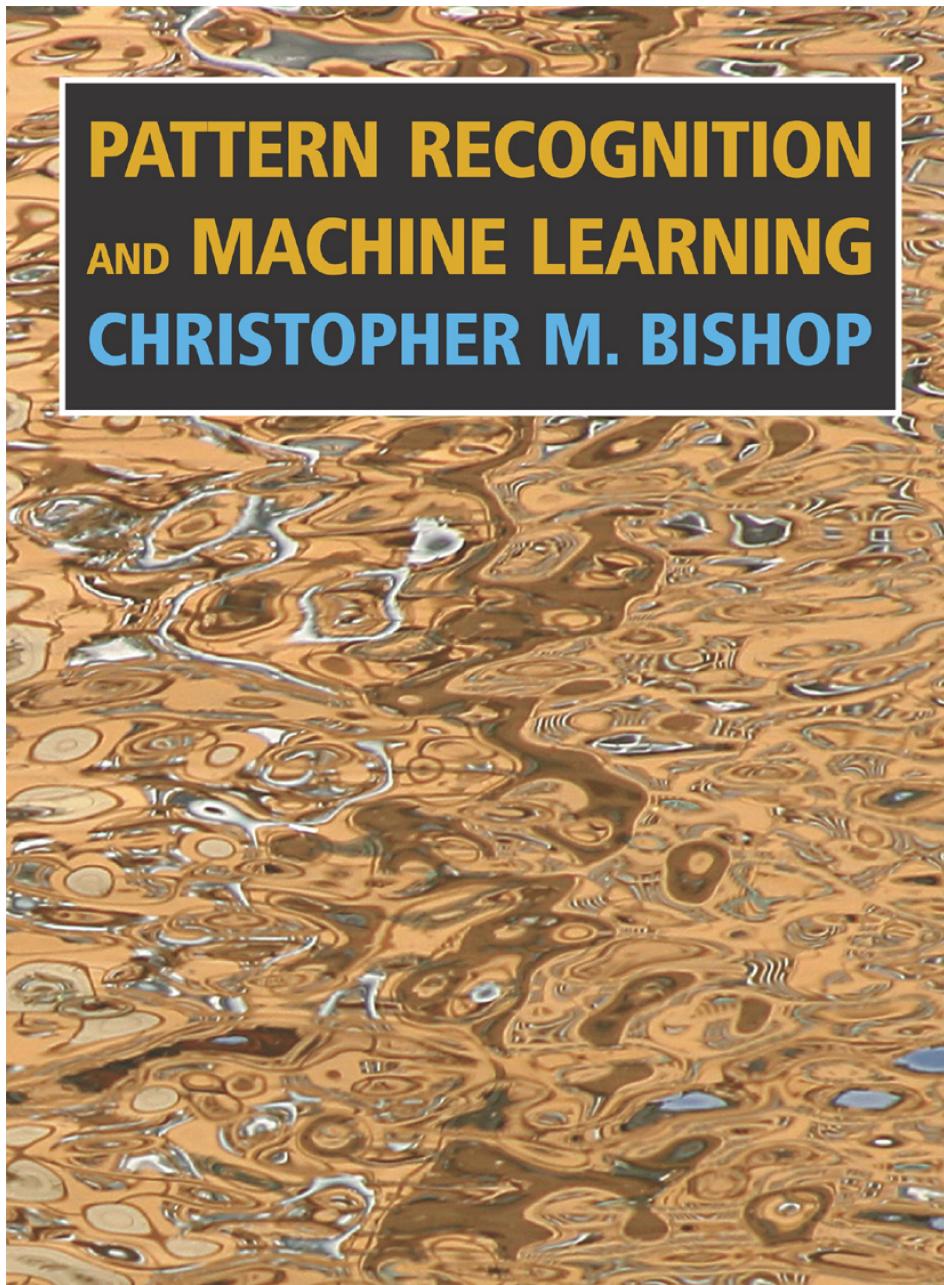


SOLUTIONS FOR



(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*" (8th 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 its 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 $y(x, \mathbf{w})$ 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

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

Solving for $dE(\mathbf{w})/dw_i = 0$, we obtain

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

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

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

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

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

Solving for $d\tilde{E}(\mathbf{w})/dw_i = 0$, we obtain

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

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

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

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

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

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

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

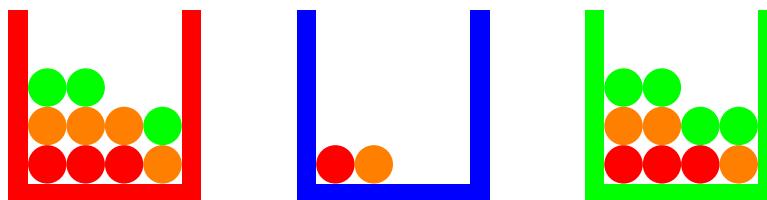


Figure 1.1: Illustration of the boxes described in Exercise 1.3.

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

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

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

$$\frac{dp_Y(y)}{dy} = \begin{cases} \frac{dp_X(g(y))}{dx} \left(\frac{dg(y)}{dy} \right)^2 + p_X(g(y)) \frac{d^2g(y)}{dy^2} & \text{if } \frac{dg(y)}{dy} > 0, \\ \frac{dp_X(g(y))}{dx} \left(\frac{dg(y)}{dy} \right)^2 - p_X(g(y)) \frac{d^2g(y)}{dy^2} & \text{if } \frac{dg(y)}{dy} < 0. \end{cases}$$

The context where $\frac{dg(y)}{dy} = 0$ 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

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

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 $\frac{dp_Y(\tilde{y})}{dy} = 0$, we find that

$$(1.3) \quad \frac{dp_Y(\tilde{y})}{dy} = 0 \iff \frac{d^2g(\tilde{y})}{dy^2} = 0.$$

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

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

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

$$\frac{dp_Y(y)}{dy} = \frac{dp_X(y + a)}{dy}.$$

We now solve $\frac{dp_X(y+a)}{dy} = 0$. By assumption, we know that $\frac{dp_X(\hat{x})}{dx} = 0$. It follows that, for \hat{y} such that $\hat{y} + a = \hat{x}$, we solve $\frac{dp_X(\hat{y})}{dy} = 0$. 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

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

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

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} xy p(x, y) dx dy - \left[\int_{\mathbb{R}} xp(x) dx \right] \left[\int_{\mathbb{R}} yp(y) dy \right] \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} xy p(x)p(y) dx dy - \left[\int_{\mathbb{R}} xp(x) dx \right] \left[\int_{\mathbb{R}} yp(y) dy \right] \quad (\text{Independence}) \\ &= \int_{\mathbb{R}} xp(x) \left[\int_{\mathbb{R}} yp(y) dy \right] dx - \left[\int_{\mathbb{R}} xp(x) dx \right] \left[\int_{\mathbb{R}} yp(y) dy \right] \\ &= \left[\int_{\mathbb{R}} xp(x) dx \right] \left[\int_{\mathbb{R}} yp(y) dy \right] - \left[\int_{\mathbb{R}} xp(x) dx \right] \left[\int_{\mathbb{R}} yp(y) dy \right] \\ \text{Cov}[X, Y] &= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] \quad (\text{Apply (1.34)}).\end{aligned}$$

We thereby conclude

$$(1.4) \quad \text{Cov}[X, Y] = 0.$$

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:

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

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

$$\begin{aligned} I^2 &= \int_{-\pi}^{\pi} \int_0^{\infty} \exp\left\{-\frac{(r \cos \theta)^2 + (r \sin \theta)^2}{2\sigma^2}\right\} \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} dr d\theta \\ &= \int_{-\pi}^{\pi} \int_0^{\infty} |r(\cos \theta)^2 + r(\sin \theta)^2| \exp\left\{-\frac{r^2}{2\sigma^2}\right\} dr d\theta \\ &= \int_{-\pi}^{\pi} \int_0^{\infty} |r| \exp\left\{-\frac{r^2}{2\sigma^2}\right\} dr d\theta \\ &= \int_{-\pi}^{\pi} \int_0^{\infty} \frac{1}{2} \exp\left\{-\frac{u}{2\sigma^2}\right\} du d\theta \quad (\text{Set } u = r^2) \\ &= \frac{1}{2} \int_{-\pi}^{\pi} 2\sigma^2 d\theta \\ I^2 &= 2\pi\sigma^2 \\ I &= \sqrt{2\pi\sigma^2}. \end{aligned}$$

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 $\int_{\mathbb{R}} p(x|\mu, \sigma^2) dx = 1$. It follows that

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

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

$$\begin{aligned}
 \mathbb{E}[X] &= \int_{-\infty}^{\infty} xp(x|\mu, \sigma^2) dx && \text{(Apply (1.34))} \\
 &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx && \text{(Apply (1.46))} \\
 &= \int_{-\infty}^{\infty} \frac{u+\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du && \text{(Set } u = x - \mu\text{)} \\
 &= \int_{-\infty}^{\infty} \frac{u}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du + \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du \\
 &= \int_0^{\infty} \frac{1}{2\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}v\right\} dv - \int_{-\infty}^0 \frac{1}{2\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}v\right\} dv + \\
 &\quad + \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du && \text{(Set } v = u^2\text{)} \\
 &= \frac{2\sigma^2}{2\sqrt{2\pi\sigma^2}} - \frac{2\sigma^2}{2\sqrt{2\pi\sigma^2}} + \frac{\mu}{\sqrt{2\pi\sigma^2}} \cdot \sqrt{2\pi\sigma^2} && \text{(Apply (1.124))} \\
 \mathbb{E}[X] &= \mu.
 \end{aligned}$$

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

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

It follows that

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

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

$$\begin{aligned}
 \text{Var}[X] &= \sigma^2 \\
 \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 &= \sigma^2 \\
 \mathbb{E}[X^2] - \mu^2 &= \sigma^2 \\
 \mathbb{E}[X^2] &= \sigma^2 + \mu^2. \tag{1.5}
 \end{aligned}$$

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 $\mu \in \mathbb{R}^D$ and covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$, the maximum may be determined as follows

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$(1.6) \quad \frac{d^2 \log p(\mathbf{x}|\mu, \sigma^2)}{d(\sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{n=1}^N (x_n - \mu)^2.$$

Evaluating (1.6) at $\sigma^2 = \sum_{n=1}^N (x_n - \mu)^2/N$ gives

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

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

$$(1.7) \quad \begin{aligned} \mathbb{E}[X_n X_m] &= \text{Cov}[X_n, X_m] + \mathbb{E}[X_n]\mathbb{E}[X_m] \\ &= \begin{cases} \mu \cdot \mu + \text{Var}[X_n] & \text{if } n = m, \\ \mu \cdot \mu & \text{otherwise.} \end{cases} \\ \mathbb{E}[X_n X_m] &= \mu^2 + \sigma^2 I_{n,m}, \end{aligned}$$

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

$$\begin{aligned}
 \mathbb{E}[\sigma_{\text{ML}}^2] &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N (X_n - \mu_{\text{ML}})^2\right] \\
 &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}[X_n^2 - 2X_n\mu_{\text{ML}} + \mu_{\text{ML}}^2] \\
 &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}\left[X_n^2 - 2X_n \frac{1}{N} \sum_{m=1}^N X_m + \left(\frac{1}{N} \sum_{m=1}^N X_m\right)^2\right] \quad (\text{Apply (1.55)}) \\
 &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}[X_n^2] - \frac{2}{N^2} \sum_{n=1}^N \mathbb{E}\left[X_n \sum_{m=1}^N X_m\right] + \frac{1}{N^2} \mathbb{E}\left[\left(\sum_{m=1}^N X_m\right)^2\right] \\
 &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}[X_n^2] - \frac{2}{N^2} \sum_{n=1}^N \sum_{m=1}^N \mathbb{E}[X_n X_m] + \frac{1}{N^2} \sum_{n=1}^N \sum_{m=1}^N \mathbb{E}[X_n X_m] \\
 &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}[X_n^2] - \frac{1}{N^2} \sum_{n=1}^N \sum_{m=1}^N \mathbb{E}[X_n X_m] \\
 &= \frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) - \frac{1}{N^2} \sum_{n=1}^N \sum_{m=1}^N (\mu^2 + \sigma^2 I_{n,m}) \quad (\text{Apply (1.7)}) \\
 &= \frac{1}{N} \cdot N \cdot \mu^2 + \frac{1}{N} \cdot N \cdot \sigma^2 - \frac{1}{N^2} \cdot N^2 \cdot \mu^2 - \frac{1}{N^2} \cdot N \cdot \sigma^2 \\
 &= \sigma^2 - \frac{\sigma^2}{N} \\
 \mathbb{E}[\sigma_{\text{ML}}^2] &= \frac{N-1}{N} \sigma^2
 \end{aligned}$$

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 $\tilde{\sigma}_{\text{ML}}^2$ becomes

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

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 $\{w_{i,j}^S\}$ such that

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

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

$$\begin{aligned} w_{i,j}^A &= \frac{w_{i,j} - w_{j,i}}{2} \\ &= -\frac{w_{j,i} - w_{i,j}}{2} \\ w_{i,j}^A &= -w_{j,i}; \end{aligned} \tag{1.8}$$

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

$$\begin{aligned} w_{i,j}^S + w_{i,j}^A &= \frac{w_{i,j} + w_{j,i}}{2} + \frac{w_{i,j} - w_{j,i}}{2} \\ &= \frac{w_{i,j} + w_{j,i} + w_{i,j} - w_{j,i}}{2} \\ &= \frac{2w_{i,j}}{2} \\ w_{i,j}^S + w_{i,j}^A &= w_{i,j}, \end{aligned} \tag{1.9}$$

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

$$\begin{aligned}
 \sum_{i=1}^D \sum_{j=1}^D w_{i,j} x_i x_j &= \sum_{i=1}^D \sum_{j=1}^D (w_{i,j}^S + w_{i,j}^A) x_i x_j && \text{(Apply (1.9))} \\
 &= \sum_{i=1}^D \sum_{j>i}^D (w_{i,j}^S + w_{i,j}^A) x_i x_j + \sum_{i=1}^D (w_{i,i}^S + w_{i,i}^A) (x_i)^2 + \\
 &\quad + \sum_{i=1}^D \sum_{j< i}^D (w_{i,j}^S + w_{i,j}^A) x_i x_j \\
 &= \sum_{i=1}^D \sum_{j>i}^D (w_{i,j}^S + w_{i,j}^A) x_i x_j + \sum_{i=1}^D w_{i,i}^S (x_i)^2 + \\
 &\quad + \sum_{i=1}^D \sum_{j>i}^D (w_{j,i}^S + w_{j,i}^A) x_j x_i && \text{(Apply } w_{i,i}^A = 0\text{)} \\
 &= \sum_{i=1}^D \sum_{j>i}^D (w_{i,j}^S + w_{i,j}^A) x_i x_j + \sum_{i=1}^D w_{i,i}^S (x_i)^2 + \\
 &\quad + \sum_{i=1}^D \sum_{j>i}^D (w_{j,i}^S - w_{i,j}^A) x_j x_i && \text{(Apply (1.8))} \\
 &= \sum_{i=1}^D \sum_{j>i}^D w_{i,j}^S x_i x_j + \sum_{i=1}^D w_{i,i}^S (x_i)^2 + \sum_{i=1}^D \sum_{j>i}^D w_{j,i}^S x_j x_i + \\
 &\quad + \sum_{i=1}^D \sum_{j>i}^D (w_{i,j}^A - w_{i,j}^A) x_i x_j \\
 &= \sum_{i=1}^D \sum_{j>i}^D w_{i,j}^S x_i x_j + \sum_{i=1}^D w_{i,i}^S (x_i)^2 + \sum_{i=1}^D \sum_{j< i}^D w_{i,j}^S x_i x_j \\
 \sum_{i=1}^D \sum_{j=1}^D w_{i,j} x_i x_j &= \sum_{i=1}^D \sum_{j=1}^D w_{i,j}^S x_i x_j.
 \end{aligned}$$

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

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

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, \dots, i_M} is equal to the number of unique unordered sequences of the form $\{j_1, j_2, \dots, j_M\}$, where $j_k \in \{1, \dots, D\} \forall k \in \{1, \dots, M\}$. We may define a class which contains all such sequences, and only said such sequences, by ordering the indexes so that $j_M \leq j_{M-1} \leq \dots \leq j_1$, i.e., the following

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

for some array $\tilde{w}_{i_1, i_2, \dots, i_M}$. The number of independent elements for the M^{th} order term may be computed similarly to (1.10) as follows

$$\begin{aligned} n(D, M) &= \sum_{i_1=1}^D \sum_{i_2=1}^{i_1} \cdots \sum_{i_{M-1}=1}^{i_{M-2}} \sum_{i_M=1}^{i_{M-1}} 1 \\ &= \sum_{i_1=1}^D \left[\sum_{i_2=1}^{i_1} \cdots \sum_{i_{M-1}=1}^{i_{M-2}} \sum_{i_M=1}^{i_{M-1}} 1 \right] \\ (1.11) \quad n(D, M) &= \sum_{i_1=1}^D n(i_1, M-1). \end{aligned}$$

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

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

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 (1.10). Subsequently, assuming it holds for $M - 1$, we seek to prove it holds for M , utilizing the recurrence relation seen in (1.11), 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

$$\begin{aligned}
 \Gamma(x + 1) &= \int_0^\infty u^{x+1-1} e^{-u} du \\
 &= \int_0^\infty u^x e^{-u} du \\
 &= - \left[u^x e^{-u} \right]_0^\infty + \int_0^\infty x u^{x-1} e^{-u} du \quad (\text{Integration by parts}) \\
 &= 0 - \lim_{u \rightarrow \infty} u^x e^{-u} + x \int_0^\infty u^{x-1} e^{-u} du \\
 (1.12) \quad \Gamma(x + 1) &= x\Gamma(x) \quad (\text{Apply (1.141)}).
 \end{aligned}$$

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

$$\begin{aligned}
 \Gamma(1) &= \int_0^\infty u^{1-1} e^{-u} du \\
 &= \int_0^\infty u^0 e^{-u} du \\
 &= \int_0^\infty e^{-u} du \\
 &= - \left[e^{-u} \right]_0^\infty \\
 &= 1 - 0 \\
 (1.13) \quad \Gamma(1) &= 1.
 \end{aligned}$$

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 (1.13). 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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \frac{\pi^{D/2}}{D 2^{D-1} \Gamma(D/2)} &\approx \frac{\pi^{D/2}}{D 2^{D-1} (2\pi)^{1/2} e^{-D/2} (D/2)^{(D+1)/2}} \\ &= \frac{2^{1/2} \pi^{D/2} 2^{D/2} e^{D/2}}{D 2^{D-1} 2^{1/2} \pi^{1/2} D^{D/2} D^{1/2}} \\ &= \frac{2}{\pi^{1/2}} \frac{\pi^{D/2} 2^{D/2} e^{D/2}}{D^{D/2+3/2} 2^D} \\ &= \frac{2}{\pi^{1/2}} \frac{\pi^{D/2} e^{D/2}}{D^{D/2+3/2} 2^{D/2}} \\ &= \frac{2}{\pi^{1/2} D^{3/2}} \frac{\pi^{D/2} e^{D/2}}{D^{D/2} 2^{D/2}} \\ (1.14) \quad \frac{\pi^{D/2}}{D 2^{D-1} \Gamma(D/2)} &= \frac{2}{\pi^{1/2} D^{3/2}} \left(\frac{(\pi e)/2}{D} \right)^{D/2}. \end{aligned}$$

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 $D \rightarrow \infty$ to both sides, we obtain

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

We conclude that the ratio in (1.14) approaches zero as $D \rightarrow \infty$. 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 $D \rightarrow \infty$.

Exercise 1.20

Consider the D dimensional normal distribution with mean $\mu = \mathbf{0}$ and covariance $\Sigma = \sigma^2 I$, 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

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

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

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

Solving $\frac{dp(r)}{dr} = 0$, 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

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

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{dp(\hat{r})}{dr}\varepsilon + \frac{1}{2}\frac{d^2p(\hat{r})}{dr^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. \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 \right. \quad (\text{Note } \frac{dp(\hat{r})}{dr} = 0) \\
 &= 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. \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 \right. \\
 &= p(\hat{r}) + \frac{1}{2} \left[\frac{(D-2)}{\hat{r}^2} \frac{dp(\hat{r})}{dr} - \frac{\hat{r}}{\sigma^2} \frac{dp(\hat{r})}{dr} + \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} \quad (\text{Note } \frac{dp(\hat{r})}{dr} = 0) \\
 &= 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\} \quad (\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

$$\begin{aligned} a &= a^{1/2}a^{1/2} \\ &\leq a^{1/2}b^{1/2} \\ (1.15) \quad &= (ab)^{1/2} \end{aligned}$$

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 (1.15) (an analogous result is valid over \mathcal{R}_2). We can consequently rewrite the probability of committing a mistake as

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

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:

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

That is, in order to minimize (1.16), we 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

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

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

$$(1.17) \quad \hat{j} = \begin{cases} \arg \min_{j \in \{0,1,\dots,K\}} \left[\sum_{k=1}^K L_{k,j} p(\mathcal{C}_k | \mathbf{x}) \mathbf{1}_{\{1,\dots,K\}}(j) + \lambda \mathbf{1}_{\{0\}}(j) \right] \\ \begin{cases} \arg \min_{j \in \{1,\dots,K\}} \sum_{k=1}^K L_{k,j} p(\mathcal{C}_k | \mathbf{x}) & \text{if } \min_{j \in \{1,\dots,K\}} \sum_{k=1}^K L_{k,j} p(\mathcal{C}_k | \mathbf{x}) < \lambda, \\ 0 & \text{if } \min_{j \in \{1,\dots,K\}} \sum_{k=1}^K L_{k,j} p(\mathcal{C}_k | \mathbf{x}) \geq \lambda. \end{cases} \end{cases}$$

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 (1.17), we find that

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

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

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

It is straightforward to conclude that solving $d\mathbb{E}[L(\mathbf{T}, \mathbf{y}(\mathbf{X}))]/d\mathbf{y}(\mathbf{x}) = 0$ 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

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

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}))] &= \iint ||\mathbf{y}(\mathbf{x}) - \mathbf{t}||^2 p(\mathbf{x}, \mathbf{t}) \, d\mathbf{x}d\mathbf{t} \\
 &= \iint ||\mathbf{y}(\mathbf{x}) - \mathbf{t}||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \, d\mathbf{x}d\mathbf{t} \quad (\text{Apply (1.32)}) \\
 &= \iint ||\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}] + \\
 &\quad + \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}] - \mathbf{t}||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \, d\mathbf{x}d\mathbf{t} \\
 &= \iint ||\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \, d\mathbf{x}d\mathbf{t} + \\
 &\quad + \iint ||\mathbf{t} - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \, d\mathbf{x}d\mathbf{t} + \\
 &\quad + 2 \iint (\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]) \times \\
 &\quad \times (\mathbf{t} - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]) p(\mathbf{t}|\mathbf{x})p(\mathbf{x}) \, d\mathbf{x}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}) \, d\mathbf{t} \right] \, d\mathbf{x} + \\
 &\quad + \int p(\mathbf{x}) \left[\int ||\mathbf{t} - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]||^2 p(\mathbf{t}|\mathbf{x}) \, d\mathbf{t} \right] \, d\mathbf{x} + \\
 &\quad + 2 \int \left[(\mathbf{y}(\mathbf{x}) - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]) p(\mathbf{x}) \times \right. \\
 &\quad \left. \times \int (\mathbf{t} - \mathbb{E}[\mathbf{T}|\mathbf{X} = \mathbf{x}]) p(\mathbf{t}|\mathbf{x}) \, d\mathbf{t} \right] \, 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}) \, d\mathbf{x} + \\
 &\quad + \int \text{Var}^*[\mathbf{T}|\mathbf{X} = \mathbf{x}] p(\mathbf{x}) \, d\mathbf{x} \quad (\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

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

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

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

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

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

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

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

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

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

That is $\lim_{q \rightarrow 0} L_q$ 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

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

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

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

As the above relation is valid for all $x > 0$, we may apply $x \rightarrow 0$ on both sides. Assuming the limit $\lim_{x \rightarrow 0} \frac{dh(p^x)}{dp} p^x$ exists, note that the right hand side is constant with respect

to x , yielding the following

$$\begin{aligned}\lim_{x \rightarrow 0} \frac{dh(p^x)}{dp} p^x \log p &= \lim_{x \rightarrow 0} h(p) \\ \log p \lim_{x \rightarrow 0} \frac{dh(p^x)}{dp} p^x &= h(p) \\ h(p) &\propto \log p.\end{aligned}$$

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

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

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

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

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 $I[\mathbf{X}, \mathbf{Y}] = H[\mathbf{Y}] - H[\mathbf{Y}|\mathbf{X}] \geq 0$ - from (1.121) -, we can infer that $H[\mathbf{Y}] \geq H[\mathbf{Y}|\mathbf{X}]$, which implies

$$(1.18) \quad H[\mathbf{X}, \mathbf{Y}] \leq H[\mathbf{Y}] + H[\mathbf{X}].$$

Assuming \mathbf{X} and \mathbf{Y} are independent, $I[\mathbf{X}, \mathbf{Y}] = 0$, and consequently $H[\mathbf{Y}|\mathbf{X}] = H[\mathbf{Y}]$. Applying this result in (1.112) we obtain (1.18). On the other hand, assuming the strict equality holds in (1.18), it holds that

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

It follows that $I[\mathbf{X}, \mathbf{Y}] = 0$, and consequently $KL(p(\mathbf{x}, \mathbf{y}) || p(\mathbf{x})p(\mathbf{y})) = 0$ - 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 $p_{\mathbf{X}}(\mathbf{x})$, it follows from (1.27) that the density associated with \mathbf{Y} is given by

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

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

$$\begin{aligned}
 H[\mathbf{Y}] &= - \int p_{\mathbf{Y}}(\mathbf{y}) \log p_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \\
 &= - \int \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} \log \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} d\mathbf{y} \quad (\text{Apply (1.19)}) \\
 &= - \int \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} \log p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y}) d\mathbf{y} + \int \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}\mathbf{y})}{|\mathbf{A}|} \log |\mathbf{A}| d\mathbf{y} \\
 &= - \int p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} + \int p_{\mathbf{X}}(\mathbf{x}) \log |\mathbf{A}| d\mathbf{x} \\
 H[\mathbf{Y}] &= H[\mathbf{X}] + \log |\mathbf{A}| \quad (\text{Apply (1.104)}).
 \end{aligned}$$

Exercise 1.33

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

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

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 (1.20) 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

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

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

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

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

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

Solving for $\frac{dg(p)}{dp} = 0$, we find that

$$\begin{aligned} -\log p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2 &= 0 \\ \log p(x) &= -1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2 \\ (1.21) \quad p(x) &= \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2\}. \end{aligned}$$

Substituting into the first constraint, we find

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

Substituting into the second constraint, we find

$$\begin{aligned}
 \mu &= \int_{-\infty}^{\infty} xp(x) dx \\
 \mu &= \int_{-\infty}^{\infty} x \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2\} dx \\
 \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\} dx \\
 \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\} dx \\
 \mu &= \exp\left\{\left(\frac{\lambda_2}{4\lambda_3} - \mu\right)\lambda_2\right\} \exp\{\lambda_3\mu^2\} \times \\
 &\quad \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\} dx \\
 \mu &= \exp\left\{\left(\frac{\lambda_2}{4\lambda_3} - \mu\right)\lambda_2\right\} \exp\{\lambda_3\mu^2\} \times \\
 &\quad \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\} dx \\
 \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 \\
 &\quad \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\} dx \\
 \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\} \left[\mu - \frac{\lambda_2}{2\lambda_3}\right] \\
 \mu &= \mu - \frac{\lambda_2}{2\lambda_3} \\
 0 &= \lambda_2.
 \end{aligned}$$

Finally, substituting into the third constraint, we find

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

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 (1.21), 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

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

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 $h \rightarrow 0$ on both sides, we find

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

Exercise 1.37

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

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

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

$$\begin{aligned} f(\lambda_1 a) &\leq \lambda_1 f(a) \\ f(a) &\leq f(a). \end{aligned}$$

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 $M + 1$

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

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 $H[X]$ as

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

We compute $H[Y]$ as

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

We compute $H[X, Y]$ as

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

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

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

and $H[Y|X]$

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

Lastly, the mutual information is

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

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	$1/3$

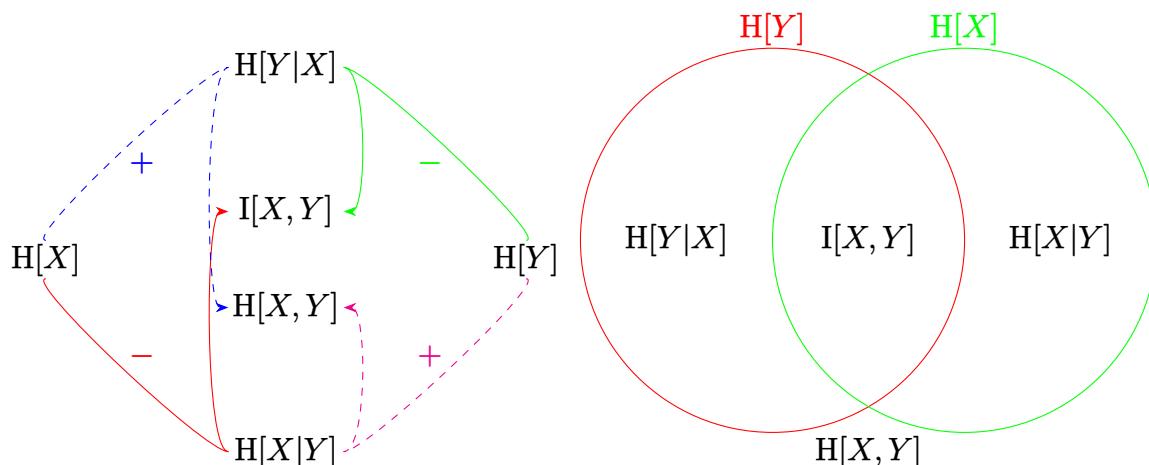


Figure 1.2: Diagrams representing the relationship between different forms of entropy.

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

$$(1.24) \quad \sum_{i=1}^M \frac{1}{M} \log a_i \leq \log \left(\sum_{i=1}^M \frac{1}{M} a_i \right).$$

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 (1.24), 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 $I[\mathbf{X}, \mathbf{Y}]$ is computed as

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

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

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

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

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

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

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

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

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

Exercise 2.3

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

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

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

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

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

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

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

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

Consequently

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

Exercise 2.5

We desire to show that

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

We write

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

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

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

It follows also that

$$\begin{aligned}
 \mathbb{E}[X^2] &= \int_0^1 x^2 p(x|a,b) dx && \text{(Apply (1.34))} \\
 &= \int_0^1 x^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} dx && \text{(Apply (2.13))} \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+2-1} (1-x)^{b-1} dx \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} && \text{(Apply (2.5))} \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{(a+1)a\Gamma(a)\Gamma(b)}{(a+b+1)(a+b)\Gamma(a+b)} && \text{(Apply (1.12))} \\
 \mathbb{E}[X^2] &= \frac{(a+1)a}{(a+b+1)(a+b)},
 \end{aligned}$$

Consequently, the variance of X is computed as

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

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

$$\begin{aligned}\frac{d}{dx} \log p(x|a,b) &= \frac{d}{dx} \left[\log \Gamma(a+b) - \log \Gamma(a) - \log \Gamma(b) + \right. \\ &\quad \left. + (a-1)\log x + (b-1)\log(1-x) \right] \\ &= \frac{a-1}{x} - \frac{b-1}{1-x}.\end{aligned}$$

Solving $d \log p(x|a,b)/dx = 0$, we obtain

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

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 . It 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

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

Moreover, we have that

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

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

$$\int_0^{1-\sum_{k=1}^{M-2} x_k} p_M(x_1, \dots, x_{M-1}) dx_{M-1}$$

$$= \int_0^{1-\sum_{k=1}^{M-2} x_k} C_M x_{M-1}^{\alpha_{M-1}-1} \prod_{k=1}^{M-2} x_k^{\alpha_k-1} \left(1 - \sum_{k=1}^{M-2} x_k - x_{M-1}\right)^{\alpha_M-1} dx_{M-1}.$$

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 [\(2.5\)](#):

$$\int_0^1 C_M t^{\alpha_{M-1}-1} \left(1 - \sum_{k=1}^{M-2} x_k\right)^{\alpha_{M-1}} (1-t)^{\alpha_M-1} \left(1 - \sum_{k=1}^{M-2} x_k\right)^{\alpha_M-1} \prod_{k=1}^{M-2} x_k^{\alpha_k-1} dt$$

$$= \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} C_m \left(1 - \sum_{k=1}^{M-2} x_k\right)^{\alpha_M + \alpha_{M-1}-1} \prod_{k=1}^{M-2} x_k^{\alpha_k-1}$$

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

$$\frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} C_m = C_{M-1}$$

$$= \frac{\Gamma(\alpha_1 + \dots + \alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \dots \Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1} + \alpha_M)}$$

$$C_M = \frac{\Gamma(\alpha_1 + \dots + \alpha_{M-1} + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \dots \Gamma(\alpha_{M-2})\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}.$$

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

$$\begin{aligned}
 \mathbb{E}[X_j] &= \int_{\mathbb{R}^M} x_j p(\mathbf{x}|\boldsymbol{\alpha}) dx_1 \dots dx_M && \text{(Apply (1.34))} \\
 &= \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_M)} x_j \prod_{k=1}^M x_k^{\alpha_k-1} dx_1 \dots dx_M && \text{(Apply (2.38))} \\
 &= \frac{\alpha_j}{\alpha_0} \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0+1)}{\Gamma(\alpha_j+1) \prod_{\substack{k=1 \\ k \neq j}}^M \Gamma(\alpha_k)} x_j^{\alpha_j+1-1} \prod_{\substack{k=1 \\ k \neq j}}^M x_k^{\alpha_k-1} dx_1 \dots dx_M && \text{(Apply (1.12))} \\
 (2.7) \quad \mathbb{E}[X_j] &= \frac{\alpha_j}{\alpha_0} && \text{(Apply (1.26)).}
 \end{aligned}$$

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

$$\begin{aligned}
 \mathbb{E}[X_j X_l] &= \int_{\mathbb{R}^M} x_j x_l p(\mathbf{x}|\boldsymbol{\alpha}) dx_1 \dots dx_M && \text{(Apply (1.34))} \\
 &= \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_M)} x_j x_l \prod_{k=1}^M x_k^{\alpha_k-1} dx_1 \dots dx_M && \text{(Apply (2.38))} \\
 &= \frac{\alpha_j \alpha_l}{\alpha_0 (\alpha_0+1)} \int_{\mathbb{R}^M} \left[\frac{\Gamma(\alpha_0+2)}{\Gamma(\alpha_j+1) \Gamma(\alpha_l+1) \prod_{\substack{k=1 \\ k \neq j \\ k \neq l}}^M \Gamma(\alpha_k)} \right. \\
 &\quad \times \left. x_j^{\alpha_j+1-1} x_l^{\alpha_l+1-1} \prod_{\substack{k=1 \\ k \neq j \\ k \neq l}}^M x_k^{\alpha_k-1} \right] dx_1 \dots dx_M && \text{(Apply (1.12))} \\
 (2.8) \quad \mathbb{E}[X_j X_l] &= \frac{\alpha_j \alpha_l}{\alpha_0 (\alpha_0+1)} && \text{(Apply (1.26)).}
 \end{aligned}$$

The expected value for X_j^2 is given by

$$\begin{aligned}
 \mathbb{E}[X_j^2] &= \int_{\mathbb{R}^M} x_j^2 p(\mathbf{x}|\boldsymbol{\alpha}) dx_1 \dots dx_M && \text{(Apply (1.34))} \\
 &= \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_M)} x_j^2 \prod_{k=1}^M x_k^{\alpha_k-1} dx_1 \dots dx_M && \text{(Apply (2.38))} \\
 &= \frac{\alpha_j (\alpha_j+1)}{\alpha_0 (\alpha_0+1)} \int_{\mathbb{R}^M} \left[\frac{\Gamma(\alpha_0+2)}{\Gamma(\alpha_j+2) \prod_{\substack{k=1 \\ k \neq j}}^M \Gamma(\alpha_k)} \right. \\
 &\quad \times \left. x_j^{\alpha_j+2-1} \prod_{\substack{k=1 \\ k \neq j}}^M x_k^{\alpha_k-1} \right] dx_1 \dots dx_M && \text{(Apply (1.12))} \\
 (2.9) \quad \mathbb{E}[X_j^2] &= \frac{\alpha_j (\alpha_j+1)}{\alpha_0 (\alpha_0+1)} && \text{(Apply (1.26)).}
 \end{aligned}$$

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

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

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

$$\begin{aligned}
 0 &= \int_{\mathbb{R}^M} \frac{d}{d\alpha_j} \left[\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_M)} \prod_{k=1}^M x_k^{\alpha_k-1} \right] dx_1 \dots dx_M \\
 0 &= \int_{\mathbb{R}^M} \left[\prod_{\substack{k=1 \\ k \neq j}}^M \frac{1}{\Gamma(\alpha_k)} \right] \left[\frac{\Gamma'(\alpha_0)\Gamma(\alpha_j) - \Gamma(\alpha_0)\Gamma'(\alpha_j)}{[\Gamma(\alpha_j)]^2} x_j^{\alpha_j-1} + \right. \\
 &\quad \left. + \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)} x_j^{\alpha_j-1} \log x_j \right] dx_1 \dots dx_M \\
 0 &= \frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)} \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0) \prod_{j=1}^M x_j^{\alpha_j-1}}{\prod_{k=1}^M \Gamma(\alpha_k)} dx_1 \dots dx_M + \\
 &\quad - \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0) \prod_{j=1}^M x_j^{\alpha_j-1}}{\prod_{k=1}^M \Gamma(\alpha_k)} dx_1 \dots dx_M + \\
 &\quad + \int_{\mathbb{R}^M} \frac{\Gamma(\alpha_0) \prod_{j=1}^M x_j^{\alpha_j-1}}{\prod_{k=1}^M \Gamma(\alpha_k)} \log x_j dx_1 \dots dx_M \\
 \mathbb{E}[\log X_j] &= \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} - \frac{\Gamma'(\alpha_0)}{\Gamma(\alpha_0)} \tag{Apply (1.26) and (1.34)} \\
 \mathbb{E}[\log X_j] &= \psi(\alpha_j) - \psi(\alpha_0) \tag{Apply (2.277)}.
 \end{aligned}$$

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

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

The expected value of U is computed as

$$\begin{aligned}\mathbb{E}[U] &= \int_a^b u p(u|a, b) \, du \quad (\text{Apply (1.30)}) \\ &= \int_a^b \frac{u}{b-a} \, du \quad (\text{Apply (2.278)}) \\ &= \left. \frac{v^2}{2(b-a)} \right|_{v=a}^{v=b} \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} \\ \mathbb{E}[U] &= \frac{b+a}{2}.\end{aligned}$$

Lastly, the variance of U is

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

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

Exercise 2.14

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

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

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

$$\begin{aligned} g(p) = & - \int_{\mathbb{R}^D} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} + \lambda_1 \left(\int_{\mathbb{R}^D} p(\mathbf{x}) d\mathbf{x} - 1 \right) \\ & + \lambda_2^\top \left(\int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\mu} \right) + \text{tr} \left(\Lambda_3 \left[\int_{\mathbb{R}^D} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x}) d\mathbf{x} - \Sigma \right] \right). \end{aligned}$$

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

$$\frac{dg(p)}{dp} = -\log p(\mathbf{x}) - 1 + \lambda_1 + \lambda_2^\top \mathbf{x} + \text{tr}[\Lambda_3(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top].$$

Solving for $\frac{dg(p)}{dp} = 0$, we find that

$$(2.10) \quad p(\mathbf{x}) = \exp\{-1 + \lambda_1 + \lambda_2^\top \mathbf{x} + \text{tr}[\Lambda_3(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]\}.$$

Substituting into the first constraint, we obtain

$$\begin{aligned} 1 &= \int_{\mathbb{R}^D} p(\mathbf{x}) d\mathbf{x} \\ 1 &= \int_{\mathbb{R}^D} \exp\{-1 + \lambda_1 + \lambda_2^\top \mathbf{x} + \text{tr}[\Lambda_3(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]\} d\mathbf{x} \\ \exp\{1\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\{\text{tr}(\mathbf{x}\lambda_2^\top) + \text{tr}[\Lambda_3\mathbf{x}\mathbf{x}^\top - 2\Lambda_3\mathbf{x}\boldsymbol{\mu}^\top + \Lambda_3\boldsymbol{\mu}\boldsymbol{\mu}^\top]\} d\mathbf{x} \\ \exp\{1 - \text{tr}(\Lambda_3\boldsymbol{\mu}\boldsymbol{\mu}^\top)\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\{\text{tr}(\Lambda_3\mathbf{x}\lambda_2^\top\Lambda_3^{-1}) + \text{tr}[\Lambda_3\mathbf{x}\mathbf{x}^\top - 2\Lambda_3\mathbf{x}\boldsymbol{\mu}^\top]\} d\mathbf{x} \\ \exp\{1 - \text{tr}(\Lambda_3\boldsymbol{\mu}\boldsymbol{\mu}^\top)\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp\{\text{tr}[\Lambda_3(\mathbf{x}\mathbf{x}^\top - 2\mathbf{x}(\boldsymbol{\mu}^\top - \lambda_2^\top\Lambda_3^{-1}/2))]\} d\mathbf{x} \\ \exp\{1 - \text{tr}(\Lambda_3\boldsymbol{\mu}\boldsymbol{\mu}^\top)\} &= \exp\{\lambda_1\} \int_{\mathbb{R}^D} \exp \left\{ -\frac{1}{2} \text{tr}[(-2\Lambda_3) \times \right. \\ &\quad \times (\mathbf{x} - \boldsymbol{\mu} + \Lambda_3^{-1}\lambda_2/2)(\mathbf{x} - \boldsymbol{\mu} + \lambda_2\Lambda_3^{-1}/2)^\top] + \\ &\quad \left. - \text{tr}[\Lambda_3(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)^\top] \right\} d\mathbf{x} \\ \exp\{\lambda_1\} 2^{-D/2} |\Lambda_3|^{-1/2} (2\pi)^{D/2} &= \exp\{1 - \text{tr}(\Lambda_3\boldsymbol{\mu}\boldsymbol{\mu}^\top) + \text{tr}[\Lambda_3(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)^\top]\} \\ \exp\{\lambda_1\} &= |\Lambda_3|^{1/2} \pi^{-D/2} \exp\{1 - \text{tr}(\Lambda_3\boldsymbol{\mu}\boldsymbol{\mu}^\top) + \\ &\quad + \text{tr}[\Lambda_3(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)^\top]\} \\ \lambda_1 &= \frac{1}{2} \log(|-\Lambda_3|) - \frac{D}{2} \log \pi + 1 - \text{tr}[\Lambda_3(\boldsymbol{\mu}\boldsymbol{\mu}^\top)] + \\ &\quad + \text{tr}[\Lambda_3(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)(\boldsymbol{\mu} - \Lambda_3^{-1}\lambda_2/2)^\top] \\ \lambda_1 &= \frac{1}{2} \log(|-\Lambda_3|) - \frac{D}{2} \log \pi + 1 - \boldsymbol{\mu}^\top \lambda_2 + \frac{\lambda_2^\top \Lambda_3^{-1} \lambda_2}{4}. \end{aligned}$$

Substituting into the second constraint, we obtain

$$\begin{aligned}
 \mu &= \int_{\mathbb{R}^D} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\
 \mu &= \int_{\mathbb{R}^D} \mathbf{x} \exp\{-1 + \lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{x} + \text{tr}[\Lambda_3(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top]\} d\mathbf{x} \\
 \mu &= \int_{\mathbb{R}^D} \mathbf{x} \exp \left\{ -1 + \frac{1}{2} \log(|-\Lambda_3|) - \frac{D}{2}\pi + 1 - \mu^\top \boldsymbol{\lambda}_2 + \right. \\
 &\quad \left. + \frac{\boldsymbol{\lambda}_2^\top \Lambda_3^{-1} \boldsymbol{\lambda}_2}{4} + \boldsymbol{\lambda}_2^\top \mathbf{x} + \text{tr}[\Lambda_3(\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top] \right\} d\mathbf{x} \\
 \mu &= \exp \left\{ \frac{\boldsymbol{\lambda}_2^\top \Lambda_3^{-1} \boldsymbol{\lambda}_2}{4} - \mu^\top \boldsymbol{\lambda}_2 \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\Lambda_3|^{1/2}}{\pi^{D/2}} \times \\
 &\quad \times \exp \left\{ \text{tr}(\mathbf{x} \boldsymbol{\lambda}_2^\top) + \text{tr}[\Lambda_3 \mathbf{x} \mathbf{x}^\top - 2\Lambda_3 \mu \mathbf{x}^\top + \Lambda_3 \mu \mu^\top] \right\} d\mathbf{x} \\
 \mu &= \exp \left\{ \frac{\boldsymbol{\lambda}_2^\top \Lambda_3^{-1} \boldsymbol{\lambda}_2}{4} - \mu^\top \boldsymbol{\lambda}_2 \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\Lambda_3|^{1/2}}{\pi^{D/2}} \times \\
 &\quad \times \exp \left\{ \text{tr} \left[\Lambda_3 (\mathbf{x} \mathbf{x}^\top - 2\mu \mathbf{x}^\top + \Lambda_3^{-1} \boldsymbol{\lambda}_2 \mathbf{x}^\top + \mu \mu^\top) \right] \right\} d\mathbf{x} \\
 \mu &= \exp \left\{ \frac{\boldsymbol{\lambda}_2^\top \Lambda_3^{-1} \boldsymbol{\lambda}_2}{4} + \text{tr}[\Lambda_3 \mu \mu^\top] - \mu^\top \boldsymbol{\lambda}_2 \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\Lambda_3|^{1/2}}{\pi^{D/2}} \times \\
 &\quad \times \exp \left\{ \text{tr} \left[\Lambda_3 (\mathbf{x} \mathbf{x}^\top - (\mu - \Lambda_3^{-1} \boldsymbol{\lambda}_2/2) \mathbf{x}^\top) \right] \right\} d\mathbf{x} \\
 \mu &= \exp \left\{ \frac{\boldsymbol{\lambda}_2^\top \Lambda_3^{-1} \boldsymbol{\lambda}_2}{4} + \text{tr}[\Lambda_3 \mu \mu^\top] - \mu^\top \boldsymbol{\lambda}_2 + \right. \\
 &\quad \left. - \text{tr} \left[\Lambda_3 (\mu - \Lambda_3^{-1} \boldsymbol{\lambda}_2/2) (\mu - \Lambda_3^{-1} \boldsymbol{\lambda}_2/2)^\top \right] \right\} \int_{\mathbb{R}^D} \mathbf{x} \frac{|-\Lambda_3|^{1/2}}{\pi^{D/2}} \times \\
 &\quad \times \exp \left\{ -\frac{1}{2} \text{tr} \left[(-2\Lambda_3)(\mathbf{x} - \mu + \Lambda_3^{-1} \boldsymbol{\lambda}_2/2)(\mathbf{x} - \mu + \Lambda_3^{-1} \boldsymbol{\lambda}_2/2)^\top \right] \right\} d\mathbf{x} \\
 \mu &= \mu - \frac{\Lambda_3^{-1} \boldsymbol{\lambda}_2}{2} \\
 \boldsymbol{\lambda}_2 &= \mathbf{0}.
 \end{aligned}$$

Finally, substituting into the third constraint, we obtain

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

Substituting the Lagrangian values into (2.10), we obtain

$$\begin{aligned}
 p(\mathbf{x}) &= \exp\{-1 + \lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{x} + \text{tr}[\Lambda_3(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]\} \\
 &= \exp \left\{ -1 + \frac{1}{2} \log(|\Sigma^{-1}/2|) - \frac{D}{2} \log \pi + 1 - \text{tr}[\Sigma^{-1}/2(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \right\} \\
 p(\mathbf{x}) &= \frac{\exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}}{(2\pi)^{D/2} |\Sigma|^{1/2}}.
 \end{aligned}$$

Thereby concluding that a D dimensional multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance $\Sigma \in \mathbb{R}^{D \times D}$ 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

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

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) dx_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\} dx_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\} dx_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_1x_2(x - \mu_1) \right. \\
&\quad \left. - \frac{\tau_2}{2}x_2^2 + \tau_2x_2\mu_2\right\} dx_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. \\
&\quad \left. + x_2[\tau_1x - \tau_1\mu_1 + \tau_2\mu_2]\right\} dx_2 \\
&= \exp\left\{-\frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2\right\} \\
&\quad \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_1x - \tau_1\mu_1 + \tau_2\mu_2}{\tau_1 + \tau_2}\right]\right\} dx_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\} \\
&\quad \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_1x - \tau_1\mu_1 + \tau_2\mu_2}{\tau_1 + \tau_2}\right]^2\right\} dx_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. \\
&\quad \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:

$$\begin{aligned}
 &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\tau_1 \tau_2}{\tau_1 + \tau_2}} \exp \left\{ -\frac{1}{2} \frac{1}{\tau_1 + \tau_2} \left[\tau_1 \tau_2 (x - \mu_1)^2 \right. \right. \\
 &\quad \left. \left. + \tau_1 \tau_2 \mu_2^2 - 2\tau_1(x - \mu_1)\tau_2\mu_2 \right] \right\} \\
 p(x) &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\tau_1 \tau_2}{\tau_1 + \tau_2}} \exp \left\{ -\frac{1}{2} \frac{\tau_1 \tau_2}{\tau_1 + \tau_2} (x - \mu_1 - \mu_2)^2 \right\}.
 \end{aligned}$$

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 (1.23), that the differential entropy associated with X is of the form

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

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 $\Lambda \in \mathbb{R}^{D \times D}$. Consider that the precision matrix $\Lambda \in \mathbb{R}^{D \times D}$ may be rewritten as the sum of anti-symmetric and a symmetric matrix, as in $\Lambda = \Lambda^A + \Lambda^S$. It follows that the density may be rewritten as

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

Write the symmetric and anti-symmetric matrices as follows

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

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

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

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

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

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

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

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

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

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 $\Sigma \in \mathbb{R}^{D \times D}$ 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

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

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

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

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

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

Exercise 2.20

Let $\Sigma \in \mathbb{R}^{D \times D}$ 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

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

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

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

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

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

Hence, we conclude that all eigenvalues being positive is a sufficient condition for the matrix Σ 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 $\mathbf{a}^\top \Sigma \mathbf{a} > 0$ 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

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

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

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

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

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

We thereby conclude that, if Σ 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

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

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

$$dz = S_D r^{D-1} dr.$$

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

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

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

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

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 $\Sigma \in \mathbb{R}^{D \times D}$, which may be decomposed as in (2.288). We likewise decompose $\mathbf{X} = (\mathbf{X}_a, \mathbf{X}_b, \mathbf{X}_c)^\top$. We seek to determine the distribution of $\mathbf{X}_a | \mathbf{X}_b = \mathbf{x}_b$. In order to do so, first we determine the distribution of $(\mathbf{X}_a, \mathbf{X}_c) | \mathbf{X}_b = \mathbf{x}_b$. Utilizing previously established results, we find that it is a multivariate normal distribution, whose mean is

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

Similarly, the associated covariance matrix is

$$\begin{aligned}\Sigma_{a,c|b} &= \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,c} \\ \Sigma_{c,a} & \Sigma_{c,c} \end{pmatrix} - \begin{pmatrix} \Sigma_{a,b} \\ \Sigma_{c,b} \end{pmatrix} \Sigma_{b,b}^{-1} \begin{pmatrix} \Sigma_{b,a} & \Sigma_{b,c} \end{pmatrix} \quad (\text{Apply (2.82)}) \\ &= \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,c} \\ \Sigma_{c,a} & \Sigma_{c,c} \end{pmatrix} - \begin{pmatrix} \Sigma_{a,b} \Sigma_{b,b}^{-1} \\ \Sigma_{c,b} \Sigma_{b,b}^{-1} \end{pmatrix} \begin{pmatrix} \Sigma_{b,a} & \Sigma_{b,c} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,c} \\ \Sigma_{c,a} & \Sigma_{c,c} \end{pmatrix} - \begin{pmatrix} \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{b,a} & \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{b,c} \\ \Sigma_{c,b} \Sigma_{b,b}^{-1} \Sigma_{b,a} & \Sigma_{c,b} \Sigma_{b,b}^{-1} \Sigma_{b,c} \end{pmatrix} \\ \Sigma_{a,c|b} &= \begin{pmatrix} \Sigma_{a,a} - \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{b,a} & \Sigma_{a,c} - \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{b,c} \\ \Sigma_{c,a} - \Sigma_{c,b} \Sigma_{b,b}^{-1} \Sigma_{b,a} & \Sigma_{c,c} - \Sigma_{c,b} \Sigma_{b,b}^{-1} \Sigma_{b,c} \end{pmatrix}.\end{aligned}$$

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

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

And covariance matrix

$$\Sigma_{a|b} = \Sigma_{a,a} - \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{b,a} \quad (\text{Apply (2.93)}).$$

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

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

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

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

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

$$\begin{aligned}\text{Var}[\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}) d\mathbf{x}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}) d\mathbf{x}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}) d\mathbf{x}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}) d\mathbf{x}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}) d\mathbf{x}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}) d\mathbf{x}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}) d\mathbf{x} \right] 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}) d\mathbf{z} \right] d\mathbf{x} \\ &\quad + \left[\int_{\mathbb{R}^D} (\mathbf{x} - \mathbb{E}[\mathbf{X}]) p(\mathbf{x}) d\mathbf{x} \right] \left[\int_{\mathbb{R}^D} (\mathbf{z} - \mathbb{E}[\mathbf{Z}])^\top p(\mathbf{z}) d\mathbf{z} \right] \\ &\quad + \left[\int_{\mathbb{R}^D} (\mathbf{z} - \mathbb{E}[\mathbf{Z}]) p(\mathbf{z}) d\mathbf{z} \right] \left[\int_{\mathbb{R}^D} (\mathbf{x} - \mathbb{E}[\mathbf{X}])^\top p(\mathbf{x}) d\mathbf{x} \right]\end{aligned}$$

$$\text{Var}[\mathbf{X} + \mathbf{Z}] = \text{Var}[\mathbf{X}] + \text{Var}[\mathbf{Z}] \quad \text{(Apply (1.30) and (1.34)).}$$

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 is 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

$$\text{Var}[\mathbf{X}] = \boldsymbol{\Lambda}^{-1}$$

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

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

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

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

The upper-right of the inverse is

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

The lower-left of the inverse is

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

The lower-right of the inverse is

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

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

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

Thereby reaching the desired conclusion.

Exercise 2.31

Let \mathbf{X} and \mathbf{Z} be D -dimensional multivariate normal random variables with means $\mu_{\mathbf{X}} \in \mathbb{R}^D$ and $\mu_{\mathbf{Z}} \in \mathbb{R}^D$ respectively, and covariances matrices $\Sigma_{\mathbf{X}} \in \mathbb{R}^{D \times D}$ and $\Sigma_{\mathbf{Z}} \in \mathbb{R}^{D \times D}$, 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:

$$\begin{aligned} p(\mathbf{x}|\mu_{\mathbf{X}}, \Sigma_{\mathbf{x}}) &= \text{MULTIVARIATE NORMAL}(\mu_{\mathbf{X}}, \Sigma_{\mathbf{x}}) \\ p(\mathbf{y}|\mathbf{x}) &= \text{MULTIVARIATE NORMAL}(\mu_{\mathbf{Z}} + \mathbf{x}, \Sigma_{\mathbf{z}}). \end{aligned}$$

I.e., $\mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}] = \mathbf{Ax} + \mathbf{b}$, where \mathbf{A} is the D -dimensional identity matrix, and $\mathbf{b} = \mu_{\mathbf{Z}}$. 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

$$\mathbb{E}[\mathbf{Y}] = \mu_{\mathbf{X}} + \mu_{\mathbf{Z}}.$$

And covariance matrix

$$\text{Var}[\mathbf{Y}] = \Sigma_{\mathbf{X}} + \Sigma_{\mathbf{Z}}.$$

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

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

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

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

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

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

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

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

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}[\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}[\Lambda + \mathbf{A}^\top \mathbf{L}\mathbf{A}]^{-1} \Lambda \mu \\
 &= -\frac{1}{2}(\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y} - \mathbf{b}) + \\
 &\quad - (\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1}(\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)\mathbf{L}\mathbf{A}(\Lambda + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1} \Lambda \mu \\
 &= -\frac{1}{2}(\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y} - \mathbf{b}) + \\
 &\quad - (\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1} \times \\
 &\quad \times [\mathbf{A}(\Lambda + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1} + \mathbf{A}\Lambda^{-1}\Lambda + \mathbf{A}^\top \mathbf{L}\mathbf{A}^{-1} + \\
 &\quad - \mathbf{A}\Lambda^{-1}\Lambda(\Lambda + \mathbf{A}^\top \mathbf{L}\mathbf{A})^{-1}]\Lambda \mu \\
 &= -\frac{1}{2}(\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y} - \mathbf{b}) + \\
 &\quad - (\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1} \mathbf{A}\Lambda^{-1}\Lambda \mu \\
 &= -\frac{1}{2}(\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y} - \mathbf{b}) + \\
 &\quad - (\mathbf{y} - \mathbf{b})^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1} \mathbf{A} \mu \\
 \log p(\mathbf{y}) &\propto -\frac{1}{2}(\mathbf{y} - \mathbf{b} - \mathbf{A} \mu)^\top (\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top)^{-1}(\mathbf{y} - \mathbf{b} - \mathbf{A} \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 $\mathbf{A}\mu + \mathbf{b} \in \mathbb{R}^M$ and covariance matrix $\mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top \in \mathbb{R}^{M \times M}$.

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

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

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

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

where $\Sigma = [\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A}]^{-1}$. 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 $\Sigma[\Lambda \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L}\{\mathbf{y} - \mathbf{b}\}] \in \mathbb{R}^D$ and variance $\Sigma \in \mathbb{R}^{D \times D}$.

Exercise 2.34

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

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

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

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

Note that this solution is dependent on the parameter μ . From (2.121), we have that the maximum likelihood estimator of μ 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 Σ . Thereafter, we conclude that the maximum likelihood estimator for the covariance matrix is

$$(2.11) \quad \Sigma_{\text{ML}} = \frac{\sum_{i=1}^N (\mathbf{x}_i - \mu_{\text{ML}})(\mathbf{x}_i - \mu_{\text{ML}})^\top}{N},$$

which is the sample covariance.

Exercise 2.35

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

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

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

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

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 Σ . It follows that

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

We hence conclude that the maximum likelihood estimator for Σ 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

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

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

$$\begin{aligned} \sigma_{\text{ML}}^{2,(N)} &= \sum_{i=1}^N \frac{(x_i - \mu)^2}{N} \\ &= \frac{(x_N - \mu)^2}{N} + \sum_{i=1}^{N-1} \frac{(x_i - \mu)^2}{N} \\ &= \frac{(x_N - \mu)^2}{N} + \frac{N-1}{N} \sum_{i=1}^{N-1} \frac{(x_i - \mu)^2}{N-1} \\ &= \frac{(x_N - \mu)^2}{N} + \frac{N-1}{N} \sigma_{\text{ML}}^{2,(N-1)} \quad (\text{Apply (2.13)}) \\ (2.14) \quad \sigma_{\text{ML}}^{2,(N)} &= \sigma_{\text{ML}}^{2,(N-1)} + \frac{1}{N} [(x_N - \mu)^2 - \sigma_{\text{ML}}^{2,(N-1)}]. \end{aligned}$$

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

$$\begin{aligned} \sigma_{\text{ML}}^{2,(N)} &= \sigma_{\text{ML}}^{2,(N-1)} - a_{N-1} \frac{d[-\log p(x_N | \sigma_{\text{ML}}^{2,(N-1)})]}{d\sigma^2} \\ (2.15) \quad &= \sigma_{\text{ML}}^{2,(N-1)} - a_{N-1} \frac{d}{d\sigma^2} \left[\frac{1}{\sigma^2} (x_N - \mu)^2 + \right. \\ &\quad \left. + \frac{1}{2} \log(2\pi) + \frac{1}{2} \sigma^2 \right] \Big|_{\sigma^2=\sigma_{\text{ML}}^{2,(N-1)}} \quad (\text{Apply (1.46)}) \\ &= \sigma_{\text{ML}}^{2,(N-1)} - a_{N-1} \left[-\frac{1}{2\sigma_{\text{ML}}^{4,(N-1)}} (x_N - \mu)^2 + \frac{1}{2\sigma_{\text{ML}}^{2,(N-1)}} \right] \\ (2.16) \quad \sigma_{\text{ML}}^{2,(N)} &= \sigma_{\text{ML}}^{2,(N-1)} + \frac{a_{N-1}}{2\sigma_{\text{ML}}^{4,(N-1)}} \left[(x_N - \mu)^2 - \sigma_{\text{ML}}^{2,(N-1)} \right]. \end{aligned}$$

We conclude that, by choosing $a_N = 2(N+1)^{-1}\sigma_{\text{ML}}^{4,(N)}$, 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 $\mu \in \mathbb{R}^D$ 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 μ , is

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

We decompose (2.17) as follows

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

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

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

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

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 \mathbf{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 $\Theta|\mathbf{X} = \mathbf{x}$ as follows

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

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

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

and variance

$$\begin{aligned}
 \sigma_N^2 &= \text{Var}[\Theta|\mathbf{X} = \mathbf{x}] \\
 &= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2} \\
 &= \left[\frac{N\sigma_0^2 + \sigma^2}{\sigma^2\sigma_0^2}\right]^{-1} \\
 (2.20) \quad \sigma_N^2 &= \left[\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right]^{-1}.
 \end{aligned}$$

Exercise 2.39

From the results seen in [Exercise 2.38](#), we can dissect the form of μ_N as

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

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 $\Theta | \mathbf{X}_{(-N)} = \mathbf{x}_{(-N)}$ 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

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

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

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

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 \mathbf{X}) which, conditioned by $\Theta = \theta$, are independent D -dimensional multivariate normals with mean $\theta \in \mathbb{R}^D$ and known covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$, whilst Θ is a D -dimensional multivariate normal with mean $\mu_0 \in \mathbb{R}^D$ and covariance matrix $\Sigma_0 \in \mathbb{R}^{D \times D}$. We aim herein to determine the distribution of $\Theta|\mathbf{X} = \mathbf{x}$ as follows

$$\begin{aligned}
 p(\theta|\mathbf{x}) &\propto p(\mathbf{x}|\theta)p(\theta) \\
 &\propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^N (\mathbf{x}_i - \theta)^\top \Sigma^{-1} (\mathbf{x}_i - \theta) \right\} \times \\
 &\quad \times \exp \left\{ -\frac{1}{2} (\theta - \mu_0)^\top \Sigma_0^{-1} (\theta - \mu_0) \right\} \tag{Apply (2.43)} \\
 &\propto \exp \left\{ -\frac{N}{2} \theta^\top \Sigma^{-1} \theta + N \theta^\top \Sigma^{-1} \mu_{ML} - \frac{1}{2} \theta^\top \Sigma_0^{-1} \theta + \theta^\top \Sigma_0^{-1} \mu_0 \right\} \\
 &= \exp \left\{ -\frac{1}{2} \theta^\top (N\Sigma^{-1} + \Sigma_0^{-1}) \theta + \theta^\top (N\Sigma^{-1} + \Sigma_0^{-1}) \times \right. \\
 &\quad \times \left. (N\Sigma^{-1} + \Sigma_0^{-1})^{-1} [N\Sigma^{-1} \mu_{ML} + \Sigma_0^{-1} \mu_0] \right\} \\
 p(\theta|\mathbf{x}) &\propto \exp \left\{ -\frac{1}{2} (\theta - (N\Sigma^{-1} + \Sigma_0^{-1})^{-1} [N\Sigma^{-1} \mu_{ML} + \Sigma_0^{-1} \mu_0])^\top \times \right. \\
 &\quad \times \left. (N\Sigma^{-1} + \Sigma_0^{-1}) (\theta - (N\Sigma^{-1} + \Sigma_0^{-1})^{-1} [N\Sigma^{-1} \mu_{ML} + \Sigma_0^{-1} \mu_0]) \right\}.
 \end{aligned}$$

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

$$\begin{aligned}
 \mu_N &= \mathbb{E}[\Theta|\mathbf{X} = \mathbf{x}] \\
 &= (N\Sigma^{-1} + \Sigma_0^{-1})^{-1} [N\Sigma^{-1} \mu_{ML} + \Sigma_0^{-1} \mu_0] \\
 \mu_N &= (\Sigma^{-1} + \Sigma_0^{-1}/N)^{-1} [\Sigma^{-1} \mu_{ML} + \Sigma_0^{-1} \mu_0/N].
 \end{aligned}$$

And covariance

$$\begin{aligned}
 \Sigma_N &= (N\Sigma^{-1} + \Sigma_0^{-1})^{-1} \\
 &= \frac{1}{N} (\Sigma^{-1} + \Sigma_0^{-1}/N)^{-1}.
 \end{aligned}$$

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

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

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 :

$$\begin{aligned}
 \mathbb{E}[X] &= \int_0^\infty xp(x|a, b) dx && \text{(Apply (1.34))} \\
 &= \int_0^\infty x \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\} dx && \text{(Apply (2.146))} \\
 &= \int_0^\infty \frac{b^a}{\Gamma(a)} x^{a+1-1} \exp\{-bx\} dx \\
 &= \frac{\Gamma(a+1)}{b\Gamma(a)} \int_0^\infty \frac{b^{a+1}}{\Gamma(a+1)} x^{a-1} \exp\{-bx\} dx \\
 (2.21) \quad \mathbb{E}[X] &= \frac{a}{b} && \text{(Apply (1.12) and (1.26)).}
 \end{aligned}$$

Thereafter, the variance is computed as

$$\begin{aligned}
 \text{Var}[X] &= \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 && \text{(Apply (1.39))} \\
 &= \int_0^\infty x^2 p(x|a, b) dx - \frac{a^2}{b^2} && \text{(Apply (1.34) and (2.21))} \\
 &= \int_0^\infty x^2 \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\} dx \\
 &= \frac{\Gamma(a+2)}{b^2 \Gamma(a)} \int_0^\infty \frac{b^a}{\Gamma(a+2)} x^{a+2-1} \exp\{-bx\} dx - \frac{a^2}{b^2} && \text{(Apply (1.30))} \\
 &= \frac{(a+1)a}{b^2} - \frac{a^2}{b^2} && \text{(Apply (1.12))} \\
 \text{Var}[X] &= \frac{a}{b^2}.
 \end{aligned}$$

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

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

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 $d \log p(x|a, b)/dx = 0$, 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

$$\begin{aligned}
 \int_{-\infty}^{\infty} p(x|\sigma^2, q) dx &= \int_{-\infty}^{\infty} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{|x|^q}{2\sigma^2}\right\} dx \\
 &= \int_{-\infty}^{0} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{(-x)^q}{2\sigma^2}\right\} dx \\
 &\quad + \int_{0}^{\infty} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{x^q}{2\sigma^2}\right\} dx \\
 &= \int_{0}^{\infty} \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left\{-\frac{x^q}{2\sigma^2}\right\} dx \\
 &= \int_{0}^{\infty} \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \frac{(2\sigma^2)(2\sigma)^{1/q-1}y^{1/q-1}}{q} \exp\{-y\} dy \quad (\text{Set } y = x^q/(2\sigma^2)) \\
 &= \frac{1}{\Gamma(1/y)} \int_0^\infty y^{1/q-1} \exp\{-y\} dy \\
 &= \frac{1}{\Gamma(1/y)} \Gamma(1/y) \tag{Apply (1.141)}
 \end{aligned}$$

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = 1.$$

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 $(\Theta, \Omega)^\top | X = x$ as follows

$$\begin{aligned}
 p(\theta, \omega | x) &\propto p(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 \\
 &\quad \times \omega^{1/2} \exp \left\{ -\frac{\beta\omega}{2} (\theta - \mu_0)^2 \right\} \omega^{a-1} \exp\{-b\omega\} \quad (\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_{ML})^2 + 2b \right] \right\} \times \\
 &\quad \times \exp \left\{ -\frac{\omega}{2} \left[N\theta^2 - 2N\theta\mu_{ML} + N\mu_{ML}^2 + \beta\theta^2 + \right. \right. \\
 &\quad \left. \left. - 2\beta\theta\mu_0 + \beta\mu_0^2 \right] \right\} \\
 &\propto \omega^{(N+1)/2+a-1} \exp \left\{ -\frac{\omega}{2} \left[\sum_{i=1}^N (x_i - \mu_{ML})^2 + 2b \right] \right\} \times \\
 &\quad \times \exp \left\{ -\frac{\omega}{2} \left[\{N + \beta\}\theta^2 + \right. \right. \\
 &\quad \left. \left. - 2\theta\{N\mu_{ML} + \beta\mu_0\} + N\mu_{ML}^2 + \beta\mu_0^2 \right] \right\} \\
 &= \omega^{(N+1)/2+a-1} \exp \left\{ -\frac{\omega}{2} \left[N\sigma_{ML}^2 + 2b \right] \right\} \times \\
 &\quad \times \exp \left\{ -\frac{\omega}{2} \left[\{N + \beta\} \left(\theta - \frac{N\mu_{ML} + \beta\mu_0}{N + \beta} \right)^2 + \right. \right. \\
 &\quad \left. \left. - \frac{N^2\mu_{ML}^2 + 2N\beta\mu_{ML}\mu_0 + \beta^2\mu_0^2}{N + \beta} \right. \right. \\
 &\quad \left. \left. + \frac{N^2\mu_{ML}^2 + N\beta\mu_{ML}^2 + N\beta\mu_0^2 + \beta^2\mu_0^2}{N + \beta} \right] \right\} \\
 p(\theta, \omega | x) &\propto \omega^{N/2+a-1} \exp \left\{ -\omega \left[\frac{1}{2} N\sigma_{ML}^2 + \right. \right. \\
 &\quad \left. \left. + \frac{1}{2} \frac{N\beta(\mu_{ML} - \mu_0)^2}{N + \beta} + b \right] \right\} \\
 &\quad \times (\omega\{N + \beta\})^{1/2} \exp \left\{ -\frac{\omega\{N + \beta\}}{2} \times \right. \\
 &\quad \left. \times \left(\theta - \frac{N\mu_{ML} + \beta\mu_0}{N + \beta} \right)^2 \right\}.
 \end{aligned}$$

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

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

and variance given by

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

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

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

Exercise 2.45

Let us consider that we observe a sample of random variables (denoted by \mathbf{X}) which, conditional on $\Lambda = \mathbf{S}$, 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 Σ 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 $\Lambda | \mathbf{X} = \mathbf{x}$ as follows

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

We note by the functional form of the distribution of $\Lambda | \mathbf{X} = \mathbf{x}$ that it follows a Wishart distribution, with updated parameters defined by $\nu_N = \nu + N$ and $\mathbf{W}_N = (\mathbf{W}^{-1} + N\Sigma_{ML})$. 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

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

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

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

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

Exercise 2.47

We now aim to demonstrate that, for $\nu \rightarrow \infty$, 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

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

We therefore conclude that, applying the limit $\nu \rightarrow \infty$ 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 $\omega\Lambda \in \mathbb{R}^{D \times D}$, 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

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

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

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

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 $\mu \in \mathbb{R}^D$, precision $\Lambda \in \mathbb{R}^{D \times D}$ 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

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

The variance of \mathbf{X} is determined as

$$\begin{aligned}\mathbb{V}\text{ar}[\mathbf{X}] &= \int_{\mathbb{R}^D} (\mathbf{x} - \mathbb{E}[\mathbf{X}]) (\mathbf{x} - \mathbb{E}[\mathbf{X}])^\top p(\mathbf{x}|\mu, \Lambda, \nu) d\mathbf{x} && \text{(Apply (1.42))} \\ &= \int_{\mathbb{R}^D} \int_0^\infty (\mathbf{x} - \mu) (\mathbf{x} - \mu)^\top p(\mathbf{x}|\mu, \omega^{-1}\Lambda^{-1}) p(\omega|\nu) d\omega d\mathbf{x} && \text{(Apply (1.32))} \\ &= \int_0^\infty \left[\int_{\mathbb{R}^D} (\mathbf{x} - \mu) (\mathbf{x} - \mu)^\top p(\mathbf{x}|\mu, \omega^{-1}\Lambda^{-1}) d\mathbf{x} \right] p(\omega|\nu) d\omega \\ &= \int_0^\infty \omega^{-1} \Lambda^{-1} p(\omega|\nu) d\omega && \text{(Apply (2.64))} \\ &= \int_0^\infty \omega^{-1} \Lambda^{-1} \frac{(\frac{\nu}{2})^{\nu/2}}{\Gamma(\nu/2)} \omega^{\nu/2-1} \exp\{-\nu\omega/2\} d\omega && \text{(Apply (2.146))} \\ &= \Lambda^{-1} \frac{\Gamma(\nu/2-1)(\frac{\nu}{2})}{\Gamma(\nu/2)} \int_0^\infty \frac{(\frac{\nu}{2})^{\nu/2-1}}{\Gamma(\nu/2-1)} \omega^{\nu/2-2} \exp\{-\nu\omega/2\} d\omega \\ &= \frac{\frac{\nu}{2}}{\frac{\nu}{2}-1} \Lambda^{-1} && \text{(Apply (1.12) and (1.26))} \\ \mathbb{V}\text{ar}[\mathbf{X}] &= \frac{\nu}{\nu-2} \Lambda^{-1} \quad \nu > 2.\end{aligned}$$

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

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

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

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 $\nu \rightarrow \infty$, 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

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

Observing the final functional form in the right-hand-side, we conclude that for $\nu \rightarrow \infty$, \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

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

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

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

Lastly, we also have that

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

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

$$(2.28) \quad \begin{aligned} \cos x &= 1 - \frac{x^2}{2} + O(x^4) \\ 1 - \cos x &= \frac{x^2}{2} + O(x^4). \end{aligned}$$

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

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

By inspection of the term in the exponent, it thereby follows that, for $m \rightarrow \infty$, $\Xi = m^{1/2}(\Theta - \theta_0)$ 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

$$\begin{aligned}
 & \sum_{n=1}^N \sin(\theta_n - \theta_0) = 0 \\
 & \sum_{n=1}^N (\sin \theta_n \cos \theta_0 - \sin \theta_0 \cos \theta_n) = 0 \quad (\text{Apply (2.27)}) \\
 & \cos \theta_0 \sum_{n=1}^N \sin \theta_n - \sin \theta_0 \sum_{n=1}^N \cos \theta_n = 0 \\
 & \sum_{n=1}^N \sin \theta_n - \frac{\sin \theta_0}{\cos \theta_0} \sum_{n=1}^N \cos \theta_n = 0 \\
 & \tan \theta_0 = \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \\
 & \theta_0 = \arctan \left\{ \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right\}.
 \end{aligned}$$

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

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

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 $\frac{d \log p(\theta|\theta_0, m)}{d\theta} = 0$, 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

$$(2.30) \quad \theta \in \{\theta_0 - \pi, \theta_0, \theta_0 + \pi\}.$$

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

$$\begin{aligned} \frac{d^2 \log p(\theta|\theta_0, m)}{d\theta^2} &< 0 \\ -m \cos(\theta - \theta_0) &< 0 \\ \cos(\theta - \theta_0) &> 0 \\ \theta - \theta_0 &\in \left(\frac{(2k-1)\pi}{2}, \frac{(2k+1)\pi}{2} \right) \quad k \in \{\dots, -2, 0, 2, \dots\} \\ (2.31) \quad \theta &\in \left(\theta_0 + \frac{(2k-1)\pi}{2}, \theta_0 + \frac{(2k+1)\pi}{2} \right) \quad k \in \{\dots, -2, 0, 2, \dots\}. \end{aligned}$$

Similarly to before, we may restrict our candidate points to

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

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

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

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

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

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 $\mu \in \mathbb{R}^D$ and variance $\Sigma \in \mathbb{R}^{D \times D}$ belongs to the exponential family. In order to do so, we rewrite (2.43) as follows

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

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

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

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 η , obtaining the following

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

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

from which we obtain the following

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

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

Differentiating again both sides of (2.195) with respect to η , we obtain

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

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

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

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

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

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

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

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

$$(2.35) \quad \sum_{i \geq 1} h_i \Delta_i = 1.$$

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)

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

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 in this context is done via

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

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

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

(3.1) $\tanh(a) = 2\sigma(2a) - 1$ (Apply (3.6)).

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

$$\begin{aligned}
 y(x, \mathbf{w}) &= y(x, \mathbf{u}) \\
 w_0 + \sum_{j=1}^M w_j \sigma\left(\frac{x - \mu_j}{s}\right) &= u_0 + \sum_{j=1}^M u_j \tanh\left(\frac{x - \mu_j}{2s}\right) && \text{(Apply (3.101) and (3.102))} \\
 w_0 + \sum_{j=1}^M w_j \sigma\left(\frac{x - \mu_j}{s}\right) &= u_0 - \sum_{j=1}^M u_j + \sum_{j=1}^M 2u_j \sigma\left(\frac{x - \mu_j}{s}\right) && \text{(Apply (3.1)).}
 \end{aligned}$$

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

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

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

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

We denote the column elements of Φ as φ_j , such that

$$(3.3) \quad \Phi = \begin{pmatrix} \varphi_1 & \dots & \varphi_M \end{pmatrix}.$$

We thereby define the following

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

We may finally therefore rewrite (3.2) as follows

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

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 Φ , operations of the form (3.2) project vectors \mathbf{v} into

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

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

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

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

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 Φ , denoted by \mathcal{S} . Figure 3.1 provides a rough sketch on the geometric intuition behind (3.2).

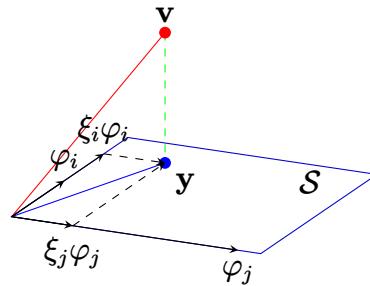


Figure 3.1: Illustration of the least-squares solution provided in (3.2).

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

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

Solving for $dE_D(\mathbf{w})/d\mathbf{w} = \mathbf{0}$, we find

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

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 $\mathbf{w}^* = (\Phi^\top \mathbf{R} \Phi)^{-1} \Phi^\top \mathbf{R} \mathbf{t}$, 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 $\{y_n, \phi(\mathbf{x}_n)\}$ ($r_n \propto \{\text{Var}[\epsilon_n]\}^{-1}$), 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 $\{y_n, \phi(\mathbf{x}_n)\}$, 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 $\text{Var}[\epsilon_i] = \sigma^2$. We desire to minimize the squared-loss error function averaged over the input noise. We write the following

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

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, \dots, 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

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

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

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

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

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

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 $g(\mathbf{w}^*) < 0$, 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 $g(\mathbf{w}^*) = 0$, and moreover that

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

Exercise 3.6

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

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

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

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

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

$$\begin{aligned} \mathbf{W}_{ML} &= (\Phi^\top \Phi)^{-1} \Phi^\top (T_1 \ \dots \ T_D) \\ &= ((\Phi^\top \Phi)^{-1} \Phi^\top T_1 \ \dots \ (\Phi^\top \Phi)^{-1} \Phi^\top T_D). \end{aligned}$$

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

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

We thereby conclude that the maximum likelihood point, with respect to Σ , occurs at

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

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

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

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 $\mathbf{w}|T$. It is as follows

$$\begin{aligned}
 p(\mathbf{w}|T) &\propto p(T|\mathbf{w})p(\mathbf{w}) \\
 &\propto \exp \left\{ -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 \right\} \times \\
 &\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{w} - \mathbf{m}_0)^\top S_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} \quad (\text{Apply (2.43) and (2.137)}) \\
 &\propto \exp \left\{ \beta \sum_{n=1}^N t_n \mathbf{w}^\top \phi(\mathbf{x}_n) + \right. \\
 &\quad \left. - \frac{\beta}{2} \sum_{n=1}^N \mathbf{w}^\top \phi(\mathbf{x}_n) \{\phi(\mathbf{x}_n)\}^\top \mathbf{w} + \right. \\
 &\quad \left. - \frac{1}{2} \mathbf{w}^\top S_0^{-1} \mathbf{w} + \mathbf{w}^\top S_0^{-1} \mathbf{m}_0 \right\} \\
 &= \exp \left\{ \mathbf{w}^\top \left[\beta \sum_{n=1}^N t_n \phi(\mathbf{x}_n) + S_0^{-1} \mathbf{m}_0 \right] + \right. \\
 &\quad \left. - \frac{1}{2} \mathbf{w}^\top \left[\beta \sum_{n=1}^N \phi(\mathbf{x}_n) \{\phi(\mathbf{x}_n)\}^\top + S_0^{-1} \right] \mathbf{w} \right\} \\
 &= \exp \left\{ \mathbf{w}^\top \left[\beta \Phi^\top \mathbf{t} + S_0^{-1} \mathbf{m}_0 \right] + \right. \\
 &\quad \left. - \frac{1}{2} \mathbf{w}^\top \left[\beta \Phi^\top \Phi + S_0^{-1} \right] \mathbf{w} \right\} \\
 p(\mathbf{w}|T) &\propto \exp \left\{ -\frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top S_N^{-1} (\mathbf{w} - \mathbf{m}_N) \right\},
 \end{aligned}$$

where

$$(3.12) \quad \mathbf{m}_N = S_N \left[\beta \Phi^\top \mathbf{t} + S_0^{-1} \mathbf{m}_0 \right]$$

$$(3.13) \quad \mathbf{S}_N = \left[\beta \Phi^\top \Phi + S_0^{-1} \right]^{-1}.$$

Hence, via the method of completing the squares, we conclude that $\mathbf{w}|T$ 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

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

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

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

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

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

And

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

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

$$\begin{aligned} p(\mathbf{w}|\mathbf{T}) &= \text{MULTIVARIATE NORMAL}(\mathbf{m}_N, \mathbf{S}_N) \\ p(t_{N+1}|\mathbf{w}) &= \text{NORMAL}(\{\phi(\mathbf{x}_{N+1})\}^\top \mathbf{w}, \beta^{-1}). \end{aligned}$$

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

$$p(\mathbf{w}|\mathbf{T}, t_{N+1}) = \text{MULTIVARIATE NORMAL}(\mathbf{m}^*, \mathbf{S}^*)$$

where

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

Note that the formulae for \mathbf{m}^* and $\tilde{\mathbf{m}}$ (also \mathbf{S}^* 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:

$$\begin{aligned} p(\mathbf{w}|\mathbf{T}) &= \text{MULTIVARIATE NORMAL}(\mathbf{m}_N, \mathbf{S}_N) \\ p(t|\mathbf{w}) &= \text{NORMAL}(\{\phi(\mathbf{x})\}^\top \mathbf{w}, \beta^{-1}). \end{aligned}$$

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

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

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

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

Exercise 3.11

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

$$(3.16) \quad \begin{aligned} \mathbf{S}_{N+1} &= [\mathbf{S}_N^{-1} + \beta\phi(\mathbf{x}_{N+1})\{\phi(\mathbf{x}_{N+1})\}^\top]^{-1} \\ \mathbf{S}_{N+1} &= \mathbf{S}_N - \frac{\beta\mathbf{S}_N\phi(\mathbf{x}_{N+1})\{\phi(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N}{1 + \beta\{\phi(\mathbf{x}_{N+1})\}^\top\mathbf{S}_N\phi(\mathbf{x}_{N+1})} \end{aligned} \quad (\text{Apply (2.289)}).$$

It follows that

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

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

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}).$$

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

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

where

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

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

$$(3.19) \quad a_N = \frac{N}{2} + a_0$$

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

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 $\mathbf{w}|\mathbf{T}$, we utilize the linear-Gaussian model framework to analyse the Bayesian linear regression model. Consider the following hierarchical model:

$$\begin{aligned} p(\mathbf{w}|\mathbf{T}, \beta) &= \text{MULTIVARIATE NORMAL}(\mathbf{m}_N, \beta^{-1}\mathbf{S}_N) \\ p(t|\mathbf{w}, \beta) &= \text{NORMAL}(\{\phi(\mathbf{x})\}^\top \mathbf{w}, \beta^{-1}). \end{aligned}$$

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

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

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

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

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

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

It follows that

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

$$\sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) = 1.$$

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

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

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

$$\begin{aligned} p(\mathbf{w}|\alpha) &= \text{MULTIVARIATE NORMAL}(\mathbf{0}, \alpha^{-1}\mathbf{I}) \\ p(\mathbf{t}|\mathbf{w}, \beta) &= \text{MULTIVARIATE NORMAL}(\Phi\mathbf{w}, \beta^{-1}\mathbf{I}). \end{aligned}$$

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

$$p(\mathbf{t}|\alpha, \beta) = \text{MULTIVARIATE NORMAL}(\mathbf{0}, \beta^{-1}\mathbf{I} + \alpha^{-1}\Phi\Phi^\top).$$

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

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

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

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

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

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

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

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

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} - \Phi \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} \\
 &= \frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^\top (\mathbf{t} - \Phi \mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} \\
 &= \frac{\beta}{2} \mathbf{t}^\top \mathbf{t} - \beta \mathbf{w}^\top \Phi^\top \mathbf{t} + \frac{\beta}{2} \mathbf{w}^\top \Phi^\top \Phi \mathbf{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 \mathbf{I} + \beta \Phi^\top \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} - \Phi \mathbf{m}_N)^\top (\mathbf{t} - \Phi \mathbf{m}_N) + \beta \mathbf{t}^\top \Phi \mathbf{m}_N - \frac{\beta}{2} \mathbf{m}_N^\top \Phi^\top \Phi \mathbf{m}_N + \\
 &\quad - \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} - \Phi \mathbf{m}_N)^\top (\mathbf{t} - \Phi \mathbf{m}_N) + \mathbf{m}_N^\top \mathbf{A} \mathbf{m}_N - \frac{\beta}{2} \mathbf{m}_N^\top \Phi^\top \Phi \mathbf{m}_N + \\
 &\quad - \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} - \Phi \mathbf{m}_N\|^2 + \frac{1}{2} \mathbf{m}_N^\top (\alpha \mathbf{I} + \beta \Phi^\top \Phi) \mathbf{m}_N + \\
 &\quad - \frac{\beta}{2} \mathbf{m}_N^\top \Phi^\top \Phi \mathbf{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} - \Phi \mathbf{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

$$\begin{aligned}
 \int \exp\{-E(w)\} dw &= \int \exp \left\{ -E(\mathbf{m}_N) + \right. \\
 &\quad \left. - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T \mathbf{A}(\mathbf{w} - \mathbf{m}_N) \right\} dw \quad (\text{Apply (3.80)}) \\
 &= \exp\{-E(\mathbf{m}_N)\} \times \\
 &\quad \times \int \exp \left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T \mathbf{A}(\mathbf{w} - \mathbf{m}_N) \right\} dw \quad (\text{Apply (3.80)}) \\
 &= \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \exp\{-E(\mathbf{m}_N)\} \times \\
 &\quad \times \int \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \times \\
 &\quad \times \exp \left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T \mathbf{A}(\mathbf{w} - \mathbf{m}_N) \right\} dw \\
 (3.25) \quad \int \exp\{-E(w)\} dw &= \exp\{-E(\mathbf{m}_N)\} (2\pi)^{M/2} |\mathbf{A}|^{-1/2} \quad (\text{Apply (1.30)}).
 \end{aligned}$$

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

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

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

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

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

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

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 $d \log p(\mathbf{t}|\alpha, \beta)/d\alpha = 0$, obtaining the following

$$\begin{aligned} \frac{d \log p(\mathbf{t}|\alpha, \beta)}{d\alpha} &= 0 \\ \frac{M}{2\alpha} - \sum_{i=1}^M \frac{1}{2(\alpha + \lambda_i)} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N &= 0 \\ M - \sum_{i=1}^M \frac{\alpha}{\alpha + \lambda_i} - \alpha \mathbf{m}_N^\top \mathbf{m}_N &= 0 \\ \sum_{i=1}^M \left(1 - \frac{\alpha}{\alpha + \lambda_i}\right) &= \alpha \mathbf{m}_N^\top \mathbf{m}_N \\ \sum_{i=1}^M \frac{\lambda_i}{\alpha + \lambda_i} &= \alpha \mathbf{m}_N^\top \mathbf{m}_N \\ \frac{\gamma}{\mathbf{m}_N^\top \mathbf{m}_N} &= \alpha && (\text{Apply (3.91)}). \end{aligned}$$

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

$$\begin{aligned}\frac{d}{d\alpha} \log|\mathbf{A}| &= \frac{d}{d\alpha} \log \prod_{i=1}^M \lambda_i \quad (\text{Apply (C.47)}) \\ &= \frac{d}{d\alpha} \sum_{i=1}^M \log \lambda_i \\ \frac{d}{d\alpha} \log|\mathbf{A}| &= \sum_{i=1}^M \frac{1}{\lambda_i} \frac{d\lambda_i}{d\alpha}.\end{aligned}$$

Now, in order to determine the value of $\text{tr}(\mathbf{A}^{-1} d/d\alpha \mathbf{A})$, we perform as follows

$$\begin{aligned}\mathbf{A} &= \sum_{i=1}^M \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \quad (\text{Apply (2.48)}) \\ \frac{d}{d\alpha} \mathbf{A} &= \sum_{i=1}^M \frac{d\lambda_i}{d\alpha} \mathbf{u}_i \mathbf{u}_i^\top + \sum_{i=1}^M \lambda_i \frac{d}{d\alpha} (\mathbf{u}_i \mathbf{u}_i^\top) \\ \mathbf{A}^{-1} \left(\frac{d}{d\alpha} \mathbf{A} \right) &= \mathbf{A}^{-1} \sum_{i=1}^M \frac{d\lambda_i}{d\alpha} \mathbf{u}_i \mathbf{u}_i^\top + \mathbf{A}^{-1} \sum_{i=1}^M \lambda_i \frac{d}{d\alpha} (\mathbf{u}_i \mathbf{u}_i^\top) \\ \mathbf{A}^{-1} \left(\frac{d}{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{d\lambda_i}{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{d}{d\alpha} (\mathbf{u}_i \mathbf{u}_i^\top) \right\} \quad (\text{Apply (2.49)}) \\ &= \sum_{j=1}^M \sum_{i=1}^M \frac{1}{\lambda_j} \frac{d\lambda_i}{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{d\mathbf{u}_i}{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{d\mathbf{u}_i^\top}{d\alpha} \\ &= \sum_{j=1}^M \sum_{i=1}^M \frac{1}{\lambda_j} \frac{d\lambda_i}{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{d\mathbf{u}_i}{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{d\mathbf{u}_i^\top}{d\alpha} \quad (\text{Apply (2.46)}) \\ \text{tr} \left\{ \mathbf{A}^{-1} \left(\frac{d}{d\alpha} \mathbf{A} \right) \right\} &= \sum_{i=1}^M \frac{1}{\lambda_i} \frac{d\lambda_i}{d\alpha} \text{tr}(\mathbf{u}_i^\top \mathbf{u}_i) + \\ &\quad + \sum_{j=1}^M \sum_{i=1}^M \frac{\lambda_i}{\lambda_j} \text{tr} \left\{ \mathbf{u}_j^\top \frac{d\mathbf{u}_i}{d\alpha} (\mathbf{u}_i^\top \mathbf{u}_j) \right\} + \sum_{i=1}^M \text{tr} \left\{ \frac{d\mathbf{u}_i^\top}{d\alpha} \mathbf{u}_i \right\} \quad (\text{Apply (C.9)}) \\ &= \sum_{i=1}^M \frac{1}{\lambda_i} \frac{d\lambda_i}{d\alpha} + \sum_{i=1}^M \text{tr} \left\{ \mathbf{u}_i^\top \frac{d\mathbf{u}_i}{d\alpha} \right\} + \sum_{i=1}^M \text{tr} \left\{ \frac{d\mathbf{u}_i^\top}{d\alpha} \mathbf{u}_i \right\} \quad (\text{Apply (2.46)}) \\ &= \sum_{i=1}^M \frac{1}{\lambda_i} \frac{d\lambda_i}{d\alpha} + \sum_{i=1}^M \text{tr} \left\{ \frac{d}{d\alpha} (\mathbf{u}_i^\top \mathbf{u}_i) \right\} \\ \text{tr} \left\{ \mathbf{A}^{-1} \left(\frac{d}{d\alpha} \mathbf{A} \right) \right\} &= \sum_{i=1}^M \frac{1}{\lambda_i} \frac{d\lambda_i}{d\alpha} \quad (\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

$$\begin{aligned}
 \frac{d}{d\alpha} \log p(\mathbf{t}|\alpha, \beta) &= \frac{d}{d\alpha} \left[\frac{M}{2} \log \alpha + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) + \right. \\
 &\quad \left. - \frac{1}{2} \log |\mathbf{A}| - E(\mathbf{m}_N) \right] \\
 &= \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N - \frac{1}{2} \frac{d}{d\alpha} \log |\mathbf{A}| \quad (\text{Apply (3.82)}) \\
 &= \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N - \frac{1}{2} \text{tr} \left(\mathbf{A}^{-1} \frac{d}{d\alpha} \mathbf{A} \right) \quad (\text{Apply (3.117)}) \\
 &= \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N - \frac{1}{2} \text{tr} \left(\mathbf{A}^{-1} \frac{d}{d\alpha} [\alpha \mathbf{I} + \beta \Phi^\top \Phi] \right) \\
 &= \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N - \frac{1}{2} \text{tr}(\mathbf{A}^{-1}) \quad (\text{Apply (C.19)}) \\
 \frac{d}{d\alpha} \log p(\mathbf{t}|\alpha, \beta) &= \frac{M}{2\alpha} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N - \frac{1}{2} \sum_{i=1}^M \frac{1}{\alpha + \lambda_i} \quad (\text{Apply (C.48)}).
 \end{aligned}$$

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 $d \log p(\mathbf{t}|\alpha, \beta) / d\alpha = 0$ 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:

$$\begin{aligned}
 \frac{d}{d\beta} \log|\mathbf{A}| &= \text{tr}\left(\mathbf{A}^{-1} \frac{d}{d\beta} \mathbf{A}\right) && \text{(Apply (3.117))} \\
 &= \text{tr}\left([\alpha \mathbf{I} + \beta \Phi^\top \Phi]^{-1} \frac{d}{d\beta} [\alpha \mathbf{I} + \beta \Phi^\top \Phi]\right) && \text{(Apply (3.81))} \\
 &= \text{tr}\left([\alpha \mathbf{I} + \beta \Phi^\top \Phi]^{-1} \Phi^\top \Phi\right) && \text{(Apply (C.19))} \\
 &= \text{tr}\left(\beta^{-1} [\alpha \mathbf{I} + \beta \Phi^\top \Phi]^{-1} (\alpha \mathbf{I} + \beta \Phi^\top \Phi) + \right. \\
 &\quad \left. - [\alpha \mathbf{I} + \beta \Phi^\top \Phi]^{-1} \alpha \beta^{-1} \mathbf{I}\right) \\
 &= \frac{1}{\beta} \text{tr}(\mathbf{I}) - \frac{\alpha}{\beta} \text{tr}([\alpha \mathbf{I} + \beta \Phi^\top \Phi]^{-1}) \\
 &= \frac{M}{\beta} - \frac{\alpha}{\beta} \sum_{i=1}^M \frac{1}{\lambda_i + \alpha} && \text{(Apply (C.48))} \\
 &= \frac{1}{\beta} \left[M - \sum_{i=1}^M \frac{\lambda_i + \alpha - \lambda_i}{\lambda_i + \alpha} \right] \\
 &= \frac{1}{\beta} \left[M - \sum_{i=1}^M \frac{\lambda_i + \alpha}{\lambda_i + \alpha} + \sum_{i=1}^M \frac{\lambda_i}{\lambda_i + \alpha} \right] \\
 &= \frac{1}{\beta} \sum_{i=1}^M \frac{\lambda_i}{\lambda_i + \alpha} \\
 (3.27) \quad \frac{d}{d\beta} \log|\mathbf{A}| &= \frac{\gamma}{\beta} && \text{(Apply (3.91)).}
 \end{aligned}$$

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

$$\begin{aligned}
 \frac{dp(\mathbf{t}|\alpha, \beta)}{d\beta} &= \frac{d}{d\beta} \left[\frac{M}{2} \log \alpha + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) + \right. \\
 &\quad \left. - \frac{1}{2} \log |\mathbf{A}| - E(\mathbf{m}_N) \right] \\
 &= \frac{d}{d\beta} \left[\frac{M}{2} \log \alpha + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) + \right. \\
 &\quad \left. - \frac{1}{2} \log |\mathbf{A}| - \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 - \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N \right] \quad (\text{Apply (3.82)}) \\
 \frac{dp(\mathbf{t}|\alpha, \beta)}{d\beta} &= \frac{N}{2\beta} - \frac{\gamma}{2\beta} - \frac{1}{2} \sum_{n=1}^N \{t_N - \mathbf{m}_N^\top \phi(\mathbf{x}_n)\}^2 \quad (\text{Apply (3.27)}).
 \end{aligned}$$

Taking $\frac{dp(\mathbf{t}|\alpha, \beta)}{d\beta} = 0$ and solving for β , we obtain

$$\begin{aligned}
 \frac{dp(\mathbf{t}|\alpha, \beta)}{d\beta} &= 0 \\
 \frac{N}{2\beta} - \frac{\gamma}{2\beta} - \frac{1}{2} \sum_{n=1}^N \{t_N - \mathbf{m}_N^\top \phi(\mathbf{x}_n)\}^2 &= 0 \\
 \frac{N - \gamma}{\beta} &= \sum_{n=1}^N \{t_N - \mathbf{m}_N^\top \phi(\mathbf{x}_n)\}^2 \\
 \frac{1}{\beta} &= \frac{1}{N - \gamma} \sum_{n=1}^N \{t_N - \mathbf{m}_N^\top \phi(\mathbf{x}_n)\}^2
 \end{aligned}$$

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 $p(\mathbf{t}|\beta)$

$$\begin{aligned}
 p(\mathbf{t}|\beta) &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\beta) d\mathbf{w} && \text{(Apply (1.31))} \\
 &= \int \frac{\beta^{N/2}}{(2\pi)^{N/2}} \times \\
 &\quad \times \exp \left\{ -\frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^\top (\mathbf{t} - \Phi \mathbf{w}) \right\} \times \\
 &\quad \times \frac{\beta^{M/2}}{(2\pi)^{M/2} |\mathbf{S}_0|^{1/2}} \times \\
 &\quad \times \exp \left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} && \text{(Apply (2.43))} \\
 &= \frac{\beta^{N/2} |\mathbf{S}_N|^{1/2}}{(2\pi)^{N/2} |\mathbf{S}_0|^{1/2}} \times \\
 &\quad \times \exp \left\{ -\frac{\beta}{2} \mathbf{t}^\top \mathbf{t} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N \mathbf{m}_N \right\} \times \\
 &\quad \times \int \frac{\beta^{M/2}}{(2\pi)^{M/2} |\mathbf{S}_N|^{1/2}} \times \\
 &\quad \times \exp \left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) \right\} d\mathbf{w} && \text{(Apply (3.17) and (3.18))} \\
 (3.28) \quad p(\mathbf{t}|\beta) &= \frac{\beta^{N/2} |\mathbf{S}_N|^{1/2}}{(2\pi)^{N/2} |\mathbf{S}_0|^{1/2}} \times \\
 &\quad \times \exp \left\{ -\frac{\beta}{2} \mathbf{t}^\top \mathbf{t} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N \mathbf{m}_N \right\} && \text{(Apply (1.30)).}
 \end{aligned}$$

It follows thereafter that the marginal distribution $p(\mathbf{t})$ is given as follows

$$\begin{aligned}
 p(\mathbf{t}) &= \int_0^\infty p(\mathbf{t}|\beta)p(\beta) d\beta && \text{(Apply (1.31))} \\
 &= \int_0^\infty \frac{b_0^{a_0}}{\Gamma(a_0)} \beta^{a_0-1} \exp\{-b_0\beta\} \times \\
 &\quad \times \frac{\beta^{N/2} |\mathbf{S}_N|^{1/2}}{(2\pi)^{N/2} |\mathbf{S}_0|^{1/2}} \times \\
 &\quad \times \exp \left\{ -\frac{\beta}{2} \mathbf{t}^\top \mathbf{t} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N \mathbf{m}_N \right\} d\beta && \text{(Apply (2.146) and (3.28))} \\
 &= \frac{1}{(2\pi)^{N/2}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \times \\
 &\quad \times \int_0^\infty \frac{b_N^{a_N}}{\Gamma(a_N)} \beta^{a_N-1} \times \\
 &\quad \times \exp \left\{ -b_0\beta - \frac{\beta}{2} \mathbf{t}^\top \mathbf{t} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N \mathbf{m}_N \right\} d\beta && \text{(Apply (3.19) and (3.20))} \\
 p(\mathbf{t}) &= \frac{1}{(2\pi)^{N/2}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} && \text{(Apply (1.26)).}
 \end{aligned}$$

Thereby deriving the desired result.

Exercise 3.24

We now aim to 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{|\mathbf{S}_N|^{1/2}}{|\mathbf{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{|\mathbf{S}_N|^{1/2}}{|\mathbf{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{|\mathbf{S}_N|^{1/2}}{|\mathbf{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 \Phi^\top \Phi \mathbf{w} - \mathbf{w}^\top \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{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \frac{p(\mathbf{t} | \mathbf{w}, \beta)}{\beta^{N/2}} \times \\
 &\quad \times \exp \left\{ \frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^\top (\mathbf{t} - \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{|\mathbf{S}_N|^{1/2}}{|\mathbf{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

$$(4.1) \quad \mathbf{z} = \sum_{n=1}^N \alpha_n \mathbf{x}_n$$

$$(4.2) \quad \mathbf{z} = \sum_{m=1}^M \beta_m \mathbf{y}_m,$$

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

$$\begin{aligned} \hat{\mathbf{w}}^\top \mathbf{z} + w_0 &= \hat{\mathbf{w}}^\top \sum_{n=1}^N \alpha_n \mathbf{x}_n + w_0 \\ &= \sum_{n=1}^N \alpha_n \hat{\mathbf{w}}^\top \mathbf{x}_n + \sum_{n=1}^N \alpha_n w_0 && (\text{Apply } \sum_{n=1}^N \alpha_n = 1) \\ &= \sum_{n=1}^N \alpha_n (\hat{\mathbf{w}}^\top \mathbf{x}_n + w_0) \\ (4.3) \quad \mathbf{z} > 0 && & (\text{By assumption}). \end{aligned}$$

However, If we left-multiply (4.2) by $\hat{\mathbf{w}}$ and sum w_0 to it, we find

$$\begin{aligned} \hat{\mathbf{w}}^\top \mathbf{z} + w_0 &= \hat{\mathbf{w}}^\top \sum_{m=1}^M \beta_m \mathbf{y}_m + w_0 \\ &= \sum_{m=1}^M \beta_m \hat{\mathbf{w}}^\top \mathbf{y}_m + \sum_{m=1}^M \beta_m w_0 && (\text{Apply } \sum_{m=1}^M \beta_m = 1) \\ &= \sum_{m=1}^M \beta_m (\hat{\mathbf{w}}^\top \mathbf{y}_m + w_0) \\ (4.4) \quad \mathbf{z} < 0 && & (\text{By assumption}). \end{aligned}$$

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 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

$$\begin{aligned}
 \mathbf{y}(\mathbf{x}) &= \tilde{\mathbf{W}}\phi(\mathbf{x}) \\
 &= \mathbf{T}^\top \Phi(\Phi^\top \Phi)^{-1}\phi(\mathbf{x}) && \text{(Apply (4.16))} \\
 &= (\mathbf{t}_1 \dots \mathbf{t}_N) \Phi(\Phi^\top \Phi)^{-1}\phi(\mathbf{x}) \\
 \mathbf{a}^\top \mathbf{y}(\mathbf{x}) &= (\mathbf{a}^\top \mathbf{t}_1 + b - b \dots \mathbf{a}^\top \mathbf{t}_N + b - b) \Phi(\Phi^\top \Phi)^{-1}\phi(\mathbf{x}) \\
 (4.5) \quad \mathbf{a}^\top \mathbf{y}(\mathbf{x}) &= -b\mathbf{1}^\top \Phi(\Phi^\top \Phi)^{-1}\phi(\mathbf{x}).
 \end{aligned}$$

We rewrite our input variable matrix as

$$(4.6) \quad \Phi = (\mathbf{1} \ \mathbf{P}),$$

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

$$\begin{aligned}
 \Phi^\top \Phi &= \begin{pmatrix} \mathbf{1}^\top \\ \mathbf{P}^\top \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{P} \end{pmatrix} && \text{(Apply (4.6))} \\
 &= \begin{pmatrix} N & \mathbf{1}^\top \mathbf{P} \\ \mathbf{P}^\top \mathbf{1} & \mathbf{P}^\top \mathbf{P} \end{pmatrix} \\
 (\Phi^\top \Phi)^{-1} &= \begin{pmatrix} N & \mathbf{1}^\top \mathbf{P} \\ \mathbf{P}^\top \mathbf{1} & \mathbf{P}^\top \mathbf{P} \end{pmatrix}^{-1} \\
 (\Phi^\top \Phi)^{-1} &= \ell \begin{pmatrix} 1 & -\mathbf{1}^\top \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \\ -(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top \mathbf{1} & \ell^{-1}(\mathbf{P}^\top \mathbf{P})^{-1} + (\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top \mathbf{1} \mathbf{1}^\top \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \end{pmatrix} && \text{(Apply (2.76)).}
 \end{aligned}$$

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 Φ , we obtain

$$\begin{aligned}
 \Phi(\Phi^\top \Phi)^{-1} &= \ell \begin{pmatrix} \{\mathbf{1} - \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top \mathbf{1}\}^\top \\ \{-\mathbf{1} \mathbf{1}^\top \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} + \ell^{-1} \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} + \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top \mathbf{1} \mathbf{1}^\top \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
 &= \ell \begin{pmatrix} \{[\mathbf{I} - \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top] \mathbf{1}\}^\top \\ \{[\ell^{-1} \mathbf{I} - \{\mathbf{I} - \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top\} \mathbf{1} \mathbf{1}^\top] \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
 \mathbf{1}^\top \Phi(\Phi^\top \Phi)^{-1} &= \ell \begin{pmatrix} \mathbf{1}^\top [\mathbf{I} - \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top] \mathbf{1} \\ \{[\ell^{-1} \mathbf{1}^\top - \mathbf{1}^\top \{\mathbf{I} - \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1} \mathbf{P}^\top\} \mathbf{1} \mathbf{1}^\top] \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
 &= \ell \begin{pmatrix} \ell^{-1} \\ \{[\ell^{-1} \mathbf{1}^\top - \ell^{-1} \mathbf{1}^\top] \mathbf{P}(\mathbf{P}^\top \mathbf{P})^{-1}\}^\top \end{pmatrix}^\top \\
 (4.7) \quad \mathbf{1}^\top \Phi(\Phi^\top \Phi)^{-1} &= (1 \ \mathbf{0}^\top).
 \end{aligned}$$

Consider now that we rewrite our input vector for prediction $\phi(\mathbf{x})$ as

$$(4.8) \quad \phi(\mathbf{x}) = \begin{pmatrix} \phi_0(\mathbf{x}) \\ \phi(\mathbf{x}) \end{pmatrix}.$$

Note that $\phi_0(\mathbf{x}) = 1$. It follows that, by substituting (4.7) and (4.8) into (4.5), we obtain

$$\begin{aligned} \mathbf{a}^\top \mathbf{y}(\mathbf{x}) &= -b \begin{pmatrix} 1 & \mathbf{0}^\top \end{pmatrix} \begin{pmatrix} 1 \\ \phi(\mathbf{x}) \end{pmatrix} \\ &= -b\{1 + \mathbf{0}^\top \phi(\mathbf{x})\} \\ &= -b \\ \mathbf{a}^\top \mathbf{y}(\mathbf{x}) + b &= 0. \end{aligned}$$

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

$$\begin{aligned} \mathbf{y}(\mathbf{x}) &= \tilde{\mathbf{W}}\phi(\mathbf{x}) \\ &= \mathbf{T}^\top \Phi (\Phi^\top \Phi)^{-1} \phi(\mathbf{x}) && \text{(Apply (4.16))} \\ &= \begin{pmatrix} \mathbf{t}_1 & \dots & \mathbf{t}_N \end{pmatrix} \Phi (\Phi^\top \Phi)^{-1} \phi(\mathbf{x}) \\ \mathbf{A}\mathbf{y}(\mathbf{x}) &= \begin{pmatrix} \mathbf{A}\mathbf{t}_1 + \mathbf{b} - \mathbf{b} & \dots & \mathbf{A}\mathbf{t}_N + \mathbf{b} - \mathbf{b} \end{pmatrix} \Phi (\Phi^\top \Phi)^{-1} \phi(\mathbf{x}) \\ (4.9) \quad \mathbf{A}\mathbf{y}(\mathbf{x}) &= -\mathbf{b}\mathbf{1}^\top \Phi (\Phi^\top \Phi)^{-1} \phi(\mathbf{x}). \end{aligned}$$

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

$$(4.10) \quad L(\mathbf{w}, \lambda) = \mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1) + \lambda(\mathbf{w}^\top \mathbf{w} - 1).$$

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

$$\begin{aligned} \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} &= \mathbf{0} \\ (\mathbf{m}_2 - \mathbf{m}_1) + 2\lambda \mathbf{w} &= \mathbf{0} \quad (\text{Apply (C.19)}) \\ (4.11) \quad \mathbf{w} &= -\frac{\mathbf{m}_2 - \mathbf{m}_1}{2\lambda}. \end{aligned}$$

Substituting into the constraint, we find

$$\begin{aligned} \mathbf{w}^\top \mathbf{w} - 1 &= 0 \\ \frac{(\mathbf{m}_2 - \mathbf{m}_1)^\top (\mathbf{m}_2 - \mathbf{m}_1)}{4\lambda^2} &= 1 \quad (\text{Apply (4.11)}) \\ (4.12) \quad \lambda &= \frac{\sqrt{(\mathbf{m}_2 - \mathbf{m}_1)^\top (\mathbf{m}_2 - \mathbf{m}_1)}}{2}. \end{aligned}$$

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

$$\begin{aligned}
 s_1^2 + s_2^2 &= \sum_{n \in \mathcal{C}_1} (y_n - m_1)^2 + \sum_{n \in \mathcal{C}_2} (y_n - m_2)^2 && \text{(Apply (4.24))} \\
 &= \sum_{n \in \mathcal{C}_1} (\mathbf{w}^\top \mathbf{x}_n - \mathbf{w}^\top \mathbf{m}_1)^2 + \\
 &\quad + \sum_{n \in \mathcal{C}_2} (\mathbf{w}^\top \mathbf{x}_n - \mathbf{w}^\top \mathbf{m}_2)^2 && \text{(Apply (4.20) and (4.23))} \\
 &= \sum_{n \in \mathcal{C}_1} \{\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_1)\}^2 + \\
 &\quad + \sum_{n \in \mathcal{C}_2} \{\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_2)\}^2 \\
 &= \sum_{n \in \mathcal{C}_1} \mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top \mathbf{w} + \\
 &\quad + \sum_{n \in \mathcal{C}_2} \mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \mathbf{w} \\
 &= \mathbf{w}^\top \left[\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \right. \\
 &\quad \left. + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \right] \mathbf{w} \\
 (4.13) \quad s_1^2 + s_2^2 &= \mathbf{w}^\top \mathbf{S}_{\mathbf{W}} \mathbf{w} && \text{(Apply (4.28)).}
 \end{aligned}$$

It follows that we may now rewrite (4.25) as

$$\begin{aligned}
 J(\mathbf{w}) &= \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} && \text{(Apply (4.25))} \\
 &= \frac{\{\mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1)\}^2}{\mathbf{w}^\top \mathbf{S}_{\mathbf{W}} \mathbf{w}} && \text{(Apply (4.13) and (4.22))} \\
 &= \frac{\mathbf{w}^\top (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_{\mathbf{W}} \mathbf{w}} \\
 J(\mathbf{w}) &= \frac{\mathbf{w}^\top \mathbf{S}_{\mathbf{B}} \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_{\mathbf{W}} \mathbf{w}} && \text{(Apply (4.27)).}
 \end{aligned}$$

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}_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} \tag{Apply (4.28)} \\
 N(\mathbf{m}_1 - \mathbf{m}_2) &= \mathbf{S}_W \mathbf{w} + \frac{N_1 N_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_1 N_2}{N} \mathbf{m}_2 \mathbf{m}_1^\top \mathbf{w} + \\
 &\quad + \frac{N_1 N_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_1 N_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}_W \mathbf{w} + \frac{N_1 N_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}_W \mathbf{w} + \frac{N_1 N_2}{N} \mathbf{S}_B \mathbf{w} \tag{Apply (4.27)} \\
 \left(\mathbf{S}_W + \frac{N_1 N_2}{N} \mathbf{S}_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

$$\begin{aligned}
 \sigma(a) &= \frac{1}{1 + \exp\{-a\}} && \text{(Apply (3.6))} \\
 1 - \sigma(a) &= 1 - \frac{1}{1 + \exp\{-a\}} \\
 &= \frac{\exp\{-a\}}{1 + \exp\{-a\}} \\
 &= \frac{1}{1 + \exp\{a\}} \\
 (4.14) \quad 1 - \sigma(a) &= \sigma(-a) && \text{(Apply (3.6)).}
 \end{aligned}$$

We aim now to determine the inverse of (3.6). It follows that

$$\begin{aligned}
 \sigma(\sigma^{-1}(y)) &= \frac{1}{1 + \exp\{-\sigma^{-1}(y)\}} && \text{(Apply (3.6))} \\
 y &= \frac{1}{1 + \exp\{-\sigma^{-1}(y)\}} \\
 1 + \exp\{-\sigma^{-1}(y)\} &= \frac{1}{y} \\
 \exp\{-\sigma^{-1}(y)\} &= \frac{1-y}{y} \\
 -\sigma^{-1}(y) &= \log \left\{ \frac{1-y}{y} \right\} \\
 (4.15) \quad \sigma^{-1}(y) &= \log \left\{ \frac{y}{1-y} \right\}.
 \end{aligned}$$

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

$$\begin{aligned}
 a &= \log \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
 &= \log p(\mathbf{x}|\mathcal{C}_1) - \log p(\mathbf{x}|\mathcal{C}_2) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\
 &= -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| + \\
 &\quad - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \\
 &\quad + \frac{M}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma| + \\
 &\quad + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \quad (\text{Apply (2.43)}) \\
 &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1 \Sigma^{-1} \boldsymbol{\mu}_1 + \\
 &\quad + \frac{1}{2} \boldsymbol{\mu}_2 \Sigma^{-1} \boldsymbol{\mu}_2 + \log \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\
 (4.16) \quad a &= \mathbf{w}^\top \mathbf{x} + w_0 \quad (\text{Apply (4.66) and (4.67)}).
 \end{aligned}$$

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)

$$(4.17) \quad L(\boldsymbol{\pi}, \lambda) = \log p(\mathbf{T}|\boldsymbol{\pi}) + \lambda \cdot \left(\sum_{j=1}^K \pi_j - 1 \right).$$

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

$$\begin{aligned} \frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \pi_k} &= 0 \\ \frac{\sum_{n=1}^N t_{n,k}}{\pi_k} + \lambda &= 0 \\ (4.18) \quad \pi_k &= -\frac{N_k}{\lambda}. \end{aligned}$$

Substituting (4.18) into the constraint $\sum_{j=1}^K \pi_j = 1$, we find

$$\begin{aligned} \sum_{j=1}^K \pi_j - 1 &= 0 \\ - \sum_{j=1}^K \frac{N_j}{\lambda} &= 1 \\ (4.19) \quad \lambda &= -N. \end{aligned}$$

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

$$\begin{aligned}
 p(\mathbf{T}|\pi, \boldsymbol{\mu}, \Sigma) &= \prod_{n=1}^N \prod_{j=1}^K \left[\frac{\pi_j}{(2\pi)^{M/2} |\Sigma|^{1/2}} \times \right. \\
 &\quad \left. \times \exp \left\{ -\frac{1}{2} (\phi_n - \mu_j)^\top \Sigma^{-1} (\phi_n - \mu_j) \right\} \right]^{t_{n,j}} \quad (\text{Apply (2.43)}) \\
 (4.20) \quad \log p(\mathbf{T}|\pi, \boldsymbol{\mu}, \Sigma) &= \sum_{n=1}^N \sum_{j=1}^K t_{n,j} \left[\log \pi_j - \frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| \right. \\
 &\quad \left. - \frac{1}{2} (\phi_n - \mu_j)^\top \Sigma^{-1} (\phi_n - \mu_j) \right].
 \end{aligned}$$

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

$$\begin{aligned}
 \frac{\partial \log p(\mathbf{T}|\pi, \boldsymbol{\mu}, \Sigma)}{\partial \mu_k} &= \mathbf{0} \\
 \sum_{n=1}^N t_{n,k} \left[-\frac{1}{2} \frac{d}{d\mu_k} \left\{ \phi_n^\top \Sigma^{-1} \phi_n - 2\phi_n^\top \Sigma^{-1} \mu_k + \mu_k^\top \Sigma^{-1} \mu_k \right\} \right] &= \mathbf{0} \\
 -2 \sum_{n=1}^N t_{n,k} \Sigma^{-1} \phi_n + 2 \sum_{n=1}^N t_{n,k} \Sigma^{-1} \mu_k &= \mathbf{0} \quad (\text{Apply (C.19)}) \\
 N_k \mu_k &= \sum_{n=1}^N t_{n,k} \phi_n \\
 \mu_k &= \frac{1}{N_k} \sum_{n=1}^N t_{n,k} \phi_n.
 \end{aligned}$$

Hence, we conclude that the maximum likelihood estimator for μ_k , for arbitrary k , is

$$(4.21) \quad \mu_k^{\text{ML}} = \frac{1}{N_k} \sum_{n=1}^N t_{n,k} \phi_n.$$

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

$$\begin{aligned}
 \frac{\partial \log p(\mathbf{T}|\pi, \boldsymbol{\mu}, \Sigma)}{\partial \Sigma} &= \mathbf{0} \\
 \sum_{j=1}^K \sum_{n=1}^N t_{n,j} \left[\Sigma^{-1} + \frac{d}{d\Sigma} \left\{ (\phi_n - \mu_j)^\top \Sigma^{-1} (\phi_n - \mu_j) \right\} \right] &= \mathbf{0} \quad (\text{Apply (C.28)}) \\
 N \Sigma^{-1} - \sum_{j=1}^K \sum_{n=1}^N t_{n,j} (\phi_n - \mu_j) (\phi_n - \mu_j)^\top \Sigma^{-2} &= \mathbf{0} \quad (\text{Apply (C.24)}) \\
 (4.22) \quad \frac{1}{N} \sum_{j=1}^K \sum_{n=1}^N t_{n,j} (\phi_n - \mu_j) (\phi_n - \mu_j)^\top &= \Sigma.
 \end{aligned}$$

Note that the form in (4.22) is dependent on μ_k , however, the maximum likelihood estimator of μ_k , determined in (4.21), does not depend on Σ . Therefore, we may plug our maximum likelihood estimator (4.21) onto (4.22). This yields the following maximum likelihood estimator for Σ

$$\begin{aligned}\Sigma_{\text{ML}} &= \frac{1}{N} \sum_{j=1}^K \sum_{n=1}^N t_{n,j} (\phi_n - \boldsymbol{\mu}_j^{\text{ML}})(\phi_n - \boldsymbol{\mu}_j^{\text{ML}})^{\top} \\ &= \sum_{j=1}^K \frac{N_j}{N} \frac{1}{N_j} \sum_{n=1}^N t_{n,j} (\phi_n - \boldsymbol{\mu}_j^{\text{ML}})(\phi_n - \boldsymbol{\mu}_j^{\text{ML}})^{\top} \\ \Sigma_{\text{ML}} &= \sum_{j=1}^K \frac{N_j}{N} \mathbf{S}_j.\end{aligned}$$

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

$$\begin{aligned} a_k &= \log\{p(\phi | \mathcal{C}_k)p(\mathcal{C}_k)\} \\ &= \log p(\phi | \mathcal{C}_k) + \log p(\mathcal{C}_k) \\ &= \log \left[\prod_{n=1}^N \prod_{i=1}^M \prod_{j=1}^L \pi_{n,i,j,k}^{\phi_{n,i,j}} \right] + \log p(\mathcal{C}_k) \\ (4.23) \quad &= \sum_{n=1}^N \sum_{i=1}^M \sum_{j=1}^L \phi_{n,i,j} \log \pi_{n,i,j,k} + \log p(\mathcal{C}_k). \end{aligned}$$

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

$$\begin{aligned}
 \sigma(a) &= \frac{1}{1 + \exp\{-a\}} && \text{(Apply (3.6))} \\
 \frac{d\sigma(a)}{da} &= \frac{\exp\{-a\}}{(1 + \exp\{-a\})^2} \\
 &= \frac{1 + \exp\{-a\} - 1}{(1 + \exp\{-a\})^2} \\
 &= \frac{1}{1 + \exp\{-a\}} - \frac{1}{(1 + \exp\{-a\})^2} \\
 &= \frac{1}{1 + \exp\{-a\}} \left(1 - \frac{1}{1 + \exp\{-a\}}\right) \\
 \frac{d\sigma(a)}{da} &= \sigma(a)\{1 - \sigma(a)\} && \text{(Apply (3.6)).}
 \end{aligned}$$

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

$$\begin{aligned}
 E(\mathbf{w}) &= -\sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} \\
 &= -\sum_{n=1}^N \{t_n \log[y_n/(1 - y_n)] + \log(1 - y_n)\} \\
 &= -\sum_{n=1}^N \{t_n \sigma^{-1}(y_n) + \log(1 - y_n)\} \quad (\text{Apply (4.15)}) \\
 &= -\sum_{n=1}^N \{t_n \sigma^{-1}(\sigma(-\mathbf{w}^\top \phi)) + \log(1 - \sigma(\mathbf{w}^\top \phi))\} \quad (\text{Apply } y_n = \sigma(\mathbf{w}^\top \phi)) \\
 (4.24) \quad E(\mathbf{w}) &= -\sum_{n=1}^N \{t_n \mathbf{w}^\top \phi_n + \log(\sigma(-\mathbf{w}^\top \phi_n))\} \quad (\text{Apply (4.14)}).
 \end{aligned}$$

Thereafter, differentiating (4.24) with respect to \mathbf{w} we obtain

$$\begin{aligned}
 \nabla E(\mathbf{w}) &= -\sum_{n=1}^N \left\{ t_n \phi_n - \frac{\sigma(-\mathbf{w}^\top \phi_n) \{1 - \sigma(-\mathbf{w}^\top \phi_n)\}}{\sigma(-\mathbf{w}^\top \phi_n)} \phi_n \right\} \quad (\text{Apply (4.88) and (C.19)}) \\
 &= \sum_{n=1}^N \{\sigma(\mathbf{w}^\top \phi_n) - t_n\} \phi_n \quad (\text{Apply (4.14)}) \\
 \nabla E(\mathbf{w}) &= \sum_{n=1}^N \{y_n - t_n\} \phi_n \quad (\text{Apply } y_n = \sigma(\mathbf{w}^\top \phi)).
 \end{aligned}$$

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 $\{\phi_n\}_{n=1}^N$ 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

$$\begin{aligned}
 p(\mathbf{t}|\mathbf{w}) &= \prod_{n=1}^N \left(\frac{\sigma(\mathbf{w}^\top \phi_n)}{1 - \sigma(\mathbf{w}^\top \phi_n)} \right)^{t_n} \{1 - \sigma(\mathbf{w}^\top \phi_n)\} \\
 &= \left[\prod_{n \in \mathcal{C}_1} \sigma(\mathbf{w}^\top \mathbf{x}_n) \right] \left[\prod_{n \in \mathcal{C}_2} \{1 - \sigma(\mathbf{w}^\top \mathbf{y}_n)\} \right] \\
 &= \left[\prod_{n \in \mathcal{C}_1} \sigma(\mathbf{w}^\top \mathbf{x}_n) \right] \left[\prod_{n \in \mathcal{C}_2} \sigma(-\mathbf{w}^\top \mathbf{y}_n) \right] \quad (\text{Apply (4.14)}) \\
 p(\mathbf{t}|\mathbf{w}) &= \left[\prod_{n \in \mathcal{C}_1} \sigma(\|\mathbf{w}\| \|\mathbf{x}_n\| \cos \alpha_n) \right] \times \\
 &\quad \times \left[\prod_{n \in \mathcal{C}_2} \sigma(-\|\mathbf{w}\| \|\mathbf{y}_n\| \cos \beta_n) \right] \\
 (4.25) \quad \log p(\mathbf{t}|\mathbf{w}) &= \sum_{n \in \mathcal{C}_1} \log \sigma(\|\mathbf{w}\| \|\mathbf{x}_n\| \cos \alpha_n) + \\
 &\quad + \sum_{n \in \mathcal{C}_2} \log \sigma(-\|\mathbf{w}\| \|\mathbf{y}_n\| \cos \beta_n).
 \end{aligned}$$

Let \mathbf{w}_0 be an arbitrary vector which linearly separates our data set, it follows that

$$\begin{aligned}
 \mathbf{w}_0^\top \mathbf{x}_m &> 0 \\
 \|\mathbf{w}_0\| \|\mathbf{x}_m\| \cos \alpha_m &> 0 \\
 (4.26) \quad \|\mathbf{x}_m\| \cos \alpha_m &> 0,
 \end{aligned}$$

where α_m denotes the angle between \mathbf{w}_0 and \mathbf{x}_m . Likewise

$$\begin{aligned}
 \mathbf{w}_0^\top \mathbf{y}_k &< 0 \\
 \|\mathbf{w}_0\| \|\mathbf{y}_k\| \cos \beta_k &< 0 \\
 (4.27) \quad -\|\mathbf{y}_k\| \cos \beta_k &> 0,
 \end{aligned}$$

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

$$\begin{aligned}
 \frac{\partial \log p(\mathbf{t}|\mathbf{w})}{\partial \|\mathbf{w}\|} &= \sum_{n \in \mathcal{C}_1} \|\mathbf{x}_n\| \cos \alpha_n \{1 - \sigma(\|\mathbf{w}\| \|\mathbf{x}_n\| \cos \alpha_n)\} + \\
 &\quad - \sum_{n \in \mathcal{C}_2} \|\mathbf{y}_n\| \cos \beta_n \{1 - \sigma(-\|\mathbf{w}\| \|\mathbf{y}_n\| \cos \beta_n)\} \quad (\text{Apply (4.88)}) \\
 &= \sum_{n \in \mathcal{C}_1} \|\mathbf{x}_n\| \cos \alpha_n \sigma(-\|\mathbf{w}\| \|\mathbf{x}_n\| \cos \alpha_n) + \\
 &\quad - \sum_{n \in \mathcal{C}_2} \|\mathbf{y}_n\| \cos \beta_n \sigma(\|\mathbf{w}\| \|\mathbf{y}_n\| \cos \beta_n) \quad (\text{Apply (4.14)}) \\
 \frac{\partial \log p(\mathbf{t}|\mathbf{w})}{\partial \|\mathbf{w}\|} &> 0.
 \end{aligned}$$

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 α and β which ensures linear separability. We conclude that, in order to maximize (4.25) with respect to $\|\mathbf{w}\|$, we must take $\|\mathbf{w}\| \rightarrow \infty$ (given a fixed set of angles α and β 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

$$\begin{aligned}\mathbf{a}^\top \mathbf{H} \mathbf{a} &= \mathbf{a}^\top \left[\sum_{n=1}^N y_n(1-y_n) \phi_n \phi_n^\top \right] \mathbf{a} && \text{(Apply (4.97))} \\ &= \sum_{n=1}^N y_n(1-y_n) \mathbf{a}^\top \phi_n \phi_n^\top \mathbf{a} \\ &= \sum_{n=1}^N y_n(1-y_n) \{\mathbf{a}^\top \phi_n\}^2 \\ \mathbf{a}^\top \mathbf{H} \mathbf{a} &> 0 && \text{(Apply } y_n(1-y_n) \in (0, 1)\text{)}.\end{aligned}$$

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

$$\begin{aligned} \sum_{n=1}^N \log p(s_n|\phi) &= \sum_{n=1}^N s_n \log y_n + \sum_{n=1}^N (1-s_n) \log(1-y_n) + \\ &\quad - \sum_{n=1}^N \log[y_n^{\pi_n}(1-y_n)^{1-\pi_n} + y_n^{1-\pi_n}(1-y_n)^{\pi_n}]. \end{aligned}$$

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

$$\begin{aligned}
 \frac{\partial p(\mathcal{C}_k|\phi)}{\partial a_j} &= \frac{\partial}{\partial a_j} \left[\frac{\exp\{a_k\}}{\sum_{i=1}^K \exp\{a_i\}} \right] \\
 &= \frac{\partial}{\partial a_j} \left[\frac{\exp\{a_k\}}{\exp\{a_j\} + \sum_{\substack{i=1 \\ i \neq j}}^K \exp\{a_i\}} \right] \\
 &= \begin{cases} \frac{\partial}{\partial a_j} \left[1 - \frac{\sum_{\substack{i=1 \\ i \neq j}}^K \exp\{a_i\}}{\exp\{a_j\} + \sum_{\substack{i=1 \\ i \neq j}}^K \exp\{a_i\}} \right] & \text{if } k = j, \\ \frac{\partial}{\partial a_j} \left[\frac{\exp\{a_k\}}{\exp\{a_j\} + \sum_{\substack{i=1 \\ i \neq j}}^K \exp\{a_i\}} \right] & \text{if } k \neq j. \end{cases} \\
 &= \begin{cases} \frac{[\sum_{i=1}^K \exp\{a_i\} - \exp\{a_j\}] \exp\{a_j\}}{(\sum_{i=1}^K \exp\{a_i\})^2} & \text{if } k = j, \\ -\frac{\exp\{a_k\} \exp\{a_j\}}{(\sum_{i=1}^K \exp\{a_i\})^2} & \text{if } k \neq j. \end{cases} \\
 &= \begin{cases} y_j(1 - y_j) & \text{if } k = j, \\ -y_j y_k & \text{if } k \neq j. \end{cases} \quad (\text{Apply (4.104)})
 \end{aligned}$$

$$\frac{\partial p(\mathcal{C}_k|\phi)}{\partial a_j} = y_k(I_{j,k} - y_j),$$

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

$$(4.28) \quad \begin{aligned} \frac{da_k}{d\mathbf{w}_j} &= \frac{d}{d\mathbf{w}_j} [\mathbf{w}_k^\top \phi] \\ \frac{da_k}{d\mathbf{w}_j} &= \phi I_{j,k} \end{aligned} \quad (\text{Apply (C.19)}),$$

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

$$\begin{aligned} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) &= \nabla_{\mathbf{w}_j} \left[- \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \log y_{n,k} \right] && (\text{Apply (4.108)}) \\ &= - \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \frac{1}{y_{n,k}} \frac{\partial y_{n,k}}{\partial a_{n,j}} \frac{da_j}{d\mathbf{w}_j} \\ &= - \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \frac{1}{y_{n,k}} y_{n,k} (I_{j,k} - y_{n,j}) \phi_n I_{j,j} && (\text{Apply (4.106) and (4.28)}) \\ &= - \sum_{n=1}^N t_{n,j} \phi_n + \sum_{n=1}^N \sum_{k=1}^K t_{n,k} y_{n,j} \phi_n \\ \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) &= \sum_{n=1}^N \{y_{n,j} - t_{n,j}\} \phi_j && (\text{Apply } \sum_{k=1}^K t_{n,k} = 1). \end{aligned}$$

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

$$(4.29) \quad p(\mathbf{T}|\mathbf{w}) = \prod_{n=1}^N \{\Phi(a_n)\}^{t_n} \{1 - \Phi(a_n)\}^{1-t_n}$$

$$\log p(\mathbf{T}|\mathbf{w}) = \sum_{n=1}^N t_n \log \Phi(a_n) + \sum_{n=1}^N (1-t_n) \log \{1 - \Phi(a_n)\}$$

where $\Phi(a)$ is as defined in (4.114), and $a_n = \mathbf{w}^\top \boldsymbol{\phi}_n$. First, we consider the derivative of (4.114) with respect to a . It follows that

$$\frac{d\Phi(a)}{da} = \frac{d}{da} \left[\int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{t^2}{2} \right\} dt \right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{a^2}{2} \right\}$$

$$\frac{d\Phi(a)}{da} = \psi(a),$$

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

$$\frac{d^2\Phi(a)}{da^2} = \psi'(a).$$

Differentiating (4.29) with respect to \mathbf{w} , we obtain

$$(4.30) \quad \begin{aligned} \frac{d \log p(\mathbf{T}|\mathbf{w})}{d\mathbf{w}} &= \sum_{n=1}^N t_n \frac{\psi(a_n)}{\Phi(a_n)} \boldsymbol{\phi}_n - \sum_{n=1}^N (1-t_n) \frac{\psi(a_n)}{1-\Phi(a_n)} \boldsymbol{\phi}_n \\ &= \sum_{n=1}^N \frac{t_n \{1 - \Phi(a_n)\} - (1-t_n)\Phi(a_n)}{\Phi(a_n)\{1 - \Phi(a_n)\}} \psi(a_n) \boldsymbol{\phi}_n \\ \frac{d \log p(\mathbf{T}|\mathbf{w})}{d\mathbf{w}} &= \sum_{n=1}^N \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1 - \Phi(a_n)\}} \psi(a_n) \boldsymbol{\phi}_n. \end{aligned}$$

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

$$\begin{aligned}
 \frac{d^2 \log p(\mathbf{T}|\mathbf{w})}{d\mathbf{w}d\mathbf{w}^\top} &= \frac{d}{d\mathbf{w}} \left[\sum_{n=1}^N \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1 - \Phi(a_n)\}} \psi(a_n) \phi_n^\top \right] \\
 &= \frac{d}{d\mathbf{w}} \left[\sum_{n=1}^N \left\{ \frac{t_n}{\Phi(a_n)\{1 - \Phi(a_n)\}} - \frac{1}{1 - \Phi(a_n)} \right\} \psi(a_n) \phi_n^\top \right] \\
 &= \sum_{n=1}^N \left[-\frac{t_n \psi(a_n)\{1 - \Phi(a_n)\} - t_n \Phi(a_n) \psi(a_n)}{[\Phi(a_n)\{1 - \Phi(a_n)\}]^2} \phi_n + \right. \\
 &\quad \left. - \frac{\psi(a_n)}{[1 - \Phi(a_n)]^2} \phi_n \right] \psi(a_n) \phi_n^\top + \\
 &\quad + \sum_{n=1}^N \left\{ \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1 - \Phi(a_n)\}} \right\} \psi'(a_n) \phi_n \phi_n^\top \\
 &= \sum_{n=1}^N \left[\frac{2t_n \Phi(a_n) - t_n - \{\Phi(a_n)\}^2}{[\Phi(a_n)\{1 - \Phi(a_n)\}]^2} \right] \{\psi(a_n)\}^2 \phi_n \phi_n^\top + \\
 &\quad + \sum_{n=1}^N \left\{ \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1 - \Phi(a_n)\}} \right\} \psi'(a_n) \phi_n \phi_n^\top \\
 (4.31) \quad \frac{d^2 \log p(\mathbf{T}|\mathbf{w})}{d\mathbf{w}d\mathbf{w}^\top} &= - \sum_{n=1}^N \left[\frac{[t_n - \Phi(a_n)]^2}{[\Phi(a_n)\{1 - \Phi(a_n)\}]^2} \right] \frac{1}{2\pi} \exp\{-a_n^2\} \phi_n \phi_n^\top + \\
 &\quad - \sum_{n=1}^N a_n \left\{ \frac{t_n - \Phi(a_n)}{\Phi(a_n)\{1 - \Phi(a_n)\}} \right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{a_n^2}{2}\right\} \phi_n \phi_n^\top.
 \end{aligned}$$

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

$$\begin{aligned} \mathbf{u}^\top \mathbf{H} \mathbf{u} &= \sum_{j=1}^K \sum_{k=1}^K \mathbf{u}_j^\top \nabla_{\mathbf{w}_j} \nabla_{\mathbf{w}_k}^\top E(\tilde{\mathbf{w}}) \mathbf{u}_k \\ &= \sum_{j=1}^K \sum_{k=1}^K \mathbf{u}_j^\top \left[\sum_{n=1}^N y_{n,k} (I_{j,k} - y_{n,j}) \phi_n \phi_n^\top \right] \mathbf{u}_k \quad (\text{Apply (4.110)}) \\ &= \sum_{n=1}^N \left[\sum_{j=1}^K \sum_{k=1}^K y_{n,k} (I_{j,k} - y_{n,j}) \{ \mathbf{u}_j^\top \phi_n \} \{ \mathbf{u}_k^\top \phi_n \} \right] \\ &= \sum_{n=1}^N \sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \}^2 + \\ &\quad - \sum_{n=1}^N \sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \} \sum_{j=1}^K y_{n,j} \{ \mathbf{u}_j^\top \phi_n \} \\ (4.32) \quad \mathbf{u}^\top \mathbf{H} \mathbf{u} &= \sum_{n=1}^N \sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \}^2 - \sum_{n=1}^N \left[\sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \} \right]^2. \end{aligned}$$

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

$$\begin{aligned} \left[\sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \} \right]^2 &\leq \sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \}^2 \quad \forall n \in \{1, \dots, N\} \\ (4.33) \quad 0 &\leq \sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \}^2 - \left[\sum_{k=1}^K y_{n,k} \{ \mathbf{u}_k^\top \phi_n \} \right]^2 \quad \forall n \in \{1, \dots, N\}. \end{aligned}$$

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

$$\begin{aligned}
 \Phi(a) &= \int_{-\infty}^a \phi(t) dt \\
 &= \int_{-\infty}^0 \phi(t) dt + \int_0^a \phi(t) dt \\
 &= \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt + \\
 &\quad + \int_0^a \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt \\
 &= \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt + \\
 &\quad + \int_0^{a/\sqrt{2}} \frac{1}{\sqrt{\pi}} \exp\{-s^2\} ds \quad (\text{Symmetry of } \phi(t) \text{ and set } s = t/\sqrt{2}) \\
 &= \frac{1}{2} \frac{1}{\sqrt{2\pi}} I + \\
 &\quad + \frac{1}{2} \frac{2}{\pi} \int_0^{a/\sqrt{2}} \frac{1}{\sqrt{\pi}} \exp\{-s^2\} ds \quad (\text{Apply (1.124)}) \\
 \Phi(a) &= \frac{1}{2} \left\{ 1 + \operatorname{erf}\left\{\frac{a}{\sqrt{2}}\right\} \right\} \quad (\text{Apply (4.115)}).
 \end{aligned}$$

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 $p(\mathcal{D}) = \int p(\mathcal{D}, \theta) d\theta$, where the parameters θ belong to an M -dimensional space. Consider moreover that we have determined the maximum density point of $p(\theta|\mathcal{D})$, denoted as θ_{MAP} , which is consequently also the maximum point of $p(\mathcal{D}|\theta)p(\theta)$ (see (1.32)). We thereby adopt the Laplace approximation in (4.135) for $p(\mathcal{D})$, which yields

$$p(\mathcal{D}) \approx p(\mathcal{D}|\theta_{\text{MAP}})p(\theta_{\text{MAP}}) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}$$

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\theta_{\text{MAP}}) + \log p(\theta_{\text{MAP}}) + \frac{M}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{A}|,$$

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 θ 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

$$(4.34) \quad \begin{aligned} \log p(\mathcal{D}) &\approx \log p(\mathcal{D}|\theta_{\text{MAP}}) + \log p(\theta_{\text{MAP}}) + \frac{M}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{A}| \\ &\approx \log p(\mathcal{D}|\theta_{\text{MAP}}) - \frac{M}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{V}_0| + \\ &\quad - \frac{1}{2} (\theta_{\text{MAP}} - \mathbf{m})^\top \mathbf{V}_0^{-1} (\theta_{\text{MAP}} - \mathbf{m}) + \frac{M}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{A}| \quad (\text{Apply (2.118)}). \end{aligned}$$

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

$$(4.35) \quad \begin{aligned} \mathbf{A} &= -\nabla \nabla^\top \log \{p(\mathcal{D}|\theta_{\text{MAP}})p(\theta_{\text{MAP}})\} \\ &= -\nabla \nabla^\top \log p(\mathcal{D}|\theta_{\text{MAP}}) - \nabla \nabla^\top \log p(\theta_{\text{MAP}}) \\ &= -\sum_{n=1}^N \nabla \nabla^\top \log p(d_n|\theta_{\text{MAP}}) + \\ &\quad - \nabla \nabla^\top \left[\frac{M}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{V}_0| + \right. \\ &\quad \left. - \frac{1}{2} (\theta_{\text{MAP}} - \mathbf{m})^\top \mathbf{V}_0^{-1} (\theta_{\text{MAP}} - \mathbf{m}) \right] \quad (\text{Apply (2.118)}) \\ &= -\sum_{n=1}^N \nabla \nabla^\top \log p(d_n|\theta_{\text{MAP}}) + \mathbf{V}_0^{-1} \quad (\text{Apply (C.19)}) \\ &\quad \mathbf{A} \approx \mathbf{H} \quad (\mathbf{V}_0^{-1} \text{ small}). \end{aligned}$$

We denote $\mathbf{H} = -\sum_{n=1}^N \nabla \nabla^\top \log p(d_n|\theta_{\text{MAP}})$. We return therefore to [\(4.34\)](#)

$$\begin{aligned} \log p(\mathcal{D}) &\approx \log p(\mathcal{D}|\theta_{\text{MAP}}) - \frac{1}{2} \log |\mathbf{H}| - \frac{1}{2} \log |\mathbf{V}_0| \quad (\text{Apply (4.35) and } \mathbf{V}_0^{-1} \text{ small}) \\ &= \log p(\mathcal{D}|\theta_{\text{MAP}}) - \frac{1}{2} \log |N\mathbf{H}/N| - \frac{1}{2} \log |\mathbf{V}_0| \\ &= \log p(\mathcal{D}|\theta_{\text{MAP}}) - \frac{M}{2} \log |N| + \\ &\quad - \frac{1}{2} \log \left| \frac{\mathbf{H}}{N} \right| - \frac{1}{2} \log |\mathbf{V}_0|. \end{aligned}$$

Note that \mathbf{H}/N constitutes an approximation of $\mathbb{E}_{\mathcal{D}}[-\nabla \nabla^\top \log p(\mathcal{D}|\theta_{\text{MAP}})]$ in the sense of [\(1.35\)](#). Hence, under mild conditions, $\log |\mathbf{H}/N|$ is bounded, and we have that

$$\log p(\mathcal{D}) \approx \log p(\mathcal{D}|\theta_{\text{MAP}}) - \frac{M}{2} \log |N|.$$

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

$$\text{Var}[\mathbf{w}] = \mathbf{S}_N \quad (\text{Apply (2.64) and (4.144)}),$$

where \mathbf{S}_N is as in (4.143). Moreover

$$\begin{aligned} \text{Cov}[\mathbf{w}, \mathbf{w}^\top \phi] &= \text{Cov}[\mathbf{w}, \phi^\top \mathbf{w}] \\ &= \phi^\top \text{Var}[\mathbf{w}] \\ \text{Cov}[\mathbf{w}, \mathbf{w}^\top \phi] &= \phi^\top \mathbf{S}_n \quad (\text{Apply (2.64) and (4.144)}) \\ \text{Cov}[\mathbf{w}^\top \phi, \mathbf{w}^\top] &= \mathbf{S}_n \phi \quad (\text{Apply (2.64) and (4.144)}) \end{aligned}$$

and

$$\text{Var}[\mathbf{w}^\top \phi] = \phi^\top \mathbf{S}_n \phi \quad (\text{Apply (2.64) and (4.144)}).$$

Therefore, we conclude that

$$\mathbb{E}\left[\begin{pmatrix} \mathbf{w} \\ \mathbf{w}^\top \phi \end{pmatrix}\right] = \begin{pmatrix} \mathbf{w}_{\text{MAP}} \\ \mathbf{w}_{\text{MAP}}^\top \phi \end{pmatrix} \quad \text{and} \quad \text{Var}\left[\begin{pmatrix} \mathbf{w} \\ \mathbf{w}^\top \phi \end{pmatrix}\right] = \begin{pmatrix} \mathbf{S}_N & \mathbf{S}_N \phi \\ \phi^\top \mathbf{S}_N & \phi^\top \mathbf{S}_N \phi \end{pmatrix}$$

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 $\text{Var}[a] = \phi^\top \mathbf{S}_N \phi$. We thereby return to (4.151), and find that

$$\begin{aligned} p(\mathcal{C}_1 | t) &\approx \int \sigma(a) p(a) da \\ &= \int \sigma(a) \phi(a | \mathbf{w}_{\text{MAP}}^\top \phi, \phi^\top \mathbf{S}_N \phi) da, \end{aligned}$$

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

$$\begin{aligned}
 \frac{d\sigma(a)}{da} \Big|_{a=0} &= \sigma(a)\{1 - \sigma(a)\} \Big|_{a=0} && \text{(Apply (4.88))} \\
 &= \frac{1}{1 + e^{-a}} \left(1 - \frac{1}{1 + e^{-a}}\right) && \text{(Apply (3.6))} \\
 (4.36) \quad \frac{d\sigma(a)}{da} \Big|_{a=0} &= \frac{1}{4}.
 \end{aligned}$$

The derivative of the scaled probit function at the origin is

$$\begin{aligned}
 \frac{d\Phi(\lambda a)}{da} \Big|_{a=0} &= \frac{d}{da} \left[\int_{-\infty}^{\lambda a} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}t^2 \right\} dt \right] \Big|_{a=0} && \text{(Apply (4.114))} \\
 &= \frac{d}{da} \left[\int_{-\infty}^a \frac{\lambda}{\sqrt{2\pi}} \exp \left\{ -\frac{\lambda^2}{2}s^2 \right\} ds \right] \Big|_{a=0} && \text{(Set } s = t/\lambda \text{)} \\
 &= \frac{\lambda}{\sqrt{2\pi}} \exp \left\{ -\frac{\lambda^2}{2}0 \right\} \\
 (4.37) \quad \frac{d\Phi(\lambda a)}{da} \Big|_{a=0} &= \frac{\lambda}{\sqrt{2\pi}}.
 \end{aligned}$$

Equating (4.36) and (4.37), we find

$$\begin{aligned}
 \frac{d\sigma(a)}{da} \Big|_{a=0} &= \frac{d\Phi(\lambda a)}{da} \Big|_{a=0} \\
 \frac{1}{4} &= \frac{\lambda}{\sqrt{2\pi}} \\
 \lambda &= \frac{\sqrt{2\pi}}{4} \\
 \lambda^2 &= \frac{\pi}{8}.
 \end{aligned}$$

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

$$\begin{aligned}
 \frac{d\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}\right)}{d\mu} &= \frac{d}{d\mu} \left[\int_{-\infty}^{\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}t^2\right\} dt \right] && \text{(Apply (4.114))} \\
 &= \frac{d}{d\mu} \left[\int_{-\infty}^{\mu} \frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}\sqrt{2\pi}} \times \right. \\
 &\quad \times \exp\left\{-\frac{1}{2}\frac{s^2}{\lambda^{-2}+\sigma^2}\right\} ds \left. \right] && \text{(Set } s = \sqrt{\lambda^{-2}+\sigma^2}t) \\
 (4.38) \quad \frac{d\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}}\right)}{d\mu} &= \frac{1}{(\lambda^{-2}+\sigma^2)^{1/2}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\lambda^{-2}+\sigma^2}\right\}.
 \end{aligned}$$

For the left-hand-side, we have that

$$\begin{aligned}
 \frac{d}{d\mu} \left[\int \Phi(\lambda a)\phi(a|\mu, \sigma^2) da \right] &= \frac{d}{d\mu} \left[\int \sigma\Phi(\lambda\{\mu + \sigma z\}) \times \right. \\
 &\quad \times \phi(\mu + \sigma z|\mu, \sigma^2) dz \left. \right] && \text{(Set } a = \mu + \sigma z) \\
 &= \frac{d}{d\mu} \left[\int \sigma \left(\int_{-\infty}^{\lambda(\mu+\sigma z)} \frac{1}{\sqrt{2\pi}} \times \right. \right. \\
 &\quad \times \exp\left\{-\frac{1}{2}t^2\right\} dt \left. \right) \times \\
 &\quad \times \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{1}{2}z^2\right\} dz \left. \right] && \text{(Apply (1.46) and (4.114))} \\
 &= \frac{d}{d\mu} \left[\left(\int_{-\infty}^{\mu} \frac{\lambda}{\sqrt{2\pi}} \times \right. \right. \\
 &\quad \times \exp\left\{-\frac{\lambda^2}{2}(s+\sigma z)^2\right\} ds \left. \right) \times \\
 &\quad \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} dz \left. \right] && \text{(Set } s = \frac{t-\lambda\sigma z}{\lambda}) \\
 &= \int \frac{\lambda}{\sqrt{2\pi}} \exp\left\{-\frac{\lambda^2\sigma^2}{2}\left(\frac{\mu}{\sigma}+z\right)^2\right\} \times \\
 &\quad \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} dz.
 \end{aligned}$$

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{d}{d\mu} \left[\int \Phi(\lambda a) \phi(a|\mu, \sigma^2) da \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\} dz \\
 &= \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\} dz \\
 &= \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. + \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. \right. \\
 &\quad \left. \left. - \frac{\mu^2}{\sigma^2} \frac{\lambda^2 \sigma^2}{\lambda^2 \sigma^2 + 1} \right) \right\} dz \\
 &= \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 \left. + \frac{\lambda^2 \sigma^2}{2} \frac{\mu^2}{\sigma^2} \frac{\lambda^2 \sigma^2}{\lambda^2 \sigma^2 + 1} + \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\} dz \\
 &= \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\} dz \\
 &= \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\} dz \\
 (4.39) \quad \frac{d}{d\mu} \left[\int \Phi(\lambda a) \phi(a|\mu, \sigma^2) da \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\} \quad (\text{Apply (1.48)}).
 \end{aligned}$$

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

$$\begin{aligned}
 (4.40) \quad & \int \left[\frac{d\Phi\left(\frac{\mu}{\sqrt{\lambda^{-2} + \sigma^2}}\right)}{d\mu} \right] d\mu + C = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) \\
 & \int \frac{1}{\sqrt{2\pi}(\sigma^2 + \lambda^{-2})^{1/2}} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma^2 + \lambda^{-2}}\right\} d\mu + C = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) \\
 & \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) + C = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) \\
 & C = 0.
 \end{aligned}$$

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

$$\begin{aligned}
 (4.41) \quad & \int \left[\frac{d}{d\mu} \left\{ \int \Phi(\lambda a) \phi(a|\mu, \sigma^2) da \right\} \right] d\mu + L = \int \Phi(\lambda a) \phi(a|\mu, \sigma^2) da \\
 & \int \frac{1}{\sqrt{2\pi}(\sigma^2 + \lambda^{-2})^{1/2}} \exp\left\{-\frac{1}{2}\frac{\mu^2}{\sigma^2 + \lambda^{-2}}\right\} d\mu + L = \int \Phi(\lambda a) \phi(a|\mu, \sigma^2) da \\
 & \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) + L = \int \Phi(\lambda a) \phi(a|\mu, \sigma^2) da.
 \end{aligned}$$

By taking the limit $\mu \rightarrow -\infty$ 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

$$(5.1) \quad u_{k,j}^{(2)} = \frac{w_{k,j}^{(2)}}{2}, \quad u_{k,0}^{(2)} = \sum_{j=1}^M \frac{w_{k,j}^{(2)}}{2} + w_{k,0}^{(2)}, \quad u_{j,i}^{(1)} = \frac{w_{j,i}^{(1)}}{2} \quad \text{and} \quad u_{j,0}^{(1)} = \frac{w_{j,0}^{(1)}}{2}.$$

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 $\beta \mathbf{I} \in \mathbb{R}^{D \times D}$, 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

$$\begin{aligned}
 \log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1} \mathbf{I}) &= -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\beta^{-1} \mathbf{I}| + \\
 &\quad - \frac{\beta}{2} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))^\top (\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w})) \quad (\text{Apply (2.118)}) \\
 &\propto -\frac{1}{2} \|\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w})\|^2 \\
 &= -\frac{1}{2} \|\mathbf{y}(\mathbf{x}, \mathbf{w}) - \mathbf{t}_n\|^2 \\
 (5.2) \quad \log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1} \mathbf{I}) &\propto -E(\mathbf{w}) \quad (\text{Apply (5.11)}).
 \end{aligned}$$

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

$$\begin{aligned}
 (5.3) \quad \log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \Sigma) &= -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma| + \\
 &\quad -\frac{1}{2}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))^\top \Sigma^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w})) \quad (\text{Apply (2.118)}) \\
 &\propto -\frac{1}{2}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))^\top \Sigma^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w})) \\
 \log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \Sigma) &\propto -\frac{1}{2}\text{tr}[(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))^\top \Sigma^{-1}] \\
 \log p(\mathbf{T}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \Sigma) &\propto -E_\Sigma(\mathbf{w})
 \end{aligned}$$

where

$$(5.4) \quad E_\Sigma(\mathbf{w}) = \frac{1}{2}\text{tr}[(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))(\mathbf{t}_n - \mathbf{y}(\mathbf{x}, \mathbf{w}))^\top \Sigma^{-1}].$$

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

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

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

$$\begin{aligned} p(\mathbf{s}|\mathbf{w}, \epsilon) &= \{p(s_n = 1|\mathbf{x}, \mathbf{w}, \epsilon)\}^{s_n} \{p(s_n = 0|\mathbf{x}, \mathbf{w}, \epsilon)\}^{1-s_n} \\ &= \prod_{n=1}^N \{\epsilon + (1 - 2\epsilon)y(\mathbf{x}_n, \mathbf{w})\}^{s_n} \{1 - \epsilon - (1 - 2\epsilon)y(\mathbf{x}_n, \mathbf{w})\}^{1-s_n} \\ (5.6) \quad -\log p(\mathbf{s}|\mathbf{w}, \epsilon) &= -\sum_{n=1}^N s_n \log\{\epsilon + (1 - 2\epsilon)y(\mathbf{x}_n, \mathbf{w})\} + \\ &\quad -\sum_{n=1}^N (1 - s_n) \log\{1 - \epsilon - (1 - 2\epsilon)y(\mathbf{x}_n, \mathbf{w})\}. \end{aligned}$$

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

$$\begin{aligned}
 p(\mathbf{T}|\mathbf{X}, \mathbf{w}) &= \prod_{n=1}^N \prod_{k=1}^K \{y_k(\mathbf{x}_n, \mathbf{w})\}^{t_{n,k}} \{1 - y_k(\mathbf{x}_n, \mathbf{w})\}^{1-t_{n,k}} \\
 \log p(\mathbf{T}|\mathbf{X}, \mathbf{w}) &= \log \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \log\{y_k(\mathbf{x}_n, \mathbf{w})\} + \\
 &\quad + \sum_{n=1}^N \sum_{k=1}^K (1 - t_{n,k}) \log\{1 - y_k(\mathbf{x}_n, \mathbf{w})\}^{1-t_{n,k}} \\
 (5.7) \quad \log p(\mathbf{T}|\mathbf{X}, \mathbf{w}) &= -E(\mathbf{w}) \quad (\text{Apply (5.23)}).
 \end{aligned}$$

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

$$\begin{aligned}
 E(\mathbf{w}) &= -\sum_{n=1}^N t_n \log y_n - \sum_{n=1}^N (1-t_n) \log \{1-y_n\} \\
 &= -\sum_{n=1}^N t_n \log \{\sigma(a_n)\} - \sum_{n=1}^N (1-t_n) \log \{1-\sigma(a_n)\} \\
 &= -\sum_{n=1}^N t_n \log \{\sigma(a_n)\} - \sum_{n=1}^N (1-t_n) \log \{\sigma(-a_n)\} \quad (\text{Apply (4.14)}) \\
 \frac{\partial E(\mathbf{w})}{\partial a_k} &= -t_k \frac{\sigma(a_k)\{1-\sigma(a_k)\}}{\sigma(a_k)} + (1-t_k) \frac{\sigma(-a_k)\{1-\sigma(-a_k)\}}{\sigma(-a_k)} \quad (\text{Apply (4.88)}) \\
 &= -t_k\{1-\sigma(a_k)\} + (1-t_k)\sigma(a_k) \\
 \frac{\partial E(\mathbf{w})}{\partial a_k} &= \sigma(a_k) - t_k.
 \end{aligned}$$

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 $p(t_k = 1|\mathbf{x}) = y_k(\mathbf{x}, \mathbf{w})$. 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

$$\begin{aligned}
 \frac{\partial E(\mathbf{w})}{\partial a_{n,k}} &= \frac{\partial}{\partial a_{n,k}} \left[- \sum_{m=1}^N \sum_{r=1}^K t_{m,r} \log y_r(\mathbf{x}_m, \mathbf{w}) \right] \\
 &= \frac{\partial}{\partial a_{n,k}} \left[- \sum_{m=1}^N \sum_{r=1}^K t_{m,r} \log \left\{ \frac{\exp\{a_{m,r}\}}{\sum_{j=1}^K \exp\{a_{m,j}\}} \right\} \right] \\
 &= - \frac{\partial}{\partial a_{n,k}} \left[\sum_{m=1}^N \sum_{r=1}^K t_{m,r} a_{m,r} - \sum_{m=1}^N \sum_{r=1}^K t_{m,r} \log \left\{ \sum_{j=1}^K \exp\{a_{m,j}\} \right\} \right] \\
 &= - \left[t_{n,k} - \sum_{r=1}^K t_{n,r} \frac{\exp\{a_{n,k}\}}{\sum_{j=1}^K \exp\{a_{n,j}\}} \right] \\
 &= - \left[t_{n,k} - y_{n,k} \sum_{r=1}^K t_{n,r} \right] \quad (\text{Apply (5.25)}) \\
 \frac{\partial E(\mathbf{w})}{\partial a_{n,k}} &= y_{n,k} - t_{n,k}.
 \end{aligned}$$

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

$$\begin{aligned}
 \frac{d \tanh(a)}{da} &= \frac{d}{da} \left[\frac{e^a - e^{-a}}{e^a + e^{-a}} \right] \\
 &= \frac{d}{da} \left[\frac{e^a}{e^a + e^{-a}} - \frac{e^{-a}}{e^a + e^{-a}} \right] \\
 &= \frac{d}{da} \left[\frac{e^a}{e^a + e^{-a}} - 1 + \frac{e^a}{e^a + e^{-a}} \right] \\
 &= \frac{d}{da} \left[\frac{2}{1 + e^{-2a}} - 1 \right] \\
 &= \frac{4e^{-2a}}{(1 + e^{-2a})^2} \\
 &= \frac{4 + (e^a - e^{-a})^2}{(e^a + e^{-a})^2} - \left(\frac{e^a - e^{-a}}{e^a + e^{-a}} \right)^2 \\
 &= \frac{4 + (e^a + e^{-a})^2 - 4e^a e^{-a}}{(e^a + e^{-a})^2} - \{\tanh(a)\}^2 \quad (\text{Apply (5.59)}) \\
 \frac{d \tanh(a)}{da} &= 1 - \{\tanh(a)\}^2.
 \end{aligned}$$

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

$$\begin{aligned}
 p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= \prod_{n=1}^N p(t_n|\mathbf{x}_n, \mathbf{w}) \\
 &= \prod_{n=1}^N \left(\frac{1 - y(\mathbf{x}_n, \mathbf{w})}{2} \right)^{(1-t_n)/2} \left(\frac{1 + y(\mathbf{x}_n, \mathbf{w})}{2} \right)^{(1+t_n)/2} \quad (\text{Apply (2.261)}) \\
 -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= -\sum_{n=1}^N \frac{1 - t_n}{2} \log \left\{ \frac{1 - y(\mathbf{x}_n, \mathbf{w})}{2} \right\} + \\
 &\quad -\sum_{n=1}^N \frac{1 + t_n}{2} \log \left\{ \frac{1 + y(\mathbf{x}_n, \mathbf{w})}{2} \right\} \\
 (5.8) \quad -\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}) &= -\frac{1}{2} \left[\sum_{n=1}^N (1 - t_n) \log \{1 - y(\mathbf{x}_n, \mathbf{w})\} + \right. \\
 &\quad \left. + \sum_{n=1}^N (1 + t_n) \log \{1 + y(\mathbf{x}_n, \mathbf{w})\} \right] + N \log 2.
 \end{aligned}$$

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

$$\begin{aligned}
 \tilde{\sigma}(a) &= 2\sigma(a) - 1 \\
 &= 2\sigma(2a/2) - 1 \\
 \tilde{\sigma}(a) &= \tanh(a/2) \quad \text{Apply (3.1).}
 \end{aligned}$$

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}^* , 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

$$E(\mathbf{w}) \approx E(\mathbf{w}^*) + \frac{1}{2} \sum_{i=1}^D \lambda_i \alpha_i^2.$$

Note, from (5.35), that the values α_i are scalars which convert $\mathbf{w} - \mathbf{w}^*$ to linear combinations of the eigenvectors \mathbf{u}_i , hence they are aligned with the eigenvectors. Let us determine a fixed value $C > E(\mathbf{w}^*)$ such that

$$\begin{aligned} E(\mathbf{w}^*) + \frac{1}{2} \sum_{i=1}^D \lambda_i \alpha_i^2 &= C \\ \sum_{i=1}^D \left(\frac{\alpha_i}{1/\sqrt{\lambda_i}} \right)^2 &= 2C - 2E(\mathbf{w}^*) \\ \sum_{i=1}^D \left(\frac{\alpha_i}{1/\sqrt{\lambda_i}} \right)^2 &= \left\{ \sqrt{C^*} \right\}^2 \\ \sum_{i=1}^D \left(\frac{\alpha_i}{\sqrt{C^*/\lambda_i}} \right)^2 &= 1, \end{aligned} \tag{5.9}$$

Where $C^* = 2C - 2E(\mathbf{w}^*)$. It follows that (5.9) is the formula for a D -dimensional ellipsoid with semi-axes lengths $\sqrt{C^*/\lambda_i}$. 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}^* 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}^* is a local minima. It follows that, for all \mathbf{w} in the neighbourhood of \mathbf{w}^* , it is true that

$$(5.10) \quad \begin{aligned} E(\mathbf{w}) &\geq E(\mathbf{w}^*) \\ E(\mathbf{w}) - E(\mathbf{w}^*) &\geq 0. \end{aligned}$$

From (5.32), we find that

$$(5.11) \quad \begin{aligned} E(\mathbf{w}) &\approx E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \\ 2\{E(\mathbf{w}) - E(\mathbf{w}^*)\} &\approx (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \\ 0 &\leq (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) && \text{(Apply (5.10))} \\ 0 &\leq \sum_{i=1}^D \sum_{j=1}^D \alpha_i \alpha_j \mathbf{u}_i^\top \mathbf{H} \mathbf{u}_j && \text{(Apply (5.35))} \\ 0 &\leq \sum_{i=1}^D \sum_{j=1}^D \alpha_i \alpha_j \lambda_j \mathbf{u}_i^\top \mathbf{u}_j && \text{(Apply (2.45))} \\ 0 &\leq \sum_{i=1}^D \alpha_i^2 \lambda_i && \text{(Orthonormality of } \mathbf{u}_i\text{).} \end{aligned}$$

Note that (5.10) is only valid for \mathbf{w} in the neighbourhood of \mathbf{w}^* . 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}^* must be nonnegative, hence the Hessian must be positive semi-definite when evaluated at \mathbf{w}^* . Now, we assume that \mathbf{H} is positive definite when evaluated at \mathbf{w}^* . It follows that, for any choice of $\mathbf{w} - \mathbf{w}^*$ we find that

$$(5.12) \quad \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) > 0.$$

Hence, from (5.32) it follows that

$$(5.13) \quad \begin{aligned} E(\mathbf{w}) &\approx E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \\ E(\mathbf{w}) - E(\mathbf{w}^*) &\approx \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \\ E(\mathbf{w}) - E(\mathbf{w}^*) &> 0 && \text{(Apply (5.12)).} \end{aligned}$$

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}^* is a minimum for the error function $E(\mathbf{w})$. We therefore conclude that, given the quadratic approximation in (5.32), \mathbf{w}^* is a local minimum for the error function $E(\mathbf{w})$ if, and only if, the associated Hessian, evaluated at \mathbf{w}^* , 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

$$(5.14) \quad E_n(w_{j,i} + \epsilon) = E(w_{j,i}) + \frac{\partial E(w_{j,i})}{\partial w_{j,i}}\epsilon + \frac{\partial^2 E(w_{j,i})}{\partial w_{j,i}^2}\epsilon^2 + O(\epsilon^3).$$

Conversely, when evaluated at $w_{j,i} - \epsilon$, for $\epsilon > 0$, we obtain

$$(5.15) \quad E_n(w_{j,i} - \epsilon) = E(w_{j,i}) - \frac{\partial E(w_{j,i})}{\partial w_{j,i}}\epsilon + \frac{\partial^2 E(w_{j,i})}{\partial w_{j,i}^2}\epsilon^2 + O(\epsilon^3)$$

We may thereby rewrite (5.69) as

$$\begin{aligned} \frac{\partial E_n(w_{j,i})}{\partial w_{j,i}} &= \frac{E_n(w_{j,i} + \epsilon) - E_n(w_{j,i} - \epsilon)}{2\epsilon} + O(\epsilon^2) \\ &= \frac{E(w_{j,i}) + \frac{\partial E(w_{j,i})}{\partial w_{j,i}}\epsilon + \frac{\partial^2 E(w_{j,i})}{\partial w_{j,i}^2}\epsilon^2 + O(\epsilon^3)}{2\epsilon} + O(\epsilon^2) \\ &\quad + \frac{-E(w_{j,i}) + \frac{\partial E(w_{j,i})}{\partial w_{j,i}}\epsilon - \frac{\partial^2 E(w_{j,i})}{\partial w_{j,i}^2}\epsilon^2 + O(\epsilon^3)}{2\epsilon} \quad (\text{Apply (5.14) and (5.15)}) \\ \frac{\partial E_n(w_{j,i})}{\partial w_{j,i}} &= \frac{\partial E_n(w_{j,i})}{\partial w_{j,i}} + O(\epsilon^2). \end{aligned}$$

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

$$\begin{aligned}
 J_{k,i} &= \frac{\partial y_k}{\partial x_i} \\
 &= \frac{\partial y_k}{\partial a_k} \frac{\partial a_k}{\partial x_i} \\
 &= \frac{d\sigma(a_k)}{da_k} \sum_j \frac{\partial a_k}{\partial a_j} \frac{\partial a_j}{\partial x_i} \\
 (5.16) \quad J_{k,i} &= \frac{d\sigma(a_k)}{da_k} \sum_j w_{k,j} \frac{dh(a_j)}{da_j} \frac{\partial a_j}{\partial x_i}
 \end{aligned}$$

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

$$\begin{aligned}
 \frac{\partial a_j}{\partial x_i} &= \sum_r \frac{\partial a_j}{\partial a_r} \frac{\partial a_r}{\partial x_i} \\
 (5.17) \quad \frac{\partial a_j}{\partial x_i} &= \sum_r w_{j,r} \frac{dh(a_r)}{da_r} \frac{\partial a_r}{\partial x_i}
 \end{aligned}$$

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

$$(5.18) \quad \frac{\partial a_r}{\partial x_i} = w_{r,i}.$$

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 squared-error loss function, as in (5.11). The corresponding Hessian is as follows

$$\begin{aligned}
 \mathbf{H} &= \nabla \nabla^\top E(\mathbf{w}) \\
 &= \nabla \nabla^\top \left[\frac{1}{2} \sum_{n=1}^N \|\mathbf{y}_n - \mathbf{t}_n\|^2 \right] && \text{(Apply (5.11))} \\
 &= \frac{1}{2} \sum_{n=1}^N \nabla \nabla^\top \{\mathbf{y}_n - \mathbf{t}_n\}^\top \{\mathbf{y}_n - \mathbf{t}_n\} \\
 &= \frac{1}{2} \sum_{n=1}^N \nabla \nabla^\top \{\mathbf{y}_n^\top - \mathbf{t}_n^\top\} \{\mathbf{y}_n - \mathbf{t}_n\} \\
 &= \frac{1}{2} \sum_{n=1}^N \nabla \nabla^\top \{\mathbf{y}_n^\top \mathbf{y}_n - 2\mathbf{t}_n^\top \mathbf{y}_n + \mathbf{t}_n^\top \mathbf{t}_n\} \\
 &= \frac{1}{2} \sum_{n=1}^N \nabla \left\{ 2\mathbf{y}_n^\top \nabla^\top \mathbf{y}_n - 2\mathbf{t}_n^\top \nabla^\top \mathbf{y}_n \right\} \\
 &= \sum_{n=1}^N \left\{ \nabla \mathbf{y}_n^\top \nabla^\top \mathbf{y}_n + \mathbf{y}_n^\top \nabla \nabla^\top \mathbf{y}_n - \mathbf{t}_n^\top \nabla \nabla^\top \mathbf{y}_n \right\} \\
 (5.19) \quad \mathbf{H} &= \sum_{n=1}^N \nabla \mathbf{y}_n^\top \nabla^\top \mathbf{y}_n + \sum_{n=1}^N (\mathbf{y}_n - \mathbf{t}_n)^\top \nabla \nabla^\top \mathbf{y}_n.
 \end{aligned}$$

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

$$\mathbf{H} \approx \sum_{n=1}^N \mathbf{B}_n \mathbf{B}_n^\top,$$

where $\mathbf{B}_n = \nabla \mathbf{y}_n^\top$.

Exercise 5.17

We can proceed by analogy to the demonstration in [Exercise 1.26](#), by which we conclude that [\(5.193\)](#) may be rewritten as

$$(5.20) \quad E = \frac{1}{2} \int (y(\mathbf{x}, \mathbf{w}) - \mathbb{E}[T|\mathbf{x}])^2 p(\mathbf{x}) d\mathbf{x} + \frac{1}{2} \int \text{Var}[T|\mathbf{x}] p(\mathbf{x}) d\mathbf{x}.$$

It is obvious that only the first term on the right-hand-side of [\(5.20\)](#) is dependent on the network parameters, hence differentiating [\(5.20\)](#) with respect to w_r , and thereafter with respect to w_s , we obtain

$$(5.21) \quad \begin{aligned} \frac{\partial^2 E}{\partial w_r \partial w_s} &= \int \frac{\partial y(\mathbf{x}, \mathbf{w})}{\partial w_r} \frac{\partial y(\mathbf{x}, \mathbf{w})}{\partial w_s} p(\mathbf{x}) d\mathbf{x} + \\ &+ \int (y(\mathbf{x}, \mathbf{w}) - \mathbb{E}[T|\mathbf{x}]) \frac{\partial^2 y(\mathbf{x}, \mathbf{w})}{\partial w_r \partial w_s} p(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

We proceed again analogously by considering that as $y(\mathbf{x}, \mathbf{w}) = \mathbb{E}[T|\mathbf{x}]$ minimizes [\(1.151\)](#) (as demonstrated in [Exercise 1.26](#)), the second term on the right-hand-side of [\(5.21\)](#) vanishes, and we obtain

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \int \frac{\partial y(\mathbf{x}, \mathbf{w})}{\partial w_r} \frac{\partial y(\mathbf{x}, \mathbf{w})}{\partial w_s} p(\mathbf{x}),$$

which in the [\(1.35\)](#) sense, may be approximated as [\(5.84\)](#).

Exercise 5.18

We consider a two-layer neural network, including also skip-layer connections connecting the inputs directly to the outputs. Consequently, for arbitrary networks functions $\sigma(a)$ and $h(a)$, we have that the k -th output unit is computed as

$$(5.22) \quad y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\sum_{j=1}^M w_{k,j}^{(2)} h \left(\sum_{i=1}^D w_{j,i}^{(1)} x_i + w_{j,0}^{(1)} \right) + \sum_{r=1}^D \alpha_{k,r} x_r + \alpha_{k,0} + w_{k,0}^{(2)} \right),$$

where $\alpha_{k,r}$ are parameters which refer to the skip-layer connections. Assuming that the output units have linear activation functions, the hidden units are submitted to the hyperbolic tangent activation function, and the sum-of-squares error function is adopted, we aim to delineate a procedure to determine the derivatives of the error function, evaluated at a pattern \mathbf{x}_n , with respect to $w_{j,i}^{(1)}$, $w_{k,j}^{(2)}$ and $\alpha_{k,r}$. It is easy to see that the forms of the derivatives with respect to $w_{j,i}^{(1)}$ and $w_{k,j}^{(2)}$ remains unchanged from (5.67). By contrast, the derivative with respect to $\alpha_{k,r}$ is simply taken as

$$\frac{\partial E_n}{\partial \alpha_{k,r}} = \delta_k x_r,$$

where δ_k is as in (5.65).

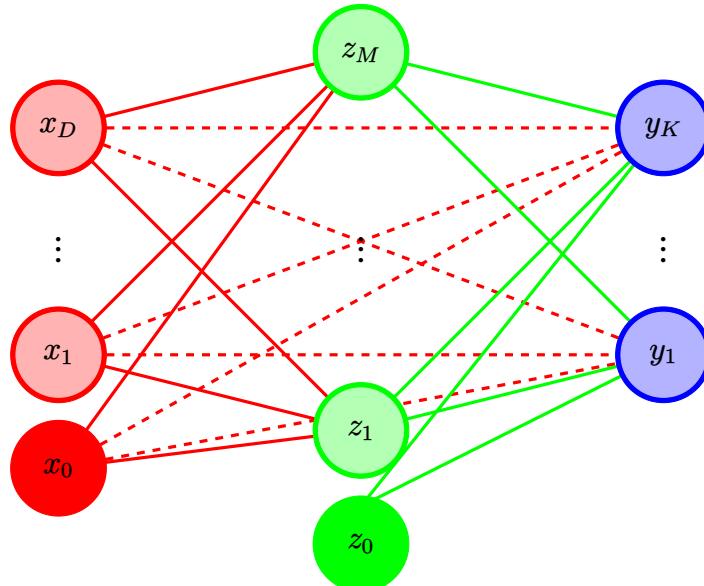


Figure 5.1: Illustration of a two-layer neural network with skip-layer connections.

Exercise 5.19

Let a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$ be observed, under the context of a classification problem, wherein we adopt the neural network framework to compute the output values. Moreover, by adopting the cross-entropy error function, as in (4.90), and utilizing the sigmoidal activation function for the output units, we determine the corresponding Hessian as

$$\begin{aligned}
 \mathbf{H} &= \nabla \nabla^\top E(\mathbf{w}) \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} \right] && \text{(Apply (4.90))} \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \{t_n \log[y_n/(1 - y_n)] + \log(1 - y_n)\} \right] \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \{t_n \sigma^{-1}(y_n) + \log(1 - y_n)\} \right] && \text{(Apply (4.15))} \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \{t_n \sigma^{-1}(\sigma(a_n)) + \log(1 - \sigma(a_n))\} \right] \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \{t_n a_n + \log(\sigma(-a_n))\} \right] && \text{(Apply (4.14))} \\
 &= \nabla \left[- \sum_{n=1}^N \left\{ t_n - \frac{1}{\sigma(-a_n)} \sigma(-a_n) \{1 - \sigma(-a_n)\} \right\} \nabla^\top a_n \right] && \text{(Apply (4.88))} \\
 &= \nabla \left[- \sum_{n=1}^N \{t_n - \sigma(a_n)\} \nabla^\top a_n \right] && \text{(Apply (4.14))} \\
 &= \sum_{n=1}^N \sigma(a_n) \{1 - \sigma(a_n)\} (\nabla a_n) (\nabla^\top a_n) + \\
 &\quad - \sum_{n=1}^N \{t_n - \sigma(a_n)\} \nabla \nabla^\top a_n && \text{(Apply (4.88))} \\
 (5.23) \quad \mathbf{H} &= \sum_{n=1}^N y_n \{1 - y_n\} (\nabla a_n) (\nabla^\top a_n) + \\
 &\quad - \sum_{n=1}^N \{t_n - y_n\} \nabla \nabla^\top a_n.
 \end{aligned}$$

We argue that, similarly to what was seen in Exercise 5.16 and Exercise 5.17, the second term on the right-hand-side of (5.23) is negligible under certain assumptions, hence we arrive at the approximation

$$\mathbf{H} \approx \sum_{n=1}^N y_n \{1 - y_n\} \mathbf{b}_n \mathbf{b}_n^\top,$$

where $\mathbf{b}_n = \nabla a_n$, as desired.

Exercise 5.20

We consider now a similar framework to that of [Exercise 5.19](#), except there are now K output units per observation, such that the output unit present softmax output unit activation function, as in [\(4.104\)](#), and we consider the cross entropy error function as in [\(4.108\)](#). It follows that the corresponding Hessian is computed as

$$\begin{aligned}
 \mathbf{H} &= \nabla \nabla^\top E(\mathbf{w}) \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \log y_{n,k} \right] && \text{(Apply (4.108))} \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \left\{ a_{n,k} - \log \left(\sum_{j=1}^K \exp\{a_{n,j}\} \right) \right\} \right] && \text{(Apply (4.104))} \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \sum_{k=1}^K t_{n,k} a_{n,k} + \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \log \left(\sum_{j=1}^K \exp\{a_{n,j}\} \right) \right] \\
 &= \nabla \nabla^\top \left[- \sum_{n=1}^N \sum_{k=1}^K t_{n,k} a_{n,k} + \sum_{n=1}^N \log \left(\sum_{j=1}^K \exp\{a_{n,j}\} \right) \right] \\
 &= \nabla \left[- \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \nabla^\top a_{n,k} + \sum_{n=1}^N \sum_{k=1}^K \frac{\exp\{a_{n,k}\}}{\sum_{j=1}^K \exp\{a_{n,j}\}} \nabla^\top a_{n,k} \right] \\
 &= - \sum_{n=1}^N \sum_{k=1}^K t_{n,k} \nabla \nabla^\top a_{n,k} + \sum_{n=1}^N \sum_{k=1}^K \frac{\exp\{a_{n,k}\}}{\sum_{j=1}^K \exp\{a_{n,j}\}} \nabla \nabla^\top a_{n,k} + \\
 &\quad - \sum_{n=1}^N \sum_{k=1}^K \sum_{\substack{r=1 \\ r \neq k}}^K \frac{\exp\{a_{n,k}\} \exp\{a_{n,r}\}}{(\sum_{j=1}^K \exp\{a_{n,j}\})^2} \nabla a_{n,r} \nabla^\top a_{n,k} + \\
 &\quad + \sum_{n=1}^N \sum_{k=1}^K \frac{(\sum_{j=1}^K \exp\{a_{n,j}\}) \exp\{a_{n,k}\} - \exp\{2a_{n,k}\}}{(\sum_{j=1}^K \exp\{a_{n,j}\})^2} \nabla a_{n,k} \nabla^\top a_{n,k} \\
 &= \sum_{n=1}^N \sum_{k=1}^K (y_{n,k} - t_{n,k}) \nabla \nabla^\top a_{n,k} - \sum_{n=1}^N \sum_{k=1}^K \sum_{\substack{r=1 \\ r \neq k}}^K y_{n,k} y_{n,r} \nabla a_{n,r} \nabla^\top a_{n,k} + \\
 &\quad + \sum_{n=1}^N \sum_{k=1}^K y_{n,k} (1 - y_{n,k}) \nabla a_{n,k} \nabla^\top a_{n,k} && \text{(Apply (4.104))} \\
 (5.24) \quad \mathbf{H} &= \sum_{n=1}^N \sum_{k=1}^K (y_{n,k} - t_{n,k}) \nabla \nabla^\top a_{n,k} + \\
 &\quad + \sum_{n=1}^N \sum_{k=1}^K \sum_{r=1}^K y_{n,k} (I_{k,r} - y_{n,r}) \nabla a_{n,k} \nabla^\top a_{n,k}
 \end{aligned}$$

Similarly to previous Exercises, we argue that, if the neural network is properly trained, the term $(y_{n,k} - t_{n,k})$ is small, hence we may approximate [\(5.24\)](#) as

$$\mathbf{H} \approx \sum_{n=1}^N \sum_{k=1}^K \sum_{r=1}^K y_{n,k} (I_{k,r} - y_{n,r}) \mathbf{b}_{n,r} \mathbf{b}_{n,k}^\top,$$

where $\mathbf{b}_{n,k} = \nabla a_{n,k}$.

Exercise 5.21

We consider the outer product approximation to the Hessian matrix as seen in (5.86), wherein we possess $K > 1$ output units per observation, such that we define

$$\begin{aligned}\mathbf{B}_n &= \nabla \mathbf{a}_n^\top \\ &= (\nabla a_{n,1}^\top \quad \dots \quad \nabla a_{n,K}^\top).\end{aligned}$$

It follows that we rewrite (5.86) as

$$\begin{aligned}\mathbf{H}_{L+1} &= \sum_{n=1}^{L+1} \mathbf{B}_n \mathbf{B}_n^\top \\ &= \sum_{n=1}^L \mathbf{B}_n \mathbf{B}_n^\top + \mathbf{B}_{L+1} \mathbf{B}_{L+1}^\top \\ (5.25) \quad \mathbf{H}_{L+1} &= \mathbf{H}_L + \mathbf{B}_{L+1} \mathbf{B}_{L+1}^\top \quad (\text{Apply (5.86)}).\end{aligned}$$

In order to compute the inverse of \mathbf{H}_{L+1} we proceed as follows

$$\begin{aligned}\mathbf{H}_{L+1}^{-1} &= (\mathbf{H}_L + \mathbf{B}_{L+1} \mathbf{B}_{L+1}^\top)^{-1} \\ (5.26) \quad \mathbf{H}_{L+1}^{-1} &= \mathbf{H}_L^{-1} - \mathbf{H}_L^{-1} \mathbf{B}_{L+1} (\mathbf{I} + \mathbf{B}_{L+1}^\top \mathbf{H}_L^{-1} \mathbf{B}_{L+1})^{-1} \mathbf{B}_{L+1}^\top \mathbf{H}_L^{-1} \quad (\text{Apply (2.289)}).\end{aligned}$$

Hence (5.26) delineates a procedure to sequentially incorporate contributions to the Hessian inverse as data points are gathered.

Exercise 5.22

We consider a two-layer neural network, with linear output unit activation function and arbitrary hidden layer activation function $h(a)$. We aim to thereafter verify (5.93), (5.94) and (5.95). First, we verify (5.93) as follows

$$\begin{aligned}
 \frac{\partial^2 E_n}{\partial w_{k,j}^{(2)} \partial w_{k',j'}^{(2)}} &= \frac{\partial}{\partial w_{k',j'}^{(2)}} \left[\frac{\partial E_n}{\partial w_{k,j}^{(2)}} \right] \\
 &= \frac{\partial}{\partial w_{k',j'}^{(2)}} \left[\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{k,j}^{(2)}} \right] \\
 &= \frac{\partial}{\partial w_{k',j'}^{(2)}} \left[\frac{\partial E_n}{\partial a_k} z_j \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial w_{k',j'}^{(2)}} z_j \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k',j'}^{(2)}} z_j \right] \\
 &= z_j z_{j'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \\
 \frac{\partial^2 E_n}{\partial w_{k,j}^{(2)} \partial w_{k',j'}^{(2)}} &= z_j z_{j'} M_{k,k'} \quad (\text{Apply (5.92)}).
 \end{aligned}$$

Hence we reach the desired result. We subsequently verify (5.93) as

$$\begin{aligned}
 \frac{\partial^2 E_n}{\partial w_{j,i}^{(1)} \partial w_{j',i'}^{(1)}} &= \frac{\partial}{\partial w_{j',i'}^{(1)}} \left[\frac{\partial E_n}{\partial w_{j,i}^{(1)}} \right] \\
 &= \frac{\partial}{\partial w_{j',i'}^{(1)}} \left[\sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{j,i}^{(1)}} \right] \\
 &= \frac{\partial}{\partial w_{j',i'}^{(1)}} \left[\sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial a_j} \frac{\partial a_j}{\partial w_{j,i}^{(1)}} \right] \\
 &= \frac{\partial}{\partial w_{j',i'}^{(1)}} \left[\sum_k \frac{\partial E_n}{\partial a_k} h'(a_j) w_{k,j}^{(2)} x_i \right] \\
 &= \sum_k \frac{\partial}{\partial a_k} \frac{\partial E_n}{\partial w_{j',i'}^{(1)}} h'(a_j) w_{k,j}^{(2)} x_i + I_{j,j'} \sum_k \frac{\partial E_n}{\partial a_k} h''(a_j) w_{k,j}^{(2)} x_i x_{i'} \\
 &= \sum_k \sum_{k'} \frac{\partial}{\partial a_k} \frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial a_{j'}} \frac{\partial a_{j'}}{\partial w_{j',i'}^{(1)}} h'(a_j) w_{k,j}^{(2)} x_i + \\
 &\quad + I_{j,j'} \sum_k \frac{\partial E_n}{\partial a_k} h''(a_j) w_{k,j}^{(2)} x_i x_{i'} \\
 &= \sum_k \sum_{k'} \frac{\partial}{\partial a_k} \frac{\partial E_n}{\partial a_{k'}} h'(a_{j'}) w_{k',j'}^{(2)} x_{i'} h'(a_j) w_{k,j}^{(2)} x_i + \\
 &\quad + I_{j,j'} \sum_k \frac{\partial E_n}{\partial a_k} h''(a_j) w_{k,j}^{(2)} x_i x_{i'} \\
 &= \sum_k \sum_{k'} \frac{\partial^2 E_n}{\partial a_{k'} \partial a_k} h'(a_{j'}) w_{k',j'}^{(2)} x_{i'} h'(a_j) w_{k,j}^{(2)} x_i + \\
 &\quad + I_{j,j'} \sum_k \frac{\partial E_n}{\partial a_k} h''(a_j) w_{k,j}^{(2)} x_i x_{i'} \\
 &= \sum_k \sum_{k'} M_{k,k'} h'(a_{j'}) w_{k',j'}^{(2)} x_{i'} h'(a_j) w_{k,j}^{(2)} x_i + \\
 &\quad + I_{j,j'} \sum_k \delta_k h''(a_j) w_{k,j}^{(2)} x_i x_{i'} \tag{Apply (5.92)} \\
 \frac{\partial^2 E_n}{\partial w_{j,i}^{(1)} \partial w_{j',i'}^{(1)}} &= x_i x_{i'} h'(a_{j'}) h'(a_j) \sum_k \sum_{k'} w_{k',j'}^{(2)} w_{k,j}^{(2)} M_{k,k'} + \\
 &\quad + x_i x_{i'} h''(a_j) I_{j,j'} \sum_k w_{k,j}^{(2)} \delta_k.
 \end{aligned}$$

Note that the sum over k above runs over output units to which the j -th hidden layer unit sends connections. Similarly, the sum over k' runs over output units to which the j' -th hidden layer unit sends connections. Again we reach the desired result. Lastly, we

verify (5.94) as

$$\begin{aligned}
 \frac{\partial^2 E_n}{\partial w_{j,i}^{(1)} \partial w_{k,j'}^{(2)}} &= \frac{\partial}{\partial w_{k,j'}^{(2)}} \left[\frac{\partial E_n}{\partial w_{j,i}^{(1)}} \right] \\
 &= \frac{\partial}{\partial w_{k,j'}^{(2)}} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{j,i}^{(1)}} \right] \\
 &= \frac{\partial}{\partial w_{k,j'}^{(2)}} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial a_j} \frac{\partial a_j}{\partial w_{j,i}^{(1)}} \right] \\
 &= \frac{\partial}{\partial w_{k,j'}^{(2)}} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} h'(a_j) w_{k',j}^{(2)} x_i \right] \\
 &= \sum_{k'} \frac{\partial}{\partial w_{k,j'}^{(2)}} \frac{\partial E_n}{\partial a_{k'}} h'(a_j) w_{k',j}^{(2)} x_i + I_{j,j'} \frac{\partial E_n}{\partial a_k} h'(a_j) x_i \\
 &= \sum_{k'} \frac{\partial}{\partial a_{k'}} \frac{\partial E_n}{\partial w_{k,j'}^{(2)}} h'(a_j) w_{k',j}^{(2)} x_i + I_{j,j'} \frac{\partial E_n}{\partial a_k} h'(a_j) x_i \\
 &= \sum_{k'} \frac{\partial}{\partial a_{k'}} \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{k,j'}^{(2)}} h'(a_j) w_{k',j}^{(2)} x_i + I_{j,j'} \frac{\partial E_n}{\partial a_k} h'(a_j) x_i \\
 &= \sum_{k'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} z_{j'} h'(a_j) w_{k',j}^{(2)} x_i + I_{j,j'} \frac{\partial E_n}{\partial a_k} h'(a_j) x_i \\
 &= \sum_{k'} M_{k,k'} z_{j'} h'(a_j) w_{k',j}^{(2)} x_i + I_{j,j'} \delta_k h'(a_j) x_i \quad (\text{Apply (5.92)}) \\
 \frac{\partial^2 E_n}{\partial w_{j,i}^{(1)} \partial w_{k,j'}^{(2)}} &= x_i h'(a_j) \left\{ \delta_k I_{j,j'} + z_{j'} \sum_{k'} w_{k',j}^{(2)} M_{k,k'} \right\}.
 \end{aligned}$$

We therefore verify the result (5.95). Note that the sum above in k' runs over output units to which the j -th hidden layer unit sends connections to.

Exercise 5.23

We consider now the same context as in [Exercise 5.22](#), except we now also include the skip-layer connections as defined in [Exercise 5.18](#), hence we adopt the same symbols as used therein and in [\(5.22\)](#). The results [\(5.93\)](#), [\(5.94\)](#) and [\(5.95\)](#) are not significantly altered, with the exception that the values a_k now also include contributions provided by the skip-layer connections. Thus, in order to extend the Hessian computations, we only consider three cases: first, the Hessian when both term are skip-layer connections, obtained as

$$\begin{aligned}
 \frac{\partial^2 E_n}{\partial \alpha_{k,i} \partial \alpha_{k',i'}} &= \frac{\partial}{\partial \alpha_{k',i'}} \left[\frac{\partial E_n}{\partial \alpha_{k,i}} \right] \\
 &= \frac{\partial}{\partial \alpha_{k',i'}} \left[\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial \alpha_{k,i}} \right] \\
 &= \frac{\partial}{\partial \alpha_{k',i'}} \left[\frac{\partial E_n}{\partial a_k} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial \alpha_{k',i'}} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial \alpha_{k',i'}} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial a_{k'}} x_{i'} x_i \right] \\
 &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} x_{i'} x_i \\
 \frac{\partial^2 E_n}{\partial \alpha_{k,i} \partial \alpha_{k',i'}} &= x_{i'} x_i M_{k,k'} \quad (\text{Apply } \textcolor{red}{(5.92)}).
 \end{aligned}$$

We thereafter consider the case wherein one term is a skip-layer connection and another is a second-layer weight, with Hessian as follows

$$\begin{aligned}
 \frac{\partial^2 E_n}{\partial \alpha_{k,i} \partial w_{k',j}^{(2)}} &= \frac{\partial}{\partial w_{k',j}^{(2)}} \left[\frac{\partial E_n}{\partial \alpha_{k,i}} \right] \\
 &= \frac{\partial}{\partial w_{k',j}^{(2)}} \left[\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial \alpha_{k,i}} \right] \\
 &= \frac{\partial}{\partial w_{k',j}^{(2)}} \left[\frac{\partial E_n}{\partial a_k} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial w_{k',j}^{(2)}} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k',j}^{(2)}} x_i \right] \\
 &= \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} z_j x_i \\
 \frac{\partial^2 E_n}{\partial \alpha_{k,i} \partial w_{k',j}^{(2)}} &= z_j x_i M_{k,k'} \quad (\text{Apply (5.92)}).
 \end{aligned}$$

Lastly, we consider the case wherein one term is a skip-layer connection and another a first-layer weight, with Hessian as follows

$$\begin{aligned}
 \frac{\partial^2 E_n}{\partial \alpha_{k,i} \partial w_{j,i'}^{(1)}} &= \frac{\partial}{\partial w_{j,i'}^{(1)}} \left[\frac{\partial E_n}{\partial \alpha_{k,i}} \right] \\
 &= \frac{\partial}{\partial w_{j,i'}^{(1)}} \left[\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial \alpha_{k,i}} \right] \\
 &= \frac{\partial}{\partial w_{j,i'}^{(1)}} \left[\frac{\partial E_n}{\partial a_k} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\frac{\partial E_n}{\partial w_{j,i'}^{(1)}} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial a_j} \frac{\partial a_j}{\partial w_{j,i'}^{(1)}} x_i \right] \\
 &= \frac{\partial}{\partial a_k} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} w_{k',j}^{(2)} h'(a_j) x_{i'} x_i \right] \\
 &= h'(a_j) x_{i'} x_i \sum_{k'} \frac{\partial^2 E_n}{\partial a_{k'} \partial a_k} w_{k',j}^{(2)} \\
 \frac{\partial^2 E_n}{\partial \alpha_{k,i} \partial w_{j,i'}^{(1)}} &= h'(a_j) x_{i'} x_i \sum_{k'} M_{k,k'} w_{k',j}^{(2)} \quad (\text{Apply (5.92)}).
 \end{aligned}$$

Wherein the above sum with respect k' runs over output units to which the j -th hidden unit sends connections to.

Exercise 5.24

We aim to demonstrate first that, for a neural network whose first hidden layers is of the form (5.113), and whose output layer is of the form (5.114), if the input units are subject to transformations of the form (5.115), by transforming the first layer weights according to (5.116) and (5.117) the hidden units z_j remain unmodified. We substitute (5.115), (5.116) and (5.117) into the right-hand-side of (5.113), obtaining the following

$$\begin{aligned} h\left(\sum_i \tilde{w}_{j,i} \tilde{x}_i + \tilde{w}_{j,0}\right) &= h\left(\sum_i \frac{1}{a} w_{j,i} (ax_i + b) + w_{j,0} - \frac{b}{a} \sum_i w_{j,i}\right) && \text{(Apply (5.115), (5.116) and (5.117))} \\ &= h\left(\sum_i w_{j,i} x_i + \frac{b}{a} \sum_i w_{j,i} + w_{j,0} - \frac{b}{a} \sum_i w_{j,i}\right) \\ &= h\left(\sum_i w_{j,i} x_i + w_{j,0}\right) \\ h\left(\sum_i \tilde{w}_{j,i} \tilde{x}_i + \tilde{w}_{j,0}\right) &= z_j && \text{(Apply (5.113)).} \end{aligned}$$

Hence, we conclude that the hidden units z_j remain unchanged. Conversely, we seek to now demonstrate that modifying the output weight parameters according to (5.119) and (5.120) is equivalent to submitting the output units to the transformation (5.118). By substituting (5.119) and (5.120) into the right-hand-side of (5.118) we obtain

$$\begin{aligned} \sum_j \tilde{w}_{k,j} z_j + \tilde{w}_{k,0} &= \sum_j cw_{k,j} z_j + cw_{k,0} + d && \text{(Apply (5.119) and (5.120))} \\ &= c \left[\sum_j w_{k,j} z_j + w_{k,0} \right] + d \\ \sum_j \tilde{w}_{k,j} z_j + \tilde{w}_{k,0} &= cy_k + d && \text{(Apply (5.114)).} \end{aligned}$$

Thereby reaching the desired result.

Exercise 5.25

We consider an optimization problem, wherein we aim to minimize a quadratic error function of form (5.195) by way of the iterative procedure delineated by (5.196) (starting at the origin), wherein τ denotes the step number and ρ is the learning rate. We aim to demonstrate that the components of $\mathbf{w}^{(\tau)}$ parallel to the eigenvectors of \mathbf{H} are such that (5.197). First, we must demonstrate that this holds for the first step ($\tau = 1$). For that purpose, we compute ∇E as follows

$$\begin{aligned}\nabla E &= \nabla \left[E_0 + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \right] \\ (5.27) \quad \nabla E &= \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \end{aligned}\quad (\text{Apply (C.19)})$$

We subsequently proceed to demonstrate that (5.197) holds for $\tau = 1$ as follows

$$\begin{aligned}\mathbf{w}^{(1)} &= \mathbf{w}^{(0)} - \rho \nabla E \\ &= \mathbf{w}^{(0)} - \rho \mathbf{H}(\mathbf{w}^{(0)} - \mathbf{w}^*) \quad (\text{Apply (5.27)}) \\ &= \rho \mathbf{H} \mathbf{w}^* \quad (\mathbf{w}^{(0)} = 0) \\ \mathbf{u}_j^\top \mathbf{w}^{(1)} &= \rho \mathbf{u}_j^\top \mathbf{H} \mathbf{w}^* \quad (\text{Left multiply by } \mathbf{u}_j^\top) \\ \mathbf{u}_j^\top \mathbf{w}^{(1)} &= \rho \eta_j \mathbf{u}_j^\top \mathbf{w}^* \quad (\text{Apply (5.198)}) \\ w_j^{(1)} &= \rho \eta_j w_j^* \\ w_j^{(1)} &= \{1 - (1 - \rho \eta_j)^1\} w_j^*,\end{aligned}$$

where $w_j^{(\tau)} = (\mathbf{w}^{(\tau)})^\top \mathbf{u}_j$. We conclude that (5.197) holds for $\tau = 1$. We seek thereafter to demonstrate that if (5.197) holds for τ , it holds for $\tau + 1$. It follows that

$$\begin{aligned}\mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \rho \nabla E \\ &= \mathbf{w}^{(\tau)} - \rho \mathbf{H}(\mathbf{w}^{(\tau)} - \mathbf{w}^*) \quad (\text{Apply (5.27)}) \\ \mathbf{u}_j^\top \mathbf{w}^{(\tau+1)} &= \mathbf{u}_j^\top \mathbf{w}^{(\tau)} - \rho \mathbf{u}_j^\top \mathbf{H}(\mathbf{w}^{(\tau)} - \mathbf{w}^*) \quad (\text{Left multiply by } \mathbf{u}_j^\top) \\ &= \mathbf{u}_j^\top \mathbf{w}^{(\tau)} - \rho \eta_j \mathbf{u}_j^\top (\mathbf{w}^{(\tau)} - \mathbf{w}^*) \quad (\text{Apply (5.198)}) \\ w_j^{(\tau+1)} &= (1 - \rho \eta_j) w_j^{(\tau)} + \rho \eta_j w_j^* \\ &= (1 - \rho \eta_j) \{1 - (1 - \rho \eta_j)^\tau\} w_j^* + \rho \eta_j w_j^* \\ &= \{1 - \rho \eta_j - (1 - \rho \eta_j)^{\tau+1} + \rho \eta_j\} w_j^* \\ w_j^{(\tau+1)} &= \{1 - (1 - \rho \eta_j)^{\tau+1}\} w_j^*.\end{aligned}$$

Hence, we conclude that if (5.197) holds for τ , it holds for $\tau + 1$. As we previously demonstrated that (5.197) holds for $\tau = 1$, we thereby conclude, by induction, that (5.197) is true for all $\tau \in \mathbb{N}$. If we assume that $|1 - \rho \eta_j| < 1$ for all j , it follows by direct application onto (5.197) that

$$\begin{aligned}\lim_{\tau \rightarrow \infty} w_j^{(\tau)} &= \lim_{\tau \rightarrow \infty} \{1 - (1 - \rho \eta_j)^\tau\} w_j^* \\ &= \{1 - 0\} w_j^* \\ \lim_{\tau \rightarrow \infty} w_j^{(\tau)} &= w_j^* \\ (5.28) \quad \lim_{\tau \rightarrow \infty} (\mathbf{w}^{(\tau)})^\top \mathbf{u}_j &= (\mathbf{w}^*)^\top \mathbf{u}_j.\end{aligned}$$

From (5.28) we can conclude that

$$\begin{aligned}
 \lim_{\tau \rightarrow \infty} (\mathbf{w}^{(\tau)})^\top \mathbf{U} &= (\mathbf{w}^*)^\top \mathbf{U} \\
 \lim_{\tau \rightarrow \infty} (\mathbf{w}^{(\tau)})^\top \mathbf{U} \mathbf{U}^\top &= (\mathbf{w}^*)^\top \mathbf{U} \mathbf{U}^\top \quad (\text{Right multiply by } \mathbf{U}^\top) \\
 \lim_{\tau \rightarrow \infty} (\mathbf{w}^{(\tau)})^\top &= (\mathbf{w}^*)^\top \\
 (5.29) \quad \lim_{\tau \rightarrow \infty} \mathbf{w}^{(\tau)} &= \mathbf{w}^*
 \end{aligned}$$

wherein \mathbf{U} is a matrix whose columns are composed by the eigenvectors \mathbf{u}_j . Hence we conclude that as $\tau \rightarrow \infty$, we obtain $\mathbf{w}^{(\tau)} \rightarrow \mathbf{w}^*$. Assuming still that $|1 - \rho\eta_j| < 1$ for all j , we consider now two cases wherein τ is finite: first, for $\eta_j \gg (\rho\tau)^{-1}$, we note that as $|1 - \rho\eta_j| < 1$ holds for all j , we also find that

$$\begin{aligned}
 |1 - \rho\eta_j| &< 1 \\
 -1 < 1 - \rho\eta_j &< 1 \\
 (5.30) \quad 0 &< \rho\eta_j < 2.
 \end{aligned}$$

As $\eta_j \gg (\rho\tau)^{-1}$ implies also that $\eta_j\rho \gg \tau^{-1}$, and from (5.30) we find that $\eta_j\rho \in (0, 2)$, we conclude that τ^{-1} must be very small, that is, that τ must be very large. For sufficiently large, yet finite, τ , we conclude that (5.29) holds approximately, hence $w_j^{(\tau)} \approx w_j^*$, as in (5.199). We now consider $\eta_j \ll (\rho\tau)^{-1}$: we obtain

$$\begin{aligned}
 w_j^{(\tau)} &= \{1 - (1 - \rho\eta_j)^\tau\} w_j^* \\
 &\approx \{1 - (1 - \rho\eta_j\tau)\} w_j^* \quad (\text{Apply } (1 - x)^k \approx (1 - kx)) \\
 &= \eta_j \rho \tau w_j^* \\
 w_j^{(\tau)} &\approx \frac{\eta_j}{(\rho\tau)^{-1}} w_j^* \\
 |w_j^{(\tau)}| &\approx \frac{\eta_j}{(\rho\tau)^{-1}} |w_j^*| \quad (\text{Apply } |\cdot| \text{ on both sides}) \\
 |w_j^{(\tau)}| &\ll |w_j^*| \quad (\text{Apply } \eta_j \ll (\rho\tau)^{-1}).
 \end{aligned}$$

Hence we verify (5.200). We may compare the early stopping procedure to the regularization induced by imposing the prior (3.52) on \mathbf{w} , whereupon we consider equivalent the roles of $(\rho\tau)^{-1}$ and α (alternatively denoted as λ), such that eigenvalues larger than $(\rho\tau)^{-1}$ (or the commensurate regularization parameter) yield estimates which are close to the maximum likelihood estimate (note that as our error function (5.195) is quadratic with respect to the parameters, it functions in an analog fashion to error functions induced by normally distributed output data). Conversely, if the eigenvalue is smaller than $(\rho\tau)^{-1}$ (or the commensurate regularization parameter), the absolute value of the resulting estimate is shrunk towards zero. As $(\rho\tau)^{-1}$ is inversely proportional to τ , it is easy to conclude that early stopping may be considered tantamount to adopting a larger value of α , whilst $\tau \rightarrow \infty$ equates to $\alpha = 0$, i.e., no regularization.

Exercise 5.26

We consider a neural network with arbitrary feed-forward topology which is trained by minimizing (5.127), with regularization term as defined in (5.128), arbitrary hidden layer activation function $h(a)$ and output activation function $\sigma(a)$. First, we rewrite the regularization term (5.128) as a sum over all observed patterns, as follows

$$\begin{aligned}
 \Omega &= \frac{1}{2} \sum_n \sum_k \left(\frac{\partial y_{n,k}}{\partial \xi} \Big|_{\xi=0} \right)^2 && \text{(Apply (5.128))} \\
 &= \sum_n \frac{1}{2} \sum_k \left(\sum_i \frac{\partial y_{n,k}}{\partial x_i} \frac{\partial x_i}{\partial \xi} \Big|_{\xi=0} \right)^2 && \text{(Apply (5.126))} \\
 &= \sum_n \frac{1}{2} \sum_k \left(\sum_i \tau_i \frac{\partial}{\partial x_i} y_{n,k} \right)^2 && \text{(Apply } \tau_i = \frac{\partial x_i}{\partial \xi} \Big|_{\xi=0} \text{)} \\
 &= \sum_n \frac{1}{2} \sum_k (\mathcal{G}y_{n,k})^2 && \text{(Apply (5.202))} \\
 \Omega &= \sum_n \Omega_n && \text{(Apply (5.201)).}
 \end{aligned}$$

We now aim to demonstrate that Ω_n can be computed by forward propagation utilizing α_j and β_j as defined in (5.205). Note that, as Ω_n is a function of the sum of squares of $\mathcal{G}y_{n,k}$, we need to demonstrate only that $\mathcal{G}y_{n,k}$ may be computed by forward propagation utilizing α_j and β_j . It follows that

$$\begin{aligned}
 \mathcal{G}y_{n,k} &= \sum_i \tau_i \frac{\partial}{\partial x_i} y_{n,k} && \text{(Apply (5.202))} \\
 &= \sum_i \tau_i \sum_j \frac{\partial y_{n,k}}{\partial z_{n,j}} \frac{\partial z_{n,j}}{\partial x_i} \\
 &= \sigma'(a_{n,k}) \sum_i \tau_i \frac{\partial}{\partial x_i} \sum_j w_{k,j} z_{n,j} \\
 &= \sigma'(a_{n,k}) \mathcal{G}a_{n,k} && \text{(Apply (5.202))} \\
 (5.31) \quad \mathcal{G}y_{n,k} &= \sigma'(a_{n,k}) \beta_k && \text{(Apply (5.205)),}
 \end{aligned}$$

wherein the sum over j is made over units which send connections to the k -th output unit. From (5.204), the terms β_k may be computed from α_i , hence we may compute (5.31) by forward propagation, and consequently so too Ω_n . The forward propagation procedure would start at the nodes which receive exclusively connections from input units, such that the corresponding α_i are computed, and thereafter one would iterate between computing α values and β values, until the propagation arrives the output units, by which point (5.204) is effectively computed. Lastly, we seek to verify that the derivative of (5.201)

with respect to $w_{r,s}$ is as in (5.206), as follows

$$\begin{aligned}
 \frac{\partial \Omega_n}{\partial w_{r,s}} &= \frac{\partial}{\partial w_{r,s}} \frac{1}{2} \sum_k (\mathcal{G}y_{n,k})^2 \\
 &= \sum_k \mathcal{G}y_{n,k} \frac{\partial \mathcal{G}y_{n,k}}{\partial w_{r,s}} \\
 &= \sum_k \mathcal{G}y_{n,k} \mathcal{G} \left(\frac{\partial y_{n,k}}{\partial a_{n,r}} \frac{\partial a_{n,r}}{\partial w_{r,s}} \right) \\
 &= \sum_k \mathcal{G}y_{n,k} \mathcal{G} \left(\frac{\partial y_{n,k}}{\partial a_{n,r}} z_{n,s} \right) \\
 &= \sum_k \mathcal{G}y_{n,k} \left\{ \mathcal{G} \frac{\partial y_{n,k}}{\partial a_{n,r}} z_{n,s} + \frac{\partial y_{n,k}}{\partial a_{n,r}} \mathcal{G} z_{n,s} \right\} \\
 \frac{\partial \Omega_n}{\partial w_{r,s}} &= \sum_k \alpha_k \{ \phi_{k,r} z_{n,s} + \delta_{k,r} \alpha_s \} \quad (\text{Apply (5.205) and (5.207)}),
 \end{aligned}$$

wherein we have extended the definition of α_j in (5.205) to account for the output activation function when the node over which we are applying \mathcal{G} is an output unit. Hence we have verified (5.206).

Exercise 5.27

We aim to demonstrate herein that, if we consider the transformation approach to invariance, particularly such that the inputs are subject to the transformation $\mathbf{x} \rightarrow \mathbf{x} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a multivariate Gaussian random variable with mean zero and identity covariance, we obtain (5.135). We utilize the squared error function (5.193), which may be decomposed as in (5.127). Let $\mathbf{s}(\mathbf{x}, \boldsymbol{\xi}) = \mathbf{x} + \boldsymbol{\xi}$, we compute the corresponding derivatives as

$$(5.32) \quad \frac{\partial s(\mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^\top} = \mathbf{I}.$$

Trivially, the second derivatives are all zero. We construct a third order Taylor polynomial approximation of $y(\mathbf{s}(\mathbf{x}, \boldsymbol{\xi}))$ around $\boldsymbol{\xi} = \mathbf{0}$ as follows

$$\begin{aligned}
 y(\mathbf{s}(\mathbf{x}, \boldsymbol{\xi})) &\approx y(\mathbf{s}(\mathbf{x}, \mathbf{0})) + \boldsymbol{\xi}^\top \frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial \boldsymbol{\xi}} \Big|_{\boldsymbol{\xi}=\mathbf{0}} + \\
 &+ \frac{1}{2} \boldsymbol{\xi}^\top \frac{\partial^2 y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^\top} \Big|_{\boldsymbol{\xi}=\mathbf{0}} \boldsymbol{\xi} \\
 &= y(\mathbf{x}) + \boldsymbol{\xi}^\top \frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial s(\mathbf{x}, \boldsymbol{\xi})} \frac{\partial s(\mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^\top} \Big|_{\boldsymbol{\xi}=\mathbf{0}} + \\
 &+ \frac{1}{2} \boldsymbol{\xi}^\top \frac{\partial}{\partial \boldsymbol{\xi}} \left[\frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial [s(\mathbf{x}, \boldsymbol{\xi})]^\top} \frac{\partial [s(\mathbf{x}, \boldsymbol{\xi})]^\top}{\partial \boldsymbol{\xi}} \Big|_{\boldsymbol{\xi}=\mathbf{0}} \right] \boldsymbol{\xi} \quad (\text{Apply } \mathbf{s}(\mathbf{x}, \mathbf{0}) = \mathbf{x}) \\
 &= y(\mathbf{x}) + \boldsymbol{\xi}^\top \frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial s(\mathbf{x}, \boldsymbol{\xi})} \Big|_{\boldsymbol{\xi}=\mathbf{0}} + \\
 &+ \frac{1}{2} \boldsymbol{\xi}^\top \frac{\partial}{\partial s(\mathbf{x}, \boldsymbol{\xi})} \left[\frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial \boldsymbol{\xi}^\top} \Big|_{\boldsymbol{\xi}=\mathbf{0}} \right] \boldsymbol{\xi} \quad (\text{Apply (5.32)}) \\
 &= y(\mathbf{x}) + \boldsymbol{\xi}^\top \frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial s(\mathbf{x}, \boldsymbol{\xi})} \Big|_{\boldsymbol{\xi}=\mathbf{0}} + \\
 &+ \frac{1}{2} \boldsymbol{\xi}^\top \frac{\partial}{\partial s(\mathbf{x}, \boldsymbol{\xi})} \left[\frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial [s(\mathbf{x}, \boldsymbol{\xi})]^\top} \frac{\partial [s(\mathbf{x}, \boldsymbol{\xi})]^\top}{\partial \boldsymbol{\xi}} \Big|_{\boldsymbol{\xi}=\mathbf{0}} \right] \boldsymbol{\xi} \\
 &= y(\mathbf{x}) + \boldsymbol{\xi}^\top \frac{\partial y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial s(\mathbf{x}, \boldsymbol{\xi})} \Big|_{\boldsymbol{\xi}=\mathbf{0}} + \\
 &+ \frac{1}{2} \boldsymbol{\xi}^\top \left[\frac{\partial^2 y(s(\mathbf{x}, \boldsymbol{\xi}))}{\partial s(\mathbf{x}, \boldsymbol{\xi}) \partial [s(\mathbf{x}, \boldsymbol{\xi})]^\top} \Big|_{\boldsymbol{\xi}=\mathbf{0}} \right] \boldsymbol{\xi} \quad (\text{Apply (5.32)}) \\
 (5.33) \quad y(\mathbf{s}(\mathbf{x}, \boldsymbol{\xi})) &\approx y(\mathbf{x}) + \boldsymbol{\xi}^\top \frac{\partial y(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{2} \boldsymbol{\xi}^\top \frac{\partial^2 y(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} \boldsymbol{\xi} \quad (\text{Apply } \mathbf{s}(\mathbf{x}, \mathbf{0}) = \mathbf{x}).
 \end{aligned}$$

We thereafter substitute (5.33) into $\{y(\mathbf{s}(\mathbf{x}, \xi)) - t\}^2$, with $\nabla y(\mathbf{x}) = \partial y(\mathbf{x})/\partial \mathbf{x}$ and $\nabla \nabla^\top y(\mathbf{x}) = \partial^2 y(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}^\top$, as follows

$$\begin{aligned}
 \{y(\mathbf{s}(\mathbf{x}, \xi)) - t\}^2 &\approx \left\{ y(\mathbf{x}) - t + \xi^\top \nabla y(\mathbf{x}) + \frac{1}{2} \xi^\top \nabla \nabla^\top y(\mathbf{x}) \xi \right\}^2 \\
 &= \{y(\mathbf{x}) - t\}^2 + \{\xi^\top \nabla y(\mathbf{x})\}^2 + \left\{ \frac{1}{2} \xi^\top \nabla \nabla^\top y(\mathbf{x}) \xi \right\}^2 + \\
 &\quad + 2\{y(\mathbf{x}) - t\} \xi^\top \nabla y(\mathbf{x}) + 2\{y(\mathbf{x}) - t\} \left\{ \frac{1}{2} \xi^\top \nabla \nabla^\top y(\mathbf{x}) \xi \right\} + \\
 &\quad + 2\{\xi^\top \nabla y(\mathbf{x})\} \left\{ \frac{1}{2} \xi^\top \nabla \nabla^\top y(\mathbf{x}) \xi \right\} \\
 (5.34) \quad \{y(\mathbf{s}(\mathbf{x}, \xi)) - t\}^2 &\approx \{y(\mathbf{x}) - t\}^2 + \xi^\top \left[\nabla y(\mathbf{x}) \nabla^\top y(\mathbf{x}) + \{y(\mathbf{x}) - t\} \nabla \nabla^\top y(\mathbf{x}) \right] \xi + \\
 &\quad + 2\{y(\mathbf{x}) - t\} \xi^\top \nabla y(\mathbf{x}).
 \end{aligned}$$

Note that, as ξ is assumed to be small, we consider ξ 's of order greater than two as negligible. It follows that, by substituting (5.34) into (5.130), we obtain

$$\begin{aligned}
 \tilde{E} &= \frac{1}{2} \int \int \int \{y(\mathbf{s}(\mathbf{x}, \xi)) - t\}^2 p(t|\mathbf{x}) p(\mathbf{x}) p(\xi) d\mathbf{x} dt d\xi \\
 &\approx \frac{1}{2} \int \int \int \{y(\mathbf{x}) - t\}^2 p(t|\mathbf{x}) p(\mathbf{x}) p(\xi) d\mathbf{x} dt d\xi + \\
 &\quad + \frac{1}{2} \int \int \int \xi^\top \left[\nabla y(\mathbf{x}) \nabla^\top y(\mathbf{x}) + \right. \\
 &\quad \left. + \{y(\mathbf{x}) - t\} \nabla \nabla^\top y(\mathbf{x}) \right] \xi p(t|\mathbf{x}) p(\mathbf{x}) p(\xi) d\mathbf{x} dt d\xi + \\
 &\quad + \int \int \int \{y(\mathbf{x}) - t\} \xi^\top \nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) p(\xi) d\mathbf{x} dt d\xi \quad (\text{Apply (5.34)}) \\
 &= \frac{1}{2} \int \int \{y(\mathbf{x}) - t\}^2 p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt + \\
 &\quad + \frac{1}{2} \int \text{tr} \left(\nabla y(\mathbf{x}) \nabla^\top y(\mathbf{x}) + \right. \\
 &\quad \left. + \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\} \nabla \nabla^\top y(\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x} \quad (\text{Apply (1.30), (1.37), (2.59) and (2.62)}) \\
 (5.35) \quad \tilde{E} &\approx E + \\
 &\quad + \frac{1}{2} \int \|\nabla y(\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \\
 &\quad + \frac{1}{2} \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\} \text{tr}(\nabla \nabla^\top y(\mathbf{x})) p(\mathbf{x}) d\mathbf{x} \quad (\text{Apply (5.193)}).
 \end{aligned}$$

Similarly to previously exposed reasoning, as the third term on the right-hand-side is averaged over $\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}$, E is the squared error function, which is minimized at $y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$ (as seen in Exercise 1.26), we consider it to be negligible. Hence, we obtain

$$(5.36) \quad \tilde{E} \approx E + \frac{1}{2} \int \|\nabla y(\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}.$$

Considering the decomposition (5.127), we obtain the regularizer

$$\Omega = \frac{1}{2} \int \|\nabla y(\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x},$$

thereby verifying (5.135). Note that, in this context we have $\lambda = 1$. By taking $\text{Cov}[\xi] = \lambda I$ we obtain a more general result.

Exercise 5.28

In the convolutional neural network framework we restrict all units within a feature map to share the same weights given an input i . That is, the weight of the contribution of an input i must be the same across all j inputs in the \mathcal{M} -th feature map. Let $a_j^{\mathcal{M}}$ denote the j -th unit in the \mathcal{M} -th feature map, its value is computed as

$$a_j^{\mathcal{M}} = \sum_i w_i^{\mathcal{M}} z_{j,i}^{\mathcal{M}},$$

wherein the sum in i above runs over all units which send connections to the j -th unit in the \mathcal{M} -th feature map. Hence, we may rewrite the derivative (5.53) in the context of convolutional neural networks as

$$\begin{aligned} \frac{\partial E_n}{\partial w_i^{\mathcal{M}}} &= \sum_j \frac{\partial E_n}{\partial a_j^{\mathcal{M}}} \frac{\partial a_j^{\mathcal{M}}}{\partial w_i^{\mathcal{M}}} \\ &= \sum_j \delta_j^{\mathcal{M}} z_{j,i}, \end{aligned}$$

where $\delta_j^{\mathcal{M}} = \partial E_n / \partial a_j^{\mathcal{M}}$.

Exercise 5.29

We aim to verify the validity of (5.141), in the context of soft weight-sharing for invariance in neural networks. It follows that

$$\begin{aligned}
 \frac{\partial \tilde{E}}{\partial w_i} &= \frac{\partial}{\partial w_i} \left[E + \lambda \Omega(\mathbf{w}) \right] && \text{(Apply (5.127))} \\
 &= \frac{\partial E}{\partial w_i} - \lambda \frac{\partial}{\partial w_i} \sum_k \log \left[\sum_{j=1}^M \pi_j p(w_k | \mu_j, \sigma_j^2) \right] && \text{(Apply (5.138))} \\
 &= \frac{\partial E}{\partial w_i} - \lambda \frac{1}{\sum_{j'=1}^M \pi_{j'} p(w_i | \mu_{j'}, \sigma_{j'}^2)} \sum_{j=1}^M \pi_j \frac{\partial p(w_i | \mu_j, \sigma_j^2)}{\partial w_i} \\
 &= \frac{\partial E}{\partial w_i} - \lambda \frac{1}{\sum_{j'=1}^M \pi_{j'} p(w_i | \mu_{j'}, \sigma_{j'}^2)} \times \\
 &\quad \times \sum_{j=1}^M \pi_j \frac{\partial}{\partial w_i} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} \right] && \text{(Apply (1.46))} \\
 &= \frac{\partial E}{\partial w_i} - \lambda \frac{1}{\sum_{j'=1}^M \pi_{j'} p(w_i | \mu_{j'}, \sigma_{j'}^2)} \times \\
 &\quad \times \sum_{j=1}^M \pi_j \frac{(\mu_j - w_i)}{\sigma_j^2 \sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} \\
 &= \frac{\partial E}{\partial w_i} + \lambda \sum_{j=1}^M \frac{(w_i - \mu_j)}{\sigma_j^2} \frac{\pi_j p(w_i | \mu_j, \sigma_j^2)}{\sum_{j'=1}^M \pi_{j'} p(w_i | \mu_{j'}, \sigma_{j'}^2)} && \text{(Apply (1.46))} \\
 \frac{\partial \tilde{E}}{\partial w_i} &= \frac{\partial E}{\partial w_i} + \lambda \sum_{j=1}^M \gamma_j(w_i) \frac{(w_i - \mu_j)}{\sigma_j^2} && \text{(Apply (5.140)).}
 \end{aligned}$$

Hence, we derive (5.141).

Exercise 5.30

We aim to verify the validity of (5.142), in the context of soft weight-sharing for invariance in neural networks. It follows that

$$\begin{aligned}
 \frac{\partial \tilde{E}}{\partial \mu_j} &= \frac{\partial}{\partial \mu_j} \left[E + \lambda \Omega(\mathbf{w}) \right] && \text{(Apply (5.127))} \\
 &= -\lambda \frac{\partial}{\partial \mu_j} \sum_i \log \left[\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2) \right] && \text{(Apply (5.138))} \\
 &= -\lambda \sum_i \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \pi_j \frac{\partial p(w_i | \mu_j, \sigma_j^2)}{\partial \mu_j} \\
 &= -\lambda \sum_i \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \times \\
 &\quad \times \pi_j \frac{\partial}{\partial \mu_j} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} \right] && \text{(Apply (1.46))} \\
 &= \lambda \sum_i \frac{\sigma_j^{-2} (\mu_j - w_i) \pi_j}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} \\
 &= \lambda \sum_i \frac{(\mu_j - w_i)}{\sigma_j^2} \frac{\pi_j p(w_i | \mu_j, \sigma_j^2)}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} && \text{(Apply (1.46))} \\
 \frac{\partial \tilde{E}}{\partial \mu_j} &= \lambda \sum_i \gamma_j(w_i) \frac{(\mu_j - w_i)}{\sigma_j^2} && \text{(Apply (5.140)).}
 \end{aligned}$$

Hence, we derive (5.142).

Exercise 5.31

We aim to verify the validity of (5.143), in the context of soft weight-sharing for invariance in neural networks. It follows that

$$\begin{aligned}
 \frac{\partial \tilde{E}}{\partial \sigma_j} &= \frac{\partial}{\partial \sigma_j} \left[E + \lambda \Omega(\mathbf{w}) \right] && \text{(Apply (5.127))} \\
 &= -\lambda \frac{\partial}{\partial \sigma_j} \sum_i \log \left[\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2) \right] && \text{(Apply (5.138))} \\
 &= -\lambda \sum_i \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \pi_j \frac{\partial p(w_i | \mu_j, \sigma_j^2)}{\partial \sigma_j} \\
 &= -\lambda \sum_i \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \times \\
 &\quad \times \pi_j \frac{\partial}{\partial \sigma_j} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} \right] && \text{(Apply (1.46))} \\
 &= -\lambda \sum_i \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \times \\
 &\quad \times \pi_j \left[\frac{1}{\sigma_j^3} \frac{(w_i - \mu_j)^2}{\sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} + \right. \\
 &\quad \left. - \frac{1}{\sigma_j \sqrt{2\pi\sigma_j^2}} \exp \left\{ -\frac{(w_i - \mu_j)^2}{2\sigma_j^2} \right\} \right] \\
 &= \lambda \sum_i \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \times \\
 &\quad \times \pi_j p(w_i | \mu_j, \sigma_j^2) \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) && \text{(Apply (1.46))} \\
 \frac{\partial \tilde{E}}{\partial \sigma_j} &= \lambda \sum_i \gamma_j(w_i) \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) && \text{(Apply (5.140)).}
 \end{aligned}$$

Hence, we derive (5.143).

Exercise 5.32

We aim to verify the validity of (5.147), in the context of soft weight-sharing for invariance in neural networks. For that purpose, we aim to verify that (5.208) is satisfied, for π_j as defined in (5.146). Note that the form (5.146) is essentially the same as the softmax activation function form in (4.104), hence demonstrating the validity of (5.208) is tantamount to determining the derivative of (4.104), which as seen in Exercise 4.17 is (4.106), hence (5.208) is valid. It follows that

$$\begin{aligned}
 \frac{\partial \tilde{E}}{\partial \eta_j} &= \frac{\partial}{\partial \eta_j} \left[E + \lambda \Omega(\mathbf{w}) \right] && \text{(Apply (5.127))} \\
 &= -\lambda \frac{\partial}{\partial \eta_j} \sum_i \log \left[\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2) \right] && \text{(Apply (5.138))} \\
 &= -\lambda \sum_i \sum_{k'=1}^M \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} p(w_i | \mu_{k'}, \sigma_{k'}^2) \frac{\partial \pi_{k'}}{\partial \eta_j} \\
 &= -\lambda \sum_i \sum_{k'=1}^M \frac{1}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} p(w_i | \mu_{k'}, \sigma_{k'}^2) (I_{j,k'} \pi_j - \pi_j \pi_{k'}) && \text{(Apply (5.208))} \\
 &= \lambda \sum_i \sum_{k'=1}^M \left[\pi_j \gamma_{k'}(w_i) - \frac{p(w_i | \mu_{k'}, \sigma_{k'}^2) I_{j,k'} \pi_j}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \right] && \text{(Apply (5.140))} \\
 &= \lambda \sum_i \left[\pi_j - \frac{p(w_i | \mu_j, \sigma_j^2)}{\sum_{k=1}^M \pi_k p(w_i | \mu_k, \sigma_k^2)} \right] && \text{(Apply } \sum_{k=1}^M \gamma_k(w) = 1\text{)} \\
 \frac{\partial \tilde{E}}{\partial \eta_j} &= \lambda \sum_i \{ \pi_j - \gamma_j(w_i) \} && \text{(Apply (5.140)).}
 \end{aligned}$$

Hence, we derive (5.147).

Exercise 5.33

Figure 5.2 provides an idea of how we may determine the position of x as a function of the joint angles θ_1 and θ_2 , and arm lengths L_1 and L_2 : utilizing (y_1, y_2) as an intermediate point (the end point of the first arm), we first determine its form as

$$(y_1, y_2) = (L_1 \sin \alpha, L_1 \cos \alpha).$$

We thereafter adopt (y_1, y_2) as the origin of the second arm, and the value of (x_1, x_2) is obtained as

$$(5.37) \quad (x_1, x_2) = (L_2 \cos \gamma + L_1 \sin \alpha, L_2 \sin \gamma + L_1 \cos \alpha).$$

It suffices now to determine the values of α and γ . Note that $\beta + \beta + \theta_1 + \theta_1 = 2\pi$, as these angles form a parallelogram, hence $\beta = \pi - \theta_1$. It follows that

$$(5.38) \quad \begin{aligned} \theta_1 &= \alpha + \frac{\pi}{2} \\ \alpha &= \frac{\pi}{2} - \theta_1, \end{aligned}$$

and

$$(5.39) \quad \begin{aligned} \theta_2 &= \gamma + \beta \\ &= \gamma + (\pi - \theta_1) \\ \gamma &= \theta_1 + \theta_2 - \pi. \end{aligned}$$

Substituting (5.38) and (5.39) into (5.37), we find that

$$\begin{aligned} (x_1, x_2) &= (L_2 \cos \gamma + L_1 \sin \alpha, L_2 \sin \gamma + L_1 \cos \alpha) \\ &= (L_2 \cos(\theta_1 + \theta_2 - \pi) + L_1 \sin(\pi/2 - \theta_1), \\ &\quad L_2 \sin(\theta_1 + \theta_2 - \pi) + L_1 \cos(\pi/2 - \theta_1)) \\ (x_1, x_2) &= (L_1 \cos(\theta_1) - L_2 \cos(\theta_1 + \theta_2), L_1 \sin(\theta_1) - L_2 \sin(\theta_1 + \theta_2)), \end{aligned}$$

hence deriving the forward kinematics of the robot arm.

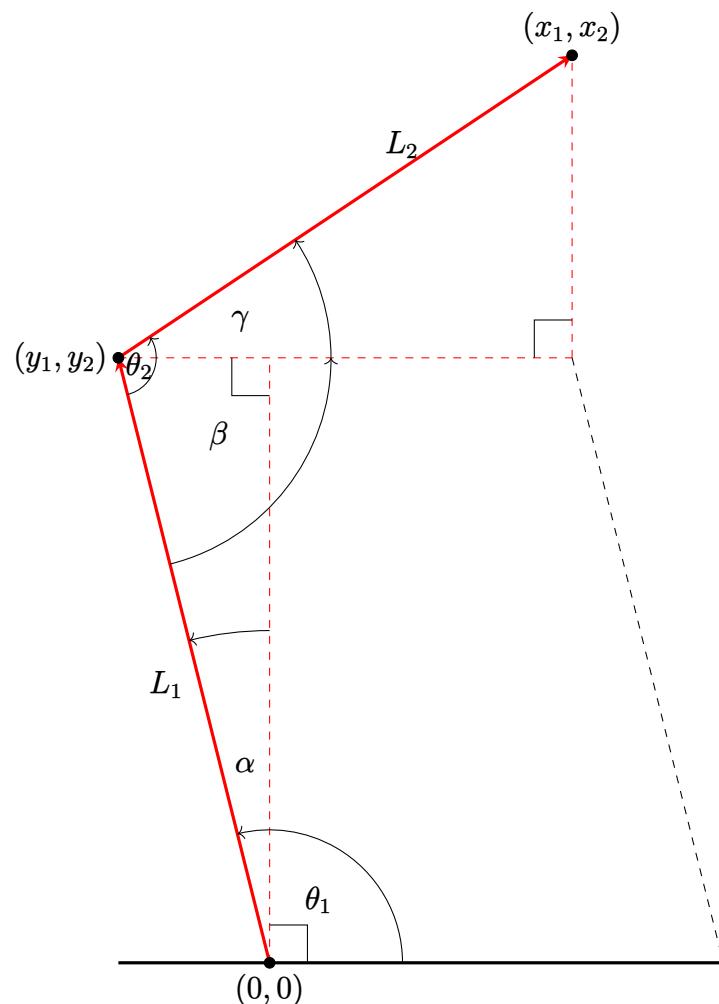


Figure 5.2: Illustration of a two-link robot arm.

Exercise 5.34

We aim to verify (5.155) in the context of neural density networks. Note that the form of the dependence of the mixing coefficients on the activation units draws from the softmax function in (4.104). It follows that

$$\begin{aligned}
 \frac{\partial E_n}{\partial a_k^\pi} &= \frac{\partial}{\partial a_k^\pi} \left[-\log \left\{ \sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2) \right\} \right] && \text{(Apply (5.153))} \\
 &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \sum_{j'=1}^K \frac{\partial \pi_{j'}}{\partial a_k^\pi} p(\mathbf{t}_n | \boldsymbol{\mu}_{j'}, \sigma_{j'}^2) \\
 &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \sum_{j'=1}^K (I_{k,j'} \pi_{j'} - \pi_k \pi_{j'}) p(\mathbf{t}_n | \boldsymbol{\mu}_{j'}, \sigma_{j'}^2) \\
 &= -\frac{\pi_k p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2) - \pi_k \sum_{j'=1}^K \pi_{j'} p(\mathbf{t}_n | \boldsymbol{\mu}_{j'}, \sigma_{j'}^2)}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} && \text{(Apply (4.106))} \\
 \frac{\partial E_n}{\partial a_k^\pi} &= \pi_k - \gamma_{n,k} && \text{(Apply (5.154)).}
 \end{aligned}$$

Hence, we verify (5.155).

Exercise 5.35

We aim to verify (5.156) in the context of neural density networks. Note that the form of the dependence of the mean components on the activation units is linear. It follows that

$$\begin{aligned}
 \frac{\partial E_n}{\partial a_{k,l}^\mu} &= \frac{\partial}{\partial a_{k,l}^\mu} \left[-\log \left\{ \sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2) \right\} \right] && \text{(Apply (5.153))} \\
 &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \pi_k \frac{\partial p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2)}{\partial a_{k,l}^\mu} \\
 &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \times \\
 &\quad \times \pi_k \frac{\partial}{\partial a_{k,l}^\mu} \left[(2\pi\sigma_k^2)^{-L/2} \exp \left\{ -\frac{1}{2\sigma_k^2} (\mathbf{t}_n - \boldsymbol{\mu}_k)^\top (\mathbf{t}_n - \boldsymbol{\mu}_k) \right\} \right] && \text{(Apply (2.43))} \\
 &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \times \\
 &\quad \times \pi_k \frac{\partial}{\partial a_{k,l}^\mu} \left[(2\pi\sigma_k^2)^{-L/2} \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{l'=1}^L (t_{n,l'} - \mu_{k,l'})^2 \right\} \right] \\
 &= \frac{\mu_{k,l} - t_{n,l}}{\sigma_k^2 \sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \times \\
 &\quad \times \pi_k (2\pi\sigma_k^2)^{-L/2} \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{l'=1}^L (t_{n,l'} - \mu_{k,l'})^2 \right\} \\
 &= \frac{\mu_{k,l} - t_{n,l}}{\sigma_k^2 \sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \pi_k p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2) && \text{(Apply (2.43))} \\
 \frac{\partial E_n}{\partial a_{k,l}^\mu} &= \gamma_{n,k} \left\{ \frac{\mu_{k,l} - t_{n,l}}{\sigma_k^2} \right\} && \text{(Apply (5.154)).}
 \end{aligned}$$

Hence, we verify (5.156).

Exercise 5.36

We aim to verify (5.157) in the context of neural density networks. Note that the form of the dependence of the variance components on the activation units is (5.151), such that

$$(5.40) \quad \begin{aligned} \frac{\partial \sigma_k(\mathbf{x})}{\partial a_k^\sigma} &= \frac{\partial \exp\{a_k^\sigma\}}{\partial a_k^\sigma} \\ &= \exp\{a_k^\sigma\} \\ \frac{\partial \sigma_k(\mathbf{x})}{\partial a_k^\sigma} &= \sigma_k(\mathbf{x}). \end{aligned}$$

It follows that

$$\begin{aligned} \frac{\partial E_n}{\partial a_k^\sigma} &= \frac{\partial}{\partial a_k^\sigma} \left[-\log \left\{ \sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2) \right\} \right] && \text{(Apply (5.153))} \\ &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \pi_k \frac{\partial p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2)}{\partial a_k^\sigma} \\ &= -\frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \pi_k \frac{\partial p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2)}{\partial \sigma_k} \frac{\partial \sigma_k}{\partial a_k^\sigma} \\ &= -\frac{\sigma_k}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \pi_k \frac{\partial p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2)}{\partial \sigma_k} && \text{(Apply (5.40))} \\ &= -\frac{\sigma_k}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \times \\ &\quad \times \pi_k \frac{\partial}{\partial \sigma_k} \left[(2\pi\sigma_k^2)^{-L/2} \exp \left\{ -\frac{1}{2\sigma_k^2} (\mathbf{t}_n - \boldsymbol{\mu}_k)^\top (\mathbf{t}_n - \boldsymbol{\mu}_k) \right\} \right] && \text{(Apply (2.43))} \\ &= -\frac{\sigma_k}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \times \\ &\quad \times \pi_k \left[\frac{\|\mathbf{t}_n - \boldsymbol{\mu}\|^2}{\sigma_k^3} (2\pi\sigma_k^2)^{-L/2} \exp \left\{ -\frac{1}{2\sigma_k^2} (\mathbf{t}_n - \boldsymbol{\mu}_k)^\top (\mathbf{t}_n - \boldsymbol{\mu}_k) \right\} + \right. \\ &\quad \left. - L\sigma_k^{-1} (2\pi\sigma_k^2)^{-L/2} \exp \left\{ -\frac{1}{2\sigma_k^2} (\mathbf{t}_n - \boldsymbol{\mu}_k)^\top (\mathbf{t}_n - \boldsymbol{\mu}_k) \right\} \right] \\ &= \frac{1}{\sum_{j=1}^K \pi_j p(\mathbf{t}_n | \boldsymbol{\mu}_j, \sigma_j^2)} \left[L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}\|^2}{\sigma_k^2} \right] \pi_k p(\mathbf{t}_n | \boldsymbol{\mu}_k, \sigma_k^2) && \text{(Apply (2.43))} \\ \frac{\partial E_n}{\partial a_k^\sigma} &= \gamma_{n,k} \left\{ L - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}\|^2}{\sigma_k^2} \right\} && \text{(Apply (5.154)).} \end{aligned}$$

Hence, we verify (5.157).

Exercise 5.37

We aim to verify the form of the conditional expectation and covariance, respectively (5.158) and (5.160), in the neural density network context. Firstly, the conditional expectation is easily obtained as

$$\begin{aligned}
 \mathbb{E}[\mathbf{t}|\mathbf{x}] &= \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} && \text{(Apply (1.37))} \\
 &= \int \mathbf{t} \sum_{k=1}^K \pi_k(\mathbf{x}) p(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})) d\mathbf{t} && \text{(Apply (5.148))} \\
 &= \sum_{k=1}^K \pi_k(\mathbf{x}) \int \mathbf{t} p(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})) d\mathbf{t} \\
 (5.41) \quad \mathbb{E}[\mathbf{t}|\mathbf{x}] &= \sum_{k=1}^K \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}) && \text{(Apply (2.59)).}
 \end{aligned}$$

Hence we verify (5.158). We now aim to verify (5.160), as follows

$$\begin{aligned}
 s^2(\mathbf{x}) &= \mathbb{E}[||\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]||^2|\mathbf{x}] \\
 &= \mathbb{E}[(\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}])^\top (\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}])|\mathbf{x}] \\
 &= \mathbb{E}[\mathbf{t}^\top \mathbf{t} - 2\mathbf{t}^\top \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}^\top |\mathbf{x}] \mathbb{E}[\mathbf{t}|\mathbf{x}]|\mathbf{x}] \\
 &= \int \{\mathbf{t}^\top \mathbf{t} - 2\mathbf{t}^\top \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}^\top |\mathbf{x}] \mathbb{E}[\mathbf{t}|\mathbf{x}]\} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} && \text{(Apply (1.37))} \\
 &= \int \{\mathbf{t}^\top \mathbf{t} - 2\mathbf{t}^\top \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}^\top |\mathbf{x}] \mathbb{E}[\mathbf{t}|\mathbf{x}]\} \sum_{k=1}^K \pi_k(\mathbf{x}) p(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})) d\mathbf{t} && \text{(Apply (5.148))} \\
 &= \sum_{k=1}^K \pi_k(\mathbf{x}) \int \{\mathbf{t}^\top \mathbf{t} - 2\mathbf{t}^\top \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}^\top |\mathbf{x}] \mathbb{E}[\mathbf{t}|\mathbf{x}]\} p(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})) d\mathbf{t} \\
 &= \sum_{k=1}^K \pi_k(\mathbf{x}) \{\boldsymbol{\mu}_k^\top(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}) + L\sigma_k^2(\mathbf{x}) - 2\boldsymbol{\mu}_k^\top(\mathbf{x}) \mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}^\top |\mathbf{x}] \mathbb{E}[\mathbf{t}|\mathbf{x}]\} && \text{(Apply (2.59) and (2.62))} \\
 &= \sum_{k=1}^K \pi_k(\mathbf{x}) \{L\sigma_k^2(\mathbf{x}) + (\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}])^\top (\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}])\} \\
 &= \sum_{k=1}^K \pi_k(\mathbf{x}) \{L\sigma_k^2(\mathbf{x}) + ||\boldsymbol{\mu}_k(\mathbf{x}) - \mathbb{E}[\mathbf{t}|\mathbf{x}]||^2\} \\
 s^2(\mathbf{x}) &= \sum_{k=1}^K \pi_k(\mathbf{x}) \left\{ L\sigma_k^2(\mathbf{x}) + \left\| \boldsymbol{\mu}_k(\mathbf{x}) - \sum_{l=1}^K \pi_l(\mathbf{x}) \boldsymbol{\mu}_l(\mathbf{x}) \right\|^2 \right\} && \text{(Apply (5.41)).}
 \end{aligned}$$

Thereby obtaining the desired result.

Exercise 5.38

Consider the Bayesian neural network estimation context, particular wherein we seek to determine the predictive distribution of a new observation t , given a training data set \mathcal{D} , and that we have chosen as prior distribution over the weights the form (3.52), thereafter obtaining an approximate posterior according to (5.167). Moreover, let the conditional distribution $p(t|\mathbf{x}, \mathbf{w}, \beta)$ be approximated as (5.171). By application of the Gaussian linear model result (2.115), with (5.167) in (2.113) and (5.171) in (2.114), we obtain a mean function of the form

$$\begin{aligned}\mathbb{E}[t|\mathbf{x}, \mathcal{D}] &\approx y(\mathbf{w}, \mathbf{w}_{\text{MAP}}) + \mathbf{g}^\top \mathbf{w}_{\text{MAP}} - \mathbf{g}^\top \mathbf{w}_{\text{MAP}} \\ &= y(\mathbf{x}, \mathbf{w}_{\text{MAP}}),\end{aligned}$$

and input-dependent variance as in

$$\text{Var}[t|\mathbf{x}, \mathcal{D}] = \beta^{-1} + \mathbf{g}^\top \mathbf{A}^{-1} \mathbf{g}.$$

Thereby obtaining the desired result.

Exercise 5.39

Consider that, in the context of Bayesian neural networks, we adopt herein the Laplace approximation to the model evidence, as seen in (4.135), to determine an approximate form of $p(\mathcal{D}|\alpha, \beta)$, setting $\mathbf{z}_0 = \mathbf{w}_{\text{MAP}}$, $f(\mathbf{z}_0) = p(\mathcal{D}|\mathbf{z}_0, \beta)p(\mathbf{z}_0|\alpha)$, the total number of parameters as W , and \mathbf{A} as

$$(5.42) \quad \mathbf{A} = -\nabla\nabla \log\{p(\mathcal{D}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\} \Big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}.$$

Note that, as $p(\mathbf{w}|\mathcal{D}, \alpha, \beta) \propto p(\mathcal{D}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)$, it follows from (5.42) that

$$\mathbf{A} = -\nabla\nabla \log p(\mathbf{w}|\mathcal{D}, \alpha, \beta) \Big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}.$$

Hence, by applying (4.135), we obtain

$$\begin{aligned} p(\mathcal{D}|\alpha, \beta) &\approx p(\mathcal{D}|\mathbf{w}_{\text{MAP}}, \beta)p(\mathbf{w}_{\text{MAP}}|\alpha) \frac{(2\pi)^{W/2}}{|\mathbf{A}|} \\ \log p(\mathcal{D}|\alpha, \beta) &\approx \log p(\mathcal{D}|\mathbf{w}_{\text{MAP}}, \beta) + \log p(\mathbf{w}_{\text{MAP}}|\alpha) + \\ &\quad - \frac{1}{2} \log |\mathbf{A}| + \frac{W}{2} \log(2\pi) \\ &= -\frac{\beta}{2} \sum_{n=1}^N \{y(\mathbf{x}_n, \mathbf{w}_{\text{MAP}}) - t_n\}^2 - \frac{N}{2} \log(2\pi) + \\ &\quad + \frac{N}{2} \log \beta - \frac{\alpha}{2} \mathbf{w}_{\text{MAP}}^\top \mathbf{w}_{\text{MAP}} - \frac{W}{2} \log(2\pi) + \\ &\quad + \frac{W}{2} \log \alpha - \frac{1}{2} \log |\mathbf{A}| + \frac{W}{2} \log(2\pi) \quad (\text{Apply (1.46)}) \\ p(\mathcal{D}|\alpha, \beta) &\approx -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2} |\mathbf{A}| + \frac{W}{2} \log \alpha + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) \quad (\text{Apply (5.176)}). \end{aligned}$$

Hence, we obtain the desired approximation to the evidence function.

Exercise 5.40

Consider that, in the context of classification with Bayesian neural networks, we aim to determine a framework for multiclass classification procedures, with our network having softmax output unit activation functions. It follows that the corresponding error function would be of the form (5.24). Hence, we could determine a multiclass procedure by modifying the binary classification procedure as follows: in initializing the hyperparameters α , the regularized error function would present a component of the form (5.24), as opposed to (5.23), from which we would determine \mathbf{w}_{MAP} . For the Laplace approximation of the posterior distribution of \mathbf{w} , as well as the model evidence $p(\mathcal{D}|\alpha)$ - and subsequent hyperparameter optimization - the procedure would remain similar to the the binary classification procedure, however the error function Hessian \mathbf{H} (whether exact or approximated) would be modified to the Hessian of (5.24). The predictive distribution of a new observation t , however, may not be as simply extended. Note that, in the binary case, an approximate distribution for the output unit activation function is determined, and thereafter we compute $p(t = 1|\mathbf{x}, \mathcal{D}) \approx \int \sigma(a)p(a|\mathbf{x}, \mathcal{D}) da$, a integral whose evaluation is made possible via the approximate convolution of a Gaussian with a logistic sigmoid (see [Exercise 4.25](#) and [Exercise 4.26](#)). While we may still compute the approximate distribution of the output unit activations similarly, the convolution of the softmax function with the distribution of the output unit activations is not as simply estimated, potentially requiring other approximations.

Exercise 5.41

Consider that, in the context of classification with Bayesian neural networks, we seek to obtain the expression (5.183) as the Laplace approximation to the model evidence. Let a data set $\mathcal{D} = \{\mathbf{x}_n, t_n\}_{n=1}^N$, where $t_n \in \{0, 1\}$, be observed, and assume that we attribute to our weights a prior distribution as in (3.52). We then determine the value of \mathbf{w}_{MAP} as that which minimizes the following

$$(5.43) \quad \begin{aligned} \mathbf{w}_{\text{MAP}} &= \arg \min_{\mathbf{w}} E(\mathbf{w}) \\ &= \arg \min_{\mathbf{w}} \left\{ -\log p(\mathcal{D}|\mathbf{w}) + \alpha \frac{\mathbf{w}^\top \mathbf{w}}{2} \right\}. \end{aligned}$$

Where $-\log p(\mathcal{D}|\mathbf{w})$ is as in (5.23). We thereafter define \mathbf{A} as

$$\mathbf{A} = -\nabla \nabla \log \{p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\alpha)\} \Big|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}.$$

We can now apply Laplace's approximation to the model evidence, as seen in (4.135), setting $z_0 = \mathbf{w}_{\text{MAP}}$, $f(z_0) = p(\mathcal{D}|\mathbf{z}_0)p(\mathbf{z}_0|\alpha)$, with total number of parameters given as W , obtaining

$$\begin{aligned} p(\mathcal{D}|\alpha) &\approx p(\mathcal{D}|\mathbf{w}_{\text{MAP}})p(\mathbf{w}_{\text{MAP}}|\alpha) \frac{(2\pi)^{W/2}}{|\mathbf{A}|} \\ \log p(\mathcal{D}|\alpha) &\approx \log p(\mathcal{D}|\mathbf{w}_{\text{MAP}}) + \log p(\mathbf{w}_{\text{MAP}}|\alpha) - \frac{1}{2} \log |\mathbf{A}| + \frac{W}{2} \log(2\pi) \\ &= \sum_{n=1}^N \{t_n \log y_n + (1-t_n) \log(1-y_n)\} - \frac{\alpha}{2} \mathbf{w}_{\text{MAP}}^\top \mathbf{w}_{\text{MAP}} + \\ &\quad - \frac{W}{2} \log(2\pi) + \frac{W}{2} \log \alpha - \frac{1}{2} \log |\mathbf{A}| + \frac{W}{2} \log(2\pi) \\ p(\mathcal{D}|\alpha) &\approx -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2} \log |\mathbf{A}| + \frac{W}{2} \log \alpha \end{aligned} \quad (\text{Apply (5.184)}).$$

Hence, we derive (5.183).

Chapter 6

Kernel Methods

Exercise 6.1

We aim to demonstrate herein that the solution \mathbf{a} of the dual formulation least squares linear regression problem is such that \mathbf{a} is a linear combination of $\{\phi_j(\mathbf{x})\}_{j=1}^M$. We follow the same procedure as is seen in [Exercise 6.16](#): first, we rewrite \mathbf{a} as

$$(6.1) \quad \begin{aligned} \mathbf{a} &= \sum_{j=1}^M u_j \phi_j(\mathbf{x}) + \mathbf{a}_\perp \\ \mathbf{a} &= \Phi \mathbf{u} + \mathbf{a}_\perp, \end{aligned}$$

where $\mathbf{a}_\perp^\top \phi_j(\mathbf{x}) = 0, \forall j \in \{1, \dots, M\}$. Note that

$$(6.2) \quad \begin{aligned} \Phi &= (\phi_1(\mathbf{x}) \ \dots \ \phi_M(\mathbf{x})) \\ \mathbf{a}^\top \Phi &= (0 \ \dots \ 0). \end{aligned}$$

It follows that, substituting (6.1) into (6.5), we obtain

$$\begin{aligned} J(\mathbf{a}) &= \frac{1}{2} \mathbf{a}^\top \Phi \Phi^\top \Phi \Phi^\top \mathbf{a} - \mathbf{a}^\top \Phi \Phi^\top \mathbf{t} + \frac{1}{2} \mathbf{t}^\top \mathbf{t} + \frac{\lambda}{2} \mathbf{a}^\top \Phi \Phi^\top \mathbf{a} \\ J(\mathbf{u}, \mathbf{a}_\perp) &= \frac{1}{2} \{\Phi \mathbf{u} + \mathbf{a}_\perp\}^\top \Phi \Phi^\top \Phi \Phi^\top \{\Phi \mathbf{u} + \mathbf{a}_\perp\} + \\ &\quad - \{\Phi \mathbf{u} + \mathbf{a}_\perp\}^\top \Phi \Phi^\top \mathbf{t} + \frac{1}{2} \mathbf{t}^\top \mathbf{t} + \frac{\lambda}{2} \{\Phi \mathbf{u} + \mathbf{a}_\perp\}^\top \Phi \Phi^\top \{\Phi \mathbf{u} + \mathbf{a}_\perp\} \quad (\text{Apply (6.1)}) \\ &= \frac{1}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \Phi^\top \Phi \mathbf{u} - \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \mathbf{t} + \frac{1}{2} \mathbf{t}^\top \mathbf{t} + \frac{\lambda}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \mathbf{u} + \\ &\quad + \frac{1}{2} \mathbf{a}_\perp^\top \Phi^\top \Phi \Phi^\top \Phi \Phi^\top \Phi \mathbf{a}_\perp - \mathbf{a}_\perp^\top \Phi^\top \Phi \Phi^\top \mathbf{t} + \frac{\lambda}{2} \mathbf{a}_\perp^\top \Phi^\top \Phi \Phi^\top \Phi \mathbf{a}_\perp + \\ &\quad + \frac{1}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \Phi^\top \Phi \mathbf{a}_\perp + \frac{\lambda}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \mathbf{a}_\perp + \\ &\quad + \frac{1}{2} \mathbf{a}_\perp^\top \Phi^\top \Phi \Phi^\top \Phi \Phi^\top \Phi \mathbf{u} + \frac{\lambda}{2} \mathbf{a}_\perp^\top \Phi^\top \Phi \Phi^\top \Phi \mathbf{u} \\ &= \frac{1}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \Phi^\top \Phi \mathbf{u} - \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \mathbf{t} + \frac{1}{2} \mathbf{t}^\top \mathbf{t} + \frac{\lambda}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \mathbf{u} \quad (\text{Apply (6.2)}) \\ (6.3) \quad J(\mathbf{u}, \mathbf{a}_\perp) &= \frac{1}{2} \{\Phi \Phi^\top \Phi \mathbf{u} - \mathbf{t}\}^\top \{\Phi \Phi^\top \Phi \mathbf{u} - \mathbf{t}\} + \frac{\lambda}{2} \mathbf{u}^\top \Phi^\top \Phi \Phi^\top \Phi \mathbf{u}. \end{aligned}$$

We observe from (6.3) that the regularized sum-of-squares error function is not dependent on \mathbf{a}_\perp . Hence, when minimizing $J(\mathbf{u}, \mathbf{a}_\perp)$ with respect to \mathbf{u} and \mathbf{a}_\perp (which is equivalent to minimizing $J(\mathbf{a})$ with respect to \mathbf{a} , given (6.1)) we may choose, or otherwise enforce, without loss of generality, that $\mathbf{a}_\perp = \mathbf{0}$. Consequently, returning to (6.1), we have that

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

Therefore, \mathbf{a} is a linear combination of $\{\phi_j(\mathbf{x})\}_{j=1}^M$. We define $\mathbf{w} = \Phi^\top \Phi \mathbf{u}$, and rewrite such in (6.3), yielding

$$\mathbf{J}(\mathbf{w}) = \frac{1}{2} \{\Phi \mathbf{w} - \mathbf{w}\}^\top \{\Phi \mathbf{w} - \mathbf{w}\} + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w}.$$

Hence, we show that the dual of the dual formulation returns to the original formulation of the least-squares linear regression problem.

Exercise 6.2

We aim to develop a dual formulation for the perceptron learning algorithm, recasting it in the context of kernels methods. For that purpose, first we rewrite the learning rule in (4.55) (adopting $\eta = 1$ for convenience) as

$$(6.4) \quad \begin{aligned} \mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} + \sum_{n \in \mathcal{M}} \phi_n t_n \\ \mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} + \sum_{n=1}^N \beta_n^{(\tau)} \phi_n t_n, \end{aligned}$$

where

$$(6.5) \quad \beta_n^{(\tau)} = \begin{cases} 0 & \text{if } (\mathbf{w}^{(\tau)})^\top \phi_n t_n > 0, \\ 1 & \text{if } (\mathbf{w}^{(\tau)})^\top \phi_n t_n \leq 0. \end{cases}$$

Intuitively, (6.5) implies that we have $\beta_n^{(\tau)} = 1$ if the n -th data point is misclassified at the τ -th step of the weight estimating procedure, and $\beta_n^{(\tau)} = 0$ otherwise. We apply (6.4) recursively to the right-hand-side of (6.4), obtaining

$$(6.6) \quad \begin{aligned} \mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau-1)} + \sum_{n=1}^N \beta_n^{(\tau-1)} \phi_n t_n + \sum_{n=1}^N \beta_n^{(\tau)} \phi_n t_n \\ &= \mathbf{w}^{(0)} + \sum_{n=1}^N \sum_{l=0}^{\tau} \beta_n^{(l)} \phi_n t_n \\ \mathbf{w}^{(\tau+1)} &= \sum_{n=1}^N \sum_{l=0}^{\tau} \beta_n^{(l)} \phi_n t_n, \end{aligned}$$

where, for convenience, we have adopted $\mathbf{w}^{(0)} = \mathbf{0}$ as the initial point for the estimating procedure. We may thereby define the coefficients

$$(6.7) \quad \alpha_n^{(\tau)} = \sum_{l=0}^{\tau} \beta_n^{(l)}.$$

Substituting (6.7) into (6.6), we obtain

$$(6.8) \quad \mathbf{w}^{(\tau+1)} = \sum_{n=1}^N \alpha_n^{(\tau)} \phi_n t_n.$$

Applying (6.7) in the right-hand-side of (6.7), we obtain

$$\begin{aligned} \alpha_n^{(\tau)} &= \beta_n^{(\tau)} + \sum_{l=0}^{\tau-1} \beta_n^{(l)} \\ &= \beta_n^{(\tau)} + \alpha_n^{(\tau-1)} && \text{(Apply (6.7))} \\ &= \begin{cases} \alpha_n^{(\tau-1)} & \text{if } (\mathbf{w}^{(\tau)})^\top \phi_n t_n > 0, \\ 1 + \alpha_n^{(\tau-1)} & \text{if } (\mathbf{w}^{(\tau)})^\top \phi_n t_n \leq 0. \end{cases} && \text{(Apply (6.5))} \\ &= \begin{cases} \alpha_n^{(\tau-1)} & \text{if } \sum_{j=1}^N \alpha_j^{(\tau)} \phi_j^\top \phi_n t_n > 0, \\ 1 + \alpha_n^{(\tau-1)} & \text{if } \sum_{j=1}^N \alpha_j^{(\tau)} \phi_j^\top \phi_n t_n \leq 0. \end{cases} && \text{(Apply (6.8))} \\ \alpha_n^{(\tau)} &= \begin{cases} \alpha_n^{(\tau-1)} & \text{if } \sum_{j=1}^N \alpha_j^{(\tau)} k(\mathbf{x}_j, \mathbf{x}_n) t_n > 0, \\ 1 + \alpha_n^{(\tau-1)} & \text{if } \sum_{j=1}^N \alpha_j^{(\tau)} k(\mathbf{x}_j, \mathbf{x}_n) t_n \leq 0. \end{cases} && \text{(Apply (6.1))} \end{aligned}$$

This provides a simple formulation for the learning procedure of the coefficients $\alpha^{(\tau)}$: at the τ -th step, we attribute the class of the n -th data point as the sign of $\sum_{j=1}^N \alpha_j^{(\tau)} k(\mathbf{x}_j, \mathbf{x}_n)$. If this results in a misclassification, we define $\alpha^{(\tau+1)} = \alpha^{(\tau)} + 1$, otherwise $\alpha^{(\tau+1)} = \alpha^{(\tau)}$. Hence, $\alpha^{(\tau)}$ acts as a counter for the number of times the n -th data point is misclassified during the learning procedure (starting at the zeroth step until the $(\tau - 1)$ -th step). For every step, we run through all data points only once. We assume now that the data is linearly separable, such that this procedure converges at the M -th step (hence, at a certain point $\alpha_n^{(\tau)} = \alpha_n, \forall n, \forall \tau \geq M$). We construct our prediction of a new observation as in (4.53) as

$$\begin{aligned} f(\mathbf{x}) &= \begin{cases} 1 & \text{if } \mathbf{w}^\top \phi(\mathbf{x}) \geq 0, \\ -1 & \text{if } \mathbf{w}^\top \phi(\mathbf{x}) < 0. \end{cases} \\ &= \begin{cases} 1 & \text{if } \sum_{j=1}^N \alpha_j \phi^\top(\mathbf{x}_j) \phi(\mathbf{x}) \geq 0, \\ -1 & \text{if } \sum_{j=1}^N \alpha_j \phi^\top(\mathbf{x}_j) \phi(\mathbf{x}) < 0. \end{cases} \quad (\text{Apply (6.8)}) \\ f(\mathbf{x}) &= \begin{cases} 1 & \text{if } \sum_{j=1}^N \alpha_j k(\mathbf{x}_j, \mathbf{x}) \geq 0, \\ -1 & \text{if } \sum_{j=1}^N \alpha_j k(\mathbf{x}_j, \mathbf{x}) < 0. \end{cases} \quad (\text{Apply (6.1)}) \end{aligned}$$

Exercise 6.3

Consider the context of a nearest neighbour model, such that we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, where \mathbf{x}_n are our input variables and t_n are our target variables. If we adopt the Euclidean distance metric as a criteria for classification of a new observation \mathbf{x} , we define its corresponding output according to the following rule

$$(6.9) \quad f(\mathbf{x}) = t_{n^*} \quad \text{where} \quad n^* = \arg \min_n \|\mathbf{x} - \mathbf{x}_n\|^2.$$

We now attempt to rewrite the criteria as a valid kernel. Note that the argument that minimizes $h(s)$ is the same as that which maximizes $-h(s)$. Additionally, as $g(s) = \exp\{s\}$ is a strictly increasing monotone function, we find that the argument which maximizes $-h(s)$ must also maximize $\exp\{-h(s)\}$. We thereby rewrite (6.9) as

$$(6.10) \quad f(\mathbf{x}) = t_{n^*} \quad \text{where} \quad n^* = \arg \max_n \exp\{-\|\mathbf{x} - \mathbf{x}_n\|^2\}.$$

We rewrite then rewrite

$$\begin{aligned} \exp\{-\|\mathbf{x} - \mathbf{x}_n\|^2\} &= \exp\{-(\mathbf{x} - \mathbf{x}_n)^\top (\mathbf{x} - \mathbf{x}_n)\} \\ &= \exp\{-\mathbf{x}^\top \mathbf{x} + 2\mathbf{x}^\top \mathbf{x}_n - \mathbf{x}_n^\top \mathbf{x}_n\} \\ &= \exp\{-\mathbf{x}^\top \mathbf{x}\} \exp\{2\mathbf{x}^\top \mathbf{x}_n\} \exp\{-\mathbf{x}_n^\top \mathbf{x}_n\} \\ &= \psi(\mathbf{x}) \exp\{2\mathbf{x}^\top \mathbf{x}_n\} \psi(\mathbf{x}_n) \\ &= \psi(\mathbf{x}) \exp\{k_a(\mathbf{x}, \mathbf{x}_n)\} \psi(\mathbf{x}_n) \quad (\text{Apply (6.1)}) \\ &= \psi(\mathbf{x}) k_b(\mathbf{x}, \mathbf{x}_n) \psi(\mathbf{x}_n) \quad (\text{Apply (6.16)}) \\ (6.11) \quad \exp\{-\|\mathbf{x} - \mathbf{x}_n\|^2\} &= k_c(\mathbf{x}, \mathbf{x}_n) \quad (\text{Apply (6.14)}). \end{aligned}$$

By substituting (6.11) into (6.10) we obtain

$$(6.12) \quad f(\mathbf{x}) = t_{n^*} \quad \text{where} \quad n^* = \arg \max_n k_c(\mathbf{x}, \mathbf{x}_n).$$

From (6.12) we obtain a more general rule for the nearest neighbour model, where the output value is drawn from an observation which maximizes an arbitrary kernel function, having the minimization of the Euclidean distance as a particular case (when the Gaussian kernel, as in (6.23), is adopted).

Exercise 6.4

Consider the matrix

$$(6.13) \quad \begin{pmatrix} 1 & 1-\epsilon \\ 1-\epsilon & 1 \end{pmatrix}.$$

We determine the eigenvalues via the corresponding characteristic equation, as in (C.30):

$$\begin{aligned} (1-\lambda)^2 - (1-\epsilon)^2 &= 0 \\ (1-\lambda)^2 &= (1-\epsilon)^2. \end{aligned}$$

Therefore, the eigenvalues are

$$\begin{cases} 1 - \lambda_1 = 1 - \epsilon \iff \lambda_1 = \epsilon, \\ 1 - \lambda_2 = \epsilon - 1 \iff \lambda_2 = 2 - \epsilon. \end{cases}$$

Hence, for $\epsilon \in (1, 2)$, we have that the eigenvalues are positive, however the off-diagonal elements in (6.13) are negative.

Exercise 6.5

We seek to demonstrate the results (6.13) and (6.14) for the construction of valid kernel functions. It follows that, for a valid kernel function $k_1(\mathbf{x}, \mathbf{x}')$, a constant $c > 0$ and an arbitrary function $f(\mathbf{x})$, we obtain

$$\begin{aligned} ck_1(\mathbf{x}, \mathbf{x}') &= c\phi^\top(\mathbf{x})\phi(\mathbf{x}') && (\text{Apply (6.1)}) \\ &= \sqrt{c}\phi^\top(\mathbf{x})\sqrt{c}\phi(\mathbf{x}') && (\text{As } c > 0) \\ &= \varphi^\top(\mathbf{x})\varphi(\mathbf{x}') && (\text{Set } \varphi(\mathbf{x}) = \sqrt{c}\phi(\mathbf{x})) \\ ck_1(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') && (\text{Apply (6.1)}). \end{aligned}$$

Hence, we derive (6.13). Subsequently, we have

$$\begin{aligned} f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') &= f(\mathbf{x})\phi^\top(\mathbf{x})\phi(\mathbf{x}')f(\mathbf{x}') && (\text{Apply (6.1)}) \\ &= \varphi^\top(\mathbf{x})\varphi(\mathbf{x}') && (\text{Set } \varphi(\mathbf{x}) = f(\mathbf{x})\phi(\mathbf{x})) \\ f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') && (\text{Apply (6.1)}). \end{aligned}$$

Thereby deriving (6.14).

Exercise 6.6

We seek to demonstrate the results (6.15) and (6.16) for the construction of valid kernel functions. It follows that, for a valid kernel function $k_1(\mathbf{x}, \mathbf{x}')$ and a polynomial function $q(s)$ with nonnegative coefficients

$$\begin{aligned}
q(k_1(\mathbf{x}, \mathbf{x}')) &= \sum_{j=0}^p c_j (k_1(\mathbf{x}, \mathbf{x}'))^j \\
&= c_0 + \sum_{j=1}^p c_j (k_1(\mathbf{x}, \mathbf{x}'))^j \\
&= c_0 + \sum_{j=0}^p c_j \prod_{m=1}^j \{k_1(\mathbf{x}, \mathbf{x}')\} \\
&= c_0 \frac{1}{M} \mathbf{1}^\top \mathbf{1} + \sum_{j=1}^p c_j k_j^*(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.18)}) \\
&= c_0 k_0^*(\mathbf{x}, \mathbf{x}') + \sum_{j=1}^p c_j k_j^*(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.1)}) \\
&= k_0^{**}(\mathbf{x}, \mathbf{x}') + \sum_{j=1}^p k_j^{**}(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.13)}) \\
q(k_1(\mathbf{x}, \mathbf{x}')) &= k(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.17)}).
\end{aligned}$$

Hence we prove the validity of (6.15). Subsequently, we have that

$$\begin{aligned}
\exp\{k_1(\mathbf{x}, \mathbf{x}')\} &= \sum_{j=0}^{\infty} \frac{\{k_1(\mathbf{x}, \mathbf{x}')\}^j}{j!} \quad (\text{Apply } e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!}) \\
&= \lim_{J \rightarrow \infty} \sum_{j=0}^J \frac{\{k_1(\mathbf{x}, \mathbf{x}')\}^j}{j!} \\
&= \lim_{J \rightarrow \infty} k_J^*(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.15)}) \\
\exp\{k_1(\mathbf{x}, \mathbf{x}')\} &= k(\mathbf{x}, \mathbf{x}').
\end{aligned}$$

Hence, we verify (6.16).

Exercise 6.7

We seek to demonstrate the results (6.17) and (6.18) for the construction of valid kernel functions. Let $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ be valid kernel functions. First, we utilize the property that for any valid kernel function, the corresponding Gram matrix is positive semidefinite. Define a function $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$ (with respective Gram matrices \mathbf{K} , \mathbf{K}_1 and \mathbf{K}_2), we aim to verify if \mathbf{K} is positive semidefinite as follows: let \mathbf{u} be an arbitrary N dimensional vector

$$\begin{aligned}\mathbf{u}^\top \mathbf{K} \mathbf{u} &= \mathbf{u}^\top (\mathbf{K}_1 + \mathbf{K}_2) \mathbf{u} \\ &= \mathbf{u}^\top \mathbf{K}_1 \mathbf{u} + \mathbf{u}^\top \mathbf{K}_2 \mathbf{u} \\ \mathbf{u}^\top \mathbf{K} \mathbf{u} &\geq 0 \quad (\text{Positive semidefiniteness of } \mathbf{K}_1, \mathbf{K}_2).\end{aligned}$$

As we have concluded that the Gram matrix \mathbf{K} is positive semidefinite, we conclude that $k(\mathbf{x}, \mathbf{x}')$ is a valid kernel function, hence proving (6.17). Subsequently, we have that

$$\begin{aligned}k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') &= [\boldsymbol{\alpha}^\top(\mathbf{x})\boldsymbol{\alpha}(\mathbf{x})][\boldsymbol{\beta}^\top(\mathbf{x})\boldsymbol{\beta}(\mathbf{x})] \quad (\text{Apply (6.1)}) \\ &= \left[\sum_{j=1}^{M_1} \alpha_j(\mathbf{x})\alpha_j(\mathbf{x}') \right] \left[\sum_{i=1}^{M_2} \beta_i(\mathbf{x})\beta_i(\mathbf{x}') \right] \\ (6.14) \quad k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') &= \sum_{j=1}^{M_1} \sum_{i=1}^{M_2} \alpha_j(\mathbf{x})\beta_i(\mathbf{x})\alpha_j(\mathbf{x}')\beta_i(\mathbf{x}').\end{aligned}$$

We now define an ordering $(j, i)(r) : \{1, \dots, M_1 M_2\} \rightarrow \{1, \dots, M_1\} \times \{1, \dots, M_2\}$ as

$$\begin{aligned}i(r) &= \left\lfloor \frac{r-1}{M_1} \right\rfloor + 1, \\ j(r) &= (r-1) \bmod M_1 + 1.\end{aligned}$$

We thereby define

$$(6.15) \quad \phi_r(\mathbf{x}) = \alpha_{j(r)}(\mathbf{x})\beta_{i(r)}(\mathbf{x})$$

Let $R = M_1 M_2$, substituting (6.15) into (6.14) we obtain

$$\begin{aligned}k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') &= \sum_{r=1}^R \phi_r(\mathbf{x})\phi_r(\mathbf{x}') \\ &= \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}') \quad (\text{Apply (6.1)}).\end{aligned}$$

Hence we prove (6.18).

Exercise 6.8

We seek to demonstrate the results (6.19) and (6.20) for the construction of valid kernel functions. Let $k_3(\mathbf{x}, \mathbf{x}')$ be a valid kernel function, $\phi(\mathbf{x}) \in \mathbb{R}^M$ and $\mathbf{A} \in \mathbb{R}^{M \times M}$ be a positive semidefinite matrix with $D \leq N$ nonzero eigenvalues. It follows that

$$\begin{aligned} k_3(\phi(\mathbf{x}), \phi(\mathbf{x}')) &= \varphi^\top(\phi(\mathbf{x}))\varphi(\phi(\mathbf{x}')) && \text{(Apply (6.1))} \\ &= \rho^\top(\mathbf{x})\rho(\mathbf{x}') && \text{(Set } \rho(\mathbf{x}) = \varphi(\phi(\mathbf{x}))) \\ k_3(\phi(\mathbf{x}), \phi(\mathbf{x}')) &= k(\mathbf{x}, \mathbf{x}') && \text{(Apply (6.1)).} \end{aligned}$$

We have therefore proven (6.19). Let \mathbf{A} be eigendecomposed as in (2.48), it follows that

$$\begin{aligned} \mathbf{x}^\top \mathbf{A} \mathbf{x}' &= \mathbf{x}^\top \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^\top \mathbf{x}' && \text{(Apply (2.48))} \\ &= \sum_{i=1}^D \sqrt{\lambda_i} \mathbf{x}^\top \mathbf{u}_i \sqrt{\lambda_i} \mathbf{u}_i^\top \mathbf{x}' \\ &= \sum_{i=1}^D \phi_i(\mathbf{x}) \phi_i(\mathbf{x}') && \text{(Set } \phi_i(\mathbf{x}) = \sqrt{\lambda_i} \mathbf{u}_i^\top \mathbf{x}) \\ &= \phi^\top(\mathbf{x}) \phi(\mathbf{x}') \\ \mathbf{x}^\top \mathbf{A} \mathbf{x}' &= k(\mathbf{x}, \mathbf{x}') && \text{(Apply (6.1)).} \end{aligned}$$

Thereby concluding our demonstration.

Exercise 6.9

We seek to demonstrate the results (6.21) and (6.22) for the construction of valid kernel functions. Let \mathbf{x} be decomposed as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix},$$

and let $k_a(\mathbf{x}_a, \mathbf{x}'_a)$ and $k_b(\mathbf{x}_b, \mathbf{x}'_b)$ be valid kernel functions on their respective spaces. We define $f_a(\mathbf{x})$ and $f_b(\mathbf{x})$ as

$$(6.16) \quad f_a(\mathbf{x}) = \mathbf{x}_a \quad \text{and} \quad f_b(\mathbf{x}) = \mathbf{x}_b.$$

Therefore, for (6.21) we have that

$$\begin{aligned} k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) &= k_a(f_a(\mathbf{x}), f_a(\mathbf{x}')) + k_b(f_b(\mathbf{x}), f_b(\mathbf{x}')) \quad (\text{Apply (6.16)}) \\ &= k_a^*(\mathbf{x}, \mathbf{x}') + k_b^*(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.19)}) \\ k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) &= k(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.17)}). \end{aligned}$$

Thereby proving (6.21). For (6.22), we find that

$$\begin{aligned} k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b) &= k_a(f_a(\mathbf{x}), f_a(\mathbf{x}'))k_b(f_b(\mathbf{x}), f_b(\mathbf{x}')) \quad (\text{Apply (6.16)}) \\ &= k_a^*(\mathbf{x}, \mathbf{x}')k_b^*(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.19)}) \\ k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b) &= k(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.18)}), \end{aligned}$$

through which we have proven (6.22).

Exercise 6.10

Consider that we have observed a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$ of input and output units, and construct a linear kernel learner which aims to estimate the function $f(\mathbf{x})$, and that we utilize as kernel function $k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})f(\mathbf{x}')$, such that

$$(6.17) \quad \mathbf{k}(\mathbf{x}) = \begin{pmatrix} f(\mathbf{x})f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x})f(\mathbf{x}_N). \end{pmatrix}$$

Therefore, for an input unit \mathbf{x} , our learner is as in (6.9):

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{k}^\top(\mathbf{x})(\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{t} \\ &= \mathbf{k}^\top(\mathbf{x}) \mathbf{a} \quad (\text{Apply (6.9)}) \\ &= \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) a_n \\ &= \sum_{n=1}^N f(\mathbf{x})f(\mathbf{x}_n) a_n \quad (\text{Apply (6.17)}) \\ &= f(\mathbf{x}) \sum_{n=1}^N f(\mathbf{x}_n) a_n \\ &= f(\mathbf{x}) c \quad (\text{Set } c = \sum_{n=1}^N f(\mathbf{x}_n) a_n) \\ y(\mathbf{x}) &\propto f(\mathbf{x}). \end{aligned}$$

We therefore conclude that the solution to this kernel learner is proportional to $f(\mathbf{x})$.

Exercise 6.11

Consider the Gaussian kernel function, as in (6.23): we aim to demonstrate it may be rewritten as the inner product of an infinite-dimensional feature vector. We rewrite the middle term in the right-hand-side of the expansion (6.25) as

$$\begin{aligned}\exp\{\mathbf{x}^\top \mathbf{x}' / \sigma^2\} &= \sum_{j=0}^{\infty} \frac{(\mathbf{x}^\top \mathbf{x}')^j}{j!} \quad (\text{Apply } e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!}) \\ &= \sum_{j=0}^{\infty} \frac{(\mathbf{x}^\top \mathbf{x}')^j}{\sigma^{2j} j!}.\end{aligned}$$

Trivially, $\mathbf{x}^\top \mathbf{x}'$ is a valid kernel function. From iterated applications of (6.18), we find that $(\mathbf{x}^\top \mathbf{x}')^j$ is, likewise, a valid kernel function. Hence, there exists $k_j^*(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^j$. Additionally, from (6.1), there exists a feature vector $\phi_j(\mathbf{x})$ such that $k_j^*(\mathbf{x}, \mathbf{x}') = \phi_j^\top(\mathbf{x}) \phi_j(\mathbf{x}')$ - in particular, these feature vectors are comprised of all monomials of order j . Consider that the corresponding feature vectors are M_j -dimensional, such that

$$\begin{aligned}\exp\{\mathbf{x}^\top \mathbf{x}' / \sigma^2\} &= \sum_{j=0}^{\infty} \frac{\phi_j^\top(\mathbf{x}) \phi_j(\mathbf{x}')}{\sigma^{2j} j!} \\ &= \sum_{j=0}^{\infty} \sum_{r=1}^{M_j} \frac{1}{\sigma^{2j} j!} \phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}').\end{aligned}$$

Take $\phi_{r,q}(\mathbf{x}) = 0, \forall q > M_j$, such that

$$\begin{aligned}\exp\{\mathbf{x}^\top \mathbf{x}' / \sigma^2\} &= \sum_{j=0}^{\infty} \sum_{r=1}^{M_j} \frac{1}{\sigma^{2j} j!} \phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}') \\ &= \sum_{j=0}^{\infty} \sum_{r=1}^{\infty} \frac{1}{\sigma^{2j} j!} \phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}') \\ (6.18) \quad \exp\{\mathbf{x}^\top \mathbf{x}' / \sigma^2\} &= \sum_{r=1}^{\infty} \sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}') \quad \left(\sum_{j=0}^{\infty} \sum_{r=1}^{\infty} \frac{1}{\sigma^{2j} j!} |\phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}')| < \infty \right).\end{aligned}$$

We may justify the argument that the double absolute sum is bounded as follows: firstly, consider that

$$\sum_{j=0}^{\infty} \sum_{r=1}^{\infty} \frac{1}{\sigma^{2j} j!} |\phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}')| = \sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \sum_{r=1}^{M_j} |\phi_{r,j}(\mathbf{x}) \phi_{r,j}(\mathbf{x}')|.$$

As all components on the sum over r on the-right-hand side are finite, and it is a finite sum, the sum over r is finite. We thereafter have that, as previously stated, the feature vector is composed of monomials of order j . In particular

$$(6.19) \quad M_j = \frac{(N+j-1)!}{(N-1)!j!},$$

where N is the dimensionality of \mathbf{x} . By taking $\varepsilon = \max|\phi_{r,j}(\mathbf{x})\phi_{r,j}(\mathbf{x}')|$, we have that

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \sum_{r=1}^{M_j} |\phi_{r,j}(\mathbf{x})\phi_{r,j}(\mathbf{x}')| &\leq \sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \sum_{r=1}^{M_j} \varepsilon \\ &= \sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \frac{(N+j-1)!}{(N-1)!j!} \varepsilon \quad (\text{Apply (6.19)}) \\ &= \frac{1}{(N-1)!} \sum_{j=0}^{\infty} \frac{(N+j-1)!}{\sigma^{2j} (j!)^2} \varepsilon \\ \sum_{j=0}^{\infty} \frac{1}{\sigma^{2j} j!} \sum_{r=1}^{M_j} |\phi_{r,j}(\mathbf{x})\phi_{r,j}(\mathbf{x}')| &< \infty. \end{aligned}$$

Hence, we continue. We rewrite the expansion in (6.25) as

$$e^{-\mathbf{x}^\top \mathbf{x}/2\sigma^2} e^{\mathbf{x}^\top \mathbf{x}'/\sigma^2} e^{-(\mathbf{x}')^\top \mathbf{x}'/2\sigma^2} = \sum_{r=1}^{\infty} \sum_{j=0}^{\infty} \frac{e^{-\mathbf{x}^\top \mathbf{x}/2\sigma^2}}{\sigma^j \sqrt{j!}} \phi_{r,j}(\mathbf{x}) \frac{e^{-(\mathbf{x}')^\top \mathbf{x}'/2\sigma^2}}{\sigma^j \sqrt{j!}} \phi_{r,j}(\mathbf{x}') \quad (\text{Apply (6.18)}),$$

by setting

$$\varphi_{r,j}(\mathbf{x}) = \frac{e^{-\mathbf{x}^\top \mathbf{x}/2\sigma^2}}{\sigma^j \sqrt{j!}} \phi_{r,j}(\mathbf{x})$$

and taking a bijective function $s : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}_0$ (such as a modification of Cantor's pairing function), and defining

$$(6.20) \quad \alpha_s(\mathbf{x}) = \varphi_{r(s),j(s)}(\mathbf{x}),$$

we are able to rewrite the Gaussian kernel as

$$k(\mathbf{x}, \mathbf{x}') = \sum_{s=1}^{\infty} \alpha_s(\mathbf{x}) \alpha_s(\mathbf{x}').$$

We express the Gaussian kernel as the inner product of an infinite-dimensional feature vector, in the form of (6.20).

Exercise 6.12

Consider a set D , of cardinality $|D|$, and the corresponding powerset $\mathcal{P}(D)$, of cardinality $2^{|D|}$. We define the feature vector $\phi(A)$, where $A \subseteq D$, as

$$(6.21) \quad \phi(A) = (\phi_U(A))_{U \in \mathcal{P}(D)},$$

where $\phi_U(A)$ is as in (6.95). Let $A_1, A_2 \subseteq D$, it follows that

$$(6.22) \quad \begin{aligned} \phi_U(A_1)\phi_U(A_2) &= \begin{cases} 1 & \text{if } U \subseteq A_1 \text{ and } U \subseteq A_2, \\ 0 & \text{otherwise.} \end{cases} \\ \phi_U(A_1)\phi_U(A_2) &= \begin{cases} 1 & \text{if } U \subseteq A_1 \cap A_2, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that $\phi_U(A_1)\phi_U(A_2) = 1$ if and only if $U \subseteq A_1 \cap A_2$, that is, if and only if $U \in \mathcal{P}(A_1 \cap A_2)$. Moreover, $\mathcal{P}(A_1 \cap A_2) \subseteq \mathcal{P}(D)$. Thus, we write

$$\begin{aligned} \phi^\top(A_1)\phi(A_2) &= \sum_{U \in \mathcal{P}(D)} \phi_U(A_1)\phi_U(A_2) \\ &= \sum_{U \in [\mathcal{P}(D) \cap \mathcal{P}(A_1 \cap A_2)]} \phi_U(A_1)\phi_U(A_2) + \\ &\quad + \sum_{U \in [\mathcal{P}(D) \setminus \mathcal{P}(A_1 \cap A_2)]} \phi_U(A_1)\phi_U(A_2) \quad (\text{Apply } A = (A \cap B) \cup (A \cap B^c)) \\ &= \sum_{U \in \mathcal{P}(A_1 \cap A_2)} \phi_U(A_1)\phi_U(A_2) \quad (\text{Apply } \mathcal{P}(A_1 \cap A_2) \subseteq \mathcal{P}(D)) \\ &= \sum_{U \in \mathcal{P}(A_1 \cap A_2)} 1 \quad (\text{Apply (6.22)}) \\ &= 2^{|A_1 \cap A_2|} \\ \phi^\top(A_1)\phi(A_2) &= k(A_1, A_2) \quad (\text{Apply (6.27)}). \end{aligned}$$

Hence, we see that (6.27) is an inner product in $\mathcal{P}(D)$ with feature vector as in (6.21).

Exercise 6.13

Consider the Fisher kernel as seen in (6.33). Define a nonlinear transformation of the parameter θ as $\psi(\theta) = \alpha$, which is invertible and differentiable, such that $\theta = \psi^{-1}(\alpha)$. Note that, in this context, the gradient in (6.32) will be taken with respect to α , whilst $p(\mathbf{x}|\theta) = p(\mathbf{x}|\psi^{-1}(\alpha))$. We compute (6.32) as

$$\begin{aligned}
 \mathbf{g}(\alpha, \mathbf{x}) &= \nabla_\alpha \log p(\mathbf{x}|\psi^{-1}(\alpha)) \\
 &= \frac{\partial \log p(\mathbf{x}|\psi^{-1}(\alpha))}{\partial \alpha} \\
 &= \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \frac{\partial \log p(\mathbf{x}|\psi^{-1}(\alpha))}{\partial \psi^{-1}(\alpha)} \\
 &= \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta} \\
 &= \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \nabla_\theta \log p(\mathbf{x}|\theta) \\
 (6.23) \quad \mathbf{g}(\alpha, \mathbf{x}) &= \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbf{g}(\theta, \mathbf{x}) \quad (\text{Apply (6.32)}).
 \end{aligned}$$

The Fisher information matrix is similarly modified as

$$\begin{aligned}
 \mathbf{F}_\alpha &= \mathbb{E}_{\mathbf{x}}[\mathbf{g}(\alpha, \mathbf{x})\mathbf{g}^\top(\alpha, \mathbf{x})] \quad (\text{Apply (6.34)}) \\
 &= \mathbb{E}_{\mathbf{x}} \left[\frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbf{g}(\theta, \mathbf{x}) \mathbf{g}^\top(\theta, \mathbf{x}) \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \right] \quad (\text{Apply (6.23)}) \\
 &= \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbb{E}_{\mathbf{x}}[\mathbf{g}(\theta, \mathbf{x})\mathbf{g}^\top(\theta, \mathbf{x})] \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \\
 (6.24) \quad \mathbf{F}_\alpha &= \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbf{F} \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \quad (\text{Apply (6.34)}).
 \end{aligned}$$

The kernel function under the parameter transformation is computed as

$$\begin{aligned}
 k_\alpha(\mathbf{x}, \mathbf{x}') &= \mathbf{g}^\top(\alpha, \mathbf{x}) \mathbf{F}_\alpha^{-1} \mathbf{g}(\alpha, \mathbf{x}) \quad (\text{Apply (6.33)}) \\
 &= \mathbf{g}^\top(\theta, \mathbf{x}) \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \left\{ \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbf{F} \times \right. \\
 &\quad \left. \times \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \right\}^{-1} \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbf{g}(\theta, \mathbf{x}) \quad (\text{Apply (6.23) and (6.24)}) \\
 &= \mathbf{g}^\top(\theta, \mathbf{x}) \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \left\{ \frac{\partial \psi^{-1}(\alpha)}{\partial \alpha^\top} \right\}^{-1} \mathbf{F}^{-1} \times \\
 &\quad \times \left\{ \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \right\}^{-1} \frac{[\partial \psi^{-1}(\alpha)]^\top}{\partial \alpha} \mathbf{g}(\theta, \mathbf{x}) \quad (\text{Apply (C.3)}) \\
 &= \mathbf{g}^\top(\theta, \mathbf{x}) \mathbf{F}^{-1} \mathbf{g}(\theta, \mathbf{x}) \\
 (6.25) \quad k_\alpha(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') \quad (\text{Apply (6.33)}).
 \end{aligned}$$

Hence, under the transformation $\alpha = \psi(\theta)$ as outlined previously, the Fisher kernel function remains unchanged, hence we conclude it is invariant to nonlinear invertible and differentiable transformations.

Exercise 6.14

Consider that, for the Fisher kernel function, we take $p(\mathbf{x}|\boldsymbol{\mu})$ as the probability density function of a D -dimensional multivariate normal random variable, with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and fixed covariance $\mathbf{S} \in \mathbb{R}^{D \times D}$. Hence, we compute (6.32) as

$$\begin{aligned}
 \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) &= \nabla_{\boldsymbol{\mu}} \log p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) \\
 &= \frac{\partial \log p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S})}{\partial \boldsymbol{\mu}} \\
 &= \frac{\partial}{\partial \boldsymbol{\mu}} \left[-\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{S}| + \right. \\
 &\quad \left. - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \quad (\text{Apply (2.118)}) \\
 (6.26) \quad \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) &= \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \quad (\text{Apply (C.19)}).
 \end{aligned}$$

Thereafter, we compute (6.34) as

$$\begin{aligned}
 \mathbf{F} &= \mathbb{E}_{\mathbf{x}}[\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) \mathbf{g}^\top(\boldsymbol{\mu}, \mathbf{x})] \\
 &= \mathbb{E}_{\mathbf{x}}[\mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}] \\
 &= \mathbf{S}^{-1} \mathbb{E}_{\mathbf{x}}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \mathbf{S}^{-1} \\
 &= \mathbf{S}^{-1} \mathbf{S} \mathbf{S}^{-1} \quad (\text{Apply (2.64)}) \\
 (6.27) \quad \mathbf{F} &= \mathbf{S}^{-1}.
 \end{aligned}$$

Finally, the kernel function (6.33) is

$$\begin{aligned}
 k(\mathbf{x}, \mathbf{x}') &= \mathbf{g}^\top(\boldsymbol{\mu}, \mathbf{x}) \mathbf{F} \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}') \\
 &= (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} \mathbf{S} \mathbf{S}^{-1} (\mathbf{x}' - \boldsymbol{\mu}) \quad (\text{Apply (6.26) and (6.27)}) \\
 k(\mathbf{x}, \mathbf{x}') &= (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\mathbf{x}' - \boldsymbol{\mu}).
 \end{aligned}$$

Exercise 6.15

Consider a two-dimensional Gram matrix, as in (6.6). We aim to demonstrate that square of the off-diagonal elements is lesser than or equal to the product of the diagonal elements, i.e., that the Gram matrix satisfies the Cauchy-Schwartz inequality. It follows that

$$\begin{aligned} |\mathbf{K}| &= \left| \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) \\ k(x_1, x_2) & k(x_2, x_2) \end{pmatrix} \right| \\ &= k(x_1, x_1)k(x_2, x_2) - \{k(x_1, x_2)\}^2 \\ 0 &\leq k(x_1, x_1)k(x_2, x_2) - \{k(x_1, x_2)\}^2 \quad (\text{Positive semidefiniteness of } \mathbf{K}) \\ \{k(x_1, x_2)\}^2 &\leq k(x_1, x_1)k(x_2, x_2). \end{aligned}$$

As we aimed to demonstrate.

Exercise 6.16

Consider that we possess a model, governed by a parameter vector \mathbf{w} and our input data $\{\mathbf{x}_n\}_{n=1}^N$, that we adopt a feature vector $\phi(\mathbf{x})$, and that the error function corresponding to this model is as in (6.97). We note that we may decompose the vector \mathbf{w} as in (6.98), i.e., as the sum between a linear combination of $\{\phi(\mathbf{x}_n)\}_{n=1}^N$ and a vector \mathbf{w}_\perp which is orthogonal to all the elements of $\{\phi(\mathbf{x}_n)\}_{n=1}^N$. Hence, we rewrite $\mathbf{w}^\top \phi(\mathbf{x}_i)$ as

$$\begin{aligned}
 \mathbf{w}^\top \phi(\mathbf{x}) &= \left[\sum_{n=1}^N \alpha_n \phi(\mathbf{x}_n) + \mathbf{w}_\perp \right]^\top \phi(\mathbf{x}_i) && \text{(Apply (6.98))} \\
 &= \left[\sum_{n=1}^N \alpha_n \phi^\top(\mathbf{x}_n) \phi(\mathbf{x}_i) + \mathbf{w}_\perp^\top \phi(\mathbf{x}_i) \right] && \text{(Apply } \mathbf{w}_\perp^\top \phi(\mathbf{x}_i) = 0\text{)} \\
 &= \sum_{n=1}^N \alpha_n \phi^\top(\mathbf{x}_n) \phi(\mathbf{x}_i) \\
 (6.28) \quad \mathbf{w}^\top \phi(\mathbf{x}_i) &= \boldsymbol{\alpha}^\top \Phi \phi(\mathbf{x}_i),
 \end{aligned}$$

where

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} \phi^\top(\mathbf{x}_1) \\ \vdots \\ \phi^\top(\mathbf{x}_N) \end{pmatrix}.$$

We now rewrite $\mathbf{w}^\top \mathbf{w}$ as

$$\begin{aligned}
 \mathbf{w}^\top \mathbf{w} &= \left[\sum_{n=1}^N \alpha_n \phi(\mathbf{x}_n) + \mathbf{w}_\perp \right]^\top \left[\sum_{m=1}^N \alpha_m \phi(\mathbf{x}_m) + \mathbf{w}_\perp \right] && \text{(Apply (6.98))} \\
 &= \left[\sum_{n=1}^N \alpha_n \phi^\top(\mathbf{x}_n) + \mathbf{w}_\perp^\top \right] \left[\sum_{m=1}^N \alpha_m \phi(\mathbf{x}_m) + \mathbf{w}_\perp \right] \\
 &= \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m \phi^\top(\mathbf{x}_n) \phi(\mathbf{x}_m) + \sum_{n=1}^N \alpha_n \phi^\top(\mathbf{x}_n) \mathbf{w}_\perp + \\
 &\quad + \sum_{m=1}^N \alpha_m \mathbf{w}_\perp^\top \phi(\mathbf{x}_m) + \mathbf{w}_\perp^\top \mathbf{w}_\perp \\
 &= \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m \phi^\top(\mathbf{x}_n) \phi(\mathbf{x}_m) + \mathbf{w}_\perp^\top \mathbf{w}_\perp && \text{(Apply } \mathbf{w}_\perp^\top \phi(\mathbf{x}_i) = 0\text{)} \\
 (6.29) \quad \mathbf{w}^\top \mathbf{w} &= \boldsymbol{\alpha}^\top \Phi \Phi^\top \boldsymbol{\alpha} + \mathbf{w}_\perp^\top \mathbf{w}_\perp.
 \end{aligned}$$

Finally, we substitute (6.28) and (6.29) into (6.97), as follows

$$\begin{aligned}
 J(\mathbf{w}) &= f(\mathbf{w}^\top \phi(\mathbf{x}_1), \dots, \mathbf{w}^\top \phi(\mathbf{x}_N)) + g(\mathbf{w}^\top \mathbf{w}) \\
 (6.30) \quad J(\boldsymbol{\alpha}, \mathbf{w}_\perp) &= f(\boldsymbol{\alpha}^\top \Phi \phi(\mathbf{x}_1), \dots, \boldsymbol{\alpha}^\top \Phi \phi(\mathbf{x}_N)) + g(\boldsymbol{\alpha}^\top \Phi \Phi^\top \boldsymbol{\alpha} + \mathbf{w}_\perp^\top \mathbf{w}_\perp).
 \end{aligned}$$

Note that, as $g(x)$ is a strictly increasing monotone function, it follows that if $a \leq b$, therefore $g(a) \leq g(b)$. Moreover, we have that

$$\begin{aligned}
 0 \leq \mathbf{w}_\perp^\top \mathbf{w} &\implies \boldsymbol{\alpha}^\top \Phi \Phi^\top \boldsymbol{\alpha} \leq \boldsymbol{\alpha}^\top \Phi \Phi^\top \boldsymbol{\alpha} + \mathbf{w}_\perp^\top \mathbf{w}_\perp \\
 &\implies g(\boldsymbol{\alpha}^\top \Phi \Phi^\top \boldsymbol{\alpha}) \leq g(\boldsymbol{\alpha}^\top \Phi \Phi^\top \boldsymbol{\alpha} + \mathbf{w}_\perp^\top \mathbf{w}_\perp).
 \end{aligned}$$

Therefore, we find that, for all $\alpha \in \mathbb{R}^N$, $J(\alpha, \mathbf{w}_\perp) \geq J(\alpha, \mathbf{0})$. We thereby conclude that, if we intend to minimize (6.97) with respect to α and \mathbf{w}_\perp (which is equivalent to the minimization of (6.97) with respect to \mathbf{w} , given (6.98)) then $\mathbf{w}_\perp = \mathbf{0}$. Thereafter, α^* is obtained as

$$\begin{aligned}\alpha^* &= \arg \min_{\alpha} J(\alpha, \mathbf{0}) \\ &= \arg \min_{\alpha} \{f(\alpha^\top \Phi \phi(\mathbf{x}_1), \dots, \alpha^\top \Phi \phi(\mathbf{x}_N)) + g(\alpha^\top \Phi \Phi^\top \alpha)\},\end{aligned}$$

hence, from (6.98), we write \mathbf{w}^* as

$$\begin{aligned}\mathbf{w}^* &= \sum_{n=1}^N \alpha_n^* \phi(\mathbf{x}_n) \\ &= (\alpha^*)^\top \Phi.\end{aligned}$$

Therefore, we show that \mathbf{w}^* which minimizes (6.98) is a linear combination of $\{\phi(\mathbf{x}_n)\}_{n=1}^N$.

Exercise 6.17

Consider that we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, of input and target variables, and consider that for the learning procedure we adopt the sum-of-squares error function with noise on the input variables, with distribution $\nu(\mathbf{x})$, as in (6.39). We aim to determine, by calculus of variations, the function $y(\mathbf{x})$ which minimizes (6.39). It follows that

$$\begin{aligned}\frac{\partial E}{\partial y} &= \frac{\partial}{\partial y} \left[\frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}^2 \nu(\boldsymbol{\xi}) d\boldsymbol{\xi} \right] \\ &= \frac{\partial}{\partial y} \left[\frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{x}) - t_n\}^2 \nu(\mathbf{x} - \mathbf{x}_n) d\mathbf{x} \right] \quad (\text{Set } \mathbf{x} = \mathbf{x}_n + \boldsymbol{\xi}) \\ (6.31) \quad \frac{\partial E}{\partial y} &= \sum_{n=1}^N \{y(\mathbf{x}) - t_n\} \nu(\mathbf{x} - \mathbf{x}_n).\end{aligned}$$

By equating (6.31) to zero and solving for $y(\mathbf{x})$, we obtain

$$\begin{aligned}\frac{\partial E}{\partial y} &= 0 \\ \sum_{n=1}^N \{y(\mathbf{x}) - t_n\} \nu(\mathbf{x} - \mathbf{x}_n) &= 0 \\ y(\mathbf{x}) \sum_{n=1}^N \nu(\mathbf{x} - \mathbf{x}_n) - \sum_{n=1}^N t_n \nu(\mathbf{x} - \mathbf{x}_n) &= 0 \\ y(\mathbf{x}) \sum_{n=1}^N \nu(\mathbf{x} - \mathbf{x}_n) &= \sum_{n=1}^N t_n \nu(\mathbf{x} - \mathbf{x}_n) \\ y(\mathbf{x}) &= \frac{\sum_{n=1}^N t_n \nu(\mathbf{x} - \mathbf{x}_n)}{\sum_{n=1}^N \nu(\mathbf{x} - \mathbf{x}_n)} \\ y(\mathbf{x}) &= \sum_{n=1}^N t_n h(\mathbf{x} - \mathbf{x}_n) \quad (\text{Apply (6.41)}).\end{aligned}$$

We thereby derive (6.40).

Exercise 6.18

Consider the Nadaraya-Watson estimator for a one dimensional input variable x and a one dimensional target variable t , such that we adopt as joint distribution $p(t, x | \sigma^2)$ a multivariate normal distribution with covariance matrix $\sigma^2 I$, $\sigma^2 > 0$. Given an observed data set $\{x_n, t_n\}_{n=1}^N$, we write the conditional density of t , as in (6.48), as

$$\begin{aligned}
 p(t|\mathbf{x}, \sigma^2) &= \frac{\sum_{n=1}^N p(x - x_n, t - t_n | \sigma^2)}{\sum_{m=1}^N \int_{-\infty}^{\infty} p(x - x_n, s - t_n | \sigma^2) ds} \\
 &= \frac{\sum_{n=1}^N p(x - x_n | 0, \sigma^2) p(t | t_n, \sigma^2)}{\sum_{m=1}^N \int_{-\infty}^{\infty} p(x - x_m | 0, \sigma^2) p(s | t_m, \sigma^2) ds} \quad (\text{Independence of } x \text{ and } t) \\
 &= \frac{\sum_{n=1}^N p(x - x_n | 0, \sigma^2) p(t | t_n, \sigma^2)}{\sum_{m=1}^N p(x - x_m | 0, \sigma^2)} \quad (\text{Apply (1.26)}) \\
 (6.32) \quad p(t|\mathbf{x}, \sigma^2) &= \sum_{n=1}^N k(x, x_n) p(t | t_n, \sigma^2) \quad (\text{Apply (6.46)}).
 \end{aligned}$$

We thereafter compute the conditional expectation as

$$\begin{aligned}
 \mathbb{E}[t|\mathbf{x}] &= \int_{-\infty}^{\infty} s p(s|\mathbf{x}, \sigma^2) ds \quad (\text{Apply (1.37)}) \\
 &= \int_{-\infty}^{\infty} s \left[\sum_{n=1}^N k(x, x_n) p(s | t_n, \sigma^2) \right] ds \quad (\text{Apply (6.32)}) \\
 &= \sum_{n=1}^N k(x, x_n) \left[\int_{-\infty}^{\infty} s p(s | t_n, \sigma^2) ds \right] \\
 (6.33) \quad \mathbb{E}[t|\mathbf{x}] &= \sum_{n=1}^N k(x, x_n) t_n \quad (\text{Apply (1.49)}).
 \end{aligned}$$

In order to compute the conditional variance of t , we compute the second conditional moment of t as follows

$$\begin{aligned}
 \mathbb{E}[t^2|\mathbf{x}] &= \int_{-\infty}^{\infty} s^2 p(s|\mathbf{x}) ds \quad (\text{Apply (1.34)}) \\
 &= \int_{-\infty}^{\infty} s^2 \left[\sum_{n=1}^N k(x, x_n) p(s | t_n, \sigma^2) \right] ds \quad (\text{Apply (6.32)}) \\
 &= \sum_{n=1}^N k(x, x_n) \left[\int_{-\infty}^{\infty} s^2 p(s | t_n, \sigma^2) ds \right] \\
 &= \sum_{n=1}^N k(x, x_n) \{t_n^2 + \sigma^2\} \quad (\text{Apply (1.5)}) \\
 (6.34) \quad \mathbb{E}[t^2|\mathbf{x}] &= \sum_{n=1}^N k(x, x_n) t_n^2 + \sigma^2 \quad (\text{Apply } \sum_{n=1}^N k(x, x_n) = 1).
 \end{aligned}$$

We thereby compute the conditional variance as

$$\begin{aligned}
 \text{Var}[t|\mathbf{x}] &= \mathbb{E}[t^2|\mathbf{x}] - \{\mathbb{E}[t|\mathbf{x}]\}^2 \quad (\text{Apply (1.39)}) \\
 &= \sum_{n=1}^N k(x, x_n) t_n^2 + \sigma^2 - \left[\sum_{n=1}^N k(x, x_n) t_n \right] \left[\sum_{m=1}^N k(x, x_m) t_m \right] \quad (\text{Apply (6.33) and (6.34)}) \\
 \text{Var}[t|\mathbf{x}] &= \sum_{n=1}^N k(x, x_n) t_n^2 + \sigma^2 - \sum_{n=1}^N \sum_{m=1}^N k(x, x_n) k(x, x_m) t_n t_m.
 \end{aligned}$$

Exercise 6.19

We consider a framework similar to that which was seen previously in [Exercise 6.17](#), where we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, of input and target variables, and for the learning procedure adopt the sum-of-squares error function with noise on the input variables, with distribution $g(\xi)$. Additionally, the input variables are corrupted such that $\mathbf{x}_n = \mathbf{z}_n + \xi_n$, where \mathbf{z}_n is the true and latent value of the input, \mathbf{x}_n is the observed value, and ξ_n is noise. Moreover, the distribution of the target variables is such that, for \mathbf{z}_n, t_n is distributed as a normal random variable with mean $\mathbb{E}[t_n] = y(\mathbf{z}_n)$ and variance $\sigma^2 > 0$ (the choice of σ^2 is inconsequential to this Exercise), hence justifying the sum-of-squares error. We aim therefore to determine $y(\mathbf{x})$ which minimizes [\(6.99\)](#), for which we first take the derivative of [\(6.99\)](#) with respect to y as

$$\begin{aligned}
 \frac{\partial E}{\partial y} &= \frac{\partial}{\partial y} \left[\frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{x}_n - \xi_n) - t_n\}^2 g(\xi_n) d\xi_n \right] \\
 &= \frac{\partial}{\partial y} \left[-\frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{z}) - t_n\}^2 g(\mathbf{x}_n - \mathbf{z}) d\mathbf{z} \right] \quad (\text{Set } \mathbf{z} = \mathbf{x}_n - \xi_n) \\
 &= \frac{\partial}{\partial y} \left[-\frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{z}) - t_n\}^2 g(\mathbf{z} - \mathbf{x}_n) d\mathbf{z} \right] \quad (\text{Symmetry of } g(\mathbf{z})) \\
 (6.35) \quad \frac{\partial E}{\partial y} &= - \sum_{n=1}^N \{y(\mathbf{z}) - t_n\} g(\mathbf{x} - \mathbf{x}_n).
 \end{aligned}$$

By equating [\(6.35\)](#) with zero, we obtain

$$\begin{aligned}
 \frac{\partial E}{\partial y} &= 0 \\
 - \sum_{n=1}^N \{y(\mathbf{z}) - t_n\} g(\mathbf{x} - \mathbf{x}_n) &= 0 \\
 y(\mathbf{x}) \sum_{n=1}^N g(\mathbf{x} - \mathbf{x}_n) - \sum_{n=1}^N t_n g(\mathbf{x} - \mathbf{x}_n) &= 0 \\
 y(\mathbf{x}) \sum_{n=1}^N g(\mathbf{x} - \mathbf{x}_n) &= \sum_{n=1}^N t_n g(\mathbf{x} - \mathbf{x}_n) \\
 y(\mathbf{x}) &= \frac{\sum_{n=1}^N t_n g(\mathbf{x} - \mathbf{x}_n)}{\sum_{n=1}^N g(\mathbf{x} - \mathbf{x}_n)} \\
 y(\mathbf{x}) &= \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n \quad (\text{Apply } (6.46)),
 \end{aligned}$$

hence, we derive [\(6.45\)](#).

Exercise 6.20

Let $\mathbf{t}_{N+1} = (t_1, \dots, t_{N+1})^\top$ be distributed as an $(N+1)$ -dimensional multivariate normal random vector with mean $\mathbf{0}_{N+1}$ and covariance matrix \mathbf{C}_{N+1} , where $c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}$ and $\mathbf{k} = (k(\mathbf{x}_1, \mathbf{x}_{N+1}), \dots, k(\mathbf{x}_N, \mathbf{x}_{N+1}))^\top$. By partitioning \mathbf{t}_{N+1} as

$$\mathbf{t}_{N+1} = \begin{pmatrix} \mathbf{t}_N \\ t_{N+1} \end{pmatrix},$$

it follows from previous results (see (2.81) and (2.82)) that the distribution of $t_{N+1} | \mathbf{t}_N$ is a one-dimensional normal with mean

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{t}$$

and variance

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k}.$$

Exercise 6.21

We aim to demonstrate herein that, in the context of Gaussian process regression, the predictive distribution of t_{N+1} is identical to the predictive distribution in Bayesian linear regression, as determined in [Exercise 3.10](#), assuming we define the kernel in terms of a basis function as in

$$(6.36) \quad k(\mathbf{x}, \mathbf{x}') = \frac{1}{\alpha} \phi^\top(\mathbf{x}) \phi(\mathbf{x}') \quad \text{such that} \quad \mathbf{K} = \frac{1}{\alpha} \Phi \Phi^\top \quad \text{and} \quad \mathbf{k} = \frac{1}{\alpha} \Phi \phi(\mathbf{x}_{N+1}).$$

Firstly, we note that, as in both cases the predictive distribution is a one-dimensional normal (as seen in [Exercise 3.10](#) and [Exercise 6.20](#)), we need now only verify that the mean and variances match. Firstly, we rewrite the mean of the predictive distribution of the kernel case, as in [\(6.66\)](#), as

$$\begin{aligned} m(\mathbf{x}_{N+1}) &= \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{t} \\ &= \frac{1}{\alpha} \phi^\top(\mathbf{x}_{N+1}) \Phi^\top [\mathbf{K} + \beta^{-1} \mathbf{I}_N]^{-1} \mathbf{t} && (\text{Apply (6.62) and (6.36)}) \\ &= \frac{1}{\alpha} \phi^\top(\mathbf{x}_{N+1}) \Phi^\top [\alpha^{-1} \Phi \Phi^\top + \beta^{-1} \mathbf{I}_N]^{-1} \mathbf{t} && (\text{Apply (6.36)}) \\ &= \frac{1}{\alpha} \phi^\top(\mathbf{x}_{N+1}) [\alpha^{-1} \Phi^\top \Phi + \beta^{-1} \mathbf{I}_N]^{-1} \Phi^\top \mathbf{t} && (\text{Apply (C.6)}) \\ &= \beta \phi^\top(\mathbf{x}_{N+1}) [\beta \Phi^\top \Phi + \alpha \mathbf{I}_N]^{-1} \Phi^\top \mathbf{t} \\ &= \phi^\top(\mathbf{x}_{N+1}) \beta \mathbf{S}_N \Phi^\top \mathbf{t} && (\text{Apply (3.54)}) \\ &= \phi^\top(\mathbf{x}_{N+1}) \mathbf{m}_N && (\text{Apply (3.84)}) \\ m(\mathbf{x}_{N+1}) &= \mathbf{m}_N^\top \phi(\mathbf{x}_{N+1}). \end{aligned}$$

We thereby observe the mean is the same as that which is obtained in [Exercise 3.10](#). We rewrite the predictive variance, as in [\(6.67\)](#), as

$$\begin{aligned} \sigma^2(\mathbf{x}_{N+1}) &= c - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k} \\ &= \frac{1}{\beta} + \frac{1}{\alpha} \phi^\top(\mathbf{x}_{N+1}) \phi(\mathbf{x}_{N+1}) + \\ &\quad - \frac{1}{\alpha^2} \phi^\top(\mathbf{x}_{N+1}) \Phi^\top [\alpha^{-1} \Phi \Phi^\top + \beta^{-1} \mathbf{I}_N]^{-1} \Phi \phi(\mathbf{x}_{N+1}) && (\text{Apply (6.36)}) \\ &= \frac{1}{\beta} + \frac{1}{\alpha} \phi^\top(\mathbf{x}_{N+1}) \{ \mathbf{I}_N - \alpha^{-1} \Phi^\top [\alpha^{-1} \Phi \Phi^\top + \beta^{-1} \mathbf{I}_N]^{-1} \Phi \} \phi(\mathbf{x}_{N+1}) \\ &= \frac{1}{\beta} + \frac{1}{\alpha} \phi^\top(\mathbf{x}_{N+1}) [\mathbf{I} + \alpha^{-1} \beta \Phi^\top \Phi]^{-1} \phi(\mathbf{x}_{N+1}) && (\text{Apply (2.289)}) \\ &= \frac{1}{\beta} + \phi^\top(\mathbf{x}_{N+1}) [\alpha \mathbf{I} + \beta \Phi^\top \Phi]^{-1} \phi(\mathbf{x}_{N+1}) \\ \sigma^2(\mathbf{x}_{N+1}) &= \frac{1}{\beta} + \phi^\top(\mathbf{x}_{N+1}) \mathbf{S}_N \phi(\mathbf{x}_{N+1}) && (\text{Apply (3.54)}). \end{aligned}$$

This result matches the input-dependent variance on [\(3.15\)](#). We consequently conclude that Bayesian linear regression procedure provides a identical predictive distribution to that under this framework of Gaussian process regression.

Exercise 6.22

Consider, in the context of Gaussian processes for regression, that we partition our observations into two sets, the training data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, of size N , and a test data set $\{\mathbf{x}_n, t_n\}_{n=N+1}^{N+L}$, of size L . We aim to determine herein the distribution of $\tilde{\mathbf{t}} = (t_{N+1}, \dots, t_{N+L})^\top$ conditional on \mathbf{t} , i.e., the distribution of the test data conditional on the training data. We define herein the following quantities

$$(6.37) \quad \begin{aligned} \mathbf{K} &= \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \\ k(\mathbf{x}_1, \mathbf{x}_2) & k(\mathbf{x}_2, \mathbf{x}_2) & \dots & k(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_1, \mathbf{x}_N) & k(\mathbf{x}_2, \mathbf{x}_N) & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} \\ \tilde{\mathbf{K}} &= \begin{pmatrix} k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) & k(\mathbf{x}_{N+1}, \mathbf{x}_{N+2}) & \dots & k(\mathbf{x}_{N+1}, \mathbf{x}_{N+L}) \\ k(\mathbf{x}_{N+1}, \mathbf{x}_{N+2}) & k(\mathbf{x}_{N+2}, \mathbf{x}_{N+2}) & \dots & k(\mathbf{x}_{N+2}, \mathbf{x}_{N+L}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_{N+1}, \mathbf{x}_{N+L}) & k(\mathbf{x}_{N+2}, \mathbf{x}_{N+L}) & \dots & k(\mathbf{x}_{N+L}, \mathbf{x}_{N+L}) \end{pmatrix} \end{aligned}$$

and

$$(6.38) \quad \tilde{\mathbf{k}} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_{N+1}) & k(\mathbf{x}_1, \mathbf{x}_{N+2}) & \dots & k(\mathbf{x}_1, \mathbf{x}_{N+L}) \\ k(\mathbf{x}_2, \mathbf{x}_{N+1}) & k(\mathbf{x}_2, \mathbf{x}_{N+2}) & \dots & k(\mathbf{x}_2, \mathbf{x}_{N+L}) \\ \vdots & \vdots & \vdots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_{N+1}) & k(\mathbf{x}_N, \mathbf{x}_{N+2}) & \dots & k(\mathbf{x}_N, \mathbf{x}_{N+L}) \end{pmatrix}.$$

Note that $(\mathbf{t}^\top, \tilde{\mathbf{t}}^\top)^\top$ is jointly distributed as a $(N+L)$ -dimensional multivariate normal random vector with mean

$$\mathbb{E} \left[\begin{pmatrix} \mathbf{t} \\ \tilde{\mathbf{t}} \end{pmatrix} \right] = \mathbf{0}_{N+L}$$

and covariance

$$\begin{aligned} \mathbb{V}\text{ar} \left[\begin{pmatrix} \mathbf{t} \\ \tilde{\mathbf{t}} \end{pmatrix} \right] &= \begin{pmatrix} \mathbf{K} + \beta^{-1} \mathbf{I}_N & \tilde{\mathbf{k}} \\ \tilde{\mathbf{k}}^\top & \tilde{\mathbf{K}} + \beta^{-1} \mathbf{I}_L \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{C}_N & \tilde{\mathbf{k}} \\ \tilde{\mathbf{k}}^\top & \tilde{\mathbf{C}}_L \end{pmatrix}. \end{aligned}$$

From previous results (see (2.81) and (2.82)), we have that the distribution of $\tilde{\mathbf{t}}|\mathbf{t}$ is an L -dimensional multivariate normal with mean

$$(6.39) \quad \mathbb{E}[\tilde{\mathbf{t}}|\mathbf{t}] = \tilde{\mathbf{k}}^\top \mathbf{C}_N^{-1} \mathbf{t},$$

and covariance matrix

$$(6.40) \quad \mathbb{V}\text{ar}[\tilde{\mathbf{t}}|\mathbf{t}] = \tilde{\mathbf{C}}_L - \tilde{\mathbf{k}}^\top \mathbf{C}_N^{-1} \tilde{\mathbf{k}}.$$

We now take $j \in \{N+1, \dots, N+L\}$, and aim to determine the marginal distribution of $\tilde{t}_j|\mathbf{t}$. We may write, without loss of generality, a partition of $\tilde{\mathbf{t}}$ as

$$\tilde{\mathbf{t}} = \begin{pmatrix} \tilde{t}_j \\ \tilde{\mathbf{t}}_{-j} \end{pmatrix}$$

wherein \tilde{t}_j indicates the j -th component of the vector $\tilde{\mathbf{t}}$, and $\tilde{\mathbf{t}}_{-j}$ indicates all other components, except for the j -th. From (2.92) and (2.93), as well as other previous results, we know that

the marginal distribution of $\tilde{t}_j|\mathbf{t}$ is a one dimensional normal, whose mean is given by the j -th coordinate of the mean vector of $\tilde{\mathbf{t}}|\mathbf{t}$, and whose variance is the (j,j) -th coordinate of the covariance matrix of $\tilde{\mathbf{t}}|\mathbf{t}$. Firstly, we rewrite (6.38) as

$$\tilde{\mathbf{k}} = \begin{pmatrix} \kappa_{N+1}^\top & \kappa_{N+2}^\top & \dots & \kappa_{N+L}^\top \end{pmatrix},$$

where $\kappa_j = (k(\mathbf{x}_1, \mathbf{x}_j), \dots, k(\mathbf{x}_N, \mathbf{x}_j))^\top$. Thereafter, we write the expected value (6.39) as

$$(6.41) \quad \begin{aligned} \mathbb{E}[\tilde{\mathbf{t}}|\mathbf{t}] &= \begin{pmatrix} \kappa_{N+1}^\top \\ \kappa_{N+2}^\top \\ \vdots \\ \kappa_{N+L}^\top \end{pmatrix} \mathbf{C}_N^{-1} \mathbf{t} \\ \mathbb{E}[\tilde{\mathbf{t}}|\mathbf{t}] &= \begin{pmatrix} \kappa_{N+1}^\top \mathbf{C}_N^{-1} \mathbf{t} \\ \kappa_{N+2}^\top \mathbf{C}_N^{-1} \mathbf{t} \\ \vdots \\ \kappa_{N+L}^\top \mathbf{C}_N^{-1} \mathbf{t} \end{pmatrix}. \end{aligned}$$

Conversely, the covariance matrix in (6.40) is rewritten as

$$(6.42) \quad \begin{aligned} \text{Var}[\tilde{\mathbf{t}}|\mathbf{t}] &= \tilde{\mathbf{C}}_L - \begin{pmatrix} \kappa_{N+1}^\top \\ \kappa_{N+2}^\top \\ \vdots \\ \kappa_{N+L}^\top \end{pmatrix} \mathbf{C}_N^{-1} \begin{pmatrix} \kappa_{N+1} & \kappa_{N+2} & \dots & \kappa_{N+L} \end{pmatrix} \\ \text{Var}[\tilde{\mathbf{t}}|\mathbf{t}] &= \tilde{\mathbf{C}}_L - \begin{pmatrix} \kappa_{N+1}^\top \mathbf{C}_N^{-1} \kappa_{N+1} & \kappa_{N+1}^\top \mathbf{C}_N^{-1} \kappa_{N+2} & \dots & \kappa_{N+1}^\top \mathbf{C}_N^{-1} \kappa_{N+L} \\ \kappa_{N+2}^\top \mathbf{C}_N^{-1} \kappa_{N+1} & \kappa_{N+2}^\top \mathbf{C}_N^{-1} \kappa_{N+2} & \dots & \kappa_{N+2}^\top \mathbf{C}_N^{-1} \kappa_{N+L} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{N+L}^\top \mathbf{C}_N^{-1} \kappa_{N+L} & \kappa_{N+L}^\top \mathbf{C}_N^{-1} \kappa_{N+L} & \dots & \kappa_{N+L}^\top \mathbf{C}_N^{-1} \kappa_{N+L} \end{pmatrix}. \end{aligned}$$

We therefore note that the j -th component of the mean vector in (6.41) is $\kappa_j^\top \mathbf{C}_N^{-1} \mathbf{t}$. Note that $\{\tilde{\mathbf{C}}_L\}_{j,j} = \beta^{-1} + k(\mathbf{x}_j, \mathbf{x}_j)$, hence the (j,j) -th component of the covariance matrix in (6.42) is $\beta^{-1} + k(\mathbf{x}_j, \mathbf{x}_j) - \kappa_j^\top \mathbf{C}_N^{-1} \kappa_j$. We thereby conclude that $\tilde{t}_j|\mathbf{t}$ is distributed as a one-dimensional normal random variable with mean

$$\mathbb{E}[\tilde{t}_j|\mathbf{t}] = \kappa_j^\top \mathbf{C}_N^{-1} \mathbf{t},$$

and variance

$$\text{Var}[\tilde{t}_j|\mathbf{t}] = \beta^{-1} + k(\mathbf{x}_j, \mathbf{x}_j) - \kappa_j^\top \mathbf{C}_N^{-1} \kappa_j.$$

It can be seen that these results match the usual marginal predictive distribution results seen in (6.66) and (6.67), as seen in [Exercise 6.20](#).

Exercise 6.23

Consider the usual Gaussian process regression framework. We aim to extend it herein to the context where the target variables are themselves D -dimensional, such that we observe the data set $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^N$, respectively of input and target variables. Firstly, we write

$$\boldsymbol{\tau} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \vdots \\ \mathbf{t}_N \end{pmatrix},$$

such that $\boldsymbol{\tau}$ is a ND -dimensional random vector, with mean zero, i.e.,

$$\mathbb{E}[\boldsymbol{\tau}] = \mathbf{0}_{ND}.$$

In this context, we consider two sources of covariance: the usual source, induced by the Gram matrix \mathbf{K} , which occurs between observed target vectors, whilst the covariance of the terms within a target vector are induced by a precision matrix $\Lambda^{-1} \in \mathbb{R}^{D \times D}$. This is expressed via the following covariance matrix:

$$\begin{aligned} \text{Var}[\boldsymbol{\tau}] &= \mathbf{C}_N \\ &= \mathbf{I}_N \otimes \Lambda^{-1} + \mathbf{K} \otimes (\mathbf{1}_D \mathbf{1}_D^\top), \end{aligned}$$

where \otimes denotes the Kronecker product and \mathbf{K} is as in (6.6). Consider an additional observation $\{\mathbf{x}_{N+1}, \mathbf{t}_{N+1}\}$, whose conditional distribution we aim to determine. We note that $\text{Var}[\mathbf{t}_{N+1}] = \Lambda^{-1}$ and $\text{Cov}[\boldsymbol{\tau}, \mathbf{t}_{N+1}] = \mathbf{k}$, where

$$\mathbf{k} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_{N+1}) \otimes (\mathbf{1}_D \mathbf{1}_D^\top) \\ k(\mathbf{x}_2, \mathbf{x}_{N+1}) \otimes (\mathbf{1}_D \mathbf{1}_D^\top) \\ \vdots \\ k(\mathbf{x}_N, \mathbf{x}_{N+1}) \otimes (\mathbf{1}_D \mathbf{1}_D^\top) \end{pmatrix}.$$

Hence, utilizing (6.66) and (6.67), we find that the predictive distribution of \mathbf{t}_{N+1} is a D -dimensional multivariate normal, with mean

$$\mathbb{E}[\mathbf{t}_{N+1} | \boldsymbol{\tau}] = \mathbf{k}^\top \mathbf{C}_N^{-1} \boldsymbol{\tau}$$

and covariance

$$\text{Var}[\mathbf{t}_{N+1}] = \Lambda^{-1} - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k}.$$

Exercise 6.24

Let \mathbf{W} be a diagonal matrix whose elements satisfy $W_{j,j} \in (0, 1)$. We can prove \mathbf{W} is positive definite by definition, taking an arbitrary real vector \mathbf{u} of same dimension as \mathbf{W} (take M), and demonstrating that

$$\begin{aligned}\mathbf{u}^\top \mathbf{W} \mathbf{u} &= \sum_{j=1}^M u_j^2 W_{j,j} \\ \mathbf{u}^\top \mathbf{W} \mathbf{u} &\geq 0,\end{aligned}$$

for all $\mathbf{u} \in \mathbb{R}^M$. Hence, \mathbf{W} is positive definite, by definition. Now, let \mathbf{A}_1 and \mathbf{A}_2 be positive definite matrices of same dimension (again, taking M). Let $\mathbf{u} \in \mathbb{R}^M$, it follows that

$$\begin{aligned}\mathbf{u}^\top (\mathbf{A}_1 + \mathbf{A}_2) \mathbf{u} &= \mathbf{u}^\top \mathbf{A}_1 \mathbf{u} + \mathbf{u}^\top \mathbf{A}_2 \mathbf{u} \\ &\geq 0 \quad (\text{Positive definite-ess of } \mathbf{A}_1, \mathbf{A}_2).\end{aligned}$$

for all values of $\mathbf{u} \in \mathbb{R}^M$. We conclude that if \mathbf{A}_1 and \mathbf{A}_2 are positive definite matrices, therefore $\mathbf{A}_1 + \mathbf{A}_2$ are positive definite.

Exercise 6.25

We aim to derive the result (6.83) via Newton-Raphson's formula (4.92) as follows:

$$\begin{aligned}
 \mathbf{a}_N^{(\text{new})} &= \mathbf{a}_N^{(\text{old})} - \{\nabla \nabla^\top \Psi(\mathbf{a}_N^{(\text{old})})\}^{-1} \nabla \Psi(\mathbf{a}_N^{(\text{old})}) \\
 &= \mathbf{a}_N^{(\text{old})} - (-\mathbf{W}_N - \mathbf{C}_N^{-1})^{-1} (\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N^{(\text{old})}) \quad (\text{Apply (6.81) and (6.82)}) \\
 &= \mathbf{a}_N^{(\text{old})} + ([\mathbf{W}_N \mathbf{C}_N + \mathbf{I}] \mathbf{C}_N^{-1})^{-1} (\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N^{(\text{old})}) \\
 &= \mathbf{a}_N^{(\text{old})} + \mathbf{C}_N (\mathbf{W}_N \mathbf{C}_N + \mathbf{I})^{-1} (\mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N^{(\text{old})}) \quad (\text{Apply (C.3)}) \\
 &= \mathbf{C}_N (\mathbf{W}_N \mathbf{C}_N + \mathbf{I})^{-1} \times \\
 &\quad \times ([\mathbf{C}_N (\mathbf{W}_N \mathbf{C}_N + \mathbf{I})^{-1}]^{-1} \mathbf{a}_N^{(\text{old})} + \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N^{(\text{old})}) \\
 &= \mathbf{C}_N (\mathbf{W}_N \mathbf{C}_N + \mathbf{I})^{-1} \times \\
 &\quad \times ([([\mathbf{W}_N \mathbf{C}_N + \mathbf{I}] \mathbf{C}_N^{-1}] \mathbf{a}_N^{(\text{old})} + \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N^{(\text{old})}) \quad (\text{Apply (C.3)}) \\
 &= \mathbf{C}_N (\mathbf{W}_N \mathbf{C}_N + \mathbf{I})^{-1} \times \\
 &\quad \times ([\mathbf{W}_N + \mathbf{C}_N^{-1}] \mathbf{a}_N^{(\text{old})} + \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N^{(\text{old})}) \\
 \mathbf{a}_N^{(\text{new})} &= \mathbf{C}_N (\mathbf{W}_N \mathbf{C}_N + \mathbf{I})^{-1} (\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N \mathbf{a}_N^{(\text{old})}).
 \end{aligned}$$

Thereby deriving (6.83).

Exercise 6.26

Consider the result that the distribution of a_{N+1} conditional on \mathbf{a}_N is as in (6.78), and that we adopt the Laplace approximation to the posterior of \mathbf{a}_N as in (6.86). Note that the mean of a_{N+1} conditional on \mathbf{a}_N is a linear function of \mathbf{a}_N , hence we may utilize the results in (2.115) for linear Gaussian models, such that $a_{N+1}|\mathbf{a}_N$ is distributed as a one-dimensional normal random variable, with mean

$$\begin{aligned}\mathbb{E}[a_{N+1}|\mathbf{a}_N] &= \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{a}_N^* \\ &= \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{C}_N (\mathbf{t}_N - \boldsymbol{\sigma}_N) \quad (\text{Apply (6.84)}) \\ \mathbb{E}[a_{N+1}|\mathbf{a}_N] &= \mathbf{k}^\top (\mathbf{t}_N - \boldsymbol{\sigma}_N),\end{aligned}$$

and with variance

$$\begin{aligned}\text{Var}[a_{N+1}|\mathbf{a}_N] &= c - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k} + \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{H}^{-1} \mathbf{C}_N^{-1} \mathbf{k} \\ &= c - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k} + \mathbf{k}^\top \mathbf{C}_N^{-1} \{ \mathbf{W}_N + \mathbf{C}_N^{-1} \}^{-1} \mathbf{C}_N^{-1} \mathbf{k} \quad (\text{Apply (6.85)}) \\ &= c - \mathbf{k}^\top [\mathbf{C}_N^{-1} - \mathbf{C}_N^{-1} \{ \mathbf{W}_N + \mathbf{C}_N^{-1} \}^{-1} \mathbf{C}_N^{-1}] \mathbf{k} \\ \text{Var}[a_{N+1}|\mathbf{a}_N] &= c - \mathbf{k}^\top \{ \mathbf{W}_N^{-1} + \mathbf{C}_N \} \mathbf{k} \quad (\text{Apply (2.289)}).\end{aligned}$$

Exercise 6.27

Consider that, in the context of Gaussian processes for classification, we utilize Laplace's approximation to the model evidence, as in (4.135), where $\mathbf{z}_0 = \mathbf{a}_N^*$ and $f(\mathbf{z}_0) = p(\mathbf{t}_N|\mathbf{z}_0)p(\mathbf{z}_0|\theta)$, and where θ are the covariance parameters, obtaining

$$\begin{aligned} p(\mathbf{t}_N|\theta) &\approx p(\mathbf{t}_N|\mathbf{a}_N^*)p(\mathbf{a}_N^*)\frac{(2\pi)^{N/2}}{|\mathbf{H}|^{1/2}} \\ &= \Psi(\mathbf{a}_N^*)\frac{(2\pi)^{N/2}}{|\mathbf{W}_N + \mathbf{C}_N^{-1}|^{1/2}} \quad (\text{Apply (6.80)}) \\ \log p(\mathbf{t}_N|\theta) &\approx \log \Psi(\mathbf{a}_N^*) - \frac{1}{2} \log |\mathbf{W}_N + \mathbf{C}_N^{-1}| + \frac{N}{2} \log(2\pi). \end{aligned}$$

We thereby derive (6.90). First, we aim to differentiate (6.90) with θ_j , holding \mathbf{a}_N^* as constant (i.e., considering only explicit dependences on θ_j), obtaining the following

$$\begin{aligned} \frac{\partial \log p(\mathbf{t}_N|\theta)}{\partial \theta_j} &\approx \frac{\partial}{\partial \theta_j} \left[\log \Psi(\mathbf{a}_N^*) - \frac{1}{2} \log |\mathbf{W}_N + \mathbf{C}_N^{-1}| + \frac{N}{2} \log(2\pi) \right] \\ &= \frac{\partial}{\partial \theta_j} \left[-\frac{1}{2} (\mathbf{a}_N^*)^\top \mathbf{C}_N^{-1} \mathbf{a}_N^* - \frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{C}_N| + \right. \\ &\quad \left. + \mathbf{t}_N^\top \mathbf{a}_N^* - \sum_{n=1}^N \log(1 + e^{a_n^*}) - \frac{1}{2} \log |\mathbf{W}_N + \mathbf{C}_N^{-1}| + \frac{N}{2} \log(2\pi) \right] \quad (\text{Apply (6.80)}) \\ &= \frac{\partial}{\partial \theta_j} \left[-\frac{1}{2} (\mathbf{a}_N^*)^\top \mathbf{C}_N^{-1} \mathbf{a}_N^* - \frac{N}{2} \log(2\pi) + \right. \\ &\quad \left. + \mathbf{t}_N^\top \mathbf{a}_N^* - \sum_{n=1}^N \log(1 + e^{a_n^*}) - \frac{1}{2} \log |\mathbf{I} + \mathbf{C}_N \mathbf{W}_N| + \frac{N}{2} \log(2\pi) \right] \\ &= \frac{1}{2} (\mathbf{a}_N^*)^\top \mathbf{C}_N^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j} \mathbf{C}_N^{-1} \mathbf{a}_N^* + \\ &\quad - \frac{1}{2} \text{tr} \left[(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1} \frac{\partial (\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)}{\partial \theta_j} \right] \quad (\text{Apply (C.21) and (C.22)}) \\ &= \frac{1}{2} (\mathbf{a}_N^*)^\top \mathbf{C}_N^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j} \mathbf{C}_N^{-1} \mathbf{a}_N^* + \\ &\quad - \frac{1}{2} \text{tr} \left[(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j} \mathbf{W}_n \right] \quad (\text{Apply (C.20)}) \\ \frac{\partial \log p(\mathbf{t}_N|\theta)}{\partial \theta_j} &\approx \frac{1}{2} (\mathbf{a}_N^*)^\top \mathbf{C}_N^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j} \mathbf{C}_N^{-1} \mathbf{a}_N^* + \\ &\quad - \frac{1}{2} \text{tr} \left[(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1} \mathbf{W}_n \frac{\partial \mathbf{C}_N}{\partial \theta_j} \right] \quad (\text{Diagonality of } \mathbf{W}_N). \end{aligned}$$

We thereby derive (6.91). We aim to consider now the implicit dependences on θ_j in (6.90), that is, those resulting of the dependence of \mathbf{a}_N^* on θ_j . Note that, by definition, $\nabla_{\mathbf{a}_N} \Psi(\mathbf{a}_N^*) = \mathbf{0}_N$, thus we need only consider the contributions from $|\mathbf{W}_N + \mathbf{C}_N^{-1}|$. Prior to doing so, let us

compute the derivative of \mathbf{W}_N with respect to a_n^* . It follows that

$$\begin{aligned}
 \frac{\partial \mathbf{W}_N}{\partial a_n^*} &= \frac{\partial}{\partial a_n^*} \begin{pmatrix} \sigma(a_1^*)(1 - \sigma(a_1^*)) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma(a_N^*)(1 - \sigma(a_N^*)) \end{pmatrix} \\
 &= \left[\frac{\partial \sigma(a_n^*)}{\partial a_n^*} (1 - \sigma(a_n^*)) - \sigma(a_n^*) \frac{\partial \sigma(a_n^*)}{\partial a_n^*} \right] \mathbf{I}^{(n)} \\
 &= \left[\frac{\partial \sigma(a_n^*)}{\partial a_n^*} (1 - 2\sigma(a_n^*)) \right] \mathbf{I}^{(n)} \\
 (6.43) \quad \frac{\partial \mathbf{W}_N}{\partial a_n^*} &= \sigma(a_n^*)(1 - \sigma(a_n^*))(1 - 2\sigma(a_n^*)) \mathbf{I}^{(n)} \quad (\text{Apply (4.88)}),
 \end{aligned}$$

where $\mathbf{I}^{(j)}$ is a $N \times N$ -dimensional matrix composed entirely of zeroes, except at the (j, j) -th coordinate, wherein $\mathbf{I}_{j,j}^{(j)} = 1$. For brevity, we hereafter write $\sigma(a_n^*) = \sigma_n^*$. It follows that

$$\begin{aligned}
 \frac{\partial \log|\mathbf{W}_N + \mathbf{C}_N^{-1}|}{\partial a_n^*} &= \text{tr} \left([\mathbf{W}_N + \mathbf{C}_N^{-1}]^{-1} \frac{\partial [\mathbf{W}_N + \mathbf{C}_N^{-1}]}{\partial a_n^*} \right) \quad (\text{Apply (C.22)}) \\
 &= \text{tr}([\mathbf{C}_N^{-1} \{ \mathbf{C}_N \mathbf{W}_N + \mathbf{I} \}]^{-1} \times \\
 &\quad \times \sigma_n^*(1 - \sigma_n^*)(1 - 2\sigma_n^*) \mathbf{I}^{(n)}) \quad (\text{Apply (6.43)}) \\
 &= \sigma_n^*(1 - \sigma_n^*)(1 - 2\sigma_n^*) \times \\
 &\quad \times \text{tr}([(I + \mathbf{C}_N \mathbf{W}_N)^{-1} \mathbf{C}_N] \mathbf{I}^{(n)}) \quad (\text{Apply (C.3)}) \\
 (6.44) \quad \frac{\partial \log|\mathbf{W}_N + \mathbf{C}_N^{-1}|}{\partial a_n^*} &= \sigma_n^*(1 - \sigma_n^*)(1 - 2\sigma_n^*) \times \\
 &\quad \times [(I + \mathbf{C}_N \mathbf{W}_N)^{-1} \mathbf{C}_N]_{n,n}.
 \end{aligned}$$

We subsequently obtain from (6.44)

$$-\frac{1}{2} \sum_{n=1}^N \frac{\partial \log|\mathbf{W}_N + \mathbf{C}_N^{-1}|}{\partial a_n^*} \frac{\partial a_n^*}{\partial \theta_j} = -\frac{1}{2} \sum_{n=1}^N [(I + \mathbf{C}_N \mathbf{W}_N)^{-1} \mathbf{C}_N]_{n,n} \sigma_n^*(1 - \sigma_n^*)(1 - 2\sigma_n^*) \frac{\partial a_n^*}{\partial \theta_j},$$

hence, we verify (6.92). We now compute the gradient of $(\sigma^*)^\top$ with respect to \mathbf{a}_N^* as follows

$$\begin{aligned}
 \frac{\partial (\sigma^*)^\top}{\partial \mathbf{a}_N^*} &= \frac{\partial}{\partial \mathbf{a}_N^*} (\sigma(a_1^*) \ \dots \ \sigma(a_N^*)) \\
 &= \begin{pmatrix} \sigma(a_1^*)(1 - \sigma(a_1^*)) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma(a_N^*)(1 - \sigma(a_N^*)) \end{pmatrix} \quad (\text{Apply (4.88)}) \\
 (6.45) \quad \frac{\partial (\sigma^*)^\top}{\partial \mathbf{a}_N^*} &= \mathbf{W}_N.
 \end{aligned}$$

Differentiating both sides of (6.84) with respect to θ_j , we obtain

$$\begin{aligned}
 \frac{\partial \mathbf{a}_n^*}{\partial \theta_j} &= \frac{\partial \mathbf{C}_N}{\partial \theta_j}(\mathbf{t}_N - \boldsymbol{\sigma}^*) + \mathbf{C}_N \frac{\partial \boldsymbol{\sigma}^*}{\partial \theta_j} \\
 &= \frac{\partial \mathbf{C}_N}{\partial \theta_j}(\mathbf{t}_N - \boldsymbol{\sigma}^*) + \mathbf{C}_N \frac{\partial (\boldsymbol{\sigma}^*)^\top}{\partial \mathbf{a}_N^*} \frac{\partial \mathbf{a}_N^*}{\partial \theta_j} \\
 &= \frac{\partial \mathbf{C}_N}{\partial \theta_j}(\mathbf{t}_N - \boldsymbol{\sigma}^*) + \mathbf{C}_N \mathbf{W}_N \frac{\partial \mathbf{a}_N^*}{\partial \theta_j} \quad (\text{Apply (6.45)}) \\
 \frac{\partial \mathbf{a}_n^*}{\partial \theta_j} - \mathbf{C}_N \mathbf{W}_N \frac{\partial \mathbf{a}_n^*}{\partial \theta_j} &= \frac{\partial \mathbf{C}_N}{\partial \theta_j}(\mathbf{t}_N - \boldsymbol{\sigma}^*) \\
 \frac{\partial \mathbf{a}_n^*}{\partial \theta_j} &= \{\mathbf{I} - \mathbf{C}_N \mathbf{W}_N\}^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j}(\mathbf{t}_N - \boldsymbol{\sigma}^*).
 \end{aligned}$$

Thereby verifying (6.94). We hence have obtained all relevant quantities for computing the derivative of the model evidence with respect to the covariance parameters θ .

Chapter 7

Sparse Kernel Machines

Exercise 7.1

Consider that we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, composed of input variables and target variables, where $t_n \in \{-1, 1\}$. We model the distribution of the input variables \mathbf{x} , conditional on the class t_n to which they belong, via a Parzen kernel density estimator as in (2.249), with arbitrary kernel function $k(\mathbf{x}, \mathbf{x}')$. We attribute a prior distribution to our target variables of the form $p(t_n = -1) = p(t_n = 1) = 1/2$. It follows that

$$(7.1) \quad p(\mathbf{x}|t_n = 1) = \frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} k(\mathbf{x}, \mathbf{x}_n) \quad \text{and} \quad p(\mathbf{x}|t_n = -1) = \frac{1}{N^{(-1)}} \sum_{n \in I^{(-1)}} k(\mathbf{x}, \mathbf{x}_n),$$

where $N^{(j)}$ indicates the number of observed data points belonging to the class $t_n = j$, for $j \in \{-1, 1\}$, and $I^{(j)} = \{n \in \{1, \dots, N\} : t_n = j\}$, $j \in \{-1, 1\}$. We likewise assume (7.1) are properly normalized and nonnegative. We determine the posterior distribution of t_n as follows

$$\begin{aligned} p(t_n = 1|\mathbf{x}) &= \frac{p(\mathbf{x}|t_n = 1)p(t_n = 1)}{p(\mathbf{x}|t_n = 1)p(t_n = 1) + p(\mathbf{x}|t_n = -1)p(t_n = -1)} \\ &= \frac{p(\mathbf{x}|t_n = 1)/2}{p(\mathbf{x}|t_n = 1)/2 + p(\mathbf{x}|t_n = -1)/2} \\ &= \frac{p(\mathbf{x}|t_n = 1)}{p(\mathbf{x}|t_n = 1) + p(\mathbf{x}|t_n = -1)} \\ &= \frac{\frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} k(\mathbf{x}, \mathbf{x}_n)}{\frac{1}{N^{(1)}N^{(-1)}} [N^{(-1)} \sum_{n \in I^{(1)}} k(\mathbf{x}, \mathbf{x}_n) + N^{(1)} \sum_{n \in I^{(-1)}} k(\mathbf{x}, \mathbf{x}_n)]} \quad (\text{Apply (7.1)}) \\ (7.2) \quad p(t_n = 1|\mathbf{x}) &= \frac{N^{(-1)} \sum_{n \in I^{(1)}} k(\mathbf{x}, \mathbf{x}_n)}{N^{(-1)} \sum_{n \in I^{(1)}} k(\mathbf{x}, \mathbf{x}_n) + N^{(1)} \sum_{n \in I^{(-1)}} k(\mathbf{x}, \mathbf{x}_n)}. \end{aligned}$$

Let us observe an additional data point $\{\mathbf{x}_{N+1}, t_{N+1}\}$. From previous results, we have that the minimum misclassification-rate decision rule is given by

$$(7.3) \quad y(\mathbf{x}_{N+1}) = \begin{cases} 1 & \text{if } p(t_n = 1|\mathbf{x}_{N+1}) \geq p(t_n = -1|\mathbf{x}_{N+1}), \\ -1 & \text{otherwise.} \end{cases}$$

$$y(\mathbf{x}_{N+1}) = \begin{cases} 1 & \text{if } \frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} k(\mathbf{x}_{N+1}, \mathbf{x}_n) \geq \frac{1}{N^{(-1)}} \sum_{n \in I^{(-1)}} k(\mathbf{x}_{N+1}, \mathbf{x}_n), \\ -1 & \text{otherwise.} \end{cases} \quad (\text{Apply (7.2)})$$

We define herein

$$(7.4) \quad \bar{\mathbf{x}}^{(-1)} = \frac{1}{N^{(-1)}} \sum_{n \in I^{(-1)}} \mathbf{x}_n \quad \text{and} \quad \bar{\mathbf{x}}^{(1)} = \frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} \mathbf{x}_n.$$

By adopting $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$ in (7.3), we obtain

$$\begin{aligned} y(\mathbf{x}_{N+1}) &= \begin{cases} 1 & \text{if } \frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} \mathbf{x}_{N+1}^\top \mathbf{x}_n \geq \frac{1}{N^{(-1)}} \sum_{n \in I^{(-1)}} \mathbf{x}_{N+1}^\top \mathbf{x}_n, \\ -1 & \text{otherwise.} \end{cases} \\ &= \begin{cases} 1 & \text{if } \mathbf{x}_{N+1}^\top \mathbf{x}^{(1)} \geq \mathbf{x}_{N+1}^\top \mathbf{x}^{(-1)}, \\ -1 & \text{otherwise.} \end{cases} \quad (\text{Apply (7.4)}) \\ y(\mathbf{x}_{N+1}) &= \begin{cases} 1 & \text{if } k(\mathbf{x}_{N+1}, \mathbf{x}^{(1)}) \geq k(\mathbf{x}_{N+1}, \mathbf{x}^{(-1)}), \\ -1 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, we assign to the new input \mathbf{x}_{N+1} the class whose mean is closest to \mathbf{x}_{N+1} with respect to the kernel function $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$. If we now consider that $k(\mathbf{x}, \mathbf{x}') = \phi^\top(\mathbf{x})\phi(\mathbf{x}')$, such that we define the quantities

$$\{\bar{\phi}(\mathbf{x})\}^{(-1)} = \frac{1}{N^{(-1)}} \sum_{n \in I^{(-1)}} \phi(\mathbf{x}_n) \quad \text{and} \quad \{\bar{\phi}(\mathbf{x})\}^{(1)} = \frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} \phi(\mathbf{x}_n).$$

Hence, returning to (7.3), we obtain

$$(7.5) \quad \begin{aligned} y(\mathbf{x}_{N+1}) &= \begin{cases} 1 & \text{if } \frac{1}{N^{(1)}} \sum_{n \in I^{(1)}} \phi^\top(\mathbf{x}_{N+1})\phi(\mathbf{x}_n) \geq \frac{1}{N^{(-1)}} \sum_{n \in I^{(-1)}} \phi^\top(\mathbf{x}_{N+1})\phi(\mathbf{x}_n), \\ -1 & \text{otherwise.} \end{cases} \\ &= \begin{cases} 1 & \text{if } \phi^\top(\mathbf{x}_{N+1})\phi(\mathbf{x}^{(1)}) \geq \phi^\top(\mathbf{x}_{N+1})\phi(\mathbf{x}^{(-1)}), \\ -1 & \text{otherwise.} \end{cases} \end{aligned}$$

From (7.5), we find that we attribute to our new input \mathbf{x}_{N+1} again the class whose mean is closest to \mathbf{x}_{N+1} with respect to the inner product kernel function.

Exercise 7.2

Consider the optimization problem defined by (7.6) and (7.5) as

$$(7.6) \quad \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{subject to} \quad t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1, \quad n \in \{1, \dots, N\}.$$

Let \mathbf{w}^* and b^* denote the solution of (7.6). Let $\gamma > 0$, consider that we substitute the right-hand-side value of (7.5) with γ , such that the optimization problem in (7.6) is rewritten as

$$(7.7) \quad \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{subject to} \quad t_n(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq \gamma, \quad n \in \{1, \dots, N\}.$$

We note that $\arg \min 2^{-1} \|\mathbf{w}\|^2$ satisfies the following

$$(7.8) \quad \begin{aligned} \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\gamma \mathbf{w}\|^2 &= \arg \min_{\mathbf{w}, b} \frac{\gamma^2}{2} \|\mathbf{w}\|^2 \\ \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\gamma \mathbf{w}\|^2 &= \arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \end{aligned} \quad (\text{As } \gamma^2 > 0).$$

Hence, setting $\mathbf{w} = \gamma \mathbf{w}_0$ and $b = \gamma b_0$, and manipulating the constraints in (7.7), we obtain

$$(7.9) \quad \begin{aligned} t_n(\gamma \mathbf{w}_0^\top \phi(\mathbf{x}_n) + \gamma b_0) &\geq \gamma \\ t_n(\mathbf{w}_0^\top \phi(\mathbf{x}_n) + b_0) &\geq 1. \end{aligned}$$

From the results (7.8) and (7.9) we obtain

$$(7.10) \quad \arg \min_{\mathbf{w}_0, b_0} \frac{1}{2} \|\mathbf{w}_0\|^2 \quad \text{subject to} \quad t_n(\mathbf{w}_0^\top \phi(\mathbf{x}_n) + b_0) \geq 1, \quad n \in \{1, \dots, N\}.$$

As (7.10) is equivalent (7.6), we find that $\mathbf{w}_0 = \mathbf{w}^*$ and $b_0 = b^*$ is the solution to (7.10), and hence, $\gamma \mathbf{w}_0 = \gamma \mathbf{w}^*$ and $\gamma b_0 = \gamma b^*$ is the solution to (7.7). Note that the hyperplane defined by the solution in (7.10) is of the form

$$(7.11) \quad \left\{ \mathbf{x} \in \mathbb{R}^N : (\gamma \mathbf{w}^*)^\top \phi(\mathbf{x}) + \gamma b^* = 0 \right\}.$$

As $(\gamma \mathbf{w}^*)^\top \phi(\mathbf{x}) + \gamma b^* = 0 \iff (\mathbf{w}^*)^\top \phi(\mathbf{x}) + b^* = 0$, the hyperplane defined in (7.11) is equivalent to the hyperplane

$$\left\{ \mathbf{x} \in \mathbb{R}^N : (\mathbf{w}^*)^\top \phi(\mathbf{x}) + b^* = 0 \right\},$$

and hence the solution remains unchanged after applying the modification to the constraints.

Exercise 7.3

Consider that we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^2$, such that, $t_1 = -t_2$, i.e., the data set consists of two data points of distinct classes. We rewrite (7.7) as

$$(7.12) \quad \begin{aligned} L(\mathbf{w}, b, \mathbf{a}) &= \frac{1}{2} \|\mathbf{w}\|^2 - a_1 \{t_1(\mathbf{w}^\top \phi(\mathbf{x}_1) + b) - 1\} - \\ &\quad - a_2 \{t_2(\mathbf{w}^\top \phi(\mathbf{x}_2) + b) - 1\} \\ L(\mathbf{w}, b, \mathbf{a}) &= \frac{1}{2} \|\mathbf{w}\|^2 + a_1 \{t_2(\mathbf{w}^\top \phi(\mathbf{x}_1) + b) - 1\} + \\ &\quad - a_2 \{t_2(\mathbf{w}^\top \phi(\mathbf{x}_2) + b) - 1\} \end{aligned} \quad (\text{Apply } t_1 = -t_2).$$

Differentiating (7.12) with respect to b and solving for 0, we obtain

$$(7.13) \quad \begin{aligned} \frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial b} &= 0 \\ a_1 t_2 - a_2 t_2 &= 0 \\ a_1 &= a_2. \end{aligned}$$

Note that, from previous results, we know that when the margin is maximized, we shall have that $t_1(\mathbf{w}^\top \phi(\mathbf{x}_1) + b) = t_2(\mathbf{w}^\top \phi(\mathbf{x}_2) + b) = 1$. Hence, (7.12) will not be dependent on a_1 or a_2 , and we may choose their values arbitrarily. For simplicity, we write $\phi(\mathbf{x}_1) = \phi_1$ and $\phi(\mathbf{x}_2) = \phi_2$. Consequently, differentiating (7.12) with respect to \mathbf{w} and solving for $\mathbf{0}$, we obtain

$$(7.14) \quad \begin{aligned} \frac{\partial L(\mathbf{w}, b, \mathbf{1})}{\partial \mathbf{w}} &= \mathbf{0} \\ \mathbf{w} + a_1 t_2 \phi_1 - a_2 t_2 \phi_2 &= \mathbf{0} \quad (\text{Apply (C.19)}) \\ \mathbf{w} + a_2 t_2 \phi_1 - a_2 t_2 \phi_2 &= \mathbf{0} \quad (\text{Apply (7.13)}) \\ \mathbf{w} &= a_2 t_2 \{\phi_2 - \phi_1\}. \end{aligned}$$

As previously stated, once the margin is maximized, the following conditions will be satisfied:

$$\begin{cases} t_1(\mathbf{w}^\top \phi_1 + b) &= 1, \\ t_2(\mathbf{w}^\top \phi_2 + b) &= 1. \end{cases}$$

Note that these conditions are equivalent to

$$(7.15) \quad \begin{cases} \mathbf{w}^\top \phi_1 + b &= t_1, \\ \mathbf{w}^\top \phi_2 + b &= t_2. \end{cases}$$

By summing the conditions in (7.15), we obtain

$$(7.16) \quad \begin{aligned} \mathbf{w}^\top \{\phi_1 + \phi_2\} + 2b &= t_1 + t_2 \\ a_2 t_2 \{\phi_2 - \phi_1\}^\top \{\phi_1 + \phi_2\} + 2b &= t_2 - t_1 \quad (\text{Apply (7.14) and } t_1 = -t_2) \\ b &= -\frac{a_2 t_2}{2} \{\phi_2 - \phi_1\}^\top \{\phi_1 + \phi_2\} \\ b &= -\frac{a_2 t_2}{2} \{\phi_2^\top \phi_2 - \phi_1^\top \phi_1\}. \end{aligned}$$

We now aim to determine the value a_2 which satisfies (7.15) by substituting (7.14) and (7.16) into the first constraint in (7.15), obtaining the following

$$\begin{aligned}
 a_2 t_2 \{\phi_2 - \phi_1\}^\top \phi_1 - \frac{a_2 t_2}{2} \{\phi_2^\top \phi_2 - \phi_1^\top \phi_1\} &= t_1 \\
 a_2 t_1 \left[\frac{1}{2} \{\phi_2^\top \phi_2 - \phi_1^\top \phi_1\} - \{\phi_2 - \phi_1\}^\top \phi_1 \right] &= t_1 \quad (\text{Apply } t_1 = -t_2) \\
 a_2 \left[\frac{1}{2} \phi_2^\top \phi_2 - \phi_1^\top \phi_2 + \frac{1}{2} \phi_2^\top \phi_2 \right] &= 1 \\
 \frac{a_2}{2} [\phi_2^\top \phi_2 - 2\phi_1^\top \phi_2 + \phi_2^\top \phi_2] &= 1 \\
 \frac{a_2}{2} \{\phi_1 - \phi_2\}^\top \{\phi_1 - \phi_2\} &= 1 \\
 a_2 &= \frac{2}{\{\phi_1 - \phi_2\}^\top \{\phi_1 - \phi_2\}}.
 \end{aligned}$$

Hence, as long as $\phi_1 \neq \phi_2$, we find that a data set consisting of two data points of distinct classes has its maximum-margin hyperplane defined by $(\mathbf{w}^*)^\top \phi(\mathbf{x}) + b^* = 0$, where the parameters are as follows

$$\begin{aligned}
 \mathbf{w}^* &= t_2 \frac{2\{\phi_2 - \phi_1\}}{\{\phi_1 - \phi_2\}^\top \{\phi_1 - \phi_2\}} \\
 b^* &= -t_2 \frac{\{\phi_2^\top \phi_2 - \phi_1^\top \phi_1\}}{\{\phi_1 - \phi_2\}^\top \{\phi_1 - \phi_2\}}.
 \end{aligned}$$

Interestingly, if ϕ_1 and ϕ_2 are orthonormal, it follows that $b^* = 0$ and $\mathbf{w}^* = t_2 \{\phi_2 - \phi_1\}$.

Exercise 7.4

Consider that we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, such that $\{\mathbf{x}_n\}_{n=1}^N$ are linearly separable, and we train a support vector machine on this data, obtaining the corresponding model parameters. We aim to demonstrate that the margin corresponding to the maximum-margin hyperplane (denoted as ρ) is such that (7.123) holds. Take an arbitrary data point $\{\mathbf{x}_j, t_j\}$ such that the constraint (7.5) is met with equality (we may do this since a properly trained model possesses at least two data points that satisfy these conditions). It follows that

$$\begin{aligned}
 \rho &= \frac{t_j(\mathbf{w}^\top \phi(\mathbf{x}_j) + b)}{\|\mathbf{w}\|} && \text{(Apply (7.2))} \\
 &= \frac{1}{\|\mathbf{w}\|} && (t_j(\mathbf{w}^\top \phi(\mathbf{x}_j) + b) = 1) \\
 \rho^2 &= \frac{1}{\|\mathbf{w}\|^2} \\
 (7.17) \quad \frac{1}{\rho^2} &= \|\mathbf{w}\|^2 \\
 &= \left[\sum_{n=1}^N a_n t_n \phi(\mathbf{x}_n) \right]^\top \left[\sum_{m=1}^N a_m t_m \phi(\mathbf{x}_m) \right] && \text{(Apply (7.8))} \\
 &= \sum_{n=1}^N \sum_{m=1}^N a_n t_n a_m t_m \phi^\top(\mathbf{x}_n) \phi(\mathbf{x}_m) \\
 &= \sum_{n=1}^N a_n t_n \left[\sum_{m=1}^N a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b - b \right] && \text{(Apply (6.1))} \\
 &= \sum_{n=1}^N a_n t_n [y(\mathbf{x}_n) - b] && \text{(Apply (7.13))} \\
 &= \sum_{n=1}^N a_n t_n y(\mathbf{x}_n) - b \sum_{n=1}^N a_n t_n \\
 &= \sum_{n=1}^N a_n \{t_n y(\mathbf{x}_n) - 1 + 1\} && \text{(Apply (7.12))} \\
 &= \sum_{n=1}^N a_n \{t_n y(\mathbf{x}_n) - 1\} + \sum_{n=1}^N a_n \\
 &\frac{1}{\rho^2} = \sum_{n=1}^N a_n && \text{(Apply (7.16)).}
 \end{aligned}$$

Hence, we obtain (7.123).

Exercise 7.5

Considering the same framework as [Exercise 7.4](#), we aim to demonstrate the validity of [\(7.124\)](#) and [\(7.125\)](#). First, note that under those conditions, we may rewrite [\(7.10\)](#) as

$$\begin{aligned}
 \tilde{L}(\mathbf{a}) &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m) \\
 &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N a_n t_n \left[\sum_{m=1}^N a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right] \\
 &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N a_n t_n \left[\sum_{m=1}^N a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b - b \right] \\
 &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N a_n t_n [y(\mathbf{x}_n) - b] \quad (\text{Apply (7.13)}) \\
 &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N a_n t_n y(\mathbf{x}_n) + b \frac{1}{2} \sum_{n=1}^N a_n t_n \\
 &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N a_n \{t_n y(\mathbf{x}_n) - 1 + 1\} \quad (\text{Apply (7.12)}) \\
 &= \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N a_n \{t_n y(\mathbf{x}_n) - 1\} - \frac{1}{2} \sum_{n=1}^N a_n \\
 \tilde{L}(\mathbf{a}) &= \frac{1}{2} \sum_{n=1}^N a_n \quad (\text{Apply (7.16)}) \\
 2\tilde{L}(\mathbf{a}) &= \frac{1}{2} \sum_{n=1}^N a_n \\
 2\tilde{L}(\mathbf{a}) &= \frac{1}{\rho^2} \quad (\text{Apply (7.123)}).
 \end{aligned}$$

Hence, we prove [\(7.124\)](#) holds. [\(7.125\)](#) holds, as seen in [\(7.17\)](#), in [Exercise 7.4](#)

Exercise 7.6

Let us consider the logistic regression context, where our target variables are such that $t_n \in \{-1, 1\}$, and $p(t_n = 1|y_n) = \sigma(y_n)$, consequently $p(t_n = -1|y_n) = \sigma(-y_n)$, and y_n is as in (7.1). From (7.46), we may write the likelihood function associated with a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, and corresponding logarithm, as

$$\begin{aligned}
 p(\mathbf{t}|\mathbf{y}) &= \prod_{n=1}^N p(t_n|y_n) \\
 &= \prod_{n=1}^N \sigma(y_n t_n) \\
 \log p(\mathbf{t}|\mathbf{y}) &= \sum_{n=1}^N \log \sigma(y_n t_n) \\
 &= \sum_{n=1}^N \log \left(\frac{1}{1 + e^{-y_n t_n}} \right) \quad (\text{Apply (3.6)}) \\
 &= - \sum_{n=1}^N \log(1 + e^{-y_n t_n}) \\
 -\log p(\mathbf{t}|\mathbf{y}) &= \sum_{n=1}^N \log(1 + e^{-y_n t_n}) \\
 (7.18) \quad -\log p(\mathbf{t}|\mathbf{y}) &= \sum_{n=1}^N E_{\text{LR}}(y_n t_n) \quad (\text{Apply (7.48)}).
 \end{aligned}$$

By adding $\lambda||\mathbf{w}||^2$ on both sides of (7.18), we obtain (7.47), as desired.

Exercise 7.7

We aim to obtain the dual Lagrangian in (7.61) by differentiating (7.56) with respect to \mathbf{w} , b , ξ_n and $\hat{\xi}_n$, and solving for 0, and substituting the corresponding solutions in (7.56). First, we obtain the derivative of (7.56) with respect to \mathbf{w} as

$$\begin{aligned}
 \mathbf{0} &= \frac{\partial L}{\partial \mathbf{w}} \\
 \mathbf{0} &= \frac{\partial}{\partial \mathbf{w}} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n) \right] \\
 &= \frac{\partial}{\partial \mathbf{w}} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + \mathbf{w}^\top \phi(\mathbf{x}_n) + b - t_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - \mathbf{w}^\top \phi(\mathbf{x}_n) - b + t_n) \right] \quad (\text{Apply (7.1)}) \\
 &= \mathbf{w} - \sum_{n=1}^N a_n \phi(\mathbf{x}_n) + \sum_{n=1}^N \hat{a}_n \phi(\mathbf{x}_n) \quad (\text{Apply (C.19)}) \\
 \mathbf{w} &= \sum_{n=1}^N a_n \phi(\mathbf{x}_n) - \sum_{n=1}^N \hat{a}_n \phi(\mathbf{x}_n) \\
 (7.19) \quad \mathbf{w} &= \sum_{n=1}^N (a_n - \hat{a}_n) \phi(\mathbf{x}_n).
 \end{aligned}$$

We now obtain the derivative of (7.56) with respect to b as

$$\begin{aligned}
 0 &= \frac{\partial L}{\partial b} \\
 0 &= \frac{\partial}{\partial b} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n) \right] \\
 &= \frac{\partial}{\partial b} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + \mathbf{w}^\top \phi(\mathbf{x}_n) + b - t_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - \mathbf{w}^\top \phi(\mathbf{x}_n) - b + t_n) \right] \quad (\text{Apply (7.1)}) \\
 &= \sum_{n=1}^N \hat{a}_n - \sum_{n=1}^N a_n \\
 (7.20) \quad 0 &= \sum_{n=1}^N (\hat{a}_n - a_n).
 \end{aligned}$$

We obtain the derivative of (7.56) with respect to ξ_n as

$$\begin{aligned}
 0 &= \frac{\partial L}{\partial \xi_n} \\
 0 &= \frac{\partial}{\partial \xi_n} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n) \right] \\
 &= \frac{\partial}{\partial \xi_n} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + \mathbf{w}^\top \phi(\mathbf{x}_n) + b - t_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - \mathbf{w}^\top \phi(\mathbf{x}_n) - b + t_n) \right] \quad (\text{Apply (7.1)}) \\
 &= C - \mu_n - a_n \\
 (7.21) \quad \mu_n + a_n &= C.
 \end{aligned}$$

Lastly, we obtain the derivative of (7.56) with respect to $\hat{\xi}_n$ as

$$\begin{aligned}
 0 &= \frac{\partial L}{\partial \hat{\xi}_n} \\
 0 &= \frac{\partial}{\partial \hat{\xi}_n} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n) \right] \\
 &= \frac{\partial}{\partial \hat{\xi}_n} \left[C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N a_n (\epsilon + \xi_n + \mathbf{w}^\top \phi(\mathbf{x}_n) + b - t_n) + \right. \\
 &\quad \left. - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - \mathbf{w}^\top \phi(\mathbf{x}_n) - b + t_n) \right] \quad (\text{Apply (7.1)}) \\
 &= C - \hat{\mu}_n - \hat{a}_n \\
 (7.22) \quad \hat{\mu}_n + \hat{a}_n &= C.
 \end{aligned}$$

We now rewrite (7.56) as follows

$$\begin{aligned}
 L &= C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) + \\
 &\quad - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n) \\
 &= \sum_{n=1}^N (C \xi_n + C \hat{\xi}_n - \mu_n \xi_n - \hat{\mu}_n \hat{\xi}_n - a_n \{\epsilon + \xi_n + y_n - t_n\} + \\
 &\quad - \hat{a}_n \{\epsilon + \hat{\xi}_n - y_n + t_n\}) + \frac{1}{2} \|\mathbf{w}\|^2 \\
 &= \sum_{n=1}^N (C \xi_n + C \hat{\xi}_n - (\mu_n + a_n) \xi_n - (\hat{\mu}_n + \hat{a}_n) \hat{\xi}_n + \\
 &\quad - a_n \{\epsilon + y_n - t_n\} - \hat{a}_n \{\epsilon - y_n + t_n\}) + \frac{1}{2} \|\mathbf{w}\|^2 \quad (\text{Apply (7.21) and (7.22)}) \\
 L &= \sum_{n=1}^N (\hat{a}_n - a_n) y_n + \frac{1}{2} \|\mathbf{w}\|^2 - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n) t_n.
 \end{aligned}$$

We thereafter proceed as

$$\begin{aligned}
 L &= \sum_{n=1}^N (\hat{a}_n - a_n)(\mathbf{w}^\top \phi(\mathbf{x}_n) + b) + \frac{1}{2} \|\mathbf{w}\|^2 + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n && \text{(Apply (7.1))} \\
 &= \sum_{n=1}^N (\hat{a}_n - a_n)\mathbf{w}^\top \phi(\mathbf{x}_n) + b \sum_{n=1}^N (\hat{a}_n - a_n) + \frac{1}{2} \|\mathbf{w}\|^2 + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n \\
 &= \sum_{n=1}^N (\hat{a}_n - a_n)\mathbf{w}^\top \phi(\mathbf{x}_n) + \frac{1}{2} \|\mathbf{w}\|^2 + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n && \text{(Apply (7.20))} \\
 &= - \sum_{n=1}^N (a_n - \hat{a}_n)\mathbf{w}^\top \phi(\mathbf{x}_n) + \frac{1}{2} \|\mathbf{w}\|^2 + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n \\
 &= - \sum_{n=1}^N (a_n - \hat{a}_n) \left[\sum_{m=1}^N (a_m - \hat{a}_m)\phi(\mathbf{x}_m) \right]^\top \phi(\mathbf{x}_n) + \\
 &\quad + \frac{1}{2} \left[\sum_{m=1}^N (a_m - \hat{a}_m)\phi(\mathbf{x}_m) \right]^\top \left[\sum_{n=1}^N (a_n - \hat{a}_n)\phi(\mathbf{x}_n) \right] + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n && \text{(Apply (7.19))} \\
 &= - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N (a_n - \hat{a}_n)(a_m - \hat{a}_m)\phi^\top(\mathbf{x}_n)\phi(\mathbf{x}_m) + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n \\
 L &= - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N (a_n - \hat{a}_n)(a_m - \hat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) + \\
 &\quad - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n && \text{(Apply (6.1)).}
 \end{aligned}$$

Hence, we obtain (7.61).

Exercise 7.8

From previous results, we find that the parameters which define the solution to the regression support vector machine satisfy a number of constraints, amongst which we have (7.67) and (7.68). It follows directly from (7.67) that for a training data point n

$$(a_n - C)\xi_n = 0 \quad \text{and} \quad \xi_n > 0 \quad \Rightarrow \quad a_n = C.$$

Conversely, from (7.68), it follows also that for the training data point n

$$(\hat{a}_n - C)\hat{\xi}_n = 0 \quad \text{and} \quad \hat{\xi}_n > 0 \quad \Rightarrow \quad \hat{a}_n = C.$$

Exercise 7.9

Consider the regression relevance vector machine framework, such that we observe a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$, where we attribute to \mathbf{t} , conditional on \mathbf{w} , an N -dimensional normal distribution with mean $\Phi\mathbf{w} \in \mathbb{R}^N$ and precision matrix $\beta I \in \mathbb{R}^{N \times N}$, where $\beta > 0$, and we attribute to \mathbf{w} an M -dimensional multivariate normal distribution prior, with mean $\mathbf{0} \in \mathbb{R}^M$ and precision matrix $\mathbf{A} = \text{diag}(\alpha_i) \in \mathbb{R}^{M \times M}$, where $\alpha_i > 0$. Note that as the mean of \mathbf{t} conditional on \mathbf{w} is a linear function of \mathbf{w} , we may utilize the linear-Gaussian model result (2.116) to determine the posterior of \mathbf{w} given \mathbf{t} as an M -dimensional multivariate normal with mean vector

$$\mathbb{E}[\mathbf{w}|\mathbf{t}] = \beta\{\mathbf{A} + \beta\Phi^\top\Phi\}^{-1}\Phi^\top\mathbf{t},$$

and covariance matrix

$$\text{Var}[\mathbf{w}|\mathbf{t}] = \{\mathbf{A} + \beta\Phi^\top\Phi\}^{-1}.$$

Exercise 7.10

Consider the same framework as in [Exercise 7.9](#). We aim to determine the marginal distribution of \mathbf{t} , firstly by directly completing the squares, as follows

$$\begin{aligned}\log p(\mathbf{t}|\mathbf{X}, \alpha, \beta) &\propto \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) + \log p(\mathbf{w}|\alpha) \\ &\propto -\frac{\beta}{2}(\mathbf{t} - \Phi\mathbf{w})^\top(\mathbf{t} - \Phi\mathbf{w}) - \frac{1}{2}\mathbf{w}^\top \mathbf{A}\mathbf{w} \\ &= -\frac{\beta}{2}\mathbf{t}^\top \mathbf{t} + \beta\mathbf{w}^\top \Phi^\top \mathbf{t} - \frac{\beta}{2}\mathbf{w}^\top \Phi^\top \Phi\mathbf{w} - \frac{1}{2}\mathbf{w}^\top \mathbf{A}\mathbf{w} \\ &= -\frac{1}{2}\mathbf{w}^\top [\beta\Phi^\top \Phi + \mathbf{A}]\mathbf{w} - \frac{\beta}{2}\mathbf{t}^\top \mathbf{t} + \beta\mathbf{w}^\top [\beta\Phi^\top \Phi + \mathbf{A}][\beta\Phi^\top \Phi + \mathbf{A}]^{-1}\Phi^\top \mathbf{t} \\ \log p(\mathbf{t}|\mathbf{X}, \alpha, \beta) &\propto -\frac{1}{2}\{\mathbf{w} - [\beta\Phi^\top \Phi + \mathbf{A}]^{-1}\mathbf{w}\}^\top [\beta\Phi^\top \Phi + \mathbf{A}]\{\mathbf{w} - [\beta\Phi^\top \Phi + \mathbf{A}]^{-1}\mathbf{w}\} + \\ &\quad - \frac{\beta}{2}\mathbf{t}^\top \mathbf{t} + \frac{\beta^2}{2}\mathbf{t}^\top \Phi[\beta\Phi^\top \Phi + \mathbf{A}]^{-1}\Phi^\top \mathbf{t}.\end{aligned}$$

Having found a quadratic form for \mathbf{w} , we integrate it out, yielding

$$\begin{aligned}\log p(\mathbf{t}|\mathbf{X}, \alpha, \beta) &\propto -\frac{\beta}{2}\mathbf{t}^\top \mathbf{t} + \frac{\beta^2}{2}\mathbf{t}^\top \Phi[\beta\Phi^\top \Phi + \mathbf{A}]^{-1}\Phi^\top \mathbf{t} \\ &= -\frac{1}{2}\mathbf{t}^\top [\beta\mathbf{I} - \beta^2\Phi[\beta\Phi^\top \Phi + \mathbf{A}]^{-1}\Phi^\top]\mathbf{t} \\ (7.23) \quad \log p(\mathbf{t}|\mathbf{X}, \alpha, \beta) &\propto -\frac{1}{2}\mathbf{t}^\top [\beta^{-1}\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^\top]^{-1}\mathbf{t} \quad (\text{Apply (2.289)}).\end{aligned}$$

From the form in (7.23), we find that the marginal distribution of \mathbf{t} is an N -dimensional multivariate normal with mean $\mathbf{0} \in \mathbb{R}$ and covariance as in (7.86).

Exercise 7.11

We consider the same framework as in [Exercise 7.9](#) once more, and that we again aim to determine the marginal distribution of \mathbf{t} . Once more utilizing a linear-Gaussian model result, yet this time [\(2.115\)](#), we find that the marginal distribution of \mathbf{t} is given by an N -dimensional multivariate normal with mean

$$\mathbb{E}[\mathbf{t}] = \mathbf{0},$$

and covariance matrix

$$\text{Var}[\mathbf{t}] = \beta^{-1} \mathbf{I} + \Phi \mathbf{A}^{-1} \Phi^\top.$$

Exercise 7.12

Returning to the framework of [Exercise 7.9](#), we aim to determine re-estimation equations for the parameters β and α , by direct maximization of the marginal distribution of \mathbf{t} . Prior to this, note that

$$\begin{aligned} \mathbf{C}^{-1} &= (\beta^{-1}\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^\top)^{-1} \\ &= \beta\mathbf{I} - \beta^2\Phi\{\beta\Phi^\top\Phi + \mathbf{A}\}^{-1}\Phi^\top && \text{(Apply (2.289))} \\ \mathbf{C}^{-1} &= \beta\mathbf{I} - \beta^2\Phi\Sigma\Phi^\top && \text{(Apply (7.83))} \end{aligned} \quad (7.24)$$

$$\begin{aligned} \mathbf{C}^{-1}\mathbf{C}^{-1} &= [\beta\mathbf{I} - \beta^2\Phi\Sigma\Phi^\top][\beta\mathbf{I} - \beta^2\Phi\Sigma\Phi^\top] \\ &= \beta^2\mathbf{I} - 2\beta^3\Phi\Sigma\Phi^\top + \\ &\quad + \beta^4\Phi\Sigma\Phi^\top\Phi\Sigma\Phi^\top \\ \mathbf{C}^{-1}\mathbf{C}^{-1} &= \beta^2[\mathbf{I} - 2\beta\Sigma\Phi^\top + \beta^2\Phi\Sigma\Phi^\top\Phi\Sigma\Phi^\top]. \end{aligned} \quad (7.25)$$

Hence, we differentiate [\(7.85\)](#) with respect to β and solve for 0, as follows

$$\begin{aligned} \frac{\partial \log p(\mathbf{t}|\mathbf{C})}{\partial \beta} &= 0 \\ \frac{\partial}{\partial \beta} \left[-\frac{1}{2}\{N \log(2\pi) + \log|\mathbf{C}| + \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}\} \right] &= 0 \\ -\frac{1}{2} \text{tr} \left(\mathbf{C}^{-1} \frac{\partial [\beta^{-1}\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^\top]}{\partial \beta} \right) + \\ + \frac{1}{2} \mathbf{t}^\top \mathbf{C}^{-1} \frac{\partial [\beta^{-1}\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^\top]}{\partial \beta} \mathbf{C}^{-1} \mathbf{t} &= 0 && \text{(Apply (7.86), (C.21) and (C.22))} \\ \frac{1}{2\beta^2} \text{tr}(\mathbf{C}^{-1}) - \frac{1}{2\beta^2} \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{C}^{-1} \mathbf{t} &= 0 \\ \frac{1}{2\beta^2} \text{tr}(\beta\mathbf{I} - \beta^2\Phi\Sigma\Phi^\top) + \\ - \frac{1}{2\beta^2} \mathbf{t}^\top \beta^2 [\mathbf{I} - 2\beta\Sigma\Phi^\top + \beta^2\Phi\Sigma\Phi^\top\Phi\Sigma\Phi^\top] \mathbf{t} &= 0 && \text{(Apply (7.24) and (7.25))} \\ \frac{1}{\beta}(N - \text{tr}(\Sigma\beta\Phi^\top\Phi)) - (\mathbf{t}^\top \mathbf{t} - 2\mathbf{t}^\top \mathbf{m} + \mathbf{m}^\top \mathbf{m}) &= 0 && \text{(Apply (7.82))} \\ \frac{1}{\beta}(N - \text{tr}(\Sigma[\beta\Phi^\top\Phi + \mathbf{A} - \mathbf{A}])) + \\ - (\mathbf{t} - \Phi\mathbf{m})^\top (\mathbf{t} - \Phi\mathbf{m}) &= 0 \\ \frac{1}{\beta}(N - \text{tr}(\Sigma[\Sigma^{-1} + \mathbf{A} - \mathbf{A}])) - \|\mathbf{t} - \Phi\mathbf{m}\|^2 &= 0 && \text{(Apply (7.83))} \\ \frac{1}{\beta}(N - \text{tr}(\mathbf{I} - \Sigma\mathbf{A})) - \|\mathbf{t} - \Phi\mathbf{m}\|^2 &= 0 \\ \frac{1}{\beta} \left(N - \sum_{j=1}^M \{\mathbf{I} - \Sigma\mathbf{A}\}_{j,j} \right) - \|\mathbf{t} - \Phi\mathbf{m}\|^2 &= 0 \\ \frac{1}{\beta} \left(N - \sum_{j=1}^M \gamma_j \right) - \|\mathbf{t} - \Phi\mathbf{m}\|^2 &= 0 && \text{(Apply (7.89))} \\ \frac{\|\mathbf{t} - \Phi\mathbf{m}\|^2}{N - \sum_{j=1}^M \gamma_j} &= \frac{1}{\beta}. \end{aligned}$$

Hence, we derive (7.88). We now dissect two terms of (7.85) which are dependent on α , and thereafter obtain their derivative with respect to α_j , as follows

$$\begin{aligned}
 \log|\mathbf{C}| &= \log|\beta^{-1}\mathbf{I} + \Phi\mathbf{A}^{-1}\Phi^\top| && \text{(Apply (7.86))} \\
 &= \log|\beta^{-1}\mathbf{I} + \mathbf{A}^{-1}\Phi^\top\Phi| && \text{(Apply (C.14))} \\
 &= \log|\beta^{-1}\mathbf{A}^{-1}[\mathbf{A} + \beta\Phi^\top\Phi]| \\
 &= \log|\beta^{-1}\mathbf{A}^{-1}| + \log|\mathbf{A} + \beta\Phi^\top\Phi| \\
 \log|\mathbf{C}| &= -\frac{M}{2}\log\beta - \sum_{i=1}^M \log\alpha_i + \log|\mathbf{A} + \beta\Phi^\top\Phi| && \text{(Apply (C.13))} \\
 \frac{\partial \log|\mathbf{C}|}{\partial \alpha_j} &= -\frac{1}{\alpha_j} + \text{tr}\left(\{\mathbf{A} + \beta\Phi^\top\Phi\}^{-1}\frac{\partial \mathbf{A} + \beta\Phi^\top\Phi}{\partial \alpha_j}\right) && \text{(Apply (C.22))} \\
 &= -\frac{1}{\alpha_j} + \text{tr}(\Sigma \mathbf{I}^{(j)}) && \text{(Apply (7.83))} \\
 (7.26) \quad \frac{\partial \log|\mathbf{C}|}{\partial \alpha_j} &= -\frac{1}{\alpha_j} + \Sigma_{j,j},
 \end{aligned}$$

where $\mathbf{I}^{(j)}$ indicates an $M \times M$ -dimensional matrix composed of zeroes, except at the (j, j) -th coordinate, which is such that $\mathbf{I}_{j,j}^{(j)} = 1$. It follows also that

$$\begin{aligned}
 \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} &= \mathbf{t}^\top \{\beta\mathbf{I} - \beta^2\Phi\{\beta\Phi^\top\Phi + \mathbf{A}\}^{-1}\Phi^\top\} \mathbf{t} && \text{(Apply (7.24))} \\
 &= \beta\mathbf{t}^\top \mathbf{t} - \beta^2\mathbf{t}^\top\Phi\{\beta\Phi^\top\Phi + \mathbf{A}\}^{-1}\Phi^\top\mathbf{t} \\
 \frac{\partial \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}}{\partial \alpha_j} &= \beta^2\mathbf{t}^\top\Phi\{\beta\Phi^\top\Phi + \mathbf{A}\}^{-1}\mathbf{I}^{(j)}\{\beta\Phi^\top\Phi + \mathbf{A}\}^{-1}\Phi^\top\mathbf{t} && \text{(Apply (C.21))} \\
 &= \beta^2\mathbf{t}^\top\Phi\Sigma\mathbf{I}^{(j)}\Sigma\Phi^\top\mathbf{t} && \text{(Apply (7.83))} \\
 &= \mathbf{m}^\top \mathbf{I}^{(j)} \mathbf{m} && \text{(Apply (7.82))} \\
 (7.27) \quad \frac{\partial \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}}{\partial \alpha_j} &= m_j^2.
 \end{aligned}$$

It follows that, differentiating (7.85) with respect to α_j , we obtain

$$\begin{aligned}
 \frac{\partial \log p(\mathbf{t}|\mathbf{C})}{\partial \alpha_j} &= 0 \\
 \frac{\partial}{\partial \alpha_j} \left[-\frac{1}{2}\{N\log(2\pi) + \log|\mathbf{C}| + \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}\} \right] &= 0 \\
 \frac{1}{2\alpha_j} - \frac{1}{2}\Sigma_{j,j} - \frac{1}{2}m_j^2 &= 0 \quad \text{(Apply (7.26) and (7.27))} \\
 \frac{1 - \alpha_j\Sigma_{j,j}}{\alpha_j} &= m_j^2 \\
 \frac{\gamma_j}{m_j^2} &= \alpha_j \quad \text{(Apply (7.89)).}
 \end{aligned}$$

Hence, we derive (7.87), and thereby conclude the demonstration.

Exercise 7.13

Consider again the framework of [Exercise 7.9](#), modified in the following manner: we now attribute Gamma hyperpriors to the variance in the precision matrix, particularly we have that α_i follows a Gamma distribution with parameters $a > 0$ and $b > 0$, $\forall i \in \{1, \dots, M\}$ (the "hyperhyperparameters" are the same across all i), and that we attribute to β , similarly, a Gamma distribution with parameters $a_0 > 0$ and $b_0 > 0$. Moreover, in order to determine the estimation formulae of α and β , we maximize $p(\mathbf{t}, \alpha, \beta | \mathbf{X})$ directly, with respect to α and β . It follows that we write $p(\mathbf{t}, \alpha, \beta | \mathbf{X})$ as

$$(7.28) \quad p(\mathbf{t}, \alpha, \beta | \mathbf{X}) = p(\mathbf{t} | \alpha, \beta, \mathbf{X}) p(\alpha) p(\beta).$$

Wherein the from of $p(\mathbf{t} | \alpha, \beta, \mathbf{X})$ is equivalent to the marginal distribution of \mathbf{t} determined in [Exercise 7.11](#). It follows that

$$\begin{aligned} p(\mathbf{t}, \alpha, \beta | \mathbf{X}) &= (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} \right\} p(\alpha) p(\beta) \quad (\text{Apply (2.118)}) \\ &= (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} \right\} \left[\prod_{j=1}^M p(\alpha_j) \right] \times \\ &\quad \times \frac{b_0^{a_0}}{\Gamma(a_0)} \beta^{a_0-1} \exp\{-b_0\beta\} \quad (\text{Apply (2.146)}) \\ (7.29) \quad p(\mathbf{t}, \alpha, \beta | \mathbf{X}) &= (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} \right\} \times \\ &\quad \times \left[\prod_{j=1}^M \frac{b^a}{\Gamma(a)} \alpha_j^{a-1} \exp\{-b\alpha_j\} \right] \times \\ &\quad \times \frac{b_0^{a_0}}{\Gamma(a_0)} \beta^{a_0-1} \exp\{-b_0\beta\} \quad (\text{Apply (2.146)}) \end{aligned}$$

Applying the logarithm in (7.29) and discarding terms independent of α or β , we obtain

$$\begin{aligned} (7.30) \quad \log p(\mathbf{t}, \alpha, \beta | \mathbf{X}) &\propto -\frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} + \\ &\quad + \sum_{j=1}^M (a-1) \log \alpha_j - b \sum_{j=1}^M \log \alpha_j + (a_0-1) \log \beta - b_0 \beta. \end{aligned}$$

Note that (7.30) may be rewritten as

$$(7.31) \quad \log p(\mathbf{t}, \alpha, \beta | \mathbf{X}) \propto \log p(\mathbf{t} | \mathbf{C}) + \sum_{j=1}^M (a-1) \log \alpha_j - b \sum_{j=1}^M \log \alpha_j + (a_0-1) \log \beta - b_0 \beta,$$

where $p(\mathbf{t} | \mathbf{C})$ is the marginal likelihood function seen in [Exercise 7.12 \(\(7.85\)\)](#). Note that, as the derivatives of $p(\mathbf{t} | \mathbf{C})$ with respect to α and β were computed in that Exercise, we may simply herein utilize the results therein that

$$(7.32) \quad \frac{\partial \log p(\mathbf{t} | \mathbf{C})}{\partial \beta} = \frac{1}{2\beta} \left(N - \sum_{j=1}^M \gamma_j \right) - \frac{1}{2} \|\mathbf{t} - \Phi \mathbf{m}\|^2.$$

$$(7.33) \quad \frac{\partial \log p(\mathbf{t} | \mathbf{C})}{\partial \alpha_j} = \frac{1}{2\alpha_j} - \frac{1}{2} \Sigma_{j,j} - \frac{1}{2} m_j^2.$$

It follows that, differentiating (7.31) with respect to β and solving for 0, we obtain

$$\begin{aligned} \frac{\partial \log p(\mathbf{t}, \boldsymbol{\alpha}, \beta | \mathbf{X})}{\partial \beta} &= 0 \\ \frac{\partial \log p(\mathbf{t} | \mathbf{C})}{\partial \beta} + \frac{a_0 - 1}{\beta} - b_0 &= 0 \quad (\text{Apply (7.31)}) \\ \frac{1}{2\beta} \left(N - \sum_{j=1}^M \gamma_j \right) - \frac{1}{2} \|\mathbf{t} - \Phi \mathbf{m}\|^2 + \frac{a_0 - 1}{\beta} - b_0 &= 0 \quad (\text{Apply (7.32)}) \\ \frac{\frac{1}{2} \|\mathbf{t} - \Phi \mathbf{m}\|^2 + b_0}{\frac{1}{2}(N - \sum_{j=1}^M \gamma_j) + a_0 - 1} &= \frac{1}{\beta} \\ \frac{\|\mathbf{t} - \Phi \mathbf{m}\|^2 + 2b_0}{(N - \sum_{j=1}^M \gamma_j) + 2(a_0 - 1)} &= \frac{1}{\beta}. \end{aligned}$$

Hence, we derive a new re-estimation equation. Note that, if $a_0 = 1$ and $b_0 \rightarrow 0$, which corresponds to an improper prior over the positive real line for β , we return to the re-estimation equation (7.88). We now differentiate (7.31) with respect to α_j and solve for 0, obtaining

$$\begin{aligned} \frac{\partial \log p(\mathbf{t}, \boldsymbol{\alpha}, \beta | \mathbf{X})}{\partial \alpha_j} &= 0 \\ \frac{\partial \log p(\mathbf{t} | \mathbf{C})}{\partial \alpha_j} + \frac{a - 1}{\alpha_j} - b &= 0 \quad (\text{Apply (7.31)}) \\ \frac{1}{2\alpha_j} - \frac{1}{2} \Sigma_{j,j} - \frac{1}{2} m_j^2 + \frac{a - 1}{\alpha_j} - b &= 0 \quad (\text{Apply (7.33)}) \\ \frac{1 - \alpha_j \Sigma_{j,j} + 2(a - 1)}{2\alpha_j} - \frac{m_j^2 + 2b}{2} &= 0 \\ \frac{\gamma_j + 2(a - 1)}{\alpha_j} - (m_j^2 + 2b) &= 0 \quad (\text{Apply (7.89)}) \\ \frac{\gamma_j + 2(a - 1)}{m_j^2 + 2b} &= \alpha_j. \end{aligned}$$

Hence, we derive a new re-estimation equation for α_j . Note, once more, that in this context, if $a = 1$ and $b \rightarrow 0$, we again return to re-estimation equation (7.87).

Exercise 7.14

Returning once more to the context of [Exercise 7.9](#), let us consider a new data point $\{\mathbf{x}_{N+1}, t_{N+1}\}$, not utilized as part of the training procedure for the regression relevance vector machine. We aim to determine the predictive distribution of t_{N+1} , once more utilizing the linear-Gaussian model result [\(2.115\)](#), considering the posterior for \mathbf{w} determined in [Exercise 7.9](#), as well as the parameter values α and β which maximize the marginal likelihood, as obtained in [Exercise 7.12](#). It follows that the predictive distribution of t_{N+1} is a one-dimensional normal with mean

$$\mathbb{E}[t_{N+1}|\mathbf{w}, \mathbf{t}] = \beta^* \mathbf{t}^\top \Phi \{ \mathbf{A}^* + \beta^* \Phi^\top \Phi \}^{-1} \phi(\mathbf{x}_{N+1}),$$

and input dependent variance

$$\text{Var}[t_{N+1}|\mathbf{w}, \mathbf{t}] = (\beta^*)^{-1} + \phi^\top(\mathbf{x}_{N+1}) \{ \mathbf{A}^* + \beta^* \Phi^\top \Phi \}^{-1} \phi(\mathbf{x}_{N+1}).$$

Exercise 7.15

Consider the same framework as that of [Exercise 7.9](#) once more. We aim to demonstrate that we may rewrite [\(7.85\)](#), the marginal likelihood of \mathbf{t} , as [\(7.96\)](#). See that

$$\begin{aligned}
 \log p(\mathbf{t}|\mathbf{C}) &= -\frac{1}{2}\{N \log(2\pi) + \log|\mathbf{C}| + \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t}\} \\
 &= -\frac{1}{2}\left\{N \log(2\pi) + \log[|\mathbf{C}_{-i}^{-1}|(1 + \alpha_i^{-1} \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i)] + \right. \\
 &\quad \left. + \mathbf{t}^\top \left[\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1} \varphi_i \varphi_i^\top \mathbf{C}_{-i}^{-1}}{\alpha_i + \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i} \right] \mathbf{t}\right\} \quad (\text{Apply (7.94) and (7.95)}) \\
 &= -\frac{1}{2}\left\{N \log(2\pi) + \log|\mathbf{C}_{-i}^{-1}| + \log(\alpha_i + \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i) + \right. \\
 &\quad \left. - \log \alpha_i + \mathbf{t}^\top \mathbf{C}_{-i}^{-1} \mathbf{t} - \frac{\mathbf{t}^\top \mathbf{C}_{-i}^{-1} \varphi_i \varphi_i^\top \mathbf{C}_{-i}^{-1} \mathbf{t}}{\alpha_i + \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i}\right\} \quad (\text{Apply (7.98) and (7.99)}) \\
 &= -\frac{1}{2}\left\{N \log(2\pi) + \log|\mathbf{C}_{-i}^{-1}| + \log(\alpha_i + s_i) + \right. \\
 &\quad \left. - \log \alpha_i + \mathbf{t}^\top \mathbf{C}_{-i}^{-1} \mathbf{t} - \frac{q_i^2}{\alpha_i + s_i}\right\} \\
 &= \frac{1}{2}\left\{N \log(2\pi) + \log|\mathbf{C}_{-i}^{-1}| + \mathbf{t}^\top \mathbf{C}_{-i}^{-1} \mathbf{t}\right\} + \\
 &\quad + \frac{1}{2}\left[\log \alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right] \\
 \log p(\mathbf{t}|\mathbf{C}) &= \log p(\mathbf{t}|\mathbf{C}_{-i}) + \lambda(\alpha_i) \quad (\text{Apply (7.97)}).
 \end{aligned}$$

Hence, we prove the validity of [\(7.96\)](#).

Exercise 7.16

We aim to demonstrate that (7.101) is a local minimum of (7.85), with respect to α_j . For that purpose, we differentiate (7.85) twice with respect to α_j , as follows

$$\begin{aligned}
 \frac{\partial L(\alpha)}{\partial \alpha_j} &= \frac{\partial}{\partial \alpha_j} \left[L(\alpha_{-j}) + \lambda(\alpha_j) \right] && \text{(Apply (7.96))} \\
 &= \frac{\partial \lambda(\alpha_j)}{\partial \alpha_j} \\
 &= \frac{1}{2} \frac{\partial}{\partial \alpha_j} \left[\log \alpha_j - \log(\alpha_j + s_j) + \frac{q_j^2}{\alpha_j + s_j} \right] && \text{(Apply (7.97))} \\
 &= \frac{1}{2} \left[\frac{1}{\alpha_j} - \frac{1}{\alpha_j + s_j} - \frac{q_j^2}{(\alpha_j + s_j)^2} \right] \\
 &= \frac{1}{2} \left[\frac{(\alpha_j + s_j)^2 - \alpha_j(\alpha_j + s_j) - \alpha_j q_j^2}{\alpha_j(\alpha_j + s_j)^2} \right] \\
 &= \frac{1}{2} \left[\frac{\alpha_j s_j + s_j^2 - \alpha_j q_j^2}{\alpha_j(\alpha_j + s_j)^2} \right] \\
 \frac{\partial L(\alpha)}{\partial \alpha_j} &= \frac{1}{2(\alpha_j + s_j)^2} \left[s_j - q_j^2 + \frac{s_j^2}{\alpha_j} \right] \\
 \frac{\partial^2 L(\alpha)}{\partial \alpha_j^2} &= -\frac{1}{(\alpha_j + s_j)^3} \left[s_j - q_j^2 + \frac{s_j^2}{\alpha_j} \right] - \frac{1}{2(\alpha_j + s_j)^2} \frac{s_j^2}{\alpha_j^2} \\
 (7.34) \quad \frac{\partial^2 L(\alpha)}{\partial \alpha_j^2} &= -\frac{1}{(\alpha_j + s_j)} \frac{\partial L(\alpha)}{\partial \alpha_j} - 2 \left[\frac{s_j^2}{(\alpha_j + s_j)\alpha_j} \right]^2.
 \end{aligned}$$

Assuming $q_j^2 > s_j$, by evaluating the right-hand-side of (7.34) at (7.101), we find that the first term is zero, since $\partial L(s_j^2/(q_j^2 - s_j))/\partial \alpha_j = 0$, as seen previously. Assuming s_j and α_j are real values, the second term is negative (as it is the negative of a squared value). We thereby conclude that

$$\frac{\partial^2 L(s_j^2/(q_j^2 - s_j))}{\partial \alpha_j^2} < 0.$$

Therefore, (7.101) is a stationary point of (7.85).

Exercise 7.17

Consider again the quantities Q_i and S_i , utilized to compute respectively the quality and sparsity of a relevance vector machine for regression. It follows that we may rewrite Q_i as

$$\begin{aligned} Q_i &= \varphi_i^\top \mathbf{C}^{-1} \mathbf{t} && \text{(Apply (7.102))} \\ &= \varphi_i^\top \{\beta \mathbf{I} - \beta^2 \Phi \Sigma \Phi^\top\} \mathbf{t} && \text{(Apply (7.24))} \\ Q_i &= \beta \varphi_i^\top \mathbf{t} - \beta^2 \varphi_i^\top \Phi \Sigma \Phi^\top \mathbf{t}. \end{aligned}$$

Hence, we obtain (7.106). We may also rewrite S_i as

$$\begin{aligned} S_i &= \varphi_i^\top \mathbf{C}^{-1} \varphi_i && \text{(Apply (7.103))} \\ &= \varphi_i^\top \{\beta \mathbf{I} - \beta^2 \Phi \Sigma \Phi^\top\} \varphi_i && \text{(Apply (7.24))} \\ Q_i &= \beta \varphi_i^\top \varphi_i - \beta^2 \varphi_i^\top \Phi \Sigma \Phi^\top \varphi_i. \end{aligned}$$

Hence, we obtain (7.107).

Exercise 7.18

Consider the context of relevance vector machines for classification. We aim to compute the gradient and Hessian of the logarithm of the posterior distribution of the weights \mathbf{w} , as seen in (7.109), where we have that $y(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\top \phi(\mathbf{x}))$. It follows that the gradient is computed as

$$\begin{aligned}
 \nabla \log p(\mathbf{w}|\mathbf{t}, \alpha) &= \nabla \left[\sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} + \right. \\
 &\quad \left. - \frac{1}{2} \mathbf{w}^\top \mathbf{A} \mathbf{w} + \text{const} \right] \\
 &= \nabla \left[\sum_{n=1}^N \{t_n \log \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n)) + \right. \\
 &\quad \left. + (1 - t_n) \log(1 - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n)))\} \right] - \mathbf{A} \mathbf{w} \\
 &= \sum_{n=1}^N \left\{ t_n \frac{\sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))[1 - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))]}{\sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))} \phi(\mathbf{x}_n) + \right. \\
 &\quad \left. - (1 - t_n) \frac{\sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))[1 - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))]}{1 - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))} \phi(\mathbf{x}_n) \right\} - \mathbf{A} \mathbf{w} \quad (\text{Apply (C.19)}) \\
 &= \sum_{n=1}^N \{t_n[1 - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))] \phi(\mathbf{x}_n) + \\
 &\quad - (1 - t_n)\sigma(\mathbf{w}^\top \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)\} - \mathbf{A} \mathbf{w} \quad (\text{Apply (4.88) and (C.19)}) \\
 &= \sum_{n=1}^N \{t_n - y_n\} \phi(\mathbf{x}_n) - \mathbf{A} \mathbf{w} \\
 (7.35) \quad \nabla \log p(\mathbf{w}|\mathbf{t}, \alpha) &= \Phi^\top (\mathbf{t} - \mathbf{y}) - \mathbf{A} \mathbf{w}.
 \end{aligned}$$

We have therefore derived the gradient of (7.109). We now compute the Hessian as

$$\begin{aligned}
 \nabla \nabla^\top \log p(\mathbf{w}|\mathbf{t}, \alpha) &= \nabla^\top \left[\sum_{n=1}^N \{t_n - y_n\} \phi(\mathbf{x}_n) - \mathbf{A} \mathbf{w} \right] \quad (\text{Apply (7.35)}) \\
 &= \nabla^\top \left[\sum_{n=1}^N \{t_n - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))\} \phi(\mathbf{x}_n) \right] - \mathbf{A} \quad (\text{Apply (C.19)}) \\
 &= - \sum_{n=1}^N \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))[1 - \sigma(\mathbf{w}^\top \phi(\mathbf{x}_n))] \times \\
 &\quad \times \phi(\mathbf{x}_n) \phi^\top(\mathbf{x}_n) - \mathbf{A} \quad (\text{Apply (4.88) and (C.19)}) \\
 &= - \sum_{n=1}^N y_n(1 - y_n) \phi(\mathbf{x}_n) \phi^\top(\mathbf{x}_n) - \mathbf{A} \\
 \nabla \nabla^\top \log p(\mathbf{w}|\mathbf{t}, \alpha) &= -\{\Phi^\top \mathbf{B} \Phi + \mathbf{A}\},
 \end{aligned}$$

where B is an $N \times N$ -dimensional diagonal matrix such that $B_{j,j} = y_j\{1 - y_j\}$. We therefore obtain the desired Hessian.

Exercise 7.19

Consider the context of relevance vector machines utilized for the purpose of classification, and that we have observed a data set $\{\mathbf{x}_n, t_n\}_{n=1}^N$ of input and target variables such that $t_n \in \{0, 1\}$. By constructing a Laplace approximation to the model evidence, as in (7.114), so that we have the following logarithm of the marginal likelihood

$$\begin{aligned}
 \log p(\mathbf{t}|\boldsymbol{\alpha}) &\approx p(\mathbf{t}|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha}) + \frac{M}{2} \log(2\pi) + \frac{1}{2} \log|\boldsymbol{\Sigma}| \\
 &\propto \sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} + \\
 &\quad - \frac{1}{2} (\mathbf{w}^*)^\top \mathbf{A} \mathbf{w}^* + \frac{1}{2} \log|\mathbf{A}| + \frac{M}{2} \log(2\pi) + \\
 &\quad - \frac{1}{2} \log|\boldsymbol{\Sigma}^{-1}| \\
 (7.36) \quad \log p(\mathbf{t}|\boldsymbol{\alpha}) &\propto \sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} + \\
 &\quad - \frac{1}{2} (\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \boldsymbol{\Phi}^\top (\mathbf{t} - \mathbf{y}) + \\
 &\quad + \frac{M}{2} \log(2\pi) - \frac{1}{2} \log|\boldsymbol{\Phi}^\top \mathbf{B} \boldsymbol{\Phi} + \mathbf{A}| + \\
 &\quad + \frac{1}{2} \sum_{j=1}^M \log \alpha_j \tag{Apply (7.112) and (7.113)}.
 \end{aligned}$$

Prior to differentiating (7.36) with respect to α_j , first consider that the derivative of $(\mathbf{w}^*)^\top \phi(\mathbf{x})$ with respect to α_j is

$$\begin{aligned}
 \frac{\partial[(\mathbf{w}^*)^\top \phi(\mathbf{x})]}{\partial \alpha_j} &= \frac{\partial[(\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \phi(\mathbf{x})]}{\partial \alpha_j} \\
 &= (\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \frac{\partial \mathbf{A}^{-1}}{\partial \alpha_j} \phi(\mathbf{x}) \\
 (7.37) \quad \frac{\partial[(\mathbf{w}^*)^\top \phi(\mathbf{x})]}{\partial \alpha_j} &= -(\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \mathbf{I}^{(j)} \mathbf{A}^{-1} \phi(\mathbf{x}) \tag{Apply (C.21)},
 \end{aligned}$$

where $\mathbf{I}^{(j)}$ is an $M \times M$ -dimensional matrix composed of zeroes, except at the (j, j) -th coordinate, wherein $\mathbf{I}_{j,j}^{(j)} = 1$. We compute the derivative of $\log p(\mathbf{t}|\mathbf{w}^*)$ with respect to α_j as

$$\begin{aligned}
 \frac{\partial \log p(\mathbf{t}|\mathbf{w}^*)}{\partial \alpha_j} &= \frac{\partial}{\partial \alpha_j} \left[\sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} \right] \\
 &= \sum_{n=1}^N \left\{ t_n \frac{y_n(1 - y_n)}{y_n} - (1 - t_n) \frac{y_n(1 - y_n)}{1 - y_n} \right\} \frac{\partial[(\mathbf{w}^*)^\top \phi(\mathbf{x}_n)]}{\partial \alpha_j} \tag{Apply (4.88)} \\
 &= - \sum_{n=1}^N \{t_n - y_n\} (\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \mathbf{I}^{(j)} \mathbf{A}^{-1} \phi(\mathbf{x}_n) \tag{Apply (7.37)} \\
 &= -(\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \mathbf{I}^{(j)} \mathbf{A}^{-1} \sum_{n=1}^N \{t_n - y_n\} \phi(\mathbf{x}_n) \\
 &= -(\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \mathbf{I}^{(j)} \mathbf{A}^{-1} \boldsymbol{\Phi}^\top (\mathbf{t} - \mathbf{y}) \\
 &= -\mathbf{w}^* \mathbf{I}^{(j)} \mathbf{w}^* \tag{Apply (7.112)} \\
 (7.38) \quad \frac{\partial \log p(\mathbf{t}|\mathbf{w}^*)}{\partial \alpha_j} &= -(w_j^*)^2.
 \end{aligned}$$

Finally, differentiating (7.36) with respect to α_j and solving for 0, we obtain

$$\begin{aligned}
 & \frac{\partial \log p(\mathbf{t}|\boldsymbol{\alpha})}{\partial \alpha_j} = 0 \\
 & \frac{\partial \log p(\mathbf{t}|\mathbf{w}^*)}{\partial \alpha_j} + \\
 & + \frac{1}{2} (\mathbf{t} - \mathbf{y})^\top \boldsymbol{\Phi} \mathbf{A}^{-1} \mathbf{I}^{(j)} \mathbf{A}^{-1} \boldsymbol{\Phi}^\top (\mathbf{t} - \mathbf{y}) + \\
 & + \frac{1}{2\alpha_j} - \frac{1}{2} \text{tr}(\{\boldsymbol{\Phi}^\top \mathbf{B} \boldsymbol{\Phi} + \mathbf{A}\}^{-1} \mathbf{I}^{(j)}) = 0 \quad (\text{Apply (C.21) and (C.22)}) \\
 & -(w_j^*)^2 + \frac{1}{2} (\mathbf{w}^*)^\top \mathbf{I}^{(j)} \mathbf{w}^* + \frac{1}{2\alpha_j} - \frac{1}{2} \text{tr}(\boldsymbol{\Sigma} \mathbf{I}^{(j)}) = 0 \quad (\text{Apply (7.112), (7.113) and (7.38)}) \\
 & -\frac{1}{2} (w_j^*)^2 + \frac{1}{2\alpha_j} - \frac{1}{2} \Sigma_{j,j} = 0 \\
 & \frac{1 - \alpha_j \Sigma_{j,j}}{(w_j^*)^2} = \alpha_j \\
 & \frac{\gamma_j}{(w_j^*)^2} = \alpha_j \quad (\text{Apply (7.89)}).
 \end{aligned}$$

Hence, we have derived the update equation for α_j .

Chapter 8

Graphical Models

Exercise 8.1

Consider a joint distribution of variables $\mathbf{x} = \{x_1, \dots, x_K\}$ characterized by (8.5), and assume that for all $k \in \{1, \dots, K\}$, $\int p(x_k | \text{pa}_k) dx_k = 1$. Moreover, we assume that the directed graph which assigns this joint distribution is a directed acyclic graph, such that we may assume all variables receive links only from variables whose index is below theirs, and that x_1 acts as the root of the graph, having no parent variables. It follows that

$$\begin{aligned}
 \int p(\mathbf{x}) d\mathbf{x} &= \int \left[\prod_{k=1}^K p(x_k | \text{pa}_k) \right] d\mathbf{x} && \text{(Apply (8.5))} \\
 &= \int \cdots \int \left[\prod_{k=1}^K p(x_k | \text{pa}_k) \right] dx_1 \dots dx_{K-1} dx_K \\
 &= \int p(x_K | \text{pa}_K) \times \\
 &\quad \times \int \cdots \int \left[\prod_{k=1}^{K-1} p(x_k | \text{pa}_k) \right] dx_1 \dots dx_{K-1} dx_K \\
 (8.1) \quad \int p(\mathbf{x}) d\mathbf{x} &= \int \cdots \int \left[\prod_{k=1}^{K-1} p(x_k | \text{pa}_k) \right] dx_1 \dots dx_{K-1} && \text{(Apply } \int p(x_K | \text{pa}_K) dx_K = 1).
 \end{aligned}$$

We repeat (8.1) for $x_{K-1}, x_{K-2}, \dots, x_2$, whereupon we obtain

$$\begin{aligned}
 \int p(\mathbf{x}) d\mathbf{x} &= \int p(x_1 | \text{pa}_1) dx_1 \\
 &= \int p(x_1) dx_1 && \text{(Apply } \text{pa}_1 = \emptyset) \\
 \int p(\mathbf{x}) d\mathbf{x} &= 1 && \text{(Apply (1.30)).}
 \end{aligned}$$

Hence, we conclude that (8.5) is properly normalized.

Exercise 8.2

Assume we have a directed graph such that all its comprising variables receive links only from variables whose index is below theirs. Assume additionally that there exists a directed cycle within this graph, for example of the form $\{x_{k_1}, x_{k_2}, \dots, x_{k_M}, x_{k_1}\}$ (that is, we have that x_{k_1} sends links to x_{k_2} , which sends links to x_{k_3} , and so forth, until we arrive at variable x_{k_M} , which sends links to x_{k_1}). From the assumption that all links must be sent from lower indexed variables to higher indexed variables, we have that $k_1 < k_2 < \dots < k_M < k_1$, which is a clear contradiction. We thereby conclude that, if the graph satisfies the initial indexing condition for the links, it cannot possess any directed cycle, and therefore it must be a directed acyclic graph.

Exercise 8.3

Consider the joint distribution of variables a , b and c seen in Table 8.1. We aim to demonstrate that $p(a)p(b) \neq p(a, b)$, i.e., a and b are marginally dependent. We compute $p(a = 0)$ as follows:

$$\begin{aligned} p(a = 0) &= p(a = 0, b = 0, c = 0) + p(a = 0, b = 0, c = 1) + \\ &\quad + p(a = 0, b = 1, c = 0) + p(a = 0, b = 1, c = 1) \quad (\text{Apply (1.10)}) \\ &= 0.192 + 0.144 + 0.048 + 0.216 \\ (8.2) \quad p(a = 0) &= 0.6. \end{aligned}$$

We now compute $p(b = 0)$ as

$$\begin{aligned} p(b = 0) &= p(a = 0, b = 0, c = 0) + p(a = 0, b = 0, c = 1) + \\ &\quad + p(a = 1, b = 0, c = 0) + p(a = 1, b = 0, c = 1) \quad (\text{Apply (1.10)}) \\ &= 0.192 + 0.144 + 0.192 + 0.064 \\ (8.3) \quad p(b = 0) &= 0.592. \end{aligned}$$

Lastly, we compute $p(a = 0, b = 0)$ as

$$\begin{aligned} p(a = 0, b = 0) &= p(a = 0, b = 0, c = 0) + p(a = 0, b = 0, c = 1) \quad (\text{Apply (1.10)}) \\ &= 0.192 + 0.144 \\ (8.4) \quad p(a = 0, b = 0) &= 0.336. \end{aligned}$$

From (8.2) and (8.3), we have $p(a = 0)p(b = 0) = 0.3552 \neq 0.336 = p(a = 0, b = 0)$, hence we conclude that a and b are not marginally independent. Before computing the conditional distribution of a and c , we first compute $p(c = 0)$ as

$$\begin{aligned} p(c = 0) &= p(a = 0, b = 0, c = 0) + p(a = 0, b = 1, c = 0) + \\ &\quad + p(a = 1, b = 0, c = 0) + p(a = 1, b = 1, c = 0) \quad (\text{Apply (1.10)}) \\ &= 0.192 + 0.048 + 0.192 + 0.048 \\ (8.5) \quad p(c = 0) &= 0.48. \end{aligned}$$

We now compute $p(a = 0|c = 0)$ as

$$\begin{aligned} p(a = 0|c = 0) &= \frac{p(a = 0, c = 0)}{p(c = 0)} \quad (\text{Apply (1.12)}) \\ &= \frac{p(a = 0, b = 0, c = 0) + p(a = 0, b = 1, c = 0)}{p(c = 0)} \quad (\text{Apply (1.10)}) \\ &= \frac{p(a = 0, b = 0, c = 0) + p(a = 0, b = 1, c = 0)}{0.48} \quad (\text{Apply (8.5)}) \\ &= \frac{0.192 + 0.048}{0.48} \\ (8.6) \quad p(a = 0|c = 0) &= \frac{1}{2}. \end{aligned}$$

We compute $p(a = 0|c = 1)$ as

$$\begin{aligned} p(a = 0|c = 1) &= \frac{p(a = 0, c = 1)}{p(c = 1)} \quad (\text{Apply (1.12)}) \\ &= \frac{p(a = 0, b = 0, c = 1) + p(a = 0, b = 1, c = 1)}{1 - p(c = 0)} \quad (\text{Apply (1.10)}) \\ &= \frac{0.144 + 0.216}{0.52} \quad (\text{Apply (8.5)}) \\ (8.7) \quad p(a = 0|c = 1) &= \frac{9}{13}. \end{aligned}$$

We compute $p(b = 0|c = 0)$ as

$$\begin{aligned}
 p(b = 0|c = 0) &= \frac{p(a = 0, c = 0)}{p(c = 0)} && \text{(Apply (1.12))} \\
 p(b = 0|c = 0) &= \frac{p(a = 0, b = 0, c = 0) + p(a = 1, b = 0, c = 0)}{p(c = 0)} && \text{(Apply (1.10))} \\
 &= \frac{0.192 + 0.192}{0.48} && \text{(Apply (8.5))} \\
 (8.8) \quad p(b = 0|c = 0) &= \frac{4}{5}.
 \end{aligned}$$

We compute $p(b = 0|c = 1)$ as

$$\begin{aligned}
 p(b = 0|c = 1) &= \frac{p(b = 0, c = 1)}{p(c = 1)} && \text{(Apply (1.12))} \\
 p(b = 0|c = 1) &= \frac{p(a = 0, b = 0, c = 1) + p(a = 1, b = 0, c = 1)}{1 - p(c = 0)} && \text{(Apply (1.10))} \\
 &= \frac{0.144 + 0.064}{0.52} && \text{(Apply (8.5))} \\
 (8.9) \quad p(b = 0|c = 1) &= \frac{2}{5}.
 \end{aligned}$$

We compute $p(a = 0, b = 0|c = 0)$ as

$$\begin{aligned}
 p(a = 0, b = 0|c = 0) &= \frac{p(a = 0, b = 0, c = 0)}{p(c = 0)} && \text{(Apply (1.12))} \\
 &= \frac{0.192}{0.48} && \text{(Apply (8.5))} \\
 (8.10) \quad p(a = 0, b = 0|c = 0) &= \frac{2}{5}.
 \end{aligned}$$

We compute $p(a = 0, b = 1|c = 0)$ as

$$\begin{aligned}
 p(a = 0, b = 1|c = 0) &= \frac{p(a = 0, b = 1, c = 0)}{p(c = 0)} && \text{(Apply (1.12))} \\
 &= \frac{0.048}{0.48} && \text{(Apply (8.5))} \\
 (8.11) \quad p(a = 0, b = 1|c = 0) &= \frac{1}{10}.
 \end{aligned}$$

We compute $p(a = 1, b = 0|c = 0)$ as

$$\begin{aligned}
 p(a = 1, b = 0|c = 0) &= \frac{p(a = 1, b = 0, c = 0)}{p(c = 0)} && \text{(Apply (1.12))} \\
 &= \frac{0.192}{0.48} && \text{(Apply (8.5))} \\
 (8.12) \quad p(a = 1, b = 0|c = 0) &= \frac{2}{5}.
 \end{aligned}$$

We compute $p(a = 1, b = 1|c = 0)$ as

$$\begin{aligned}
 p(a = 1, b = 1|c = 0) &= \frac{p(a = 1, b = 1, c = 0)}{p(c = 0)} && \text{(Apply (1.12))} \\
 &= \frac{0.048}{0.48} && \text{(Apply (8.5))} \\
 (8.13) \quad p(a = 1, b = 1|c = 0) &= \frac{1}{10}.
 \end{aligned}$$

We compute $p(a = 0, b = 0|c = 1)$ as

$$\begin{aligned}
 p(a = 0, b = 0|c = 1) &= \frac{p(a = 0, b = 0, c = 1)}{p(c = 1)} && \text{(Apply (1.12))} \\
 &= \frac{0.144}{1 - p(c = 0)} \\
 &= \frac{0.144}{0.52} && \text{(Apply (8.5))} \\
 (8.14) \quad p(a = 0, b = 0|c = 1) &= \frac{18}{65}.
 \end{aligned}$$

We compute $p(a = 0, b = 1|c = 1)$ as

$$\begin{aligned}
 p(a = 0, b = 1|c = 1) &= \frac{p(a = 0, b = 1, c = 1)}{p(c = 1)} && \text{(Apply (1.12))} \\
 &= \frac{0.216}{1 - p(c = 0)} \\
 &= \frac{0.216}{0.52} && \text{(Apply (8.5))} \\
 (8.15) \quad p(a = 0, b = 1|c = 1) &= \frac{27}{65}.
 \end{aligned}$$

We compute $p(a = 1, b = 0|c = 1)$ as

$$\begin{aligned}
 p(a = 1, b = 0|c = 1) &= \frac{p(a = 1, b = 0, c = 1)}{p(c = 1)} && \text{(Apply (1.12))} \\
 &= \frac{0.064}{1 - p(c = 0)} \\
 &= \frac{0.064}{0.52} && \text{(Apply (8.5))} \\
 (8.16) \quad p(a = 1, b = 0|c = 1) &= \frac{8}{65}.
 \end{aligned}$$

We compute $p(a = 1, b = 1|c = 1)$ as

$$\begin{aligned}
 p(a = 1, b = 1|c = 1) &= \frac{p(a = 1, b = 1, c = 1)}{p(c = 1)} && \text{(Apply (1.12))} \\
 &= \frac{0.096}{1 - p(c = 0)} \\
 &= \frac{0.096}{0.52} && \text{(Apply (8.5))} \\
 (8.17) \quad p(a = 1, b = 1|c = 1) &= \frac{12}{65}.
 \end{aligned}$$

Reuniting the results (8.6), (8.7), (8.8), (8.9), (8.10), (8.11), (8.12), (8.13), (8.14), (8.15), (8.16), (8.17), we write Table 8.2, which includes the values of $p(a|c)p(b|c)$ and $p(a, b|c)$. As such, it is easy to see that $p(a, b|c) = p(a|c)p(b|c)$, hence a and b are, conditional on any value of c , independent, whilst, as previously shown, being marginally dependent.

Table 8.1: Joint distribution of a , b and c .

a	b	c	$p(a, b, c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Table 8.2: Distribution of a and b , marginal and joint, conditional on c .

a	b	c	$p(a c)$	$p(b c)$	$p(a c)p(b c)$	$p(a, b c)$
0	0	0	1/2	4/5	2/5	2/5
0	0	1	9/13	2/5	18/65	18/65
0	1	0	1/2	1/5	1/10	1/10
0	1	1	9/13	3/5	27/65	27/65
1	0	0	1/2	4/5	2/5	2/5
1	0	1	4/13	2/5	8/65	8/65
1	1	0	1/2	1/5	1/10	1/10
1	1	1	4/13	3/5	12/65	12/65

Exercise 8.4

Consider the same context as in [Exercise 8.3](#). We aim to directly demonstrate that $p(a, b, c) = p(a)p(c|a)p(b|c)$. As we have already computed $p(a)$ in [\(8.2\)](#), and $p(b|c)$ in [\(8.8\)](#) and [\(8.9\)](#), hence, we need only compute $p(c|a)$. It follows that we compute $p(c = 0|a = 0)$ as

$$\begin{aligned}
 p(c = 0|a = 0) &= \frac{p(a = 0, c = 0)}{p(a = 0)} && \text{(Apply (1.12))} \\
 &= \frac{p(a = 0, b = 0, c = 0) + p(a = 0, b = 1, c = 0)}{p(a = 0)} && \text{(Apply (1.10))} \\
 &= \frac{0.192 + 0.048}{0.6} && \text{(Apply (8.2))} \\
 (8.18) \quad p(c = 0|a = 0) &= \frac{2}{5}.
 \end{aligned}$$

We compute $p(c = 0|a = 1)$ as

$$\begin{aligned}
 p(c = 0|a = 1) &= \frac{p(a = 1, c = 0)}{p(a = 1)} && \text{(Apply (1.12))} \\
 &= \frac{p(a = 1, b = 0, c = 0) + p(a = 1, b = 1, c = 0)}{1 - p(a = 0)} && \text{(Apply (1.10))} \\
 &= \frac{0.192 + 0.048}{0.4} && \text{(Apply (8.2))} \\
 (8.19) \quad p(c = 0|a = 1) &= \frac{3}{5}.
 \end{aligned}$$

Utilizing the results [\(8.2\)](#), [\(8.8\)](#), [\(8.9\)](#), [\(8.18\)](#) and [\(8.19\)](#), we write Table [8.3](#), which showcases the values of $p(a, b, c)$ and $p(a)p(c|a)p(b|c)$. We thereby conclude that the joint distribution of a , b and c may be factored accordingly. The directed graph corresponding to this factorization is depicted in [Figure 8.1](#).

Table 8.3: Joint distribution of a , b and c , and factored conditional distributions.

a	b	c	$p(a)$	$p(c a)$	$p(b c)$	$p(a)p(c a)p(b c)$	$p(a, b, c)$
0	0	0	$3/5$	$2/5$	$4/5$	$24/125$	0.192
0	0	1	$3/5$	$3/5$	$2/5$	$18/125$	0.144
0	1	0	$3/5$	$2/5$	$1/5$	$6/125$	0.048
0	1	1	$3/5$	$3/5$	$3/5$	$27/125$	0.216
1	0	0	$2/5$	$3/5$	$4/5$	$24/125$	0.192
1	0	1	$2/5$	$2/5$	$2/5$	$8/125$	0.064
1	1	0	$2/5$	$3/5$	$1/5$	$6/125$	0.048
1	1	1	$2/5$	$2/5$	$3/5$	$12/125$	0.096

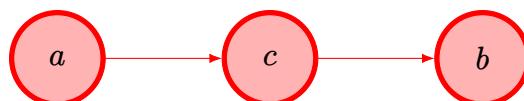


Figure 8.1: Directed graph corresponding to the factorization $p(a, b, c)$ displayed by the joint distribution in [Table 8.1](#).

Exercise 8.5

We return to the context of relevance vector machines for regression. Consider that we write the joint distribution of our target variables as in (7.79), and assign to our weight parameters the prior (7.80), we find the probabilistic directed graphical model which corresponds to this model is as in Figure 8.2

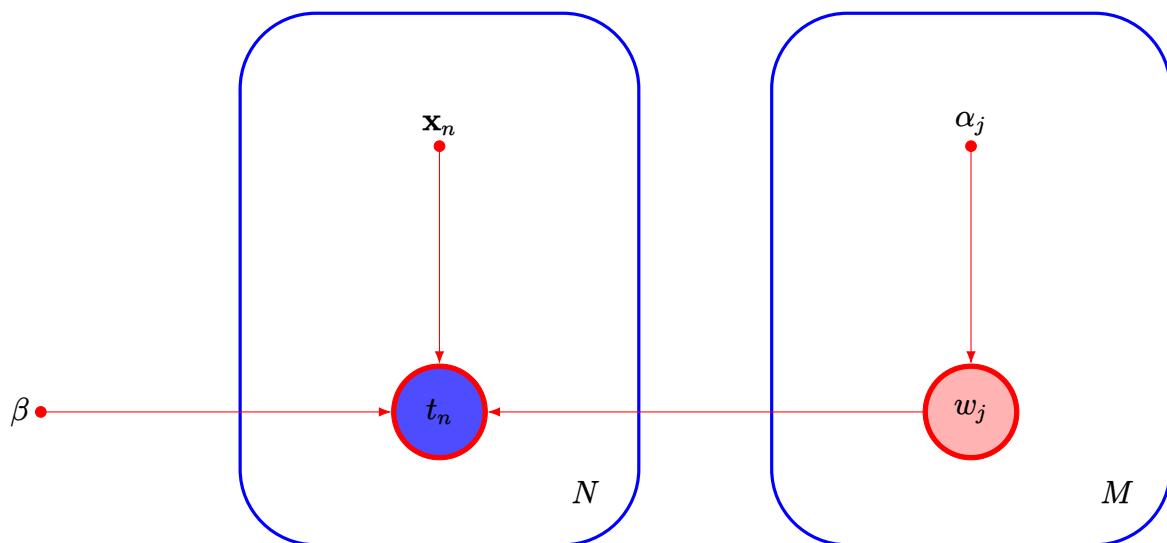


Figure 8.2: Probabilistic directed graphical model corresponding to the relevance vector machine for regression.

Exercise 8.6

Consider a set of random variables x_1, \dots, x_M such that $x_i \in \{0, 1\}$ and $p(x_i = 1) = \mu_i \in [0, 1], \forall i \in \{1, \dots, M\}$. Define a random variable y such that

$$(8.20) \quad \begin{aligned} p(y = 0|x_1, \dots, x_M) &= \prod_{i=1}^M \{p(x_i = 0)\}^{x_i} \\ &= \prod_{i=1}^M \{1 - p(x_i = 1)\}^{x_i} \\ p(y = 0|x_1, \dots, x_M) &= \prod_{i=1}^M \{1 - \mu_i\}^{x_i}. \end{aligned}$$

It follows that $p(y = 1|x_1, \dots, x_M)$ is

$$(8.21) \quad \begin{aligned} p(y = 1|x_1, \dots, x_M) &= 1 - p(y = 0|x_1, \dots, x_M) \\ p(y = 1|x_1, \dots, x_M) &= 1 - \prod_{i=1}^M \{1 - \mu_i\}^{x_i} \quad (\text{Apply (8.20)}). \end{aligned}$$

Note that, if we consider that any of the random variables x_j are deterministically 1 (for example, $p(x_j = 1) = \mu_j = 1$, i.e., $x_j = 1$ almost certainly), we have that $p(y = 1|x_1, \dots, x_M) = 1$, i.e., $y = 1$ almost certainly, therein corresponding to the logical OR function. Conversely, (8.21) demonstrates that the following holds

$$(8.22) \quad \begin{aligned} p(y = 1|x_1, \dots, x_M) &= 1 - \prod_{i=1}^M \{1 - \mu_i\}^{x_i} \\ &= 1 - \{1 - \mu_k\} \prod_{\substack{i=1 \\ i \neq k}}^M \{1 - \mu_i\}^{x_i} \\ &\geq 1 - 1 + \mu_k \quad (\text{Apply } \prod_{\substack{i=1 \\ i \neq k}}^M \{1 - \mu_i\}^{x_i} \leq 1) \end{aligned}$$

$$p(y = 1|x_1, \dots, x_M) \geq p(x_k = 1),$$

where

$$k = \arg \max_{i \in \{1, \dots, M\}} \{p(x_i = 1)x_i\}.$$

That is, (8.22) implies that the probability of $y = 1$, conditional on x_1, \dots, x_M , is greater than or equal to the maximum probability of $p(x_j = 1)$ amongst the variables such that $x_j = 1$. We conclude therefore that (8.21) works as a probabilistic approximation to the logical OR function. The formulation in (8.104) adds an additional variable $\mu_0 \in [0, 1]$ which incorporates prior beliefs to the probability $p(y = 1|x_1, \dots, x_M)$. Particularly, following the same exercise as in (8.22), it establishes a minimum value for $p(y = 1|x_1, \dots, x_M)$, as $p(y = 1|x_1, \dots, x_M) \geq \mu_0$.

Exercise 8.7

Consider the directed graph portrayed in Figure 8.3, such that the joint distribution of x_1, x_2, x_3 is described by that graph, in unison with the formulation in (8.11). We may determine the form of the joint mean $\mathbb{E}[\mathbf{x}]$ by inspecting its components $\mathbb{E}[x_1]$, $\mathbb{E}[x_2]$ and $\mathbb{E}[x_3]$ and utilizing (8.15) recursively. First, $\mathbb{E}[x_1]$ trivially is

$$(8.23) \quad \mathbb{E}[x_1] = b_1.$$

It follows that $\mathbb{E}[x_2]$ is

$$(8.24) \quad \begin{aligned} \mathbb{E}[x_2] &= w_{2,1}\mathbb{E}[x_1] + b_2 && \text{(Apply (8.15))} \\ \mathbb{E}[x_2] &= w_{2,1}b_1 + b_2 && \text{(Apply (8.23)).} \end{aligned}$$

It follows that $\mathbb{E}[x_3]$ is

$$\begin{aligned} \mathbb{E}[x_3] &= w_{3,2}\mathbb{E}[x_2] + b_3 && \text{(Apply (8.15))} \\ &= w_{3,2}\{w_{2,1}b_1 + b_2\} + b_3 && \text{(Apply (8.24))} \\ \mathbb{E}[x_3] &= w_{3,2}w_{2,1}b_1 + w_{3,2}b_2 + b_3. \end{aligned}$$

The computation of the covariance matrix associated with x_1, x_2 and x_3 is analogously obtained by application of (8.16). It follows that $\text{Cov}[x_1, x_1]$ is trivially obtained as

$$(8.25) \quad \text{Cov}[x_1, x_1] = v_1$$

$\text{Cov}[x_1, x_2]$ is computed as

$$(8.26) \quad \begin{aligned} \text{Cov}[x_1, x_2] &= w_{2,1}\text{Cov}[x_1, x_1] && \text{(Apply (8.15))} \\ \text{Cov}[x_1, x_2] &= w_{2,1}v_1 && \text{(Apply (8.25))} \end{aligned}$$

$\text{Cov}[x_1, x_3]$ is computed as

$$\begin{aligned} \text{Cov}[x_1, x_3] &= w_{3,2}\text{Cov}[x_1, x_2] && \text{(Apply (8.15))} \\ \text{Cov}[x_1, x_3] &= w_{3,2}w_{2,1}v_1 && \text{(Apply (8.26))} \end{aligned}$$

$\text{Cov}[x_2, x_2]$ is computed as

$$(8.27) \quad \begin{aligned} \text{Cov}[x_2, x_2] &= w_{2,1}\text{Cov}[x_2, x_1] + v_2 && \text{(Apply (8.15))} \\ \text{Cov}[x_2, x_2] &= w_{2,1}^2v_1 + v_2 && \text{(Apply (8.26))} \end{aligned}$$

$\text{Cov}[x_2, x_3]$ is computed as

$$(8.28) \quad \begin{aligned} \text{Cov}[x_2, x_3] &= w_{3,2}\text{Cov}[x_2, x_2] && \text{(Apply (8.15))} \\ &= w_{3,2}\{w_{2,1}^2v_1 + v_2\} && \text{(Apply (8.27))} \\ \text{Cov}[x_2, x_3] &= w_{3,2}w_{2,1}^2v_1 + w_{3,2}v_2 \end{aligned}$$

$\text{Cov}[x_3, x_3]$ is computed as

$$\begin{aligned} \text{Cov}[x_3, x_3] &= w_{3,2}\text{Cov}[x_3, x_2] + v_3 && \text{(Apply (8.15))} \\ &= w_{3,2}\{w_{3,2}w_{2,1}^2v_1 + w_{3,2}v_2\} + v_3 && \text{(Apply (8.28))} \\ \text{Cov}[x_3, x_3] &= w_{3,2}^2w_{2,1}^2v_1 + w_{3,2}^2v_2 + v_3. \end{aligned}$$

We can therefore conclude that

$$\mathbb{E}[\mathbf{x}] = \begin{pmatrix} b_1 \\ w_{2,1}b_1 + b_2 \\ w_{3,2}w_{2,1}b_1 + w_{3,2}b_2 + b_3 \end{pmatrix}$$

and

$$\text{Cov}[\mathbf{x}] = \begin{pmatrix} v_1 & w_{2,1}v_1 & w_{3,2}w_{2,1}v_1 \\ w_{2,1}v_1 & w_{2,1}^2v_1 + v_2 & w_{3,2}w_{2,1}^2v_1 + w_{3,2}v_2 \\ w_{3,2}w_{2,1}v_1 & w_{3,2}w_{2,1}^2v_1 + w_{3,2}v_2 & w_{3,2}^2w_{2,1}^2v_1 + w_{3,2}^2v_2 + v_3 \end{pmatrix}.$$

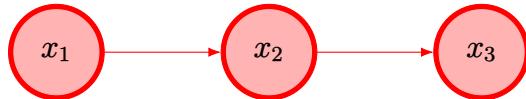


Figure 8.3: Directed graph corresponding to the model utilized in [Exercise 8.7](#).

Exercise 8.8

Consider a set of random variables a, b, c and d such that

$$(8.29) \quad p(a, b, c|d) = p(a|d)p(b, c|d).$$

It follows that, by integrating both sides of (8.29) with respect to c , we obtain

$$\int p(a, b, c|d) dc = p(a|d) \int p(b, c|d) dc$$
$$p(a, b|d) = p(a|d)p(b|d) \quad (\text{Apply (1.31)}).$$

Hence, we conclude that $a \perp\!\!\!\perp b|d$. Consequently, we have that $a \perp\!\!\!\perp b, c|d \Rightarrow a \perp\!\!\!\perp b|d$.

Exercise 8.9

We aim to demonstrate that, for any graph containing a random variable \mathbf{x} , it follows that the distribution of \mathbf{x} conditional on its Markov blanket is independent of all remaining random variables in this graph. We adopt a simplified graph, depicted in Figure 8.4, where the variables \mathbf{y} , \mathbf{w} , \mathbf{z} , \mathbf{s} and \mathbf{r} exhibit the five distinct forms through which another random variable may be connected to \mathbf{x} , whilst not belonging to its Markov blanket: as a parent of one, or more, parents of \mathbf{x} , exemplified by \mathbf{w} ; as a child of one, or more, parents of \mathbf{x} , exemplified by \mathbf{s} ; as a child of one, or more, children of \mathbf{x} , exemplified by \mathbf{z} ; as a parent of one, or more, co-parent of \mathbf{x} , as exemplified by \mathbf{y} ; and as a child of one, or more, co-parent of \mathbf{x} , as exemplified by \mathbf{r} (note that for all these definitions, the variable must also not be a child or parent of \mathbf{x}). Utilizing the d-separation criterion, we shall now demonstrate why each of these variables, conditioned by the Markov blanket (a set which we will denote as C), are independent:

1. For links of the form exemplified by \mathbf{w} , we find that the path connecting \mathbf{w} and \mathbf{x} must contain a head-to-tail link at a variable contained within the conditioning set C , hence, all paths of this form are blocked, and $\mathbf{x} \perp\!\!\!\perp \mathbf{w}|C$.
2. For links of the form exemplified by \mathbf{s} , we find that the path connecting \mathbf{s} and \mathbf{x} must contain a tail-to-tail link at a variable contained within the conditioning set C , hence, all paths of this form are blocked, and $\mathbf{x} \perp\!\!\!\perp \mathbf{s}|C$.
3. For links of the form exemplified by \mathbf{z} , we find that the path connecting \mathbf{z} and \mathbf{x} must contain a head-to-tail link at a variable contained within the conditioning set C , hence, all paths of this form are blocked, and $\mathbf{x} \perp\!\!\!\perp \mathbf{z}|C$.
4. For links of the form exemplified by \mathbf{y} , we find that the path connecting \mathbf{y} and \mathbf{x} must contain a head-to-tail link at a variable contained within the conditioning set C , hence, all paths of this form are blocked, and $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|C$.
5. For links of the form exemplified by \mathbf{r} , we find that the path connecting \mathbf{y} and \mathbf{x} must contain a tail-to-tail link at a variable contained within the conditioning set C , hence, all paths of this form are blocked, and $\mathbf{x} \perp\!\!\!\perp \mathbf{r}|C$.

As we have demonstrated that all possible forms of paths to \mathbf{x} are blocked, when we condition \mathbf{x} on C , we may conclude that it is independent of all remaining variables.

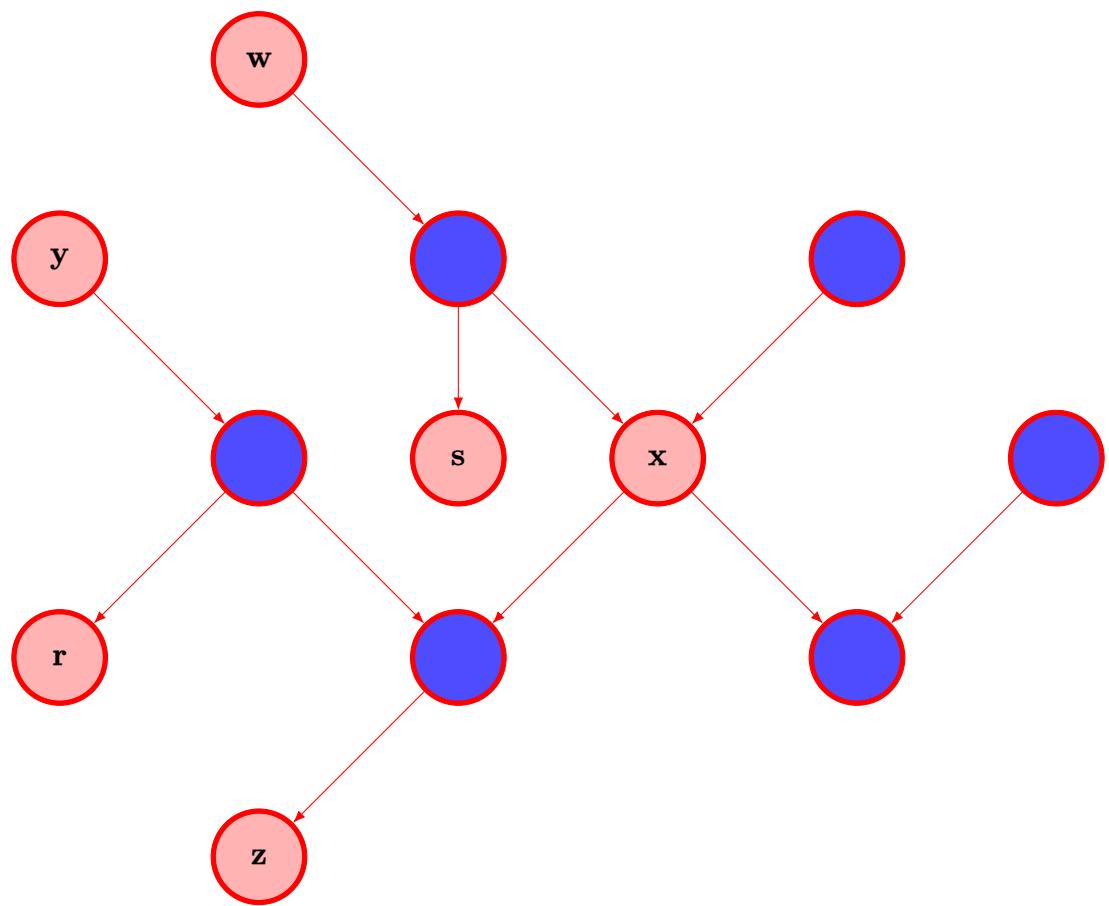


Figure 8.4: Illustration of the Markov blanket for a variable x .

Exercise 8.10

Consider a set of random variables a, b, c and d such that

$$(8.30) \quad p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c).$$

It follows that, by integrating both sides of (8.30), first with respect to d and thereafter with respect to c , we obtain

$$\begin{aligned} \int \int p(a, b, c, d) \, ddc &= \int \int p(a)p(b)p(c|a, b)p(d|c) \, ddc \\ \int p(a, b, c) \, dc &= \int p(a)p(b)p(c|a, b) \, dc && \text{(Apply (1.31))} \\ p(a, b) &= p(a)p(b) && \text{(Apply (1.31)).} \end{aligned}$$

Hence, we conclude that $a \perp\!\!\!\perp b|\emptyset$. Conversely, if we condition (8.30) on d , we have that

$$p(a, b, c|d) = \frac{p(a)p(b)p(c|a, b)p(d|c)}{p(d)} \quad \text{(Apply (1.12))},$$

which does not necessarily imply that a and b are, when conditioned on d , independent. That is, $a \not\perp\!\!\!\perp b|d$.

Exercise 8.11

Consider a probabilistic model for the fuel system of a car, such that B , F , G and D are random variables, respectively, denoting the state of the car's battery ($B = 0$ if the battery is flat, $B = 1$ if the battery is charged), the state of the car's fuel tank ($G = 0$ if the fuel tank is empty, $G = 1$ if the fuel tank is full), the reading on the car's fuel gauge ($G = 0$ if the gauge indicates the tank is empty, $G = 1$ if the Gauge indicates the tank is full) and the driver's report on the gauge's reading ($D = 0$ if the driver reports the gauge reads that the tank is empty, $D = 1$ if the driver reports that the gauge reads that the tank is full). We assign to B and F the following marginal probabilities

$$(8.31) \quad p(B = 1) = 0.9 \quad \text{and} \quad p(F = 1) = 0.9.$$

To G we assign the following conditional probabilities

$$(8.32) \quad p(G = 1|B = 1, F = 1) = 0.8$$

$$(8.33) \quad p(G = 1|B = 1, F = 0) = 0.2$$

$$(8.34) \quad p(G = 1|B = 0, F = 1) = 0.2$$

$$(8.35) \quad p(G = 1|B = 0, F = 0) = 0.1.$$

And to D we assign the following conditional probabilities

$$(8.36) \quad p(D = 1|G = 1) = 0.9$$

$$(8.37) \quad p(D = 0|G = 0) = 0.9.$$

We aim to compute $p(F = 0|D = 0)$. For that purpose, utilizing the results (8.31), (8.32), (8.33), (8.34) and (8.35), we compute $p(G = 0)$ as follows

$$\begin{aligned} p(G = 0) &= p(G = 0, B = 0, F = 0) + p(G = 0, B = 0, F = 1) + \\ &\quad + p(G = 0, B = 1, F = 0) + p(G = 0, B = 1, F = 1) \quad (\text{Apply (1.10)}) \\ &= p(G = 0|B = 0, F = 0)p(B = 0)p(F = 0) + \\ &\quad + p(G = 0|B = 0, F = 1)p(B = 0)p(F = 1) + \\ &\quad + p(G = 0|B = 1, F = 0)p(B = 1)p(F = 0) + \\ &\quad + p(G = 0|B = 1, F = 1)p(B = 1)p(F = 1) \quad (\text{Apply (1.32)}) \\ &= 0.9 \cdot 0.1 \cdot 0.1 + 0.8 \cdot 0.1 \cdot 0.9 + \\ &\quad + 0.8 \cdot 0.9 \cdot 0.1 + 0.2 \cdot 0.9 \cdot 0.9 \\ (8.38) \quad p(G = 0) &= 0.315. \end{aligned}$$

We now compute $p(D = 0)$ as follows

$$\begin{aligned} p(D = 0) &= p(D = 0, G = 0) + p(D = 0, G = 1) \quad (\text{Apply (1.10)}) \\ &= p(D = 0|G = 0)p(G = 0) + \\ &\quad + p(D = 0|G = 1)p(G = 1) \quad (\text{Apply (1.32)}) \\ &= 0.9 \cdot 0.315 + 0.1 \cdot 0.685 \quad (\text{Apply (8.36), (8.37) and (8.38)}) \end{aligned}$$

$$(8.39) \quad p(D = 0) = 0.352.$$

It follows that, utilizing the results (8.31), (8.33), (8.35), (8.36), (8.37) and (8.39), we compute $p(F = 0|D = 0)$ as

$$\begin{aligned}
 p(F = 0|D = 0) &= \frac{p(F = 0, D = 0)}{p(D = 0)} && \text{(Apply (1.12))} \\
 &= \frac{p(F = 0, D = 0, B = 0, G = 0)}{p(D = 0)} + \\
 &\quad + \frac{p(F = 0, D = 0, B = 0, G = 1)}{p(D = 0)} + \\
 &\quad + \frac{p(F = 0, D = 0, B = 1, G = 0)}{p(D = 0)} + \\
 &\quad + \frac{p(F = 0, D = 0, B = 1, G = 1)}{p(D = 0)} && \text{(Apply (1.10))} \\
 &= \frac{p(D = 0|G = 0)p(G = 0|B = 0, F = 0)p(B = 0)p(F = 0)}{p(D = 0)} + \\
 &\quad + \frac{p(D = 0|G = 1)p(G = 1|B = 0, F = 0)p(B = 0)p(F = 0)}{p(D = 0)} + \\
 &\quad + \frac{p(D = 0|G = 0)p(G = 0|B = 1, F = 0)p(B = 1)p(F = 0)}{p(D = 0)} + \\
 &\quad + \frac{p(D = 0|G = 1)p(G = 1|B = 1, F = 0)p(B = 1)p(F = 0)}{p(D = 0)} && \text{(Apply (1.32))} \\
 &= \frac{0.9 \cdot 0.9 \cdot 0.1 \cdot 0.1 + 0.1 \cdot 0.1 \cdot 0.1 \cdot 0.1}{0.352} + \\
 &\quad + \frac{0.9 \cdot 0.8 \cdot 0.9 \cdot 0.1 + 0.1 \cdot 0.2 \cdot 0.9 \cdot 0.1}{0.352}
 \end{aligned}$$

$$(8.40) \quad p(F = 0|D = 0) = 0.2125.$$

Utilizing (1.10), (1.32), (8.31), (8.34), (8.35), (8.36) and (8.37), we compute $p(D = 0, B = 0)$ as

$$\begin{aligned}
 p(D = 0, B = 0) &= p(D = 0, F = 0, G = 0, B = 0) + p(D = 0, F = 0, G = 1, B = 0) + \\
 &\quad + p(D = 0, F = 1, G = 0, B = 0) + p(D = 0, F = 1, G = 1, B = 0) \\
 &= p(D = 0|G = 0)p(G = 0|F = 0, B = 0)p(B = 0)p(F = 0) + \\
 &\quad + p(D = 0|G = 1)p(G = 1|F = 0, B = 0)p(B = 0)p(F = 0) + \\
 &\quad + p(D = 0|G = 0)p(G = 0|F = 1, B = 0)p(B = 0)p(F = 1) + \\
 &\quad + p(D = 0|G = 1)p(G = 1|F = 1, B = 0)p(B = 0)p(F = 1) \\
 &= 0.9 \cdot 0.9 \cdot 0.1 \cdot 0.1 + 0.1 \cdot 0.1 \cdot 0.1 \cdot 0.1 + \\
 &\quad + 0.9 \cdot 0.8 \cdot 0.1 \cdot 0.9 + 0.1 \cdot 0.2 \cdot 0.1 \cdot 0.9
 \end{aligned}$$

$$(8.41) \quad p(D = 0, B = 0) = 0.0748.$$

We proceed, utilizing (8.31), (8.35), (8.36), (8.37) and (8.41), by computing $p(F = 0|D = 0, B = 0)$ as

$$\begin{aligned}
 p(F = 0|D = 0, B = 0) &= \frac{p(F = 0, D = 0, B = 0)}{p(D = 0, B = 0)} && \text{(Apply (1.12))} \\
 &= \frac{p(F = 0, D = 0, B = 0, G = 0)}{p(D = 0, B = 0)} \\
 &\quad + \frac{p(F = 0, D = 0, B = 0, G = 1)}{p(D = 0, B = 0)} && \text{(Apply (1.10))} \\
 &= \frac{p(D = 0|G = 0)p(G = 0|F = 0, B = 0)p(F = 0)p(B = 0)}{p(D = 0, B = 0)} + \\
 &\quad + \frac{p(D = 0|G = 1)p(G = 1|F = 0, B = 0)p(F = 0)p(B = 0)}{p(D = 0, B = 0)} && \text{(Apply (1.32))} \\
 &= \frac{0.9 \cdot 0.9 \cdot 0.1 \cdot 0.1 + 0.1 \cdot 0.1 \cdot 0.1 \cdot 0.1}{0.0748}
 \end{aligned}$$

$$(8.42) \quad p(F = 0|D = 0, B = 0) \approx 0.1096..$$

Note therefore, from (8.40) is greater than (8.42), that is $p(F = 0|D = 0) > p(F = 0|D = 0, B = 0)$. This result is intuitive: the fuel tank, jointly with the battery, informs the gauge, which informs the driver, which then informs us. As we only have the information provided by the driver, who reports that the fuel tank is empty, we attribute a higher probability to the event that the tank is empty. When informed that the battery is flat, as we are aware that the flatness of the battery induces a higher probability for the gauge to read empty, whilst not necessarily being so, we may believe to some degree that the information the gauge provides to the driver is misleading, hence we attribute a lower probability to the fuel tank being empty than previously. Note that, conditional to the information obtained from the drivers report, we will attribute a dependence between the state of the fuel tank F and the battery B , as we will attribute a higher probability to events which jointly agree with the driver's report. This is reminiscent of the result studied in [Exercise 8.10](#), wherein we observed two variables which are marginally independent, that are also dependent on one another when conditioned by a mutual descendent.

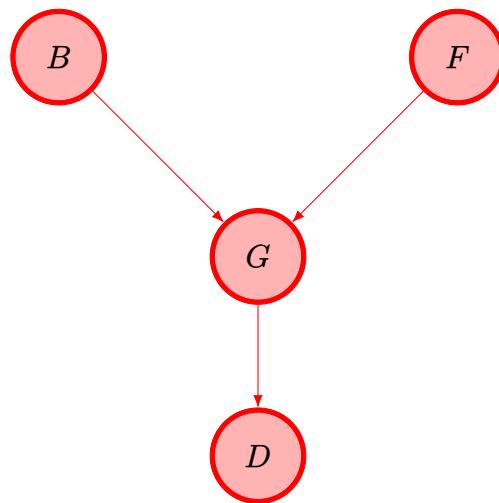


Figure 8.5: Graphical illustration of the car fuel system probabilistic model.

Exercise 8.12

Consider an undirected graph composed of M random variables. We may define a binary matrix Ω , which is $M \times M$ dimensional, which is such that $\Omega_{i,j} = 1$ if there exists a link between variables x_i and x_j , and zero otherwise. For simplicity, consider that $\Omega_{i,i} = 0$, $\forall i \in \{1, \dots, M\}$. Moreover, we may consider that all values above the diagonal are zero, such that Ω is a lower diagonal binary matrix. The number of total variable cells in Ω is therefore

$$\sum_{i=1}^M \sum_{j=1}^{i-1} 1 = \frac{M(M-1)}{2}.$$

As all variable cells in Ω are such that $\Omega_{i,j} \in \{0, 1\}$, that is, they assume one of two states, it suffices for us to take that number of states, and power it by the number of variable cells, that is $2^{M(M-1)/2}$. Hence, for $M = 3$, we have 8 possible graph arrangements, displayed in Figure 8.6.

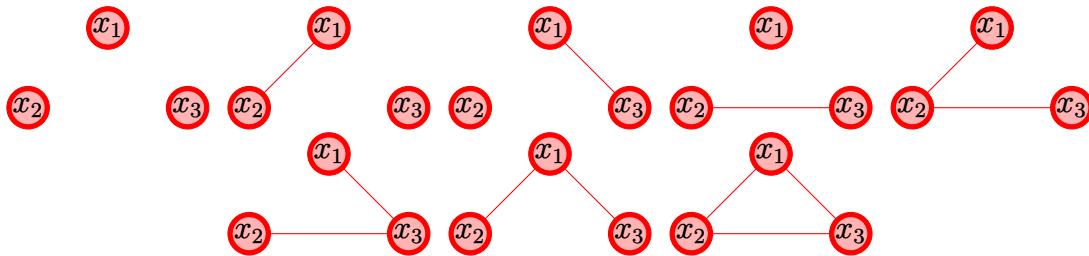


Figure 8.6: All possible arrangements for an undirected graph with $M = 3$ random variables.

Exercise 8.13

We consider the implementation of iterated conditional modes (ICM) to the minimization of the energy function in (8.42). Consider that we observe a data set $\{x_i, y_i\}_{i=1}^D$ of data points, such that $x_i \in \{-1, 1\}$ and $y_i \in \{-1, 1\}$, and we aim to update the value of x_j . For that purpose, we define $\mathbf{x}_{(-j)} = \{x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_D\}$, and evaluate the energy function (8.42), for $\mathbf{x}_{(-j)}$ fixed, and $x_j = -1$ or $x_j = 1$, as follows

$$(8.43) \quad \begin{aligned} E(1, \mathbf{x}_{(-j)}, \mathbf{y}) &= h \sum_{i=1}^D x_i - \beta \sum_{\{i,k\}} x_i x_k - \eta \sum_{i=1}^D x_i y_i \\ E(-1, \mathbf{x}_{(-j)}, \mathbf{y}) &= h \sum_{\substack{i=1 \\ i \neq j}}^D x_i + h - \beta \sum_{\{k'\}} x_{k'} - \beta \sum_{\{i,k\} \setminus \{j,k\}} x_i x_k - \eta y_j - \eta \sum_{\substack{i=1 \\ i \neq j}}^D x_i y_i \end{aligned}$$

and

$$(8.44) \quad \begin{aligned} E(-1, \mathbf{x}_{(-j)}, \mathbf{y}) &= h \sum_{i=1}^D x_i - \beta \sum_{\{i,k\}} x_i x_k - \eta \sum_{i=1}^D x_i y_i \\ E(1, \mathbf{x}_{(-j)}, \mathbf{y}) &= h \sum_{\substack{i=1 \\ i \neq j}}^D x_i - h + \beta \sum_{\{k'\}} x_{k'} - \beta \sum_{\{i,k\} \setminus \{j,k\}} x_i x_k + \eta y_j - \eta \sum_{\substack{i=1 \\ i \neq j}}^D x_i y_i. \end{aligned}$$

Where the sums over k' run over all neighbours of the variable x_j , whilst the sums over $\{i, k\} \setminus \{j, k\}$ runs over all neighbouring pixels, except the pairs which include x_j . We subtract (8.43) by (8.44), obtaining the following

$$E(1, \mathbf{x}_{(-j)}, \mathbf{y}) - E(-1, \mathbf{x}_{(-j)}, \mathbf{y}) = 2h - 2\beta \sum_{\{k'\}} x_{k'} - 2\eta y_j.$$

Note that the difference between (8.43) and (8.44) depends only on the values of y_j and the neighbours of x_j , all of which belong to cliques containing x_j , and hence are all local to x_j in an undirected graph.

Exercise 8.14

We consider once more the Ising model, with energy function as in (8.42), yet now we fix $\beta = h = 0$, such that it is rewritten as

$$(8.45) \quad E(\mathbf{x}, \mathbf{y}) = -\eta \sum_{i=1}^D x_i y_i.$$

As our aim is to minimize (8.45) with respect to \mathbf{x} , for every coordinate x_j of \mathbf{x} we must pick values of x_j such that $x_j y_j > 0$. Trivially, this is satisfied for $x_j = y_j$, as, since $y_j \in \{-1, 1\}$, we have that $y_j^2 = 1 > 0$. So, under this framework, the minimum energy configuration for the Markov field is obtained by setting $x_i = y_i, \forall i \in \{1, \dots, D\}$.

Exercise 8.15

Consider an undirected graph whose underlying distribution is as in (8.49). We obtain the marginal distribution of x_{n-1} and x_n as follows

$$\begin{aligned}
 p(x_{n-1}, x_n) &= \sum_{x_1} \cdots \sum_{x_{n-2}} \sum_{x_{n+1}} \cdots \sum_{x_N} p(\mathbf{x}) && \text{(Apply (1.10))} \\
 &= \sum_{x_1} \cdots \sum_{x_{n-2}} \sum_{x_{n+1}} \cdots \sum_{x_N} \frac{1}{Z} \psi_{1,2}(x_1, x_2) \dots \psi_{N-1,N}(x_{N-1}, x_N) \\
 &= \frac{1}{Z} \psi_{n-1,n}(x_{n-1}, x_n) \times \\
 &\quad \times \left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \dots \left[\sum_{x_1} \psi_{1,2}(x_1, x_2) \right] \right] \times \\
 &\quad \left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \dots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N) \right] \right] \\
 p(x_{n-1}, x_n) &= \frac{1}{Z} \psi_{n-1,n}(x_{n-1}, x_n) \mu_\alpha(x_{n-1}) \mu_\beta(x_n) && \text{(Apply (8.55) and (8.57)).}
 \end{aligned}$$

We thereby reach the desired result.

Exercise 8.16

Consider an undirected graph whose underlying distribution is as in (8.49). We obtain the marginal distribution of x_n and x_N as follows

$$\begin{aligned}
 p(x_n, x_N) &= \sum_{x_1} \cdots \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_{N-1}} p(\mathbf{x}) \\
 &= \sum_{x_1} \cdots \sum_{x_{n-1}} \sum_{x_{n+1}} \cdots \sum_{x_{N-1}} \frac{1}{Z} \psi_{1,2}(x_1, x_2) \dots \psi_{N-1,N}(x_{N-1}, x_N) \\
 &= \frac{1}{Z} \left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \dots \left[\sum_{x_1} \psi_{1,2}(x_1, x_2) \right] \right] \times \\
 &\quad \times \left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \times \right. \\
 &\quad \left. \times \dots \left[\sum_{x_{N-1}} \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \psi_{N-1,N}(x_{N-1}, x_N) \right] \right] \\
 (8.46) \quad p(x_n, x_N) &= \frac{1}{Z} \mu_\alpha(x_n) \bar{\mu}_\beta(x_n, x_N) \tag{Apply (8.55)},
 \end{aligned}$$

where

$$\begin{aligned}
 (8.47) \quad \bar{\mu}_\beta(x_n, x_N) &= \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \left[\sum_{x_{n+2}} \psi_{n+1,n+2}(x_{n+1}, x_{n+2}) \times \right. \\
 &\quad \left. \times \dots \left[\sum_{x_{N-1}} \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \psi_{N-1,N}(x_{N-1}, x_N) \right] \right]
 \end{aligned}$$

$$(8.48) \quad \bar{\mu}_\beta(x_n, x_N) = \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \bar{\mu}_\beta(x_{n+1}, x_N) \tag{Apply (8.47)}.$$

Note we will briefly ignore the need to compute the normalizing factor Z . For $\bar{\mu}_\beta(x_{N-1}, x_N)$ we set

$$\bar{\mu}_\beta(x_{N-1}, N) = \psi_{N-1,N}(x_{N-1}, x_N).$$

By analogy to previous results, we may interpret $p(x_n, x_N)$ from the framework of local message passing, where the message passed forwards through the network (starting at x_1) is not modified from (8.55), whilst the message passed backwards (starting at x_N) is modified to start at x_{N-1} , and is of the form (8.48), as opposed to the form (8.57). This occurs because the sum over x_N is not performed in (8.48), while it is for (8.57). We may relate (8.48) and (8.57) by

$$\mu_\beta(x_n) = \sum_{x_N} \bar{\mu}_\beta(x_n, x_N).$$

Hence, the computation of the joint marginal of x_n and x_N may be adapted very simply from the usual framework of local message passing, resulting in $(N-2)K^2$ operations for the computation of $p(x_n, x_N)$, which, similarly to the previous algorithm, is linear with respect to the number of variables. If we intend to compute $p(x_n, x_N)$ for all variables $n \in \{1, \dots, N-1\}$, we note that, for example, there are several redundancies when computing $p(x_2, x_N)$ and $p(x_3, x_N)$. As $p(x_2, x_N) = \mu_\alpha(x_2) \bar{\mu}_\beta(x_2, x_N)/Z$, $p(x_3, x_N) = \mu_\alpha(x_3) \bar{\mu}_\beta(x_3, x_N)/Z$, and from (8.55) $\mu_\alpha(x_3) = \sum_{x_2} \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2)$ and from (8.48) $\bar{\mu}_\beta(x_2) = \sum_{x_3} \psi_{2,3}(x_2, x_3) \bar{\mu}_\beta(x_3, x_N)$, if we have transmitted all the messages backwards from x_{N-1} to x_3 , we need only perform a sum over x_3 to obtain the message transmitted backwards to x_2 . Conversely, if we have transmitted all the messages forwards from x_1 to x_2 , we need only compute a sum over x_2 to obtain the message transmitted forwards to x_3 , hence, we need only one additional computation for these

scenarios. We conclude that, if we store all computed values of $\mu_\alpha(x_n)$ and $\bar{\mu}_\beta(x_n, x_N)$, we need only compute $\mu_\alpha(x_{N-1})$ and $\bar{\mu}_\beta(x_1, x_N)$, which is $2(N - 2)K^2$ operations, to be able to determine the the marginal distributions $p(x_n, x_N)$, $\forall n \in \{1, \dots, N - 1\}$, as in (8.46). Consider now that we aim to compute the conditional distributions $p(x_n|x_N)$, $\forall n \in \{1, \dots, N - 1\}$. First, consider that we need to compute the marginal distribution $p(x_N)$ as

$$\begin{aligned}
 p(x_N) &= \sum_{x_1} \cdots \sum_{x_{N-1}} p(\mathbf{x}) \\
 &= \sum_{x_1} \cdots \sum_{x_{N-1}} \frac{1}{Z} \psi_{1,2}(x_1, x_2) \dots \psi_{N-1,N}(x_{N-1}, x_N) \\
 &= \frac{1}{Z} \sum_{x_{N-1}} \psi_{N-1,N}(x_{N-1}, x_N) \times \\
 &\quad \times \left[\sum_{x_{N-2}} \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \dots \left[\sum_{x_1} \psi_{1,2}(x_1, x_2) \right] \right] \quad (\text{Apply (8.55)}) \\
 (8.49) \quad p(x_N) &= \frac{1}{Z} \mu_\alpha(x_N).
 \end{aligned}$$

Note that, by computing $p(x_{N-1}, x_N)$, we will have obtained $\mu_\alpha(x_{N-1})$. Assuming this value has been stored, we can obtain $\mu_\alpha(x_N) = \sum_{x_{N-1}} \psi_{N-1,N}(x_{N-1}, x_N) \mu_\alpha(x_{N-1})$ utilizing (8.55), needing only K^2 operations. Alternatively, we may sum $p(x_n, x_N)$ over x_n , which would likewise would cost K^2 operations. Having obtained $p(x_N)$ and $p(x_n, x_N)$, we thereafter compute $p(x_n|x_N)$ utilizing (1.12) and (8.49) as

$$\begin{aligned}
 p(x_n|x_N) &= \frac{p(x_n, x_N)}{p(x_N)} \\
 &= \frac{\frac{1}{Z} \mu_\alpha(x_n) \bar{\mu}_\beta(x_n, x_N)}{\frac{1}{Z} \mu_\alpha(x_N)} \\
 (8.50) \quad p(x_n|x_N) &= \frac{\mu_\alpha(x_n) \bar{\mu}_\beta(x_n, x_N)}{\mu_\alpha(x_N)}.
 \end{aligned}$$

Note that (8.50) is not dependent on the normalizing constant Z , hence if we intend to study the form of the conditional distribution $p(x_n|x_N)$, we need not compute the normalizing constant.

Exercise 8.17

We consider the probabilistic model defined by the undirected graph in Figure 8.7. Utilizing the d-separation criterion, we see that the conditioning set $\{x_3\}$ blocks the only existing path between variables x_2 and x_5 , hence we conclude that $x_2 \perp\!\!\!\perp x_5 | x_3$. We now obtain the conditional distribution $p(x_2|x_3, x_5)$ using (1.12) as

$$\begin{aligned}
 p(x_2|x_3, x_5) &= \frac{p(x_2, x_3, x_5)}{p(x_3, x_5)} \\
 &= \frac{\sum_{x_1} \sum_{x_4} p(\mathbf{x})}{\sum_{x_1} \sum_{x_1} \sum_{x_4} p(\mathbf{x})} \\
 &= \frac{\sum_{x_1} \sum_{x_4} \frac{1}{Z} \psi_{1,2}(x_1, x_2) \dots \psi_{4,5}(x_4, x_5)}{\sum_{x_1} \sum_{x_2} \sum_{x_4} \frac{1}{Z} \psi_{1,2}(x_1, x_2) \dots \psi_{4,5}(x_4, x_5)} \quad (\text{Apply (8.49)}) \\
 &= \frac{\sum_{x_1} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3) [\sum_{x_4} \psi_{3,4}(x_3, x_4) \psi_{4,5}(x_4, x_5)]}{\sum_{x_1} \sum_{x_2} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3) [\sum_{x_4} \psi_{3,4}(x_3, x_4) \psi_{4,5}(x_4, x_5)]} \\
 p(x_2|x_3, x_5) &= \frac{\sum_{x_1} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3)}{\sum_{x_1} \sum_{x_2} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3)}.
 \end{aligned}$$

We conclude that the value of $p(x_2|x_3, x_5)$ is, consequently, independent of the value of x_5 .

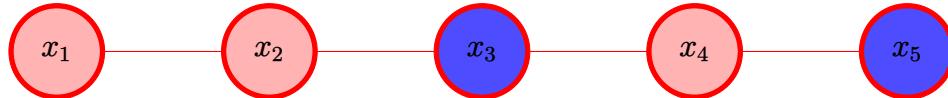


Figure 8.7: Undirected graph corresponding to the model utilized in Exercise 8.17.

Exercise 8.18

Consider that we have a set of random variables x_1, \dots, x_N whose distribution is determined by a directed tree graph. Without loss of generality, we adopt x_1 as the root node, such that the distribution of x_1 has no parents. Let us represent ch_i as the set of children of the i -th random variable. Note that we may do this as the variables of a directed tree possess only one parent. Hence, we conclude that the maximal cliques of the corresponding undirected graph are of the form $\{x_i, x_j\}$, where x_j denotes a child variable of x_i . By repeated applications of (1.32), we may write the joint distribution of all random variables as the product of the probability functions corresponding to each variable conditioned on their respective parents, also multiplied by the marginal distribution of the root variable x_1 . Hence, we define the potentials

$$\psi_{j,i}(x_j, x_i) = p(x_j | x_i)$$

where x_j is not the root (x_1 , in this case), x_j is a child of x_i , and x_i is not a leaf variable. In particular, let $x_{j'}$ denote one, and just one, arbitrary child variable of the root, we set

$$\psi_{j',1}(x_{j'}, x_1) = p(x_{j'} | p_1)p(x_1).$$

Lastly, assuming the conditional distributions are all properly normalized, so too is the product of all potentials, hence their normalizing constant is $Z = 1$. Thus, we write

$$p(\mathbf{x}) = \prod_{\substack{i=1 \\ \text{ch}_i \neq \emptyset}}^N \prod_{j \in \text{ch}_i} \psi_{j,i}(x_j, x_i).$$

We conclude that we may write the joint distribution determined by a directed tree as an equivalent distribution determined by a undirected tree. We now aim to determine how a distribution determined by an undirected tree may be converted into one determined by a directed tree. Note that as an undirected tree graph must have no loops, its maximal cliques are likewise of the form $\{x_i, x_j\}$. Take x_k as the root variable, such that ch_k are its corresponding children. Assuming the potentials of the undirected graph are properly normalized, we define the conditional distributions

$$p(x_j | x_k) = \psi_{j,k}(x_j, x_k) \quad \forall j \in \text{ch}_k \setminus \{j'\}$$

and take one, and just one, child j' , such that joint distribution of $x_{j'}$ and x_k is

$$p(x_{j'}, x_k) = \psi_{j',k}(x_{j'}, x_k).$$

Note that, for a directed graph to be a tree, all its variables must have a single parent. Hence, if we consider a variable x_j in an undirected tree graph that has K links, if it's links were modified to fit the directed tree graph framework, assuming it was not attributed the role of root, once the link which x_j receives from its parent is determined, all other links must be sent to its children. Hence, we find that once we attributed to the variable x_k the role of root, all the children of x_k had their own children determined. We may repeated this argument recursively until we arrive at the leaves of the tree, such that for an arbitrary non-root variable x_j and arbitrary non-leave variable x_i we determine the conditional distribution

$$p(x_j | x_i) = \psi_{j,i}(x_j, x_i).$$

We thereby conclude that, once we attribute to a variable x_k in an undirected tree graph the role of root, the direction of all links in the corresponding directed tree graph is determined. Hence, for an undirected tree graph with N variables, we have N possible corresponding directed tree graphs.

Exercise 8.19

Consider a joint probability function described by the factor graph in Figure 8.8. We attribute to f_0 the role of the root variable, such that f_N is the sole leaf variable. Consider that we aim to determine the marginal distribution $p(x_n)$, it follows from (8.63) that the form of $p(x_n)$ is

$$(8.51) \quad p(x_n) = \mu_{f_{n-1} \rightarrow x_n}(x_n) \mu_{f_n \rightarrow x_n}(x_n).$$

We start the sum-product algorithm at the leaf variable utilizing (8.71), yielding

$$\mu_{f_N \rightarrow x_N}(x_N) = f_N(x_N).$$

We now recursively alternate between (8.66) and (8.69) until we derive $\mu_{f_n \rightarrow x_n}(x_n)$ as follows

$$(8.52) \quad \begin{aligned} \mu_{x_N \rightarrow f_{N-1}}(x_N) &= \mu_{f_N \rightarrow x_N}(x_N) \\ \mu_{f_{N-1} \rightarrow x_{N-1}}(x_{N-1}) &= \sum_{x_N} f_{N-1}(x_{N-1}, x_N) \mu_{x_N \rightarrow f_{N-1}}(x_N) \\ \mu_{x_{N-1} \rightarrow f_{N-2}}(x_{N-1}) &= \mu_{f_{N-1} \rightarrow x_{N-1}}(x_{N-1}) \\ \mu_{f_{N-2} \rightarrow x_{N-2}}(x_{N-2}) &= \sum_{x_{N-1}} f_{N-2}(x_{N-2}, x_{N-1}) \mu_{x_{N-1} \rightarrow f_{N-2}}(x_N) \\ &\vdots \\ \mu_{x_{n+1} \rightarrow f_n}(x_{n+1}) &= \mu_{f_{n+1} \rightarrow x_{n+1}}(x_{n+1}) \\ \mu_{f_n \rightarrow x_n}(x_n) &= \sum_{x_{n+1}} f_n(x_n, x_{n+1}) \mu_{x_{n+1} \rightarrow f_n}(x_{n+1}). \end{aligned}$$

Thereafter, we start the sum-product algorithm at the root variable utilizing (8.70), yielding

$$\mu_{f_0 \rightarrow x_1}(x_1) = f_0(x_1).$$

Again, we alternate between (8.66) and (8.69) until we derive $\mu_{f_{n-1} \rightarrow x_n}(x_n)$, as follows

$$(8.53) \quad \begin{aligned} \mu_{x_1 \rightarrow f_1}(x_1) &= \mu_{f_0 \rightarrow x_1}(x_1) \\ \mu_{f_1 \rightarrow x_2}(x_2) &= \sum_{x_1} f_1(x_1, x_2) \mu_{x_1 \rightarrow f_1}(x_1) \\ \mu_{x_2 \rightarrow f_2}(x_2) &= \mu_{f_1 \rightarrow x_2}(x_2) \\ \mu_{f_2 \rightarrow x_3}(x_3) &= \sum_{x_2} f_2(x_2, x_3) \mu_{x_2 \rightarrow f_2}(x_2) \\ &\vdots \\ \mu_{x_{n-1} \rightarrow f_{n-1}}(x_{n-1}) &= \mu_{f_{n-2} \rightarrow x_{n-1}}(x_{n-1}) \\ \mu_{f_{n-1} \rightarrow x_n}(x_n) &= \sum_{x_{n-1}} f_{n-1}(x_{n-1}, x_n) \mu_{x_{n-1} \rightarrow f_{n-1}}(x_{n-1}). \end{aligned}$$

By backsubstituting the messages in (8.52) and (8.53), we obtain

$$(8.54) \quad \begin{aligned} \mu_{f_n \rightarrow x_n}(x_n) &= \sum_{x_{n+1}} f_n(x_n, x_{n+1}) \left[\sum_{x_{n+2}} f_{n+1}(x_{n+1}, x_{n+2}) \times \right. \\ &\quad \left. \times \dots \left[\sum_{x_N} f_{N-1}(x_{N-1}, x_N) f_N(x_N) \right] \right] \end{aligned}$$

and

$$(8.55) \quad \mu_{f_{n-1} \rightarrow x_n}(x_n) = \sum_{x_{n-1}} f_{n-1}(x_{n-1}, x_n) \left[\sum_{x_{n-2}} f_{n-2}(x_{n-2}, x_{n-1}) \dots \left[\sum_{x_1} f_1(x_1, x_2) f_0(x_1) \right] \right].$$

We now set

$$(8.56) \quad \frac{1}{\sqrt{Z}} \psi_{1,2}(x_1, x_2) = f_0(x_1) f_1(x_1, x_2)$$

$$(8.57) \quad \frac{1}{\sqrt{Z}} \psi_{N-1,N}(x_{N-1}, x_N) = f_N(x_N) f_{N-1}(x_{N-1}, x_N)$$

$$(8.58) \quad \psi_{n,n+1}(x_n, x_{n+1}) = f_n(x_n, x_{n+1}) \quad \forall n \in \{2, \dots, N-1\},$$

where

$$Z = \sum_{x_1} \dots \sum_{x_N} \psi_{1,2}(x_1, x_2) \dots \psi_{N-1,N}(x_{N-1}, x_N).$$

Substituting (8.56), (8.57) and (8.58) into (8.54), we obtain

$$\begin{aligned} \mu_{f_n \rightarrow x_n}(x_n) &= \frac{1}{\sqrt{Z}} \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \left[\sum_{x_{n+2}} \psi_{n+1,n+2}(x_{n+1}, x_{n+2}) \times \right. \\ &\quad \times \dots \left. \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N) \right] \right] \\ (8.59) \quad \mu_{f_n \rightarrow x_n}(x_n) &= \frac{1}{\sqrt{Z}} \mu_\beta(x_n) \end{aligned} \quad (\text{Apply (8.57)})$$

and substituting (8.56), (8.57) and (8.58) into (8.55), we obtain

$$\begin{aligned} \mu_{f_{n-1} \rightarrow x_n}(x_n) &= \frac{1}{\sqrt{Z}} \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \left[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \times \right. \\ &\quad \times \dots \left. \left[\sum_{x_1} \psi_{1,2}(x_1, x_2) \right] \right] \\ (8.60) \quad \mu_{f_{n-1} \rightarrow x_n}(x_n) &= \frac{1}{\sqrt{Z}} \mu_\alpha(x_n) \end{aligned} \quad (\text{Apply (8.55)}).$$

Utilizing (8.51), (8.59) and (8.60), we obtain

$$\begin{aligned} p(x_n) &= \mu_{f_{n-1} \rightarrow x_n}(x_n) \mu_{f_n \rightarrow x_n}(x_n) \\ p(x_n) &= \frac{1}{Z} \mu_\alpha(x_n) \mu_\beta(x_n). \end{aligned}$$

We hence derive (8.54) utilizing the sum-product algorithm.



Figure 8.8: Factor graph corresponding to the model utilized in [Exercise 8.19](#).

Exercise 8.20

We aim to demonstrate, by induction, that the message passing protocol for the sum-product algorithm is valid in the sense that at every step where a variable or factor must send an outgoing message to another variable or node, it has received the necessary incoming messages to do so. First, we take a simple factor graph, comprised of a single factor and a single variable, as in Figure 8.9. Note that this graph is a tree. We may attribute to one node the role of root, so that the other corresponds to the leaf, and as there are no other links, the nodes don't need to receive any incoming messages before sending their outgoing message. We may simply perform the message passing from the leaf to the root as in (8.70), and the root to the leaf as in (8.71), or vice-versa, dependent on which node is considered the root. Hence, we conclude that the sum-product algorithm is valid for a factor graph comprised of only two nodes (one variable and one factor). We now assume the message passing protocol holds for a tree factor graph comprised of N total nodes. If we add an additional node, which may be a factor or a variable, as under our framework the graph must remain a tree, this node will be a leaf, hence, when we start the message passing protocol from the leaf to the root, this new node has no incoming messages, and so it will send an outgoing message to its one neighbouring node as in (8.70) and (8.71). Once its neighbouring node receives the incoming message, it may send its own outgoing message, and as we assume the message passing protocol holds for the remaining nodes in the graph, the stage of message passing from the leaves to the root is complete. When the message passing protocol from the root to the leaves occurs, as we assume it holds for the tree factor graph, the neighbouring node to the new node will receive all its incoming messages, and thereafter the new node will receive its only incoming message as in (8.66) or (8.69). Hence, if we assume the message passing protocol holds for a factor tree graph comprised of N nodes, it must hold for a factor tree graph comprised of $N + 1$ total nodes. As we also find it holds for $N = 2$, we have proven by induction that the sum-product message passing protocol holds for any tree factor graph.

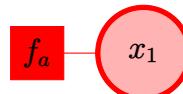


Figure 8.9: Factor graph comprised of a single factor and a single variable.

Exercise 8.21

We aim to determine, for a factor tree graph, how we may determine the marginal of \mathbf{x}_s , which indicates the set of variables linked to the s -th factor in the graph. It follows that

$$\begin{aligned}
 p(\mathbf{x}_s) &= \sum_{\mathbf{x} \setminus \mathbf{x}_s} p(\mathbf{x}) && \text{(Apply (1.10))} \\
 &= \sum_{\mathbf{x} \setminus \mathbf{x}_s} \prod_r f_r(\mathbf{x}_r) && \text{(Apply (8.59))} \\
 &= \sum_{\mathbf{x} \setminus \mathbf{x}_s} f_s(\mathbf{x}_s) \prod_{r \setminus s} f_r(\mathbf{x}_r) \\
 &= \sum_{\mathbf{x} \setminus \mathbf{x}_s} f_s(\mathbf{x}_s) \prod_{i \in \text{ne}(f_s)} \prod_{j \in \text{ne}(x_i) \setminus f_s} F_j(x_i, X_{j,i}) \\
 &= f_s(\mathbf{x}_s) \prod_{i \in \text{ne}(f_s)} \sum_{\mathbf{x} \setminus \mathbf{x}_s} G_i(x_i, X_{s,i}) && \text{(Apply (8.68))} \\
 &= f_s(\mathbf{x}_s) \prod_{i \in \text{ne}(f_s)} \sum_{X_{s,j}} G_i(x_i, X_{s,i}) \\
 p(\mathbf{x}_s) &= f_s(\mathbf{x}_s) \prod_{i \in \text{ne}(f_s)} \mu_{x_i \rightarrow f_s}(x_i) && \text{(Apply (8.67)).}
 \end{aligned}$$

Hence, we derive (8.72).

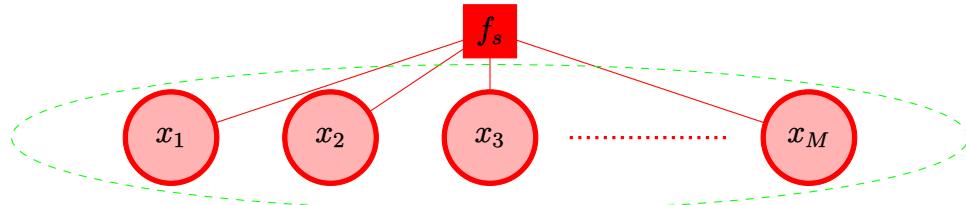


Figure 8.10: Factor graph corresponding to the model utilized in Exercise 8.21.

Exercise 8.22

Consider now a context similar to that of Exercise [Exercise 8.21](#), except we now aim to determine the joint marginal of $\mathbf{x}_a, \dots, \mathbf{x}_z$, which are sets of variables linked respectively to factors f_a, \dots, f_z . Note that we assume that these sets of variables are disjoint (this follows from the assumption that any one variable connects to another via a single node). Let $S = \{a, \dots, z\}$ denote the index set of factors under consideration, we compute the corresponding marginal distribution as

$$\begin{aligned} p(\mathbf{x}_a, \dots, \mathbf{x}_z) &= \sum_{\mathbf{x} \setminus \{\mathbf{x}_a, \dots, \mathbf{x}_z\}} p(\mathbf{x}) && \text{(Apply (1.10))} \\ &= \sum_{\mathbf{x} \setminus \{\mathbf{x}_a, \dots, \mathbf{x}_z\}} f_a(\mathbf{x}_a) \dots f_z(\mathbf{x}_z) \prod_{s \in S} \prod_{x_j \in \mathbf{x}_s} F_s(x_j, X_{s,j}) \\ &= f_a(\mathbf{x}_a) \dots f_z(\mathbf{x}_z) \prod_{s \in S} \prod_{x_j \in \mathbf{x}_s} \sum_{X_{s,j}} F_s(x_j, X_{s,j}) \\ p(\mathbf{x}_a, \dots, \mathbf{x}_z) &= f_a(\mathbf{x}_a) \dots f_z(\mathbf{x}_z) \prod_{s \in S} \prod_{x_j \in \mathbf{x}_s} \mu_{f_s \rightarrow x_j}(x_j) && \text{(Apply (8.63)).} \end{aligned}$$

Hence, we may compute the marginal distribution of $\mathbf{x}_a, \dots, \mathbf{x}_z$ by applying the sum-product algorithm, starting at leaf nodes outside of $\mathbf{x}_a, \dots, \mathbf{x}_z$ and the corresponding factors f_a, \dots, f_z .

Exercise 8.23

We aim to demonstrate that the marginal distribution of a variable x_i in a factor graph may be written as the product of an incoming message received by the variable from one of its factors and the outgoing message x_i sends to that same factor. It follows that

$$\begin{aligned} p(x_i) &= \prod_{j \in \text{ne}(x_i)} \mu_{f_j \rightarrow x_i}(x_i) && \text{(Apply (8.63))} \\ &= \mu_{f_k \rightarrow x_i}(x_i) \prod_{j \in \text{ne}(x_i) \setminus f_k} \mu_{f_j \rightarrow x_i}(x_i) \\ p(x_i) &= \mu_{f_k \rightarrow x_i}(x_i) \mu_{x_i \rightarrow f_k}(x_i) && \text{(Apply (8.69)),} \end{aligned}$$

where $k \in \text{ne}(x_i)$. Hence, we derive the desired result.

Exercise 8.24

This Exercise follows as a restatement of [Exercise 8.21](#).

Exercise 8.25

We aim to apply the sum-product algorithm to the factor graph in Figure 8.11, with corresponding unnormalized distribution function as in (8.73), such that we obtain the marginal distribution of x_1 and x_3 , and the joint marginal distribution of x_1 and x_2 . It follows that

$$\begin{aligned}
 p(x_1) &= \mu_{f_a \rightarrow x_1}(x_1) && \text{(Apply (8.63))} \\
 &= \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \rightarrow f_a}(x_2) && \text{(Apply (8.83))} \\
 &= \sum_{x_2} f_a(x_1, x_2) \mu_{f_b \rightarrow x_2}(x_2) \mu_{f_c \rightarrow x_2}(x_2) && \text{(Apply (8.82))} \\
 &= \sum_{x_2} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) && \text{(Apply (8.77) and (8.81))} \\
 p(x_1) &= \sum_{x_2} \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4).
 \end{aligned}$$

Which corresponds to the marginal distribution of x_1 as computed via (1.10). It follows also that

$$\begin{aligned}
 p(x_3) &= \mu_{f_b \rightarrow x_3}(x_3) && \text{(Apply (8.63))} \\
 &= \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \rightarrow f_b}(x_2) && \text{(Apply (8.79))} \\
 &= \sum_{x_2} f_b(x_2, x_3) \mu_{f_a \rightarrow x_2}(x_2) \mu_{f_c \rightarrow x_2}(x_2) && \text{(Apply (8.78))} \\
 &= \sum_{x_2} f_b(x_2, x_3) \sum_{x_1} f_a(x_1, x_2) \sum_{x_4} f_c(x_2, x_4) && \text{(Apply (8.75) and (8.77))} \\
 p(x_3) &= \sum_{x_1} \sum_{x_2} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4).
 \end{aligned}$$

Which corresponds to the marginal distribution of x_3 as computed via (1.10). Lastly, we have that

$$\begin{aligned}
 p(x_1, x_2) &= f_a(x_1, x_2) \mu_{x_1 \rightarrow f_a}(x_1) \mu_{x_2 \rightarrow f_a}(x_2) && \text{(Apply (8.72))} \\
 &= f_a(x_1, x_2) \mu_{f_b \rightarrow x_2}(x_2) \mu_{f_c \rightarrow x_2}(x_2) && \text{(Apply (8.74) and (8.82))} \\
 &= f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) && \text{(Apply (8.77) and (8.81))} \\
 p(x_1, x_2) &= \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4).
 \end{aligned}$$

Which corresponds to the marginal joint distribution of x_1 and x_2 as computed via (1.10).

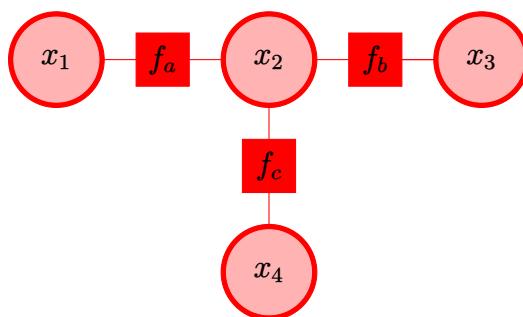


Figure 8.11: Factor graph corresponding to the model utilized in Exercise 8.25.

Exercise 8.26

Consider a factor graph of variables \mathbf{x} , which includes discrete variables x_a and x_b , respectively possessing K_a and K_b possible states. We aim to determine a procedure to obtain the joint marginal distribution of x_a and x_b . Firstly, consider from (1.32) that we may write

$$p(x_a, x_b) = p(x_a|x_b)p(x_b).$$

We suggest the following procedure: first, we run the usual sum-product algorithm, and obtain the marginal distribution $p(x_b)$. We thereafter clamp the values of x_b at each of its possible states by multiplying the joint distribution $p(\mathbf{x})$ by $I(x_b, \hat{x}_b)$, where

$$I(x_b, \hat{x}_b) = \begin{cases} 1 & \text{if } x_b = \hat{x}_b, \\ 0 & \text{otherwise.} \end{cases}$$

By running the the sum-product algorithm on $p(\mathbf{x})I(x_b, \hat{x}_b)$, we obtain a nonnormalized version of $p(x_a|x_b = \hat{x}_b)$. By simply summing over the K_a states of x_a , we may normalize this, and properly obtain $p(x_a|x_b = \hat{x}_b)$. By repeating this procedure for each of the K_b states of x_b , we obtain the joint marginal distribution $p(x_a, x_b)$.

Exercise 8.27

Consider two random variables x, y such that $x, y \in \{0, 1, 2\}$. We propose the joint distribution of x and y in Table 8.4, such that the marginal distribution of x is determined as

$$\begin{aligned} p(x = 0) &= p(x = 0, y = 0) + p(x = 0, y = 1) + p(x = 0, y = 2) \quad (\text{Apply (1.31)}) \\ &= \frac{1}{10} + \frac{1}{5} + 0 \\ p(x = 0) &= \frac{3}{10} \\ p(x = 1) &= p(x = 1, y = 0) + p(x = 1, y = 1) + p(x = 1, y = 2) \quad (\text{Apply (1.31)}) \\ &= \frac{1}{5} + 0 + \frac{1}{5} \\ p(x = 1) &= \frac{2}{5} \\ p(x = 2) &= p(x = 2, y = 0) + p(x = 2, y = 1) + p(x = 2, y = 2) \quad (\text{Apply (1.31)}) \\ &= 0 + \frac{1}{5} + \frac{1}{10} \\ p(x = 2) &= \frac{3}{10}. \end{aligned}$$

In performing analogous computation for the marginal distribution of y , we find that it is identical to the marginal distribution of x . Hence, we find that the value of \hat{x} (equivalently \hat{y}) which maximizes the $p(x)$ (equivalently $p(y)$) is $\hat{x} = 1$ (equivalently $\hat{y} = 1$). However, from Table 8.4, we have that $p(\hat{x}, \hat{y}) = p(1, 1) = 0$.

x	y	$p(x, y)$
0	0	1/10
0	1	1/5
0	2	0
1	0	1/5
1	1	0
1	2	1/5
2	0	0
2	1	1/5
2	2	1/10

Table 8.4: Proposed joint distribution of x and y for Exercise 8.27.

Exercise 8.28

Consider an arbitrary factor graph, in which there is a cycle represented by a path within it of the form $A = \{x_1, f_1, x_2, f_2, \dots, x_M, f_M, x_1\}$ (where the order of the links are as presented in the set A). We can show that, within this very cycle, there will always be a pending message: it follows that we will ignore the messages pending to nodes outside this cycle, as they will not be relevant. Assume that x_1 has received messages from all its links, including its link to node f_M . Hence, there is a pending message from x_1 to f_1 . Once that message is sent, there will be a pending message from f_1 to x_2 . We may proceed as thus recursively, until we arrive at f_M , which will have a message pending to x_1 . Once it sends that message, x_1 will have a pending message to f_1 , and we return to the initial state, having travelled only once across every link in the path A . We conclude hence that, at any point, there must be at least one pending message within this loop.

Exercise 8.29

We aim to demonstrate by induction that by running the sum-product procedure on a factor graph tree, in a finite number of steps there will be no pending messages. First, we consider a factor graph tree with just two nodes, one factor node and one variable node, as in Figure 8.9. Note that this graph is a tree. Similarly to the context of Exercise 8.20, we find that neither node needs to receive any incoming message before sending its outgoing message, and after the messages are sent they result in no pending messages. We therefore find that after the message travels across both directions there are no pending messages necessary, hence we conclude that for $N = 2$ nodes the number of steps before there are no more pending messages is finite. We assume for a factor graph tree with N nodes that the number of steps before there are no more pending messages is finite. If we add an additional node (which we may denote as x_{N+1}), as the graph must remain a tree, the node must be a leaf. Denote the node in the original graph which is linked to the new node as x_k . The node x_k receives all incoming messages in a finite number of steps, by assumption, and thereafter sends its outgoing message to x_{N+1} , generating no pending messages. Conversely, if x_{N+1} sends its outgoing message to x_k , it results in x_k having a pending message. As we assume the remaining graph of N nodes has no pending messages within a finite number of steps, likewise in a finite number of steps there will be no pending messages on the original graph. Hence, we conclude that if a factor tree graph with N nodes has no pending messages within a finite number of steps, a factor tree graph with $N + 1$ nodes must also have no pending messages within a finite number of steps. As we have also demonstrated that a factor tree graph with $N = 2$ nodes has no pending messages within a finite number of steps, we conclude by induction that the sum-product algorithm has no pending messages when applied to any tree factor tree graph.

Chapter 9

Mixture Models and EM

Exercise 9.1

Consider the K -means algorithm obtained by minimizing the distortion measure in (9.1). Let I be the set of all possible assignments to all data points in a set $\{\mathbf{x}_n\}_{n=1}^N$. It follows that I is composed of K^N elements. For every element $i \in I$ we may define $\boldsymbol{\mu}^{(i)} = \{\boldsymbol{\mu}_1^{(i)}, \dots, \boldsymbol{\mu}_K^{(i)}\}$ as

$$\boldsymbol{\mu}^{(i)} = \arg \min_{\boldsymbol{\mu}} \sum_{n=1}^N \sum_{k=1}^K r_{n,k}^{(i)} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2,$$

where the assignments $r_{n,k}^{(i)}$ are drawn from i . That is, $\boldsymbol{\mu}^{(i)}$ is the set of centre vectors which minimizes (9.1) for the set of assignments $i \in I$. We may hereafter determine the set of centre vectors $\boldsymbol{\mu}^*$ that minimizes the distortion measure J across all possible assignments as

$$(9.1) \quad \boldsymbol{\mu}^* = \boldsymbol{\mu}^{(i^*)} \quad \text{where} \quad i^* = \arg \min_{i \in I} \sum_{n=1}^N \sum_{k=1}^K r_{n,k}^{(i)} \|\mathbf{x}_n - \boldsymbol{\mu}_k^{(i)}\|^2.$$

Hence, the distortion measure J possesses a global minimum. With respect to the EM algorithm applied to the K -means problem, note that the E step, in which the assignments are updated as in (9.2), will reduce the distortion measure J , whilst, by definition, in the M step the cluster centres are updated by minimizing the distortion measure J , conditional to the previous assignments. Hence, J is always reduced at every step of the EM algorithm. Once the assignments $i^* \in I$ (as defined in (9.1)) are attributed to the data points, which occurs in a finite amount of steps as the number of assignments is $K^N < \infty$, the centre vectors are computed as in (9.4) (yielding $\boldsymbol{\mu}^*$ as in (9.1)), and when the E step is next initiated, the assignments will not change (by (9.2)). As the assignments are not changed, from (9.4) the centre vectors are also unchanged.

Exercise 9.2

We consider applying the Robbins-Monro sequential estimation procedure for the centre vectors in the K -means algorithm procedure. Consider that our data set has already processed the data set $\{\mathbf{x}_n\}_{n=1}^{N-1}$, and we aim to incorporate the data point \mathbf{x}_N to the estimate of the centre vectors, such that the individual contribution of the \mathbf{x}_N data point is

$$(9.2) \quad J_N = \sum_{j=1}^K r_{N,j} \|\mathbf{x}_N - \boldsymbol{\mu}_j^{(N-1)}\|^2.$$

From (9.2), we define $r_{N,j}$ as

$$r_{N,j} = \begin{cases} 1 & \text{if } j = \arg \min_i \|\mathbf{x}_N - \boldsymbol{\mu}_i^{(N-1)}\|^2, \\ 0 & \text{otherwise.} \end{cases}$$

That is, $r_{N,j}$ determines the class of the data point \mathbf{x}_N based on the centre vector estimates of the previously processed data. Consider, without loss of generality, that N -th data point is attributed to the k -th class, such that $r_{N,k} = 1$, and $r_{N,j} = 0, j \neq k$. We rewrite (9.2) as

$$(9.3) \quad J_N = \|\mathbf{x}_N - \boldsymbol{\mu}_k^{N-1}\|^2.$$

As only the value of the centre vector associated with the k -th class is modified when the \mathbf{x}_N data point is added, we substitute $\boldsymbol{\mu}_k^{(N-1)}$ and (9.3) into (2.135), obtaining the following

$$\begin{aligned} \boldsymbol{\mu}_k^{(N)} &= \boldsymbol{\mu}_k^{(N-1)} - a_{N-1} \frac{\partial J_N}{\partial \boldsymbol{\mu}_k} \\ &= \boldsymbol{\mu}_k^{(N-1)} - a_{N-1} \frac{\partial}{\partial \boldsymbol{\mu}_k} \|\mathbf{x}_N - \boldsymbol{\mu}_k^{N-1}\|^2 && \text{(Apply (9.3))} \\ &= \boldsymbol{\mu}_k^{(N-1)} - a_{N-1} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\mathbf{x}_N - \boldsymbol{\mu}_k^{N-1})^\top (\mathbf{x}_N - \boldsymbol{\mu}_k^{N-1}) \\ \boldsymbol{\mu}_k^{(N)} &= \boldsymbol{\mu}_k^{(N-1)} + 2a_{N-1} (\mathbf{x}_N - \boldsymbol{\mu}_k^{(N-1)}) && \text{(Apply (C.19) and (C.20)).} \end{aligned}$$

By taking $2a_{N-1} = \eta_N$, we derive (9.5).

Exercise 9.3

We aim to demonstrate that, if we adopt a Gaussian conditional distribution for \mathbf{x} of the form (9.11), where the latent variables \mathbf{z} are distributed as (9.10), the marginal distribution of \mathbf{x} is as in (9.7). It follows that

$$\begin{aligned}
 p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) && \text{(Apply (1.10))} \\
 &= \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) && \text{(Apply (1.32))} \\
 &= \sum_{k=1}^K \sum_{\{z_k=1\}} \prod_{j=1}^K \pi_j^{z_j} p(\mathbf{x}|\mathbf{z}) && \text{(Apply (9.10))} \\
 &= \sum_{k=1}^K \pi_k p(\mathbf{x}|z_k = 1) \\
 &= \sum_{k=1}^K \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && \text{(Apply (9.10)).}
 \end{aligned}$$

Hence, we derive (9.7).

Exercise 9.4

Consider that we have a model comprised of observed variables \mathbf{X} , latent variables \mathbf{Z} and parameters θ , such that we attribute to the parameters θ a prior $p(\theta)$, and a joint conditional distribution for the observed and latent variables of the form $p(\mathbf{X}, \mathbf{Z}|\theta)$. We aim to modify the EM algorithm such that the parameter values are updated according to the maximum posterior values of $p(\theta|\mathbf{X})$. First, consider from (1.12) that we may rewrite $p(\theta|\mathbf{X})$ as

$$(9.4) \quad \begin{aligned} p(\theta|\mathbf{X})p(\mathbf{X}) &= p(\mathbf{X}|\theta)p(\theta) \\ p(\theta|\mathbf{X}) &\propto p(\mathbf{X}|\theta)p(\theta). \end{aligned}$$

As the prior distribution of θ is not directly dependent on the latent variables, we conclude that the E step need not be modified. Conversely, for the M step, we aim to maximize the logarithm of the posterior distribution. From (9.4), we have that the maximization of the posterior $p(\theta|\mathbf{X})$ with respect to θ is equivalent to the maximization of $p(\mathbf{X}|\theta)p(\theta)$ with respect to θ (as they are proportional). Hence, when constructing the complete-data posterior likelihood $\tilde{\mathcal{Q}}(\theta, \theta^{\text{old}})$ we may take the expectation with respect to \mathbf{Z} of $p(\mathbf{X}|\theta)p(\theta)$, as follows

$$(9.5) \quad \begin{aligned} \tilde{\mathcal{Q}}(\theta, \theta^{\text{old}}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \log\{p(\mathbf{X}, \mathbf{Z}|\theta)p(\theta)\} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\theta) + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \log p(\theta) \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\theta) + \log p(\theta) \quad (\text{Apply (1.26)}) \\ \tilde{\mathcal{Q}}(\theta, \theta^{\text{old}}) &= \mathcal{Q}(\theta, \theta^{\text{old}}) + \log p(\theta). \end{aligned}$$

We hence modify the M step so that we obtain $\theta^{\text{new}} = \arg \max \tilde{\mathcal{Q}}(\theta, \theta^{\text{old}})$. We have thereby derived the desired result.

Exercise 9.5

Consider the probabilistic graph model which delineates a Gaussian mixture model. Note that for any two latent variables \mathbf{z}_i and \mathbf{z}_j , where $i \neq j$, all possible paths between those must be similar to that seen in Figure 9.1. We note that parameter nodes (corresponding to π , μ and Σ), for the purpose of the d-separation criterion, are similar to observed nodes in the sense they may block (or unblock) paths through them. As the paths meet at the parameter nodes as tail to tail, and the parameter nodes are technically on the conditioning set, we have that $\mathbf{z}_i \perp\!\!\!\perp \mathbf{z}_j | \mu, \Sigma, \pi$, for any two nodes \mathbf{z}_i and \mathbf{z}_j . Hence, we may factorize \mathbf{Z} across all different data points.

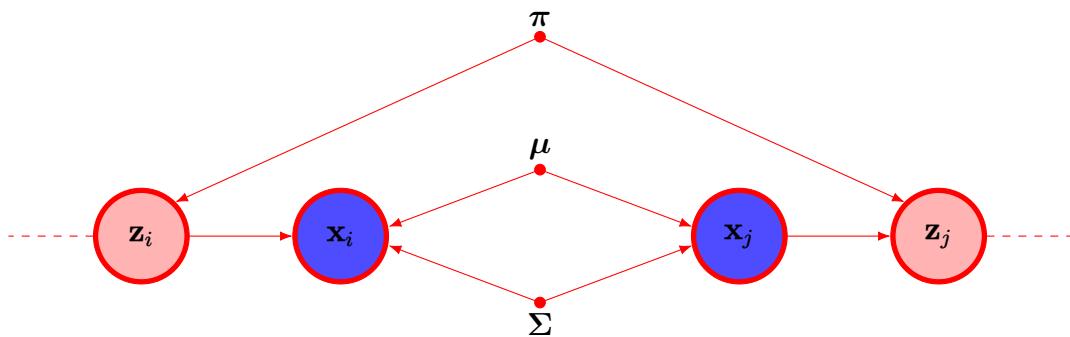


Figure 9.1: Illustration of the form of all possible path between two nodes \mathbf{z}_i and \mathbf{z}_j in the Gaussian mixture model.

Exercise 9.6

$$\begin{aligned}
 \frac{\partial \log p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} &= \mathbf{0} \\
 \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[\sum_{n=1}^N \log \left\{ \sum_{k=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right\} \right] &= \mathbf{0} \\
 \sum_{n=1}^N \frac{1}{\sum_{j=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \sum_{k=1}^K \frac{\partial p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} &= \mathbf{0} \\
 \sum_{n=1}^N \frac{1}{\sum_{j=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \sum_{k=1}^K \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[(2\pi)^{-D/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right\} \right] &= \mathbf{0} \\
 \sum_{n=1}^N \frac{1}{\sum_{j=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \sum_{k=1}^K \frac{\partial}{\partial \boldsymbol{\Sigma}} \left[(2\pi)^{-D/2} \exp \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right\} \right] &= \mathbf{0} \\
 \sum_{n=1}^N \frac{1}{\sum_{j=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \sum_{k=1}^K (2\pi)^{-D/2} \exp \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right\} \times & \\
 \times \left[-\frac{1}{2} \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} \right] &= \mathbf{0} \\
 \sum_{n=1}^N \frac{1}{\sum_{j=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \sum_{k=1}^K p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} \right] &= \mathbf{0} \\
 \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{n,k}) \left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} \right] &= \mathbf{0} \\
 N \boldsymbol{\Sigma} - \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{n,k}) (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top &= \mathbf{0} \\
 \sum_{n=1}^N \sum_{k=1}^K \frac{\gamma(z_{n,k})}{N} (\mathbf{x}_n - \boldsymbol{\mu}_k) (\mathbf{x}_n - \boldsymbol{\mu}_k)^\top &= \boldsymbol{\Sigma}
 \end{aligned}$$

Exercise 9.7

Exercise 9.8

Exercise 9.9

Exercise 9.10

$$\begin{aligned}
 p(\mathbf{x}_b | \mathbf{x}_a = \hat{\mathbf{x}}_a) &= \frac{p(\mathbf{x}_a = \hat{\mathbf{x}}_a, \mathbf{x}_b)}{p(\mathbf{x}_a = \hat{\mathbf{x}}_a)} \\
 &= \frac{p(\mathbf{x}_a = \hat{\mathbf{x}}_a, \mathbf{x}_b)}{\int p(\mathbf{x}) d\mathbf{x}_b} \\
 &= \frac{\sum_{k=1}^K \pi_k p(\mathbf{x}_a = \hat{\mathbf{x}}_a, \mathbf{x}_b | k)}{\int \sum_{j=1}^K \pi_j p(\mathbf{x}_a = \hat{\mathbf{x}}_a, \mathbf{x}_b | j) d\mathbf{x}_b} \\
 &= \frac{\sum_{k=1}^K \pi_k p(\mathbf{x}_a = \hat{\mathbf{x}}_a, \mathbf{x}_b | k)}{\sum_{j=1}^K \pi_j p(\mathbf{x}_a = \hat{\mathbf{x}}_a | j)} \\
 p(\mathbf{x}_b | \mathbf{x}_a = \hat{\mathbf{x}}_a) &= \sum_{k=1}^K \tilde{\pi}_k p(\mathbf{x}_a = \hat{\mathbf{x}}_a, \mathbf{x}_b | k)
 \end{aligned}$$

where we have set

$$\tilde{\pi}_k = \frac{\pi_k}{\sum_{j=1}^K \pi_j p(\mathbf{x}_a = \hat{\mathbf{x}}_a | j)}$$

Exercise 9.11

Exercise 9.12

$$\begin{aligned}
 \mathbb{E}[\mathbf{x}] &= \sum_{\mathbf{x}} \mathbf{x} \sum_{k=1}^K \pi_k p(\mathbf{x}|k) \\
 &= \sum_{k=1}^K \pi_k \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|k) \\
 &= \sum_{k=1}^K \pi_k \mathbb{E}[\mathbf{x}|k] \\
 \mathbb{E}[\mathbf{x}] &= \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k
 \end{aligned}$$

$$\begin{aligned}
 \text{Cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top] \\
 &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^\top] \\
 &= \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^\top \sum_{k=1}^K \pi_k p(\mathbf{x}|k) - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^\top] \\
 &= \sum_{k=1}^K \pi_k \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^\top p(\mathbf{x}|k) - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^\top] \\
 \text{Cov}[\mathbf{x}] &= \sum_{k=1}^K \pi_k \{\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}^\top]
 \end{aligned}$$

Exercise 9.13

Exercise 9.14

Exercise 9.15

Exercise 9.16

Exercise 9.17

Exercise 9.18

$$\tilde{\mathcal{Q}}(\mu, \pi, \mu^{\text{old}}, \pi^{\text{old}}) = \mathcal{Q}(\mu, \pi, \mu^{\text{old}}, \pi^{\text{old}}) + \sum_{k=1}^K \sum_{i=1}^D \log p(\mu_{k,i} | a_k, b_k) + \log p(\pi | \alpha)$$

Exercise 9.19

Exercise 9.20

Exercise 9.21

Exercise 9.22

Exercise 9.23

Exercise 9.24

Exercise 9.25

Exercise 9.26

Exercise 9.27

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](#).

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^M w_jx^j.$$

(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](#).

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2.$$

(1.10) Page 14. Referenced in Exercises: [1.3](#), [1.39](#), [8.3](#), [8.4](#), [8.11](#), [8.21](#), [8.22](#), [8.25](#), [9.3](#).

$$p(X) = \sum_Y p(X, Y).$$

(1.12) Page 15. Referenced in Exercises: [1.3](#), [3.24](#), [8.3](#), [8.4](#), [8.10](#), [8.11](#), [8.16](#), [9.4](#).

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}.$$

(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](#), [6.18](#), [9.4](#).

$$\int_{-\infty}^{\infty} p(x) dx = 1.$$

(1.27) Page 18. Referenced in Exercises: [1.4](#), [1.32](#).

$$\begin{aligned} p_y(y) &= p_x(y) \left| \frac{dx}{dy} \right| \\ &= p_x(g(y)) |g'(y)|. \end{aligned}$$

(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](#), [5.27](#), [8.1](#).

$$\int p(\mathbf{x}) d\mathbf{x} = 1.$$

(1.31) Page 19. Referenced in Exercises: [1.37](#), [2.8](#), [3.23](#), [8.8](#), [8.10](#), [8.27](#).

$$p(x) = \int p(x, y) dy.$$

- (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, 8.11, 8.18, 8.26, 9.3.

$$p(x, y) = p(y|x)p(x).$$

- (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, 6.18.

$$\mathbb{E}[f] = \int p(x)f(x) dx.$$

- (1.35) Page 19. Referenced in Exercises: 4.23, 5.17.

$$\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^N f(x_n).$$

- (1.37) Page 20. Referenced in Exercises: 1.25, 1.26, 2.8, 5.27, 5.37, 6.18.

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x).$$

- (1.38) Page 20. Referenced in Exercises: 1.5, 1.8, 1.13, 2.1, 2.2, 2.8, 2.27.

$$\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2].$$

- (1.39) Page 20. Referenced in Exercises: 1.5, 1.8, 2.4, 2.6, 2.10, 2.12, 2.42, 6.18.

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2.$$

- (1.41) Page 20. Referenced in Exercises: 1.6, 2.10.

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \end{aligned}$$

- (1.42) Page 20. Referenced in Exercises: 2.49, 2.58.

$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y} - \mathbb{E}[\mathbf{y}]\}] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\mathbf{xy}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]. \end{aligned}$$

- (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, 5.29, 5.30, 5.31, 5.39.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

- (1.48) Page 25. Referenced in Exercises: 1.8, 4.26.

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1.$$

- (1.49) Page 25. Referenced in Exercises: 1.8, 6.18.

$$\int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) dx = \mu.$$

(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](#).

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n.$$

(1.56) Page 27. Referenced in Exercises: [1.11](#).

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2.$$

(1.78) Page 39. Referenced in Exercises: [1.21](#).

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

(1.91) Page 48. Referenced in Exercises: [1.27](#).

$$\mathbb{E}[L_q(y(\mathbf{X}), T)] = \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) \, d\mathbf{x} dt.$$

(1.98) Page 51. Referenced in Exercises: [1.28](#), [1.29](#), [2.1](#), [2.2](#).

$$H[p] = - \sum_i p(x_i) \ln p(x_i).$$

(1.104) Page 53. Referenced in Exercises: [1.32](#), [1.35](#), [1.37](#), [1.41](#), [2.15](#).

$$H[\mathbf{x}] = - \int p(\mathbf{x}) \ln p(\mathbf{x}) \, d\mathbf{x}.$$

(1.111) Page 54. Referenced in Exercises: [1.33](#), [1.37](#), [1.41](#).

$$H[\mathbf{y}|\mathbf{x}] = - \iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} d\mathbf{x}.$$

(1.112) Page 55. Referenced in Exercises: [1.31](#).

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}].$$

(1.113) Page 56. Referenced in Exercises: [1.30](#), [1.41](#), [2.13](#).

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

(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](#).

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &= \text{KL}(p(\mathbf{x}, \mathbf{y}) || p(\mathbf{x})p(\mathbf{y})) \\ &= - \iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right) d\mathbf{x}d\mathbf{y}. \end{aligned}$$

(1.121) Page 57. Referenced in Exercises: [1.31](#).

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}].$$

(1.124) Page 59. Referenced in Exercises: [1.7](#), [1.8](#), [1.18](#), [4.21](#).

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx.$$

(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](#).

$$\sum_{i_1=1}^D \sum_{i_2=1}^D \cdots \sum_{i_M=1}^D w_{i_1, i_2, \dots, i_M} x_{i_1} x_{i_2} \cdots x_{i_M}.$$

(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](#).

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du.$$

(1.142) Page 62. Referenced in Exercises: 1.18.

$$\prod_{i=1}^D \int_{-\infty}^{\infty} e^{-x_i^2} dx_i = S_D \int_0^{\infty} e^{-r^2} r^{D-1} dr.$$

(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, 5.17.

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

(2.2) Page 69. Referenced in Exercises: 2.1, 2.7.

$$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

(2.13) Page 71. Referenced in Exercises: 2.6, 2.7, 2.56.

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

(2.38) Page 76. Referenced in Exercises: 2.10.

$$\text{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k-1}$$

(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, 5.35, 5.36.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}.$$

(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](#), [6.8](#), [6.14](#).

$$\Sigma = \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^\top.$$

(2.49) Page 80. Referenced in Exercises: [3.2](#), [3.21](#).

$$\Sigma^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top.$$

(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](#).

$$|\Sigma| = \prod_{j=1}^D \lambda_j^{1/2}.$$

(2.59) Page 82. Referenced in Exercises: [2.13](#), [2.35](#), [2.49](#), [4.24](#), [5.27](#), [5.37](#).

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

(2.62) Page 83. Referenced in Exercises: [2.13](#), [5.27](#), [5.37](#).

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \Sigma.$$

(2.64) Page 83. Referenced in Exercises: [2.13](#), [2.15](#), [2.35](#), [2.49](#), [4.24](#).

$$\text{cov}[\mathbf{x}] = \Sigma.$$

(2.76) Page 87. Referenced in Exercises: [2.24](#), [4.2](#).

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{pmatrix}.$$

(2.77) Page 87. Referenced in Exercises: [4.2](#).

$$\mathbf{M} = (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}.$$

(2.81) Page 87. Referenced in Exercises: [2.25](#), [2.28](#), [6.20](#), [6.22](#).

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b).$$

(2.82) Page 87. Referenced in Exercises: [2.25](#), [2.28](#), [6.20](#), [6.22](#).

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.$$

(2.92) Page 89. Referenced in Exercises: [2.25](#), [2.28](#), [4.24](#), [6.22](#).

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

(2.93) Page 89. Referenced in Exercises: [2.25](#), [2.28](#), [4.24](#), [6.22](#).

$$\text{cov}[\mathbf{x}_a] = \Sigma_{aa}.$$

(2.104) Page 92. Referenced in Exercises: [2.29](#), [2.30](#).

$$\mathbf{R} = \begin{pmatrix} \Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{pmatrix}.$$

(2.105) Page 92. Referenced in Exercises: [2.28](#), [2.29](#), [2.30](#).

$$\text{cov}[\mathbf{z}] = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1} \mathbf{A}^\top \\ \mathbf{A} \Lambda^{-1} & \mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top \end{pmatrix}$$

(2.107) Page 92. Referenced in Exercises: [2.28](#), [2.30](#).

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \Lambda \mu - \mathbf{A}^\top \mathbf{L} \mathbf{b} \\ \mathbf{L} \mathbf{b} \end{pmatrix}$$

(2.108) Page 92. Referenced in Exercises: [2.28](#), [2.30](#).

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \mu \\ \mathbf{A} \mu + \mathbf{b} \end{pmatrix}$$

(2.113) Page 93. Referenced in Exercises: [2.31](#), [3.9](#), [3.10](#), [3.13](#), [3.16](#), [5.38](#).

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \mu, \Lambda^{-1}).$$

(2.114) Page 93. Referenced in Exercises: [2.31](#), [3.9](#), [3.10](#), [3.13](#), [3.16](#), [5.38](#).

$$p(\mathbf{y} | \mathbf{x}) = \mathcal{N}(\mathbf{y} | \mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}).$$

(2.115) Page 93. Referenced in Exercises: [2.31](#), [3.10](#), [3.16](#), [5.38](#), [6.26](#), [7.11](#), [7.14](#).

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\mu + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^\top).$$

(2.116) Page 93. Referenced in Exercises: [3.9](#), [3.13](#), [7.9](#).

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\mathbf{x} | \Sigma \{\mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b}) + \Lambda \mu\}, \Sigma).$$

(2.118) Page 93. Referenced in Exercises: [2.34](#), [4.23](#), [5.2](#), [6.14](#), [7.13](#).

$$\ln p(\mathbf{X} | \mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \mu)^\top \Sigma^{-1} (\mathbf{x}_n - \mu).$$

(2.121) Page 93. Referenced in Exercises: [2.34](#), [2.35](#).

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

(2.135) Page 96. Referenced in Exercises: [2.36](#), [2.37](#), [2.39](#), [9.2](#).

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} [-\ln p(x_N | \theta^{(N-1)})].$$

(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](#), [7.13](#).

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

(2.155) Page 102. Referenced in Exercises: [2.45](#).

$$\mathcal{W}(\Lambda|\mathbf{W}, \nu) = B|\Lambda|^{(\nu-D-1)/2} \exp \left(-\frac{1}{2} \text{Tr}(\mathbf{W}^{-1}\Lambda) \right).$$

(2.168) Page 106. Referenced in Exercises: [2.55](#).

$$\bar{x}_1 = \bar{r} \cos \bar{\theta} = \frac{1}{N} \sum_{n=1}^N \cos \theta_n \quad \bar{x}_2 = \bar{r} \sin \bar{\theta} = \frac{1}{N} \sum_{n=1}^N \sin \theta_n.$$

(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](#).

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

(2.185) Page 109. Referenced in Exercises: [2.55](#).

$$A(m_{\text{ML}}) = \frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}}).$$

(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](#).

$$g(\boldsymbol{\eta}) \int h(\mathbf{x})g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} d\mathbf{x} = 1.$$

(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](#).

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

(2.249) Page 123. Referenced in Exercises: [7.1](#).

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right).$$

(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](#).

$$p_M(x_1, \dots, x_{M-1}) = C_M \prod_{k=1}^{M-1} x_k^{\alpha_k-1} \left(1 - \sum_{k=1}^{M-1} x_k\right)^{\alpha_M-1}.$$

(2.277) Page 130. Referenced in Exercises: [2.11](#).

$$\psi(a) = \frac{d}{da} \ln \Gamma(a).$$

(2.278) Page 130. Referenced in Exercises: [2.12](#).

$$U(x|a, b) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

(2.288) Page 132. Referenced in Exercises: [2.25](#).

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,b} & \Sigma_{a,c} \\ \Sigma_{b,a} & \Sigma_{b,b} & \Sigma_{b,c} \\ \Sigma_{c,a} & \Sigma_{c,b} & \Sigma_{c,c} \end{pmatrix}.$$

(2.289) Page 132. Referenced in Exercises: [2.26](#), [3.11](#), [5.21](#), [6.21](#), [6.26](#), [7.10](#), [7.12](#).

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}$$

(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](#), [7.6](#).

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

(3.11) Page 141. Referenced in Exercises: [3.17](#).

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^N \ln \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}). \end{aligned}$$

(3.52) Page 153. Referenced in Exercises: [3.17](#), [5.25](#), [5.38](#).

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}).$$

(3.54) Page 153. Referenced in Exercises: [3.14](#), [6.21](#).

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^\top \Phi.$$

(3.62) Page 159. Referenced in Exercises: [3.14](#).

$$k(\mathbf{x}, \mathbf{x}') = \beta \phi(\mathbf{x})^\top \mathbf{S}_N \phi(\mathbf{x}').$$

(3.77) Page 166. Referenced in Exercises: [3.17](#).

$$p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha) d\mathbf{w}.$$

(3.78) Page 166. Referenced in Exercises: [3.17](#), [3.19](#).

$$p(\mathbf{t}|\alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}.$$

(3.79) Page 167. Referenced in Exercises: [3.17](#), [3.18](#).

$$\begin{aligned} E(\mathbf{w}) &= \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\ &= \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \end{aligned}$$

(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](#).

$$\mathbf{A} = \alpha \mathbf{I} + \beta \Phi^\top \Phi.$$

(3.82) Page 167. Referenced in Exercises: [3.15](#), [3.16](#), [3.18](#), [3.20](#), [3.21](#), [3.22](#).

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N.$$

(3.84) Page 167. Referenced in Exercises: [3.16](#), [3.18](#), [6.21](#).

$$\mathbf{m}_N = \beta \mathbf{A}^{-1} \Phi^\top \mathbf{t}.$$

(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](#).

$$(\beta \Phi^\top \Phi) \mathbf{u}_i = \lambda_i \mathbf{u}_i.$$

(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](#).

$$\frac{d}{d\alpha} \ln |\mathbf{A}| = \text{Tr}\left(\mathbf{A}^{-1} \frac{d}{d\alpha} \mathbf{A}\right).$$

(4.15) Page 185. Referenced in Exercises: [4.2](#).

$$E_D(\tilde{\mathbf{W}}) = \frac{1}{2} \text{Tr}\{(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \tilde{\mathbf{T}})^\top (\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \tilde{\mathbf{T}})\}.$$

(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](#).

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^\top \tilde{\mathbf{x}} = \mathbf{T}^\top (\tilde{\mathbf{X}}^\dagger)^\top \tilde{\mathbf{x}}.$$

(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](#).

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n.$$

(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.

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}.$$

(4.27) Page 189. Referenced in Exercises: 4.5, 4.6.

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top.$$

(4.28) Page 189. Referenced in Exercises: 4.5, 4.6.

$$\mathbf{S}_W = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top.$$

(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.

$$\left(\mathbf{S}_W + \frac{N_1 N_2}{N} \mathbf{S}_B \right) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2).$$

(4.53) Page 193. Referenced in Exercises: 6.2.

$$f(a) = \begin{cases} +1, & a \geq 0, \\ -1, & a < 0. \end{cases}$$

(4.55) Page 194. Referenced in Exercises: 6.2.

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_P(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \phi_n t_n.$$

(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, 5.19, 6.27, 7.18, 7.19.

$$\frac{d\sigma}{da} = \sigma(1 - \sigma).$$

(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, 5.19.

$$E(\mathbf{w}) = - \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}.$$

(4.91) Page 206. Referenced in Exercises: 4.13.

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n.$$

(4.92) Page 207. Referenced in Exercises: 6.25.

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1} \nabla E(\mathbf{w}).$$

(4.97) Page 207. Referenced in Exercises: 4.15.

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n (1 - y_n) \phi_n \phi_n^\top = \Phi^\top \mathbf{R} \Phi.$$

(4.104) Page 209. Referenced in Exercises: 4.17, 4.20, 5.32, 5.34.

$$p(\mathcal{C}_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}.$$

(4.105) Page 209. Referenced in Exercises: 4.17, 4.18, 5.32.

$$a_k = \mathbf{w}_k^\top \phi$$

(4.106) Page 209. Referenced in Exercises: 4.17, 5.32, 5.34.

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j).$$

(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](#).

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta.$$

(4.115) Page 211. Referenced in Exercises: [4.21](#).

$$\text{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp(-\theta^2) d\theta.$$

(4.116) Page 211. Referenced in Exercises: [4.21](#).

$$\Phi(a) = \frac{1}{2} \left\{ 1 + \text{erf}\left(\frac{a}{\sqrt{2}}\right) \right\}.$$

(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](#), [5.39](#), [6.27](#).

$$\begin{aligned} Z &= \int f(\mathbf{z}) d\mathbf{z} \\ &\approx f(\mathbf{z}_0) \int \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^\top \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\} \\ &= f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}. \end{aligned}$$

(4.137) Page 216. Referenced in Exercises: [4.22](#), [4.23](#).

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\theta_{\text{MAP}}) + \ln p(\theta_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|.$$

(4.138) Page 217. Referenced in Exercises: [4.22](#), [4.23](#).

$$\mathbf{A} = -\nabla \nabla \ln p(\mathcal{D}|\theta_{\text{MAP}}) p(\theta_{\text{MAP}}) = -\nabla \nabla \ln p(\theta_{\text{MAP}}|\mathcal{D}).$$

(4.139) Page 217. Referenced in Exercises: [4.23](#).

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\theta_{\text{MAP}}) - \frac{1}{2} M \ln N.$$

(4.143) Page 218. Referenced in Exercises: 4.24.

$$\mathbf{S}_N^{-1} = -\nabla \nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^N y_n(1-y_n)\phi_n\phi_n^\top.$$

(4.144) Page 218. Referenced in Exercises: 4.24.

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{S}_N).$$

(4.151) Page 219. Referenced in Exercises: 4.24.

$$p(\mathcal{C}_1|\mathbf{t}) = \int \sigma(a)p(a) da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2) da.$$

(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.

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^N t_{nk}(\phi_n - \mu_k)(\phi_n - \mu_k)^\top.$$

(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, 5.16.

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2.$$

(5.16) Page 234. Referenced in Exercises: 5.2.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1} \mathbf{I}).$$

(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, 5.40.

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K \{t_{nk} \ln y_{nk} + (1-t_{nk}) \ln(1-y_{nk})\}.$$

(5.24) Page 235. Referenced in Exercises: 5.7, 5.40.

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}(\mathbf{x}_n, \mathbf{w}).$$

(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.

$$E(\mathbf{w}) \approx E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}^*).$$

(5.35) Page 238. Referenced in Exercises: 5.11, 5.12.

$$\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i.$$

(5.53) Page 243. Referenced in Exercises: 5.27.

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_{ji}.$$

(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.

$$h'(a) = 1 - h(a)^2.$$

(5.65) Page 246. Referenced in Exercises: 5.18.

$$\delta_k = y_k - t_k.$$

(5.67) Page 246. Referenced in Exercises: 5.18.

$$\frac{\partial E_n}{\partial w_{ji}^{(1)}} = \delta_j x_i, \quad \frac{\partial E_n}{\partial w_{kj}^{(2)}} = \delta_k z_j.$$

(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, 5.26.

$$J_{ki} = \frac{\partial y_k}{\partial x_i}.$$

(5.84) Page 251. Referenced in Exercises: 5.17.

$$\mathbf{H} \approx \sum_{n=1}^N \mathbf{b}_n \mathbf{b}_n^\top.$$

(5.86) Page 252. Referenced in Exercises: 5.21.

$$\mathbf{H}_N = \sum_{n=1}^N \mathbf{b}_n \mathbf{b}_n^\top.$$

(5.92) Page 253. Referenced in Exercises: 5.22, 5.23.

$$\delta_k = \frac{\partial E_n}{\partial a_k}, \quad M_{kk'} = \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}}.$$

(5.93) Page 253. Referenced in Exercises: 5.22, 5.23.

$$\frac{\partial^2 E_n}{\partial w_{kj}^{(2)} \partial w_{k'j'}^{(2)}} = z_j z_{j'} M_{kk'}.$$

(5.94) Page 254. Referenced in Exercises: 5.22, 5.23.

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} = x_i x_{i'} h''(a_{j'}) I_{jj'} \sum_k w_{kj'}^{(2)} \delta_k + x_i x_{i'} h'(a_{j'}) h'(a_j) \sum_k \sum_{k'} w_{k'j'}^{(2)} w_{kj}^{(2)} M_{kk'}.$$

(5.95) Page 254. Referenced in Exercises: 5.22, 5.23.

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

(5.113) Page 258. Referenced in Exercises: 5.24.

$$z_j = h \left(\sum_i w_{ji} x_i + w_{j0} \right).$$

(5.114) Page 258. Referenced in Exercises: 5.24.

$$y_k = \sum_j w_{kj} z_j + w_{k0}.$$

(5.115) Page 258. Referenced in Exercises: 5.24.

$$x_i \rightarrow \tilde{x}_i = ax_i + b.$$

(5.116) Page 258. Referenced in Exercises: 5.24.

$$w_{ji} \rightarrow \tilde{w}_{ji} = \frac{1}{a} w_{ji}.$$

(5.117) Page 258. Referenced in Exercises: 5.24.

$$w_{j0} \rightarrow \tilde{w}_{j0} = w_{j0} - \frac{b}{a} \sum_i w_{ji}.$$

(5.118) Page 258. Referenced in Exercises: 5.24.

$$y_k \rightarrow \tilde{y}_k = cy_k + d.$$

(5.119) Page 258. Referenced in Exercises: 5.24.

$$w_{kj} \rightarrow \tilde{w}_{kj} = cw_{kj}.$$

(5.120) Page 258. Referenced in Exercises: 5.24.

$$w_{k0} \rightarrow \tilde{w}_{k0} = cw_{k0} + d.$$

(5.126) Page 264. Referenced in Exercises: 5.26.

$$\frac{\partial y_k}{\partial \xi} \Bigg|_{\xi=0} = \sum_{i=1}^D \frac{\partial y_k}{\partial x_i} \frac{\partial x_i}{\partial \xi} \Bigg|_{\xi=0} = \sum_{i=1}^D J_{ki} \tau_i.$$

(5.127) Page 264. Referenced in Exercises: 5.26, 5.27, 5.29, 5.30, 5.31, 5.32.

$$\tilde{E} = E + \lambda \Omega.$$

(5.128) Page 264. Referenced in Exercises: 5.26.

$$\Omega = \frac{1}{2} \sum_n \sum_k \left(\frac{\partial y_{nk}}{\partial \xi} \Bigg|_{\xi=0} \right)^2 = \frac{1}{2} \sum_n \sum_k \left(\sum_{i=1}^D J_{nki} \tau_{ni} \right)^2.$$

(5.130) Page 266. Referenced in Exercises: 5.27.

$$\tilde{E} = \frac{1}{2} \int \int \int \{y(\mathbf{s}(\mathbf{x}, \xi)) - t\}^2 p(t|\mathbf{x}) p(\mathbf{x}) p(\xi) \, d\mathbf{x} dt d\xi.$$

(5.135) Page 267. Referenced in Exercises: 5.27.

$$\Omega = \frac{1}{2} \int ||\nabla \mathbf{y}(\mathbf{x})||^2 p(\mathbf{x}) \, d\mathbf{x}.$$

(5.138) Page 270. Referenced in Exercises: 5.29, 5.30, 5.31, 5.32.

$$\Omega(\mathbf{w}) = - \sum_i \ln \left(\sum_{j=1}^M \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right).$$

(5.140) Page 270. Referenced in Exercises: 5.29, 5.30, 5.31.

$$\gamma_j(w) = \frac{\pi_j \mathcal{N}(w | \mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w | \mu_k, \sigma_k^2)}.$$

(5.141) Page 270. Referenced in Exercises: 5.29.

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \lambda \sum_j \gamma_j(w_i) \frac{(w_i - \mu_j)}{\sigma_j^2}.$$

(5.142) Page 271. Referenced in Exercises: 5.30.

$$\frac{\partial \tilde{E}}{\partial \mu_j} = \lambda \sum_i \gamma_j(w_i) \frac{(\mu_j - w_i)}{\sigma_j^2}$$

(5.143) Page 271. Referenced in Exercises: 5.31.

$$\frac{\partial \tilde{E}}{\partial \sigma_j} = \lambda \sum_i \gamma_j(w_i) \left(\frac{1}{\sigma_j} - \frac{(\mu_j - w_i)^2}{\sigma_j^3} \right)$$

(5.146) Page 271. Referenced in Exercises: 5.32.

$$\pi_j = \frac{\exp(\eta_j)}{\sum_{k=1}^M \exp(\eta_k)}.$$

(5.147) Page 272. Referenced in Exercises: 5.32.

$$\frac{\partial \tilde{E}}{\partial \eta_j} = \lambda \sum_i \{\pi_j - \gamma_j(w_i)\}.$$

(5.148) Page 273. Referenced in Exercises: 5.37.

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{x}) \mathcal{N}(\mathbf{t}|\mu_k(\mathbf{x}), \sigma_k^2(\mathbf{x})).$$

(5.151) Page 274. Referenced in Exercises: 5.36.

$$\sigma_k(\mathbf{x}) = \exp(a_k^\sigma).$$

(5.153) Page 275. Referenced in Exercises: 5.34, 5.35, 5.36.

$$E(\mathbf{w}) = - \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k(\mathbf{x}_n, \mathbf{w}) \mathcal{N}(\mathbf{t}_n|\mu_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w})) \right\}.$$

(5.154) Page 275. Referenced in Exercises: 5.34, 5.35, 5.36.

$$\gamma_{nk} = \gamma_k(\mathbf{t}_n|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}_{nk}}{\sum_{l=1}^K \pi_l \mathcal{N}_{nl}}.$$

(5.155) Page 275. Referenced in Exercises: 5.34.

$$\frac{\partial E_n}{\partial a_k^\pi} = \pi_k - \gamma_{nk}.$$

(5.156) Page 275. Referenced in Exercises: 5.35.

$$\frac{\partial E_n}{\partial a_{kl}^\mu} = \gamma_{nk} \left\{ \frac{\mu_{kl} - t_{nl}}{\sigma_k^2} \right\}.$$

(5.157) Page 275. Referenced in Exercises: 5.36.

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

(5.158) Page 276. Referenced in Exercises: 5.37.

$$\mathbb{E}[\mathbf{t}|\mathbf{x}] = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \sum_{k=1}^K \pi_k(\mathbf{x}) \mu_k(\mathbf{x}).$$

(5.160) Page 277. Referenced in Exercises: 5.37.

$$\begin{aligned} s^2(\mathbf{x}) &= \mathbb{E}[||\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]||^2|\mathbf{x}] \\ &= \sum_{k=1}^K \pi_k(\mathbf{x}) \left\{ \sigma^2(\mathbf{x}) + \left\| \mu_k - \sum_{l=1}^K \pi_l \mu_l(\mathbf{x}) \right\|^2 \right\}. \end{aligned}$$

(5.167) Page 279. Referenced in Exercises: 5.38.

$$q(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}).$$

(5.171) Page 279. Referenced in Exercises: 5.38.

$$p(t|\mathbf{x}, \mathbf{w}, \beta) \approx \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}_{\text{MAP}}) + \mathbf{g}^\top(\mathbf{w} - \mathbf{w}_{\text{MAP}}), \beta^{-1}).$$

(5.176) Page 280. Referenced in Exercises: 5.39.

$$E(\mathbf{w}_{\text{MAP}}) = \frac{\beta}{2} \sum_{n=1}^N \{y(\mathbf{x}_n, \mathbf{w}_{\text{MAP}}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}_{\text{MAP}}^\top \mathbf{w}_{\text{MAP}}.$$

(5.183) Page 282. Referenced in Exercises: 5.41.

$$\ln p(\mathcal{D}|\alpha) \approx -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2} \ln |\mathbf{A}| + \frac{W}{2} \ln \alpha.$$

(5.184) Page 282. Referenced in Exercises: 5.41.

$$E(\mathbf{w}_{\text{MAP}}) = - \sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} + \frac{\alpha}{2} \mathbf{w}_{\text{MAP}}^\top \mathbf{w}_{\text{MAP}}.$$

(5.193) Page 286. Referenced in Exercises: 5.17, 5.27.

$$E = \frac{1}{2} \int \int \{y(\mathbf{x}, \mathbf{w}) - t\}^2 p(\mathbf{x}, \mathbf{t}) \, d\mathbf{x} dt.$$

(5.195) Page 287. Referenced in Exercises: 5.25.

$$E = E_0 + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*).$$

(5.196) Page 287. Referenced in Exercises: 5.25.

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho \nabla E.$$

(5.197) Page 287. Referenced in Exercises: 5.25.

$$w_j^{(\tau)} = \{1 - (1 - \rho \eta_j)^\tau\} w_j^*.$$

(5.198) Page 287. Referenced in Exercises: 5.25.

$$\mathbf{H} \mathbf{u}_j = \eta_j \mathbf{u}_j.$$

(5.199) Page 288. Referenced in Exercises: 5.25.

$$w_j^{(\tau)} \approx w_j^* \quad \text{when } \eta_j \gg (\rho \tau)^{-1}.$$

(5.200) Page 288. Referenced in Exercises: 5.25.

$$|w_j^{(\tau)}| \ll |w_j^*| \quad \text{when } \eta_j \ll (\rho \tau)^{-1}.$$

(5.201) Page 288. Referenced in Exercises: 5.26.

$$\Omega_n = \frac{1}{2} \sum_k (\mathcal{G} y_k)^2.$$

(5.202) Page 288. Referenced in Exercises: 5.26.

$$\mathcal{G} = \sum_i \tau_i \frac{\partial}{\partial x_i}.$$

(5.204) Page 288. Referenced in Exercises: [5.26](#).

$$\alpha_j = h'(a_j)\beta_j, \quad \beta_j = \sum_i w_{ji}\alpha_i.$$

(5.205) Page 288. Referenced in Exercises: [5.26](#).

$$\alpha_j = \mathcal{G}z_j, \quad \beta_j = \mathcal{G}a_j.$$

(5.206) Page 288. Referenced in Exercises: [5.26](#).

$$\frac{\partial \Omega_n}{\partial w_{rs}} = \sum_k \alpha_k \{\phi_{kr}z_s + \delta_{kr}\alpha_s\}.$$

(5.207) Page 288. Referenced in Exercises: [5.26](#).

$$\delta_{kr} = \frac{\partial y_k}{\partial a_r}, \quad \phi_{kr} = \mathcal{G}\delta_{kr}.$$

(5.208) Page 289. Referenced in Exercises: [5.32](#).

$$\frac{\partial \pi_k}{\partial \eta_j} = \delta_{jk}\pi_j - \pi_j\pi_k.$$

(6.1) Page 292. Referenced in Exercises: [6.2](#), [6.3](#), [6.5](#), [6.6](#), [6.7](#), [6.8](#), [6.11](#), [7.4](#), [7.7](#).

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}').$$

(6.5) Page 293. Referenced in Exercises: [6.1](#).

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^\top \Phi \Phi^\top \Phi \Phi^\top \mathbf{a} - \mathbf{a}^\top \Phi \Phi^\top \mathbf{t} + \frac{1}{2}\mathbf{t}^\top \mathbf{t} + \frac{\lambda}{2}\mathbf{a}^\top \Phi \Phi^\top \mathbf{a}.$$

(6.6) Page 293. Referenced in Exercises: [6.15](#), [6.23](#).

$$K_{nm} = \phi(\mathbf{x}_n)^\top \phi(\mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m).$$

(6.8) Page 293. Referenced in Exercises: [6.10](#).

$$\mathbf{a} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1}.$$

(6.9) Page 294. Referenced in Exercises: [6.10](#).

$$y(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) = \mathbf{a}^\top \Phi \phi(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \mathbf{t}.$$

(6.13) Page 296. Referenced in Exercises: [6.5](#), [6.6](#).

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}').$$

(6.14) Page 296. Referenced in Exercises: [6.3](#), [6.5](#).

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}').$$

(6.15) Page 296. Referenced in Exercises: [6.3](#), [6.5](#), [6.6](#).

$$k(\mathbf{x}, \mathbf{x}') = q(k_1(\mathbf{x}, \mathbf{x}')).$$

(6.16) Page 296. Referenced in Exercises: [6.3](#), [6.6](#).

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}')).$$

(6.17) Page 296. Referenced in Exercises: [6.6](#), [6.7](#), [6.9](#).

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}').$$

(6.18) Page 296. Referenced in Exercises: [6.6](#), [6.7](#), [6.9](#), [6.11](#).

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}').$$

(6.19) Page 296. Referenced in Exercises: [6.8](#), [6.9](#).

$$k(\mathbf{x}, \mathbf{x}') = k_3(\phi(\mathbf{x}), \phi(\mathbf{x}')).$$

(6.20) Page 296. Referenced in Exercises: [6.8](#).

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A} \mathbf{x}'.$$

(6.21) Page 296. Referenced in Exercises: [6.9](#).

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b).$$

(6.22) Page 296. Referenced in Exercises: [6.9](#).

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b).$$

(6.23) Page 296. Referenced in Exercises: [6.3](#), [6.11](#).

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2).$$

(6.25) Page 297. Referenced in Exercises: [6.11](#).

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\mathbf{x}^\top \mathbf{x}/2\sigma^2) \exp(\mathbf{x}^\top \mathbf{x}/\sigma^2) \exp(-(\mathbf{x}')^\top \mathbf{x}'/2\sigma^2).$$

(6.27) Page 297. Referenced in Exercises: [6.12](#).

$$k(A_1, A_2) = 2^{|A_1 \cap A_2|}.$$

(6.32) Page 298. Referenced in Exercises: [6.13](#), [6.14](#).

$$\mathbf{g}(\theta, \mathbf{x}) = \nabla_\theta \ln \ln p(\mathbf{x} | \theta).$$

(6.33) Page 298. Referenced in Exercises: [6.13](#), [6.14](#).

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{g}(\theta, \mathbf{x})^\top \mathbf{F}^{-1} \mathbf{g}(\theta, \mathbf{x}).$$

(6.34) Page 298. Referenced in Exercises: [6.13](#), [6.14](#).

$$\mathbf{F} = \mathbb{E}_{\mathbf{x}}[\mathbf{g}(\theta, \mathbf{x}) \mathbf{g}(\theta, \mathbf{x})^\top].$$

(6.39) Page 300. Referenced in Exercises: [6.17](#).

$$E = \frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{x}_n + \xi) - t_n\}^2 \nu(\xi) d\xi.$$

(6.40) Page 300. Referenced in Exercises: [6.17](#).

$$y(\mathbf{x}) = \sum_{n=1}^N t_n h(\mathbf{x} - \mathbf{x}_n).$$

(6.41) Page 300. Referenced in Exercises: 6.17.

$$h(\mathbf{x} - \mathbf{x}_n) = \frac{\nu(\mathbf{x} - \mathbf{x}_n)}{\sum_{n=1}^N \nu(\mathbf{x} - \mathbf{x}_n)}.$$

(6.45) Page 302. Referenced in Exercises: 6.19.

$$y(\mathbf{x}) = \frac{\sum_n g(\mathbf{x} - \mathbf{x}_n)t_n}{\sum_m g(\mathbf{x} - \mathbf{x}_m)}.$$

(6.46) Page 302. Referenced in Exercises: 6.18, 6.19.

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

(6.48) Page 303. Referenced in Exercises: 6.18.

$$p(t|\mathbf{x}) = \frac{p(t, \mathbf{x})}{\int p(t, \mathbf{x}) dt} = \frac{\sum_n f(\mathbf{x} - \mathbf{x}_n, t - t_n)}{\sum_m \int f(\mathbf{x} - \mathbf{x}_m, t - t_m)}.$$

(6.62) Page 307. Referenced in Exercises: 6.21.

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}.$$

(6.66) Page 308. Referenced in Exercises: 6.20, 6.21, 6.22, 6.23.

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^\top \mathbf{C}^{-1} N \mathbf{t}.$$

(6.67) Page 308. Referenced in Exercises: 6.20, 6.21, 6.22, 6.23.

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^\top \mathbf{C}^{-1} N \mathbf{k}.$$

(6.78) Page 316. Referenced in Exercises: 6.26.

$$p(a_{N+1}|\mathbf{a}_N) = \mathcal{N}(a_{N+1}|\mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{a}_N, c - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k}).$$

(6.80) Page 316. Referenced in Exercises: 6.27.

$$\begin{aligned} \Psi(\mathbf{a}_N) &= \ln p(\mathbf{a}_N) + \ln p(\mathbf{t}_N|\mathbf{a}_N) \\ &= -\frac{1}{2}\mathbf{a}_N^\top \mathbf{C}_N^{-1} \mathbf{a}_N - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln|\mathbf{C}_N| + \mathbf{t}_N^\top \mathbf{a}_N \\ &\quad - \sum_{n=1}^N \ln(1 + e^{a_n}). \end{aligned}$$

(6.81) Page 316. Referenced in Exercises: 6.25.

$$\nabla \Psi(\mathbf{x}_N) = \mathbf{t}_N - \boldsymbol{\sigma}_N - \mathbf{C}_N^{-1} \mathbf{a}_N.$$

(6.82) Page 316. Referenced in Exercises: 6.25.

$$\nabla \nabla \Psi(\mathbf{x}_N) = -\mathbf{W}_N - \mathbf{C}_N^{-1}.$$

(6.83) Page 317. Referenced in Exercises: 6.25.

$$\mathbf{a}_N^{(\text{new})} = \mathbf{C}_N (\mathbf{I} + \mathbf{W}_N \mathbf{C}_N)^{-1} \{ \mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N \mathbf{a}_N \}.$$

(6.84) Page 317. Referenced in Exercises: 6.26.

$$\mathbf{a}_N^* = \mathbf{C}_N(\mathbf{t}_N - \boldsymbol{\sigma}_N).$$

(6.85) Page 317. Referenced in Exercises: 6.26.

$$\mathbf{H} = -\nabla \nabla \Psi(\mathbf{a}_N) = \mathbf{W}_N + \mathbf{C}_N^{-1}.$$

(6.86) Page 317. Referenced in Exercises: 6.26.

$$q(\mathbf{a}_N) = \mathcal{N}(\mathbf{a}_N | \mathbf{a}_N^*, \mathbf{H}^{-1}).$$

(6.90) Page 317. Referenced in Exercises: 6.27.

$$\ln p(\mathbf{t}_N | \boldsymbol{\theta}) = \ln \Psi(\mathbf{a}_N^*) - \frac{1}{2} \ln |\mathbf{W}_N + \mathbf{C}_N^{-1}| + \frac{N}{2} \ln(2\pi).$$

(6.91) Page 318. Referenced in Exercises: 6.27.

$$\begin{aligned} \frac{\partial \ln p(\mathbf{t}_N | \boldsymbol{\theta})}{\partial \theta_j} &= \frac{1}{2} \mathbf{a}_N^* \top \mathbf{C}_N^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j} \mathbf{C}_N^{-1} \mathbf{a}_N^* \\ &\quad - \frac{1}{2} \text{Tr} \left[(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1} \mathbf{W}_N \frac{\partial \mathbf{C}_N}{\partial \theta_j} \right]. \end{aligned}$$

(6.92) Page 318. Referenced in Exercises: 6.27.

$$\begin{aligned} &- \frac{1}{2} \sum_{n=1}^N \frac{\partial \ln |\mathbf{W}_N + \mathbf{C}_N^{-1}|}{\partial a_n^*} \frac{\partial a_n^*}{\partial \theta_j} \\ &= -\frac{1}{2} \sum_{n=1}^N [(\mathbf{I} + \mathbf{C}_N \mathbf{W}_N)^{-1} \mathbf{C}_N]_{nn} \sigma_n^* (1 - \sigma_n^*) (1 - 2\sigma_n^*) \frac{\partial a_n^*}{\partial \theta_j}. \end{aligned}$$

(6.94) Page 318. Referenced in Exercises: 6.27.

$$\frac{\partial a_n^*}{\partial \theta_j} = (\mathbf{I} + \mathbf{W}_N \mathbf{C}_N)^{-1} \frac{\partial \mathbf{C}_N}{\partial \theta_j} (\mathbf{t}_N - \boldsymbol{\sigma}_N).$$

(6.95) Page 321. Referenced in Exercises: 6.12.

$$\phi_U(\mathbf{A}) = \begin{cases} 1, & \text{if } U \subseteq A, \\ 0, & \text{otherwise.} \end{cases}$$

(6.97) Page 321. Referenced in Exercises: 6.16.

$$J(\mathbf{w}) = f(\mathbf{w}^\top \phi(\mathbf{x}_1), \dots, \mathbf{w}^\top \phi(\mathbf{x}_N)) + g(\mathbf{w}^\top \mathbf{w}).$$

(6.98) Page 321. Referenced in Exercises: 6.16.

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

(6.99) Page 322. Referenced in Exercises: 6.19.

$$E = \frac{1}{2} \sum_{n=1}^N \int \{y(\mathbf{x}_n - \boldsymbol{\xi}_n) - t_n\}^2 g(\boldsymbol{\xi}_n) d\boldsymbol{\xi}_n.$$

(7.1) Page 326. Referenced in Exercises: 7.6, 7.7.

$$y(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b.$$

(7.2) Page 327. Referenced in Exercises: 7.4.

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n (\mathbf{w}^\top \phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}.$$

(7.5) Page 328. Referenced in Exercises: 7.2, 7.4.

$$t_n (\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq 1, \quad n = 1, \dots, N.$$

(7.6) Page 328. Referenced in Exercises: 7.2.

$$\arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2.$$

(7.7) Page 328. Referenced in Exercises: 7.3.

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N a_n \{t_n (\mathbf{w}^\top \phi(\mathbf{x}_n) + b) - 1\}.$$

(7.8) Page 328. Referenced in Exercises: 7.4.

$$\mathbf{w} = \sum_{n=1}^N a_n t_n \phi(\mathbf{x}_n).$$

(7.10) Page 329. Referenced in Exercises: 7.5.

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m).$$

(7.12) Page 329. Referenced in Exercises: 7.4, 7.5.

$$\sum_{n=1}^N a_n t_n = 0.$$

(7.13) Page 329. Referenced in Exercises: 7.4, 7.5.

$$y(\mathbf{x}) = \sum_{n=1}^N a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b.$$

(7.16) Page 330. Referenced in Exercises: 7.4, 7.5.

$$a_n \{t_n y(\mathbf{x}_n) - 1\} = 0.$$

(7.46) Page 337. Referenced in Exercises: 7.6.

$$p(t|y) = \sigma(yt).$$

(7.47) Page 337. Referenced in Exercises: 7.6.

$$\sum_{n=1}^N E_{\text{LR}}(y_n t_n) + \lambda \|\mathbf{w}\|^2.$$

(7.48) Page 337. Referenced in Exercises: 7.6.

$$E_{\text{LR}}(yt) = \ln(1 + \exp(-yt)).$$

(7.56) Page 341. Referenced in Exercises: 7.7.

$$\begin{aligned} L = & C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 \\ & - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n). \end{aligned}$$

(7.61) Page 342. Referenced in Exercises: 7.7.

$$\begin{aligned} \tilde{L}(\mathbf{a}, \hat{\mathbf{a}}) = & -\frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N (a_n - \hat{a}_n)(a_m - \hat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m) \\ & - \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)t_n. \end{aligned}$$

(7.67) Page 342. Referenced in Exercises: 7.8.

$$(C - a_n)\xi_n = 0.$$

(7.68) Page 342. Referenced in Exercises: 7.8.

$$(C - \hat{a}_n)\hat{\xi}_n = 0.$$

(7.79) Page 346. Referenced in Exercises: 8.5.

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N p(t_n|\mathbf{x}_n, \mathbf{w}, \beta^{-1}).$$

(7.80) Page 346. Referenced in Exercises: 8.5.

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^M \mathcal{N}(w_i|0, \alpha_i^{-1}).$$

(7.82) Page 347. Referenced in Exercises: 7.12.

$$\mathbf{m} = \beta \Sigma \Phi^\top \mathbf{t}.$$

(7.83) Page 347. Referenced in Exercises: 7.12.

$$\Sigma = (\mathbf{A} + \beta \Phi^\top \Phi)^{-1}.$$

(7.85) Page 347. Referenced in Exercises: 7.12, 7.13, 7.15, 7.16.

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta) &= \ln \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}) \\ &= -\frac{1}{2} \{ N \ln(2\pi) + \ln |\mathbf{C}| + \mathbf{t}^\top \mathbf{C}^{-1} \mathbf{t} \}. \end{aligned}$$

(7.86) Page 347. Referenced in Exercises: 7.10, 7.12.

$$\mathbf{C} = \beta^{-1} \mathbf{I} + \Phi \mathbf{A}^{-1} \Phi^\top.$$

(7.87) Page 347. Referenced in Exercises: 7.12, 7.13.

$$\alpha_i^{\text{new}} = \frac{\gamma_i}{m_i^2}.$$

(7.88) Page 347. Referenced in Exercises: 7.12, 7.13.

$$(\beta^{\text{new}})^{-1} = \frac{\|\mathbf{t} - \Phi\mathbf{m}\|^2}{N - \sum_i \gamma_i}.$$

(7.89) Page 348. Referenced in Exercises: 7.12, 7.13, 7.19.

$$\gamma_i = 1 - \alpha_i \Sigma_{ii}.$$

(7.94) Page 351. Referenced in Exercises: 7.15.

$$|\mathbf{C}| = |\mathbf{C}_{-i}|(1 + \alpha_i^{-1} \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i).$$

(7.95) Page 351. Referenced in Exercises: 7.15.

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1} \varphi_i \varphi_i^\top \mathbf{C}_{-i}^{-1}}{\alpha_i + \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i}.$$

(7.96) Page 351. Referenced in Exercises: 7.15, 7.16.

$$L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i).$$

(7.97) Page 351. Referenced in Exercises: 7.15, 7.16.

$$\lambda(\alpha_i) = \frac{1}{2} \left[\ln \alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i} \right].$$

(7.98) Page 351. Referenced in Exercises: 7.15.

$$s_i = \varphi_i^\top \mathbf{C}_{-i}^{-1} \varphi_i.$$

(7.99) Page 351. Referenced in Exercises: 7.15.

$$q_i = \varphi_i^\top \mathbf{C}_{-i}^{-1} \mathbf{t}.$$

(7.101) Page 352. Referenced in Exercises: 7.16.

$$\alpha_i = \frac{s_i^2}{q_i^2 - s_i}.$$

(7.102) Page 353. Referenced in Exercises: 7.17.

$$Q_i = \varphi_i^\top \mathbf{C}^{-1} \mathbf{t}.$$

(7.103) Page 353. Referenced in Exercises: 7.17.

$$S_i = \varphi_i^\top \mathbf{C}^{-1} \varphi_i.$$

(7.106) Page 353. Referenced in Exercises: 7.17.

$$Q_i = \beta \varphi_i^\top \mathbf{t} - \beta^2 \varphi_i^\top \Phi \Sigma \Phi^\top \mathbf{t}.$$

(7.107) Page 353. Referenced in Exercises: 7.17.

$$S_i = \beta \varphi_i^\top \varphi_i - \beta^2 \varphi_i^\top \Phi \Sigma \Phi^\top \varphi_i.$$

(7.109) Page 354. Referenced in Exercises: 7.18.

$$\begin{aligned} \ln p(\mathbf{w}|\mathbf{w}, \boldsymbol{\alpha}) &= \ln\{p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})\} - \ln p(\mathbf{t}|\boldsymbol{\alpha}) \\ &= \sum_{n=1}^N \{t_n \ln y_n + (1-t_n) \ln(1-y_n)\} - \frac{1}{2} \mathbf{w}^\top \mathbf{A} \mathbf{w} + \text{const.} \end{aligned}$$

(7.112) Page 354. Referenced in Exercises: 7.19.

$$\mathbf{w}^* = \mathbf{A}^{-1} \Phi^\top (\mathbf{t} - \mathbf{y}).$$

(7.113) Page 354. Referenced in Exercises: 7.19.

$$\Sigma = (\Phi^\top \mathbf{B} \Phi + \mathbf{A})^{-1}.$$

(7.114) Page 355. Referenced in Exercises: 7.19.

$$\begin{aligned} p(\mathbf{t}|\boldsymbol{\alpha}) &= \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} \\ &\approx p(\mathbf{t}|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\Sigma|^{1/2}. \end{aligned}$$

(7.123) Page 357. Referenced in Exercises: 7.4, 7.5.

$$\frac{1}{\rho^2} = \sum_{n=1}^N a_n.$$

(7.124) Page 357. Referenced in Exercises: 7.5.

$$\frac{1}{\rho^2} = 2\tilde{L}(\tilde{\mathbf{a}}).$$

(7.125) Page 357. Referenced in Exercises: 7.5.

$$\frac{1}{\rho^2} = ||\mathbf{w}||^2.$$

(8.5) Page 362. Referenced in Exercises: 8.1.

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k|\text{pa}_k).$$

(8.11) Page 370. Referenced in Exercises: 8.7.

$$p(x_i|\text{pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j \in \text{pa}_i} w_{ij}x_j + b_i, v_i\right).$$

(8.15) Page 370. Referenced in Exercises: 8.7.

$$\mathbb{E}[x_i] = \sum_{j \in \text{pa}_i} w_{ij} \mathbb{E}[x_j] + b_i.$$

(8.16) Page 371. Referenced in Exercises: [8.7](#).

$$\begin{aligned}\text{cov}[x_i, x_j] &= \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] \\ &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\left\{\sum_{k \in \text{pa}_j} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\right\}\right] \\ &= \sum_{k \in \text{pa}_j} w_{jk}\text{cov}[x_i, x_k] + I_{ij}v_j.\end{aligned}$$

(8.42) Page 389. Referenced in Exercises: [8.13](#), [8.14](#).

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i.$$

(8.49) Page 395. Referenced in Exercises: [8.15](#).

$$p(\mathbf{x}) = \frac{1}{Z} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3) \dots \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \psi_{N-1,N}(x_{N-1}, x_N).$$

(8.54) Page 396. Referenced in Exercises: [8.19](#).

$$p(x_n) = \frac{1}{Z} \mu_\alpha(x_n) \mu_\beta(x_n).$$

(8.55) Page 397. Referenced in Exercises: [8.15](#), [8.16](#), [8.19](#).

$$\begin{aligned}\mu_\alpha(x_n) &= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \left[\sum_{x_{n-2}} \dots \right] \\ &= \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \mu_\alpha(x_{n-1}).\end{aligned}$$

(8.57) Page 397. Referenced in Exercises: [8.15](#), [8.16](#), [8.19](#).

$$\begin{aligned}\mu_\beta(x_n) &= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n) \left[\sum_{x_{n+2}} \dots \right] \\ &= \sum_{x_{n+1}} \psi_{n+1,n}(x_{n+1}, x_n) \mu_\beta(x_{n+1}).\end{aligned}$$

(8.59) Page 399. Referenced in Exercises: [8.21](#).

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s).$$

(8.63) Page 404. Referenced in Exercises: [8.19](#), [8.21](#), [8.22](#), [8.23](#).

$$\begin{aligned}p(x) &= \prod_{s \in \text{ne}(x)} \left[\sum_{X_s} F_s(x, X_s) \right] \\ &= \prod_{s \in \text{ne}(x)} \mu_{f_s \rightarrow x}(x).\end{aligned}$$

(8.66) Page 404. Referenced in Exercises: [8.19](#), [8.20](#).

$$\begin{aligned}\mu_{f_s \rightarrow x}(x) &= \sum_{x_1} \dots \sum_{x_M} f_s(x, x_1, \dots, x_M) \prod_{m \in \text{ne}(f_s) \setminus x} \left[\sum_{X_{xm}} G(x_m, X_{sm}) \right] \\ &= \sum_{x_1} \dots \sum_{x_M} f_s(x, x_1, \dots, x_M) \prod_{m \in \text{ne}(f_s) \setminus x} \mu_{x_m \rightarrow f_s}(x_m).\end{aligned}$$

(8.67) Page 405. Referenced in Exercises: [8.21](#).

$$\mu_{x_m \rightarrow f_s}(x_m) = \sum_{X_{sm}} G_m(x_m, X_{sm}).$$

(8.68) Page 405. Referenced in Exercises: [8.21](#).

$$G_m(x_m, X_{sm}) = \prod_{l \in \text{ne}(x_m) \setminus f_s} F_l(x_m, X_{ml}).$$

(8.69) Page 404. Referenced in Exercises: [8.19](#), [8.20](#), [8.23](#).

$$\begin{aligned} \mu_{x_m \rightarrow f_s}(x_m) &= \prod_{l \in \text{ne}(x_m) \setminus f_s} \left[\sum_{X_{ml}} F_l(x_m, X_{ml}) \right] \\ &= \prod_{l \in \text{ne}(x_m) \setminus f_s} \mu_{f_l \rightarrow x_m}(x_m). \end{aligned}$$

(8.70) Page 406. Referenced in Exercises: [8.19](#), [8.20](#).

$$\mu_{x \rightarrow f}(x) = 1.$$

(8.71) Page 406. Referenced in Exercises: [8.19](#), [8.20](#).

$$\mu_{f \rightarrow x}(x) = f(x).$$

(8.72) Page 408. Referenced in Exercises: [8.21](#), [8.25](#).

$$p(\mathbf{x}_s) = f_s(\mathbf{x}_s) \prod_{i \in \text{ne}(f_s)} \mu_{x_i \rightarrow f_s}(x_i).$$

(8.73) Page 409. Referenced in Exercises: [8.25](#).

$$\tilde{p}(\mathbf{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_3, x_4).$$

(8.74) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{x_1 \rightarrow f_a}(x_1) = 1.$$

(8.75) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{f_a \rightarrow x_2}(x_2) = \sum_{x_1} f_a(x_1, x_2).$$

(8.77) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{f_c \rightarrow x_2}(x_2) = \sum_{x_4} f_c(x_3, x_4).$$

(8.78) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{x_2 \rightarrow f_b}(x_2) = \mu_{f_a \rightarrow x_2}(x_2) \mu_{f_b \rightarrow x_2}(x_2).$$

(8.79) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{f_b \rightarrow x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \rightarrow f_b}(x_2).$$

(8.81) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{f_b \rightarrow x_2}(x_2) = \sum_{x_3} f_b(x_2, x_3).$$

(8.82) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{x_2 \rightarrow f_a}(x_2) = \mu_{f_b \rightarrow x_2}(x_2) \mu_{f_c \rightarrow x_2}(x_2).$$

(8.83) Page 409. Referenced in Exercises: [8.25](#).

$$\mu_{f_a \rightarrow x_1}(x_1) = \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \rightarrow f_a}(x_2).$$

(8.104) Page 419. Referenced in Exercises: [8.6](#).

$$p(y=1|x_1, \dots, x_M) = 1 - (1 - \mu_0) \prod_{i=1}^M (1 - \mu_i)^{x_i}.$$

(9.1) Page 424. Referenced in Exercises: [9.1](#).

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2.$$

(9.2) Page 425. Referenced in Exercises: [9.1, 9.2](#).

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

(9.4) Page 425. Referenced in Exercises: [9.1](#).

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}.$$

(9.5) Page 427. Referenced in Exercises: [9.2](#).

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

(9.7) Page 430. Referenced in Exercises: [9.3](#).

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

(9.10) Page 431. Referenced in Exercises: [9.3](#).

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}.$$

(9.11) Page 431. Referenced in Exercises: [9.3](#).

$$p(\mathbf{x}|z_k=1) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

(C.3) Page 695. Referenced in Exercises: [6.13, 6.25, 6.27](#).

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

(C.6) Page 696. Referenced in Exercises: [3.16](#), [6.21](#).

$$(\mathbf{I} + \mathbf{AB})^{-1}\mathbf{A} = \mathbf{A}(\mathbf{I} + \mathbf{BA})^{-1}$$

(C.9) Page 696. Referenced in Exercises: [2.17](#), [3.21](#).

$$\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{CAB}) = \text{Tr}(\mathbf{BCA}).$$

(C.13) Page 697. Referenced in Exercises: [1.32](#), [7.12](#).

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

(C.14) Page 697. Referenced in Exercises: [3.16](#), [7.12](#).

$$|\mathbf{I}_N + \mathbf{AB}^\top| = |\mathbf{I}_M + \mathbf{A}^\top \mathbf{B}|.$$

(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](#), [5.25](#), [6.14](#), [7.3](#), [7.18](#), [9.2](#).

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{a}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^\top \mathbf{x}) = \mathbf{a}.$$

(C.20) Page 697. Referenced in Exercises: [6.27](#).

$$\frac{\partial}{\partial x}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial x} \mathbf{B} + \mathbf{A} \frac{\partial \mathbf{B}}{\partial x}.$$

(C.21) Page 698. Referenced in Exercises: [2.34](#), [2.37](#), [3.6](#), [5.3](#), [6.27](#), [7.12](#), [7.19](#), [9.2](#).

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1}.$$

(C.22) Page 698. Referenced in Exercises: [6.27](#), [7.12](#), [7.19](#).

$$\frac{\partial}{\partial x} \ln|\mathbf{A}| = \text{Tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x}\right).$$

(C.24) Page 698. Referenced in Exercises: [2.34](#), [2.37](#), [3.6](#), [4.10](#), [5.3](#).

$$\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{AB}) = \mathbf{B}^\top.$$

(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](#), [6.4](#).

$$|\mathbf{A} - \lambda_i \mathbf{I}| = 0.$$

(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](#).

$$\text{Tr}(\mathbf{A}) = \sum_{i=1}^M \lambda_i.$$

(E.4) Page 708. Referenced in Exercises: [1.34](#), [2.14](#), [2.60](#), [3.5](#), [4.4](#), [4.9](#).

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}).$$