\alpha > 0$ and $\beta > 0$ in Bayesian linear regression, such that (3.92) and (3.95) are satisfied. Consider therefore the $E(\mathbf{m}_N)$ term as in (3.82). It follows that, in light of the results in (3.92) and (3.95), we may rewrite (3.82) as

$$\begin{aligned}
E(\mathbf{m}_N) &= \frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}_n)\} + \frac{\alpha}{2}\mathbf{m}_N^\top \mathbf{m}_N \\
&= \frac{1}{2}\left[\frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n - \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}_n)\}\right]^{-1} \times \\
&\times \sum_{n=1}^{N}\{t_n - \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}_n)\} + \frac{\gamma}{2\mathbf{m}_N^\top \mathbf{m}_N}\mathbf{m}_N^\top \mathbf{m}_N \quad \text{(Apply (3.92) and (3.95))} \\
&= \frac{N-\gamma}{2} + \frac{\gamma}{2} \\
&= \frac{N}{2}.
\end{aligned}$$

Hence, we conclude that, if $\alpha > 0$ and $\beta > 0$ are estimated under the empirical Bayes' framework, it follows that $2E(\mathbf{m}_N) = N$.

Exercise 3.16

We return herein to the linear-Gaussian model in order to determine the marginal distribution of a sample set $\{t_n\}_{n=1}^N$. Consider the following hierarchical model

$$\begin{aligned} p(\mathbf{w}|\alpha) &= \textsc{Multivariate Normal}(\mathbf{0}, \alpha^{-1}\boldsymbol{I}) \\ p(\mathbf{t}|\mathbf{w}, \beta) &= \textsc{Multivariate Normal}(\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\boldsymbol{I}). \end{aligned}$$

We conclude, from (2.113), (2.114) and (2.115), that

$$p(\mathbf{t}|\alpha, \beta) = \textsc{Multivariate Normal}(\mathbf{0}, \beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top).$$

It follows that the logarithm of the marginal likelihood function associated with the observed data set is

$$\begin{aligned} \log p(\mathbf{t}|\alpha, \beta) = &-\frac{1}{2}\mathbf{t}^\top(\beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}\mathbf{t} - \frac{N}{2}\log(2\pi) + \\ &- \frac{1}{2}\log|\beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top|. \end{aligned} \tag{3.22}$$

Let us briefly examine the matrix form of the terms in (3.82):

$$
\begin{aligned}
E(\mathbf{m}_N) &= \frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N||^2 + \frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N \\
&= \frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)^\top(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N) + \frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N \\
&= \frac{\beta}{2}\mathbf{t}^\top(\boldsymbol{I}-\beta\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^\top)^\top(\boldsymbol{I}-\beta\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^\top)\mathbf{t} + \\
&+ \frac{\alpha\beta^2}{2}\mathbf{t}^\top\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{t} && \text{(Apply (3.84))} \\
&= \frac{\beta}{2}\mathbf{t}^\top\Big\{(\boldsymbol{I}-\beta\mathbf{\Phi}[\alpha\boldsymbol{I}+\beta\mathbf{\Phi}^\top\mathbf{\Phi}]^{-1}\mathbf{\Phi}^\top)^\top\times \\
&\times(\boldsymbol{I}-\beta\mathbf{\Phi}[\alpha\boldsymbol{I}+\beta\mathbf{\Phi}^\top\mathbf{\Phi}]^{-1}\mathbf{\Phi}^\top) + \\
&+ \alpha\beta\mathbf{\Phi}[\alpha(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}^\top\mathbf{\Phi})]^{-2}\mathbf{\Phi}^\top\Big\}\mathbf{t} && \text{(Apply (3.81))} \\
&= \frac{\beta}{2}\mathbf{t}^\top\Big\{(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-2} + \\
&+ \frac{\beta}{\alpha}\mathbf{\Phi}(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}^\top\mathbf{\Phi})^{-2}\mathbf{\Phi}^\top\Big\}\mathbf{t} && \text{(Apply (2.289))} \\
&= \frac{\beta}{2}\mathbf{t}^\top\Big\{(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-2} + \\
&+ \frac{\beta}{\alpha}\mathbf{\Phi}(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\times \\
&\times(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\Big\}\mathbf{t} \\
&= \frac{\beta}{2}\mathbf{t}^\top\Big\{(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-2} + \frac{\beta}{\alpha}(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\mathbf{\Phi}\times \\
&\times\mathbf{\Phi}^\top(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\Big\}\mathbf{t} && \text{(Apply (C.6))} \\
&= \frac{\beta}{2}\mathbf{t}^\top\Big\{(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)\times \\
&\times(\boldsymbol{I}+\alpha^{-1}\beta\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\Big\}\mathbf{t} \\
E(\mathbf{m}_N) &= \frac{1}{2}\mathbf{t}^\top\Big\{(\beta^{-1}\boldsymbol{I}+\alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\Big\}\mathbf{t}. && (3.23)
\end{aligned}
$$

Let us now examine the form of $1/2 \log|\beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top|$. It follows that

$$\begin{aligned}
\frac{1}{2}\log|\beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top| &= \frac{1}{2}\log|\beta^{-1}(\boldsymbol{I} + \alpha^{-1}\beta\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)| \\
&= -\frac{N}{2}\log\beta + \frac{1}{2}\log|\boldsymbol{I} + \alpha^{-1}\beta\boldsymbol{\Phi}\boldsymbol{\Phi}^\top| \\
&= -\frac{N}{2}\log\beta + \frac{1}{2}\log|\boldsymbol{I} + \alpha^{-1}\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}| && \text{(Apply (C.14))} \\
&= -\frac{N}{2}\log\beta + \frac{1}{2}\log|\alpha^{-1}(\alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})| \\
&= -\frac{N}{2}\log\beta - \frac{M}{2}\log\alpha + \\
&\quad + \frac{1}{2}\log|\alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}| \\
\frac{1}{2}\log|\beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top| &= -\frac{N}{2}\log\beta - \frac{M}{2}\log\alpha + \frac{1}{2}\log|\mathbf{A}| && \text{(Apply (3.81)).}
\end{aligned} \tag{3.24}$$

Substituting (3.23) and (3.24) into (3.22), we obtain

$$\log p(\mathbf{t}|\alpha,\beta) = -E(\mathbf{m}_N) - \frac{N}{2}\log(2\pi) + \frac{N}{2}\log\beta + \frac{M}{2}\log\alpha - \frac{1}{2}\log|\mathbf{A}|.$$

Exercise 3.17

We aim herein to demonstrate that (3.77) is equal to (3.78). It follows from (3.77) that

$$
\begin{aligned}
p(\mathbf{w}|\alpha,\beta) &= \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w} \\
&= \int \exp\{\log p(\mathbf{t}|\mathbf{w},\beta) + \log p(\mathbf{w}|\alpha)\}\,\mathrm{d}\mathbf{w} \\
&= \int \exp\left\{\frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \beta E_D(\mathbf{w}) + \right. \\
&\quad \left. - \frac{M}{2}\log(2\pi) + \frac{M}{2}\log\alpha - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}\right\}\mathrm{d}\mathbf{w} \qquad \text{(Apply (3.11) and (3.52))} \\
&= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2} \times \\
&\quad \times \int \exp\left\{-\beta E_D(\mathbf{w}) - \alpha E_W(\mathbf{w})\right\}\mathrm{d}\mathbf{w} \qquad \text{(Apply } E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top\mathbf{w}\text{)} \\
p(\mathbf{w}|\alpha,\beta) &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \qquad \text{(Apply (3.79)).}
\end{aligned}
$$

Thereby reaching the desired result.

Exercise 3.18

We aim to demonstrate, by completing the squares, that (3.79) may be rewritten as (3.80). See that

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi w}||^2 + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} \\
&= \frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi w})^\top(\mathbf{t}-\mathbf{\Phi w}) + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} \\
&= \frac{\beta}{2}\mathbf{t}^\top\mathbf{t} - \beta\mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{t} + \frac{\beta}{2}\mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Phi w} + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} \\
&= \frac{\beta}{2}\mathbf{t}^\top\mathbf{t} - \mathbf{w}^\top\mathbf{A}\mathbf{m}_N + \frac{1}{2}\mathbf{w}^\top(\alpha\boldsymbol{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi})\mathbf{w} && \text{(Apply (3.84))} \\
&= \frac{\beta}{2}\mathbf{t}^\top\mathbf{t} - \mathbf{w}^\top\mathbf{A}\mathbf{m}_N + \frac{1}{2}\mathbf{w}^\top\mathbf{A}\mathbf{w} && \text{(Apply (3.81))} \\
&= \frac{\beta}{2}\mathbf{t}^\top\mathbf{t} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{A}\mathbf{m}_N + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N) \\
&= \frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi m}_N)^\top(\mathbf{t}-\mathbf{\Phi m}_N) + \beta\mathbf{t}^\top\mathbf{\Phi m}_N - \frac{\beta}{2}\mathbf{m}_N^\top\mathbf{\Phi}^\top\mathbf{\Phi m}_N + \\
&- \frac{1}{2}\mathbf{m}_N^\top\mathbf{A}\mathbf{m}_N + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N) \\
&= \frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi m}_N)^\top(\mathbf{t}-\mathbf{\Phi m}_N) + \mathbf{m}_N^\top\mathbf{A}\mathbf{m}_N - \frac{\beta}{2}\mathbf{m}_N^\top\mathbf{\Phi}^\top\mathbf{\Phi m}_N + \\
&- \frac{1}{2}\mathbf{m}_N^\top\mathbf{A}\mathbf{m}_N + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N) && \text{(Apply (3.84))} \\
&= \frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi m}_N||^2 + \frac{1}{2}\mathbf{m}_N^\top(\alpha\boldsymbol{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi})\mathbf{m}_N + \\
&- \frac{\beta}{2}\mathbf{m}_N^\top\mathbf{\Phi}^\top\mathbf{\Phi m}_N + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N) && \text{(Apply (3.81))} \\
&= \frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi m}_N||^2 + \frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N) \\
E(\mathbf{w}) &= E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N) && \text{(Apply (3.82)).}
\end{aligned}
$$

Thereby reaching the desired result.

Exercise 3.19

We desire to integrate $\exp\{-E(w)\}$ within (3.78). It follows that

$$\begin{aligned}
\int \exp\{-E(w)\}\,\mathrm{d}\mathbf{w} &= \int \exp\Bigg\{ -E(\mathbf{m}_N) + \\
&\quad - \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top \mathbf{A}(\mathbf{w}-\mathbf{m}_N)\Bigg\}\,\mathrm{d}\mathbf{w} && \text{(Apply (3.80))} \\
&= \exp\{-E(\mathbf{m}_N)\}\times \\
&\quad \times \int \exp\left\{ -\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top \mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w} && \text{(Apply (3.80))} \\
&= \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}\exp\{-E(\mathbf{m}_N)\}\times \\
&\quad \times \int \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\times \\
&\quad \times \exp\left\{ -\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top \mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w}
\end{aligned}$$

$$\int \exp\{-E(w)\}\,\mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2} \qquad \text{(Apply (1.30)).} \tag{3.25}$$

Substituting the result (3.25) into (3.78), we find that

$$\begin{aligned}
p(\mathbf{t}|\alpha,\beta) &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \\
&= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2} \\
p(\mathbf{t}|\alpha,\beta) &= \alpha^{M/2}\left(\frac{\beta}{2\pi}\right)^{N/2}|\mathbf{A}|^{-1/2}\exp\{-E(\mathbf{m}_N)\} \\
\log p(\mathbf{t}|\alpha,\beta) &= \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{A}| - E(\mathbf{m}_N).
\end{aligned}$$

Thereby reaching the desired result.

Exercise 3.20

We aim now to arrive at the maximization procedure of the marginal likelihood with respect to α as in (3.92). First, we aim to determine the eigenvalues of $\mathbf{A}$ as in (3.81). Consider the eigendecomposition in (3.87), which is equivalently determined by a set of M homogeneous linear equations, as in

$$
\begin{aligned}
|\beta\mathbf{\Phi}^\top\mathbf{\Phi} - \lambda_i\boldsymbol{I}| &= 0 && \text{(Apply (C.30))}\\
|\beta\mathbf{\Phi}^\top\mathbf{\Phi} + \alpha\boldsymbol{I} - \alpha\boldsymbol{I} - \lambda_i\boldsymbol{I}| &= 0 &&\\
|\mathbf{A} - (\alpha+\lambda_i)\boldsymbol{I}| &= 0 && \text{(Apply (3.81)).}
\end{aligned} \tag{3.26}
$$

We thereby conclude that the eigenvalues of $\mathbf{A}$ are $\lambda_i + \alpha$. We thereby rewrite (3.86) as

$$
\begin{aligned}
\log p(\mathbf{t}|\alpha,\beta) &= \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{A}| - E(\mathbf{m}_N)\\
&= \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) +\\
&- \frac{1}{2}\log\prod_{i=1}^{M}(\alpha+\lambda_i) - E(\mathbf{m}_N) && \text{(Apply (C.47))}\\
\log p(\mathbf{t}|\alpha,\beta) &= \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) +\\
&- \frac{1}{2}\sum_{i=1}^{M}\log(\alpha+\lambda_i) - \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 - \frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N && \text{(Apply (3.82)).}
\end{aligned}
$$

Consider $\mathbf{m}_N$ fixed. In order to determine a maximum of $\log p(\mathbf{t}|\alpha,\beta)$ we first differentiate it with respect to α, and solve for ${}^{\mathrm{d}\log p(\mathbf{t}|\alpha,\beta)}/_{\mathrm{d}\alpha} = 0$, obtaining the following

$$
\begin{aligned}
\frac{\mathrm{d}\log p(\mathbf{t}|\alpha,\beta)}{\mathrm{d}\alpha} &= 0\\
\frac{M}{2\alpha} - \sum_{i=1}^{M}\frac{1}{2(\alpha+\lambda_i)} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{m}_N &= 0\\
M - \sum_{i=1}^{M}\frac{\alpha}{\alpha+\lambda_i} - \alpha\mathbf{m}_N^\top\mathbf{m}_N &= 0\\
\sum_{i=1}^{M}\left(1 - \frac{\alpha}{\alpha+\lambda_i}\right) &= \alpha\mathbf{m}_N^\top\mathbf{m}_N\\
\sum_{i=1}^{M}\frac{\lambda_i}{\alpha+\lambda_i} &= \alpha\mathbf{m}_N^\top\mathbf{m}_N\\
\frac{\gamma}{\mathbf{m}_N^\top\mathbf{m}_N} &= \alpha && \text{(Apply (3.91)).}
\end{aligned}
$$

Hence we obtain the re-estimation equation in (3.92).

Exercise 3.21

We aim to demonstrate the validity of (3.117) for an arbitrary real symmetric matrix $\mathbf{A}$. First, we consider the eigendecomposition of $\mathbf{A}$ as in (2.48). Consequently, we find that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\log|\mathbf{A}| &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\log\prod_{i=1}^{M}\lambda_i \quad \text{(Apply (C.47))}\\
&= \frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{i=1}^{M}\log\lambda_i\\
\frac{\mathrm{d}}{\mathrm{d}\alpha}\log|\mathbf{A}| &= \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}.
\end{aligned}$$

Now, in order to determine the value of $\mathrm{tr}(\mathbf{A}^{-1}\mathrm{d}/\mathrm{d}\alpha\mathbf{A})$, we perform as follows

$$\begin{aligned}
\mathbf{A} &= \sum_{i=1}^{M}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top && \text{(Apply (2.48))}\\
\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A} &= \sum_{i=1}^{M}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\mathbf{u}_i\mathbf{u}_i^\top + \sum_{i=1}^{M}\lambda_i\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\mathbf{u}_i\mathbf{u}_i^\top\right)\\
\mathbf{A}^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right) &= \mathbf{A}^{-1}\sum_{i=1}^{M}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\mathbf{u}_i\mathbf{u}_i^\top + \mathbf{A}^{-1}\sum_{i=1}^{M}\lambda_i\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\mathbf{u}_i\mathbf{u}_i^\top\right)\\
\mathbf{A}^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right) &= \left\{\sum_{j=1}^{M}\frac{1}{\lambda_j}\mathbf{u}_j\mathbf{u}_j^\top\right\}\left\{\sum_{i=1}^{M}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\mathbf{u}_i\mathbf{u}_i^\top\right\}+\\
&\quad+\left\{\sum_{j=1}^{M}\frac{1}{\lambda_j}\mathbf{u}_j\mathbf{u}_j^\top\right\}\left\{\sum_{i=1}^{M}\lambda_i\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\mathbf{u}_i\mathbf{u}_i^\top\right)\right\} && \text{(Apply (2.49))}\\
&= \sum_{j=1}^{M}\sum_{i=1}^{M}\frac{1}{\lambda_j}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\mathbf{u}_j(\mathbf{u}_j^\top\mathbf{u}_i)\mathbf{u}_i^\top+\\
&\quad+\sum_{j=1}^{M}\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_j}\mathbf{u}_j\mathbf{u}_j^\top\frac{\mathrm{d}\mathbf{u}_i}{\mathrm{d}\alpha}\mathbf{u}_i^\top+\sum_{j=1}^{M}\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_j}\mathbf{u}_j(\mathbf{u}_j^\top\mathbf{u}_i)\frac{\mathrm{d}\mathbf{u}_i^\top}{\mathrm{d}\alpha}\\
&= \sum_{j=1}^{M}\sum_{i=1}^{M}\frac{1}{\lambda_j}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\mathbf{u}_j I_{j,i}\mathbf{u}_i^\top+\\
&\quad+\sum_{j=1}^{M}\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_j}\mathbf{u}_j\mathbf{u}_j^\top\frac{\mathrm{d}\mathbf{u}_i}{\mathrm{d}\alpha}\mathbf{u}_i^\top+\sum_{j=1}^{M}\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_j}\mathbf{u}_j I_{j,i}\frac{\mathrm{d}\mathbf{u}_i^\top}{\mathrm{d}\alpha} && \text{(Apply (2.46))}\\
\mathrm{tr}\left\{\mathbf{A}^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)\right\} &= \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}\mathrm{tr}(\mathbf{u}_i^\top\mathbf{u}_i)+\\
&\quad+\sum_{j=1}^{M}\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_j}\mathrm{tr}\left\{\mathbf{u}_j^\top\frac{\mathrm{d}\mathbf{u}_i}{\mathrm{d}\alpha}(\mathbf{u}_i^\top\mathbf{u}_j)\right\}+\sum_{i=1}^{M}\mathrm{tr}\left\{\frac{\mathrm{d}\mathbf{u}_i^\top}{\mathrm{d}\alpha}\mathbf{u}_i\right\} && \text{(Apply (C.9))}\\
&= \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}+\sum_{i=1}^{M}\mathrm{tr}\left\{\mathbf{u}_i^\top\frac{\mathrm{d}\mathbf{u}_i}{\mathrm{d}\alpha}\right\}+\sum_{i=1}^{M}\mathrm{tr}\left\{\frac{\mathrm{d}\mathbf{u}_i^\top}{\mathrm{d}\alpha}\mathbf{u}_i\right\} && \text{(Apply (2.46))}\\
&= \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha}+\sum_{i=1}^{M}\mathrm{tr}\left\{\frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathbf{u}_i^\top\mathbf{u}_i)\right\}\\
\mathrm{tr}\left\{\mathbf{A}^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right)\right\} &= \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\mathrm{d}\lambda_i}{\mathrm{d}\alpha} && \text{(Apply (2.46)).}
\end{aligned}$$

Thereby concluding that (3.117) is valid. We return to (3.86), and seek to arrive at (3.92) applying this result directly. Assuming $\mathbf{m}_N$ is fixed, we differentiate (3.86) with respect to α as follows

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\log p(\mathbf{t}|\alpha,\beta) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[\frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) + \right. \\
&\quad \left. - \frac{1}{2}\log|\mathbf{A}| - E(\mathbf{m}_N)\right] \\
&= \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{m}_N - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\alpha}\log|\mathbf{A}| && \text{(Apply (3.82))} \\
&= \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{m}_N - \frac{1}{2}\mathrm{tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right) && \text{(Apply (3.117))} \\
&= \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{m}_N - \frac{1}{2}\mathrm{tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}[\alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]\right) \\
&= \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{m}_N - \frac{1}{2}\mathrm{tr}(\mathbf{A}^{-1}) && \text{(Apply (C.19))} \\
\frac{\mathrm{d}}{\mathrm{d}\alpha}\log p(\mathbf{t}|\alpha,\beta) &= \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^\top\mathbf{m}_N - \frac{1}{2}\sum_{i=1}^{M}\frac{1}{\alpha+\lambda_i} && \text{(Apply (C.48))}.
\end{aligned}$$

Note that above, we utilized the result that the eigenvalues of $\mathbf{A}$ are $\alpha + \lambda_i$, as seen in (3.26), in Exercise 3.20, consequently implying that the eigenvalues of $\mathbf{A}^{-1}$ are $(\alpha + \lambda_i)^{-1}$. By setting $\mathrm{d}\log p(\mathbf{t}|\alpha,\beta)/\mathrm{d}\alpha = 0$ and solving for α, we can trivially see that procedure is analogous to that which is performed in Exercise 3.20, hence we reach the desired conclusion.

Exercise 3.22

We aim to demonstrate that the re-estimation equation for β under the empirical Bayes' framework is as seen in (3.95). As we must differentiate (3.86) with respect to β, we will first determine the form of one of the corresponding components, as follows:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\log|\mathbf{A}| &= \mathrm{tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\beta}\mathbf{A}\right) && \text{(Apply (3.117))}\\
&= \mathrm{tr}\left([\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]^{-1}\frac{\mathrm{d}}{\mathrm{d}\beta}[\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]\right) && \text{(Apply (3.81))}\\
&= \mathrm{tr}\left([\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right) && \text{(Apply (C.19))}\\
&= \mathrm{tr}\Big(\beta^{-1}[\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]^{-1}(\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})+\\
&\quad -[\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]^{-1}\alpha\beta^{-1}\boldsymbol{I}\Big)\\
&= \frac{1}{\beta}\mathrm{tr}(\boldsymbol{I})-\frac{\alpha}{\beta}\mathrm{tr}([\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}]^{-1})\\
&= \frac{M}{\beta}-\frac{\alpha}{\beta}\sum_{i=1}^{M}\frac{1}{\lambda_i+\alpha} && \text{(Apply (C.48))}\\
&= \frac{1}{\beta}\left[M-\sum_{i=1}^{M}\frac{\lambda_i+\alpha-\lambda_i}{\lambda_i+\alpha}\right]\\
&= \frac{1}{\beta}\left[M-\sum_{i=1}^{M}\frac{\lambda_i+\alpha}{\lambda_i+\alpha}+\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_i+\alpha}\right]\\
&= \frac{1}{\beta}\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_i+\alpha}
\end{aligned}$$

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\log|\mathbf{A}| = \frac{\gamma}{\beta} \qquad \text{(Apply (3.91)).} \tag{3.27}$$

Once again, for these calculations we utilized the fact that the eigenvalues associated with $\mathbf{A}^{-1}$ are of the form $\lambda_i+\alpha$, given the eigendecomposition seen in (3.87). Holding

$\mathbf{m}_N$ as fixed, we differentiate (3.86) with respect to β, as follows

$$\begin{aligned}
\frac{\mathrm{d}p(\mathbf{t}|\alpha,\beta)}{\mathrm{d}\beta} &= \frac{\mathrm{d}}{\mathrm{d}\beta}\left[\frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) + \right. \\
&\quad \left. - \frac{1}{2}\log|\mathbf{A}| - E(\mathbf{m}_N)\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}\beta}\left[\frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) + \right. \\
&\quad \left. - \frac{1}{2}\log|\mathbf{A}| - \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 - \frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N\right] \quad \text{(Apply (3.82))} \\
\frac{\mathrm{d}p(\mathbf{t}|\alpha,\beta)}{\mathrm{d}\beta} &= \frac{N}{2\beta} - \frac{\gamma}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\{t_N - \mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \qquad \text{(Apply (3.27))}.
\end{aligned}$$

Taking ${}^{\mathrm{d}p(\mathbf{t}|\alpha,\beta)}/_{\mathrm{d}\beta} = 0$ and solving for β, we obtain

$$\begin{aligned}
\frac{\mathrm{d}p(\mathbf{t}|\alpha,\beta)}{\mathrm{d}\beta} &= 0 \\
\frac{N}{2\beta} - \frac{\gamma}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\{t_N - \mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}_n)\}^2 &= 0 \\
\frac{N-\gamma}{\beta} &= \sum_{n=1}^{N}\{t_N - \mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \\
\frac{1}{\beta} &= \frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_N - \mathbf{m}_N^\top\boldsymbol{\phi}(\mathbf{x}_n)\}^2.
\end{aligned}$$

Thereby deriving the result in (3.95).

Exercise 3.23

We consider now the same framework as in Exercise 3.12, and hope to determine the marginal likelihood of our data given the model which is adopted. To prevent this work from becoming cluttered, we will first determine the marginal distribution $p(\mathbf{t}|\beta)$

$$
\begin{aligned}
p(\mathbf{t}|\beta) &= \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\beta)\,\mathrm{d}\mathbf{w} && \text{(Apply (1.31))}\\
&= \int \frac{\beta^{N/2}}{(2\pi)^{N/2}}\times \\
&\times \exp\left\{-\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi w})^\top(\mathbf{t}-\mathbf{\Phi w})\right\}\times \\
&\times \frac{\beta^{M/2}}{(2\pi)^{M/2}\mathbf{S}_0|^{1/2}}\times \\
&\times \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^\top\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\} && \text{(Apply (2.43))}\\
&= \frac{\beta^{N/2}|\mathbf{S}_N|^{1/2}}{(2\pi)^{N/2}|\mathbf{S}_0|^{1/2}}\times \\
&\times \exp\left\{-\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}-\frac{\beta}{2}\mathbf{m}_0\mathbf{S}_0\mathbf{m}_0+\frac{\beta}{2}\mathrm{m}_N^\top\mathbf{S}_N\mathbf{m}_N\right\}\times \\
&\times \int \frac{\beta^{M/2}}{(2\pi)^{M/2}|\mathbf{S}_N|^{1/2}}\times \\
&\times \exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w} && \text{(Apply (3.17) and (3.18))}
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{t}|\beta) &= \frac{\beta^{N/2}|\mathbf{S}_N|^{1/2}}{(2\pi)^{N/2}|\mathbf{S}_0|^{1/2}}\times \\
&\times \exp\left\{-\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}-\frac{\beta}{2}\mathbf{m}_0^\top\mathbf{S}_0\mathbf{m}_0+\frac{\beta}{2}\mathrm{m}_N^\top\mathbf{S}_N\mathbf{m}_N\right\} && \text{(Apply (1.30)).}
\end{aligned} \tag{3.28}
$$

It follows thereafter that the marginal distribution $p(\mathbf{t})$ is given as follows

$$
\begin{aligned}
p(\mathbf{t}) &= \int_0^\infty p(\mathbf{t}|\beta)p(\beta)\,\mathrm{d}\beta && \text{(Apply (1.31))}\\
&= \int_0^\infty \frac{b_0^{a_0}}{\Gamma(a_0)}\beta^{a_0-1}\exp\{-b_0\beta\}\times \\
&\times \frac{\beta^{N/2}|\mathbf{S}_N|^{1/2}}{(2\pi)^{N/2}|\mathbf{S}_0|^{1/2}}\times \\
&\times \exp\left\{-\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}-\frac{\beta}{2}\mathbf{m}_0^\top\mathbf{S}_0\mathbf{m}_0+\frac{\beta}{2}\mathrm{m}_N^\top\mathbf{S}_N\mathbf{m}_N\right\}\mathrm{d}\beta && \text{(Apply (2.146) and (3.28))}\\
&= \frac{1}{(2\pi)^{N/2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\times \\
&\times \int_0^\infty \frac{b_N^{a_N}}{\Gamma(a_N)}\beta^{a_N-1}\times \\
&\times \exp\left\{-b_0\beta-\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}-\frac{\beta}{2}\mathbf{m}_0^\top\mathbf{S}_0\mathbf{m}_0+\frac{\beta}{2}\mathrm{m}_N^\top\mathbf{S}_N\mathbf{m}_N\right\}\mathrm{d}\beta && \text{(Apply (3.19) and (3.20))}\\
p(\mathbf{t}) &= \frac{1}{(2\pi)^{N/2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} && \text{(Apply (1.26)).}
\end{aligned}
$$

Thereby deriving the desired result.

Exercise 3.24

We now aim the repeat Exercise 3.23, applying Bayes' theorem directly in order to determine the marginal distribution of $\mathbf{t}$. It is as follows

$$
\begin{aligned}
p(\mathbf{w},\beta|\mathbf{t}) &= \frac{p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w},\beta)}{p(\mathbf{t})} && \text{(Apply (1.12))}\\
p(\mathbf{t}) &= \frac{p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w},\beta)}{p(\mathbf{w},\beta|\mathbf{t})}\\
&= \frac{p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\beta)p(\beta)}{p(\mathbf{w}|\beta,\mathbf{t})p(\beta|\mathbf{t})} && \text{(Apply (1.32))}\\
&= \frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\beta)\exp\{(b_N-b_0)\beta\}}{\beta^{a_N-a_0}p(\mathbf{w}|\beta,\mathbf{t})} && \text{(Apply (2.146))}\\
&= \frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathrm{S}_N|^{1/2}}{|\mathrm{S}_0|^{1/2}}\frac{p(\mathbf{t}|\mathbf{w},\beta)}{\beta^{a_N-a_0}}\times\\
&\quad\times\exp\left\{\left(b_N-b_0-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^\top\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)+\right.\right.\\
&\quad\left.\left.+\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right)\beta\right\} && \text{(Apply (2.43))}\\
&= \frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathrm{S}_N|^{1/2}}{|\mathrm{S}_0|^{1/2}}\frac{p(\mathbf{t}|\mathbf{w},\beta)}{\beta^{N/2}}\times\\
&\quad\times\exp\left\{\left(\frac{1}{2}\mathbf{t}^\top\mathbf{t}+\frac{1}{2}\mathbf{w}^\top(\mathbf{S}_N^{-1}-\mathbf{S}_0^{-1})\mathbf{w}+\right.\right.\\
&\quad\left.\left.+\mathbf{w}^\top(\mathbf{S}_0^{-1}\mathbf{m}_0-\mathbf{S}_N^{-1}\mathbf{m}_N)\right)\beta\right\} && \text{(Apply (3.19) and (3.20))}\\
&= \frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathrm{S}_N|^{1/2}}{|\mathrm{S}_0|^{1/2}}\frac{p(\mathbf{t}|\mathbf{w},\beta)}{\beta^{N/2}}\times\\
&\quad\times\exp\left\{\left(\frac{1}{2}\mathbf{t}^\top\mathbf{t}+\frac{1}{2}\mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{w}-\mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{t}\right)\beta\right\} && \text{(Apply (3.17) and (3.18))}\\
&= \frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathrm{S}_N|^{1/2}}{|\mathrm{S}_0|^{1/2}}\frac{p(\mathbf{t}|\mathbf{w},\beta)}{\beta^{N/2}}\times\\
&\quad\times\exp\left\{\frac{\beta}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^\top(\mathbf{t}-\mathbf{\Phi}\mathbf{w})\right\}\\
p(\mathbf{w},\beta|\mathbf{t}) &= \frac{1}{(2\pi)^{N/2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathrm{S}_N|^{1/2}}{|\mathrm{S}_0|^{1/2}} && \text{(Apply (2.43)).}
\end{aligned}
$$

Thereby reaching the same result as in Exercise 3.23.

Chapter 4
Linear Models for Classification

Exercise 4.1

Consider two sets of points $\{x_n\}_{n=1}^N$ and $\{y_m\}_{m=1}^M$ whose convex hulls, as defined in (4.156), intersect. As these intersect, we may take an element $\mathbf{z}$ which belongs to said intersection, such that

$$\mathbf{z} = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n \tag{4.1}$$

$$\mathbf{z} = \sum_{m=1}^{M} \beta_m \mathbf{y}_m, \tag{4.2}$$

for some $\sum_{n=1}^N \alpha_n = 1$, $\alpha_n \geq 0$ and $\sum_{m=1}^M \beta_m = 1$, $\beta_m \geq 0$. We will demonstrate, by contradiction, that $\{x_n\}_{n=1}^N$ and $\{y_m\}_{m=1}^M$ are not linearly separable. For that purpose, we assume they are linearly separable, which implies there exists a vector $\hat{\mathbf{w}}$ and scalar w_0 such that $\hat{\mathbf{w}}^\top \mathbf{x}_n + w_0 > 0$ for all $n \in \{1, \dots, N\}$ and $\hat{\mathbf{w}}^\top \mathbf{y}_m + w_0 < 0$ for all $m \in \{1, \dots, M\}$. If we left-multiply (4.1) by $\hat{\mathbf{w}}$ and sum w_0 to it, we find

$$\begin{aligned}
\hat{\mathbf{w}}^\top \mathbf{z} + w_0 &= \hat{\mathbf{w}}^\top \sum_{n=1}^{N} \alpha_n \mathbf{x}_n + w_0 \\
&= \sum_{n=1}^{N} \alpha_n \hat{\mathbf{w}}^\top \mathbf{x}_n + \sum_{n=1}^{N} \alpha_n w_0 \qquad \left(\text{Apply } \sum_{n=1}^{N} \alpha_n = 1\right) \\
&= \sum_{n=1}^{N} \alpha_n (\hat{\mathbf{w}}^\top \mathbf{x}_n + w_0)
\end{aligned}$$

$$\mathbf{z} > 0 \qquad \text{(By assumption).} \tag{4.3}$$

However, If we left-multiply (4.2) by $\hat{\mathbf{w}}$ and sum w_0 to it, we find

$$\begin{aligned}
\hat{\mathbf{w}}^\top \mathbf{z} + w_0 &= \hat{\mathbf{w}}^\top \sum_{m=1}^{N} \beta_m \mathbf{y}_m + w_0 \\
&= \sum_{m=1}^{M} \beta_m \hat{\mathbf{w}}^\top \mathbf{y}_m + \sum_{m=1}^{M} \beta_m w_0 \qquad \left(\text{Apply } \sum_{m=1}^{M} \beta_m = 1\right) \\
&= \sum_{m=1}^{M} \beta_m (\hat{\mathbf{w}}^\top \mathbf{y}_m + w_0)
\end{aligned}$$

$$\mathbf{z} < 0 \qquad \text{(By assumption).} \tag{4.4}$$

We therefore arrive at an contradiction: the points in the intersection of the convex hull may be shown to be positive or negative. We thereby conclude that, if the convex hull of the data points intersect, the data sets are not linearly separable. Proving that if the the data sets are linearly separable therefore the respective convex hulls do not intersect is tantamount to that which was done prior, i.e., can also be performed by contradiction. Assume that the data sets are linearly separable and that the respective convex hulls intersect. We may therefore arrive at the same contradiction as (4.3) and (4.4), thereby concluding that if the data sets are linearly separable, their respective convex hulls do not intersect.

Exercise 4.2

Consider the solution to the least squares problem (4.15) given by (4.17). We aim to demonstrate that if (4.18) holds for all target variables, it likewise holds for predictions (4.19). From (4.17), we rewrite our predictions as

$$\begin{aligned}
\mathbf{y}(\mathbf{x}) &= \tilde{\mathbf{W}}\phi(\mathbf{x}) \\
&= \mathbf{T}^\top \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\phi(\mathbf{x}) && \text{(Apply (4.16))} \\
&= \begin{pmatrix} \mathbf{t}_1 & \dots & \mathbf{t}_N \end{pmatrix} \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\phi(\mathbf{x}) \\
\mathrm{a}^\top\mathbf{y}(\mathbf{x}) &= \begin{pmatrix} \mathrm{a}^\top\mathbf{t}_1 + b - b & \dots & \mathrm{a}^\top\mathbf{t}_N + b - b \end{pmatrix} \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\phi(\mathbf{x}) \\
\mathrm{a}^\top\mathbf{y}(\mathbf{x}) &= -b\mathbf{1}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\phi(\mathbf{x}). && (4.5)
\end{aligned}$$

We rewrite our input variable matrix as

$$\boldsymbol{\Phi} = \begin{pmatrix} \mathbf{1} & \mathbf{P} \end{pmatrix}, \tag{4.6}$$

where $\mathbf{1}$ is a vector of length N composed of ones and $\mathbf{P}$ is such that

$$\mathbf{P} = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \dots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \dots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

We thereby may write

$$\begin{aligned}
\boldsymbol{\Phi}^\top\boldsymbol{\Phi} &= \begin{pmatrix} \mathbf{1}^\top \\ \mathbf{P}^\top \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{P} \end{pmatrix} && \text{(Apply (4.6))} \\
&= \begin{pmatrix} N & \mathbf{1}^\top\mathbf{P} \\ \mathbf{P}^\top\mathbf{1} & \mathbf{P}^\top\mathbf{P} \end{pmatrix} \\
(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} &= \begin{pmatrix} N & \mathbf{1}^\top\mathbf{P} \\ \mathbf{P}^\top\mathbf{1} & \mathbf{P}^\top\mathbf{P} \end{pmatrix}^{-1} \\
(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} &= \ell \begin{pmatrix} 1 & -\mathbf{1}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1} \\ -(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{1} & \ell^{-1}(\mathbf{P}^\top\mathbf{P})^{-1} + (\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{1}\mathbf{1}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1} \end{pmatrix} && \text{(Apply (2.76))}.
\end{aligned}$$

Where $\ell = [N - \mathbf{1}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{1}]^{-1}$, as in (2.77), and $\mathbf{1}^\top\mathbf{1} = N$. Note that ℓ is a scalar value, therefore it commutes under multiplication. By left multiplying both sides by $\boldsymbol{\Phi}$, we obtain

$$\begin{aligned}
\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} &= \ell \begin{pmatrix} \{\mathbf{1} - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{1}\}^\top \\ \{-\mathbf{1}\mathbf{1}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1} + \ell^{-1}\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1} + \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\mathbf{1}\mathbf{1}^\top\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
&= \ell \begin{pmatrix} \{[I - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top]\mathbf{1}\}^\top \\ \{[\ell^{-1}I - \{I - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\}\mathbf{1}\mathbf{1}^\top]\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
\mathbf{1}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} &= \ell \begin{pmatrix} \mathbf{1}^\top[I - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top]\mathbf{1} \\ \{[\ell^{-1}\mathbf{1}^\top - \mathbf{1}^\top\{I - \mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\mathbf{P}^\top\}\mathbf{1}\mathbf{1}^\top]\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
&= \ell \begin{pmatrix} \ell^{-1} \\ \{[\ell^{-1}\mathbf{1}^\top - \ell^{-1}\mathbf{1}^\top]\mathbf{P}(\mathbf{P}^\top\mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
\mathbf{1}^\top\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} &= \begin{pmatrix} 1 & \mathbf{0}^\top \end{pmatrix}. && (4.7)
\end{aligned}$$

Consider now that we rewrite our input vector for prediction $\phi(\mathbf{x})$ as

$$\phi(\mathbf{x}) = \begin{pmatrix} \phi_0(\mathbf{x}) \\ \phi(\mathbf{x}) \end{pmatrix}. \tag{4.8}$$

Note that $\phi_0(\mathbf{x}) = 1$. It follows that, by substituting (4.7) and (4.8) into (4.5), we obtain

$$\begin{aligned} \mathrm{a}^\top \mathbf{y}(\mathbf{x}) &= -b \begin{pmatrix} 1 & \mathbf{0}^\top \end{pmatrix} \begin{pmatrix} 1 \\ \phi(\mathbf{x}) \end{pmatrix} \\ &= -b\{1 + \mathbf{0}^\top \phi(\mathbf{x})\} \\ &= -b \\ \mathrm{a}^\top \mathbf{y}(\mathbf{x}) + b &= 0. \end{aligned}$$

Thereby reaching the desired result.

Exercise 4.3

We consider now the same framework as that in Exercise 4.2, where observations are now subject to $q \leq K$ linear constraints, which may be equivalently written as

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

where $\mathbf{A} \in \mathbb{R}^{q \times K}$ and $\mathbf{b} \in \mathbb{R}^q$. We rewrite our predictions as

$$\begin{aligned}
\mathbf{y}(\mathbf{x}) &= \tilde{\mathbf{W}}\phi(\mathbf{x}) \\
&= \mathbf{T}^\top \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}\phi(\mathbf{x}) && \text{(Apply (4.16))} \\
&= \begin{pmatrix} \mathbf{t}_1 & \dots & \mathbf{t}_N \end{pmatrix} \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}\phi(\mathbf{x}) \\
\mathbf{A}\mathbf{y}(\mathbf{x}) &= \begin{pmatrix} \mathbf{A}\mathbf{t}_1 + \mathbf{b} - \mathbf{b} & \dots & \mathbf{A}\mathbf{t}_N + \mathbf{b} - \mathbf{b} \end{pmatrix} \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}\phi(\mathbf{x}) \\
\mathbf{A}\mathbf{y}(\mathbf{x}) &= -\mathbf{b}\mathbf{1}^\top \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}\phi(\mathbf{x}). && (4.9)
\end{aligned}$$

Proceeding analogously to Exercise 4.2, we may conclude by substituting (4.7) and (4.8) into (4.9), and obtaining

$$\begin{aligned}
\mathbf{A}\mathbf{y}(\mathbf{x}) &= -\mathbf{b} \begin{pmatrix} 1 & \mathbf{0}^\top \end{pmatrix} \begin{pmatrix} 1 \\ \phi(\mathbf{x}) \end{pmatrix} \\
&= -\mathbf{b}\{1 + \mathbf{0}^\top \phi(\mathbf{x})\} \\
&= -\mathbf{b} \\
\mathbf{A}\mathbf{y}(\mathbf{x}) + \mathbf{b} &= 0.
\end{aligned}$$

Thereby reaching the desired result.

Exercise 4.4

We seek to determine $\mathbf{w}$ which maximizes (4.22), constrained to $\mathbf{w}^\top \mathbf{w} = 1$. The corresponding Lagrangian, as in (E.4) is as follows

$$L(\mathbf{w}, \lambda) = \mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1) + \lambda(\mathbf{w}^\top \mathbf{w} - 1). \tag{4.10}$$

Differentiating (4.10) with respect to $\mathbf{w}$ and thereafter solving $\partial L(\mathbf{w},\lambda)/\partial \mathbf{w} = \mathbf{0}$ for $\mathbf{w}$, we determine that

$$\begin{aligned} \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} &= \mathbf{0} \\ (\mathbf{m}_2 - \mathbf{m}_1) + 2\lambda \mathbf{w} &= \mathbf{0} && \text{(Apply (C.19))} \\ \mathbf{w} &= -\frac{\mathbf{m}_2 - \mathbf{m}_1}{2\lambda}. && (4.11) \end{aligned}$$

Substituting into the constraint, we find

$$\begin{aligned} \mathbf{w}^\top \mathbf{w} - 1 &= 0 \\ \frac{(\mathbf{m}_2 - \mathbf{m}_1)^\top (\mathbf{m}_2 - \mathbf{m}_1)}{4\lambda^2} &= 1 && \text{(Apply (4.11))} \\ \lambda &= \frac{\sqrt{(\mathbf{m}_2 - \mathbf{m}_1)^\top (\mathbf{m}_2 - \mathbf{m}_1)}}{2}. && (4.12) \end{aligned}$$

By substituting (4.12) into (4.11), we find

$$\begin{aligned} \mathbf{w} &= \frac{\mathbf{m}_1 - \mathbf{m}_2}{\sqrt{(\mathbf{m}_2 - \mathbf{m}_1)^\top (\mathbf{m}_2 - \mathbf{m}_1)}} \\ &\propto \mathbf{m}_2 - \mathbf{m}_1. \end{aligned}$$

Exercise 4.5

We desire herein to rewrite (4.25) as (4.26), in order for the dependence of $J(\mathbf{w})$ on $\mathbf{w}$ to become explicit. Firstly, we study the form of the denominator of (4.25), rewriting it as follows

$$\begin{aligned}
s_1^2 + s_2^2 &= \sum_{n\in\mathcal{C}_1} (y_n - m_1)^2 + \sum_{n\in\mathcal{C}_2} (y_n - m_2)^2 && \text{(Apply (4.24))} \\
&= \sum_{n\in\mathcal{C}_1} (\mathbf{w}^\top \mathbf{x}_n - \mathbf{w}^\top \mathbf{m}_1)^2 + \\
&+ \sum_{n\in\mathcal{C}_2} (\mathbf{w}^\top \mathbf{x}_n - \mathrm{w}^\top \mathbf{m}_2)^2 && \text{(Apply (4.20) and (4.23))} \\
&= \sum_{n\in\mathcal{C}_1} \{\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_1)\}^2 + \\
&+ \sum_{n\in\mathcal{C}_2} \{\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_2)\}^2 \\
&= \sum_{n\in\mathcal{C}_1} \mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top \mathbf{w} + \\
&+ \sum_{n\in\mathcal{C}_2} \mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \mathbf{w} \\
&= \mathbf{w}^\top \left[\sum_{n\in\mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \right. \\
&\left. + \sum_{n\in\mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \right] \mathbf{w} \\
s_1^2 + s_2^2 &= \mathbf{w}^\top \mathbf{S}_\mathrm{W} \mathbf{w} && \text{(Apply (4.28)).} \qquad (4.13)
\end{aligned}$$

It follows that we may now rewrite (4.25) as

$$\begin{aligned}
J(\mathbf{w}) &= \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} && \text{(Apply (4.25))} \\
&= \frac{\{\mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1)\}^2}{\mathbf{w}^\top \mathbf{S}_\mathrm{W} \mathbf{w}} && \text{(Apply (4.13) and (4.22))} \\
&= \frac{\mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_\mathrm{W} \mathbf{w}} \\
J(\mathbf{w}) &= \frac{\mathbf{w}^\top \mathbf{S}_\mathrm{B} \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_\mathrm{W} \mathbf{w}} && \text{(Apply (4.27)).}
\end{aligned}$$

Exercise 4.6

Consider the framework wherein target variables belonging to class $\mathcal{C}_1$ are attributed the value N/N_1, whilst target variables belonging to class $\mathcal{C}_2$ are attributed the value $-N/N_2$. We aim to demonstrate that (4.33) is equivalent to (4.37). For that purpose

$$\begin{aligned}
\mathbf{0} &= \sum_{n=1}^{N}(\mathbf{w}^\top\mathbf{x}_n + w_0 - t_n)\mathbf{x}_n \\
\mathbf{0} &= \sum_{n=1}^{N}(\mathbf{w}^\top\mathbf{x}_n - \mathbf{w}^\top\mathbf{m} - t_n)\mathbf{x}_n && \text{(Apply (4.34))} \\
\mathbf{0} &= \sum_{n=1}^{N}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m})^\top\mathbf{w} - \sum_{n=1}^{N}t_n\mathbf{x}_n \\
\mathbf{0} &= \sum_{n=1}^{N}\mathbf{x}_n\left(\mathbf{x}_n - \frac{1}{N}\{N_1\mathbf{m}_1 + N_2\mathbf{m}_2\}\right)^\top\mathbf{w} + \\
&\quad - \sum_{n\in\mathcal{C}_1}\frac{N}{N_1}\mathbf{x}_n + \sum_{n\in\mathcal{C}_2}\frac{N}{N_2}\mathbf{x}_n && \text{(Apply (4.36))} \\
N(\mathbf{m}_1 - \mathbf{m}_2) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\left(N\mathbf{x}_n - N_1\mathbf{m}_1 - N_2\mathbf{m}_2\right)^\top\mathbf{w} && \text{(Apply (4.21))} \\
N(\mathbf{m}_1 - \mathbf{m}_2) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\left(N_1\mathbf{x}_n + N_2\mathbf{x}_n - N_1\mathbf{m}_1 - N_2\mathbf{m}_2\right)^\top\mathbf{w} \\
N(\mathbf{m}_1 - \mathbf{m}_2) &= \frac{N_1}{N}\sum_{n\in\mathcal{C}_1}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m}_1)^\top\mathbf{w} + \frac{N_1}{N}\sum_{n\in\mathcal{C}_2}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m}_1)^\top\mathbf{w} + \\
&\quad + \frac{N_2}{N}\sum_{n\in\mathcal{C}_2}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m}_2)^\top\mathbf{w} + \frac{N_2}{N}\sum_{n\in\mathcal{C}_1}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m}_2)^\top\mathbf{w} \\
N(\mathbf{m}_1 - \mathbf{m}_2) &= \frac{N_1}{N}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top\mathbf{w} + \frac{N_1}{N}\sum_{n\in\mathcal{C}_2}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m}_1)^\top\mathbf{w} + \\
&\quad + \frac{N_2}{N}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top\mathbf{w} + \frac{N_2}{N}\sum_{n\in\mathcal{C}_1}\mathbf{x}_n(\mathbf{x}_n - \mathbf{m}_2)^\top\mathbf{w} \\
&\quad + \frac{N_1}{N}\sum_{n\in\mathcal{C}_1}\mathbf{m}_1(\mathbf{x}_n - \mathbf{m}_1)^\top\mathbf{w} + \frac{N_2}{N}\sum_{n\in\mathcal{C}_2}\mathbf{m}_2(\mathbf{x}_n - \mathbf{m}_2)^\top\mathbf{w}.
\end{aligned}$$

Continued:

$$
\begin{aligned}
N(\mathbf{m}_1 - \mathbf{m}_2) &= \mathbf{S}_{\mathrm{W}}\mathbf{w} - \frac{N_2}{N}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n\mathbf{x}_n^\top - \mathbf{x}_n\mathbf{m}_1^\top - \mathbf{m}_1\mathbf{x}_n^\top + \mathbf{m}_1\mathbf{m}_1^\top)\mathbf{w} \\
&\quad + \frac{N_1}{N}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n\mathbf{x}_n^\top - \mathbf{x}_n\mathbf{m}_1^\top)\mathbf{w} + \\
&\quad - \frac{N_1}{N}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n\mathbf{x}_n^\top - \mathbf{x}_n\mathbf{m}_2^\top - \mathbf{m}_2\mathbf{x}_n^\top + \mathbf{m}_2\mathbf{m}_2^\top)\mathbf{w} + \\
&\quad + \frac{N_2}{N}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n\mathbf{x}_n^\top - \mathbf{x}_n\mathbf{m}_2^\top)\mathbf{w} \\
&\quad + \frac{N_1}{N}\sum_{n\in\mathcal{C}_1}(\mathbf{m}_1\mathbf{x}_n^\top - \mathbf{m}_1\mathbf{m}_1^\top)\mathbf{w} + \\
&\quad + \frac{N_2}{N}\sum_{n\in\mathcal{C}_2}(\mathbf{m}_2\mathbf{x}_n^\top - \mathbf{m}_2\mathbf{m}_2^\top)\mathbf{w} && \text{(Apply (4.28))} \\
N(\mathbf{m}_1 - \mathbf{m}_2) &= \mathbf{S}_{\mathrm{W}}\mathbf{w} + \frac{N_1N_2}{N}(\mathbf{m}_1\mathbf{m}_1^\top + \mathbf{m}_1\mathbf{m}_1^\top - \mathbf{m}_1\mathbf{m}_1^\top)\mathbf{w} \\
&\quad - \frac{N_1N_2}{N}\mathbf{m}_2\mathbf{m}_1^\top\mathbf{w} + \\
&\quad + \frac{N_1N_2}{N}(\mathbf{m}_2\mathbf{m}_2^\top + \mathbf{m}_2\mathbf{m}_2^\top - \mathbf{m}_2\mathbf{m}_2^\top)\mathbf{w} + \\
&\quad - \frac{N_1N_2}{N}\mathbf{m}_1\mathbf{m}_2^\top\mathbf{w} \\
&\quad + \frac{N_1^2}{N}(\mathbf{m}_1\mathbf{m}_1^\top - \mathbf{m}_1\mathbf{m}_1^\top)\mathbf{w} + \\
&\quad + \frac{N_2^2}{N}(\mathbf{m}_2\mathbf{m}_2^\top - \mathbf{m}_2\mathbf{m}_2^\top)\mathbf{w} \\
&= \mathbf{S}_{\mathrm{W}}\mathbf{w} + \frac{N_1N_2}{N}(\mathbf{m}_2\mathbf{m}_2^\top - \mathbf{m}_2\mathbf{m}_1^\top - \mathbf{m}_1\mathbf{m}_2^\top + \mathbf{m}_1\mathbf{m}_1^\top)\mathbf{w} \\
&= \mathbf{S}_{\mathrm{W}}\mathbf{w} + \frac{N_1N_2}{N}\mathbf{S}_{\mathrm{B}}\mathbf{w} && \text{(Apply (4.27))} \\
\left(\mathbf{S}_{\mathrm{W}} + \frac{N_1N_2}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w} &= N(\mathbf{m}_1 - \mathbf{m}_2).
\end{aligned}
$$

Hence, we derive the desired result.

Exercise 4.7

Consider the logistic-sigmoid function in (3.6). First, we find that

$$\begin{aligned}
\sigma(a) &= \frac{1}{1+\exp\{-a\}} && \text{(Apply (3.6))} \\
1-\sigma(a) &= 1 - \frac{1}{1+\exp\{-a\}} \\
&= \frac{\exp\{-a\}}{1+\exp\{-a\}} \\
&= \frac{1}{1+\exp\{a\}} \\
1-\sigma(a) &= \sigma(-a) && \text{(Apply (3.6)).} \qquad (4.14)
\end{aligned}$$

We aim now to determine the inverse of (3.6). It follows that

$$\begin{aligned}
\sigma(\sigma^{-1}(y)) &= \frac{1}{1+\exp\{-\sigma^{-1}(y)\}} && \text{(Apply (3.6))} \\
y &= \frac{1}{1+\exp\{-\sigma^{-1}(y)\}} \\
1+\exp\{-\sigma^{-1}(y)\} &= \frac{1}{y} \\
\exp\{-\sigma^{-1}(y)\} &= \frac{1-y}{y} \\
-\sigma^{-1}(y) &= \log\left\{\frac{1-y}{y}\right\} \\
\sigma^{-1}(y) &= \log\left\{\frac{y}{1-y}\right\}. \qquad (4.15)
\end{aligned}$$

Exercise 4.8

We aim herein to demonstrate that, for the two class model with multivariate Gaussian densities with same covariance matrix Σ, it follows that the posterior probability (4.57) is computed as (4.65). From (4.58), it follows that the form of a in (4.57) is

$$\begin{aligned}
a &= \log \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \log p(\mathbf{x}|\mathcal{C}_1) - \log p(\mathbf{x}|\mathcal{C}_2) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\
&= -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| + \\
&- \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \\
&+ \frac{M}{2}\log(2\pi) + + \frac{1}{2}\log|\Sigma| + \\
&+ \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} && \text{(Apply (2.43))} \\
&= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1\Sigma^{-1}\boldsymbol{\mu}_1 + \\
&+ \frac{1}{2}\boldsymbol{\mu}_2\Sigma^{-1}\boldsymbol{\mu}_2 + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}
\end{aligned}$$

$$a = \mathbf{w}^\top \mathbf{x} + w_0 \qquad \text{(Apply (4.66) and (4.67)).} \tag{4.16}$$

By substituting (4.16) into (4.57), we obtain (4.65).

Exercise 4.9

Under the framework of a generative classification model with K classes, with respective prior probabilities $p(\mathcal{C}_k) = \pi_k$, and class-conditional densities $p(\phi|\mathcal{C}_k)$, suppose we observe a training data set $\{\phi_n, \mathbf{t}_n\}_{n=1}^N$, and seek to determine a maximum likelihood estimator for π_k. It follows that the likelihood function associated with this data, and its respective logarithm, are as follows

$$\begin{aligned} p(\mathbf{T}|\boldsymbol{\pi}) &= \prod_{n=1}^N \prod_{j=1}^K \pi_j^{t_{n,j}} \\ &= \prod_{j=1}^K \pi_j^{\sum_{n=1}^N t_{n,j}} \\ \log p(\mathbf{T}|\boldsymbol{\pi}) &= \sum_{j=1}^K \sum_{n=1}^N t_{n,j} \log \pi_j. \end{aligned}$$

Note that, as the parameters $\boldsymbol{\pi}$ are such that $\sum_{j=1}^K \pi_j = 1$, determining the maximum likelihood estimator must be performed as a constrained optimization problem. We define the corresponding Lagrangian, as in (E.4)

$$L(\boldsymbol{\pi}, \lambda) = \log p(\mathbf{T}|\boldsymbol{\pi}) + \lambda \cdot \left(\sum_{j=1}^K \pi_j - 1 \right). \tag{4.17}$$

We differentiate (4.17) with respect to π_k and solve $\partial L(\boldsymbol{\pi}, \lambda)/\partial \pi_k = 0$ for π_k, for arbitrary k, obtaining the following

$$\begin{aligned} \frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \pi_k} &= 0 \\ \frac{\sum_{n=1}^N t_{n,k}}{\pi_k} + \lambda &= 0 \\ \pi_k &= -\frac{N_k}{\lambda}. \end{aligned} \tag{4.18}$$

Substituting (4.18) into the constraint $\sum_{j=1}^K \pi_j = 1$, we find

$$\begin{aligned} \sum_{j=1}^K \pi_j - 1 &= 0 \\ -\sum_{j=1}^K \frac{N_j}{\lambda} &= 1 \\ \lambda &= -N. \end{aligned} \tag{4.19}$$

Hence, substituting (4.19) into (4.18), we find the maximum likelihood estimators of π_k are of the form

$$\pi_k^{\text{ML}} = \frac{N_k}{N}.$$

Exercise 4.10

Consider the same framework as in Exercise 4.9, under the added information that the class-conditional densities $p(\phi|\mathcal{C}_k)$ are multivariate Normal with same covariance matrix Σ and varying mean μ_k. It follows that the likelihood function associated with said data, and corresponding logarithm, is

$$p(\mathbf{T}|\pi,\mu,\Sigma)=\prod_{n=1}^{N}\prod_{j=1}^{K}\left[\frac{\pi_j}{(2\pi)^{M/2}|\Sigma|^{1/2}}\times\right.$$
$$\left.\times\exp\left\{-\frac{1}{2}(\phi_n-\mu_j)^\top\Sigma^{-1}(\phi_n-\mu_j)\right\}\right]^{t_{n,j}} \qquad \text{(Apply (2.43))}$$

$$\log p(\mathbf{T}|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\sum_{j=1}^{K}t_{n,j}\left[\log\pi_j-\frac{M}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma|\right.$$
$$\left.-\frac{1}{2}(\phi_n-\mu_j)^\top\Sigma^{-1}(\phi_n-\mu_j)\right]. \tag{4.20}$$

In order to determine the maximum likelihood estimator of μ_k, for arbitrary k, we differentiate (4.20) with respect to μ_k, equal the result to 0 and solve for μ_k, as follows

$$\frac{\partial\log p(\mathbf{T}|\pi,\mu,\Sigma)}{\partial\mu_k}=\mathbf{0}$$
$$\sum_{n=1}^{N}t_{n,k}\left[-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}\mu_k}\left\{\phi_n^\top\Sigma^{-1}\phi_n-2\phi_n^\top\Sigma^{-1}\mu_k+\mu_k\Sigma^{-1}\mu_k\right\}\right]=\mathbf{0}$$
$$-2\sum_{n=1}^{N}t_{n,k}\Sigma^{-1}\phi_n+2\sum_{n=1}^{N}t_{n,k}\Sigma^{-1}\mu_k=\mathbf{0} \qquad \text{(Apply (C.19))}$$
$$N_k\mu_k=\sum_{n=1}^{N}t_{n,k}\phi_n$$
$$\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}t_{n,k}\phi_n.$$

Hence, we conclude that the maximum likelihood estimator for μ_k, for arbitrary k, is

$$\mu_k^{\mathrm{ML}}=\frac{1}{N_k}\sum_{n=1}^{N}t_{n,k}\phi_n. \tag{4.21}$$

In order to determine the maximum likelihood estimator of Σ, we differentiate (4.20) with respect to Σ and equal the result to 0, and solve for Σ, as follows

$$\frac{\partial\log p(\mathbf{T}|\pi,\mu,\Sigma)}{\partial\Sigma}=\mathbf{0}$$
$$\sum_{j=1}^{K}\sum_{n=1}^{N}t_{n,j}\left[\Sigma^{-1}+\frac{\mathrm{d}}{\mathrm{d}\Sigma}\left\{(\phi_n-\mu_j)^\top\Sigma^{-1}(\phi_n-\mu_j)\right\}\right]=\mathbf{0} \qquad \text{(Apply (C.28))}$$
$$N\Sigma^{-1}-\sum_{j=1}^{K}\sum_{n=1}^{N}t_{n,j}(\phi_n-\mu_j)(\phi_n-\mu_j)^\top\Sigma^{-2}=\mathbf{0} \qquad \text{(Apply (C.24))}$$
$$\frac{1}{N}\sum_{j=1}^{K}\sum_{n=1}^{N}t_{n,j}(\phi_n-\mu_j)(\phi_n-\mu_j)^\top=\Sigma. \tag{4.22}$$

Note that the form in (4.22) is dependent on $\boldsymbol{\mu}_k$, however, the maximum likelihood estimator of μ_k, determined in (4.21), does not depend on $\boldsymbol{\Sigma}$. Therefore, we may plug our maximum likelihood estimator (4.21) onto (4.22). This yields the following maximum likelihood estimator for $\boldsymbol{\Sigma}$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N}\sum_{j=1}^{K}\sum_{n=1}^{N} t_{n,j}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_j^{\text{ML}})(\boldsymbol{\phi}_n - \boldsymbol{\mu}_j^{\text{ML}})^\top \\
&= \sum_{j=1}^{K}\frac{N_j}{N}\frac{1}{N_j}\sum_{n=1}^{N} t_{n,j}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_j^{\text{ML}})(\boldsymbol{\phi}_n - \boldsymbol{\mu}_j^{\text{ML}})^\top \\
\boldsymbol{\Sigma}_{\text{ML}} &= \sum_{j=1}^{K}\frac{N_j}{N}\mathbf{S}_j.
\end{aligned}$$

Where $\mathbf{S}_j$ is as in (4.163), thereby reaching the desired result.

Exercise 4.11

Consider a classification problem where we observe a data set $\{\phi_n, \mathbf{t}_n\}_{n=1}^N$, where ϕ_n are M feature vectors whose coordinates, conditional on the class $\mathcal{C}_k$ to which said observations belong, are independent of each other, hence it possesses a factorized distribution. Moreover, every coordinate of $\phi_{n,i}$ is itself an L-dimensional vector, representing an 1-of-L coding scheme, as every vector $\phi_{n,i}$ may assume one of L discrete states. Therefore, we write $\phi_{n,i,j} = 1$ if the M-th feature of the n-th observed data point was allocated to the j-class, and zero otherwise. We define

$$p(\phi_{n,i,j}|\mathcal{C}_k) = \prod_{j=1}^{L} \pi_{n,i,j,k}^{\phi_{n,i,j}},$$

where $\sum_{j=1}^{L} \pi_{n,i,j,k} = 1$ and $\pi_{n,i,j,k} \geq 0$. Note that, under this definition, the probabilities associated with each of the possible classes for the feature vectors may vary according to the sampled observation, its corresponding class, and the feature index. We may therefore rewrite (4.63) as

$$\begin{aligned} a_k &= \log\{p(\phi|\mathcal{C}_k)p(\mathcal{C}_k)\} \\ &= \log p(\phi|\mathcal{C}_k) + \log p(\mathcal{C}_k) \\ &= \log\left[\prod_{n=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{L} \pi_{n,i,j,k}^{\phi_{n,i,j}}\right] + \log p(\mathcal{C}_k) \\ &= \sum_{n=1}^{N}\sum_{i=1}^{M}\sum_{j=1}^{L} \phi_{n,i,j} \log \pi_{n,i,j,k} + \log p(\mathcal{C}_k). \end{aligned} \tag{4.23}$$

It is trivial to observe that (4.23) is linear with respect to the observed features.

Exercise 4.12

Consider the logistic-sigmoid function seen in (3.6). We seek to demonstrate the validity of the relation (4.88). It follows that

$$\begin{aligned}
\sigma(a) &= \frac{1}{1+\exp\{-a\}} && \text{(Apply (3.6))} \\
\frac{\mathrm{d}\sigma(a)}{\mathrm{d}a} &= \frac{\exp\{-a\}}{(1+\exp\{-a\})^2} \\
&= \frac{1+\exp\{-a\}-1}{(1+\exp\{-a\})^2} \\
&= \frac{1}{1+\exp\{-a\}} - \frac{1}{(1+\exp\{-a\})^2} \\
&= \frac{1}{1+\exp\{-a\}}\left(1 - \frac{1}{1+\exp\{-a\}}\right) \\
\frac{\mathrm{d}\sigma(a)}{\mathrm{d}a} &= \sigma(a)\{1-\sigma(a)\} && \text{(Apply (3.6))}.
\end{aligned}$$

Thereby reaching the desired result.

Exercise 4.13

We aim to demonstrate that the gradient of the cross-entropy function may be written as in (4.91). We rewrite the cross-entropy function (4.90) as follows

$$\begin{aligned}
E(\mathbf{w}) &= -\sum_{n=1}^{N}\{t_n \log y_n + (1-t_n)\log(1-y_n)\} \\
&= -\sum_{n=1}^{N}\{t_n \log[y_n/(1-y_n)] + \log(1-y_n)\} \\
&= -\sum_{n=1}^{N}\{t_n \sigma^{-1}(y_n) + \log(1-y_n)\} && \text{(Apply (4.15))} \\
&= -\sum_{n=1}^{N}\{t_n \sigma^{-1}(\sigma(-\mathbf{w}^\top\phi)) + \log(1-\sigma(\mathbf{w}^\top\phi))\} && \text{(Apply } y_n = \sigma(\mathbf{w}^\top\phi)) \\
E(\mathbf{w}) &= -\sum_{n=1}^{N}\{t_n \mathbf{w}^\top\phi_n + \log(\sigma(-\mathbf{w}^\top\phi_n))\} && \text{(Apply (4.14))}. \qquad (4.24)
\end{aligned}$$

Thereafter, differentiating (4.24) with respect to $\mathbf{w}$ we obtain

$$\begin{aligned}
\nabla E(\mathbf{w}) &= -\sum_{n=1}^{N}\left\{t_n\phi_n - \frac{\sigma(-\mathbf{w}^\top\phi_n)\{1-\sigma(-\mathbf{w}^\top\phi_n)\}}{\sigma(-\mathbf{w}^\top\phi_n)}\phi_n\right\} && \text{(Apply (4.88) and (C.19))} \\
&= \sum_{n=1}^{N}\{\sigma(\mathbf{w}^\top\phi_n) - t_n\}\phi_n && \text{(Apply (4.14))} \\
\nabla E(\mathbf{w}) &= \sum_{n=1}^{N}\{y_n - t_n\}\phi_n && \text{(Apply } y_n = \sigma(\mathbf{w}^\top\phi)).
\end{aligned}$$

Hence, we conclude that the gradient of the cross-entropy error function may be written as in (4.91).

Exercise 4.14

Consider that we observe two data sets $\{\mathbf{x}_m\}_{m=1}^{N_1}$ and $\{\mathbf{y}_k\}_{k=1}^{N_2}$ which are linearly separable, so that there exists $\mathbf{w}$ and w_0 such that $\mathbf{w}^\top \mathbf{x}_m + w_0 > 0$ and $\mathbf{w}^\top \mathbf{y}_k + w_0 < 0$ for all $m \in \{1, \dots, N_1\}$ and $k \in \{1, \dots, N_2\}$ (without loss of generality, we take herein $w_0 = 0$; this may be justified by adding a dummy component $y_0 = x_0 = 1$ to each data point). Consider now the commensurate logistic regression problem, where these data sets are joined, and we attribute to observations belonging to $\{\mathbf{x}_m\}_{m=1}^{N_1}$ the value $t_n = 1$ (belonging to class $\mathcal{C}_1$) and points belonging to $\{\mathbf{y}_k\}_{k=1}^{N_2}$ the value $t_n = 0$ (belonging to class $\mathcal{C}_2$), yielding a complete data set of target variables $\{t_n\}_{n=1}^{N}$, where $N = N_1 + N_2$, and input vectors $\{\boldsymbol{\phi}_n\}_{n=1}^{N}$ which is the union of $\{\mathbf{x}_m\}_{m=1}^{N_1}$ and $\{\mathbf{y}_k\}_{k=1}^{N_2}$. We rewrite the likelihood function (4.89) associated with the data, and corresponding logarithm, as

$$\begin{aligned}
p(\mathbf{t}|\mathbf{w}) &= \prod_{n=1}^{N} \left(\frac{\sigma(\mathbf{w}^\top \boldsymbol{\phi}_n)}{1 - \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n)} \right)^{t_n} \{1 - \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n)\} \\
&= \left[\prod_{n \in \mathcal{C}_1} \sigma(\mathbf{w}^\top \mathbf{x}_n) \right] \left[\prod_{n \in \mathcal{C}_2} \{1 - \sigma(\mathbf{w}^\top \mathbf{y}_n)\} \right] \\
&= \left[\prod_{n \in \mathcal{C}_1} \sigma(\mathbf{w}^\top \mathbf{x}_n) \right] \left[\prod_{n \in \mathcal{C}_2} \sigma(-\mathbf{w}^\top \mathbf{y}_n) \right] \qquad \text{(Apply (4.14))} \\
p(\mathbf{t}|\mathbf{w}) &= \left[\prod_{n \in \mathcal{C}_1} \sigma(||\mathbf{w}||||\mathbf{x}_n||\cos\alpha_n) \right] \times \\
&\quad \times \left[\prod_{n \in \mathcal{C}_2} \sigma(-||\mathbf{w}||||\mathbf{y}_n||\cos\beta_n) \right] \\
\log p(\mathbf{t}|\mathbf{w}) &= \sum_{n \in \mathcal{C}_1} \log \sigma(||\mathbf{w}||||\mathbf{x}_n||\cos\alpha_n) + \\
&\quad + \sum_{n \in \mathcal{C}_2} \log \sigma(-||\mathbf{w}||||\mathbf{y}_n||\cos\beta_n). \qquad (4.25)
\end{aligned}$$

Let $\mathbf{w}_0$ be an arbitrary vector which linearly separates our data set, it follows that

$$\begin{aligned}
\mathbf{w}_0^\top \mathbf{x}_m &> 0 \\
||\mathbf{w}_0||||\mathbf{x}_m|| \cos\alpha_m &> 0 \\
||\mathbf{x}_m|| \cos\alpha_m &> 0, \qquad (4.26)
\end{aligned}$$

where α_m denotes the angle between $\mathbf{w}_0$ and $\mathbf{x}_m$. Likewise

$$\begin{aligned}
\mathbf{w}_0^\top \mathbf{y}_k &< 0 \\
||\mathbf{w}_0||||\mathbf{y}_k|| \cos\beta_k &< 0 \\
-||\mathbf{y}_k|| \cos\beta_k &> 0, \qquad (4.27)
\end{aligned}$$

where β_k denotes the angle between $\mathbf{w}_0$ and $\mathbf{y}_k$. Let us fix the angles α and β between our input vector and $\mathbf{w}$, and consider choosing a magnitude $||\mathbf{w}||$ which maximizes (4.25).

By differentiating (4.25) with respect to $||\mathbf{w}||$, we obtain

$$\begin{aligned}
\frac{\partial \log p(\mathbf{t}|\mathbf{w})}{\partial ||\mathbf{w}||} &= \sum_{n\in\mathcal{C}_1} ||\mathbf{x}_n|| \cos\alpha_n \{1-\sigma(||\mathbf{w}||||\mathbf{x}_n||\cos\alpha_n)\} + \\
&- \sum_{n\in\mathcal{C}_2} ||\mathbf{y}_n|| \cos\beta_n \{1-\sigma(-||\mathbf{w}||||\mathbf{y}_n||\cos\beta_n)\} \quad \text{(Apply (4.88))} \\
&= \sum_{n\in\mathcal{C}_1} ||\mathbf{x}_n|| \cos\alpha_n \sigma(-||\mathbf{w}||||\mathbf{x}_n||\cos\alpha_n) + \\
&- \sum_{n\in\mathcal{C}_2} ||\mathbf{y}_n|| \cos\beta_n \sigma(||\mathbf{w}||||\mathbf{y}_n||\cos\beta_n) \quad \text{(Apply (4.14))} \\
\frac{\partial \log p(\mathbf{t}|\mathbf{w})}{\partial ||\mathbf{w}||} &> 0.
\end{aligned}$$

The above result is trivial to see, by considering that (3.6) is strictly positive and applying (4.26) and (4.27). Hence, the likelihood function (4.25) is strictly increasing with respect to the magnitude $||\mathbf{w}||$, given a fixed set of angles $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which ensures linear separability. We conclude that, in order to maximize (4.25) with respect to $||\mathbf{w}||$, we must take $||\mathbf{w}|| \to \infty$ (given a fixed set of angles $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which ensures linear separability).

Exercise 4.15

Consider the Hessian of the logistic regression model as defined in (4.97). We aim to demonstrate it is positive-definite. Let $\mathbf{a}$ be an arbitrary vector, it follows that

$$\begin{aligned}
\mathbf{a}^\top \mathbf{H}\mathbf{a} &= \mathbf{a}^\top \left[\sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top\right]\mathbf{a} && \text{(Apply (4.97))} \\
&= \sum_{n=1}^{N} y_n(1-y_n)\mathbf{a}^\top\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top\mathbf{a} \\
&= \sum_{n=1}^{N} y_n(1-y_n)\{\mathbf{a}^\top\boldsymbol{\phi}_n\}^2 \\
\mathbf{a}^\top \mathbf{H}\mathbf{a} &> 0 && \text{(Apply } y_n(1-y_n) \in (0,1)\text{)}.
\end{aligned}$$

As $\mathbf{a}^\top\mathbf{H}\mathbf{a} > 0$ we conclude by definition that $\mathbf{H}$ is positive-definite. As the Hessian of the error function is positive-definite, we conclude that the error function is convex with respect to $\mathbf{w}$, and therefore that it possesses an unique minimum.

Exercise 4.16

Consider now the context wherein we observe a data set $\{\phi_n, t_n\}_{n=1}^N$ where our target variables t_n are subject to mislabelling. As such, for every observation, we instead utilize $s_n \in \{1-\pi_n, \pi_n\}$, for $\pi_n \in [0,1]$ as a proxy for the class label, where π_n close to 1 indicates we believe the n-th observation has higher probability of belonging to class 1, and π_n close to 0 indicates we believe the n-th observation has higher probability of belonging to class 0. Assuming that, under no mislabelling, we have that $p(t_n = 1|\phi_n) = y_n$, it follows that the likelihood function associated with the n-th observation is

$$\begin{aligned} p(s_n|\phi) &\propto y_n^{s_n}(1-y_n)^{1-s_n} \\ &= \frac{y_n^{s_n}(1-y_n)^{1-s_n}}{y_n^{\pi_n}(1-y_n)^{1-\pi_n} + y_n^{1-\pi_n}(1-y_n)^{\pi_n}}. \end{aligned}$$

Consequently, the sample logarithm likelihood is equal to

$$\begin{aligned} \sum_{n=1}^N \log p(s_n|\phi) = \sum_{n=1}^N s_n \log y_n + \sum_{n=1}^N (1-s_n)\log(1-y_n) + \\ - \sum_{n=1}^N \log[y_n^{\pi_n}(1-y_n)^{1-\pi_n} + y_n^{1-\pi_n}(1-y_n)^{\pi_n}]. \end{aligned}$$

Exercise 4.17

We aim to demonstrate herein that the derivative of (4.104) with respect to a_j, where a_j is as in (4.105), may be written as in (4.106), as follows

$$\begin{aligned}
\frac{\partial p(\mathcal{C}_k|\phi)}{\partial a_j} &= \frac{\partial}{\partial a_j}\left[\frac{\exp\{a_k\}}{\sum_{i=1}^K \exp\{a_i\}}\right] \\
&= \frac{\partial}{\partial a_j}\left[\frac{\exp\{a_k\}}{\exp\{a_j\} + \sum_{\substack{i=1\\ i\neq j}}^K \exp\{a_i\}}\right] \\
&= \begin{cases} \frac{\partial}{\partial a_j}\left[1 - \frac{\sum_{\substack{i=1\\ i\neq j}}^K \exp\{a_i\}}{\exp\{a_j\}+\sum_{\substack{i=1\\ i\neq j}}^K \exp\{a_i\}}\right] & \text{if } k = j, \\ \frac{\partial}{\partial a_j}\left[\frac{\exp\{a_k\}}{\exp\{a_j\}+\sum_{\substack{i=1\\ i\neq j}}^K \exp\{a_i\}}\right] & \text{if } k \neq j. \end{cases} \\
&= \begin{cases} \frac{[\sum_{i=1}^K \exp\{a_i\} - \exp\{a_j\}]\exp\{a_j\}}{(\sum_{i=1}^K \exp\{a_i\})^2} & \text{if } k = j, \\ -\frac{\exp\{a_k\}\exp\{a_j\}}{(\sum_{i=1}^K \exp\{a_i\})^2} & \text{if } k \neq j. \end{cases} \\
&= \begin{cases} y_j(1-y_j) & \text{if } k = j, \\ -y_j y_k & \text{if } k \neq j. \end{cases} \qquad \text{(Apply (4.104))} \\
\frac{\partial p(\mathcal{C}_k|\phi)}{\partial a_j} &= y_k(I_{j,k} - y_j),
\end{aligned}$$

where $I_{j,k} = 0$ if $j \neq k$ and $I_{j,j} = 1$. We thereby reach the desired result.

Exercise 4.18

We aim herein to demonstrate that the gradient of the cross-entropy loss for multiple classification (4.108) with respect to $\mathbf{w}_j$ is determined by (4.109). Firstly, we find that the gradient associated with (4.105) is

$$\begin{aligned}
&\frac{\mathrm{d}a_k}{\mathrm{d}\mathbf{w}_j} = \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}_j}[\mathbf{w}_k^\top \boldsymbol{\phi}] \\
&\frac{\mathrm{d}a_k}{\mathrm{d}\mathbf{w}_j} = \boldsymbol{\phi} I_{j,k} \qquad \text{(Apply (C.19))},
\end{aligned} \tag{4.28}$$

where $I_{j,k} = 0$ if $j \neq k$ and $I_{j,j} = 1$. Therefore, differentiating (4.108) with respect to $\mathbf{w}_j$, we find that

$$\begin{aligned}
\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) &= \nabla_{\mathbf{w}_j} \left[-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{n,k} \log y_{n,k} \right] && \text{(Apply (4.108))} \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{n,k} \frac{1}{y_{n,k}} \frac{\partial y_{n,k}}{\partial a_{n,j}} \frac{\mathrm{d}a_j}{\mathrm{d}\mathbf{w}_j} \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{n,k} \frac{1}{y_{n,k}} y_{n,k}(I_{j,k} - y_{n,j}) \boldsymbol{\phi}_n I_{j,j} && \text{(Apply (4.106) and (4.28))} \\
&= -\sum_{n=1}^{N} t_{n,j} \boldsymbol{\phi}_n + \sum_{n=1}^{N}\sum_{k=1}^{K} t_{n,k} y_{n,j} \boldsymbol{\phi}_n \\
\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) &= \sum_{n=1}^{N} \{y_{n,j} - t_{n,j}\} \boldsymbol{\phi}_j && \text{(Apply } \sum_{k=1}^{K} t_{n,k} = 1).
\end{aligned}$$

We thereby reach the desired result.

Exercise 4.19

Consider that we observe the data set $\{\phi_n, t_n\}_{n=1}^N$, and choose to adopt the probit model for classification, such that the likelihood function associated with the data set, and corresponding logarithm, is

$$p(\mathbf{T}|\mathbf{w}) = \prod_{n=1}^N \{\Phi(a_n)\}^{t_n}\{1-\Phi(a_n)\}^{1-t_n}$$

$$\log p(\mathbf{T}|\mathbf{w}) = \sum_{n=1}^N t_n \log \Phi(a_n) + \sum_{n=1}^N (1-t_n)\log\{1-\Phi(a_n)\} \tag{4.29}$$

where $\Phi(a)$ is as defined in (4.114), and $a_n = \mathbf{w}^\top \phi_n$. First, we consider the derivative of (4.114) with respect to a. It follows that

$$\begin{aligned}
\frac{\mathrm{d}\Phi(a)}{\mathrm{d}a} &= \frac{\mathrm{d}}{\mathrm{d}a}\left[\int_{-\infty}^a \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{t^2}{2}\right\}\mathrm{d}t\right] \\
&= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{a^2}{2}\right\} \\
\frac{\mathrm{d}\Phi(a)}{\mathrm{d}a} &= \psi(a),
\end{aligned}$$

where $\psi(t)$ denotes the probability density function of a Normal random variable with mean 0 and variance 1. Differentiating once more (4.114) we find that

$$\frac{\mathrm{d}^2\Phi(a)}{\mathrm{d}a^2} = \psi'(a).$$

Differentiating (4.29) with respect to $\mathbf{w}$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}\log p(\mathbf{T}|\mathbf{w})}{\mathrm{d}\mathbf{w}} &= \sum_{n=1}^N t_n \frac{\psi(a_n)}{\Phi(a_n)}\phi_n - \sum_{n=1}^N (1-t_n)\frac{\psi(a_n)}{1-\Phi(a_n)}\phi_n \\
&= \sum_{n=1}^N \frac{t_n\{1-\Phi(a_n)\} - (1-t_n)\Phi(a_n)}{\Phi(a_n)\{1-\Phi(a_n)\}}\psi(a_n)\phi_n \\
\frac{\mathrm{d}\log p(\mathbf{T}|\mathbf{w})}{\mathrm{d}\mathbf{w}} &= \sum_{n=1}^N \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1-\Phi(a_n)\}}\psi(a_n)\phi_n.
\end{aligned} \tag{4.30}$$

Concluding that the gradient the logarithm of the likelihood function with respect to $\mathbf{w}$ is as in (4.30). In order to determine the Hessian of (4.29), we differentiate the transpose

of (4.30) with respect to $\mathbf{w}$, as follows

$$
\begin{aligned}
\frac{\mathrm{d}^2 \log p(\mathbf{T}|\mathbf{w})}{\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{w}^\top} &= \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\left[\sum_{n=1}^{N} \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1-\Phi(a_n)\}}\psi(a_n)\phi_n^\top\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\left[\sum_{n=1}^{N} \left\{\frac{t_n}{\Phi(a_n)\{1-\Phi(a_n)\}} - \frac{1}{1-\Phi(a_n)}\right\}\psi(a_n)\phi_n^\top\right] \\
&= \sum_{n=1}^{N}\left[-\frac{t_n\psi(a_n)\{1-\Phi(a_n)\} - t_n\Phi(a_n)\psi(a_n)}{[\Phi(a_n)\{1-\Phi(a_n)\}]^2}\phi_n + \right. \\
&\quad \left. - \frac{\psi(a_n)}{[1-\Phi(a_n)]^2}\phi_n\right]\psi(a_n)\phi_n^\top + \\
&\quad + \sum_{n=1}^{N}\left\{\frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1-\Phi(a_n)\}}\right\}\psi'(a_n)\phi_n\phi_n^\top \\
&= \sum_{n=1}^{N}\left[\frac{2t_n\Phi(a_n) - t_n - \{\Phi(a_n)\}^2}{[\Phi(a_n)\{1-\Phi(a_n)\}]^2}\right]\{\psi(a_n)\}^2\phi_n\phi_n^\top + \\
&\quad + \sum_{n=1}^{N}\left\{\frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1-\Phi(a_n)\}}\right\}\psi'(a_n)\phi_n\phi_n^\top
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}^2 \log p(\mathbf{T}|\mathbf{w})}{\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{w}^\top} &= -\sum_{n=1}^{N}\left[\frac{[t_n - \Phi(a_n)]^2}{[\Phi(a_n)\{1-\Phi(a_n)\}]^2}\right]\frac{1}{2\pi}\exp\{-a_n^2\}\phi_n\phi_n^\top + \\
&\quad - \sum_{n=1}^{N} a_n\left\{\frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1-\Phi(a_n)\}}\right\}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{a_n^2}{2}\right\}\phi_n\phi_n^\top.
\end{aligned} \tag{4.31}
$$

Note that, above, we utilized the result that since $t_n \in \{0,1\}$, it follows that $t_n = t_n^2$. Hence, we find that (4.31) is the Hessian matrix associated with our model likelihood. Note that we defined both (4.30) and (4.31) in the maximum likelihood context, as opposed to the minimum error context, i.e., we are determining ways to maximize (4.29), not minimize its negative, albeit the result is equivalent. Note that this implies that the Hessian in (4.31) must be negative definite.

Exercise 4.20

We aim herein to demonstrate that the Hessian in (4.110) is positive semi definite. It follows that the Hessian matrix is composed of blocks, as follows

$$\mathbf{H} = \begin{pmatrix} \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_1}^\top E(\tilde{\mathbf{w}}) & \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_2}^\top E(\tilde{\mathbf{w}}) & \dots & \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_K}^\top E(\tilde{\mathbf{w}}) \\ \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_1}^\top E(\tilde{\mathbf{w}}) & \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_2}^\top E(\tilde{\mathbf{w}}) & \dots & \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_K}^\top E(\tilde{\mathbf{w}}) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_1}^\top E(\tilde{\mathbf{w}}) & \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_2}^\top E(\tilde{\mathbf{w}}) & \dots & \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_K}^\top E(\tilde{\mathbf{w}}) \end{pmatrix}.$$

We thereby compute $\mathbf{u}^\top\mathbf{H}\mathbf{u}$, where $\mathbf{u}$ is an arbitrary vector of dimension MK, itself partitioned into K sections of length M, as follows

$$\mathbf{u}^\top\mathbf{H}\mathbf{u} = \sum_{n=1}^{N} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_K \end{pmatrix}^\top \begin{pmatrix} \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_1}^\top E(\tilde{\mathbf{w}}) & \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_2}^\top E(\tilde{\mathbf{w}}) & \dots & \nabla_{\mathbf{w}_1}\nabla_{\mathbf{w}_K}^\top E(\tilde{\mathbf{w}}) \\ \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_1}^\top E(\tilde{\mathbf{w}}) & \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_2}^\top E(\tilde{\mathbf{w}}) & \dots & \nabla_{\mathbf{w}_2}\nabla_{\mathbf{w}_K}^\top E(\tilde{\mathbf{w}}) \\ \vdots & \vdots & \ddots & \vdots \\ \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_1}^\top E(\tilde{\mathbf{w}}) & \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_2}^\top E(\tilde{\mathbf{w}}) & \dots & \nabla_{\mathbf{w}_K}\nabla_{\mathbf{w}_K}^\top E(\tilde{\mathbf{w}}) \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_K \end{pmatrix}$$

We proceed thereafter as

$$\begin{aligned} \mathbf{u}^\top\mathbf{H}\mathbf{u} &= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{u}_j^\top \nabla_{\mathbf{w}_j}\nabla_{\mathbf{w}_k}^\top E(\tilde{\mathbf{w}})\mathbf{u}_k \\ &= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{u}_j^\top \left[\sum_{n=1}^{N} y_{n,k}(I_{j,k} - y_{n,j})\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top\right]\mathbf{u}_k \qquad \text{(Apply (4.110))} \\ &= \sum_{n=1}^{N}\left[\sum_{j=1}^{K}\sum_{k=1}^{K} y_{n,k}(I_{j,k} - y_{n,j})\{\mathbf{u}_j^\top\boldsymbol{\phi}_n\}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}\right] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}^2 + \\ &\quad - \sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}\sum_{j=1}^{K} y_{n,j}\{\mathbf{u}_j^\top\boldsymbol{\phi}_n\} \end{aligned}$$

$$\mathbf{u}^\top\mathbf{H}\mathbf{u} = \sum_{n=1}^{N}\sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}^2 - \sum_{n=1}^{N}\left[\sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}\right]^2. \tag{4.32}$$

Note, from (4.104), that $\sum_{k=1}^{K} y_{n,k} = 1$, and $y_{n,k} \geq 0, \forall k \in \{1,\dots,K\}$, $\forall n \in \{1,\dots,N\}$. Consider that the function $f(t) = t^2$ is convex. Therefore, from (1.115), we can conclude that

$$\left[\sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}\right]^2 \leq \sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}^2 \quad \forall n \in \{1,\dots,N\}$$

$$0 \leq \sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}^2 - \left[\sum_{k=1}^{K} y_{n,k}\{\mathbf{u}_k^\top\boldsymbol{\phi}_n\}\right]^2 \quad \forall n \in \{1,\dots,N\}. \tag{4.33}$$

By joining (4.32) and (4.33), we find that

$$\mathbf{u}^\top\mathbf{H}\mathbf{u} \geq 0,$$

for any arbitrary vector $\mathbf{u}$. We conclude that, by definition, $\mathbf{H}$ is positive semi definite.

Exercise 4.21

We aim to demonstrate the equivalence between (4.114) and (4.116). It follows that

$$
\begin{aligned}
\Phi(a) &= \int_{-\infty}^{a} \phi(t)\,\mathrm{d}t \\
&= \int_{-\infty}^{0} \phi(t)\,\mathrm{d}t + \int_{0}^{a} \phi(t)\,\mathrm{d}t \\
&= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t + \\
&+ \int_{0}^{a} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t \\
&= \frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t + \\
&+ \int_{0}^{a/\sqrt{2}} \frac{1}{\sqrt{\pi}} \exp\{-s^2\}\,\mathrm{d}s \qquad \text{(Symmetry of } \phi(t) \text{ and set } s = t/\sqrt{2}) \\
&= \frac{1}{2}\frac{1}{\sqrt{2\pi}} I + \\
&+ \frac{1}{2}\frac{2}{\pi}\int_{0}^{a/\sqrt{2}} \frac{1}{\sqrt{\pi}} \exp\{-s^2\}\,\mathrm{d}s \qquad \text{(Apply (1.124))} \\
\Phi(a) &= \frac{1}{2}\left\{1 + \mathrm{erf}\left\{\frac{a}{\sqrt{2}}\right\}\right\} \qquad \text{(Apply (4.115)).}
\end{aligned}
$$

Where herein we denoted $\phi(t)$ as the probability density function of a Normal random variable with mean 0 and variance 1. Hence we reach the desired result. Note that for this Exercise we utilized results proven in Exercise 1.7.

Exercise 4.22

Consider herein that we aim to determine an approximation to the model evidence $p(\mathcal{D}) = \int p(\mathcal{D}, \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}$, where the parameters $\boldsymbol{\theta}$ belong to an M-dimensional space. Consider moreover that we have determined the maximum density point of $p(\boldsymbol{\theta}|\mathcal{D})$, denoted as $\boldsymbol{\theta}_{\text{MAP}}$, which is consequently also the maximum point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ (see (1.32)). We thereby adopt the Laplace approximation in (4.135) for $p(\mathcal{D})$, which yields

$$\begin{aligned} p(\mathcal{D}) &\approx p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}})\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \\ \log p(\mathcal{D}) &\approx \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \log p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{A}|, \end{aligned}$$

where $\mathbf{A}$ is as in (4.138). Hence, we reach the desired result.

Exercise 4.23

We return to the context of Exercise 4.22, and aim to exploit the result in (4.137) to derive the BIC approximation result (4.139). We attribute to our parameter vector $\boldsymbol{\theta}$ an M-dimensional multivariate Normal distribution with mean $\mathbf{m} \in \mathbb{R}^M$ and covariance matrix $\mathbf{V}_0 \in \mathbb{R}^{M\times M}$. We therefore rewrite (4.137) as

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \log p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{A}|$$

$$\begin{aligned}\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) &- \frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{V}_0| + \\ &- \frac{1}{2}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m})^\top \mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}) + \frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{A}| \quad \text{(Apply (2.118))}.\end{aligned} \tag{4.34}$$

We assume that our data set $\mathcal{D}$ is composed of independent and identically distributed data points, which we denote as $\mathcal{D} = \{d_1, \dots, d_N\}$, and also that $\mathbf{V}_0^{-1}$ is approximately a matrix composed of zeroes. We rewrite $\mathbf{A}$ in (4.138), such that

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla^\top \log\{p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}})\} \\ &= -\nabla\nabla^\top \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \nabla\nabla^\top \log p(\boldsymbol{\theta}_{\text{MAP}}) \\ &= -\sum_{n=1}^N \nabla\nabla^\top \log p(d_n|\boldsymbol{\theta}_{\text{MAP}}) + \\ &\quad - \nabla\nabla^\top \left[\frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{V}_0| + \right. \\ &\quad \left. - \frac{1}{2}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m})^\top \mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m})\right] \qquad \text{(Apply (2.118))} \\ &= -\sum_{n=1}^N \nabla\nabla^\top \log p(d_n|\boldsymbol{\theta}_{\text{MAP}}) + \mathbf{V}_0^{-1} \qquad \text{(Apply (C.19))}\end{aligned}$$

$$\mathbf{A} \approx \mathbf{H} \qquad (\mathbf{V}_0^{-1} \text{ small}). \tag{4.35}$$

We denote $\mathbf{H} = -\sum_{n=1}^N \nabla\nabla^\top \log p(d_n|\boldsymbol{\theta}_{\text{MAP}})$. We return therefore to (4.34)

$$\begin{aligned}\log p(\mathcal{D}) &\approx \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\log|\mathbf{V}_0| \qquad \text{(Apply (4.35) and } \mathbf{V}_0^{-1} \text{ small)} \\ &= \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{1}{2}\log|N\mathbf{H}/N| - \frac{1}{2}\log|\mathbf{V}_0| \\ &= \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2}\log|N| + \\ &\quad - \frac{1}{2}\log\left|\frac{\mathbf{H}}{N}\right| - \frac{1}{2}\log|\mathbf{V}_0|.\end{aligned}$$

Note that $\mathbf{H}/N$ constitutes an approximation of $\mathbb{E}_{\mathcal{D}}[-\nabla\nabla^\top \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})]$ in the sense of (1.35). Hence, under mild conditions, $\log|\mathbf{H}/N|$ is bounded, and we have that

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2}\log|N|.$$

Exercise 4.24

Consider that, in the classification context, we are provided with a sample $\{\phi_n, t_n\}_{n=1}^N$ of input and target values and seek to determine the predictive distribution of a new data point given the input vector ϕ, for a Bayesian logistic regression model. We adopt a Laplace approximation for the posterior distribution of our parameters $\mathbf{w}$, as in (4.144). Under this approximation, we seek to determine the joint posterior distribution of the following vector

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{w}^\top \phi \end{pmatrix}.$$

Note that, as $\mathbf{w}$ is marginally distributed as a multivariate Normal random vector, therefore $\mathbf{w}^\top \phi$ is marginally distributed as a univariate Normal random variable. Particularly, we find that the expected values of this partition are

$$\mathbb{E}[\mathbf{w}] = \mathbf{w}_{\text{MAP}} \quad \text{(Apply (2.59) and (4.144))},$$

and

$$\begin{aligned} \mathbb{E}[\mathbf{w}^\top \phi] &= \mathbb{E}[\mathbf{w}^\top]\phi \\ &= \mathbf{w}_{\text{MAP}}^\top \phi \quad \text{(Apply (2.59) and (4.144))}. \end{aligned}$$

Whilst the covariances are

$$\mathbb{V}\text{ar}[\mathbf{w}] = \mathbf{S}_N \quad \text{(Apply (2.64) and (4.144))},$$

where $\mathbf{S}_N$ is as in (4.143). Moreover

$$\begin{aligned} \mathbb{C}\text{ov}[\mathbf{w}, \mathbf{w}^\top \phi] &= \mathbb{C}\text{ov}[\mathbf{w}, \phi^\top \mathbf{w}] \\ &= \phi^\top \mathbb{V}\text{ar}[\mathbf{w}] \\ \mathbb{C}\text{ov}[\mathbf{w}, \mathbf{w}^\top \phi] &= \phi^\top \mathbf{S}_n \qquad \text{(Apply (2.64) and (4.144))} \\ \mathbb{C}\text{ov}[\mathbf{w}^\top \phi, \mathbf{w}^\top] &= \mathbf{S}_n \phi \qquad \text{(Apply (2.64) and (4.144))} \end{aligned}$$

and

$$\mathbb{V}\text{ar}[\mathbf{w}^\top \phi] = \phi^\top \mathbf{S}_n \phi \quad \text{(Apply (2.64) and (4.144))}.$$

Therefore, we conclude that

$$\mathbb{E}\left[\begin{pmatrix} \mathbf{w} \\ \mathbf{w}^\top \phi \end{pmatrix}\right] = \begin{pmatrix} \mathbf{w}_{\text{MAP}} \\ \mathbf{w}_{\text{MAP}}^\top \phi \end{pmatrix} \quad \text{and} \quad \mathbb{V}\text{ar}\left[\begin{pmatrix} \mathbf{w} \\ \mathbf{w}^\top \phi \end{pmatrix}\right] = \begin{pmatrix} \mathbf{S}_N & \mathbf{S}_N \phi \\ \phi^\top \mathbf{S}_N & \phi^\top \mathbf{S}_n \phi. \end{pmatrix}$$

Let us denote $a = \mathbf{w}^\top \phi$. It follows from previous results (2.92) and (2.93) for the marginal distribution of joint multivariate Normal random vectors, that a is marginally distributed as an univariate Normal with mean $\mathbb{E}[a] = \mathbf{w}_{\text{MAP}}^\top \phi$ and variance $\mathbb{V}\text{ar}[a] = \phi^\top \mathbf{S}_N \phi$. We thereby return to (4.151), and find that

$$\begin{aligned} p(\mathcal{C}_1|t) &\approx \int \sigma(a) p(a) \, \mathrm{d}a \\ &= \int \sigma(a) \phi(a | \mathbf{w}_{\text{MAP}}^\top \phi, \phi^\top \mathbf{S}_N \phi) \, \mathrm{d}a, \end{aligned}$$

whence $\phi(t|\mu, \sigma^2)$ denotes the univariate Normal probability density function with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. We thereby conclude that we reach the same result as in (4.151).

Exercise 4.25

We aim to demonstrate that the derivative of the logistic-sigmoid function (3.6) evaluated at zero equals that of the scaled probit function (4.114) with $\Phi(\lambda a)$ for $\lambda = \pi^2/8$. It follows that the derivative of the logistic-sigmoid function evaluated at the origin is

$$\begin{aligned}
\left.\frac{\mathrm{d}\sigma(a)}{\mathrm{d}a}\right|_{a=0} &= \left.\sigma(a)\{1-\sigma(a)\}\right|_{a=0} && \text{(Apply (4.88))} \\
&= \frac{1}{1+e^{-0}}\left(1-\frac{1}{1+e^{-0}}\right) && \text{(Apply (3.6))}
\end{aligned}$$

$$\left.\frac{\mathrm{d}\sigma(a)}{\mathrm{d}a}\right|_{a=0} = \frac{1}{4}. \tag{4.36}$$

The derivative of the scaled probit function at the origin is

$$\begin{aligned}
\left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} &= \left.\frac{\mathrm{d}}{\mathrm{d}a}\left[\int_{-\infty}^{\lambda a}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}t^2\right\}\mathrm{d}t\right]\right|_{a=0} && \text{(Apply (4.114))} \\
&= \left.\frac{\mathrm{d}}{\mathrm{d}a}\left[\int_{-\infty}^{a}\frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2}{2}s^2\right\}\mathrm{d}s\right]\right|_{a=0} && \text{(Set } s=t/\lambda) \\
&= \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2}{2}0\right\}
\end{aligned}$$

$$\left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}. \tag{4.37}$$

Equating (4.36) and (4.37), we find

$$\begin{aligned}
\left.\frac{\mathrm{d}\sigma(a)}{\mathrm{d}a}\right|_{a=0} &= \left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} \\
\frac{1}{4} &= \frac{\lambda}{\sqrt{2\pi}} \\
\lambda &= \frac{\sqrt{2\pi}}{4} \\
\lambda^2 &= \frac{\pi}{8}.
\end{aligned}$$

Hence we reach the desired result.

Exercise 4.26

We seek to demonstrate the validity of the relation (4.152). For that purpose, first we find the form of the derivative of the right-hand-side with respect to μ, as follows

$$\begin{aligned}
\frac{\mathrm{d}\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}\right)}{\mathrm{d}\mu} &= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int_{-\infty}^{\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}t^2\right\}\mathrm{d}t\right] && \text{(Apply (4.114))}\\
&= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int_{-\infty}^{\mu} \frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}\sqrt{2\pi}}\times\right.\\
&\quad \left.\times\exp\left\{-\frac{1}{2}\frac{s^2}{\lambda^{-2}+\sigma^2}\right\}\mathrm{d}s\right] && \text{(Set } s=\sqrt{\lambda^{-2}+\sigma^2}t\text{)}
\end{aligned}$$

$$\frac{\mathrm{d}\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}\right)}{\mathrm{d}\mu} = \frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{\mu^2}{\lambda^{-2}+\sigma^2}\right\}. \tag{4.38}$$

For the left-hand-side, we have that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a\right] &= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int \sigma\Phi(\lambda\{\mu+\sigma z\})\times\right.\\
&\quad \left.\times\,\phi(\mu+\sigma z|\mu,\sigma^2)\,\mathrm{d}z\right] && \text{(Set } a=\mu+\sigma z\text{)}\\
&= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int \sigma\left(\int_{-\infty}^{\lambda(\mu+\sigma z)}\frac{1}{\sqrt{2\pi}}\times\right.\right.\\
&\quad \left.\left.\times\exp\left\{-\frac{1}{2}t^2\right\}\mathrm{d}t\right)\times\right.\\
&\quad \left.\times\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}z^2\right\}\mathrm{d}z\right] && \text{(Apply (1.46) and (4.114))}\\
&= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[\left(\int_{-\infty}^{\mu}\frac{\lambda}{\sqrt{2\pi}}\times\right.\right.\\
&\quad \left.\left.\times\exp\left\{-\frac{\lambda^2}{2}(s+\sigma z)^2\right\}\mathrm{d}s\right)\times\right.\\
&\quad \left.\times\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}z^2\right\}\mathrm{d}z\right] && \text{(Set } s=\frac{t-\lambda\sigma z}{\lambda}\text{)}\\
&= \int\frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2\sigma^2}{2}\left(\frac{\mu}{\sigma}+z\right)^2\right\}\times\\
&\quad \times\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}z^2\right\}\mathrm{d}z.
\end{aligned}$$

Note that $\phi(t|\mu,\sigma)$ denotes the probability density function of a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. We proceed as follows

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a\right] &= \frac{1}{\sqrt{2\pi}}\int \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2\sigma^2}{2}\left(\frac{\mu^2}{\sigma^2}+\right.\right.\\
&\quad \left.\left.+2\frac{\mu}{\sigma}z+z^2+\frac{1}{\lambda^2\sigma^2}z^2\right)\right\}\mathrm{d}z\\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2\sigma^2}{2}\left(\frac{\mu^2}{\sigma^2}+\right.\right.\\
&\quad \left.\left.+2\frac{\mu}{\sigma}z+\frac{\lambda^2\sigma^2+1}{\lambda^2\sigma^2}z^2\right)\right\}\mathrm{d}z\\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2\sigma^2}{2}\left(\frac{\mu^2}{\sigma^2}+\right.\right.\\
&\quad +\left[\frac{\mu}{\sigma}\sqrt{\frac{\lambda^2\sigma^2}{\lambda^2\sigma^2+1}}+\sqrt{\frac{\lambda^2\sigma^2+1}{\lambda^2\sigma^2}}z\right]^2+\\
&\quad \left.\left.-\frac{\mu^2}{\sigma^2}\frac{\lambda^2\sigma^2}{\lambda^2\sigma^2+1}\right)\right\}\mathrm{d}z\\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{\lambda^2\sigma^2}{2}\frac{\mu^2}{\sigma^2}+\right.\\
&\quad +\frac{\lambda^2\sigma^2}{2}\frac{\mu^2}{\sigma^2}\frac{\lambda^2\sigma^2}{\lambda^2\sigma^2+1}+\\
&\quad \left.-\frac{\lambda^2\sigma^2}{2}\left[\frac{\mu}{\sigma}\sqrt{\frac{\lambda^2\sigma^2}{\lambda^2\sigma^2+1}}+\sqrt{\frac{\lambda^2\sigma^2+1}{\lambda^2\sigma^2}}z\right]^2\right\}\mathrm{d}z\\
&= \frac{1}{\sqrt{2\pi}}\int \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\lambda^2\mu^2-\frac{\mu^2\lambda^4\sigma^2}{\lambda^2\sigma^2+1}\right)+\right.\\
&\quad \left.-\frac{\lambda^2\sigma^2}{2}\left[\frac{\mu}{\sigma}\sqrt{\frac{\lambda^2\sigma^2}{\lambda^2\sigma^2+1}}+\sqrt{\frac{\lambda^2\sigma^2+1}{\lambda^2\sigma^2}}z\right]^2\right\}\mathrm{d}z\\
&= \frac{\lambda}{\sqrt{2\pi}(\lambda^2\sigma^2+1)^{1/2}}\int \frac{(\lambda^2\sigma^2+1)^{1/2}}{\sqrt{2\pi}}\times\\
&\quad \times\exp\left\{-\frac{1}{2}\left(\frac{\lambda^2\mu^2}{\lambda^2\sigma^2+1}\right)+\right.\\
&\quad \left.-\frac{\lambda^2\sigma^2+1}{2}\left[\frac{\mu}{\sigma}\frac{\lambda^2\sigma^2}{\lambda^2\sigma^2+1}+z\right]^2\right\}\mathrm{d}z
\end{aligned}$$

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\left[\int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a\right] = \frac{1}{\sqrt{2\pi}(\sigma^2+\lambda^{-2})^{1/2}}\exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma^2+\lambda^{-2}}\right\} \tag{4.39}$$

(Apply (1.48)).

We conclude by comparing (4.38) and (4.39) that the derivatives of the left-hand-side and right-hand-side of (4.152) taken with respect to μ are equal. We note moreover that the derivative is given by $\phi(0|\mu,\sigma^2+\lambda^{-2})$, which is the probability density function of a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2+\lambda^{-2} > 0$ evaluated at 0. We consider now integrating both sides of (4.152) with respect to μ: first, for the

right-hand-side of (4.152) we have that

$$\int \left[\frac{\mathrm{d}\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}\right)}{\mathrm{d}\mu}\right] \mathrm{d}\mu + C = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right) \tag{4.40}$$

$$\int \frac{1}{\sqrt{2\pi}(\sigma^2+\lambda^{-2})^{1/2}} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma^2+\lambda^{-2}}\right\} \mathrm{d}\mu + C = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right)$$

$$\Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right) + C = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right)$$

$$C = 0.$$

It is therefore trivial to conclude that, for the right-hand-side, the integration coefficient $C = 0$ vanishes. For the left-hand-side, we find that

$$\int \left[\frac{\mathrm{d}}{\mathrm{d}\mu}\left\{\int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a\right\}\right] \mathrm{d}\mu + L = \int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a$$

$$\int \frac{1}{\sqrt{2\pi}(\sigma^2+\lambda^{-2})^{1/2}} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma^2+\lambda^{-2}}\right\} \mathrm{d}\mu + L = \int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a$$

$$\Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right) + L = \int \Phi(\lambda a)\phi(a|\mu,\sigma^2)\,\mathrm{d}a. \tag{4.41}$$

By taking the limit $\mu \to -\infty$ on both sides of (4.41), we find that $L = 0$. We thereby conclude that the relation (4.152) is valid.

Chapter 5
Neural Networks

Exercise 5.1

We aim to demonstrate that there if we adopt the logistic-sigmoid as an activation function in (5.7), there exists an equivalent network whose activation function is the $\tanh(a)$ function, as in (5.59). From (5.59) we may rewrite

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

We thereby rewrite (5.7) as

$$\begin{aligned}y_k(\mathbf{x},\mathbf{w}) &= \sigma\left(\sum_{j=1}^{M} w_{k,j}^{(2)}\sigma\left(\sum_{i=1}^{D} w_{j,i}^{(1)}x_i + w_{j,0}^{(1)}\right) + w_{k,0}^{(2)}\right)\\ &= \sigma\left(\sum_{j=1}^{M} w_{k,j}^{(2)}\left[\frac{1}{2}\tanh\left(\sum_{i=1}^{D}\frac{w_{j,i}^{(1)}}{2}x_i + \frac{w_{j,0}^{(1)}}{2}\right) + \frac{1}{2}\right] + w_{k,0}^{(2)}\right)\\ &= \sigma\left(\sum_{j=1}^{M}\frac{w_{k,j}^{(2)}}{2}\tanh\left(\sum_{i=1}^{D}\frac{w_{j,i}^{(1)}}{2}x_i + \frac{w_{j,0}^{(1)}}{2}\right) + \sum_{j=1}^{M}\frac{w_{k,j}^{(2)}}{2} + w_{k,0}^{(2)}\right)\\ y_k(\mathbf{x},\mathbf{u}) &= \sigma\left(\sum_{j=1}^{M} u_{k,j}^{(2)}\tanh\left(\sum_{i=1}^{D} u_{j,i}^{(1)}x_i + u_{j,0}^{(1)}\right) + u_{k,0}^{(2)}\right),\end{aligned}$$

where

$$u_{k,j}^{(2)} = \frac{w_{k,j}^{(2)}}{2}, \qquad u_{k,0}^{(2)} = \sum_{j=1}^{M}\frac{w_{k,j}^{(2)}}{2} + w_{k,0}^{(2)}, \qquad u_{j,i}^{(1)} = \frac{w_{j,i}^{(1)}}{2} \quad \text{and} \quad u_{j,0}^{(1)} = \frac{w_{j,0}^{(1)}}{2}. \tag{5.1}$$

Therefore, by choosing the network parameters for a two-layer network with $\tanh(a)$ activation to be of the form (5.1), we find that it will compute the same value as the network function in (5.7) with logistic-sigmoid activation function.

Exercise 5.2

We aim to demonstrate that, given a data set $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$ of input and target variables, and assuming that the target variables $\mathbf{t}_n$ are distributed as D-dimensional Multivariate normal with mean $\mathbf{y}(\mathbf{x}, \mathbf{w}) \in \mathbb{R}^D$ and precision matrix $\beta \boldsymbol{I} \in \mathbb{R}^{D\times D}$, where $\beta > 0$, maximizing the likelihood function (5.16) with respect to $\mathbf{w}$ corresponds to minimizing the least squares function (5.11). It follows that

$$\begin{aligned}
\log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1}\boldsymbol{I}) &= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\beta^{-1}\boldsymbol{I}| + \\
&\quad - \frac{\beta}{2}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))^\top(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w})) \qquad \text{(Apply (2.118))} \\
&\propto -\frac{1}{2}||\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w})||^2 \\
&= -\frac{1}{2}||\mathbf{y}(\mathbf{x}, \mathbf{w}) - \mathbf{t}_n||^2 \\
\log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1}\boldsymbol{I}) &\propto -E(\mathbf{w}) \qquad \text{(Apply (5.11))}.
\end{aligned} \tag{5.2}$$

Note therefore that the logarithm likelihood associated with our data set is proportional (when seen as a function of $\mathbf{w}$) exclusively on the negative least squares function. Trivially, it follows that maximizing the left-hand-side of (5.2) is equivalent to minimizing the negative of the right-hand-side.

Exercise 5.3

We now consider much the same context as in Exercise 5.2, except now the target variables possess covariance matrix $\Sigma \in \mathbb{R}^{D\times D}$: note this implies the coordinates of the target variables $\mathbf{t}$ are no longer independent. Assuming Σ is known, we write

$$\begin{aligned}
\log p(\mathbf{T}|\mathbf{y}(\mathbf{x},\mathbf{w}),\Sigma) &= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| + \\
&\quad - \frac{1}{2}(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))^\top \Sigma^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w})) \qquad \text{(Apply (2.118))} \\
&\propto -\frac{1}{2}(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))^\top \Sigma^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w})) \qquad (5.3)\\
\log p(\mathbf{T}|\mathbf{y}(\mathbf{x},\mathbf{w}),\Sigma) &\propto -\frac{1}{2}\mathrm{tr}[(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))^\top \Sigma^{-1}] \\
\log p(\mathbf{T}|\mathbf{y}(\mathbf{x},\mathbf{w}),\Sigma) &\propto -E_\Sigma(\mathbf{w})
\end{aligned}$$

where

$$E_\Sigma(\mathbf{w}) = \frac{1}{2}\mathrm{tr}[(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))^\top \Sigma^{-1}]. \tag{5.4}$$

Therefore, maximizing the likelihood function with respect to $\mathbf{w}$ is tantamount to minimizing $E_\Sigma(\mathbf{w})$, assuming Σ is known. Assuming, however, that Σ is unknown, first we attempt to determine the maximum likelihood estimator of Σ. In order to do so, we differentiate (5.3) with respect to Σ, equate the result to $\mathbf{0}$ and solve for Σ, as follows

$$\begin{aligned}
\frac{\partial p(\mathbf{T}|\mathbf{y}(\mathbf{x},\mathbf{w}),\Sigma)}{\partial \Sigma} &= \mathbf{0} \\
-\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))^\top &= \mathbf{0} \qquad \text{(Apply (C.21), (C.24) and (C.28))} \\
\frac{\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))(\mathbf{t}_n - \mathbf{y}(\mathbf{x},\mathbf{w}))^\top}{N} &= \Sigma. \qquad (5.5)
\end{aligned}$$

Note that the form for the maximum likelihood estimator in (5.5) is dependent on the network parameters $\mathbf{w}$; by contrast, the maximum likelihood estimator of $\mathbf{w}$ is determined by minimizing (5.4) with respect to $\mathbf{w}$, where the function (5.4) is itself also dependent on Σ. Hence, the maximum likelihood estimation of the parameters $\mathbf{w}$ and Σ is now a coupled procedure.

Exercise 5.4

Consider the context wherein, for a data set $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$ of input and target variables, we are performing binary classification utilizing neural networks, such that for a new data point $\{\mathbf{x}, t\}$ we determine $p(t=1|\mathbf{x}) = y(\mathbf{x}, \mathbf{w})$. Consider moreover that, for all observations, there exists a probability $\epsilon \in [0,1]$ that it has been flipped to the incorrect class, as in (4.117). We denote s as the true class to which the new data point belongs to. From (4.117) we find that

$$p(s=1|\mathbf{x}, \mathbf{w}, \epsilon) = \epsilon + (1-2\epsilon)y(\mathbf{x}, \mathbf{w}) \quad \text{and} \quad p(s=0|\mathbf{x}, \mathbf{w}, \epsilon) = 1 - \epsilon - (1-2\epsilon)y(\mathbf{x}, \mathbf{w}).$$

It follows that, under this result, the likelihood function, and corresponding negative logarithm, associated with the data set (with respect to the true class labels) is

$$\begin{aligned} p(\mathbf{s}|\mathbf{w}, \epsilon) &= \{p(s_n=1|\mathbf{x}, \mathbf{w}, \epsilon)\}^{s_n}\{p(s_n=0|\mathbf{x}, \mathbf{w}, \epsilon)\}^{1-s_n} \\ &= \prod_{n=1}^N \{\epsilon + (1-2\epsilon)y(\mathbf{x}_n, \mathbf{w})\}^{s_n}\{1-\epsilon-(1-2\epsilon)y(\mathbf{x}_n, \mathbf{w})\}^{1-s_n} \\ -\log p(\mathbf{s}|\mathbf{w}, \epsilon) &= -\sum_{n=1}^N s_n \log\{\epsilon + (1-2\epsilon)y(\mathbf{x}_n, \mathbf{w})\} + \\ &\quad -\sum_{n=1}^N (1-s_n)\log\{1-\epsilon-(1-2\epsilon)y(\mathbf{x}_n, \mathbf{w})\}. \end{aligned} \tag{5.6}$$

By taking $\epsilon = 0$ in (5.6), we write

$$-\log p(\mathbf{s}|\mathbf{w}, 0) = -\sum_{n=1}^N s_n \log\{y(\mathbf{x}_n, \mathbf{w})\} - \sum_{n=1}^N (1-s_n)\log\{1-y(\mathbf{x}_n, \mathbf{w})\}.$$

This result matches the error function (5.21), as desired.

Exercise 5.5

We consider now the multiclass classification context, wherein for a data set of $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$, we find that the target variables are K-dimensional vectors whose coordinates are such that $t_{n,k} \in \{0, 1\}$. Assuming the neural network framework is adopted, such that for a new data point $\{\mathbf{x}, \mathbf{t}\}$ we determine $p(t_k|\mathbf{x}) = y_k(\mathbf{x}, \mathbf{w})$. The likelihood function associated with this data set, and its corresponding logarithm, is

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^N \prod_{k=1}^K \{y_k(\mathbf{x}_n, \mathbf{w})\}^{t_{n,k}} \{1 - y_k(\mathbf{x}_n, \mathbf{w})\}^{1-t_{n,k}}$$

$$\begin{aligned} \log p(\mathbf{T}|\mathbf{X}, \mathbf{w}) = \log \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \log\{y_k(\mathbf{x}_n, \mathbf{w})\} + \\ + \sum_{n=1}^N \sum_{k=1}^K (1 - t_{n,k}) \log\{1 - y_k(\mathbf{x}_n, \mathbf{w})\}^{1-t_{n,k}} \end{aligned}$$

$$\log p(\mathbf{T}|\mathbf{X}, \mathbf{w}) = -E(\mathbf{w}) \qquad \text{(Apply (5.23))}. \tag{5.7}$$

Similarly to the context of Exercise 5.2, we conclude that the maximization of the left-hand-side of (5.7) is analogous to the minimization of the negative of the right-hand-side.

Exercise 5.6

Consider that we adopt as output activation function in a classification context the logistic-sigmoid function, that is $y_n = \sigma(a_n)$. We return to (5.21) as follows

$$\begin{aligned}
E(\mathbf{w}) &= -\sum_{n=1}^{N} t_n \log y_n - \sum_{n=1}^{N}(1-t_n)\log\{1-y_n\} \\
&= -\sum_{n=1}^{N} t_n \log\{\sigma(a_n)\} - \sum_{n=1}^{N}(1-t_n)\log\{1-\sigma(a_n)\} \\
&= -\sum_{n=1}^{N} t_n \log\{\sigma(a_n)\} - \sum_{n=1}^{N}(1-t_n)\log\{\sigma(-a_n)\} && \text{(Apply (4.14))} \\
\frac{\partial E(\mathbf{w})}{\partial a_k} &= -t_k \frac{\sigma(a_k)\{1-\sigma(a_k)\}}{\sigma(a_k)} + (1-t_k)\frac{\sigma(-a_k)\{1-\sigma(-a_k)\}}{\sigma(-a_k)} && \text{(Apply (4.88))} \\
&= -t_k\{1-\sigma(a_k)\} + (1-t_k)\sigma(a_k) \\
\frac{\partial E(\mathbf{w})}{\partial a_k} &= \sigma(a_k) - t_k.
\end{aligned}$$

Hence, we reach the desired result.

Exercise 5.7

Let us consider the classification context, such that we obtain a data set $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$ of input and target variables, where the target variables may belong to one of K classes, and are hence binary variables in a 1-of-K coding scheme. Let us also consider that we model this data via the neural network framework, such that for a data point $\{\mathbf{x}, \mathbf{t}\}$ we determine $p(t_k = 1|\mathrm{x}) = y_k(\mathbf{x}, \mathbf{w})$. We adopt the error function (5.24), and seek to determine the form of its derivative, taken with respect to $a_{n,k} = a_k(\mathbf{x}_n, \mathbf{w})$, wherein $y_k(\mathbf{x}_n, \mathbf{w})$ is associated with $a_{n,k}$ via (5.25). It follows that

$$
\begin{aligned}
\frac{\partial E(\mathbf{w})}{\partial a_{n,k}} &= \frac{\partial}{\partial a_{n,k}}\left[-\sum_{m=1}^{N}\sum_{r=1}^{K} t_{m,r}\log y_r(\mathbf{x}_m, \mathbf{w})\right] \\
&= \frac{\partial}{\partial a_{n,k}}\left[-\sum_{m=1}^{N}\sum_{r=1}^{K} t_{m,r}\log\left\{\frac{\exp\{a_{m,r}\}}{\sum_{j=1}^{K}\exp\{a_{m,j}\}}\right\}\right] \\
&= -\frac{\partial}{\partial a_{n,k}}\left[\sum_{m=1}^{N}\sum_{r=1}^{K} t_{m,r}a_{m,r} - \sum_{m=1}^{N}\sum_{r=1}^{K} t_{m,r}\log\left\{\sum_{j=1}^{K}\exp\{a_{m,j}\}\right\}\right] \\
&= -\left[t_{n,k} - \sum_{r=1}^{K} t_{n,r}\frac{\exp\{a_{n,k}\}}{\sum_{j=1}^{K}\exp\{a_{n,j}\}}\right] \\
&= -\left[t_{n,k} - y_{n,k}\sum_{r=1}^{K} t_{n,r}\right] && \text{(Apply (5.25))} \\
\frac{\partial E(\mathbf{w})}{\partial a_{n,k}} &= y_{n,k} - t_{n,k}.
\end{aligned}
$$

Hence reaching the desired result.

Exercise 5.8

We aim to demonstrate herein that the $\tanh(a)$ function, as in (5.59), satisfies (5.60). It follows that

$$\begin{aligned}
\frac{\mathrm{d}\tanh(a)}{\mathrm{d}a} &= \frac{\mathrm{d}}{\mathrm{d}a}\left[\frac{e^a - e^{-a}}{e^a + e^{-a}}\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}a}\left[\frac{e^a}{e^a + e^{-a}} - \frac{e^{-a}}{e^a + e^{-a}}\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}a}\left[\frac{e^a}{e^a + e^{-a}} - 1 + \frac{e^a}{e^a + e^{-a}}\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}a}\left[\frac{2}{1 + e^{-2a}} - 1\right] \\
&= \frac{4e^{-2a}}{(1 + e^{-2a})^2)} \\
&= \frac{4 + (e^a - e^{-a})^2}{(e^a + e^{-a})^2} - \left(\frac{e^a - e^{-a}}{e^a + e^{-a}}\right)^2 \\
&= \frac{4 + (e^a + e^{-a})^2 - 4e^a e^{-a}}{(e^a + e^{-a})^2} - \{\tanh(a)\}^2 \quad \text{(Apply (5.59))} \\
\frac{\mathrm{d}\tanh(a)}{\mathrm{d}a} &= 1 - \{\tanh(a)\}^2.
\end{aligned}$$

Hence we conclude the relation in (5.60) is valid.

Exercise 5.9

Consider that we adapt the usual binary classification context, with logistic-sigmoid output activation function, so that the target variables are such that $t \in \{-1, 1\}$, where $t = -1$ corresponds to class $\mathcal{C}_2$ and $t = 1$ corresponds to class $\mathcal{C}_1$. Note that, consequently, for any data point $\{\mathbf{x}, t\}$ considered in this context, the corresponding likelihood function is (2.261), with $\mu = y(\mathbf{x}, \mathbf{w})$. Hence, the likelihood function associated with the data set, and its corresponding negative logarithm, is

$$\begin{aligned}
p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}) \\
&= \prod_{n=1}^{N} \left(\frac{1 - y(\mathbf{x}_n, \mathbf{w})}{2}\right)^{(1-t_n)/2} \left(\frac{1 + y(\mathbf{x}_n, \mathbf{w})}{2}\right)^{(1+t_n)/2} \qquad \text{(Apply (2.261))} \\
-\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= -\sum_{n=1}^{N} \frac{1 - t_n}{2} \log\left\{\frac{1 - y(\mathbf{x}_n, \mathbf{w})}{2}\right\} + \\
&\quad - \sum_{n=1}^{N} \frac{1 + t_n}{2} \log\left\{\frac{1 + y(\mathbf{x}_n, \mathbf{w})}{2}\right\} \\
-\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= -\frac{1}{2}\left[\sum_{n=1}^{N} (1 - t_n) \log\{1 - y(\mathbf{x}_n, \mathbf{w})\} + \right. \\
&\quad \left. + \sum_{n=1}^{N} (1 + t_n) \log\{1 + y(\mathbf{x}_n, \mathbf{w})\}\right] + N \log 2. \qquad (5.8)
\end{aligned}$$

We conclude that, by considering in (5.8) only elements dependent on $\mathbf{w}$, we may adopt the error function

$$E(\mathbf{w}) = -\sum_{n=1}^{N} [(1 - t_n) \log\{1 - y(\mathbf{x}_n, \mathbf{w})\} + (1 + t_n) \log\{1 + y(\mathbf{x}_n, \mathbf{w})\}].$$

We adapt the logistic-sigmoid output activation function to the new range as follows

$$\begin{aligned}
\tilde{\sigma}(a) &= 2\sigma(a) - 1 \\
&= 2\sigma(2a/2) - 1 \\
\tilde{\sigma}(a) &= \tanh(a/2) \qquad \text{Apply (3.1).}
\end{aligned}$$

Exercise 5.10

Consider that we inspect a Hessian matrix $\mathbf{H}$ with eigenvalue equation as in (2.45). We aim to demonstrate that $\mathbf{H}$ is positive-definite if, and only if, all its eigenvalues are positive. For that purpose, first we assume all eigenvalues are positive, hence we take an arbitrary real vector $\mathbf{v}$ such that

$$\begin{aligned}
\mathbf{v}^\top \mathbf{H}\mathbf{v} &= \mathbf{v}^\top \left[\sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top\right] \mathbf{v} && \text{(Apply (2.48))} \\
&= \sum_{i=1}^{D} \lambda_i \{\mathbf{v}^\top \mathbf{u}_i\}^2 \\
\mathbf{v}^\top \mathbf{H}\mathbf{v} &> 0 && \text{(By assumption).}
\end{aligned}$$

Hence, we conclude that the assumption that all eigenvalues are positive implies $\mathbf{H}$ is positive definite. Conversely, if we assume $\mathbf{H}$ is positive definite, it follows that, for all real vector $\mathbf{v}$, $\mathbf{v}^\top \mathbf{H}\mathbf{v}$. Particularly, we may choose $\mathbf{v} = \mathbf{u}_i$, where $\mathbf{u}_i$ is the i-th eigenvector associated with $\mathbf{H}$. It follows that

$$\begin{aligned}
\mathbf{u}_i^\top \mathbf{H}\mathbf{u}_i &> 0 && \text{(By assumption)} \\
\lambda_i \mathbf{u}_i^\top \mathbf{u}_i &> 0 && \text{(Apply (2.45))} \\
\lambda_i &> 0 && \text{(Orthonormality of } \mathbf{u}_i\text{).}
\end{aligned}$$

As this holds for any arbitrary eigenvector associated with $\mathbf{H}$, we hence conclude that all eigenvalues must be positive, and hence conclude our demonstration.

Exercise 5.11

Consider the local quadratic approximation of an error function $E(\mathbf{w})$ around $\mathbf{w}^\star$, as in (5.32). It follows, by applying an eigendecomposition of the Hessian matrix, as in (2.48), and applying (5.35), that we obtain

$$E(\mathbf{w}) \approx E(\mathbf{w}^\star) + \frac{1}{2}\sum_{i=1}^{D} \lambda_i \alpha_i^2.$$

Note, from (5.35), that the values α_i are scalars which convert $\mathbf{w} - \mathbf{w}^\star$ to linear combinations of the eigenvectors $\mathbf{u}_i$, hence they are aligned with the eigenvectors. Let us determine a fixed value $C > E(\mathbf{w}^\star)$ such that

$$\begin{aligned}
E(\mathbf{w}^\star) + \frac{1}{2}\sum_{i=1}^{D} \lambda_i \alpha_i^2 &= C \\
\sum_{i=1}^{D} \left(\frac{\alpha_i}{1/\sqrt{\lambda_i}}\right)^2 &= 2C - 2E(\mathbf{w}^\star) \\
\sum_{i=1}^{D} \left(\frac{\alpha_i}{1/\sqrt{\lambda_i}}\right)^2 &= \left\{\sqrt{C^\star}\right\}^2 \\
\sum_{i=1}^{D} \left(\frac{\alpha_i}{\sqrt{C^\star/\lambda_i}}\right)^2 &= 1, \qquad (5.9)
\end{aligned}$$

Where $C^\star = 2C - 2E(\mathbf{w}^\star)$. It follows that (5.9) is the formula for a D-dimensional ellipsoid with semi-axes lengths $\sqrt{C^\star/\lambda_i}$. Hence, the semi-axes lengths are inversely proportional to the square-root of the eigenvalues. We thereby conclude that, approximately, the contours of constant error C constitute ellipsoids whose axes are aligned with the eigenvectors of $\mathbf{H}$, and whose lengths are inversely proportional to the square root of the eigenvalues of $\mathbf{H}$.

Exercise 5.12

We aim to demonstrate that a stationary point $\mathbf{w}^\star$ of an error function constitutes a local minima if, and only if, the Hessian matrix associated with said error function is positive definite. By adopting the quadratic approximation to the error function as in (5.32), let us first assume that $\mathbf{w}^\star$ is a local minima. It follows that, for all $\mathbf{w}$ in the neighbourhood of $\mathbf{w}^\star$, it is true that

$$\begin{aligned} E(\mathbf{w}) &\geq E(\mathbf{w}^\star) \\ E(\mathbf{w}) - E(\mathbf{w}^\star) &\geq 0. \end{aligned} \tag{5.10}$$

From (5.32), we find that

$$\begin{aligned} E(\mathbf{w}) &\approx E(\mathbf{w}^\star) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^\star) \\ 2\{E(\mathbf{w}) - E(\mathbf{w}^\star)\} &\approx (\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^\star) \\ 0 &\leq (\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^\star) && \text{(Apply (5.10))} \\ 0 &\leq \sum_{i=1}^{D}\sum_{j=1}^{D} \alpha_i \alpha_j \mathbf{u}_i^\top \mathbf{H}\mathbf{u}_j && \text{(Apply (5.35))} \\ 0 &\leq \sum_{i=1}^{D}\sum_{j=1}^{D} \alpha_i \alpha_j \lambda_j \mathbf{u}_i^\top \mathbf{u}_j && \text{(Apply (2.45))} \\ 0 &\leq \sum_{i=1}^{D} \alpha_i^2 \lambda_i && \text{(Orthonormality of } \mathbf{u}_i\text{).} \end{aligned} \tag{5.11}$$

Note that (5.10) is only valid for $\mathbf{w}$ in the neighbourhood of $\mathbf{w}^\star$. As $\alpha_i^2 > 0$, we note that in order for (5.11) to be greater than or equal to zero, it must follow that $\lambda_i \geq 0$ for all $i \in \{1, \dots, D\}$. Hence, all eigenvalues associated with the Hessian evaluated at $\mathbf{w}^\star$ must be nonnegative, hence the Hessian must be positive semi-definite when evaluated at $\mathbf{w}^\star$. Now, we assume that $\mathbf{H}$ is positive definite when evaluated at $\mathbf{w}^\star$. It follows that, for any choice of $\mathbf{w} - \mathbf{w}^\star$ we find that

$$\frac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^\star) > 0. \tag{5.12}$$

Hence, from (5.32) it follows that

$$\begin{aligned} E(\mathbf{w}) &\approx E(\mathbf{w}^\star) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^\star) \\ E(\mathbf{w}) - E(\mathbf{w}^\star) &\approx \frac{1}{2}(\mathbf{w}-\mathbf{w}^\star)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^\star) \\ E(\mathbf{w}) - E(\mathbf{w}^\star) &> 0 && \text{(Apply (5.12)).} \end{aligned} \tag{5.13}$$

That is, for all values of $\mathbf{w}$ in the neighbourhood in which the quadratic approximation (5.32) is valid, likewise (5.13) holds, that is, $\mathbf{w}^\star$ is a minimum for the error function $E(\mathbf{w})$. We therefore conclude that, given the quadratic approximation in (5.32), $\mathbf{w}^\star$ is a local minimum for the error function $E(\mathbf{w})$ if, and only if, the associated Hessian, evaluated at $\mathbf{w}^\star$, is positive semi-definite.

Exercise 5.13

Let $\mathbf{w} \in \mathbb{R}^W$ be the adaptive parameters associated with a neural network, and consider the quadratic approximation for an error function as in (5.28). Trivially, the term $\mathbf{b} \in \mathbb{R}$ is composed of W terms. As the Hessian matrix $\mathbf{H}$ is a symmetric matrix of dimensions $W \times W$, it is composed of $W(W+1)/2$ terms, as seen in Exercise 2.21. It follows that the total number of independent elements in (5.28) is

$$\begin{aligned} W + \frac{W(W+1)}{2} &= \frac{2W + W^2 + W}{2} \\ &= \frac{W(W+3)}{2}. \end{aligned}$$

Exercise 5.14

We denote the error function evaluated at the n-th data point as $E_n(\mathbf{w})$. We aim to demonstrate that for the central differences approximation to the derivative of the n-th term with respect to the (j,i)-th adaptive parameter, as in (5.69), the $O(\epsilon)$ terms vanish. The quadratic order Taylor polynomial approximation of E_n, centred at $w_{j,i}$ is such that, evaluated at $w_{j,i}+\epsilon$, for $\epsilon > 0$, we obtain

$$E_n(w_{j,i}+\epsilon) = E(w_{j,i}) + \frac{\partial E(w_{i,j})}{\partial w_{i,j}}\epsilon + \frac{\partial^2 E(w_{i,j})}{\partial w_{i,j}^2}\epsilon^2 + O(\epsilon^3). \tag{5.14}$$

Conversely, when evaluated at $w_{j,i}-\epsilon$, for $\epsilon > 0$, we obtain

$$E_n(w_{j,i}-\epsilon) = E(w_{j,i}) - \frac{\partial E(w_{i,j})}{\partial w_{i,j}}\epsilon + \frac{\partial^2 E(w_{i,j})}{\partial w_{i,j}^2}\epsilon^2 + O(\epsilon^3) \tag{5.15}$$

We may thereby rewrite (5.69) as

$$\begin{aligned}
\frac{\partial E_n(w_{j,i})}{\partial w_{j,i}} &= \frac{E_n(w_{j,i}+\epsilon) - E_n(w_{j,i}+\epsilon)}{2\epsilon} + O(\epsilon^2) \\
&= \frac{E(w_{j,i}) + \frac{\partial E(w_{i,j})}{\partial w_{i,j}}\epsilon + \frac{\partial^2 E(w_{i,j})}{\partial w_{i,j}^2}\epsilon^2 + O(\epsilon^3)}{2\epsilon} + O(\epsilon^2) \\
&\quad + \frac{-E(w_{j,i}) + \frac{\partial E(w_{i,j})}{\partial w_{i,j}}\epsilon - \frac{\partial^2 E(w_{i,j})}{\partial w_{i,j}^2}\epsilon^2 + O(\epsilon^3)}{2\epsilon} \qquad \text{(Apply (5.14) and (5.15))} \\
\frac{\partial E_n(w_{j,i})}{\partial w_{j,i}} &= \frac{\partial E_n(w_{j,i})}{\partial w_{j,i}} + O(\epsilon^2).
\end{aligned}$$

Hence, we conclude that the $O(\epsilon)$ terms vanish.

Exercise 5.15

We consider a general neural network, with fixed hidden unit activation function denoted by $h(a)$, and fixed output unit activation function denoted by $\sigma(a)$. Note that any output unit is consequently such that $y_k = \sigma(a_k)$. We aim to determine a forward propagation approach to determining the Jacobian matrix associated with our neural network, written as in (5.70). It follows that

$$\begin{aligned} J_{k,i} &= \frac{\partial y_k}{\partial x_i} \\ &= \frac{\partial y_k}{\partial a_k}\frac{\partial a_k}{\partial x_i} \\ &= \frac{\mathrm{d}\sigma(a_k)}{\mathrm{d}a_k}\sum_j \frac{\partial a_k}{\partial a_j}\frac{\partial a_j}{\partial x_i} \\ J_{k,i} &= \frac{\mathrm{d}\sigma(a_k)}{\mathrm{d}a_k}\sum_j w_{k,j}\frac{\mathrm{d}h(a_j)}{\mathrm{d}a_j}\frac{\partial a_j}{\partial x_i} \end{aligned} \tag{5.16}$$

Note that the sum in (5.16) runs over all units j which send connections to unit k. We thereafter compute $\partial a_j/\partial x_i$ as

$$\begin{aligned} \frac{\partial a_j}{\partial x_i} &= \sum_r \frac{\partial a_j}{\partial a_r}\frac{\partial a_r}{\partial x_i} \\ \frac{\partial a_j}{\partial x_i} &= \sum_r w_{j,r}\frac{\mathrm{d}h(a_r)}{\mathrm{d}a_r}\frac{\partial a_r}{\partial x_i} \end{aligned} \tag{5.17}$$

Where the sum in (5.17) runs over all units r which send connections to unit j. If the input unit i sends connections to the unit r, we find that

$$\frac{\partial a_r}{\partial x_i} = w_{r,i}. \tag{5.18}$$

Otherwise, we simply recursively apply (5.17). By analogy to the backward propagation formalism proposed prior, we may summarize the forward propagation as follows: as usual, we apply the desired input unit x_i, and forward propagate it across the network, obtaining the consequent activations of hidden and output units. Next, for each unit j which sends connections to unit k, as in (5.16), recursively compute $\partial a_j/\partial x_i$ as in (5.17), starting from those to which the input unit i sends connections to, as in (5.18). Once those values are obtained, apply (5.16).

Exercise 5.16

Let a data set $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$ be observed, and consider that we adopt the the squared-error loss function, as in (5.11). The corresponding Hessian is as follows

$$\begin{aligned}
\mathbf{H} &= \nabla\nabla^\top E(\mathbf{w}) \\
&= \nabla\nabla^\top \left[\frac{1}{2}\sum_{n=1}^N ||\mathbf{y}_n - \mathbf{t}_n||^2\right] && \text{(Apply (5.11))} \\
&= \frac{1}{2}\sum_{n=1}^N \nabla\nabla^\top \{\mathbf{y}_n - \mathbf{t}_n\}^\top \{\mathbf{y}_n - \mathbf{t}_n\} \\
&= \frac{1}{2}\sum_{n=1}^N \nabla\nabla^\top \{\mathbf{y}_n^\top - \mathbf{t}_n^\top\} \{\mathbf{y}_n - \mathbf{t}_n\} \\
&= \frac{1}{2}\sum_{n=1}^N \nabla\nabla^\top \{\mathbf{y}_n^\top \mathbf{y}_n - 2\mathbf{t}_n^\top \mathbf{y}_n + \mathbf{t}_n^\top \mathbf{t}_n\} \\
&= \frac{1}{2}\sum_{n=1}^N \nabla \left\{2\mathbf{y}_n^\top \nabla^\top \mathbf{y}_n - 2\mathbf{t}_n^\top \nabla^\top \mathbf{y}_n\right\} \\
&= \sum_{n=1}^N \left\{\nabla \mathbf{y}_n^\top \nabla^\top \mathbf{y}_n + \mathbf{y}_n^\top \nabla\nabla^\top \mathbf{y}_n - \mathbf{t}_n^\top \nabla\nabla^\top \mathbf{y}_n\right\} \\
\mathbf{H} &= \sum_{n=1}^N \nabla \mathbf{y}_n^\top \nabla^\top \mathbf{y}_n + \sum_{n=1}^N (\mathbf{y}_n - \mathbf{t}_n)^\top \nabla\nabla^\top \mathbf{y}_n. && (5.19)
\end{aligned}$$

We may adopt similar reasoning to the univariate case in order to justify the negligibility of the second term in the right-hand-side of (5.19), hence we reach the approximation

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

where $\mathbf{B}_n = \nabla \mathbf{y}_n^\top$.

Exercise 5.17

Exercise 5.18

Exercise 5.19

Exercise 5.20

Exercise 5.21

Exercise 5.22

Exercise 5.23

Exercise 5.24

Exercise 5.25

Exercise 5.26

Exercise 5.27

Exercise 5.28

Exercise 5.29

Exercise 5.30

Exercise 5.31

Exercise 5.32

Exercise 5.33

Exercise 5.34

Exercise 5.35

Exercise 5.36

Exercise 5.37

Exercise 5.38

Exercise 5.39

Exercise 5.40

Exercise 5.41

Referenced Formulae

This chapter contains a list of all equations in the original textbook which are referenced for Exercise solutions, numbered as they were originally. The Exercises which reference them are highlighted, and they are presented in the same order as they are presented on the source book. The notation adopted herein is not necessarily consistent with that which is adopted throughout the remainder of this document, and follows as closely as possible that which is adopted in the original textbook. Formulae are presented deprived of the surrounding context, so access to the original text remains imperative in order to fully understand the solutions.

(1.1) Page 5. Referenced in Exercises: 1.1.

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

(1.2) Page 5. Referenced in Exercises: 1.1.

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2.$$

(1.4) Page 10. Referenced in Exercises: 1.2.

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2$$

(1.10) Page 14. Referenced in Exercises: 1.3, 1.39.

$$p(X) = \sum_{Y} p(X, Y).$$

(1.12) Page 15. Referenced in Exercises: 1.3, 3.24.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}.$$

(1.26) Page 18. Referenced in Exercises: 1.10, 2.4, 2.10, 2.11, 2.42, 2.46, 2.48, 2.49, 3.13, 3.23.

$$\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1.$$

(1.27) Page 18. Referenced in Exercises: 1.4, 1.32.

$$\begin{aligned} p_y(y) &= p_x(y) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\ &= p_x(g(y)) |g'(y)|. \end{aligned}$$

(1.30) Page 19. Referenced in Exercises: 1.25, 1.41, 2.13, 2.15, 2.27, 2.48, 2.60, 3.19, 3.23, 3.24.

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

(1.31) Page 19. Referenced in Exercises: 1.37, 2.8, 3.23.

$$p(x) = \int p(x, y)\, \mathrm{d}y.$$

(1.32) Page 19. Referenced in Exercises: 1.25, 1.26, 1.27, 1.37, 1.41, 2.8, 2.46, 2.48, 2.49, 3.13, 4.22.

$$p(x, y) = p(y|x)p(x).$$

(1.33) Page 19. Referenced in Exercises: 2.1, 2.2, 2.4.

$$\mathbb{E}[f] = \sum_x p(x) f(x).$$

(1.34) Page 19. Referenced in Exercises: 1.6, 1.8, 1.10, 2.6, 2.8, 2.10, 2.11, 2.12, 2.27, 2.49, 2.58.

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

(1.35) Page 19. Referenced in Exercises: 4.23.

$$\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n).$$

(1.37) Page 20. Referenced in Exercises: 1.25, 1.26, 2.8.

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x).$$

(1.38) Page 20. Referenced in Exercises: 1.5, 1.8, 1.13, 2.1, 2.2, 2.8, 2.27.

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2].$$

(1.39) Page 20. Referenced in Exercises: 1.5, 1.8, 2.4, 2.6, 2.10, 2.12, 2.42.

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2.$$

(1.41) Page 20. Referenced in Exercises: 1.6, 2.10.

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \end{aligned}$$

(1.42) Page 20. Referenced in Exercises: 2.49, 2.58.

$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y} - \mathbb{E}[\mathbf{y}]\}] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{xy}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]. \end{aligned}$$

(1.46) Page 24. Referenced in Exercises: 1.7, 1.8, 1.30, 1.35, 2.16, 2.34, 2.36, 2.38, 2.39, 2.42, 2.44, 2.46, 3.8, 3.13, 4.26.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}.$$

(1.48) Page 25. Referenced in Exercises: 1.8, 4.26.

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)\, \mathrm{d}x = 1.$$

(1.49) Page 25. Referenced in Exercises: 1.8.

$$\int_{-\infty}^{\infty} x\mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x = 1.$$

(1.54) Page 27. Referenced in Exercises: 1.11.

$$\ln p(\mathbf{x}|\mu,\sigma^2) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi).$$

(1.55) Page 27. Referenced in Exercises: 1.11, 2.38, 2.39.

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

(1.56) Page 27. Referenced in Exercises: 1.11.

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2.$$

(1.78) Page 39. Referenced in Exercises: 1.21.

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

(1.91) Page 48. Referenced in Exercises: 1.27.

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

(1.98) Page 51. Referenced in Exercises: 1.28, 1.29, 2.1, 2.2.

$$\mathrm{H}[p] = -\sum_i p(x_i)\ln p(x_i).$$

(1.104) Page 53. Referenced in Exercises: 1.32, 1.35, 1.37, 1.41, 2.15.

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

(1.111) Page 54. Referenced in Exercises: 1.33, 1.37, 1.41.

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x}.$$

(1.112) Page 55. Referenced in Exercises: 1.31.

$$\mathrm{H}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}].$$

(1.113) Page 56. Referenced in Exercises: 1.30, 1.41, 2.13.

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

(1.114) Page 56. Referenced in Exercises: 1.36, 1.38.

$$f(\lambda a + (1-\lambda)b) < \lambda f(a) + (1-\lambda)f(b).$$

(1.115) Page 56. Referenced in Exercises: 1.29, 1.38, 1.40, 4.20.

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i).$$

(1.120) Page 57. Referenced in Exercises: 1.31, 1.41.

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

(1.121) Page 57. Referenced in Exercises: 1.31.

$$\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}].$$

(1.124) Page 59. Referenced in Exercises: 1.7, 1.8, 1.18, 4.21.

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) \mathrm{d}x.$$

(1.131) Page 60. Referenced in Exercises: 1.14.

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{i,j} x_i x_j.$$

(1.133) Page 60. Referenced in Exercises: 1.15.

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1,i_2,\ldots,i_M} x_{i_1} x_{i_2} \ldots x_{i_M}.$$

(1.136) Page 61. Referenced in Exercises: 1.15.

$$\sum_{i=1}^{D} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}.$$

(1.137) Page 61. Referenced in Exercises: 1.15.

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

(1.139) Page 61. Referenced in Exercises: 1.16.

$$N(D,M) = \frac{(D+M)!}{D!M!}.$$

(1.140) Page 61. Referenced in Exercises: 1.16.

$$n! \approx n^n e^{-n}.$$

(1.141) Page 62. Referenced in Exercises: 1.17, 2.5, 2.41, 2.43.

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \, \mathrm{d}u.$$

(1.142) Page 62. Referenced in Exercises: 1.18.

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \, \mathrm{d}x_i = S_D \int_0^\infty e^{-r^2} r^{D-1} \, \mathrm{d}r.$$

(1.143) Page 62. Referenced in Exercises: 1.18.

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)}.$$

(1.144) Page 62. Referenced in Exercises: 1.18.

$$V_D = \frac{S_D}{D}.$$

(1.146) Page 62. Referenced in Exercises: 1.19.

$$\Gamma(x+1) \approx (2\pi)^{1/2} e^{-x} x^{x+1/2}.$$

(1.147) Page 63. Referenced in Exercises: 1.20.

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

(1.151) Page 64. Referenced in Exercises: 1.25, 1.26.

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

(2.2) Page 69. Referenced in Exercises: 2.1, 2.7.

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

(2.13) Page 71. Referenced in Exercises: 2.6, 2.7, 2.56.

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

(2.38) Page 76. Referenced in Exercises: 2.10.

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

(2.43) Page 78. Referenced in Exercises: 2.13, 2.15, 2.17, 2.37, 2.40, 2.45, 2.48, 2.57, 3.7, 3.8, 3.12, 3.23, 3.24, 4.8, 4.10.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\}.$$

(2.44) Page 80. Referenced in Exercises: 2.23.

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

(2.45) Page 80. Referenced in Exercises: 2.18, 2.19, 5.12.

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i.$$

(2.46) Page 80. Referenced in Exercises: 3.21.

$$\mathbf{u}_i^\top \mathbf{u}_j = I_{ij}.$$

(2.48) Page 80. Referenced in Exercises: 2.18, 2.20, 3.21, 5.11.

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top.$$

(2.49) Page 80. Referenced in Exercises: 3.2, 3.21.

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top.$$

(2.50) Page 80. Referenced in Exercises: 2.23.

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i}{\lambda_i}.$$

(2.55) Page 81. Referenced in Exercises: 2.23.

$$|\boldsymbol{\Sigma}| = \prod_{j=1}^{D} \lambda_j^{1/2}.$$

(2.59) Page 82. Referenced in Exercises: 2.13, 2.35, 2.49, 4.24.

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

(2.62) Page 83. Referenced in Exercises: 2.13.

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}.$$

(2.64) Page 83. Referenced in Exercises: 2.13, 2.15, 2.35, 2.49, 4.24.

$$\operatorname{cov}[\mathbf{x}] = \boldsymbol{\Sigma}.$$

(2.76) Page 87. Referenced in Exercises: 2.24, 4.2.

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}.$$

(2.77) Page 87. Referenced in Exercises: 4.2.

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}.$$

(2.81) Page 87. Referenced in Exercises: 2.25, 2.28.

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b).$$

(2.82) Page 87. Referenced in Exercises: 2.25, 2.28.

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}.$$

(2.92) Page 89. Referenced in Exercises: 2.25, 2.28, 4.24.

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

(2.93) Page 89. Referenced in Exercises: 2.25, 2.28, 4.24.

$$\text{cov}[\mathbf{x}_a] = \boldsymbol{\Sigma}_{aa}.$$

(2.104) Page 92. Referenced in Exercises: 2.29, 2.30.

$$\mathbf{R} = \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top\mathbf{L}\mathbf{A} & -\mathbf{A}^\top\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{pmatrix}.$$

(2.105) Page 92. Referenced in Exercises: 2.28, 2.29, 2.30.

$$\text{cov}[\mathbf{z}] = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\top \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top \end{pmatrix}$$

(2.107) Page 92. Referenced in Exercises: 2.28, 2.30.

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1}\begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\top\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix}$$

(2.108) Page 92. Referenced in Exercises: 2.28, 2.30.

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}$$

(2.113) Page 93. Referenced in Exercises: 2.31, 3.9, 3.10, 3.13, 3.16.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}).$$

(2.114) Page 93. Referenced in Exercises: 2.31, 3.9, 3.10, 3.13, 3.16.

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}).$$

(2.115) Page 93. Referenced in Exercises: 2.31, 3.10, 3.16.

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top).$$

(2.116) Page 93. Referenced in Exercises: 3.9, 3.13.

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^\top\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}).$$

(2.118) Page 93. Referenced in Exercises: 2.34, 4.23, 5.2.

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).$$

(2.121) Page 93. Referenced in Exercises: 2.34, 2.35.

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

(2.135) Page 96. Referenced in Exercises: 2.36, 2.37, 2.39.

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}[-\ln p(x_N|\theta^{(N-1)})].$$

(2.137) Page 97. Referenced in Exercises: 2.38, 2.44, 3.7, 3.12.

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

(2.146) Page 100. Referenced in Exercises: 2.41, 2.42, 2.44, 2.46, 2.48, 2.49, 2.56, 3.12, 3.23, 3.24.

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

(2.155) Page 102. Referenced in Exercises: 2.45.

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right).$$

(2.168) Page 106. Referenced in Exercises: 2.55.

$$\bar{x}_1 = \bar{r}\cos\bar{\theta} = \frac{1}{N}\sum_{n=1}^{N}\cos\theta_n \qquad \bar{x}_2 = \bar{r}\sin\bar{\theta} = \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n.$$

(2.169) Page 106. Referenced in Exercises: 2.55.

$$\bar{\theta} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}.$$

(2.179) Page 108. Referenced in Exercises: 2.52, 2.54.

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

(2.182) Page 109. Referenced in Exercises: 2.53.

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

(2.184) Page 109. Referenced in Exercises: 2.53, 2.55.

$$\theta_0^{\mathrm{ML}} = \arctan\left\{\frac{\sum_{i=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}\right\}.$$

(2.185) Page 109. Referenced in Exercises: 2.55.

$$A(m_{\mathrm{ML}}) = \frac{1}{N}\sum_{n=1}^{N}\cos(\theta_n - \theta_0^{\mathrm{ML}}).$$

(2.194) Page 113. Referenced in Exercises: 2.56, 2.57.

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\top}\mathbf{u}(\mathbf{x})\}.$$

(2.195) Page 113. Referenced in Exercises: 2.58.

$$g(\boldsymbol{\eta})\int h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\top}\mathbf{u}(\mathbf{x})\}\,\mathrm{d}\mathbf{x} = 1.$$

(2.236) Page 119. Referenced in Exercises: 2.59.

$$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right).$$

(2.246) Page 122. Referenced in Exercises: 2.61.

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

(2.261) Page 127. Referenced in Exercises: 2.2, 5.9.

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

(2.262) Page 127. Referenced in Exercises: 2.3.

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}.$$

(2.263) Page 128. Referenced in Exercises: 2.3.

$$(1+x)^N = \sum_{m=0}^{N}\binom{N}{m}x^m$$

(2.264) Page 128. Referenced in Exercises: 2.3, 2.4.

$$\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = 1.$$

(2.272) Page 129. Referenced in Exercises: 2.9.

$$p_M(x_1,\ldots,x_{M-1}) = C_M\prod_{k=1}^{M-1}x_k^{\alpha_k-1}\left(1-\sum_{k=1}^{M-1}x_k\right)^{\alpha_M-1}.$$

(2.277) Page 130. Referenced in Exercises: 2.11.

$$\psi(a) = \frac{\mathrm{d}}{\mathrm{d}a}\ln\Gamma(a).$$

(2.278) Page 130. Referenced in Exercises: 2.12.

$$\mathrm{U}(x|a,b) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

(2.288) Page 132. Referenced in Exercises: 2.25.

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_\mathrm{a} \\ \boldsymbol{\mu}_\mathrm{b} \\ \boldsymbol{\mu}_\mathrm{c} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathrm{a,a}} & \boldsymbol{\Sigma}_{\mathrm{a,b}} & \boldsymbol{\Sigma}_{\mathrm{a,c}} \\ \boldsymbol{\Sigma}_{\mathrm{b,a}} & \boldsymbol{\Sigma}_{\mathrm{b,b}} & \boldsymbol{\Sigma}_{\mathrm{b,c}} \\ \boldsymbol{\Sigma}_{\mathrm{c,a}} & \boldsymbol{\Sigma}_{\mathrm{c,b}} & \boldsymbol{\Sigma}_{\mathrm{c,c}} \end{pmatrix}.$$

(2.289) Page 132. Referenced in Exercises: 2.26, 3.11.

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}.$$

(2.293) Page 134. Referenced in Exercises: 2.43.

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

(2.296) Page 135. Referenced in Exercises: 2.51.

$$\exp(iA) = \cos A + i \sin A.$$

(3.6) Page 139. Referenced in Exercises: 3.1, 4.7, 4.12, 4.14, 4.25.

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

(3.11) Page 141. Referenced in Exercises: 3.17.

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}). \end{aligned}$$

(3.52) Page 153. Referenced in Exercises: 3.17.

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}).$$

(3.54) Page 153. Referenced in Exercises: 3.14.

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}.$$

(3.62) Page 159. Referenced in Exercises: 3.14.

$$k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}').$$

(3.77) Page 166. Referenced in Exercises: 3.17.

$$p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}|\mathbf{w}, \beta) p(\mathbf{w}|\alpha) \, \mathrm{d}\mathbf{w}.$$

(3.78) Page 166. Referenced in Exercises: 3.17, 3.19.

$$p(\mathbf{t}|\alpha,\beta)=\left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w}.$$

(3.79) Page 167. Referenced in Exercises: 3.17, 3.18.

$$\begin{aligned}E(\mathbf{w})&=\beta E_D(\mathbf{w})+\alpha E_W(\mathbf{w})\\&=\frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi}\mathbf{w}||^2+\frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}.\end{aligned}$$

(3.80) Page 167. Referenced in Exercises: 3.18, 3.19.

$$E(\mathbf{w})=E(\mathbf{m}_N)+\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top\mathbf{A}(\mathbf{w}-\mathbf{m}_N).$$

(3.81) Page 167. Referenced in Exercises: 3.16, 3.18, 3.20, 3.22.

$$\mathbf{A}=\alpha\mathbf{I}+\beta\mathbf{\Phi}^\top\mathbf{\Phi}.$$

(3.82) Page 167. Referenced in Exercises: 3.15, 3.16, 3.18, 3.20, 3.21, 3.22.

$$E(\mathbf{m}_N)=\frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N||^2+\frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N.$$

(3.84) Page 167. Referenced in Exercises: 3.16, 3.18.

$$\mathbf{m}_N=\beta\mathbf{A}^{-1}\mathbf{\Phi}^\top\mathbf{t}.$$

(3.86) Page 167. Referenced in Exercises: 3.20, 3.21.

$$\ln p(\mathbf{t}|\alpha,\beta)=\frac{M}{2}\ln\alpha+\frac{N}{2}\ln\beta-E(\mathbf{m}_N)-\frac{1}{2}\ln|\mathbf{A}|-\frac{N}{2}\ln(2\pi).$$

(3.87) Page 168. Referenced in Exercises: 3.20, 3.22.

$$(\beta\mathbf{\Phi}^\top\mathbf{\Phi})\mathbf{u}_i=\lambda_i\mathbf{u}_i.$$

(3.91) Page 169. Referenced in Exercises: 3.20, 3.22.

$$\gamma=\sum_i\frac{\lambda_i}{\alpha+\lambda_i}.$$

(3.92) Page 169. Referenced in Exercises: 3.15, 3.20.

$$\alpha=\frac{\gamma}{\mathbf{m}_N^\top\mathbf{m}_N}.$$

(3.95) Page 169. Referenced in Exercises: 3.15, 3.22.

$$\frac{1}{\beta}=\frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n-\mathbf{m}_N^\top\phi(\mathbf{x}_n)\}^2.$$

(3.101) Page 173. Referenced in Exercises: 3.1.

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma \left(\frac{x - \mu_j}{s} \right).$$

(3.102) Page 173. Referenced in Exercises: 3.1.

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh \left(\frac{x - \mu_j}{2s} \right).$$

(3.104) Page 174. Referenced in Exercises: 3.3.

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2.$$

(3.105) Page 174. Referenced in Exercises: 3.4.

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i.$$

(3.115) Page 176. Referenced in Exercises: 3.14.

$$\sum_{n=1}^{N} \psi_j(\mathbf{x}_n) \psi_k(\mathbf{x}_n) = I_{j,k}.$$

(3.117) Page 177. Referenced in Exercises: 3.21, 3.22.

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln |\mathbf{A}| = \mathrm{Tr} \left(\mathbf{A}^{-1} \frac{\mathrm{d}}{\mathrm{d}\alpha} \mathbf{A} \right).$$

(4.15) Page 185. Referenced in Exercises: 4.2.

$$E_D(\tilde{\mathbf{W}}) = \frac{1}{2} \mathrm{Tr} \Big\{ (\tilde{\mathbf{X}} \tilde{\mathbf{W}} - \tilde{\mathbf{T}})^\top (\tilde{\mathbf{X}} \tilde{\mathbf{W}} - \tilde{\mathbf{T}}) \Big\}.$$

(4.16) Page 185. Referenced in Exercises: 4.2.

$$\tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{T}.$$

(4.17) Page 185. Referenced in Exercises: 4.2.

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^\top \tilde{\mathbf{x}} = \mathbf{T}^\top \left(\tilde{\mathbf{X}}^\dagger \right)^\top \tilde{\mathbf{x}}.$$

(4.18) Page 185. Referenced in Exercises: 4.2.

$$\mathbf{a}^\top \mathbf{t}_n + b = 0.$$

(4.19) Page 185. Referenced in Exercises: 4.2.

$$\mathbf{a}^\top \mathbf{y}(\mathbf{x}) + b = 0.$$

(4.20) Page 187. Referenced in Exercises: 4.5.

$$y = \mathbf{w}^\top \mathbf{x}.$$

(4.21) Page 187. Referenced in Exercises: 4.6.

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n.$$

(4.22) Page 187. Referenced in Exercises: 4.4, 4.5.

$$m_2 - m_1 = \mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1).$$

(4.23) Page 187. Referenced in Exercises: 4.5.

$$m_k = \mathbf{w}^\top \mathbf{m}_k.$$

(4.24) Page 188. Referenced in Exercises: 4.5.

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2.$$

(4.25) Page 188. Referenced in Exercises: 4.5.

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}.$$

(4.26) Page 189. Referenced in Exercises: 4.5.

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_\mathrm{B} \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_\mathrm{W} \mathbf{w}}.$$

(4.27) Page 189. Referenced in Exercises: 4.5, 4.6.

$$\mathbf{S}_\mathrm{B} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top.$$

(4.28) Page 189. Referenced in Exercises: 4.5, 4.6.

$$\mathbf{S}_\mathrm{W} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top.$$

(4.33) Page 190. Referenced in Exercises: 4.6.

$$\sum_{n=1}^{N} (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n) \mathbf{x}_n = 0.$$

(4.34) Page 190. Referenced in Exercises: 4.6.

$$w_0 = -\mathbf{w}^\top \mathbf{m}.$$

(4.36) Page 190. Referenced in Exercises: 4.6.

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{1}{N} (N_1 \mathbf{m}_1 + N_2 \mathbf{m}_2).$$

(4.37) Page 190. Referenced in Exercises: 4.6.

$$\left(\mathbf{S}_{\mathrm{W}}+\frac{N_1N_2}{N}\mathbf{S}_{\mathrm{B}}\right)\mathbf{w}=N(\mathbf{m}_1-\mathbf{m}_2).$$

(4.57) Page 197. Referenced in Exercises: 4.8.

$$\begin{aligned}p(\mathcal{C}_1|\mathbf{x})&=\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)|p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)|p(\mathcal{C}_2)}\\&=\frac{1}{1+\exp(-a)}=\sigma(a).\end{aligned}$$

(4.58) Page 197. Referenced in Exercises: 4.8.

$$a=\ln\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}.$$

(4.63) Page 198. Referenced in Exercises: 4.11.

$$a_k=\ln(p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)).$$

(4.65) Page 198. Referenced in Exercises: 4.8.

$$p(\mathcal{C}_1|\mathbf{x})=\sigma(\mathbf{w}^\top\mathbf{x}+w_0).$$

(4.66) Page 198. Referenced in Exercises: 4.8.

$$\mathbf{w}=\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).$$

(4.67) Page 198. Referenced in Exercises: 4.8.

$$w_0=-\frac{1}{2}\boldsymbol{\mu}_1^\top\Sigma^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^\top\Sigma^{-1}\boldsymbol{\mu}_2+\ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.$$

(4.88) Page 205. Referenced in Exercises: 4.12, 4.13, 4.14, 4.25.

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a}=\sigma(1-\sigma).$$

(4.89) Page 205. Referenced in Exercises: 4.14.

$$p(\mathbf{t}|\mathbf{w})=\prod_{n=1}^{N}y_n^{t_n}(1-y_n)^{1-t_n}.$$

(4.90) Page 206. Referenced in Exercises: 4.13.

$$E(\mathbf{w})=-\sum_{n=1}^{N}\{t_n\ln y_n+(1-t_n)\ln(1-y_n)\}.$$

(4.91) Page 206. Referenced in Exercises: 4.13.

$$\nabla E(\mathbf{w})=\sum_{n=1}^{N}(y_n-t_n)\phi_n.$$

(4.97) Page 207. Referenced in Exercises: 4.15.

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^\top = \mathbf{\Phi}^\top \mathbf{R}\mathbf{\Phi}.$$

(4.104) Page 209. Referenced in Exercises: 4.17, 4.20.

$$p(\mathcal{C}_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}.$$

(4.105) Page 209. Referenced in Exercises: 4.17, 4.18.

$$a_k = \mathbf{w}_k^\top \phi$$

(4.106) Page 209. Referenced in Exercises: 4.17.

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j).$$

(4.108) Page 209. Referenced in Exercises: 4.18.

$$E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}.$$

(4.109) Page 209. Referenced in Exercises: 4.18.

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_k) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\phi_n.$$

(4.110) Page 210. Referenced in Exercises: 4.20.

$$\nabla_{\mathbf{w}_k}\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_k) = \sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\phi_n\phi_n^\top.$$

(4.114) Page 211. Referenced in Exercises: 4.19, 4.21, 4.25, 4.26.

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\, \mathrm{d}\theta.$$

(4.115) Page 211. Referenced in Exercises: 4.21.

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp(\theta^2)\, \mathrm{d}\theta.$$

(4.116) Page 211. Referenced in Exercises: 4.21.

$$\Phi(a) = \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right\}.$$

(4.117) Page 212. Referenced in Exercises: 5.4.

$$\begin{aligned} p(t|\mathbf{x}) &= (1-\epsilon)\sigma(\mathbf{x}) + \epsilon(1-\sigma(x)) \\ &= \epsilon + (1-2\epsilon)\sigma(\mathbf{x}). \end{aligned}$$

(4.135) Page 216. Referenced in Exercises: 4.22.

$$\begin{aligned} Z &= \int f(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\approx f(\mathbf{z}_0)\int \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} \\ &= f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}. \end{aligned}$$

(4.137) Page 216. Referenced in Exercises: 4.22, 4.23.

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|.$$

(4.138) Page 217. Referenced in Exercises: 4.22, 4.23.

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}|\mathcal{D}).$$

(4.139) Page 217. Referenced in Exercises: 4.23.

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}M\ln N.$$

(4.143) Page 218. Referenced in Exercises: 4.24.

$$\mathbf{S}_N^{-1} = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\top.$$

(4.144) Page 218. Referenced in Exercises: 4.24.

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N).$$

(4.151) Page 219. Referenced in Exercises: 4.24.

$$p(\mathcal{C}_1|\mathbf{t}) = \int \sigma(a)p(a)\,\mathrm{d}a = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)\,\mathrm{d}a.$$

(4.152) Page 219. Referenced in Exercises: 4.26.

$$\int \Phi(\lambda a)\mathcal{N}(a|\mu, \sigma^2) = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{1/2}}\right).$$

(4.156) Page 220. Referenced in Exercises: 4.1.

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}_n.$$

(4.163) Page 222. Referenced in Exercises: 4.10.

$$\mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\top$$

(5.7) Page 228. Referenced in Exercises: 5.1.

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right).$$

(5.11) Page 233. Referenced in Exercises: 5.2.

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n||^2.$$

(5.16) Page 234. Referenced in Exercises: 5.2.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1}\boldsymbol{I}).$$

(5.21) Page 235. Referenced in Exercises: 5.4.

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\}.$$

(5.23) Page 235. Referenced in Exercises: 5.5.

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\{t_{nk} \ln y_{nk} + (1 - t_{nk})\ln(1 - y_{nk})\}.$$

(5.24) Page 235. Referenced in Exercises: 5.7.

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}(\mathbf{x}_n, \mathbf{w}).$$

(5.25) Page 236. Referenced in Exercises: 5.7.

$$y_k(\mathbf{x}, \mathbf{w}) = \frac{\exp\{a_k(\mathbf{x}, \mathbf{w})\}}{\sum_j \exp\{a_j(\mathbf{x}, \mathbf{w})\}}.$$

(5.28) Page 237. Referenced in Exercises: 5.13.

$$E(\mathbf{w}) \approx E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{b} + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{H}(\mathbf{w} - \hat{\mathbf{w}}).$$

(5.32) Page 238. Referenced in Exercises: 5.11, 5.12.

$$E(\mathbf{w}) \approx E(\mathbf{w}^\star) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^\star \mathbf{H}(\mathbf{w} - \mathbf{w}^\star).$$

(5.35) Page 238. Referenced in Exercises: 5.11, 5.12.

$$\mathbf{w} - \mathbf{w}^\star = \sum_i \alpha_i \mathbf{u}_i.$$

(5.59) Page 245. Referenced in Exercises: 3.1, 5.8.

$$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}.$$

(5.60) Page 245. Referenced in Exercises: 5.8.

$$h'(a) = 1 - h(a)^2.$$

(5.69) Page 246. Referenced in Exercises: 5.13.

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{E_n(w_{ji} + \epsilon) - E_n(w_{ji} - \epsilon)}{2\epsilon} + O(\epsilon^2).$$

(5.70) Page 247. Referenced in Exercises: 5.15.

$$J_{k,i} = \frac{\partial y_k}{\partial x_i}.$$

(C.6) Page 696. Referenced in Exercises: 3.16.

$$(\mathbf{I} + \mathbf{AB})^{-1}\mathbf{A} = \mathbf{A}(\mathbf{I} + \mathbf{BA})^{-1}.$$

(C.9) Page 696. Referenced in Exercises: 2.17, 3.21.

$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{CAB}) = \mathrm{Tr}(\mathbf{BCA}).$$

(C.13) Page 697. Referenced in Exercises: 1.32.

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

(C.14) Page 697. Referenced in Exercises: 3.16.

$$|\mathbf{I}_N + \mathbf{AB}^\top| = |\mathbf{I}_M + \mathbf{A}^\top\mathbf{B}|.$$

(C.19) Page 697. Referenced in Exercises: 1.32, 3.6, 3.21, 3.22, 4.4, 4.10, 4.13, 4.18, 4.23.

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{a}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^\top\mathbf{x}) = \mathbf{a}.$$

(C.21) Page 698. Referenced in Exercises: 2.34, 2.37, 3.6, 5.3.

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}.$$

(C.24) Page 698. Referenced in Exercises: 2.34, 2.37, 3.6, 4.10, 5.3.

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{AB}) = \mathbf{B}^\top.$$

(C.28) Page 698. Referenced in Exercises: 2.34, 2.37, 3.6, 4.10, 5.3.

$$\frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = (\mathbf{A}^{-1})^\top.$$

(C.30) Page 699. Referenced in Exercises: 3.20.

$$|\mathbf{A} - \lambda_i\mathbf{I}| = 0.$$

(C.47) Page 701. Referenced in Exercises: 3.20, 3.21.

$$|\mathbf{A}| = \prod_{i=1}^{M} \lambda_i.$$

(C.48) Page 701. Referenced in Exercises: 3.21, 3.22.

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{M} \lambda_i.$$

(E.4) Page 708. Referenced in Exercises: 1.34, 2.14, 2.60, 3.5, 4.4, 4.9.

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}).$$