

Solutions Manual to  
Pattern Recognition and Machine Learning

Hiromichi Inawashiro

January 12, 2026

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Probability Distributions</b>	<b>43</b>
<b>3</b>	<b>Linear Models for Regression</b>	<b>106</b>
<b>4</b>	<b>Linear Models for Classification</b>	<b>131</b>
<b>5</b>	<b>Neural Networks</b>	<b>158</b>
<b>6</b>	<b>Kernel Methods</b>	<b>195</b>
<b>7</b>	<b>Sparse Kernel Machines</b>	<b>224</b>

# 1 Introduction

## 1.1

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2. \quad (1.1)$$

Setting the derivative with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = \sum_{n=1}^N \frac{\partial y(x_n, \mathbf{w})}{\partial \mathbf{w}} (y(x_n, \mathbf{w}) - t_n). \quad (1.2)$$

If

$$y(x_n, \mathbf{w}) = \mathbf{w}^\top \boldsymbol{\phi}(x_n), \quad (1.3)$$

then

$$\mathbf{0} = \sum_{n=1}^N \boldsymbol{\phi}(x_n) (\mathbf{w}^\top \boldsymbol{\phi}(x_n) - t_n). \quad (1.4)$$

Then,

$$\left( \sum_{n=1}^N \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\top \right) \mathbf{w} = \sum_{n=1}^N t_n \boldsymbol{\phi}(x_n). \quad (1.5)$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} E(\mathbf{w}) = \mathbf{A}^{-1} \mathbf{v}, \quad (1.6)$$

where

$$\begin{aligned} \mathbf{A} &= \sum_{n=1}^N \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\top, \\ \mathbf{v} &= \sum_{n=1}^N t_n \boldsymbol{\phi}(x_n). \end{aligned} \quad (1.7)$$

If

$$\boldsymbol{\phi}(x_n) = \begin{bmatrix} 1 \\ x_n \\ \vdots \\ x_n^M \end{bmatrix},$$

then

$$\begin{aligned} A_{mm'} &= \sum_{n=1}^N x_n^{m+m'}, \\ v_m &= \sum_{n=1}^N t_n x_n^m. \end{aligned} \tag{1.8}$$

## 1.2

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2. \tag{1.9}$$

Setting the derivative with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = \sum_{n=1}^N \frac{\partial y(x_n, \mathbf{w})}{\partial \mathbf{w}} (y(x_n, \mathbf{w}) - t_n) + \lambda \mathbf{w}. \tag{1.10}$$

If

$$y(x_n, \mathbf{w}) = \mathbf{w}^\top \phi(x_n), \tag{1.11}$$

then

$$\mathbf{0} = \sum_{n=1}^N \phi(x_n) (\mathbf{w}^\top \phi(x_n) - t_n) + \lambda \mathbf{w}. \tag{1.12}$$

Then,

$$\left( \sum_{n=1}^N \phi(x_n) \phi(x_n)^\top + \lambda \mathbf{I} \right) \mathbf{w} = \sum_{n=1}^N t_n \phi(x_n). \tag{1.13}$$

Therefore,

$$\operatorname{argmin}_{\mathbf{w}} E(\mathbf{w}) = \mathbf{A}^{-1} \mathbf{v}, \tag{1.14}$$

where

$$\begin{aligned} \mathbf{A} &= \sum_{n=1}^N \phi(x_n) \phi(x_n)^\top + \lambda \mathbf{I}, \\ \mathbf{v} &= \sum_{n=1}^N t_n \phi(x_n). \end{aligned} \tag{1.15}$$

If

$$\phi(x_n) = \begin{bmatrix} 1 \\ x_n \\ \vdots \\ x_n^M \end{bmatrix},$$

then

$$\begin{aligned} A_{mm'} &= \sum_{n=1}^N x_n^{m+m'} + \lambda I_{mm'}, \\ v_m &= \sum_{n=1}^N t_n x_n^m. \end{aligned} \tag{1.16}$$

### 1.3

Let

$$\begin{aligned} p(a|r) &= \frac{3}{10}, p(o|r) = \frac{2}{5}, p(l|r) = \frac{3}{10}, \\ p(a|b) &= \frac{1}{2}, p(o|b) = \frac{1}{2}, \\ p(a|g) &= \frac{3}{10}, p(o|g) = \frac{3}{10}, p(l|g) = \frac{2}{5}. \end{aligned} \tag{1.17}$$

Let

$$p(r) = \frac{1}{5}, p(b) = \frac{1}{5}, p(g) = \frac{3}{5}. \tag{1.18}$$

(a)

By the Bayes' theorem,

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \tag{1.19}$$

Therefore,

$$p(a) = \frac{17}{50}. \tag{1.20}$$

(b)

By the Bayes' theorem,

$$p(g|o) = \frac{p(g,o)}{p(o)}. \tag{1.21}$$

By the Bayes' theorem, the right hand side can be written as

$$\frac{p(o|g)p(g)}{p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g)}. \quad (1.22)$$

Therefore,

$$p(g|o) = \frac{1}{2}. \quad (1.23)$$

## 1.4

Let  $x$  and  $y$  be variables such that

$$x = g(y). \quad (1.24)$$

Let  $\hat{x}$  and  $\hat{y}$  be the locations of the maximum of  $p_x$  and  $p_y$  respectively. Let us assume that there exists a positive  $\epsilon$  such that if

$$|y - \hat{y}| < \epsilon, \quad (1.25)$$

then

$$g'(y) \neq 0. \quad (1.26)$$

(a)

Since

$$\int p_x(x)dx = \int p_x(g(y)) |g'(y)|dy, \quad (1.27)$$

we have

$$p_y(y) = p_x(g(y)) |g'(y)|. \quad (1.28)$$

Taking the derivative and substituting

$$y = \hat{y} \quad (1.29)$$

gives

$$0 = g'(\hat{y})p'_x(g(\hat{y})) + p_x(g(\hat{y}))g''(\hat{y}). \quad (1.30)$$

Therefore, in general,

$$\hat{x} \neq g(\hat{y}). \quad (1.31)$$

(b)

By (a), if

$$g(y) = ay + b, \quad (1.32)$$

then

$$0 = ap'_x(g(\hat{y})). \quad (1.33)$$

Therefore,

$$\hat{x} = g(\hat{y}). \quad (1.34)$$

## 1.5

We have

$$\text{var } f(x) = E(f(x) - E f(x))^2. \quad (1.35)$$

The right hand side can be written as

$$E((f(x))^2 - 2f(x)E f(x) + (E f(x))^2) = E(f(x))^2 - (E f(x))^2. \quad (1.36)$$

Therefore,

$$\text{var } f(x) = E(f(x))^2 - (E f(x))^2. \quad (1.37)$$

## 1.6

We have

$$\text{cov}(x, y) = E((x - E x)(y - E y)). \quad (1.38)$$

The right hand side can be written as

$$E xy - E(x E y) - E(y E x) + E(E x E y) = E xy - E x E y. \quad (1.39)$$

The right hand side can be written as

$$\int xyp(x, y)dxdy - \int xp(x)dx \int yp(y)dy. \quad (1.40)$$

If  $x$  and  $y$  are independent, by the definition,

$$f(x, y) = f(x)f(y). \quad (1.41)$$

Then,

$$\int xyp(x, y)dxdy = \int p(x)dx \int p(y)dy. \quad (1.42)$$

Therefore,

$$\text{cov}(x, y) = 0. \quad (1.43)$$

## 1.7

(a)

Let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx. \quad (1.44)$$

Then,

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2 + y^2)\right) dx dy. \quad (1.45)$$

By the transformation from Cartesian coordinates  $(x, y)$  to polar coordinates  $(r, \theta)$ , the right hand side can be written as

$$\int_0^{\infty} \int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} dr d\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr. \quad (1.46)$$

By the transformation  $s = \frac{r}{\sigma}$ , the right hand side can be written as

$$2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{1}{2}s^2\right) s ds = 2\pi\sigma^2 \left[ -\exp\left(-\frac{1}{2}s^2\right) \right]_0^{\infty}. \quad (1.47)$$

Therefore,

$$I = (2\pi\sigma^2)^{\frac{1}{2}}. \quad (1.48)$$

(b)

By the definition,

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \quad (1.49)$$

Then,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \quad (1.50)$$

By the transformation  $t = x - \mu$ , the right hand side can be written as

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}t^2\right) dt = (2\pi\sigma^2)^{-\frac{1}{2}} I. \quad (1.51)$$

Therefore,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1. \quad (1.52)$$

## 1.8

(a)

Let  $x$  be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.53)$$

Then,

$$\mathbb{E} x = \int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.54)$$

The right hand side can be written as

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \quad (1.55)$$

By the transformation

$$y = x - \mu, \quad (1.56)$$

the integral can be written as

$$\begin{aligned} & \int_{-\infty}^{\infty} (y + \mu) \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy \\ &= \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy + \mu \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy. \end{aligned} \quad (1.57)$$

By 1.7(a), the right hand side can be written as

$$\mu (2\pi\sigma^2)^{\frac{1}{2}}. \quad (1.58)$$

Therefore,

$$\mathbb{E} x = \mu. \quad (1.59)$$

(b)

By 1.7(b),

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1, \quad (1.60)$$

so that

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 1. \quad (1.61)$$

Taking the derivative with respect to  $\sigma^2$  gives

$$\begin{aligned} & (2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) (\sigma^2)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ & + (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} (\sigma^2)^{-2} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 0. \end{aligned} \quad (1.62)$$

The left hand side can be written as

$$\begin{aligned} & -\frac{1}{2} (\sigma^2)^{-1} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2} (\sigma^2)^{-2} \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx \\ & = -\frac{1}{2} (\sigma^2)^{-1} + \frac{1}{2} (\sigma^2)^{-2} \text{var } x. \end{aligned} \quad (1.63)$$

Therefore,

$$\text{var } x = \sigma^2. \quad (1.64)$$

## 1.9

### (a)

Let  $x$  be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.65)$$

Setting the derivative of the right hand side with respect to  $x$  to zero gives

$$0 = (2\pi\sigma^2)^{-\frac{1}{2}} \left(-\frac{1}{\sigma^2}(x-\mu)\right) \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \quad (1.66)$$

Therefore,

$$\text{mode } x = \mu. \quad (1.67)$$

### (b)

Let  $\mathbf{x}$  be a variable such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (1.68)$$

Setting the derivative of the right hand side with respect to  $\mathbf{x}$  to zero gives

$$\mathbf{0} = -(2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x} - \boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \quad (1.69)$$

Therefore,

$$\text{mode } \mathbf{x} = \boldsymbol{\mu}. \quad (1.70)$$

## 1.10

(a)

We have

$$E(x + y) = \iint (x + y)p(x, y)dxdy. \quad (1.71)$$

The right hand side can be written as

$$\int x \left( \int p(x, y)dy \right) dx + \int y \left( \int p(x, y)dx \right) dy = \int xp(x)dx + \int yp(y)dy. \quad (1.72)$$

The right hand side can be written as

$$E x + E y. \quad (1.73)$$

Therefore,

$$E(x + y) = E x + E y. \quad (1.74)$$

(b)

We have

$$\text{var}(x + y) = E(x + y - E(x + y))^2 \quad (1.75)$$

The right hand side can be written as

$$\begin{aligned} & E(x - E x)^2 + 2 E((x - E x)(y - E y)) + E(y - E y)^2 \\ &= \text{var } x + 2 \text{cov}(x, y) + \text{var } y. \end{aligned} \quad (1.76)$$

By 1.6, if  $x$  and  $y$  are independent, then

$$\text{cov}(x, y) = 0. \quad (1.77)$$

Therefore,

$$\text{var}(x + y) = \text{var } x + \text{var } y. \quad (1.78)$$

## 1.11

Let  $x_1, \dots, x_N$  be variables such that

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (1.79)$$

Then,

$$\ln \left( \prod_{n=1}^N p(x_n) \right) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2. \quad (1.80)$$

Setting the derivatives with respect to  $\mu$  and  $\sigma^2$  to zero gives

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu), \\ 0 &= -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{n=1}^N (x_n - \mu)^2. \end{aligned} \quad (1.81)$$

Therefore, the maximum likelihood solutions for  $\mu$  and  $\sigma^2$  are given by

$$\begin{aligned} \mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N x_n, \\ \sigma_{\text{ML}}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2. \end{aligned} \quad (1.82)$$

## 1.12

(a)

Let  $x_n$  and  $x_{n'}$  be independent variables such that

$$\begin{aligned} p(x_n) &= \mathcal{N}(x_n | \mu, \sigma^2), \\ p(x_{n'}) &= \mathcal{N}(x_{n'} | \mu, \sigma^2). \end{aligned} \quad (1.83)$$

Then,

$$\mathbb{E} x_n x_{n'} = \mu^2. \quad (1.84)$$

By the property

$$\mathbb{E} x_n^2 = \text{var } x_n + (\mathbb{E} x_n)^2, \quad (1.85)$$

we have

$$\mathbb{E} x_n^2 = \sigma^2 + \mu^2. \quad (1.86)$$

Therefore,

$$\mathbb{E} x_n x_{n'} = \mu^2 + I_{nn'} \sigma^2. \quad (1.87)$$

(b)

Let  $x_1, \dots, x_N$  be independent variables such that

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (1.88)$$

By 1.11, the maximum likelihood solution for  $\mu$  is given by

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (1.89)$$

Then,

$$\mathbb{E} \mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n. \quad (1.90)$$

Therefore,

$$\mathbb{E} \mu_{\text{ML}} = \mu. \quad (1.91)$$

(c)

Let  $x_1, \dots, x_N$  be independent variables such that

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (1.92)$$

By 1.11, the maximum likelihood solution for  $\sigma^2$  is given by

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2. \quad (1.93)$$

Then,

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n - \mu_{\text{ML}})^2. \quad (1.94)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} \sum_{n=1}^N E(x_n^2 - 2\mu_{ML}x_n + \mu_{ML}^2) \\ &= \frac{1}{N} \sum_{n=1}^N E x_n^2 - \frac{2}{N} E \left( \mu_{ML} \left( \sum_{n=1}^N x_n \right) \right) + E \mu_{ML}^2. \end{aligned} \quad (1.95)$$

The first term of the right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) = \mu^2 + \sigma^2, \quad (1.96)$$

while, by 1.11, the second and third terms can be written as

$$-\frac{2}{N} E \left( N \left( \frac{1}{N} \sum_{n=1}^N x_n \right)^2 \right) + E \left( \frac{1}{N} \sum_{n=1}^N x_n \right)^2 = -E \left( \frac{1}{N} \sum_{n=1}^N x_n \right)^2. \quad (1.97)$$

By (a), the right hand side can be written as

$$\begin{aligned} & -\frac{1}{N^2} \sum_{n=1}^N E x_n^2 - \frac{2}{N^2} \sum_{1 \leq n < n' \leq N} E x_n x_{n'} \\ &= -\frac{1}{N^2} N (\mu^2 + \sigma^2) - \frac{2}{N^2} \frac{N(N-1)}{2} \mu^2. \end{aligned} \quad (1.98)$$

The right hand side can be written as

$$-\frac{1}{N} (\mu^2 + \sigma^2) - \frac{N-1}{N} \mu^2 = -\mu^2 - \frac{1}{N} \sigma^2. \quad (1.99)$$

Then,

$$E \sigma_{ML}^2 = \mu^2 + \sigma^2 - \mu^2 - \frac{1}{N} \sigma^2. \quad (1.100)$$

Therefore,

$$E \sigma_{ML}^2 = \frac{N-1}{N} \sigma^2. \quad (1.101)$$

## 1.13

Let  $x_1, \dots, x_N$  be variables such that

$$\begin{aligned} \mathbb{E} x_n &= \mu, \\ \text{var } x_n &= \sigma^2. \end{aligned} \quad (1.102)$$

We have

$$\mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \frac{1}{N^2} \sum_{n=1}^N \mathbb{E} (x_n - \mu)^2. \quad (1.103)$$

The right hand side can be written as

$$\frac{1}{N^2} \sum_{n=1}^N \text{var } x_n = \frac{\sigma^2}{N}. \quad (1.104)$$

Therefore,

$$\mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \frac{\sigma^2}{N}. \quad (1.105)$$

## 1.14

Let

$$\begin{aligned} w_{dd'}^S &= \frac{1}{2}(w_{dd'} + w_{d'd}), \\ w_{dd'}^A &= \frac{1}{2}(w_{dd'} - w_{d'd}). \end{aligned} \quad (1.106)$$

(a)

We have

$$\begin{aligned} w_{dd'} &= w_{dd'}^S + w_{dd'}^A, \\ w_{dd'}^S &= w_{d'd}, \\ w_{dd'}^A &= -w_{d'd}. \end{aligned} \quad (1.107)$$

(b)

We have

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^A x_d x_{d'} = \frac{1}{2} \sum_{d=1}^D \sum_{d'=1}^D (w_{dd'} - w_{d'd}) x_d x_{d'}. \quad (1.108)$$

The right hand side can be written as

$$\frac{1}{2} \left( \sum_{d=1}^D \sum_{d'=1}^D w_{dd'} x_d x_{d'} - \sum_{d=1}^D \sum_{d'=1}^D w_{d'd} x_d x_{d'} \right) = 0. \quad (1.109)$$

Therefore,

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^A x_d x_{d'} = 0. \quad (1.110)$$

(c)

We have

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'} x_d x_{d'} = \sum_{d=1}^D \sum_{d'=1}^D (w_{dd'}^S + w_{dd'}^A) x_d x_{d'}. \quad (1.111)$$

By (b), the right hand side can be written as

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^S x_d x_{d'} + \sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^A x_d x_{d'} = \sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^S x_d x_{d'}, \quad (1.112)$$

Therefore,

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'} x_d x_{d'} = \sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^S x_d x_{d'}. \quad (1.113)$$

(d)

Since  $\mathbf{W}^S$  is a  $D \times D$  symmetric matrix, its number of independent parameters is  $\frac{D(D+1)}{2}$ .

## 1.15

(a)

Let  $n(D, M)$  be the number of independent parameters of a polynomial in  $D$  dimensions and  $M$  orders. Then

$$n(1, M) = n(1, M-1) = 1. \quad (1.114)$$

Let us assume that

$$n(D, M) = \sum_{d=1}^D n(d, M-1). \quad (1.115)$$

The independent terms of a polynomial in  $D+1$  dimensions and  $M$  orders can be split into 1. the ones of a polynomial in  $D$  dimensions and  $M$  orders and 2. the ones generated by multiplying the ones in  $D+1$  dimensions and  $M$  orders by the  $D+1$ th variable. Then,

$$n(D+1, M) = n(D, M) + n(D+1, M-1), \quad (1.116)$$

so that

$$n(D+1, M) = \sum_{d=1}^{D+1} n(d, M-1). \quad (1.117)$$

Therefore, the assumption is proved by induction on  $D$ .

**(b)**

We have

$$\sum_{d=1}^1 \frac{(d+M-2)!}{(d-1)!(M-1)!} = 1. \quad (1.118)$$

Let us assume that

$$\sum_{d=1}^D \frac{(d+M-2)!}{(d-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}. \quad (1.119)$$

Then,

$$\sum_{d=1}^{D+1} \frac{(d+M-2)!}{(d-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!}. \quad (1.120)$$

The right hand side can be written as

$$\frac{D(D+M-1)! + M(D+M-1)!}{D!M!} = \frac{(D+M)!}{D!M!}. \quad (1.121)$$

Therefore, the assumption is proved by induction on  $D$ .

(c)

By 1.14(d),

$$n(D, 2) = \frac{D(D+1)}{2}. \quad (1.122)$$

Let us assume that

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!}. \quad (1.123)$$

By (a),

$$n(D, M+1) = \sum_{d=1}^D n(d, M). \quad (1.124)$$

By the assumption and (b), the right hand side can be written as

$$\sum_{d=1}^D \frac{(d+M-1)!}{(d-1)!M!} = \frac{(D+M)!}{(D-1)!(M+1)!}. \quad (1.125)$$

Therefore, the assumption is proved by induction on  $M$ .

## 1.16

(a)

Let  $N(D, M)$  be the number of independent parameters in all of the terms up to and including the ones of  $D$  dimensions and  $M$  orders. By 1.15,

$$N(D, M) = \sum_{m=0}^M n(D, m), \quad (1.126)$$

where

$$n(D, m) = \frac{(D+m-1)!}{(D-1)!m!}. \quad (1.127)$$

(b)

By (a),

$$N(D, 0) = 1. \quad (1.128)$$

Let us assume that

$$\sum_{m=0}^M n(D, m) = \frac{(D+M)!}{D!M!}. \quad (1.129)$$

Then,

$$\sum_{m=0}^{M+1} n(D, m) = \frac{(D+M)!}{D!M!} + \frac{(D+M)!}{(D-1)!(M+1)!}. \quad (1.130)$$

The right hand side can be written as

$$\frac{(M+1)(D+M)! + D(D+M)!}{D!(M+1)!} = \frac{(D+M+1)!}{D!(M+1)!}. \quad (1.131)$$

Then, the assumption is proved by induction on  $M$ . Therefore,

$$N(D, M) = \frac{(D+M)!}{D!M!}. \quad (1.132)$$

**(c)**

By the approximation

$$n! \simeq n^n \exp(-n), \quad (1.133)$$

we have

$$\frac{(D+M)!}{D!M!} \simeq \frac{(D+M)^{D+M}}{D^D M^M}. \quad (1.134)$$

The right hand side can be written as

$$D^M \left(1 + \frac{M}{D}\right)^D \left(\frac{1}{M} + \frac{1}{D}\right)^M = M^D \left(1 + \frac{D}{M}\right)^M \left(\frac{1}{D} + \frac{1}{M}\right)^D. \quad (1.135)$$

Therefore,

$$N(D, M) \simeq \begin{cases} D^M, & D \gg M, \\ M^D, & M \gg D. \end{cases} \quad (1.136)$$

**(d)**

By (b),

$$\begin{aligned} N(10, 3) &= 286, \\ N(100, 3) &= 176851, \\ N(1000, 3) &= 167668501. \end{aligned} \quad (1.137)$$

## 1.17

Let

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \quad (1.138)$$

(a)

We have

$$\Gamma(x+1) = \int_0^\infty u^x \exp(-u) du. \quad (1.139)$$

The right hand side can be written as

$$[-u^x \exp(-u)]_{u=0}^{u=\infty} + \int_0^\infty xu^{x-1} \exp(-u) du = x\Gamma(x). \quad (1.140)$$

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \quad (1.141)$$

(b)

We have

$$\Gamma(1) = \int_0^\infty \exp(-u) du, \quad (1.142)$$

so that

$$\Gamma(1) = 0!. \quad (1.143)$$

For a positive integer  $x$ , let us assume that

$$\Gamma(x) = (x-1)!. \quad (1.144)$$

By (a),

$$\Gamma(x+1) = x\Gamma(x), \quad (1.145)$$

so that

$$\Gamma(x+1) = x!. \quad (1.146)$$

Therefore, the assumption is proved by induction on  $x$ .

## 1.18

(a)

Let

$$\prod_{d=1}^D \int_{-\infty}^\infty \exp(-x_d^2) dx_i = S_D \int_0^\infty \exp(-r^2) r^{D-1} dr, \quad (1.147)$$

where  $S_D$  is the surface area of a sphere of unit radius in  $D$  dimensions. By 1.7, the left hand side can be written as  $\pi^{\frac{D}{2}}$ . By the transformation

$$s = r^2, \quad (1.148)$$

the right hand side can be written as

$$\frac{S_D}{2} \int_0^\infty \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \quad (1.149)$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \quad (1.150)$$

**(b)**

The volume of the sphere can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \quad (1.151)$$

Therefore,

$$V_D = \frac{S_D}{D}. \quad (1.152)$$

**(c)**

By (a) and (b),

$$\begin{aligned} S_2 &= 2\pi, \\ V_2 &= \pi. \end{aligned} \quad (1.153)$$

Similarly,

$$\begin{aligned} S_3 &= 4\pi, \\ V_3 &= \frac{4}{3}\pi. \end{aligned} \quad (1.154)$$

## 1.19

**(a)**

The volume of a cube of side 2 in  $D$  dimensions is  $2^D$ . By 1.18, the ratio of the volume of the concentric sphere of radius 1 divided by the volume of the

cube is given by

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma\left(\frac{D}{2}\right)}. \quad (1.155)$$

**(b)**

By (a) and the Sterling's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}} \exp(-x)x^{\frac{x+1}{2}}, \quad (1.156)$$

we have

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D2^{D-1}(2\pi)^{\frac{1}{2}} \exp\left(1 - \frac{D}{2}\right)\left(\frac{D}{2} - 1\right)^{\frac{D}{4}}}. \quad (1.157)$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}} \frac{1}{D} \left( \frac{e^2\pi^2}{8D-16} \right)^{\frac{D}{4}}. \quad (1.158)$$

Therefore,

$$\lim_{D \rightarrow \infty} \frac{V_D}{2^D} = 0. \quad (1.159)$$

**(c)**

The ratio of the distance from the center of the cube to one of the corners divided by the perpendicular distance to one of the sides is given by

$$\frac{\sqrt{\sum_{i=1}^D 1^2}}{1} = \sqrt{D}. \quad (1.160)$$

Therefore, the ratio goes to  $\infty$  as  $D \rightarrow \infty$ .

## 1.20

Let  $\mathbf{x}$  be a variable in  $D$  dimensions such that

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \quad (1.161)$$

(a)

We have

$$\int_{r \leq \| \mathbf{x} \| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_r^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\phi, \quad (1.162)$$

where  $\phi$  is the vector of the angular components of the polar coordinate and  $J$  is the Jacobian of the transformation from the Cartesian to polar coordinate. For a sufficiently small  $\epsilon$ , the right hand side can be approximated as

$$\begin{aligned} & (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon} \int J dr' d\phi \\ &= (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \| \mathbf{x} \| \leq r+\epsilon} d\mathbf{x}. \end{aligned} \quad (1.163)$$

Therefore,

$$\int_{r \leq \| \mathbf{x} \| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} \simeq p(r)\epsilon, \quad (1.164)$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad (1.165)$$

and  $S_D$  is the surface area of a unit sphere in  $D$  dimensions.

(b)

Setting the derivative of  $p(r)$  to zero gives

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left( (D-1)r^{D-2} - \frac{r^D}{\sigma^2} \right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \quad (1.166)$$

Therefore,  $p(r)$  is maximised at a single stationary point

$$\hat{r} = \sqrt{D-1}\sigma. \quad (1.167)$$

(c)

We have

$$\frac{p(\hat{r} + \epsilon)}{p(\hat{r})} = \left( \frac{\hat{r} + \epsilon}{\hat{r}} \right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \quad (1.168)$$

By (b), the right hand side can be written as

$$\begin{aligned} & \exp \left( (D-1) \ln \left( 1 + \frac{\epsilon}{\hat{r}} \right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2} \right) \\ &= \exp \left( \frac{\hat{r}^2}{\sigma^2} \ln \left( 1 + \frac{\epsilon}{\hat{r}} \right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2} \right). \end{aligned} \quad (1.169)$$

By the Taylor series

$$\ln(1+x) = x - \frac{1}{2}x^2 + o(x^3), \quad (1.170)$$

the right hand side can be approximated as

$$\exp \left( \frac{\hat{r}^2}{\sigma^2} \left( \frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2} \right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2} \right) = \exp \left( -\frac{\epsilon^2}{\sigma^2} \right). \quad (1.171)$$

Therefore,

$$p(\hat{r} + \epsilon) \simeq p(\hat{r}) \exp \left( -\frac{\epsilon^2}{\sigma^2} \right). \quad (1.172)$$

#### (d)

Let  $\hat{\mathbf{r}}$  be a vector of length  $\hat{r}$ . We have

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp \left( \frac{\hat{r}^2}{2\sigma^2} \right). \quad (1.173)$$

By (b), the right hand side can be written as

$$\exp \left( \frac{D-1}{2} \right). \quad (1.174)$$

Therefore,

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp \left( \frac{D-1}{2} \right). \quad (1.175)$$

### 1.21

#### (a)

If  $0 \leq a \leq b$ , then

$$0 \leq a(b-a). \quad (1.176)$$

Therefore,

$$a \leq (ab)^{\frac{1}{2}}. \quad (1.177)$$

(b)

For a two-class classification problem of  $\mathbf{x}$ , let the classes be  $\mathcal{C}_1$  and  $\mathcal{C}_2$  and let the decision regions be  $\mathcal{R}_1$  and  $\mathcal{R}_2$ . Let us choose the decision regions to minimise the probability of misclassification. Then,

$$\begin{aligned} p(\mathbf{x}, \mathcal{C}_1) &> p(\mathbf{x}, \mathcal{C}_2) \Rightarrow \mathbf{x} \in \mathcal{C}_1, \\ p(\mathbf{x}, \mathcal{C}_2) &> p(\mathbf{x}, \mathcal{C}_1) \Rightarrow \mathbf{x} \in \mathcal{C}_2. \end{aligned} \quad (1.178)$$

By (a),

$$\begin{aligned} \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} &\leq \int_{\mathcal{R}_1} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}, \\ \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} &\leq \int_{\mathcal{R}_2} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \end{aligned} \quad (1.179)$$

Therefore,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \quad (1.180)$$

## 1.22

Let

$$E L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.181)$$

If

$$L_{kj} = 1 - I_{kj}, \quad (1.182)$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \quad (1.183)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \quad (1.184)$$

Then,

$$E L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.185)$$

Therefore, minimising  $E L$  reduces to choosing the criterion to maximise the posterior probability  $p(\mathcal{C}_j|\mathbf{x})$ .

### 1.23

Let

$$E L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.186)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.187)$$

Then,

$$E L = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.188)$$

Therefore, minimising  $E L$  reduces to minimising  $\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$ .

### 1.24 (Incomplete)

Let

$$E L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} + \lambda \int_{\forall k p(\mathcal{C}_k | \mathbf{x}) < \theta} p(\mathbf{x}) d\mathbf{x}. \quad (1.189)$$

### 1.25

Let

$$E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} dt. \quad (1.190)$$

Setting the derivative with respect to  $\mathbf{y}(\mathbf{x})$  to zero gives

$$\mathbf{0} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) dt. \quad (1.191)$$

The integral of the right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t}) dt - \int \mathbf{t} p(\mathbf{x}, \mathbf{t}) dt = \mathbf{y}(\mathbf{x}) p(\mathbf{x}) - p(\mathbf{x}) \int \mathbf{t} p(\mathbf{t} | \mathbf{x}) dt. \quad (1.192)$$

The integral in the second term of the right hand side can be written as  $E_t(\mathbf{t}|\mathbf{x})$ . Then, the right hand side can be written as

$$\mathbf{0} = p(\mathbf{x}) (\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x})). \quad (1.193)$$

Therefore,

$$\operatorname{argmin}_{\mathbf{y}(\mathbf{x})} E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = E_t(\mathbf{t}|\mathbf{x}). \quad (1.194)$$

For a single target variable  $t$ , it reduces to

$$\operatorname{argmin}_{\mathbf{y}(\mathbf{x})} E L(t, \mathbf{y}(\mathbf{x})) = E_t(t|\mathbf{x}). \quad (1.195)$$

## 1.26

Let

$$E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.196)$$

The right hand side can be written as

$$\begin{aligned} & \iint \|\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x}) + E_t(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &= \iint \|\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &+ 2 \iint (\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x}))^\top (E_t(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &+ \iint \|E_t(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \end{aligned} \quad (1.197)$$

Let us look at each term of the right hand side. The first term can be written as

$$\int \|\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x})\|^2 \left( \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \quad (1.198)$$

The integral of the second term can be written as

$$\int (\mathbf{y}(\mathbf{x}) - E_t(\mathbf{t}|\mathbf{x}))^\top \left( \int (E_t(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.199)$$

Since

$$\begin{aligned}\int E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})p(\mathbf{t}|\mathbf{x})d\mathbf{t} &= E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x}, \mathbf{t})d\mathbf{t}}{p(\mathbf{x})}, \\ \int \mathbf{t}p(\mathbf{t}|\mathbf{x})d\mathbf{t} &= E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}),\end{aligned}\tag{1.200}$$

the second term is zero. The third term can be written as

$$\int \left( \int \|E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}|\mathbf{x})d\mathbf{t} \right) p(\mathbf{x})d\mathbf{x} = \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}.\tag{1.201}$$

Then,

$$E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x})d\mathbf{x} + \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}.\tag{1.202}$$

Therefore,

$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}).\tag{1.203}$$

## 1.27 (Incomplete)

(a)

Let

$$E L_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t)d\mathbf{x}dt.\tag{1.204}$$

Setting the derivative with respect to  $y(\mathbf{x})$  to zero gives

$$0 = qp(\mathbf{x}) \int |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})dt.\tag{1.205}$$

Therefore,

$$\underset{y(\mathbf{x})}{\operatorname{argmin}} E L_q = \left\{ y(\mathbf{x}) \mid \int |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})dt = 0 \right\}.\tag{1.206}$$

(b)

We have

$$\mathbb{E} L_1 = \int \left( \int \text{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.207)$$

The integral of the right hand side with respect to  $t$  can be written as

$$\int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x}) dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x}) dt. \quad (1.208)$$

Therefore,

$$\underset{y(\mathbf{x})}{\operatorname{argmin}} \mathbb{E} L_1 = \text{median}(t|\mathbf{x}). \quad (1.209)$$

(c)

We have

$$\lim_{q \rightarrow 0} \left( \underset{y(\mathbf{x})}{\operatorname{argmin}} \mathbb{E} L_q \right) = \text{mode}(t|\mathbf{x})? \quad (1.210)$$

## 1.28

(a)

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \quad (1.211)$$

Then,

$$h(p^2) = 2h(p). \quad (1.212)$$

Let us assume that, for a positive integer  $n$ ,

$$h(p^n) = nh(p). \quad (1.213)$$

Then, by the first assumption,

$$h(p^{n+1}) = h(p^n) + h(p), \quad (1.214)$$

so that

$$h(p^{n+1}) = (n+1)h(p). \quad (1.215)$$

Therefore, the second assumption is proved by induction on  $n$ .

(b)

For positive integers  $m$  and  $n$ ,

$$h(p^n) = h(p^{\frac{n}{m}m}). \quad (1.216)$$

By the second assumption in (a), the left hand side can be written as  $nh(p)$ .

By the first assumption in (a), the right hand side can be written as  $mh(p^{\frac{n}{m}})$ .

Therefore,

$$h(p^{\frac{n}{m}}) = \frac{n}{m}h(p). \quad (1.217)$$

(c)

By the continuity, for a positive real number  $a$ ,

$$h(p^a) = ah(p). \quad (1.218)$$

Taking the derivative with respect to  $a$  and substituting  $a = 1$  gives

$$(p \ln p)h'(p) = h(p). \quad (1.219)$$

Then,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + \text{const}. \quad (1.220)$$

Ignoring the constants, the left hand side can be written as  $\ln h(p)$  and the right hand side can be written as  $\ln(\ln p)$ . Therefore,

$$h(p) \propto \ln p. \quad (1.221)$$

## 1.29

Let  $x$  be an  $M$ -state discrete random variable. Then, the entropy is given by

$$H(x) = - \sum_{m=1}^M p(x_m) \ln p(x_m), \quad (1.222)$$

where

$$\sum_{m=1}^M p(x_m) = 1. \quad (1.223)$$

By the Jensen's inequality,

$$\sum_{m=1}^M p(x_i) \ln \frac{1}{p(x_m)} \leq \ln \left( \sum_{m=1}^M 1 \right). \quad (1.224)$$

Therefore,

$$H(x) \leq \ln M. \quad (1.225)$$

## 1.30

Let

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu, \sigma^2), \\ q(x) &= \mathcal{N}(x|m, s^2). \end{aligned} \quad (1.226)$$

Then, the Kullback-Leibler divergence is given by

$$KL(p||q) = - \int p(x) \ln \frac{q(x)}{p(x)} dx. \quad (1.227)$$

The right hand side can be written as

$$\begin{aligned} & - \int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\ &= - \int_{-\infty}^{\infty} p(x) \left( -\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2} \right) dx. \end{aligned} \quad (1.228)$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x) dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x) dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx. \quad (1.229)$$

The integral of the second term can be written as

$$\begin{aligned} & \int_{-\infty}^{\infty} (x - \mu + \mu - m)^2 p(x) dx \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx + 2(\mu - m) \int_{-\infty}^{\infty} (x - \mu) p(x) dx \\ & \quad + (\mu - m)^2 \int_{-\infty}^{\infty} p(x) dx. \end{aligned} \quad (1.230)$$

Therefore,

$$KL(p||q) = \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{2s^2} - \frac{1}{2}. \quad (1.231)$$

### 1.31

Let  $\mathbf{x}$  and  $\mathbf{y}$  be two variables. We have

$$\begin{aligned} H(\mathbf{x}) &= - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}, \\ H(\mathbf{y}) &= - \int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \\ H(\mathbf{x}, \mathbf{y}) &= - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x}d\mathbf{y}. \end{aligned} \quad (1.232)$$

Since

$$\begin{aligned} H(\mathbf{x}) &= - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}, \\ H(\mathbf{y}) &= - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} \right) \ln p(\mathbf{y}) d\mathbf{y}, \end{aligned} \quad (1.233)$$

we have

$$H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y}. \quad (1.234)$$

Since

$$\iint p(\mathbf{x}, \mathbf{y}) d\mathbf{x}d\mathbf{y} = 1, \quad (1.235)$$

the Jensen's inequality can be used to have

$$-\iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y} \geq -\ln \left( \iint p(\mathbf{x})p(\mathbf{y}) d\mathbf{x}d\mathbf{y} \right). \quad (1.236)$$

The right hand side can be written as

$$-\ln \left( \int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{y}) d\mathbf{y} \right) = 0. \quad (1.237)$$

Therefore,

$$H(\mathbf{x}, \mathbf{y}) \leq H(\mathbf{x}) + H(\mathbf{y}). \quad (1.238)$$

## 1.32

Let  $\mathbf{x}$  and  $\mathbf{y}$  be variables such that

$$\mathbf{y} = \mathbf{Ax}, \quad (1.239)$$

where  $\mathbf{A}$  is a nonsingular matrix. We have

$$\int p_x(\mathbf{x}) d\mathbf{x} = \int p_x(\mathbf{A}^{-1}\mathbf{y}) |\det \mathbf{A}^{-1}| d\mathbf{y}. \quad (1.240)$$

Then,

$$p_y(\mathbf{y}) = p_x(\mathbf{A}^{-1}\mathbf{y}) |\det \mathbf{A}^{-1}|. \quad (1.241)$$

We have

$$H(\mathbf{y}) = - \int p_y(\mathbf{y}) \ln p_y(\mathbf{y}) d\mathbf{y}. \quad (1.242)$$

The right hand side can be written as

$$\begin{aligned} & - \int p_y(\mathbf{y}) \ln(p_x(\mathbf{A}^{-1}\mathbf{y}) |\det \mathbf{A}^{-1}|) d\mathbf{y} \\ &= - \int p_y(\mathbf{y}) \ln p_x(\mathbf{A}^{-1}\mathbf{y}) d\mathbf{y} + \ln |\det \mathbf{A}| \int p_y(\mathbf{y}) d\mathbf{y}. \end{aligned} \quad (1.243)$$

The first term can be written as

$$-|\det \mathbf{A}^{-1}| \int p_x(\mathbf{A}^{-1}\mathbf{y}) \ln p_x(\mathbf{A}^{-1}\mathbf{y}) d\mathbf{y}. \quad (1.244)$$

By the transformation

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}, \quad (1.245)$$

it can be written as

$$- \int p_x(\mathbf{x}) \ln p_x(\mathbf{x}) d\mathbf{x} = H(\mathbf{x}). \quad (1.246)$$

Therefore,

$$H(\mathbf{y}) = H(\mathbf{x}) + \ln |\det \mathbf{A}|. \quad (1.247)$$

### 1.33

Let  $x$  and  $y$  be two discrete variables with  $K$  and  $L$  states. Then,

$$H(y|x) = - \sum_{k=1}^K \sum_{l=1}^L p(x_k, y_l) \ln p(y_l|x_k). \quad (1.248)$$

If

$$H(y|x) = 0, \quad (1.249)$$

then

$$0 = - \sum_{k=1}^K p(x_k) \sum_{l=1}^L p(y_l|x_k) \ln p(y_l|x_k). \quad (1.250)$$

Since

$$\begin{aligned} p(x_k) &\geq 0, \\ p(y_l|x_k) \ln p(y_l|x_k) &\leq 0, \end{aligned} \quad (1.251)$$

the equation reduces to

$$p(y_l|x_k) \ln p(y_l|x_k) = 0. \quad (1.252)$$

Then,

$$p(y_l|x_k) = \begin{cases} 1, \\ 0. \end{cases} \quad (1.253)$$

Since

$$\sum_{l=1}^L p(y_l|x_k) = 1, \quad (1.254)$$

we have

$$p(y_l|x_k) = \begin{cases} 1, & \text{if } K = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (1.255)$$

### 1.34

Let  $x$  be a variable. We have

$$H(x) = - \int_{-\infty}^{\infty} p(x) \ln p(x) dx. \quad (1.256)$$

In order to maximise  $H(x)$  with the constraints

$$\begin{aligned} \int_{-\infty}^{\infty} p(x)dx &= 1, \\ \int_{-\infty}^{\infty} xp(x)dx &= \mu, \\ \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx &= \sigma^2, \end{aligned} \quad (1.257)$$

let

$$\begin{aligned} L(p) = H(x) + \lambda_1 \left( \int_{-\infty}^{\infty} p(x)dx - 1 \right) + \lambda_2 \left( \int_{-\infty}^{\infty} xp(x)dx - \mu \right) \\ + \lambda_3 \left( \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx - \sigma^2 \right). \end{aligned} \quad (1.258)$$

Setting the variation with respect to  $p$  to zero gives

$$0 = -\ln p - 1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2. \quad (1.259)$$

Then,

$$p(x) = \exp \left( -1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2 \right), \quad (1.260)$$

so that

$$p(x) = c \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right), \quad (1.261)$$

where

$$c = \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right). \quad (1.262)$$

Substituting it to the constraints gives

$$\begin{aligned} c \int_{-\infty}^{\infty} \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= 1, \\ c \int_{-\infty}^{\infty} x \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= \mu, \\ c \int_{-\infty}^{\infty} (x - \mu)^2 \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= \sigma^2. \end{aligned} \quad (1.263)$$

By the transformation

$$y = (-\lambda_3)^{\frac{1}{2}} \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right), \quad (1.264)$$

they can be written as

$$\begin{aligned} c \int_{-\infty}^{\infty} \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= 1, \\ c \int_{-\infty}^{\infty} \left( (-\lambda_3)^{-\frac{1}{2}} y + \mu - \frac{\lambda_2}{2\lambda_3} \right) \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= \mu, \\ c \int_{-\infty}^{\infty} \left( (-\lambda_3)^{-\frac{1}{2}} y - \frac{\lambda_2}{2\lambda_3} \right)^2 \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= \sigma^2. \end{aligned} \quad (1.265)$$

Since

$$\begin{aligned} \int_{-\infty}^{\infty} \exp(-y^2) dy &= \Gamma\left(\frac{1}{2}\right), \\ \int_{-\infty}^{\infty} y \exp(-y^2) dy &= 0, \\ \int_{-\infty}^{\infty} y^2 \exp(-y^2) dy &= \Gamma\left(\frac{3}{2}\right), \end{aligned} \quad (1.266)$$

they can be written as

$$\begin{aligned} c(-\lambda_3)^{-\frac{1}{2}} \Gamma\left(\frac{1}{2}\right) &= 1, \\ c \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) (-\lambda_3)^{-\frac{1}{2}} \Gamma\left(\frac{1}{2}\right) &= \mu, \\ c \left( (-\lambda_3)^{-\frac{3}{2}} \Gamma\left(\frac{3}{2}\right) + (-\lambda_3)^{-\frac{1}{2}} \frac{\lambda_2^2}{4\lambda_3^2} \Gamma\left(\frac{1}{2}\right) \right) &= \sigma^2. \end{aligned} \quad (1.267)$$

Then,

$$\begin{aligned} \lambda_1 &= 1 - \frac{1}{2} \ln(2\pi\sigma^2), \\ \lambda_2 &= 0, \\ \lambda_3 &= -\frac{1}{2\sigma^2}. \end{aligned} \quad (1.268)$$

Therefore,

$$p(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.269)$$

## 1.35

Let  $x$  be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.270)$$

Then,

$$H(x) = - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \ln \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.271)$$

The right hand side can be written as

$$\begin{aligned} & - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 \right) dx \\ &= \frac{1}{2} \ln(2\pi\sigma^2) \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx. \end{aligned} \quad (1.272)$$

Therefore,

$$H(x) = \frac{1}{2} (1 + \ln(2\pi\sigma^2)). \quad (1.273)$$

## 1.36 (Incomplete)

Let  $f$  be a strictly convex function. Then,

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b), \quad (1.274)$$

where  $a \leq b$  and  $0 \leq \lambda \leq 1$ . Let

$$x = \lambda a + (1-\lambda)b. \quad (1.275)$$

Then, the inequality can be written as

$$f(x) \leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b). \quad (1.276)$$

Let

$$g(x) = \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b) - f(x). \quad (1.277)$$

Then,

$$g(x) \geq 0. \quad (1.278)$$

Additionally, for  $x > a$ ,

$$g(x) = (x - a) \left( \frac{f(b) - f(a)}{b - a} - \frac{f(x) - f(a)}{x - a} \right). \quad (1.279)$$

By the mean value theorem, there exists  $c$  and  $y$  such that  $a \leq c \leq b$ ,  $a \leq y \leq x$  and

$$\begin{aligned} f'(c) &= \frac{f(b) - f(a)}{b - a}, \\ f'(y) &= \frac{f(x) - f(a)}{x - a}. \end{aligned} \quad (1.280)$$

Then, for  $x > a$ , the inequality reduces to

$$f'(y) \leq f'(c). \quad (1.281)$$

### 1.37

Let  $\mathbf{x}$  and  $\mathbf{y}$  be two variables. We have

$$H(\mathbf{x}, \mathbf{y}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \quad (1.282)$$

The right hand side can be written as

$$\begin{aligned} & - \iint p(\mathbf{x}, \mathbf{y}) (\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})) d\mathbf{x} d\mathbf{y} \\ &= - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (1.283)$$

By the definition, the first and second terms of the right hand side can be written as  $H(\mathbf{y}|\mathbf{x})$  and  $H(\mathbf{x})$ . Therefore,

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \quad (1.284)$$

### 1.38

Let  $f$  be a strictly convex function. Then,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \quad (1.285)$$

where

$$0 \leq \lambda \leq 1. \quad (1.286)$$

Let us assume that

$$f\left(\sum_{m=1}^M \lambda_m x_m\right) \leq \sum_{m=1}^M \lambda_m f(x_m), \quad (1.287)$$

where

$$\begin{aligned} \sum_{m=1}^M \lambda_m &= 1, \\ \lambda_m &\geq 0. \end{aligned} \quad (1.288)$$

Since  $f$  is strictly convex,

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right), \quad (1.289)$$

where

$$\begin{aligned} \sum_{m=1}^{M+1} \lambda_m &= 1, \\ \lambda_m &\geq 0. \end{aligned} \quad (1.290)$$

By the assumption,

$$f\left(\sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right) \leq \sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m). \quad (1.291)$$

Then,

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m), \quad (1.292)$$

so that

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \sum_{m=1}^{M+1} \lambda_m f(x_m). \quad (1.293)$$

Therefore, the assumption is proved by induction on  $M$ .

## 1.39

Let  $x$  and  $y$  be two binary variables where

$$\begin{aligned} p(x = 0, y = 0) &= \frac{1}{3}, \\ p(x = 0, y = 1) &= \frac{1}{3}, \\ p(x = 1, y = 0) &= 0, \\ p(x = 1, y = 1) &= \frac{1}{3}. \end{aligned} \tag{1.294}$$

(a)

We have

$$H(x) = - \sum_x p(x) \ln p(x). \tag{1.295}$$

We have

$$\begin{aligned} p(x = 0) &= p(x = 0, y = 0) + p(x = 0, y = 1), \\ p(x = 1) &= p(x = 1, y = 0) + p(x = 1, y = 1). \end{aligned} \tag{1.296}$$

Then,

$$\begin{aligned} p(x = 0) &= \frac{2}{3}, \\ p(x = 1) &= \frac{1}{3}. \end{aligned} \tag{1.297}$$

Therefore,

$$H(x) = \ln 3 - \frac{2}{3} \ln 2. \tag{1.298}$$

(b)

We have

$$H(y) = - \sum_y p(y) \ln p(y). \tag{1.299}$$

We have

$$\begin{aligned} p(y = 0) &= p(x = 0, y = 0) + p(x = 1, y = 0), \\ p(y = 1) &= p(x = 0, y = 1) + p(x = 1, y = 1). \end{aligned} \tag{1.300}$$

Then,

$$\begin{aligned} p(y=0) &= \frac{1}{3}, \\ p(y=1) &= \frac{2}{3}. \end{aligned} \tag{1.301}$$

Therefore,

$$H(y) = \ln 3 - \frac{2}{3} \ln 2. \tag{1.302}$$

(c)

We have

$$H(y|x) = - \sum_{x,y} p(x,y) \ln p(y|x). \tag{1.303}$$

By the Bayes' theorem,

$$\begin{aligned} p(y=0|x=0) &= \frac{p(x=0, y=0)}{p(x=0)}, \\ p(y=0|x=1) &= \frac{p(x=1, y=0)}{p(x=1)}, \\ p(y=1|x=0) &= \frac{p(x=0, y=1)}{p(x=0)}, \\ p(y=1|x=1) &= \frac{p(x=1, y=1)}{p(x=1)}. \end{aligned} \tag{1.304}$$

Then,

$$\begin{aligned} p(y=0|x=0) &= \frac{1}{2}, \\ p(y=0|x=1) &= 0, \\ p(y=1|x=0) &= \frac{1}{2}, \\ p(y=1|x=1) &= 1. \end{aligned} \tag{1.305}$$

Therefore,

$$H(y|x) = \frac{2}{3} \ln 2. \tag{1.306}$$

(d)

We have

$$H(x|y) = - \sum_{x,y} p(x,y) \ln p(x|y). \quad (1.307)$$

By the Bayes' theorem,

$$\begin{aligned} p(x=0|y=0) &= \frac{p(x=0, y=0)}{p(y=0)}, \\ p(x=0|y=1) &= \frac{p(x=0, y=1)}{p(y=1)}, \\ p(x=1|y=0) &= \frac{p(x=1, y=0)}{p(y=0)}, \\ p(x=1|y=1) &= \frac{p(x=1, y=1)}{p(y=1)}. \end{aligned} \quad (1.308)$$

Then,

$$\begin{aligned} p(x=0|y=0) &= 1, \\ p(x=0|y=1) &= \frac{1}{2}, \\ p(x=1|y=0) &= 0, \\ p(x=1|y=1) &= \frac{1}{2}. \end{aligned} \quad (1.309)$$

Therefore,

$$H(x|y) = \frac{2}{3} \ln 2. \quad (1.310)$$

(e)

We have

$$H(x,y) = - \sum_{x,y} p(x,y) \ln p(x,y). \quad (1.311)$$

Therefore,

$$H(x,y) = \ln 3. \quad (1.312)$$

(f)

We have

$$I(x, y) = - \sum_{x,y} p(x, y) \ln \frac{p(x)p(y)}{p(x, y)}. \quad (1.313)$$

The right hand side can be written as

$$H(x) + H(y) - H(x, y). \quad (1.314)$$

Therefore,

$$I(x, y) = \ln 3 - \frac{4}{3} \ln 2. \quad (1.315)$$

## 1.40

Let  $\lambda_1, \dots, \lambda_M$  and  $x_1, \dots, x_M$  be numbers such that

$$\begin{aligned} \sum_{m=1}^M \lambda_m &= 1, \\ \lambda_m &\geq 0, \\ x_m &> 0. \end{aligned} \quad (1.316)$$

By the Jensen's inequality,

$$\sum_{m=1}^M \lambda_m \ln x_m \leq \ln \left( \sum_{m=1}^M \lambda_m x_m \right), \quad (1.317)$$

so that

$$\prod_{m=1}^M x_m^{\lambda_m} \leq \sum_{m=1}^M \lambda_m x_m. \quad (1.318)$$

Substituting

$$\lambda_m = \frac{1}{M} \quad (1.319)$$

to the inequality gives

$$\left( \prod_{m=1}^M x_m \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{m=1}^M x_m. \quad (1.320)$$

## 1.41

Let  $\mathbf{x}$  and  $\mathbf{y}$  be variables.

(a)

We have

$$I(\mathbf{x}, \mathbf{y}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y}. \quad (1.321)$$

By the Bayes' theorem, the right hand side can be written as

$$\begin{aligned} & - \iint p(\mathbf{x}, \mathbf{y}) \left( \ln p(\mathbf{x}) + \ln \frac{p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x}d\mathbf{y} \\ &= - \iint p(\mathbf{x}, \mathbf{y}) (\ln p(\mathbf{x}) - \ln p(\mathbf{x}|\mathbf{y})) d\mathbf{x}d\mathbf{y}. \end{aligned} \quad (1.322)$$

The right hand side can be written as

$$- \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x}d\mathbf{y}. \quad (1.323)$$

Therefore,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}). \quad (1.324)$$

(b)

We have

$$I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x}). \quad (1.325)$$

By (a), the right hand side can be written as

$$H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \quad (1.326)$$

Therefore,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \quad (1.327)$$

## 2 Probability Distributions

### 2.1

Let  $x$  be a variable such that

$$\begin{aligned} x &\in \{0, 1\}, \\ p(x) &= \mu^x(1 - \mu)^{1-x}. \end{aligned} \tag{2.1}$$

(a)

We have

$$\sum_x p(x) = 1 - \mu + \mu. \tag{2.2}$$

Therefore,

$$\sum_x p(x) = 1. \tag{2.3}$$

(b)

We have

$$\begin{aligned} \text{E } x &= \sum_x x p(x), \\ \text{E } x^2 &= \sum_x x^2 p(x), \end{aligned} \tag{2.4}$$

Then,

$$\begin{aligned} \text{E } x &= \mu, \\ \text{E } x^2 &= \mu. \end{aligned} \tag{2.5}$$

We have

$$\text{var } x = \text{E } x^2 - (\text{E } x)^2. \tag{2.6}$$

Therefore,

$$\text{var } x = \mu(1 - \mu). \tag{2.7}$$

(c)

We have

$$\text{H}(x) = - \sum_x p(x) \ln p(x). \tag{2.8}$$

Therefore,

$$H(x) = -\mu \ln \mu - (1 - \mu) \ln(1 - \mu). \quad (2.9)$$

## 2.2

Let  $x$  be a variable such that

$$\begin{aligned} x &\in \{-1, 1\}, \\ p(x) &= \left(\frac{1-\mu}{2}\right)^{\frac{1-x}{2}} \left(\frac{1+\mu}{2}\right)^{\frac{1+x}{2}}. \end{aligned} \quad (2.10)$$

(a)

We have

$$\sum_x p(x) = \frac{1-\mu}{2} + \frac{1+\mu}{2}. \quad (2.11)$$

Therefore,

$$\sum_x p(x) = 1. \quad (2.12)$$

(b)

We have

$$\begin{aligned} E x &= \sum_x x p(x), \\ E x^2 &= \sum_x x^2 p(x). \end{aligned} \quad (2.13)$$

The right hand sides can be written as

$$\begin{aligned} -\frac{1-\mu}{2} + \frac{1+\mu}{2} &= \mu, \\ \frac{1-\mu}{2} + \frac{1+\mu}{2} &= 1. \end{aligned} \quad (2.14)$$

Then,

$$\begin{aligned} E x &= \mu, \\ E x^2 &= 1. \end{aligned} \quad (2.15)$$

We have

$$\text{var } x = E x^2 - (E x)^2. \quad (2.16)$$

Therefore,

$$\text{var } x = 1 - \mu^2. \quad (2.17)$$

(c)

We have

$$H(x) = - \sum_x p(x|\mu) \ln p(x|\mu). \quad (2.18)$$

Therefore,

$$H(x) = -\frac{1-\mu}{2} \ln \frac{1-\mu}{2} - \frac{1+\mu}{2} \ln \frac{1+\mu}{2}. \quad (2.19)$$

## 2.3

(a)

We have

$$\begin{aligned} \binom{N}{n} &= \frac{N!}{n!(N-n)!}, \\ \binom{N}{n-1} &= \frac{N!}{(n-1)!(N-n+1)!} \end{aligned} \quad (2.20)$$

Then,

$$\binom{N}{n} + \binom{N}{n-1} = \frac{(N-n+1)N! + nN!}{n!(N-n+1)!}. \quad (2.21)$$

The right hand side can be written as

$$\frac{(N+1)!}{n!(N+1-n)!} = \binom{N+1}{n}. \quad (2.22)$$

Therefore,

$$\binom{N}{n} + \binom{N}{n-1} = \binom{N+1}{n}. \quad (2.23)$$

(b)

We have

$$1+x = \sum_{n=0}^1 \binom{1}{n} x^n. \quad (2.24)$$

Let us assume that

$$(1+x)^N = \sum_{n=0}^N \binom{N}{n} x^n. \quad (2.25)$$

Then,

$$(1+x)^{N+1} = \sum_{n=0}^N \binom{N}{n} x^n + \sum_{n=0}^N \binom{N}{n} x^{n+1}. \quad (2.26)$$

By (a), the right hand side can be written as

$$\sum_{n=0}^N \binom{N}{n} x^n + \sum_{n=1}^{N+1} \binom{N}{n-1} x^n = 1 + x^{N+1} + \sum_{n=1}^N \binom{N+1}{n} x^n. \quad (2.27)$$

Then,

$$(1+x)^{N+1} = \sum_{n=0}^{N+1} \binom{N+1}{n} x^n. \quad (2.28)$$

Therefore, the assumption is proved by induction on  $N$ .

**(c)**

Let  $n$  be a variable such that

$$p(n) = \binom{N}{n} \mu^n (1-\mu)^{N-n}. \quad (2.29)$$

Then,

$$\sum_{n=0}^N p(n) = \sum_{n=0}^N \binom{N}{n} \mu^n (1-\mu)^{N-n}. \quad (2.30)$$

By (b), the right hand side can be written as

$$(1-\mu)^N \sum_{n=0}^N \binom{N}{n} \left(\frac{\mu}{1-\mu}\right)^n = (1-\mu)^N \left(1 + \frac{\mu}{1-\mu}\right)^N. \quad (2.31)$$

Therefore,

$$\sum_{n=0}^N p(n) = 1. \quad (2.32)$$

## 2.4

Let  $n$  be a variable such that

$$p(n) = \binom{N}{n} \mu^n (1 - \mu)^{N-n}. \quad (2.33)$$

(a)

We have

$$\mathbb{E} n = \sum_{n=0}^N n \binom{N}{n} \mu^n (1 - \mu)^{N-n}. \quad (2.34)$$

By 2.3(c),

$$\sum_{n=0}^N \binom{N}{n} \mu^n (1 - \mu)^{N-n} = 1. \quad (2.35)$$

Taking the derivative with respect to  $\mu$  gives

$$\sum_{n=0}^N n \binom{N}{n} \mu^{n-1} (1 - \mu)^{N-n} - \sum_{n=0}^N (N - n) \binom{N}{n} \mu^n (1 - \mu)^{N-n-1} = 0. \quad (2.36)$$

The first term of the left hand side can be written as

$$\frac{1}{\mu} \sum_{n=0}^N np(n) = \frac{1}{\mu} \mathbb{E} n. \quad (2.37)$$

Since

$$(N - n) \binom{N}{n} = N \binom{N-1}{n}, \quad (2.38)$$

the second term can be written as

$$-N \sum_{n=0}^{N-1} \binom{N-1}{n} \mu^n (1 - \mu)^{N-n-1} = -N. \quad (2.39)$$

Therefore,

$$\mathbb{E} n = N\mu. \quad (2.40)$$

(b)

By 2.3(c),

$$\sum_{n=0}^N \binom{N}{n} \mu^n (1-\mu)^{N-n} = 1. \quad (2.41)$$

Taking the second derivative with respect to  $\mu$  gives

$$\begin{aligned} & \sum_{n=0}^N n(n-1) \binom{N}{n} \mu^{n-2} (1-\mu)^{N-n} \\ & - 2 \sum_{n=0}^N n(N-n) \binom{N}{n} \mu^{n-1} (1-\mu)^{N-n-1} \\ & + \sum_{n=0}^N (N-n)(N-n-1) \binom{N}{n} \mu^n (1-\mu)^{N-n-2} = 0. \end{aligned} \quad (2.42)$$

The first term of the left hand side can be written as

$$\frac{1}{\mu^2} \sum_{n=0}^N n(n-1)p(n) = \frac{1}{\mu^2} E n(n-1). \quad (2.43)$$

Since

$$\begin{aligned} n(N-n) \binom{N}{n} &= N(N-1) \binom{N-2}{n-1}, \\ (N-n)(N-n-1) \binom{N}{n} &= N(N-1) \binom{N-2}{n}, \end{aligned} \quad (2.44)$$

the second and third terms can be written as

$$\begin{aligned} -2N(N-1) \sum_{n=1}^{N-1} \binom{N-2}{n-1} \mu^{n-1} (1-\mu)^{N-n-1} &= -2N(N-1), \\ N(N-1) \sum_{n=0}^N \binom{N-2}{n} \mu^n (1-\mu)^{N-n-2} &= N(N-1). \end{aligned} \quad (2.45)$$

Then,

$$E n(n-1) = N(N-1)\mu^2. \quad (2.46)$$

We have

$$\text{var } n = E n(n-1) + E n - (E n)^2. \quad (2.47)$$

Therefore,

$$\text{var } n = N\mu(1-\mu). \quad (2.48)$$

## 2.5

We have

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1} \exp(-x) dx \int_0^\infty y^{b-1} \exp(-y) dy. \quad (2.49)$$

By the transformation

$$t = x + y, \quad (2.50)$$

the right hand side can be written as

$$\begin{aligned} & \int_0^\infty x^{a-1} \left( \int_x^\infty (t-x)^{b-1} \exp(-t) dt \right) dx \\ &= \int_0^\infty \left( \int_0^t x^{a-1} (t-x)^{b-1} dx \right) \exp(-t) dt. \end{aligned} \quad (2.51)$$

By the transformation

$$x = t\mu, \quad (2.52)$$

the right hand side can be written as

$$\begin{aligned} & \int_0^\infty \left( \int_0^1 (t\mu)^{a-1} t^{b-1} (1-\mu)^{b-1} t d\mu \right) \exp(-t) dt \\ &= \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu \int_0^\infty t^{a+b-1} \exp(-t) dt. \end{aligned} \quad (2.53)$$

Then,

$$\Gamma(a)\Gamma(b) = \Gamma(a+b) \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu. \quad (2.54)$$

Therefore,

$$\int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \quad (2.55)$$

## 2.6

Let  $\mu$  be a variable such that

$$p(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \quad (2.56)$$

(a)

We have

$$\begin{aligned} \text{E } \mu &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu, \\ \text{E } \mu^2 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu. \end{aligned} \quad (2.57)$$

By 2.5,

$$\begin{aligned} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}, \\ \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu &= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}. \end{aligned} \quad (2.58)$$

Therefore,

$$\begin{aligned} \text{E } \mu &= \frac{a}{a+b}, \\ \text{E } \mu^2 &= \frac{a(a+1)}{(a+b)(a+b+1)}. \end{aligned} \quad (2.59)$$

(b)

We have

$$\text{var } \mu = \text{E } \mu^2 - (\text{E } \mu)^2. \quad (2.60)$$

By (a), the right hand side can be written as

$$\frac{a(a+1)}{(a+b)(a+b+1)} - \left( \frac{a}{a+b} \right)^2 = \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)}. \quad (2.61)$$

Therefore,

$$\text{var } \mu = \frac{ab}{(a+b)^2(a+b+1)}. \quad (2.62)$$

(c)

Setting the derivative of  $p$  with respect to  $\mu$  to zero gives

$$0 = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \left( \frac{a-1}{\mu} - \frac{b-1}{1-\mu} \right). \quad (2.63)$$

Therefore,

$$\text{mode } \mu = \frac{a-1}{a+b-2}. \quad (2.64)$$

## 2.7

Let  $m$  and  $l$  be variables such that

$$\begin{aligned} p(m, l | \mu) &= \binom{m+l}{m} \mu^m (1-\mu)^l, \\ p(\mu) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \end{aligned} \quad (2.65)$$

By 2.6,

$$E\mu = \frac{a}{a+b}. \quad (2.66)$$

Setting the derivative of  $p(m, l | \mu)$  with respect to  $\mu$  to zero gives

$$0 = \binom{m+l}{m} \mu^m (1-\mu)^l \left( \frac{m}{\mu} + \frac{l}{1-\mu} \right). \quad (2.67)$$

Then, the maximum likelihood solution for  $\mu$  is given by

$$\mu_{ML} = \frac{m}{m+l}. \quad (2.68)$$

By the Bayes' theorem,

$$p(\mu | m, l) p(m, l) = p(m, l | \mu) p(\mu). \quad (2.69)$$

Then, by 2.5,

$$p(\mu | m, l) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1} (1-\mu)^{l+b-1}. \quad (2.70)$$

The, by 2.6,

$$E(\mu | m, l) = \frac{m+a}{m+l+a+b}. \quad (2.71)$$

Therefore,

$$E(\mu | m, l) = \lambda \mu_{ML} + (1-\lambda) E\mu, \quad (2.72)$$

where

$$\lambda = \frac{m+l}{m+l+a+b}. \quad (2.73)$$

## 2.8

Let  $x$  and  $y$  be variables.

(a)

By the definition,

$$\mathbb{E} x = \int xp(x)dx. \quad (2.74)$$

The right hand side can be written as

$$\int x \left( \int p(x,y)dy \right) dx = \int \left( \int xp(x|y)dx \right) p(y)dy. \quad (2.75)$$

Therefore,

$$\mathbb{E} x = \mathbb{E}_y (\mathbb{E}_x(x|y)). \quad (2.76)$$

(b)

By the definition,

$$\text{var } x = \mathbb{E} (x - \mathbb{E} x)^2. \quad (2.77)$$

By (a), the right hand side can be written as

$$\mathbb{E}_y (\mathbb{E}_x ((x - \mathbb{E} x)^2|y)) = \mathbb{E}_y (\mathbb{E}_x ((x - \mathbb{E}_x(x|y) + \mathbb{E}_x(x|y) - \mathbb{E} x)^2|y)). \quad (2.78)$$

The right hand side can be written as

$$\begin{aligned} & \mathbb{E}_y (\mathbb{E}_x ((x - \mathbb{E}_x(x|y))^2|y)) \\ & + 2\mathbb{E}_y (((\mathbb{E}_x(x|y) - \mathbb{E} x) \mathbb{E}_x (x - \mathbb{E}_x(x|y))|y)) \\ & + \mathbb{E}_y ((\mathbb{E}_x(x|y) - \mathbb{E} x)^2|y). \end{aligned} \quad (2.79)$$

Let us look at each term of the right hand side. By the definition, the first term can be written as  $\mathbb{E}_y (\text{var}_x(x|y))$ . The second term can be written as

$$2\mathbb{E}_y ((\mathbb{E}_x(x|y) - \mathbb{E} x) (\mathbb{E}_x(x|y) - \mathbb{E}_x(x|y))) = 0. \quad (2.80)$$

By (a), the third term can be written as

$$\mathbb{E}_y (\mathbb{E}_x(x|y) - \mathbb{E}_y (\mathbb{E}_x(x|y)))^2 = \text{var}_y (\mathbb{E}_x(x|y)). \quad (2.81)$$

Therefore,

$$\text{var } x = \mathbb{E}_y (\text{var}_x(x|y)) + \text{var}_y (\mathbb{E}_x(x|y)). \quad (2.82)$$

## 2.9 (Incomplete)

For a vector  $\mu$  in 2 dimensions, by 2.5,

$$\int_{\substack{\mu_1+\mu_2=1 \\ \mu_1 \geq 0, \mu_2 \geq 0}} \mu_1^{\alpha_1-1} \mu_2^{\alpha_2-1} d\mu = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}.$$

For a vector  $\mu$  in  $M$  dimensions, let us assume that

$$\int_{\substack{\sum_{m=1}^M \mu_m=1 \\ \mu_m \geq 0}} \prod_{m=1}^M \mu_m^{\alpha_m-1} d\mu = \frac{\prod_{m=1}^M \Gamma(\alpha_m)}{\Gamma(\sum_{m=1}^M \alpha_m)}.$$

Under the constraint

$$\sum_{m=1}^{M+1} \mu_m = 1, \quad (2.83)$$

we have

$$\int_0^c \prod_{m=1}^{M+1} \mu_m^{\alpha_m-1} d\mu_{M+1} = \left( \prod_{m=1}^{M-1} \mu_m^{\alpha_m-1} \right) \int_0^c \mu_{M+1}^{\alpha_{M+1}-1} (c - \mu_{M+1})^{\alpha_M-1} d\mu_{M+1}, \quad (2.84)$$

where

$$c = 1 - \sum_{m=1}^{M-1} \mu_m. \quad (2.85)$$

By the transformation

$$\mu'_{M+1} = \frac{\mu_{M+1}}{c}, \quad (2.86)$$

the integral of the right hand side can be written as

$$\begin{aligned} & \int_0^1 (c\mu'_{M+1})^{\alpha_{M+1}-1} (c(1 - \mu'_{M+1}))^{\alpha_M-1} cd\mu'_{M+1} \\ &= c^{\alpha_M + \alpha_{M+1}-1} \int_0^1 \mu'_{M+1}^{\alpha_{M+1}-1} (1 - \mu'_{M+1})^{\alpha_M-1} d\mu'_{M+1}. \end{aligned} \quad (2.87)$$

By 2.5, the integral of the right hand side can be written as

$$\frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \quad (2.88)$$

Then,

$$\int_0^c \prod_{m=1}^{M+1} \mu_m^{\alpha_m-1} d\mu_{M+1} = \left( \prod_{m=1}^{M-1} \mu_m^{\alpha_m-1} \right) c^{\alpha_M + \alpha_{M+1}-1} \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \quad (2.89)$$

For a vector  $\boldsymbol{\mu}$  in  $M$  dimensions, by the assumption,

$$\int_{\sum_{m=1}^M \mu_m=1 \atop \mu_m \geq 0} \prod_{m=1}^{M-1} \mu_m^{\alpha_m-1} \mu_M^{\alpha_M + \alpha_{M+1}-1} d\boldsymbol{\mu} = \frac{\left( \prod_{m=1}^{M-1} \Gamma(\alpha_m) \right) \Gamma(\alpha_M + \alpha_{M+1})}{\Gamma(\sum_{m=1}^{M+1} \alpha_m)}.$$

Then, for a vector  $\boldsymbol{\mu}$  in  $M+1$  dimensions,

$$\int_{\sum_{m=1}^{M+1} \mu_m=1 \atop \mu_m \geq 0} \prod_{m=1}^{M+1} \mu_m^{\alpha_m-1} d\boldsymbol{\mu} = \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})} \frac{\left( \prod_{m=1}^{M-1} \Gamma(\alpha_m) \right) \Gamma(\alpha_M + \alpha_{M+1})}{\Gamma(\sum_{m=1}^{M+1} \alpha_m)}?$$

The right hand side can be written as

$$\frac{\prod_{m=1}^{M+1} \Gamma(\alpha_m)}{\Gamma(\sum_{m=1}^{M+1} \alpha_m)}. \quad (2.90)$$

Therefore, the assumption is proved by induction on  $M$ .

## 2.10

Let  $\boldsymbol{\mu}$  be a vector such that

$$p(\boldsymbol{\mu}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k-1}. \quad (2.91)$$

Then, by the definition,

$$\begin{aligned} \mathbb{E} \mu_k &= \int \mu_k p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \\ \mathbb{E} \mu_k^2 &= \int \mu_k^2 p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \\ \mathbb{E} \mu_k \mu_{k'} &= \int \mu_k \mu_{k'} p(\boldsymbol{\mu}) d\boldsymbol{\mu}. \end{aligned} \quad (2.92)$$

Let  $k \neq k'$ . Then, by 2.9, the right hand sides can be written as

$$\begin{aligned} \frac{\Gamma(\sum_{k=1}^K \alpha_k) \frac{\Gamma(\alpha_k+1)}{\Gamma(\alpha_k)} \prod_{k=1}^K \Gamma(\alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} &= \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}, \\ \frac{\Gamma(\sum_{k=1}^K \alpha_k) \frac{\Gamma(\alpha_k+2)}{\Gamma(\alpha_k)} \prod_{k=1}^K \Gamma(\alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k) \Gamma(\sum_{k=1}^K \alpha_k + 2)} &= \frac{\alpha_k(\alpha_k + 1)}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}, \quad (2.93) \\ \frac{\Gamma(\sum_{k=1}^K \alpha_k) \frac{\Gamma(\alpha_k+1)\Gamma(\alpha_{k'}+1)}{\Gamma(\alpha_k)\Gamma(\alpha_{k'})} \prod_{k=1}^K \Gamma(\alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k) \Gamma(\sum_{k=1}^K \alpha_k + 2)} &= \frac{\alpha_k \alpha_{k'}}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}. \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E} \mu_k &= \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}, \\ \mathbb{E} \mu_k^2 &= \frac{\alpha_k(\alpha_k + 1)}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}, \quad (2.94) \\ \mathbb{E} \mu_k \mu_{k'} &= \frac{\alpha_k \alpha_{k'}}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}. \end{aligned}$$

Since

$$\begin{aligned} \text{var } \mu_k &= \mathbb{E} \mu_k^2 - (\mathbb{E} \mu_k)^2, \\ \text{cov}(\mu_k, \mu_{k'}) &= \mathbb{E} \mu_k \mu_{k'} - \mathbb{E} \mu_k \mathbb{E} \mu_{k'}, \quad (2.95) \end{aligned}$$

we have

$$\begin{aligned} \text{var } \mu_k &= \frac{\alpha_k \left( \left( \sum_{k=1}^K \alpha_k \right) - \alpha_k \right)}{(\sum_{k=1}^K \alpha_k)^2 (\sum_{k=1}^K \alpha_k + 1)}, \\ \text{cov}(\mu_k, \mu_{k'}) &= -\frac{\alpha_k \alpha_{k'}}{(\sum_{k=1}^K \alpha_k)^2 (\sum_{k=1}^K \alpha_k + 1)}. \quad (2.96) \end{aligned}$$

## 2.11

Let  $\boldsymbol{\mu}$  be a variable such that

$$p(\boldsymbol{\mu}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}. \quad (2.97)$$

Then, by the definition,

$$\mathbb{E} \ln \mu_k = \int (\ln \mu_k) p(\boldsymbol{\mu}) d\boldsymbol{\mu}. \quad (2.98)$$

Since

$$\frac{\partial}{\partial \alpha_k} p(\boldsymbol{\mu}) = \left( \frac{\Gamma' \left( \sum_{k=1}^K \alpha_k \right)}{\Gamma \left( \sum_{k=1}^K \alpha_k \right)} - \frac{\Gamma'(\alpha_k)}{\Gamma(\alpha_k)} + \ln \mu_k \right) p(\boldsymbol{\mu}), \quad (2.99)$$

we have

$$\mathbb{E} \ln \mu_k = \frac{\partial}{\partial \alpha_k} \int p(\boldsymbol{\mu}) d\boldsymbol{\mu} + \left( \psi(\alpha_k) - \psi \left( \sum_{k=1}^K \alpha_k \right) \right) \int p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \quad (2.100)$$

where

$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \quad (2.101)$$

Therefore,

$$\mathbb{E} \ln \mu_k = \psi(\alpha_k) - \psi \left( \sum_{k=1}^K \alpha_k \right). \quad (2.102)$$

## 2.12

Let  $x$  be a variable such that

$$p(x) = \frac{1}{b-a}, \quad (2.103)$$

where  $a < b$ . Then

$$\int_a^b p(x) dx = 1. \quad (2.104)$$

Then, by the definition,

$$\begin{aligned} \mathbb{E} x &= \frac{1}{b-a} \int_a^b x dx, \\ \mathbb{E} x^2 &= \frac{1}{b-a} \int_a^b x^2 dx. \end{aligned} \quad (2.105)$$

Then,

$$\begin{aligned}\mathrm{E} x &= \frac{1}{2}(a + b), \\ \mathrm{E} x^2 &= \frac{1}{3}(a^2 + ab + b^2).\end{aligned}\tag{2.106}$$

Since

$$\mathrm{var} x = \mathrm{E} x^2 - (\mathrm{E} x)^2,\tag{2.107}$$

we have

$$\mathrm{var} x = \frac{1}{12}(b - a)^2.\tag{2.108}$$

## 2.13

Let  $\mathbf{x}$  be a variable in  $D$  dimensions and let

$$\begin{aligned}p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}).\end{aligned}\tag{2.109}$$

Then, by the definition, the Kulleback-Leibler divergence is given by

$$\mathrm{KL}(p||q) = - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}.\tag{2.110}$$

Note that

$$\ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \frac{(2\pi)^{-\frac{D}{2}} |\det \mathbf{L}|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}))}{(2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))}.\tag{2.111}$$

The right hand side can be written as

$$\frac{1}{2} \ln \left| \frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}} \right| + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}).\tag{2.112}$$

Then, the integral can be written as

$$\begin{aligned}&\frac{1}{2} \ln \left| \frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}} \right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &+ \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &- \frac{1}{2} \int (\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.\end{aligned}\tag{2.113}$$

Let us look at the integral of each term. The integral of the first term is 1. Since

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x} = \Sigma, \quad (2.114)$$

we have

$$\int (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x} = \text{tr } \Sigma. \quad (2.115)$$

Then, the integral of the second term can be written as

$$\text{tr}(\Sigma^{-1}\Sigma) = D. \quad (2.116)$$

Since

$$(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) = (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m}), \quad (2.117)$$

the integral of the third term can be written as

$$\begin{aligned} & \int (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x} \\ & + 2(\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} \int (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x} \\ & + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m}) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) d\mathbf{x} \\ & = \text{tr}(\mathbf{L}^{-1}\Sigma) + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m}). \end{aligned} \quad (2.118)$$

Therefore,

$$\text{KL}(p||q) = \frac{1}{2} \left( \ln \left| \frac{\det \mathbf{L}}{\det \Sigma} \right| - D + \text{tr}(\mathbf{L}^{-1}\Sigma) + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m}) \right). \quad (2.119)$$

## 2.14

Let  $\mathbf{x}$  be a variable in  $D$  dimensions. By the definition, the entropy is given by

$$H(\mathbf{x}) = - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \quad (2.120)$$

In order to maximise  $H(x)$  with the constraints

$$\begin{aligned} \int p(\mathbf{x})d\mathbf{x} &= 1, \\ \int \mathbf{x}p(\mathbf{x})d\mathbf{x} &= \boldsymbol{\mu}, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x})d\mathbf{x} &= \boldsymbol{\Sigma}, \end{aligned} \quad (2.121)$$

let

$$\begin{aligned} L(p) = H(\mathbf{x}) + \lambda \left( \int p(\mathbf{x})d\mathbf{x} - 1 \right) + \mathbf{l}^\top \left( \int \mathbf{x}p(\mathbf{x})d\mathbf{x} - \boldsymbol{\mu} \right) \\ + \mathbf{m}^\top \left( \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x})d\mathbf{x} - \boldsymbol{\Sigma} \right) \mathbf{m}. \end{aligned} \quad (2.122)$$

Setting the variation with respect to  $p$  to zero gives

$$0 = -\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^\top \mathbf{x} + \mathbf{m}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{m}. \quad (2.123)$$

Then,

$$p(\mathbf{x}) = \exp(-1 + \lambda + \mathbf{l}^\top \mathbf{x} + \mathbf{m}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{m}), \quad (2.124)$$

so that

$$p(\mathbf{x}) = c \exp(-(\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})^\top \mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})), \quad (2.125)$$

where

$$\begin{aligned} c &= \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}), \\ \mathbf{M} &= -(\mathbf{m}\mathbf{m}^\top)^{-1}. \end{aligned} \quad (2.126)$$

Substituting it to the constraints and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l} \quad (2.127)$$

gives

$$\begin{aligned} c \int \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= 1, \\ c \int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= \boldsymbol{\mu}, \\ c \int (\mathbf{y} + \mathbf{M}\mathbf{l}) (\mathbf{y} + \mathbf{M}\mathbf{l})^\top \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= \boldsymbol{\Sigma}. \end{aligned} \quad (2.128)$$

Since

$$\begin{aligned} \int \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \left( \Gamma\left(\frac{1}{2}\right) \right)^D, \\ \int \mathbf{y} \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \mathbf{0}, \\ \int \mathbf{y} \mathbf{y}^\top \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \Gamma\left(\frac{3}{2}\right) \left( \Gamma\left(\frac{1}{2}\right) \right)^{D-1} \mathbf{I}, \end{aligned} \quad (2.129)$$

they can be written as

$$\begin{aligned} c \left( \Gamma\left(\frac{1}{2}\right) \right)^D |\det \mathbf{M}|^{\frac{1}{2}} &= 1, \\ c(\boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \left( \Gamma\left(\frac{1}{2}\right) \right)^D |\det \mathbf{M}|^{\frac{1}{2}} &= \boldsymbol{\mu}, \\ c \left( \Gamma\left(\frac{3}{2}\right) \left( \Gamma\left(\frac{1}{2}\right) \right)^{D-1} \mathbf{M} + \mathbf{M}\mathbf{l}(\mathbf{M}\mathbf{l})^\top \left( \Gamma\left(\frac{1}{2}\right) \right)^D \right) |\det \mathbf{M}|^{\frac{1}{2}} &= \boldsymbol{\Sigma}. \end{aligned} \quad (2.130)$$

Then,

$$\begin{aligned} \lambda &= 1 - \frac{D}{2} \ln \pi - \frac{1}{2} \ln |\det \mathbf{M}|, \\ \mathbf{l} &= \mathbf{0}, \\ \mathbf{M} &= 2\boldsymbol{\Sigma}. \end{aligned} \quad (2.131)$$

Therefore,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (2.132)$$

## 2.15

Let  $\mathbf{x}$  be a variable in  $D$  dimensions such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.133)$$

Then, by the definition, the entropy is given by

$$H(\mathbf{x}) = - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \quad (2.134)$$

The right hand side can be written as

$$\begin{aligned}
& - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \left( -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln |\det \boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x} \\
& = \left( \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}| \right) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\
& + \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.
\end{aligned} \tag{2.135}$$

Let us look at each integral of the right hand side. The first integral is 1. Since

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\Sigma}, \tag{2.136}$$

we have

$$\int (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \text{tr } \boldsymbol{\Sigma}. \tag{2.137}$$

Then, the second integral can be written as

$$\text{tr} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}) = D. \tag{2.138}$$

Therefore,

$$H(\mathbf{x}) = \frac{D}{2} (1 + \ln(2\pi)) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}|. \tag{2.139}$$

## 2.16

Let  $x$  be a variable such that

$$x = x_1 + x_2, \tag{2.140}$$

where

$$\begin{aligned}
p(x_1) &= \mathcal{N}(x_1|\mu_1, \tau_1^{-1}), \\
p(x_2) &= \mathcal{N}(x_2|\mu_2, \tau_2^{-1}).
\end{aligned} \tag{2.141}$$

By marginalisation,

$$p(x) = \int_{-\infty}^{\infty} p(x|x_2)p(x_2)dx_2. \tag{2.142}$$

The right hand side can be written as

$$\begin{aligned} & \int_{-\infty}^{\infty} \mathcal{N}(x|\mu_1 + x_2, \tau_1^{-1}) \mathcal{N}(x_2|\mu_2, \tau_2^{-1}) dx_2 \\ &= \int_{-\infty}^{\infty} \left( \frac{\tau_1}{2\pi} \right)^{\frac{1}{2}} \exp \left( -\frac{\tau_1}{2}(x - \mu_1 - x_2)^2 \right) \left( \frac{\tau_2}{2\pi} \right)^{\frac{1}{2}} \exp \left( -\frac{\tau_2}{2}(x_2 - \mu_2)^2 \right) dx_2. \end{aligned} \quad (2.143)$$

The logarithm of the integrand except the terms independent of  $x$  and  $z$  is given by

$$\begin{aligned} & -\frac{\tau_1 + \tau_2}{2} \left( x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 \\ &+ \frac{\tau_1 + \tau_2}{2} \left( \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 \\ &= -\frac{\tau_1 + \tau_2}{2} \left( x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2. \end{aligned} \quad (2.144)$$

Then,

$$p(x) = \mathcal{N}(x|\mu_1 + \mu_2, \tau_1^{-1} + \tau_2^{-1}). \quad (2.145)$$

Therefore, by 1.35,

$$H(x) = \frac{1}{2} (1 + \ln(2\pi) + \ln(\tau_1^{-1} + \tau_2^{-1})). \quad (2.146)$$

## 2.17

Let  $\Sigma$  be a matrix and

$$\begin{aligned} \mathbf{S} &= \frac{1}{2} (\Sigma^{-1} + (\Sigma^{-1})^\top), \\ \mathbf{A} &= \frac{1}{2} (\Sigma^{-1} - (\Sigma^{-1})^\top). \end{aligned} \quad (2.147)$$

Then,

$$\Sigma^{-1} = \mathbf{S} + \mathbf{A}, \quad (2.148)$$

so that

$$(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S} (\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.149)$$

The second term of the right hand side can be written as

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma}^{-1})^\top (\mathbf{x} - \boldsymbol{\mu}). \quad (2.150)$$

The second term of the right hand side can be written as

$$-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))^\top (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (2.151)$$

Then,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = 0. \quad (2.152)$$

Therefore,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S}(\mathbf{x} - \boldsymbol{\mu}). \quad (2.153)$$

## 2.18

(a)

Let  $\boldsymbol{\Sigma}$  be a  $D \times D$  real symmetric matrix such that

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.154)$$

where  $\mathbf{u}_1, \dots, \mathbf{u}_D$  are unit vectors. Then,

$$\overline{\mathbf{u}_d}^\top \boldsymbol{\Sigma} \mathbf{u}_d = \lambda_d, \quad (2.155)$$

where  $\overline{\mathbf{u}_d}$  is the conjugate of  $\mathbf{u}_d$ . Since  $\boldsymbol{\Sigma}$  is real and symmetric, the left hand side can be written as

$$\overline{\mathbf{u}_d}^\top \overline{\boldsymbol{\Sigma}}^\top \mathbf{u}_d = (\overline{\boldsymbol{\Sigma} \mathbf{u}_d})^\top \mathbf{u}_d. \quad (2.156)$$

The right hand side can be written as

$$\overline{\lambda_d} \overline{\mathbf{u}_d}^\top \mathbf{u}_d = \overline{\lambda_d}. \quad (2.157)$$

Therefore,

$$\lambda_d = \overline{\lambda_d}. \quad (2.158)$$

(b)

For  $d \neq d'$ , taking the inner product with  $\mathbf{u}'_d$  on both sides of

$$\Sigma \mathbf{u}_d = \lambda_d \mathbf{u}_d \quad (2.159)$$

gives

$$\mathbf{u}'_{d'} \Sigma \mathbf{u}_d = \lambda_d \mathbf{u}'_{d'} \mathbf{u}_d. \quad (2.160)$$

Since  $\Sigma$  is symmetric, the left hand side can be written as

$$\mathbf{u}'_{d'} \Sigma^\top \mathbf{u}_d = (\Sigma \mathbf{u}_{d'})^\top \mathbf{u}_d. \quad (2.161)$$

The right hand side can be written as  $\lambda_{d'} \mathbf{u}'_{d'} \mathbf{u}_d$ . Then,

$$\lambda_d \mathbf{u}'_{d'} \mathbf{u}_d = \lambda_{d'} \mathbf{u}'_{d'} \mathbf{u}_d. \quad (2.162)$$

Therefore, if  $\lambda_d \neq \lambda_{d'}$ , then

$$\mathbf{u}'_{d'} \mathbf{u}_d = 0. \quad (2.163)$$

## 2.19

Let  $\Sigma$  be a  $D \times D$  real symmetric matrix such that

$$\Sigma \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.164)$$

where  $\mathbf{u}_1, \dots, \mathbf{u}_D$  are unit vectors. Let

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \dots \mathbf{u}_D]. \end{aligned} \quad (2.165)$$

Then,

$$\Sigma \mathbf{U} = \mathbf{U} \Lambda. \quad (2.166)$$

By 2.18,

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}. \quad (2.167)$$

Then,

$$\begin{aligned} \Sigma &= \mathbf{U} \Lambda \mathbf{U}^\top, \\ \Sigma^{-1} &= \mathbf{U} \Lambda^{-1} \mathbf{U}^\top, \end{aligned} \quad (2.168)$$

Therefore,

$$\begin{aligned} \Sigma &= \sum_{d=1}^D \lambda_d \mathbf{u}_d \mathbf{u}_d^\top, \\ \Sigma^{-1} &= \sum_{d=1}^D \frac{1}{\lambda_d} \mathbf{u}_d \mathbf{u}_d^\top. \end{aligned} \quad (2.169)$$

## 2.20

Let  $\Sigma$  be a  $D \times D$  real symmetric matrix such that

$$\Sigma \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.170)$$

where  $u_1, \dots, u_D$  are unit vectors. Let

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.171)$$

By 2.19,

$$\mathbf{a}^\top \Sigma \mathbf{a} = \mathbf{b}^\top \Lambda \mathbf{b}, \quad (2.172)$$

where

$$\mathbf{b} = \mathbf{U}^\top \mathbf{a}. \quad (2.173)$$

The right hand side can be written as  $\sum_{d=1}^D \lambda_d b_d^2$ . Therefore, the necessary and sufficient condition for

$$\mathbf{a}^\top \Sigma \mathbf{a} > 0 \quad (2.174)$$

for any real vector  $\mathbf{a}$  is

$$\lambda_d > 0. \quad (2.175)$$

## 2.21

Let  $\Sigma$  be a  $D \times D$  real symmetric matrix. Then the number of independent parameters is  $\frac{D(D+1)}{2}$ .

## 2.22

Let  $\Sigma$  be a  $D \times D$  symmetric matrix and

$$\Sigma \Lambda = \mathbf{I}. \quad (2.176)$$

Taking the transpose of the both sides gives

$$\Lambda^\top \Sigma = \mathbf{I}. \quad (2.177)$$

Therefore,

$$\Lambda^\top = \Lambda. \quad (2.178)$$

## 2.23

Let  $\Sigma$  be a  $D \times D$  real symmetric matrix such that

$$\Sigma \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.179)$$

where  $u_1, \dots, u_D$  are unit vectors. Let

$$\begin{aligned} \Lambda' &= \text{diag} \left( \lambda_1^{-\frac{1}{2}}, \dots, \lambda_D^{-\frac{1}{2}} \right), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.180)$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{U} \Lambda' \Lambda'^\top \mathbf{U}^\top (\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \quad (2.181)$$

By the transformation

$$\mathbf{y} = \Lambda'^\top \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu}) \quad (2.182)$$

and the property

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}, \quad (2.183)$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left| \det \left( \mathbf{U} \Lambda'^{-1} \right) \right| d\mathbf{y} = |\det \Sigma|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \quad (2.184)$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det \Sigma|^{\frac{1}{2}} \Delta^D V_D, \quad (2.185)$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \quad (2.186)$$

## 2.24

Let  $\mathbf{A}$  be a square matrix and  $\mathbf{D}$  be an invertible matrix. We have

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{BD}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{BD}^{-1}\mathbf{C} & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}. \quad (2.187)$$

The right hand side can be written as

$$\begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}. \quad (2.188)$$

Then,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{M} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}, \quad (2.189)$$

where

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}. \quad (2.190)$$

Therefore,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & -\mathbf{B}\mathbf{D}^{-1}\mathbf{M} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}. \quad (2.191)$$

## 2.25

Let  $\mathbf{x}$  be a variable such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2.192)$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \\ \mathbf{x}_c \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} & \boldsymbol{\Sigma}_{ac} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} & \boldsymbol{\Sigma}_{bc} \\ \boldsymbol{\Sigma}_{ca} & \boldsymbol{\Sigma}_{cb} & \boldsymbol{\Sigma}_{cc} \end{bmatrix}.$$

Let

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \quad (2.193)$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} & \boldsymbol{\Lambda}_{ac} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} & \boldsymbol{\Lambda}_{bc} \\ \boldsymbol{\Lambda}_{ca} & \boldsymbol{\Lambda}_{cb} & \boldsymbol{\Lambda}_{cc} \end{bmatrix}.$$

Then, the logarithm of  $p(\mathbf{x})$  except the terms independent of  $\mathbf{x}_a$  can be written as

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa} (\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \\ & - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ac} (\mathbf{x}_c - \boldsymbol{\mu}_c) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{ba} (\mathbf{x}_a - \boldsymbol{\mu}_a) \\ & - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{ca} (\mathbf{x}_a - \boldsymbol{\mu}_a). \end{aligned} \quad (2.194)$$

Except the terms independent of  $\mathbf{x}_a$ , it can be written as

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\top \boldsymbol{\Sigma}_{a|b,c}^{-1} (\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}), \quad (2.195)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} (\mathbf{x}_c - \boldsymbol{\mu}_c), \\ \boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \quad (2.196)$$

Then,

$$p(\mathbf{x}_a | \mathbf{x}_b, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}). \quad (2.197)$$

By marginalisation,

$$p(\mathbf{x}_a | \mathbf{x}_b) = \int p(\mathbf{x}_a | \mathbf{x}_b, \mathbf{x}_c) p(\mathbf{x}_c) d\mathbf{x}_c. \quad (2.198)$$

The integrand of the right hand side except the terms independet of  $\mathbf{x}_c$  can be written as

$$\begin{aligned} &-\frac{1}{2} (\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\top \boldsymbol{\Sigma}_{a|b,c}^{-1} (\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}) - \frac{1}{2} (\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{cc} (\mathbf{x}_c - \boldsymbol{\mu}_c) \\ &= -\frac{1}{2} \mathbf{v}^\top \mathbf{M} \mathbf{v}, \end{aligned} \quad (2.199)$$

where

$$\begin{aligned} \mathbf{v} &= \begin{bmatrix} \mathbf{x}_c - \boldsymbol{\mu}_c \\ \mathbf{x}_a - \boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \end{bmatrix}, \\ \mathbf{M} &= \begin{bmatrix} \boldsymbol{\Lambda}_{cc} + \boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} & \boldsymbol{\Lambda}_{ac}^\top \\ \boldsymbol{\Lambda}_{ac} & \boldsymbol{\Lambda}_{aa} \end{bmatrix}. \end{aligned} \quad (2.200)$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{cc}^{-1} & -\boldsymbol{\Lambda}_{cc}^{-1} \boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1} \\ -\boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} \boldsymbol{\Lambda}_{cc}^{-1} & \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} \boldsymbol{\Lambda}_{cc}^{-1} \boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1} \end{bmatrix}. \quad (2.201)$$

Therefore,

$$p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \quad (2.202)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b), \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} \boldsymbol{\Lambda}_{cc}^{-1} \boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \quad (2.203)$$

## 2.26

Let  $\mathbf{A}$  be a square matrix and  $\mathbf{C}$  be an invertible matrix. By 2.24,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{D} & -\mathbf{C}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & \mathbf{MBC} \\ \mathbf{CDM} & -\mathbf{C} + \mathbf{CDMBC} \end{bmatrix}, \quad (2.204)$$

where

$$\mathbf{M} = (\mathbf{A} + \mathbf{BCD})^{-1}. \quad (2.205)$$

By 2.24,

$$\begin{bmatrix} -\mathbf{C}^{-1} & \mathbf{D} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}^{-1} = \begin{bmatrix} -\mathbf{N} & \mathbf{NDA}^{-1} \\ \mathbf{A}^{-1}\mathbf{BN} & \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BNDA}^{-1} \end{bmatrix}, \quad (2.206)$$

where

$$\mathbf{N} = (\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}. \quad (2.207)$$

Therefore,

$$\mathbf{M} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BNDA}^{-1}. \quad (2.208)$$

## 2.27

(a)

Let  $\mathbf{x}$  and  $\mathbf{z}$  be two variables. By the definition,

$$E(\mathbf{x} + \mathbf{z}) = \int \int (\mathbf{x} + \mathbf{z}) p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \quad (2.209)$$

The right hand side can be written as

$$\int \mathbf{x} \left( \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \right) d\mathbf{x} + \int \mathbf{z} \left( \int p(\mathbf{x}, \mathbf{z}) d\mathbf{x} \right) d\mathbf{z} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} + \int \mathbf{z} p(\mathbf{z}) d\mathbf{z}. \quad (2.210)$$

Therefore,

$$E(\mathbf{x} + \mathbf{z}) = E \mathbf{x} + E \mathbf{z}. \quad (2.211)$$

(b)

Let  $\mathbf{x}$  and  $\mathbf{z}$  be two independent variables. By the definition,

$$\text{cov}(\mathbf{x} + \mathbf{z}) = \int \int (\mathbf{x} + \mathbf{z} - E(\mathbf{x} + \mathbf{z})) (\mathbf{x} + \mathbf{z} - E(\mathbf{x} + \mathbf{z}))^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \quad (2.212)$$

The right hand side can be written as

$$\begin{aligned} & \int \int (\mathbf{x} - E\mathbf{x}) (\mathbf{x} - E\mathbf{x})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} + \int \int (\mathbf{x} - E\mathbf{x}) (\mathbf{z} - E\mathbf{z})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ & + \int \int (\mathbf{z} - E\mathbf{z}) (\mathbf{x} - E\mathbf{x})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} + \int \int (\mathbf{z} - E\mathbf{z}) (\mathbf{z} - E\mathbf{z})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \end{aligned} \quad (2.213)$$

Each term can be written as

$$\begin{aligned} & \int (\mathbf{x} - E\mathbf{x}) (\mathbf{x} - E\mathbf{x})^\top \left( \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \right) d\mathbf{x} = \int (\mathbf{x} - E\mathbf{x}) (\mathbf{x} - E\mathbf{x})^\top p(\mathbf{x}) d\mathbf{x}, \\ & \int (\mathbf{x} - E\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int (\mathbf{z} - E\mathbf{z})^\top p(\mathbf{z}) d\mathbf{z} = (E\mathbf{x} - E\mathbf{x})(E\mathbf{z} - E\mathbf{z})^\top, \\ & \int (\mathbf{z} - E\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \int (\mathbf{x} - E\mathbf{x})^\top p(\mathbf{x}) d\mathbf{x} = (E\mathbf{z} - E\mathbf{z})(E\mathbf{x} - E\mathbf{x})^\top, \\ & \int (\mathbf{z} - E\mathbf{z}) (\mathbf{z} - E\mathbf{z})^\top \left( \int p(\mathbf{x}, \mathbf{z}) d\mathbf{x} \right) d\mathbf{z} = \int (\mathbf{z} - E\mathbf{z}) (\mathbf{z} - E\mathbf{z})^\top p(\mathbf{z}) d\mathbf{z}. \end{aligned} \quad (2.214)$$

Therefore,

$$\text{cov}(\mathbf{x} + \mathbf{z}) = \text{cov } \mathbf{x} + \text{cov } \mathbf{z}. \quad (2.215)$$

## 2.28

Let  $\mathbf{z}$  be a variable such that

$$\begin{aligned} \mathbf{z} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \\ E\mathbf{z} &= \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}, \\ \text{cov } \mathbf{z} &= \begin{bmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^\top \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\top \end{bmatrix}, \end{aligned} \quad (2.216)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are Gaussian variables. By 2.29,

$$(\text{cov } \mathbf{z})^{-1} = \begin{bmatrix} \Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}.$$

Then,  $\ln p(\mathbf{z})$  except the terms independent of  $\mathbf{x}$  and  $\mathbf{y}$  is given by

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ & + \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ = & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu}) + \\ & -\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}))^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})) \\ & + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L} \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}). \end{aligned} \tag{2.217}$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Lambda (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{2.218}$$

Therefore,

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \Lambda^{-1}), \\ p(\mathbf{y} | \mathbf{x}) &= \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}). \end{aligned} \tag{2.219}$$

## 2.29

Let

$$\mathbf{R} = \begin{bmatrix} \Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}. \tag{2.220}$$

By 2.24,

$$\mathbf{R}^{-1} = \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1} \mathbf{A}^\top \\ \mathbf{A} \Lambda^{-1} & \mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top \end{bmatrix}. \tag{2.221}$$

## 2.30

Let

$$\mathbf{R}^{-1} = \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1} \mathbf{A}^\top \\ \mathbf{A} \Lambda^{-1} & \mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top \end{bmatrix}. \tag{2.222}$$

Then,

$$\mathbf{R}^{-1} \begin{bmatrix} \Lambda\boldsymbol{\mu} - \mathbf{A}^\top \mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \quad (2.223)$$

## 2.31

Let  $\mathbf{y}$  be a variable such that

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \quad (2.224)$$

where

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \Sigma_x), \\ p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \Sigma_z). \end{aligned} \quad (2.225)$$

By marginalisation,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \quad (2.226)$$

The right hand side can be written as

$$\int \mathcal{N}(\mathbf{y}|\mathbf{x} + \boldsymbol{\mu}_z, \Sigma_z) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \Sigma_x) d\mathbf{x}. \quad (2.227)$$

The logarithm of the integrand except the terms independent of  $\mathbf{x}$  and  $\mathbf{y}$  is given by

$$-\frac{1}{2} (\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_z)^\top \Sigma_z^{-1} (\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_z) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_x)^\top \Sigma_x^{-1} (\mathbf{x} - \boldsymbol{\mu}_x). \quad (2.228)$$

The terms except the ones independent of  $\mathbf{x}$  and  $\mathbf{y}$  can be written as

$$-\frac{1}{2} \mathbf{u}^\top \mathbf{R} \mathbf{u} + \mathbf{u}^\top \mathbf{v} = -\frac{1}{2} (\mathbf{u} - \mathbf{R}^{-1} \mathbf{v})^\top \mathbf{R} (\mathbf{u} - \mathbf{R}^{-1} \mathbf{v}) + \frac{1}{2} \mathbf{v}^\top \mathbf{R}^{-1} \mathbf{v}, \quad (2.229)$$

where

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \\ \mathbf{R} &= \begin{bmatrix} \Sigma_x^{-1} + \Sigma_z^{-1} & -\Sigma_z^{-1} \\ -\Sigma_z^{-1} & \Sigma_z^{-1} \end{bmatrix}, \\ \mathbf{v} &= \begin{bmatrix} \Sigma_x^{-1} \boldsymbol{\mu}_x - \Sigma_z^{-1} \boldsymbol{\mu}_z \\ \Sigma_z^{-1} \boldsymbol{\mu}_z \end{bmatrix}, \end{aligned} \quad (2.230)$$

By 2.29 and 2.30,

$$\begin{aligned}\mathbf{R}^{-1} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{z}} \end{bmatrix}, \\ \mathbf{R}^{-1}\mathbf{v} &= \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}} \end{bmatrix}.\end{aligned}\tag{2.231}$$

Therefore,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{z}}).\tag{2.232}$$

## 2.32

Let  $\mathbf{x}$  and  $\mathbf{y}$  be variables such that

$$\begin{aligned}p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\ p(\mathbf{y} | \mathbf{x}) &= \mathcal{N}(\mathbf{y} | \mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}).\end{aligned}\tag{2.233}$$

By the Bayes' theorem,

$$p(\mathbf{y} | \mathbf{x})p(\mathbf{x}) = p(\mathbf{x} | \mathbf{y})p(\mathbf{y}).\tag{2.234}$$

The logarithm of the left hand side except the terms independent of  $\mathbf{x}$  and  $\mathbf{y}$  can be written as

$$\begin{aligned}& -\frac{1}{2}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})^T \mathbf{L} (\mathbf{y} - \mathbf{Ax} - \mathbf{b}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) \\&= -\frac{1}{2}(\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^T \mathbf{L} (\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\&\quad - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}).\end{aligned}\tag{2.235}$$

The right hand side can be written as

$$\begin{aligned}& -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T (\boldsymbol{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{A}^T \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\&\quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^T \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\&= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^T (\boldsymbol{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) + \frac{1}{2}\mathbf{z}^T (\boldsymbol{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A}) \mathbf{z} \\&\quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^T \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}),\end{aligned}\tag{2.236}$$

where

$$\mathbf{z} = (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}). \quad (2.237)$$

The right hand side can be written as

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\top (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) \\ & -\frac{1}{2}(\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b})^\top \mathbf{M} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}), \end{aligned} \quad (2.238)$$

where

$$\mathbf{M} = \mathbf{L} - \mathbf{L} \mathbf{A} (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L}. \quad (2.239)$$

We have

$$\boldsymbol{\mu} + \mathbf{z} = (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} ((\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A}) \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b})). \quad (2.240)$$

Then,

$$\boldsymbol{\mu} + \mathbf{z} = (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b}) + \Lambda \boldsymbol{\mu}). \quad (2.241)$$

By 2.26,

$$(\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} = \Lambda^{-1} - \Lambda^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \Lambda^{-1}. \quad (2.242)$$

Then,

$$\mathbf{M} = \mathbf{L} - \mathbf{L} \mathbf{A} \left( \Lambda^{-1} - \Lambda^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \Lambda^{-1} \right) \mathbf{A}^\top \mathbf{L}. \quad (2.243)$$

The right hand side can be written as

$$\begin{aligned} & \mathbf{L} - \mathbf{L} \left( \mathbf{A} \Lambda^{-1} \mathbf{A}^\top - \mathbf{A} \Lambda^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \Lambda^{-1} \mathbf{A}^\top \right) \mathbf{L} \\ & = \mathbf{L} - \mathbf{L} \mathbf{A} \Lambda^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1} (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top - \mathbf{A} \Lambda^{-1} \mathbf{A}^\top) \mathbf{L}. \end{aligned} \quad (2.244)$$

The right hand side can be written as

$$\begin{aligned} & \mathbf{L} - \mathbf{L} \mathbf{A} \Lambda^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1} \\ & = \mathbf{L} (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top - \mathbf{A} \Lambda^{-1} \mathbf{A}^\top) (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1}. \end{aligned} \quad (2.245)$$

Then,

$$\mathbf{M} = (\mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top)^{-1}. \quad (2.246)$$

Therefore,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|(\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b}) + \Lambda \boldsymbol{\mu}), (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}), \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A} \boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A} \Lambda^{-1} \mathbf{A}^\top). \end{aligned} \quad (2.247)$$

## 2.33

Refer to 2.32, while a different approach is presented below.

Let  $\mathbf{x}$  and  $\mathbf{y}$  be variables such that

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}). \end{aligned} \quad (2.248)$$

By the Bayes' theorem,

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}). \quad (2.249)$$

The logarithm of the left hand side except the terms independent of  $\mathbf{x}$  and  $\mathbf{y}$  can be written as

$$-\frac{1}{2}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{Ax} - \mathbf{b}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.250)$$

The terms except the ones independent of  $\mathbf{x}$  and  $\mathbf{y}$  can be written as

$$-\frac{1}{2} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} + \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} -\mathbf{A}^\top \mathbf{L} \mathbf{b} + \boldsymbol{\Lambda} \boldsymbol{\mu} \\ \mathbf{L} \mathbf{b} \end{bmatrix}. \quad (2.251)$$

By 2.24,

$$\begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \\ \mathbf{A} \boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \end{bmatrix}, \quad (2.252)$$

so that

$$\begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}^{-1} \begin{bmatrix} -\mathbf{A}^\top \mathbf{L} \mathbf{b} + \boldsymbol{\Lambda} \boldsymbol{\mu} \\ \mathbf{L} \mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A} \boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \quad (2.253)$$

Then,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top). \quad (2.254)$$

By 2.25,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu} + (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}), (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}). \quad (2.255)$$

We have

$$\begin{aligned} & \boldsymbol{\mu} + (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}) \\ &= (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\Lambda \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b})). \end{aligned} \quad (2.256)$$

Therefore,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x} | (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\Lambda \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b})), (\Lambda + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}). \quad (2.257)$$

## 2.34

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.258)$$

Then,

$$\ln \left( \prod_{n=1}^N p(\mathbf{x}_n) \right) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\det \boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}). \quad (2.259)$$

By 3.21(a), setting the derivatives with respect to  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  to zero gives

$$\begin{aligned} \mathbf{0} &= \sum_{n=1}^N (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x}_n - \boldsymbol{\mu}), \\ \mathbf{O} &= -\frac{N}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \end{aligned} \quad (2.260)$$

Therefore, the maximum likelihood solutions for  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are given by

$$\begin{aligned} \boldsymbol{\mu}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \end{aligned} \quad (2.261)$$

## 2.35

(a)

Let  $\mathbf{x}$  be a variable such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.262)$$

Then,

$$E \mathbf{x} \mathbf{x}^\top = E(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^\top. \quad (2.263)$$

The right hand side can be written as

$$E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top + \boldsymbol{\mu} E(\mathbf{x} - \boldsymbol{\mu})^\top + E(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.264)$$

Since

$$\begin{aligned} E \mathbf{x} &= \boldsymbol{\mu}, \\ \text{cov } \mathbf{x} &= \boldsymbol{\Sigma}, \end{aligned} \quad (2.265)$$

The right hand side can be written as  $\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$ . Therefore,

$$E \mathbf{x} \mathbf{x}^\top = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.266)$$

(b)

Let  $\mathbf{x}_n$  and  $\mathbf{x}_m$  be independent variables such that

$$\begin{aligned} p(\mathbf{x}_n) &= \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ p(\mathbf{x}_m) &= \mathcal{N}(\mathbf{x}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned} \quad (2.267)$$

If  $n \neq m$ , then

$$E \mathbf{x}_n \mathbf{x}_m^\top = E \mathbf{x}_n E \mathbf{x}_m^\top. \quad (2.268)$$

The right hand side can be written as  $\boldsymbol{\mu}\boldsymbol{\mu}^\top$ . Therefore, by (a),

$$E \mathbf{x}_n \mathbf{x}_m^\top = I_{nm} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.269)$$

(c)

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables such that

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.270)$$

By 2.34, the maximum likelihood solutions for  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are given by

$$\begin{aligned} \boldsymbol{\mu}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \end{aligned} \quad (2.271)$$

Then,

$$E \boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^N E(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^\top. \quad (2.272)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} \sum_{n=1}^N E \mathbf{x}_n \mathbf{x}_n^\top - \frac{1}{N^2} \sum_{n=1}^N E \left( \sum_{n=1}^N \mathbf{x}_n \right) \mathbf{x}_n^\top - \frac{1}{N^2} \sum_{n=1}^N E \mathbf{x}_n \left( \sum_{n=1}^N \mathbf{x}_n \right)^\top \\ & + \frac{1}{N^3} \sum_{n=1}^N E \left( \sum_{n=1}^N \mathbf{x}_n \right) \left( \sum_{n=1}^N \mathbf{x}_n \right)^\top. \end{aligned} \quad (2.273)$$

By (b), the first term can be written as  $\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$ . By (b), the second and third terms can be written as

$$-\frac{1}{N} ((\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) + (N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\top) = -\frac{1}{N} \boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.274)$$

By (b), the fourth term can be written as

$$\frac{1}{N^2} (N(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) + N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\top) = \frac{1}{N} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.275)$$

Then,

$$E \boldsymbol{\Sigma}_{ML} = (\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) + 2 \left( -\frac{1}{N} \boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\top \right) + \frac{1}{N} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.276)$$

Therefore,

$$E \boldsymbol{\Sigma}_{ML} = \frac{N-1}{N} \boldsymbol{\Sigma}. \quad (2.277)$$

## 2.36

Let  $x_1, \dots, x_N$  be variables such that

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (2.278)$$

Let us assume that  $\mu$  is known. By 2.34, the maximum likelihood solution for  $\sigma^2$  is given by

$$\sigma_{ML}^{(N)} = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2. \quad (2.279)$$

The right hand side can be written as

$$\frac{1}{N}(x_N - \mu)^2 + \frac{1}{N} \sum_{n=1}^{N-1} (x_n - \mu)^2 = \frac{1}{N}(x_N - \mu)^2 + \frac{N-1}{N} \sigma_{\text{ML}}^{2(N-1)}. \quad (2.280)$$

Then,

$$\sigma_{\text{ML}}^{2(N)} = \sigma_{\text{ML}}^{2(N-1)} + \frac{1}{N} \left( (x_N - \mu)^2 - \sigma_{\text{ML}}^{2(N-1)} \right). \quad (2.281)$$

We have

$$\frac{\partial}{\partial \sigma^2} (-\ln p(x_n)) = \frac{1}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} (x_n - \mu)^2. \quad (2.282)$$

Therefore,

$$\sigma_{\text{ML}}^{2(N)} = \sigma_{\text{ML}}^{2(N-1)} - \frac{2 \left( \sigma_{\text{ML}}^{2(N-1)} \right)^2}{N} \left( \frac{\partial}{\partial \sigma^2} (-\ln p(x_N)) \right) \Big|_{\sigma^2 = \sigma_{\text{ML}}^{2(N-1)}}. \quad (2.283)$$

## 2.37

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables such that

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.284)$$

Let us assume that  $\boldsymbol{\mu}$  is known. By 2.34, the maximum likelihood solution for  $\boldsymbol{\Sigma}$  is given by

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \quad (2.285)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top + \frac{1}{N} \sum_{n=1}^{N-1} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top \\ &= \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top + \frac{N-1}{N} \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}. \end{aligned} \quad (2.286)$$

Then,

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} + \frac{1}{N} \left( (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top - \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} \right). \quad (2.287)$$

By 3.21(a), we have

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} (-\ln p(x_n)) = -\frac{1}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top. \quad (2.288)$$

Therefore,

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} - \frac{(\boldsymbol{\Sigma}_{\text{ML}}^{(N-1)})^2}{N} \left( \frac{\partial}{\partial \boldsymbol{\Sigma}} (-\ln p(\mathbf{x}_N)) \right) \Big|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}}. \quad (2.289)$$

## 2.38

Let  $x_1, \dots, x_N$  be variables such that

$$\begin{aligned} p(x_n|\mu) &= \mathcal{N}(x_n|\mu, \sigma^2), \\ p(\mu) &= \mathcal{N}(\mu|\mu_0, \sigma_0^2). \end{aligned} \quad (2.290)$$

By the Bayes' theorem,

$$p(\mu|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu)p(\mu). \quad (2.291)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{x}$  and  $\mu$  can be written as

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2. \quad (2.292)$$

By 2.34, the maximum likelihood solution for  $\mu$  is given by

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.293)$$

Then, the first term can be written as

$$\begin{aligned} &-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}} + \mu_{\text{ML}} - \mu)^2 \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{\mu_{\text{ML}} - \mu}{\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}}) - \frac{N}{2\sigma^2} (\mu_{\text{ML}} - \mu)^2. \end{aligned} \quad (2.294)$$

Since the second term of the right hand side is zero, the logarithm except the terms independent of  $\mathbf{x}$  and  $\mu$  can be written as

$$\begin{aligned} & -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{N}{2\sigma^2} (\mu_{\text{ML}} - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \\ & = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{1}{2\sigma_N^2} (\mu - \mu_N)^2 + \frac{\mu_N^2}{2\sigma_N^2} - \frac{N\mu_{\text{ML}}^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}, \end{aligned} \quad (2.295)$$

where

$$\begin{aligned} \mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\ \sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}. \end{aligned} \quad (2.296)$$

Therefore,

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2). \quad (2.297)$$

## 2.39

Let  $x_1, \dots, x_N$  be variables such that

$$\begin{aligned} p(x_n|\mu) &= \mathcal{N}(x_n|\mu, \sigma^2), \\ p(\mu) &= \mathcal{N}(\mu|\mu_0, \sigma_0^2). \end{aligned} \quad (2.298)$$

(a)

By 2.38,

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2), \quad (2.299)$$

where

$$\begin{aligned} \mu_N &= \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{n=1}^N x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\ \sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}. \end{aligned} \quad (2.300)$$

Then,

$$\frac{1}{\sigma_N^2} = \frac{(N-1)\sigma_0^2 + \sigma^2}{\sigma^2 \sigma_0^2} + \frac{1}{\sigma^2}. \quad (2.301)$$

Therefore,

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \quad (2.302)$$

We have

$$\frac{\mu_N}{\sigma_N^2} = \frac{1}{\sigma^2} \sum_{n=1}^N x_n + \frac{\mu_0}{\sigma_0^2}, \quad (2.303)$$

so that

$$\frac{\mu_{N-1}}{\sigma_{N-1}^2} = \frac{1}{\sigma^2} \sum_{n=1}^{N-1} x_n + \frac{\mu_0}{\sigma_0^2}. \quad (2.304)$$

Therefore,

$$\frac{\mu_N}{\sigma_N^2} = \frac{\mu_{N-1}}{\sigma_{N-1}^2} + \frac{x_N}{\sigma^2}. \quad (2.305)$$

(b)

By the Bayes' theorem,

$$p(\mu|\mathbf{x}_N)p(\mathbf{x}_N) = p(\mathbf{x}_N|\mu)p(\mu). \quad (2.306)$$

Since  $x_N$  and  $\mathbf{x}_{N-1}$  are independent, it can be written as

$$p(\mu|\mathbf{x}_N)p(x_N)p(\mathbf{x}_{N-1}) = p(x_N|\mu)p(\mathbf{x}_{N-1}|\mu)p(\mu). \quad (2.307)$$

By the Bayes' theorem, the right hand side can be written as

$$p(x_N|\mu)p(\mu|\mathbf{x}_{N-1})p(\mathbf{x}_{N-1}). \quad (2.308)$$

Then,

$$p(\mu|\mathbf{x}_N)p(x_N) = p(\mu|\mathbf{x}_{N-1})p(x_N|\mu). \quad (2.309)$$

The logarithm of the integrand of the right hand side except the terms independent of  $\mu$  or  $x_N$  can be written as

$$\begin{aligned} & -\frac{1}{2\sigma_{N-1}^2}(\mu - \mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N - \mu)^2 \\ &= -\frac{1}{2} \left( \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2} \right) \left( \mu - \frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} \mu_{N-1} - \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} x_N \right)^2 \\ &+ \frac{1}{2} \left( \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2} \right) \left( \frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} \mu_{N-1} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} x_N \right)^2 - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}. \end{aligned} \quad (2.310)$$

Then,

$$\begin{aligned}\mu_N &= \frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} \mu_{N-1} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} x_N, \\ \sigma_N^2 &= \frac{1}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}.\end{aligned}\tag{2.311}$$

Therefore,

$$\begin{aligned}\frac{\mu_N}{\sigma_N^2} &= \frac{\mu_{N-1}}{\sigma_{N-1}^2} + \frac{x_N}{\sigma^2}, \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}.\end{aligned}\tag{2.312}$$

## 2.40

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables such that

$$\begin{aligned}p(\mathbf{x}_n | \boldsymbol{\mu}) &= \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ p(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).\end{aligned}\tag{2.313}$$

By 2.34, the maximum likelihood solution for  $\boldsymbol{\mu}$  is given by

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n,\tag{2.314}$$

By the Bayes' theorem,

$$p(\boldsymbol{\mu} | \mathbf{X})p(\mathbf{X}) = p(\mathbf{X} | \boldsymbol{\mu})p(\boldsymbol{\mu}).\tag{2.315}$$

The logarithm of the right hand side except the terms independent of  $\mathbf{X}$  and  $\boldsymbol{\mu}$  can be written as

$$-\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0). \tag{2.316}$$

The first term can be written as

$$\begin{aligned}
& -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}) \\
& = -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) \\
& \quad - \frac{N}{2} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}).
\end{aligned} \tag{2.317}$$

The second term of the right hand side is zero. Then, the logarithm except the terms independent of  $\mathbf{X}$  and  $\boldsymbol{\mu}$  can be written as

$$\begin{aligned}
& -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{N}{2} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}) \\
& \quad - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\
& = -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\top \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) \\
& \quad + \frac{1}{2} \boldsymbol{\mu}_N^\top \boldsymbol{\Sigma}_N^{-1} \boldsymbol{\mu}_N - \frac{N}{2} \boldsymbol{\mu}_{\text{ML}}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\text{ML}},
\end{aligned} \tag{2.318}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_N &= (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1} (N\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0), \\
\boldsymbol{\Sigma}_N &= (N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}.
\end{aligned} \tag{2.319}$$

Therefore,

$$p(\boldsymbol{\mu} | \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N). \tag{2.320}$$

## 2.41

By the definition,

$$\text{Gam}(\lambda | a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \tag{2.321}$$

Then,

$$\int_0^\infty \text{Gam}(\lambda | a, b) d\lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a-1} \exp(-b\lambda) d\lambda. \tag{2.322}$$

By the transformation

$$\lambda' = b\lambda, \quad (2.323)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a-1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{\Gamma(a)} \int_0^\infty \lambda'^{a-1} \exp(-\lambda') d\lambda'. \quad (2.324)$$

The right hand side can be written as

$$\frac{1}{\Gamma(a)} \Gamma(a) = 1. \quad (2.325)$$

Therefore,

$$\int_0^\infty \text{Gam}(\lambda|a, b) d\lambda = 1. \quad (2.326)$$

## 2.42

Let  $\lambda$  be a variable such that

$$p(\lambda) = \text{Gam}(\lambda|a, b). \quad (2.327)$$

By the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.328)$$

(a)

We have

$$\mathbb{E} \lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right) d\lambda. \quad (2.329)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.330)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b\Gamma(a)} \int_0^\infty \lambda'^a \exp(-\lambda') d\lambda'. \quad (2.331)$$

The right hand side can be written as

$$\frac{1}{b\Gamma(a)} \Gamma(a+1) = \frac{a}{b}. \quad (2.332)$$

Therefore,

$$\mathbb{E} \lambda = \frac{a}{b}. \quad (2.333)$$

(b)

We have

$$E \lambda^2 = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1} \exp\left(-\frac{\lambda}{b}\right) d\lambda. \quad (2.334)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.335)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b^2 \Gamma(a)} \int_0^\infty \lambda'^{a+1} \exp(-\lambda') d\lambda'. \quad (2.336)$$

The right hand side can be written as

$$\frac{1}{b^2 \Gamma(a)} \Gamma(a+2) = \frac{a(a+1)}{b^2}. \quad (2.337)$$

Then,

$$E \lambda^2 = \frac{a(a+1)}{b^2}. \quad (2.338)$$

We have

$$\text{var } \lambda = E \lambda^2 - (E \lambda)^2. \quad (2.339)$$

Therefore,

$$\text{var } \lambda = \frac{a}{b^2}. \quad (2.340)$$

(c)

Setting the derivative of  $\text{Gam}(\lambda|a, b)$  with respect to  $\lambda$  to zero gives

$$0 = \frac{b^a}{\Gamma(a)} \left( \frac{a-1}{\lambda} - b \right) \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right). \quad (2.341)$$

Therefore,

$$\text{mode } \lambda = \frac{a-1}{b}. \quad (2.342)$$

## 2.43

Let

$$p(x|\sigma^2, q) = \frac{q}{2\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \exp\left(-\frac{|x|^q}{2\sigma^2}\right). \quad (2.343)$$

(a)

We have

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = \frac{q}{\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \int_0^{\infty} \exp\left(-\frac{x^q}{2\sigma^2}\right) dx. \quad (2.344)$$

By the transformation

$$x' = \frac{x^q}{2\sigma^2}, \quad (2.345)$$

the right hand side can be written as

$$\begin{aligned} & \frac{q}{\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \int_0^{\infty} \exp(-x') (2\sigma^2)^{\frac{1}{q}} \frac{1}{q} x^{\frac{1}{q}-1} dx' \\ &= \frac{1}{\Gamma(\frac{1}{q})} \int_0^{\infty} x^{\frac{1}{q}-1} \exp(-x') dx'. \end{aligned} \quad (2.346)$$

The right hand side can be written as

$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma\left(\frac{1}{q}\right) = 1. \quad (2.347)$$

Therefore,

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = 1. \quad (2.348)$$

(b)

We have

$$p(x|\sigma^2, 2) = \frac{1}{\Gamma(\frac{1}{2})} (2\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \quad (2.349)$$

Therefore,

$$p(x|\sigma^2, 2) = \mathcal{N}(x|0, \sigma^2). \quad (2.350)$$

(c)

Let  $\mathbf{t} = (t_1, \dots, t_N)^T$  and  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  such that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \quad (2.351)$$

where

$$p(\epsilon_n) = p(\epsilon_n | \sigma^2, q). \quad (2.352)$$

Then, the logarithm of  $p(\epsilon_n)$  except the terms independent of  $\mathbf{w}$  and  $\sigma^2$  can be written as

$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q} \ln(2\sigma^2). \quad (2.353)$$

Therefore, the logarithm of  $p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \sigma^2)$  except the terms independent of  $\mathbf{w}$  and  $\sigma^2$  can be written as

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln(2\sigma^2). \quad (2.354)$$

## 2.44

Let  $x_1, \dots, x_N$  be variables such that

$$\begin{aligned} p(x_n | \mu, \tau) &= \mathcal{N}(x_n | \mu, \tau^{-1}), \\ p(\mu, \tau) &= \mathcal{N}(\mu | \mu_0, (\beta\tau)^{-1}) \text{Gam}(\tau | a, b). \end{aligned} \quad (2.355)$$

By the Bayes' theorem,

$$p(\mu, \tau | \mathbf{x})p(\mathbf{x}) = p(\mathbf{x} | \mu, \tau)p(\mu, \tau). \quad (2.356)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{x}$ ,  $\mu$  and  $\tau$  can be written as

$$\begin{aligned} &\frac{N}{2} \ln \tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \mu)^2 + \frac{1}{2} \ln \tau - \frac{\beta\tau}{2} (\mu - \mu_0)^2 + (a - 1) \ln \tau - b\tau \\ &= \left(a + \frac{N-1}{2}\right) \ln \tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{\beta\tau}{2} (\mu - \mu_0)^2 - b\tau. \end{aligned} \quad (2.357)$$

Let

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.358)$$

Then,

$$\sum_{n=1}^N (x_n - \mu)^2 = \sum_{n=1}^N (x_n - \bar{x} + \bar{x} - \mu)^2. \quad (2.359)$$

The right hand side can be written as

$$\begin{aligned} & \sum_{n=1}^N (x_n - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{n=1}^N (x_n - \bar{x}) + N(\bar{x} - \mu)^2 \\ &= \sum_{n=1}^N (x_n - \bar{x})^2 + N(\bar{x} - \mu)^2. \end{aligned} \quad (2.360)$$

Then, the logarithm except the terms independent of  $\mathbf{x}$ ,  $\mu$  and  $\tau$  can be written as

$$\left( a + \frac{N-1}{2} \right) \ln \tau - \frac{N\tau}{2} (\bar{x} - \mu)^2 - \frac{\beta\tau}{2} (\mu - \mu_0)^2 - b\tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \bar{x})^2. \quad (2.361)$$

The second and third terms can be written as

$$\begin{aligned} & -\frac{N\tau}{2} (\bar{x} - \mu)^2 - \frac{\beta\tau}{2} (\mu - \mu_0)^2 \\ &= -\frac{(N+\beta)\tau}{2} \left( \mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta} \right)^2 + \frac{(N\bar{x} + \beta\mu_0)^2\tau}{2(N+\beta)} - \frac{N\tau}{2} \bar{x}^2 - \frac{\beta\tau}{2} \mu_0^2. \end{aligned} \quad (2.362)$$

The second, third and forth terms on the right hand side can be written as

$$\frac{(N\bar{x} + \beta\mu_0)^2\tau - (N+\beta)N\tau\bar{x}^2 - (N+\beta)\beta\tau\mu_0^2}{2(N+\beta)} = -\frac{N\beta\tau(\bar{x} - \mu_0)^2}{2(N+\beta)}. \quad (2.363)$$

Then, the logarithm except the terms independent of  $\mathbf{x}$ ,  $\mu$  and  $\tau$  can be written as

$$\begin{aligned} & -\frac{(N+\beta)\tau}{2} \left( \mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta} \right)^2 \\ &+ \left( a + \frac{N-1}{2} \right) \ln \tau - \left( b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N+\beta)} + \frac{1}{2} \sum_{n=1}^N (x_n - \bar{x})^2 \right) \tau. \end{aligned} \quad (2.364)$$

Therefore,

$$p(\mu, \tau | \mathbf{x}) = \mathcal{N}(\mu | \mu_N, \tau_N^{-1}) \text{Gam}(\tau | a_N, b_N), \quad (2.365)$$

where

$$\begin{aligned}\mu_N &= \frac{N\bar{x} + \beta\mu_0}{N + \beta}, \\ \tau_N &= (N + \beta)\tau, \\ a_N &= a + \frac{N + 1}{2}, \\ b_N &= b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N + \beta)} + \frac{1}{2} \sum_{n=1}^N (x_n - \bar{x})^2.\end{aligned}\tag{2.366}$$

## 2.45

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables in  $D$  dimensions such that

$$\begin{aligned}p(\mathbf{x}_n | \boldsymbol{\Lambda}) &= \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\ p(\boldsymbol{\Lambda}) &= \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu),\end{aligned}\tag{2.367}$$

where

$$\mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) = B(\mathbf{W}, \nu) |\det \boldsymbol{\Lambda}|^{\frac{\nu-D-1}{2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{W}^{-1} \boldsymbol{\Lambda})\right).\tag{2.368}$$

By the Bayes' theorem,

$$p(\boldsymbol{\Lambda} | \mathbf{X})p(\mathbf{X}) = p(\mathbf{X} | \boldsymbol{\Lambda})p(\boldsymbol{\Lambda}).\tag{2.369}$$

The logarithm of the right hand side except the terms independent of  $\boldsymbol{\Lambda}$  can be written as

$$\begin{aligned}&-\frac{N}{2} \ln |\det(\boldsymbol{\Lambda}^{-1})| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x}_n - \boldsymbol{\mu}) \\ &+ \frac{\nu - D - 1}{2} \ln |\det \boldsymbol{\Lambda}| - \frac{1}{2} \text{tr}(\mathbf{W}^{-1} \boldsymbol{\Lambda}) \\ &= \frac{\nu + N - D - 1}{2} \ln |\det \boldsymbol{\Lambda}| - \frac{1}{2} \text{tr}((\mathbf{W}^{-1} + \mathbf{S}) \boldsymbol{\Lambda}),\end{aligned}\tag{2.370}$$

where

$$\mathbf{S} = \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top.\tag{2.371}$$

Then,

$$p(\boldsymbol{\Lambda} | \mathbf{X}) = \mathcal{W}\left(\boldsymbol{\Lambda} | (\mathbf{W}^{-1} + \mathbf{S})^{-1}, \nu + N\right).\tag{2.372}$$

Therefore,  $\mathcal{W}$  is a conjugate prior distribution of  $\boldsymbol{\Lambda}$ .

## 2.46

Let  $x$  be a variable such that

$$\begin{aligned} p(x|\tau) &= \mathcal{N}(x|\mu, \tau^{-1}), \\ p(\tau) &= \text{Gam}(\tau|a, b). \end{aligned} \quad (2.373)$$

By marginalisation,

$$p(x) = \int_0^\infty p(x|\tau)p(\tau)d\tau. \quad (2.374)$$

The right hand side can be written as

$$\begin{aligned} &\int_0^\infty (2\pi\tau^{-1})^{-\frac{1}{2}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) d\tau \\ &= (2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp\left(-\left(b + \frac{(x-\mu)^2}{2}\right)\tau\right) d\tau. \end{aligned} \quad (2.375)$$

By the transformation

$$\tau' = \left(b + \frac{(x-\mu)^2}{2}\right)\tau, \quad (2.376)$$

the integral of the right hand side can be written as

$$\begin{aligned} &\int_0^\infty \left(\frac{\tau'}{b + \frac{(x-\mu)^2}{2}}\right)^{a-\frac{1}{2}} \exp(-\tau') \frac{1}{b + \frac{(x-\mu)^2}{2}} d\tau' \\ &= \Gamma\left(a + \frac{1}{2}\right) \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \end{aligned} \quad (2.377)$$

Then,

$$p(x) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \quad (2.378)$$

By the transformation

$$\begin{aligned} \nu &= 2a, \\ \lambda &= \frac{a}{b}, \end{aligned} \quad (2.379)$$

the right hand side can be written as

$$\begin{aligned} & (2\pi)^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( \frac{\nu}{2\lambda} \right)^{-\frac{1}{2}} \left( 1 + \frac{(x-\mu)^2}{\frac{\nu}{\lambda}} \right)^{-\frac{\nu+1}{2}} \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( \frac{\lambda}{\pi\nu} \right)^{\frac{1}{2}} \left( 1 + \frac{\lambda}{\nu}(x-\mu)^2 \right)^{-\frac{\nu+1}{2}}. \end{aligned} \quad (2.380)$$

Therefore,

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( \frac{\lambda}{\pi\nu} \right)^{\frac{1}{2}} \left( 1 + \frac{\lambda}{\nu}(x-\mu)^2 \right)^{-\frac{\nu+1}{2}}. \quad (2.381)$$

## 2.47

Let

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( \frac{\lambda}{\pi\nu} \right)^{\frac{1}{2}} \left( 1 + \frac{\lambda}{\nu}(x-\mu)^2 \right)^{-\frac{\nu+1}{2}}. \quad (2.382)$$

By the transformation

$$\frac{1}{y} = \frac{\lambda}{\nu}(x-\mu)^2, \quad (2.383)$$

the right hand side except the terms independent of  $x$  can be written as

$$\left( 1 + \frac{1}{y} \right)^{-\frac{\lambda(x-\mu)^2}{2} y - \frac{1}{2}}. \quad (2.384)$$

By the property

$$\lim_{x \rightarrow \infty} \left( 1 + \frac{1}{x} \right)^x = e, \quad (2.385)$$

we have

$$\lim_{y \rightarrow \infty} \left( 1 + \frac{1}{y} \right)^{-\frac{\lambda(x-\mu)^2}{2} y - \frac{1}{2}} = \exp \left( -\frac{\lambda}{2}(x-\mu)^2 \right). \quad (2.386)$$

Therefore,

$$\lim_{\nu \rightarrow \infty} \text{St}(x|\mu, \lambda, \nu) = \mathcal{N}(x|\mu, \lambda^{-1}). \quad (2.387)$$

## 2.48

Let  $\mathbf{x}$  be a variable in  $D$  dimensions such that

$$\begin{aligned} p(\mathbf{x}|\eta) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\Lambda)^{-1}), \\ p(\eta) &= \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned} \quad (2.388)$$

By marginalisation,

$$p(\mathbf{x}) = \int_0^\infty p(\mathbf{x}|\eta)p(\eta)d\eta. \quad (2.389)$$

The right hand side can be written as

$$\begin{aligned} &\int_0^\infty (2\pi)^{-\frac{D}{2}} |\det(\eta\Lambda)^{-1}|^{-\frac{1}{2}} \exp\left(-\frac{\eta}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})\right) \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \eta^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu}{2}\eta\right) d\eta \\ &= (2\pi)^{-\frac{D}{2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} |\det \Lambda|^{\frac{1}{2}} \int_0^\infty \eta^{\frac{D+\nu}{2}-1} \exp\left(-\frac{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})}{2}\eta\right) d\eta. \end{aligned} \quad (2.390)$$

By the transformation

$$\eta' = \frac{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})}{2}\eta, \quad (2.391)$$

the integral of the right hand side can be written as

$$\begin{aligned} &\int_0^\infty \left( \frac{2\eta'}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})} \right)^{\frac{D+\nu}{2}-1} \exp(-\eta') \frac{2}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})} d\eta' \\ &= \left( \frac{2}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})} \right)^{\frac{D+\nu}{2}} \int_0^\infty \eta'^{\frac{D+\nu}{2}-1} \exp(-\eta') d\eta'. \end{aligned} \quad (2.392)$$

Then,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} |\det \Lambda|^{\frac{1}{2}} \left( \frac{2}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})} \right)^{\frac{D+\nu}{2}} \Gamma\left(\frac{D+\nu}{2}\right). \quad (2.393)$$

The right hand side can be written as

$$\begin{aligned} &(2\pi)^{-\frac{D}{2}} \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} |\det \Lambda|^{\frac{1}{2}} \left(\frac{\nu}{2}\right)^{-\frac{D}{2}} \left(\frac{\nu}{2}\right)^{\frac{D+\nu}{2}} \left(\frac{2}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}} \\ &= (2\pi)^{-\frac{D}{2}} \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} |\det \Lambda|^{\frac{1}{2}} \left(\frac{\nu}{2}\right)^{-\frac{D}{2}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \Lambda(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \end{aligned} \quad (2.394)$$

Therefore,

$$p(\mathbf{x}) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\det \boldsymbol{\Lambda}|^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.395)$$

## 2.49

Let  $\mathbf{x}$  be a variable such that

$$p(\mathbf{x}) = \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu), \quad (2.396)$$

where

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.397)$$

(a)

We have

$$\mathbb{E} \mathbf{x} = \int \mathbf{x} \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \quad (2.398)$$

The right hand side can be written as

$$\begin{aligned} & \int \mathbf{x} \left( \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left( \int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \quad (2.399)$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \quad (2.400)$$

Therefore,

$$\mathbb{E} \mathbf{x} = \boldsymbol{\mu}. \quad (2.401)$$

(b)

By (a), we have

$$\text{cov } \mathbf{x} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \quad (2.402)$$

The right hand side can be written as

$$\begin{aligned} & \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \left( \int \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left( \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \quad (2.403)$$

The right hand side can be written as

$$\int (\eta \boldsymbol{\Lambda})^{-1} \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left( \int \eta^{\frac{\nu}{2}-2} \exp\left(-\frac{\nu}{2}\eta\right) d\eta \right) \boldsymbol{\Lambda}^{-1}. \quad (2.404)$$

By the transformation

$$\eta' = \frac{\nu}{2}\eta, \quad (2.405)$$

the integral of the right hand side can be written as

$$\int \left( \frac{2}{\nu} \eta' \right)^{\frac{\nu}{2}-2} \exp(-\eta') \frac{2}{\nu} d\eta' = \left( \frac{2}{\nu} \right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2} - 1\right). \quad (2.406)$$

Then,

$$\text{cov } \mathbf{x} = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left( \frac{2}{\nu} \right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2} - 1\right) \boldsymbol{\Lambda}^{-1}. \quad (2.407)$$

Therefore,

$$\text{cov } \mathbf{x} = \frac{\nu}{\nu - 2} \boldsymbol{\Lambda}^{-1}. \quad (2.408)$$

(c)

Setting the derivative of  $p(\mathbf{x})$  to zero gives

$$\mathbf{0} = -\frac{1}{2} (\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top) (\mathbf{x} - \boldsymbol{\mu}) \int \eta \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.409)$$

Therefore,

$$\text{mode } \mathbf{x} = \boldsymbol{\mu}. \quad (2.410)$$

## 2.50

Let

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.411)$$

By the transformation

$$y = \frac{\nu}{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}, \quad (2.412)$$

the right hand side except the terms independent of  $\mathbf{x}$  can be written as

$$\left(1 + \frac{1}{y}\right)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x}-\boldsymbol{\mu})}{2} y - \frac{D}{2}}. \quad (2.413)$$

By the property

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x = e, \quad (2.414)$$

we have

$$\lim_{y \rightarrow \infty} \left(1 + \frac{1}{y}\right)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x}-\boldsymbol{\mu})}{2} y - \frac{D}{2}} = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right). \quad (2.415)$$

Therefore,

$$\lim_{\nu \rightarrow \infty} \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.416)$$

## 2.51

(a)

We have

$$\exp(iA) \exp(-iA) = 1. \quad (2.417)$$

The left hand side can be written as

$$(\cos A + i \sin A)(\cos A - i \sin A) = \cos^2 A + \sin^2 A. \quad (2.418)$$

Therefore,

$$\cos^2 A + \sin^2 A = 1. \quad (2.419)$$

(b)

We have

$$\cos(A - B) = \operatorname{Re}(\exp(i(A - B))). \quad (2.420)$$

The right hand side can be written as

$$\operatorname{Re}(\exp(iA)\exp(-iB)) = \operatorname{Re}((\cos A + i \sin A)(\cos B - i \sin B)). \quad (2.421)$$

Therefore,

$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \quad (2.422)$$

(c)

We have

$$\sin(A - B) = \operatorname{Im}(\exp(i(A - B))). \quad (2.423)$$

The right hand side can be written as

$$\operatorname{Im}(\exp(iA)\exp(-iB)) = ((\cos A + i \sin A)(\cos B - i \sin B)). \quad (2.424)$$

Therefore,

$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \quad (2.425)$$

## 2.52

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.426)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \quad (2.427)$$

By the Taylor series

$$\cos \alpha = 1 - \frac{1}{2}\alpha^2 + O(\alpha^4), \quad (2.428)$$

the right hand side can be written as

$$\begin{aligned} & \frac{\exp(m(1 - \frac{1}{2}(\theta - \theta_0)^2 + O((\theta - \theta_0)^4)))}{\int_0^{2\pi} \exp(m(1 - \frac{1}{2}\theta^2 + O(\theta^4))) d\theta} \\ &= \exp\left(-\frac{m}{2}(\theta - \theta_0)^2\right) \frac{\exp(mO((\theta - \theta_0)^4))}{\int_0^{2\pi} \exp(m(-\frac{1}{2}\theta^2 + O(\theta^4))) d\theta}. \end{aligned} \quad (2.429)$$

Therefore,

$$\lim_{m \rightarrow \infty} f(\theta|\theta_0, m) = \mathcal{N}(\theta|\theta_0, m^{-1}). \quad (2.430)$$

## 2.53

Let

$$\sum_{n=1}^N \sin(\theta_n - \theta_0) = 0. \quad (2.431)$$

The left hand side can be written as

$$\sum_{n=1}^N (\sin \theta_n \cos \theta_0 - \cos \theta_n \sin \theta_0) = \cos \theta_0 \sum_{n=1}^N \sin \theta_n - \sin \theta_0 \sum_{n=1}^N \cos \theta_n. \quad (2.432)$$

Therefore,

$$\theta_0 = \arctan \left( \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right). \quad (2.433)$$

## 2.54

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.434)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \quad (2.435)$$

Setting the first and second derivatives with respect to  $\theta$  to zero gives

$$\begin{aligned} 0 &= -m \sin(\theta - \theta_0) f(\theta|\theta_0, m), \\ 0 &= (m^2 \sin^2(\theta - \theta_0) - m \cos(\theta - \theta_0)) f(\theta|\theta_0, m). \end{aligned} \quad (2.436)$$

Therefore,

$$\begin{aligned} \underset{\theta}{\operatorname{argmax}} f(\theta|\theta_0, m) &= \theta_0, \\ \underset{\theta}{\operatorname{argmin}} f(\theta|\theta_0, m) &= \theta_0 - \pi \operatorname{sgn}(\theta_0 - \pi). \end{aligned} \quad (2.437)$$

## 2.55

(a)

Let  $\theta_1, \dots, \theta_N$  be variables such that

$$p(\theta_n) = f(\theta_n|\theta_0, m), \quad (2.438)$$

where

$$\begin{aligned} f(\theta|\theta_0, m) &= \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \\ I_0(m) &= \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \end{aligned} \quad (2.439)$$

Then,

$$\ln \left( \prod_{n=1}^N p(\theta_n|\theta_0, m) \right) = -\frac{N}{2} \ln(2\pi I_0(m)) + m \sum_{n=1}^N \cos(\theta_n - \theta_0). \quad (2.440)$$

Setting the derivative with respect to  $\theta_0$  to zero gives

$$0 = m \sum_{n=1}^N \sin(\theta_n - \theta_0). \quad (2.441)$$

Therefore, by 2.53, the maximum likelihood solution for  $\theta_0$  is given by

$$\theta_0^{\text{ML}} = \arctan \left( \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right). \quad (2.442)$$

(b)

Let

$$\begin{aligned} \bar{r} \cos \bar{\theta} &= \frac{1}{N} \sum_{n=1}^N \cos \theta_n, \\ \bar{r} \sin \bar{\theta} &= \frac{1}{N} \sum_{n=1}^N \sin \theta_n. \end{aligned} \quad (2.443)$$

By (a),

$$\bar{\theta} = \theta_0^{\text{ML}}. \quad (2.444)$$

We have

$$\frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}}) = \left( \frac{1}{N} \sum_{n=1}^N \cos \theta_n \right) \cos \theta_0^{\text{ML}} + \left( \frac{1}{N} \sum_{n=1}^N \sin \theta_n \right) \sin \theta_0^{\text{ML}}. \quad (2.445)$$

The right hand side can be written as

$$\bar{r} \cos^2 \bar{\theta} + \bar{r} \sin^2 \bar{\theta} = \bar{r}. \quad (2.446)$$

Therefore,

$$\frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}}) = \bar{r}. \quad (2.447)$$

## 2.56

(a)

Let

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \quad (2.448)$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp((a-1)\ln\mu + (b-1)\ln(1-\mu)) \quad (2.449)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}.$$

(b)

Let

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.450)$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \exp((a-1)\ln\lambda - b\lambda). \quad (2.451)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ -b \end{bmatrix}.$$

(c)

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.452)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta, \quad (2.453)$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m \cos \theta_0 \cos \theta + m \sin \theta_0 \sin \theta). \quad (2.454)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} m \cos \theta_0 \\ m \sin \theta_0 \end{bmatrix}.$$

## 2.57

By the definition,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (2.455)$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})), \quad (2.456)$$

where

$$h(\mathbf{x}) = (2\pi)^{-\frac{D}{2}},$$

$$g(\boldsymbol{\eta}) = (\det(-2\boldsymbol{\eta}_2))^{-\frac{1}{2}} \exp\left(\frac{1}{4}\boldsymbol{\eta}_1^\top \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1\right),$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ -\frac{1}{2} \boldsymbol{\Sigma}^{-1} \end{bmatrix},$$

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \mathbf{x}^\top \end{bmatrix}.$$

## 2.58

Let  $\mathbf{x}$  be a variable such that

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})). \quad (2.457)$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = 1 \quad (2.458)$$

with respect to  $\boldsymbol{\eta}$  gives

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = \mathbf{0}. \quad (2.459)$$

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + E \mathbf{u}(\mathbf{x}). \quad (2.460)$$

Therefore,

$$E \mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}. \quad (2.461)$$

Thus,

$$E \mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}). \quad (2.462)$$

Taking the second derivative with respect to  $\boldsymbol{\eta}$  gives

$$\begin{aligned} & \nabla \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} + 2 \nabla g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} \\ & + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = \mathbf{O}. \end{aligned} \quad (2.463)$$

The left hand side can be written as

$$\begin{aligned} & \frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \frac{2 \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int \mathbf{u}(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} \\ & = \frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2 E \mathbf{u}(\mathbf{x}) E \mathbf{u}(\mathbf{x})^\top + E (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top). \end{aligned} \quad (2.464)$$

Therefore,

$$E(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\top) = -\frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\top}{g^2(\boldsymbol{\eta})}. \quad (2.465)$$

By the definition,

$$\text{cov } \mathbf{u}(\mathbf{x}) = E(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\top) - E\mathbf{u}(\mathbf{x})E\mathbf{u}(\mathbf{x})^\top. \quad (2.466)$$

Thus,

$$\text{cov } \mathbf{u}(\mathbf{x}) = -\frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\top}{g^2(\boldsymbol{\eta})}. \quad (2.467)$$

Hence,

$$\text{cov } \mathbf{u}(\mathbf{x}) = -\nabla \nabla \ln g(\boldsymbol{\eta}). \quad (2.468)$$

## 2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right). \quad (2.469)$$

Then

$$\int p(x|\sigma)dx = \frac{1}{\sigma} \int f\left(\frac{x}{\sigma}\right) dx. \quad (2.470)$$

By the transformation

$$x' = \frac{x}{\sigma}, \quad (2.471)$$

the right hand side can be written as

$$\frac{1}{\sigma} \int f(x')\sigma dx' = \int f(x')dx'. \quad (2.472)$$

Therefore,  $p(x|\sigma)$  will be normalised if  $f(x)$  is normalised.

## 2.60

Let  $\mathbf{x}$  be a variable such that

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \quad (2.473)$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \quad (2.474)$$

Since

$$\int p(\mathbf{x})d\mathbf{x} = 1, \quad (2.475)$$

we have

$$\sum_i h_i \Delta_i = 1. \quad (2.476)$$

Let  $N$  be the total number of observations and  $n_i$  be the number of observations which fall in  $\mathcal{R}_i$ . Then, the logarithm of the likelihood is given by

$$\ln \left( \prod_i h_i^{n_i} \right) = \sum_i n_i \ln h_i, \quad (2.477)$$

where

$$\sum_i n_i = N. \quad (2.478)$$

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda \left( \sum_i h_i \Delta_i - 1 \right) \quad (2.479)$$

with respect to  $h_i$  and  $\lambda$  to zero gives

$$\begin{aligned} \frac{n_i}{h_i} + \lambda \Delta_i &= 0, \\ \sum_i h_i \Delta_i - 1 &= 0. \end{aligned} \quad (2.480)$$

Then,

$$\begin{aligned} \lambda &= -N, \\ h_i &= \frac{n_i}{N \Delta_i}. \end{aligned} \quad (2.481)$$

Therefore, the maximum likelihood estimator for the  $\{h_i\}$  is  $\frac{n_i}{N \Delta_i}$ .

## 2.61 (Incomplete)

Let  $\mathbf{x}$  be a variable and  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be observations. Let

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})}, \quad (2.482)$$

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{x}_{(K)} - \mathbf{x}\|} d\mathbf{x}', \quad (2.483)$$

$K$  is a constant and  $\mathbf{x}_{(K)}$  is the  $K$ th nearest observation from the point  $\mathbf{x}$ .

### 3 Linear Models for Regression

#### 3.1

By the definition,

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \quad (3.1)$$

The right hand side can be written as

$$\frac{1 - \exp(-2a)}{1 + \exp(-2a)} = \frac{2}{1 + \exp(-2a)} - 1. \quad (3.2)$$

Therefore,

$$\tanh a = 2\sigma(2a) - 1, \quad (3.3)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (3.4)$$

Let

$$y(x_n, \mathbf{w}) = w_0 + \sum_{m=1}^M w_m \sigma\left(\frac{x - \mu_m}{s}\right). \quad (3.5)$$

By the result above, the right hand side can be written as

$$w_0 + \sum_{m=1}^M w_m \frac{1 + \tanh\left(\frac{x - \mu_m}{2s}\right)}{2} = w_0 + \frac{1}{2} \sum_{m=1}^M w_m + \frac{1}{2} \sum_{m=1}^M w_m \tanh\left(\frac{x - \mu_m}{2s}\right). \quad (3.6)$$

Therefore,  $y(x_n, \mathbf{w})$  is equivalent to

$$y(x_n, \mathbf{u}) = u_0 + \sum_{m=1}^M u_m \tanh\left(\frac{x - \mu_m}{2s}\right), \quad (3.7)$$

where

$$\begin{aligned} u_0 &= w_0 + \frac{1}{2} \sum_{m=1}^M w_m, \\ u_m &= \frac{1}{2} w_m. \end{aligned} \quad (3.8)$$

### 3.2 (Incomplete)

Let  $\Phi$  be an  $N \times M$  matrix. Then, for any vector  $\mathbf{v}$  in  $N$  dimensions,

$$\Phi (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{v} \quad (3.9)$$

is a projection of  $\mathbf{v}$  onto the space spanned by the columns of  $\Phi$ ?

Additionally, for a vector  $\mathbf{t}$  in  $N$  dimensions,

$$(\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{t} \quad (3.10)$$

is an orthogonal projection of  $\mathbf{t}$  onto the space spanned by the columns of  $\Phi$ ?

### 3.3

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N r_n (t_n - \mathbf{w}^\top \phi_n)^2. \quad (3.11)$$

The right hand side can be written as

$$\frac{1}{2} \|\mathbf{t}' - \Phi' \mathbf{w}\|^2, \quad (3.12)$$

where

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \vdots \\ \sqrt{r_N} t_N \end{bmatrix}, \Phi' = \begin{bmatrix} \sqrt{r_1} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \sqrt{r_N} \phi(\mathbf{x}_N)^\top \end{bmatrix}. \quad (3.13)$$

Setting the derivative with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = -\Phi'^\top (\mathbf{t}' - \Phi' \mathbf{w}). \quad (3.14)$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} E(\mathbf{w}) = (\Phi'^\top \Phi')^{-1} \Phi'^\top \mathbf{t}'. \quad (3.15)$$

### 3.4 (Incomplete)

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(\mathbf{x}_n, \mathbf{w}) - t_n)^2, \quad (3.16)$$

where

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{m=1}^M w_m(x_m + \epsilon_m), \\ p(\epsilon_m) &= \mathcal{N}(\epsilon_m | 0, \sigma^2). \end{aligned} \quad (3.17)$$

Setting the derivative with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = \sum_{n=1}^N \begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n). \quad (3.18)$$

The right hand side can be written as

$$\sum_{n=1}^N \begin{bmatrix} 1 \\ \mathbf{x}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n) + \sum_{n=1}^N \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n). \quad (3.19)$$

### 3.5

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2. \quad (3.20)$$

Then, the minimisation of  $E(\mathbf{w})$  under the constraint

$$\sum_{m=1}^M |w_m|^q \leq \eta \quad (3.21)$$

reduces to the minimisation of

$$E(\mathbf{w}) + \lambda \left( \sum_{m=1}^M |w_m|^q - \eta \right) \quad (3.22)$$

with respect to  $\mathbf{w}$  and  $\lambda$ . Then,

$$\eta = \sum_{m=1}^M |w_m^*(\lambda)|^q, \quad (3.23)$$

where

$$\mathbf{w}^*(\lambda) = \operatorname{argmin}_{\mathbf{w}} \left( E(\mathbf{w}) + \lambda \left( \sum_{m=1}^M |w_m|^q - \eta \right) \right). \quad (3.24)$$

### 3.6

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables in  $D$  dimensions such that

$$p(\mathbf{t}_n | \mathbf{y}_n) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}_n, \boldsymbol{\Sigma}), \quad (3.25)$$

where

$$\mathbf{y}_n = \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n). \quad (3.26)$$

Then,

$$\begin{aligned} \ln p(\mathbf{T} | \mathbf{Y}) &= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln(\det \boldsymbol{\Sigma}) \\ &\quad - \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)). \end{aligned} \quad (3.27)$$

By 3.21(a), setting the derivatives with respect to  $\mathbf{W}$  and  $\boldsymbol{\Sigma}$  to zero gives

$$\begin{aligned} \mathbf{O} &= -\frac{1}{2} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)) (\boldsymbol{\phi}(\mathbf{x}_n))^\top, \\ \mathbf{O} &= -\frac{N}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n))^\top. \end{aligned} \quad (3.28)$$

Therefore, the maximum likelihood solutions for  $\mathbf{W}$  and  $\boldsymbol{\Sigma}$  are given by

$$\begin{aligned} \mathbf{W}_{\text{ML}} &= (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}_{\text{ML}}^\top \boldsymbol{\phi}(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}_{\text{ML}}^\top \boldsymbol{\phi}(\mathbf{x}_n))^\top. \end{aligned} \quad (3.29)$$

### 3.7

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.30)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \quad (3.31)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{t}$  and  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 - \frac{1}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ &= -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} & -\beta \boldsymbol{\Phi}^\top \\ -\beta \boldsymbol{\Phi} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_0^{-1} \mathbf{m}_0 \\ \mathbf{0} \end{bmatrix} \\ & \quad - \frac{1}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0. \end{aligned} \quad (3.32)$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad (3.33)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.34)$$

### 3.8

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.35)$$

By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad (3.36)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.37)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}')p(\mathbf{t}') = p(\mathbf{t}'|\mathbf{w})p(\mathbf{w}), \quad (3.38)$$

where

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}. \quad (3.39)$$

Since  $\mathbf{t}$  and  $t_{N+1}$  are independent, it can be written as

$$p(\mathbf{w}|\mathbf{t}')p(t_{N+1})p(\mathbf{t}) = p(t_{N+1}|\mathbf{w})p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \quad (3.40)$$

By the Bayes' theorem, the right hand side can be written as

$$p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})p(\mathbf{t}). \quad (3.41)$$

Then,

$$p(\mathbf{w}|\mathbf{t}')p(t_{N+1}) = p(\mathbf{w}|\mathbf{t})p(t_{N+1}|\mathbf{w}). \quad (3.42)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{1}{2}(\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) - \frac{\beta}{2}(t_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}_{N+1})^2 \\ &= -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top & -\beta \boldsymbol{\phi}_{N+1} \\ -\beta \boldsymbol{\phi}_{N+1}^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}^{-1} \mathbf{m} \\ 0 \end{bmatrix} - \frac{1}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}. \end{aligned} \quad (3.43)$$

Then,

$$p(\mathbf{w}|\mathbf{t}') = \mathcal{N}(\mathbf{w}|\mathbf{m}', \mathbf{S}'), \quad (3.44)$$

where

$$\begin{aligned} \mathbf{m}' &= \mathbf{S}' (\mathbf{S}^{-1} \mathbf{m} + \beta t_{N+1} \boldsymbol{\phi}_{N+1}), \\ \mathbf{S}' &= (\mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top)^{-1}. \end{aligned} \quad (3.45)$$

We have

$$\mathbf{S}^{-1} \mathbf{m} + \beta t_{N+1} \boldsymbol{\phi}_{N+1} = \mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}'^\top \mathbf{t}', \quad (3.46)$$

and

$$\mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top = \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}'^\top \boldsymbol{\Phi}', \quad (3.47)$$

where

$$\boldsymbol{\Phi}' = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\phi}_{N+1}^\top \end{bmatrix}. \quad (3.48)$$

Therefore,

$$\begin{aligned} \mathbf{m}' &= \mathbf{S}' (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}'^\top \mathbf{t}'), \\ \mathbf{S}' &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}'^\top \boldsymbol{\Phi}')^{-1}. \end{aligned} \quad (3.49)$$

### 3.9

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.50)$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad (3.51)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.52)$$

Let

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\phi}_{N+1}^\top \end{bmatrix}. \quad (3.53)$$

Then,

$$p(\mathbf{w}|\mathbf{t}') = \mathcal{N}(\mathbf{w}|\mathbf{m}', \mathbf{t}'), \quad (3.54)$$

where

$$\begin{aligned} \mathbf{m}' &= \mathbf{S}' (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}'^\top \mathbf{t}'), \\ \mathbf{S}' &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}'^\top \boldsymbol{\Phi}')^{-1}. \end{aligned} \quad (3.55)$$

### 3.10

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.56)$$

By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad (3.57)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.58)$$

By marginalisation,

$$p(t_{N+1}|\mathbf{t}) = \int p(t_{N+1}|\mathbf{w}) p(\mathbf{w}|\mathbf{t}) d\mathbf{w}. \quad (3.59)$$

The logarithm of the integrand of the right hand side except the terms independent of  $t_{N+1}$  and  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{\beta}{2}(t_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}_{N+1})^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) \\ &= -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top & -\beta \boldsymbol{\phi}_{N+1} \\ -\beta \boldsymbol{\phi}_{N+1}^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}^{-1} \mathbf{m} \\ 0 \end{bmatrix} - \frac{1}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}. \end{aligned} \quad (3.60)$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top & -\beta \boldsymbol{\phi}_{N+1} \\ -\beta \boldsymbol{\phi}_{N+1}^\top & \beta \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S} & \mathbf{S} \boldsymbol{\phi}_{N+1} \\ \boldsymbol{\phi}_{N+1}^\top \mathbf{S} & \beta^{-1} + \boldsymbol{\phi}_{N+1}^\top \mathbf{S} \boldsymbol{\phi}_{N+1} \end{bmatrix}. \quad (3.61)$$

Then,

$$\begin{bmatrix} \mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top & \beta \boldsymbol{\phi}_{N+1} \\ \beta \boldsymbol{\phi}_{N+1}^\top & \beta \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}^{-1} \mathbf{m} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m}^\top \boldsymbol{\phi}_{N+1} \end{bmatrix}. \quad (3.62)$$

Therefore,

$$p(t_{N+1} | \mathbf{t}) = \mathcal{N}(t_{N+1} | \mathbf{m}^\top \boldsymbol{\phi}_{N+1}, \sigma^2(\boldsymbol{\phi}_{N+1})), \quad (3.63)$$

where

$$\sigma^2(\boldsymbol{\phi}) = \beta^{-1} + \boldsymbol{\phi}^\top \mathbf{S} \boldsymbol{\phi}. \quad (3.64)$$

### 3.11

Let  $t_1, \dots, t_N$  be a variable such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.65)$$

By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.66)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.67)$$

By 3.10,

$$p(t_{N+1}|\mathbf{t}) = \mathcal{N}(t_{N+1}|\mathbf{m}^\top \boldsymbol{\phi}_{N+1}, \sigma^2(\boldsymbol{\phi}_{N+1})), \quad (3.68)$$

where

$$\sigma^2(\boldsymbol{\phi}) = \beta^{-1} + \boldsymbol{\phi}^\top \mathbf{S} \boldsymbol{\phi}. \quad (3.69)$$

Let

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\phi}_{N+1}^\top \end{bmatrix}. \quad (3.70)$$

Then,

$$p(t_{N+2}|\mathbf{t}') = \mathcal{N}(t_{N+2}|\mathbf{m}'^\top \boldsymbol{\phi}_{N+2}, \sigma'^2(\boldsymbol{\phi}_{N+2})), \quad (3.71)$$

where

$$\begin{aligned} \mathbf{m}' &= \mathbf{S}' (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}'^\top \mathbf{t}'), \\ \mathbf{S}' &= (\mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}'^\top \boldsymbol{\Phi}')^{-1}, \\ \sigma'^2(\boldsymbol{\phi}) &= \beta^{-1} + \boldsymbol{\phi}^\top \mathbf{S}' \boldsymbol{\phi}. \end{aligned} \quad (3.72)$$

We have

$$\sigma^2(\boldsymbol{\phi}) - \sigma'^2(\boldsymbol{\phi}) = \boldsymbol{\phi}^\top (\mathbf{S} - \mathbf{S}') \boldsymbol{\phi}. \quad (3.73)$$

We have

$$\mathbf{S}' = (\mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top)^{-1}. \quad (3.74)$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}^{-1} & \beta \boldsymbol{\phi}_{N+1} \\ \beta \boldsymbol{\phi}_{N+1}^\top & -\beta \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}' & \mathbf{S}' \boldsymbol{\phi}_{N+1} \\ \boldsymbol{\phi}_{N+1}^\top \mathbf{S}' & -\beta^{-1} + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}' \boldsymbol{\phi}_{N+1} \end{bmatrix}, \quad (3.75)$$

and

$$\begin{bmatrix} -\beta & \beta \boldsymbol{\phi}_{N+1}^\top \\ \beta \boldsymbol{\phi}_{N+1} & \mathbf{S}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} -c & c \beta \boldsymbol{\phi}_{N+1}^\top \mathbf{S} \\ c \beta \mathbf{S} \boldsymbol{\phi}_{N+1} & \mathbf{S} - c \beta^2 \mathbf{S} \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top \mathbf{S} \end{bmatrix}, \quad (3.76)$$

where

$$c = (\beta + \beta^2 \boldsymbol{\phi}_{N+1}^\top \mathbf{S} \boldsymbol{\phi}_{N+1})^{-1}. \quad (3.77)$$

Then,

$$\mathbf{S}' = \mathbf{S} - c \beta^2 \mathbf{S} \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top \mathbf{S}. \quad (3.78)$$

Then,

$$\sigma^2(\boldsymbol{\phi}) - \sigma'^2(\boldsymbol{\phi}) = c \beta^2 (\boldsymbol{\phi}^\top \mathbf{S} \boldsymbol{\phi}_{N+1})^2. \quad (3.79)$$

Therefore,

$$\sigma^2(\boldsymbol{\phi}) \geq \sigma'^2(\boldsymbol{\phi}). \quad (3.80)$$

### 3.12

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0), \end{aligned} \quad (3.81)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions.

(a)

By the Bayes' theorem,

$$p(\mathbf{w}, \beta | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta). \quad (3.82)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{t}$ ,  $\mathbf{w}$  and  $\beta$  can be written as

$$\begin{aligned} & -\frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 \\ & - \frac{M}{2} \ln \beta^{-1} - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) + (a_0 - 1) \ln \beta - b_0 \beta. \end{aligned} \quad (3.83)$$

We have

$$\begin{aligned} & -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ & = -\frac{\beta}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi} & -\boldsymbol{\Phi}^\top \\ -\boldsymbol{\Phi} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix} + \beta \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_0^{-1} \mathbf{m}_0 \\ \mathbf{0} \end{bmatrix} \\ & - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0. \end{aligned} \quad (3.84)$$

The right hand side can be written as

$$-\frac{\beta}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}) + \frac{\beta}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0, \quad (3.85)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.86)$$

Therefore,

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \beta^{-1} \mathbf{S}) \text{Gam}(\beta | a, b), \quad (3.87)$$

where

$$\begin{aligned} a &= a_0 + \frac{N}{2}, \\ b &= b_0 + \frac{1}{2} \|\mathbf{t}\|^2 + \frac{1}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 - \frac{1}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}. \end{aligned} \quad (3.88)$$

(b)

By the expression of the Byes' theorem in (a),

$$p(\mathbf{t}) = \frac{\mathcal{N}(\mathbf{t} | \Phi \mathbf{w}, \beta^{-1} \mathbf{I}) \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0)}{\mathcal{N}(\mathbf{w} | \mathbf{m}, \beta^{-1} \mathbf{S}) \text{Gam}(\beta | a, b)}. \quad (3.89)$$

The logarithm of the right hand side can be written as

$$\begin{aligned} & -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 \\ & - \frac{M}{2} \ln(2\pi) - \frac{M}{2} \ln \beta^{-1} - \frac{1}{2} \det \mathbf{S}_0 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ & + a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1) \ln \beta - b_0 \beta \\ & + \frac{M}{2} \ln(2\pi) + \frac{M}{2} \ln \beta^{-1} + \frac{1}{2} \det \mathbf{S} + \frac{\beta}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}) \\ & - a \ln b + \ln \Gamma(a) - (a - 1) \ln \beta + b \beta \\ & = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 \\ & - \frac{1}{2} \det \mathbf{S}_0 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) + a_0 \ln b_0 - \ln \Gamma(a_0) \\ & + \frac{1}{2} \det \mathbf{S} + \frac{\beta}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}) - a \ln b + \ln \Gamma(a) \\ & - (a - a_0) \ln \beta + (b - b_0) \beta. \end{aligned} \quad (3.90)$$

The right hand side can be written as

$$\begin{aligned} & -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \det \mathbf{S}_0 + a_0 \ln b_0 - \ln \Gamma(a_0) \\ & + \frac{1}{2} \det \mathbf{S} - a \ln b + \ln \Gamma(a). \end{aligned} \quad (3.91)$$

Therefore,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left( \frac{\det \mathbf{S}}{\det \mathbf{S}_0} \right)^{\frac{1}{2}} \frac{\Gamma(a)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b^a}. \quad (3.92)$$

### 3.13

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0), \end{aligned} \quad (3.93)$$

where  $\mathbf{w}$  and  $\boldsymbol{\phi}$  are vectors in  $M$  dimensions. By 3.12(a),

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \beta^{-1} \mathbf{S}) \text{Gam}(\beta | a, b), \quad (3.94)$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S} &= (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \\ a &= a_0 + \frac{N}{2}, \\ b &= b_0 + \frac{1}{2} \|\mathbf{t}\|^2 + \frac{1}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 - \frac{1}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}. \end{aligned} \quad (3.95)$$

By marginalisation,

$$p(t_{N+1} | \mathbf{t}) = \int \int p(t_{N+1} | \mathbf{w}, \beta) p(\mathbf{w}, \beta | \mathbf{t}) d\mathbf{w} d\beta. \quad (3.96)$$

The right hand side can be written as

$$\int \left( \int \mathcal{N}(t_{N+1} | \mathbf{w}^\top \boldsymbol{\phi}_{N+1}, \beta^{-1}) \mathcal{N}(\mathbf{w} | \mathbf{m}, \beta^{-1} \mathbf{S}) d\mathbf{w} \right) \text{Gam}(\beta | a, b) d\beta. \quad (3.97)$$

The logarithm of the integrand with respect to  $\mathbf{w}$  except the terms independent of  $\mathbf{w}$  can be written as

$$-\frac{\beta}{2} (t_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}_{N+1})^2 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}). \quad (3.98)$$

It can be written as

$$\begin{aligned} &-\frac{\beta}{2} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}^{-1} + \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\top & -\boldsymbol{\phi}_{N+1} \\ -\boldsymbol{\phi}_{N+1}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix} \\ &+ \beta \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}^{-1} \mathbf{m} \\ 0 \end{bmatrix} - \frac{\beta}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}. \end{aligned} \quad (3.99)$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}^{-1} + \boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\top & -\boldsymbol{\phi}_{N+1} \\ -\boldsymbol{\phi}_{N+1}^\top & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S} & \mathbf{S}\boldsymbol{\phi}_{N+1} \\ \boldsymbol{\phi}_{N+1}^\top \mathbf{S} & 1 + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}\boldsymbol{\phi}_{N+1} \end{bmatrix}. \quad (3.100)$$

Then,

$$\begin{bmatrix} \mathbf{S}^{-1} + \boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\top & -\boldsymbol{\phi}_{N+1} \\ -\boldsymbol{\phi}_{N+1}^\top & 1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}^{-1}\mathbf{m} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m}^\top \boldsymbol{\phi}_{N+1} \end{bmatrix}. \quad (3.101)$$

Then, the integral with respect to  $\mathbf{w}$  can be written as

$$\mathcal{N}(t_{N+1} | \mathbf{m}^\top \boldsymbol{\phi}_{N+1}, \beta^{-1} (1 + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}\boldsymbol{\phi}_{N+1})). \quad (3.102)$$

Then, the logarithm of the integrand with respect to  $\beta$  except the terms independent of  $\beta$  can be written as

$$\begin{aligned} & -\frac{1}{2} \ln \beta^{-1} - \frac{\beta}{2(1 + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}\boldsymbol{\phi}_{N+1})} (t - \mathbf{m}^\top \boldsymbol{\phi}_{N+1})^2 \\ & + (a-1) \ln \beta - b\beta \\ & = \left( a + \frac{1}{2} - 1 \right) \ln \beta - \left( b + \frac{(t - \mathbf{m}^\top \boldsymbol{\phi}_{N+1})^2}{2(1 + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}\boldsymbol{\phi}_{N+1})} \right) \beta. \end{aligned} \quad (3.103)$$

Then, the integral with respect to  $\beta$  except the terms independent of  $t_{N+1}$  can be written as

$$\left( b + \frac{(t - \mathbf{m}^\top \boldsymbol{\phi}_{N+1})^2}{2(1 + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}\boldsymbol{\phi}_{N+1})} \right)^{-a-\frac{1}{2}}. \quad (3.104)$$

Therefore,

$$p(t_{N+1} | \mathbf{t}) = \text{St}(t_{N+1} | \mu, \lambda, \nu), \quad (3.105)$$

where

$$\begin{aligned} \mu &= \mathbf{m}^\top \boldsymbol{\phi}_{N+1}, \\ \lambda &= \frac{a}{b} (1 + \boldsymbol{\phi}_{N+1}^\top \mathbf{S}\boldsymbol{\phi}_{N+1})^{-1}, \\ \nu &= 2a. \end{aligned} \quad (3.106)$$

### 3.14 (Incomplete)

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.107)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions.

(a)

By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.108)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.109)$$

Let

$$y(\boldsymbol{\phi}, \mathbf{w}) = \mathbf{w}^\top \boldsymbol{\phi}. \quad (3.110)$$

Then,

$$y(\boldsymbol{\phi}, \mathbf{m}) = \beta \boldsymbol{\phi}^\top \mathbf{S} \boldsymbol{\Phi}^\top \mathbf{t}. \quad (3.111)$$

The right hand side can be written as

$$\sum_{n=1}^N k(\boldsymbol{\phi}, \boldsymbol{\phi}_n) t_n, \quad (3.112)$$

where

$$k(\boldsymbol{\phi}, \boldsymbol{\phi}') = \beta \boldsymbol{\phi}^\top \mathbf{S} \boldsymbol{\phi}'. \quad (3.113)$$

Let us suppose that  $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_M$  are linearly independent,  $N > M$  and

$$\boldsymbol{\phi}_1 = 1. \quad (3.114)$$

Then, we can construct a new basis set  $\psi_1, \dots, \psi_M$  such that

$$\boldsymbol{\Psi}^\top \boldsymbol{\Psi} = \mathbf{I} \quad (3.115)$$

where

$$\psi_1 = 1. \quad (3.116)$$

(b)

### 3.15

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.117)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.118)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.119)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}) - E(\mathbf{m}), \quad (3.120)$$

where

$$E(\mathbf{m}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}\|^2 + \frac{\alpha}{2} \mathbf{m}^\top \mathbf{m}. \quad (3.121)$$

By 3.22, setting the derivatives of  $\ln p(\mathbf{t})$  with respect to  $\alpha$  and  $\beta$  to zero gives

$$\begin{aligned} \alpha &= \frac{\gamma}{\mathbf{m}^\top \mathbf{m}}, \\ \beta &= \frac{N - \gamma}{\|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}\|^2}, \end{aligned} \quad (3.122)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}, \quad (3.123)$$

and  $\lambda_1, \dots, \lambda_M$  are the eigenvalues of  $\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ . If  $\alpha$  and  $\beta$  are set as above, then

$$E(\mathbf{m}) = \frac{N}{2}. \quad (3.124)$$

### 3.16

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.125)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \quad (3.126)$$

The logarithm of the integrand of the right hand side except the terms independent of  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} \\ &= -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\top \boldsymbol{\Phi} & -\beta\boldsymbol{\Phi}^\top \\ -\beta\boldsymbol{\Phi} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}. \end{aligned} \quad (3.127)$$

By 2.24,

$$\begin{bmatrix} \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\top \boldsymbol{\Phi} & -\beta\boldsymbol{\Phi}^\top \\ -\beta\boldsymbol{\Phi} & \beta\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^\top \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \beta^{-1}\mathbf{I} \end{bmatrix}. \quad (3.128)$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \beta^{-1}\mathbf{I}). \quad (3.129)$$

### 3.17

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.130)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \quad (3.131)$$

The logarithm of the integrand of the right hand side can be written as

$$\begin{aligned} & -\frac{N}{2} \ln(2\pi\beta^{-1}) - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 \\ & - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\alpha^{-1}\mathbf{I})) - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \end{aligned} \quad (3.132)$$

Therefore,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp(-E(\mathbf{w})) d\mathbf{w}, \quad (3.133)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.134)$$

### 3.18

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.135)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.136)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.137)$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp(-E(\mathbf{w})) d\mathbf{w}, \quad (3.138)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.139)$$

The expression of  $E(\mathbf{w})$  can be written as

$$\begin{aligned} & \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m} - \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m})\|^2 + \frac{\alpha}{2} (\mathbf{w} - \mathbf{m} + \mathbf{m})^\top (\mathbf{w} - \mathbf{m} + \mathbf{m}) \\ & = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}\|^2 - \beta (\mathbf{t} - \boldsymbol{\Phi} \mathbf{m})^\top \boldsymbol{\Phi} (\mathbf{w} - \mathbf{m}) + \frac{\beta}{2} \|\boldsymbol{\Phi} (\mathbf{w} - \mathbf{m})\|^2 \\ & + \frac{\alpha}{2} \|(\mathbf{w} - \mathbf{m})\|^2 + \alpha \mathbf{m}^\top (\mathbf{w} - \mathbf{m}) + \frac{\alpha}{2} \|\mathbf{m}\|^2. \end{aligned} \quad (3.140)$$

Here,

$$\begin{aligned} & -\beta(\mathbf{t} - \Phi\mathbf{m})^\top \Phi(\mathbf{w} - \mathbf{m}) + \alpha\mathbf{m}^\top (\mathbf{w} - \mathbf{m}) \\ & = (-\beta\Phi^\top \mathbf{t} + \beta\Phi^\top \Phi\mathbf{m} + \alpha\mathbf{m})^\top (\mathbf{w} - \mathbf{m}). \end{aligned} \quad (3.141)$$

By the definitions of  $\mathbf{m}_N$  and  $\mathbf{S}_N$  above, the right hand can be written as

$$(-\mathbf{S}^{-1}\mathbf{m} + \mathbf{S}^{-1}\mathbf{m})^\top (\mathbf{w} - \mathbf{m}) = 0. \quad (3.142)$$

Therefore,

$$E(\mathbf{w}) = E(\mathbf{m}) + \frac{1}{2}(\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}). \quad (3.143)$$

### 3.19

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.144)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad (3.145)$$

where

$$\begin{aligned} \mathbf{m} &= \beta\mathbf{S}\Phi^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha\mathbf{I} + \beta\Phi^\top \Phi)^{-1}. \end{aligned} \quad (3.146)$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp(-E(\mathbf{w})) d\mathbf{w}, \quad (3.147)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\top \mathbf{w}. \quad (3.148)$$

By 3.18,

$$E(\mathbf{w}) = E(\mathbf{m}) + \frac{1}{2}(\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}). \quad (3.149)$$

Therefore, the integral in the expression above of  $p(\mathbf{t})$  can be written as

$$\begin{aligned} & \exp(-E(\mathbf{m})) \int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1}(\mathbf{w} - \mathbf{m})\right) d\mathbf{w} \\ & = (2\pi)^{\frac{M}{2}} (\det \mathbf{S})^{\frac{1}{2}} \exp(-E(\mathbf{m})). \end{aligned} \quad (3.150)$$

Thus,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}) - E(\mathbf{m}). \quad (3.151)$$

### 3.20

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.152)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.153)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.154)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}) - E(\mathbf{m}), \quad (3.155)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.156)$$

Let  $\mathbf{u}_1, \dots, \mathbf{u}_M$  be eigenvectors of  $\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$  such that

$$\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{u}_m = \lambda_m \mathbf{u}_m. \quad (3.157)$$

Then,

$$\mathbf{S}^{-1} \mathbf{u}_m = (\alpha + \lambda_m) \mathbf{u}_m, \quad (3.158)$$

so that

$$\det \mathbf{S} = \prod_{m=1}^M \frac{1}{\alpha + \lambda_m}. \quad (3.159)$$

Setting the derivative of  $\ln p(\mathbf{t})$  with respect to  $\alpha$  to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \sum_{m=1}^M \frac{1}{\alpha + \lambda_m} - \frac{1}{2} \mathbf{m}^\top \mathbf{m}. \quad (3.160)$$

Then,

$$\alpha \mathbf{m}^\top \mathbf{m} = M - \sum_{m=1}^M \frac{\alpha}{\alpha + \lambda_m}. \quad (3.161)$$

The right hand side can be written as

$$\sum_{m=1}^M \left(1 - \frac{\alpha}{\alpha + \lambda_m}\right) = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.162)$$

Therefore,

$$\alpha = \frac{\gamma}{\mathbf{m}^\top \mathbf{m}}, \quad (3.163)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.164)$$

### 3.21

(a)

Let  $\Sigma$  be a  $M \times M$  real symmetric matrix such that

$$\Sigma \mathbf{u}_m = \lambda_m \mathbf{u}_m, \quad (3.165)$$

where  $\mathbf{u}_1, \dots, \mathbf{u}_M$  are unit vectors. Let

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_M), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_M]. \end{aligned} \quad (3.166)$$

By 2.19,

$$\begin{aligned} \Sigma &= \mathbf{U} \Lambda \mathbf{U}^\top, \\ \mathbf{U}^\top \mathbf{U} &= \mathbf{I}. \end{aligned} \quad (3.167)$$

Then,

$$\det \Sigma = \prod_{m=1}^M \lambda_m, \quad (3.168)$$

so that

$$\ln(\det \Sigma) = \sum_{m=1}^M \ln \lambda_m. \quad (3.169)$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \Sigma) = \sum_{m=1}^M \frac{\partial \lambda_m}{\partial \alpha} \frac{1}{\lambda_m}. \quad (3.170)$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \Sigma) = \text{tr} \left( \Lambda^{-1} \frac{\partial \Lambda}{\partial \alpha} \right). \quad (3.171)$$

The right hand side can be written as

$$\text{tr} \left( \mathbf{U} \Lambda^{-1} \mathbf{U}^\top \frac{\partial \mathbf{U} \Lambda \mathbf{U}^\top}{\partial \alpha} \right) = \text{tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial \alpha} \right). \quad (3.172)$$

Therefore,

$$\frac{\partial}{\partial \alpha} \ln(\det \Sigma) = \text{tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial \alpha} \right). \quad (3.173)$$

(b)

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.174)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.175)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}. \end{aligned} \quad (3.176)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}) - E(\mathbf{m}), \quad (3.177)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.178)$$

By 3.21(a),

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{S}^{-1}) = \text{tr } \mathbf{S}. \quad (3.179)$$

The right hand side can be written as

$$\sum_{m=1}^M \frac{1}{\alpha + \lambda_m}, \quad (3.180)$$

where  $\lambda_1, \dots, \lambda_M$  are eigenvalues of  $\beta\Phi^\top\Phi$ . Setting the derivative of  $\ln p(\mathbf{t})$  with respect to  $\alpha$  to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \sum_{m=1}^M \frac{1}{\alpha + \lambda_m} - \frac{1}{2} \mathbf{m}^\top \mathbf{m}, \quad (3.181)$$

Therefore,

$$\alpha = \frac{\gamma}{\mathbf{m}^\top \mathbf{m}}, \quad (3.182)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.183)$$

### 3.22

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.184)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (3.185)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \Phi^\top \mathbf{t}, \\ \mathbf{S} &= (\alpha \mathbf{I} + \beta \Phi^\top \Phi)^{-1}. \end{aligned} \quad (3.186)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}) - E(\mathbf{m}), \quad (3.187)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.188)$$

By 3.21(a),

$$\frac{\partial}{\partial \beta} \ln(\det \mathbf{S}^{-1}) = \text{tr}(\mathbf{S} \Phi^\top \Phi). \quad (3.189)$$

Since

$$\mathbf{S} \Phi^\top \Phi = \frac{1}{\beta} (\mathbf{I} - \alpha \mathbf{S}), \quad (3.190)$$

the right hand side can be written as

$$\frac{1}{\beta} \left( M - \alpha \sum_{m=1}^M \frac{1}{\alpha + \lambda_m} \right) = \frac{1}{\beta} \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}, \quad (3.191)$$

where  $\lambda_1, \dots, \lambda_M$  are eigenvalues of  $\beta \Phi^\top \Phi$ . Setting the derivative of  $\ln p(\mathbf{t})$  with respect to  $\beta$  to zero gives

$$0 = \frac{N}{2\beta} - \frac{1}{2\beta} \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m} - \frac{1}{2} \|\mathbf{t} - \Phi \mathbf{m}\|^2. \quad (3.192)$$

Therefore,

$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \Phi \mathbf{m}\|^2}, \quad (3.193)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.194)$$

### 3.23

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0), \end{aligned} \quad (3.195)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By marginalisation,

$$p(\mathbf{t}) = \int \int p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta) d\mathbf{w} d\beta. \quad (3.196)$$

The right hand side can be written as

$$\int \left( \int \left( \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}) \right) \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) d\mathbf{w} \right) \text{Gam}(\beta | a_0, b_0) d\beta. \quad (3.197)$$

The logarithm of the integrand with respect to  $\mathbf{w}$  can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 \\
& - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln \det(\beta^{-1} \mathbf{S}_0) - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\
= & -\frac{N+M}{2} \ln(2\pi) + \frac{N+M}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) \\
& - \frac{\beta}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi} & -\boldsymbol{\Phi}^\top \\ -\boldsymbol{\Phi} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_0^{-1} \mathbf{m}_0 \\ \mathbf{0} \end{bmatrix} \\
& - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0.
\end{aligned} \tag{3.198}$$

The right hand side can be written as

$$\begin{aligned}
& -\frac{N+M}{2} \ln(2\pi) + \frac{N+M}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) \\
& - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}) + \frac{\beta}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0,
\end{aligned} \tag{3.199}$$

where

$$\begin{aligned}
\mathbf{m} &= \mathbf{S} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}), \\
\mathbf{S} &= (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}.
\end{aligned} \tag{3.200}$$

Then, the logarithm of the integral with respect to  $\mathbf{w}$  can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) + \frac{N}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}) \\
& - \frac{\beta}{2} \|\mathbf{t}\|^2 + \frac{\beta}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0.
\end{aligned} \tag{3.201}$$

Then, the logarithm of the integrand with respect to  $\beta$  can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) + \frac{N}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}) \\
& + \frac{\beta}{2} \mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m} - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 \\
& - \ln \Gamma(a_0) + a_0 \ln b_0 + (a_0 - 1) \ln \beta - b_0 \beta \\
= & -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}) \\
& - \ln \Gamma(a_0) + a_0 \ln b_0 + (a - 1) \ln \beta - b \beta,
\end{aligned} \tag{3.202}$$

where

$$\begin{aligned} a &= a_0 + \frac{N}{2}, \\ b &= b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 - \frac{1}{2}\mathbf{m}^\top \mathbf{S}^{-1} \mathbf{m}. \end{aligned} \quad (3.203)$$

Then, the logarithm of the integral with respect to  $\beta$  can be written as

$$\begin{aligned} &- \frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}) \\ &- \ln \Gamma(a_0) + a_0 \ln b_0 + \ln \Gamma(a) - a \ln b. \end{aligned} \quad (3.204)$$

Therefore,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left( \frac{\det \mathbf{S}}{\det \mathbf{S}_0} \right)^{\frac{1}{2}} \frac{\Gamma(a)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b^a}. \quad (3.205)$$

### 3.24

Refer to 3.12.

## 4 Linear Models for Classification

### 4.1

Let  $x_1, \dots, x_M$  and  $y_1, \dots, y_N$  be two sets of data points. Then, the corresponding convex hulls are defined as the sets of all points  $\mathbf{x}$  and  $\mathbf{y}$  such that

$$\begin{aligned}\mathbf{x} &= \sum_{m=1}^M \alpha_m \mathbf{x}_m, \\ \mathbf{y} &= \sum_{n=1}^N \beta_n \mathbf{y}_n,\end{aligned}\tag{4.1}$$

where

$$\begin{aligned}\sum_{m=1}^M \alpha_m &= \sum_{n=1}^N \beta_n = 1, \\ \alpha_m &\geq 0, \beta_n \geq 0.\end{aligned}\tag{4.2}$$

Let us assume that  $\alpha_1, \dots, \alpha_M$  and  $\beta_1, \dots, \beta_N$  below are subject to the constraints above.

If the convex hulls intersect, then there exist  $\alpha_1, \dots, \alpha_M$  and  $\beta_1, \dots, \beta_N$  such that

$$\sum_{m=1}^M \alpha_m \mathbf{x}_m = \sum_{n=1}^N \beta_n \mathbf{y}_n.\tag{4.3}$$

Then,

$$\sum_{m=1}^M \alpha_m (\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0) = \hat{\mathbf{w}}^\top \sum_{m=1}^M \alpha_m \mathbf{x}_m + w_0 \sum_m \alpha_m,\tag{4.4}$$

for any  $\hat{\mathbf{w}}$  and  $w_0$ . The right hand side can be written as

$$\hat{\mathbf{w}}^\top \sum_{n=1}^N \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^N \beta_n = \sum_{n=1}^N \beta_n (\hat{\mathbf{w}}^\top \mathbf{y}_n + w_0).\tag{4.5}$$

Therefore, there do not exist  $\hat{\mathbf{w}}$  and  $w_0$  such that

$$\begin{aligned}\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0 &> 0, \\ \hat{\mathbf{w}}^\top \mathbf{y}_n + w_0 &< 0.\end{aligned}\tag{4.6}$$

Conversely, if there exist  $\hat{\mathbf{w}}$  and  $w_0$  such that

$$\begin{aligned}\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0 &> 0, \\ \hat{\mathbf{w}}^\top \mathbf{y}_n + w_0 &< 0,\end{aligned}\tag{4.7}$$

then

$$\begin{aligned}\sum_{m=1}^M \alpha_m (\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0) &> 0, \\ \sum_{n=1}^N \beta_n (\hat{\mathbf{w}}^\top \mathbf{y}_n + w_0) &< 0.\end{aligned}\tag{4.8}$$

The left hand sides can be written as

$$\begin{aligned}\hat{\mathbf{w}}^\top \sum_{m=1}^M \alpha_m \mathbf{x}_m + w_0 \sum_{m=1}^M \alpha_m &= \hat{\mathbf{w}}^\top \sum_{m=1}^M \alpha_m \mathbf{x}_m + w_0, \\ \hat{\mathbf{w}}^\top \sum_{n=1}^N \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^N \beta_n &= \hat{\mathbf{w}}^\top \sum_{n=1}^N \beta_n \mathbf{y}_n + w_0.\end{aligned}\tag{4.9}$$

Therefore, there do not exist  $\alpha_1, \dots, \alpha_M$  and  $\beta_1, \dots, \beta_N$  such that

$$\sum_{m=1}^M \alpha_m \mathbf{x}_m = \sum_{n=1}^N \beta_n \mathbf{y}_n.\tag{4.10}$$

Thus, the convex hulls do not intersect.

## 4.2 (Incomplete)

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  and  $\mathbf{w}_1, \dots, \mathbf{w}_K$  are variables in  $M$  dimensions and  $\mathbf{t}_1, \dots, \mathbf{t}_N$  are ones in  $K$  dimensions. Let

$$E(\tilde{\mathbf{W}}) = \frac{1}{2} \text{tr} \left( (\tilde{\mathbf{X}} \tilde{\mathbf{W}} - \mathbf{T})^\top (\tilde{\mathbf{X}} \tilde{\mathbf{W}} - \mathbf{T}) \right),\tag{4.11}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\top \end{bmatrix},$$

$$\tilde{\mathbf{W}} = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^\top \\ \vdots \\ \mathbf{t}_N^\top \end{bmatrix}.$$

Setting the derivative with respect to  $\tilde{\mathbf{W}}$  to zero gives

$$\mathbf{O} = \tilde{\mathbf{X}}^\top (\tilde{\mathbf{X}} \tilde{\mathbf{W}} - \mathbf{T}). \quad (4.12)$$

Therefore,

$$\underset{\tilde{\mathbf{W}}}{\operatorname{argmin}} E(\tilde{\mathbf{W}}) = \left( \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^\top \mathbf{T}. \quad (4.13)$$

Let  $\tilde{\mathbf{W}}^*$  denote the least-square solution above. Then,

$$(\tilde{\mathbf{W}}^*)^\top \tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{T}^\top \tilde{\mathbf{X}} \left( \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{x}} - \mathbf{t}_n, \quad (4.14)$$

where  $\tilde{\mathbf{x}}$  is a vector in  $M + 1$  dimensions whose first element is 1. The right hand side can be written as

$$\mathbf{T}^\top \left( \tilde{\mathbf{X}} \left( \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{x}} - \mathbf{v}_n \right) = \mathbf{0}? \quad (4.15)$$

where  $\mathbf{v}_n$  is a vector in  $N$  dimensions whose  $n$  th element is 1 and other elements are zero. Therefore,

$$(\tilde{\mathbf{W}}^*)^\top \tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{0}. \quad (4.16)$$

Thus, if

$$\mathbf{a}^\top \mathbf{t}_n + b = 0, \quad (4.17)$$

then

$$\mathbf{a}^\top (\tilde{\mathbf{W}}^*)^\top \tilde{\mathbf{x}} + b = 0. \quad (4.18)$$

### 4.3 (Incomplete)

#### 4.4

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \quad (4.19)$$

where  $N_k$  is the number of  $\mathbf{x}_n$  such that  $n$  is in  $\mathcal{C}_k$ . Setting the derivatives of

$$\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1) + \lambda (\|\mathbf{w}\|^2 - 1) \quad (4.20)$$

with respect to  $\mathbf{w}$  and  $\lambda$  to zero gives

$$\begin{aligned} \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} &= \mathbf{0}, \\ \|\mathbf{w}\|^2 - 1 &= 0. \end{aligned} \quad (4.21)$$

Therefore,  $\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1)$  under the constraint

$$\|\mathbf{w}\|^2 = 1 \quad (4.22)$$

is maximised if

$$\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1. \quad (4.23)$$

## 4.5

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \quad (4.24)$$

where  $N_k$  is the number of  $\mathbf{x}_n$  such that  $n$  is in  $\mathcal{C}_k$ . Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \quad (4.25)$$

where

$$\begin{aligned} s_k^2 &= \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2, \\ y_n &= \mathbf{w}^\top \mathbf{x}_n, \\ m_k &= \mathbf{w}^\top \mathbf{m}_k. \end{aligned} \quad (4.26)$$

Then,  $J(\mathbf{w})$  can be written as

$$\frac{(\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1))^2}{\sum_{n \in \mathcal{C}_1} (\mathbf{w}^\top(\mathbf{x}_n - \mathbf{m}_1))^2 + \sum_{n \in \mathcal{C}_2} (\mathbf{w}^\top(\mathbf{x}_n - \mathbf{m}_2))^2} = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}, \quad (4.27)$$

where

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top, \\ \mathbf{S}_W &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top. \end{aligned} \quad (4.28)$$

## 4.6

Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \quad (4.29)$$

where  $N_k$  is the number of  $\mathbf{x}_n$  such that  $n$  is in  $\mathcal{C}_k$ . Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \quad (4.30)$$

where

$$\begin{aligned} s_k^2 &= \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2, \\ y_n &= \mathbf{w}^\top \mathbf{x}_n, \\ m_k &= \mathbf{w}^\top \mathbf{m}_k. \end{aligned} \quad (4.31)$$

Then, by 4.5,

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}, \quad (4.32)$$

where

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top, \\ \mathbf{S}_W &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top. \end{aligned} \quad (4.33)$$

Let

$$E = \frac{1}{2} \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n)^2, \quad (4.34)$$

where

$$t_n = \begin{cases} \frac{N}{N_1}, & n \in \mathcal{C}_1, \\ -\frac{N}{N_2}, & n \in \mathcal{C}_2. \end{cases} \quad (4.35)$$

Setting the derivative with respect to  $\mathbf{w}$  and  $w_0$  gives

$$\begin{aligned} 0 &= \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n), \\ \mathbf{0} &= \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n) \mathbf{x}_n. \end{aligned} \quad (4.36)$$

The right hand side of the first equation can be written as

$$\mathbf{w}^\top \sum_{n=1}^N \mathbf{x}_n + N w_0 - \sum_{n=1}^N t_n = N (\mathbf{w}^\top \mathbf{m} + w_0), \quad (4.37)$$

where

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n. \quad (4.38)$$

Therefore,

$$w_0 = -\mathbf{w}^\top \mathbf{m}. \quad (4.39)$$

Then, the right hand side of the second equation above can be written as

$$\begin{aligned} & \sum_{n=1}^N (\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}) - t_n) \mathbf{x}_n \\ &= \sum_{n \in \mathcal{C}_1} \left( \mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}) - \frac{N}{N_1} \right) \mathbf{x}_n + \sum_{n \in \mathcal{C}_2} \left( \mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}) + \frac{N}{N_2} \right) \mathbf{x}_n. \end{aligned} \quad (4.40)$$

Since

$$\begin{aligned} \mathbf{m} &= \frac{N_1}{N} \mathbf{m}_1 + \frac{N_2}{N} \mathbf{m}_2, \\ \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1) &= \mathbf{0}, \end{aligned} \quad (4.41)$$

the first term of the right hand side can be written as

$$\begin{aligned} & \sum_{n \in \mathcal{C}_1} \left( \mathbf{w}^\top \left( \mathbf{x}_n - \mathbf{m}_1 + \frac{N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \right) - \frac{N}{N_1} \right) (\mathbf{x}_n - \mathbf{m}_1 + \mathbf{m}_1) \\ &= \left( \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^\top \mathbf{w} - N \mathbf{m}_1. \end{aligned} \quad (4.42)$$

Similarly, the second term can be written as

$$\left( \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^\top \mathbf{w} - N \mathbf{m}_2. \quad (4.43)$$

Therefore,

$$\begin{aligned} \mathbf{0} = & \left( \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^\top \mathbf{w} - N \mathbf{m}_1 \\ & + \left( \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^\top \mathbf{w} - N \mathbf{m}_2. \end{aligned} \quad (4.44)$$

Thus,

$$\left( \mathbf{S}_W + \frac{N_1 N_2}{N} \mathbf{S}_B \right) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2). \quad (4.45)$$

## 4.7

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (4.46)$$

Then,

$$\sigma(-a) = \frac{1}{1 + \exp(a)}. \quad (4.47)$$

The right hand side can be written as

$$1 - \frac{\exp(a)}{1 + \exp(a)} = 1 - \frac{1}{1 + \exp(-a)}. \quad (4.48)$$

Therefore,

$$\sigma(-a) = 1 - \sigma(a). \quad (4.49)$$

Additionally,

$$\exp(-a) = \frac{1}{\sigma(a)} - 1. \quad (4.50)$$

Then,

$$a = -\ln \left( \frac{1}{\sigma(a)} - 1 \right). \quad (4.51)$$

Therefore,

$$\sigma^{-1}(y) = \ln \left( \frac{y}{1-y} \right). \quad (4.52)$$

## 4.8

Let  $\mathbf{x}$  be a variable in  $D$  dimensions such that

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad (4.53)$$

where

$$p(\mathcal{C}_1) + p(\mathcal{C}_2) = 1. \quad (4.54)$$

By the Bayes' theorem,

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}. \quad (4.55)$$

The right hand side can be written as

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \quad (4.56)$$

where

$$a = \ln \left( \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \right). \quad (4.57)$$

Substituting the expressions above of  $p(\mathbf{x}|\mathcal{C}_k)$ , we have

$$\begin{aligned} a &= -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \ln p(\mathcal{C}_1) \\ &\quad + \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \boldsymbol{\Sigma}) + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) - \ln p(\mathcal{C}_2). \end{aligned} \quad (4.58)$$

Therefore,

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0), \quad (4.59)$$

where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ w_0 &= -\frac{1}{2} \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2). \end{aligned} \quad (4.60)$$

## 4.9

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n, \phi_n) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (4.61)$$

where

$$\begin{aligned} y_{nk} &= p(\phi_n, \mathcal{C}_k), \\ p(\mathcal{C}_k) &= \pi_k, \\ \sum_{k=1}^K \pi_k &= 1. \end{aligned} \tag{4.62}$$

Then,

$$p(\mathbf{T}, \Phi) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}, \tag{4.63}$$

By the Bayes' theorem,

$$y_{nk} = \pi_k p(\phi_n | \mathcal{C}_k). \tag{4.64}$$

Then,

$$\ln p(\mathbf{T}, \Phi) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \pi_k + \ln p(\phi_n | \mathcal{C}_k)). \tag{4.65}$$

Setting the derivatives of

$$\ln p(\mathbf{T}, \Phi) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right) \tag{4.66}$$

with respect to  $\pi_k$  and  $\lambda$  to zero gives

$$\begin{aligned} 0 &= \frac{1}{\pi_k} \sum_{n=1}^N t_{nk} + \lambda, \\ 0 &= \sum_{k=1}^K \pi_k - 1. \end{aligned} \tag{4.67}$$

Then,

$$\lambda = - \sum_{n=1}^N \sum_{k=1}^K t_{nk}. \tag{4.68}$$

Since  $\mathbf{t}_1, \dots, \mathbf{t}_N$  are from the standard basis in  $K$  dimensions, the right hand side can be written as  $-N$ . Therefore, the maximum likelihood solution for  $\pi_k$  is given by

$$\pi_{k \text{ML}} = \frac{N_k}{N}, \tag{4.69}$$

where

$$N_k = \sum_{n=1}^N t_{nk}. \quad (4.70)$$

## 4.10

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (4.71)$$

where

$$\begin{aligned} y_{nk} &= p(\boldsymbol{\phi}_n, \mathcal{C}_k), \\ p(\boldsymbol{\phi}_n | \mathcal{C}_k) &= \mathcal{N}(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \\ p(\mathcal{C}_k) &= \pi_k, \\ \sum_{k=1}^K \pi_k &= 1. \end{aligned} \quad (4.72)$$

Then,

$$p(\mathbf{T}, \boldsymbol{\Phi}) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (4.73)$$

By the Bayes' theorem,

$$y_{nk} = \pi_k \mathcal{N}(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}). \quad (4.74)$$

Then,

$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \mathcal{N}(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) + \ln \pi_k). \quad (4.75)$$

The right hand side except the terms independent of  $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}$  can be written as

$$\sum_{n=1}^N \sum_{k=1}^K t_{nk} \left( -\frac{1}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) \right). \quad (4.76)$$

By 3.21(a), setting the derivatives with respect to  $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}$  to zero gives

$$\begin{aligned}\mathbf{0} &= \frac{1}{2} \sum_{n=1}^N t_{nk} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k), \\ \mathbf{O} &= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \left( (\boldsymbol{\Sigma}^{-1})^\top - (\boldsymbol{\Sigma}^{-1})^2 (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\top \right).\end{aligned}\quad (4.77)$$

Therefore, the maximum likelihood solutions for  $\boldsymbol{\mu}_k$  and  $\boldsymbol{\Sigma}$  are given by

$$\begin{aligned}\boldsymbol{\mu}_{k\text{ML}} &= \frac{1}{N_k} \sum_{n=1}^N t_{nk} \boldsymbol{\phi}_n, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{k=1}^K N_k \mathbf{S}_k,\end{aligned}\quad (4.78)$$

where

$$\begin{aligned}N_k &= \sum_{n=1}^N t_{nk}, \\ \mathbf{S}_k &= \frac{1}{N_k} \sum_{n=1}^N t_{nk} (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\top.\end{aligned}\quad (4.79)$$

## 4.11

Let  $\boldsymbol{\phi}$  be a variable in  $M$  dimensions whose each component is a binary code with length  $L$  such that

$$p(\boldsymbol{\phi}|\mathcal{C}_k) = \prod_{m=1}^M \prod_{l=1}^L \mu_{kml}^{\phi_{ml}}, \quad (4.80)$$

where

$$\sum_{k=1}^K p(\mathcal{C}_k) = 1. \quad (4.81)$$

By the Bayes' theorem,

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = \frac{p(\boldsymbol{\phi}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{k'=1}^K p(\boldsymbol{\phi}|\mathcal{C}_{k'})p(\mathcal{C}_{k'})}. \quad (4.82)$$

Therefore,

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = \frac{\exp(a_k(\boldsymbol{\phi}))}{\sum_{k'=1}^K \exp(a_{k'}(\boldsymbol{\phi}))}, \quad (4.83)$$

where

$$a_k(\boldsymbol{\phi}) = \left( \sum_{m=1}^M \sum_{l=1}^L \phi_{ml} \ln \mu_{kml} \right) + \ln p(\mathcal{C}_k). \quad (4.84)$$

## 4.12

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (4.85)$$

Then,

$$\frac{d\sigma(a)}{da} = \frac{\exp(-a)}{(1 + \exp(-a))^2}. \quad (4.86)$$

The right hand side can be written as

$$\frac{1}{1 + \exp(-a)} - \frac{1}{(1 + \exp(-a))^2} = \sigma(a) - (\sigma(a))^2. \quad (4.87)$$

Therefore,

$$\frac{d\sigma(a)}{da} = \sigma(a) (1 - \sigma(a)). \quad (4.88)$$

## 4.13

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n} (1 - y_n)^{1-t_n}, \quad (4.89)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.90)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \quad (4.91)$$

The right hand side can be written as

$$-\ln \left( \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n} \right) = -\sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)). \quad (4.92)$$

By 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \left( \frac{t_n}{y_n} y_n (1 - y_n) \boldsymbol{\phi}_n - \frac{1 - t_n}{1 - y_n} y_n (1 - y_n) \boldsymbol{\phi}_n \right). \quad (4.93)$$

The right hand side can be written as

$$-\sum_{n=1}^N (t_n (1 - y_n) \boldsymbol{\phi}_n - (1 - t_n) y_n \boldsymbol{\phi}_n) = \sum_{n=1}^N (y_n - t_n) \boldsymbol{\phi}_n. \quad (4.94)$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \boldsymbol{\phi}_n. \quad (4.95)$$

#### 4.14

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n | \mathbf{w}) = y_n^{t_n} (1 - y_n)^{1-t_n}, \quad (4.96)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.97)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}). \quad (4.98)$$

By 4.13, setting the derivative with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = \sum_{n=1}^N (y_n - t_n) \boldsymbol{\phi}_n. \quad (4.99)$$

If  $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N$  are linearly independent, then

$$y_n = t_n. \quad (4.100)$$

Then,

$$\sigma(\mathbf{w}^\top \boldsymbol{\phi}_n) = \begin{cases} 1, & t_n = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (4.101)$$

Therefore,

$$\mathbf{w}^\top \boldsymbol{\phi}_n = \begin{cases} \infty, & t_n = 1, \\ -\infty, & \text{otherwise.} \end{cases} \quad (4.102)$$

## 4.15

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n | \mathbf{w}) = y_n^{t_n} (1 - y_n)^{1-t_n}, \quad (4.103)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.104)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}). \quad (4.105)$$

By 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \boldsymbol{\phi}_n. \quad (4.106)$$

By 4.12,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n (1 - y_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\top. \quad (4.107)$$

The right hand side can be written as

$$\mathbf{H} = \boldsymbol{\Phi}^\top \mathbf{R} \boldsymbol{\Phi}, \quad (4.108)$$

where

$$R_{nn'} = y_n (1 - y_n) I_{nn'}, \quad (4.109)$$

Then,

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} = \mathbf{v}^\top \mathbf{R} \mathbf{v}, \quad (4.110)$$

where

$$\mathbf{v} = \boldsymbol{\Phi} \mathbf{u}. \quad (4.111)$$

The right hand side can be written as

$$\sum_{n=1}^N \sum_{n'=1}^N v_n y_n (1 - y_n) I_{nn'} v_{n'} = \sum_{n=1}^N y_n (1 - y_n) v_n^2. \quad (4.112)$$

Since

$$y_n (1 - y_n) > 0, \quad (4.113)$$

we have

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} > 0. \quad (4.114)$$

Then,  $\mathbf{H}$  is positive definite. Therefore,  $E$  is a convex function of  $\mathbf{w}$  and it has a unique minimum.

## 4.16

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n = 1 | \boldsymbol{\phi}_n) = \pi_n. \quad (4.115)$$

Then,

$$p(t_n | \boldsymbol{\phi}_n) = \pi_n^{t_n} (1 - \pi_n)^{1-t_n}. \quad (4.116)$$

Then,

$$p(\mathbf{t} | \boldsymbol{\Phi}) = \prod_{n=1}^N \pi_n^{t_n} (1 - \pi_n)^{1-t_n}. \quad (4.117)$$

Therefore,

$$-\ln p(\mathbf{t} | \boldsymbol{\Phi}) = -\sum_{n=1}^N (t_n \ln \pi_n + (1 - t_n) \ln(1 - \pi_n)). \quad (4.118)$$

## 4.17

Let

$$y_k = \frac{\exp(a_k)}{\sum_{k'=1}^K \exp(a_{k'})}. \quad (4.119)$$

Then,

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k)}{\sum_{k'=1}^K \exp(a_{k'})} - \frac{\exp(2a_k)}{\left(\sum_{k'=1}^K \exp(a_{k'})\right)^2}. \quad (4.120)$$

The right hand side can be written as  $y_k(1 - y_k)$ . If  $k \neq k'$ , then

$$\frac{\partial y_k}{\partial a_{k'}} = -\frac{\exp(a_k + a_{k'})}{\left(\sum_{k''=1}^K \exp(a_{k''})\right)^2}. \quad (4.121)$$

The right hand side can be written as  $-y_k y_{k'}$ . Therefore,

$$\frac{\partial y_k}{\partial a_{k'}} = y_k(I_{kk'} - y_{k'}). \quad (4.122)$$

## 4.18

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (4.123)$$

where

$$\begin{aligned} y_{nk} &= \frac{\exp(a_{nk})}{\sum_{k'=1}^K \exp(a_{nk'})}, \\ a_{nk} &= \mathbf{w}_k^\top \boldsymbol{\phi}_n. \end{aligned} \quad (4.124)$$

Then,

$$p(\mathbf{T} | \mathbf{W}) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (4.125)$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T} | \mathbf{W}). \quad (4.126)$$

The right hand side can be written as

$$-\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (4.127)$$

By 4.17,

$$\nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = -\sum_{n=1}^N \sum_{k=1}^K y_{nk}(I_{kk'} - y_{nk'}) \frac{t_{nk}}{y_{nk}} \boldsymbol{\phi}_n. \quad (4.128)$$

Since  $\mathbf{t}_1, \dots, \mathbf{t}_N$  are variables from the standard basis in  $K$  dimensions, the right hand side can be written as

$$-\sum_{n=1}^N \left( \sum_{k=1}^K (I_{kk'} - y_{nk'}) t_{nk} \right) \boldsymbol{\phi}_n = -\sum_{n=1}^N (t_{nk'} - y_{nk'}) \boldsymbol{\phi}_n. \quad (4.129)$$

Therefore,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^N (y_{nk} - t_{nk}) \boldsymbol{\phi}_n. \quad (4.130)$$

## 4.19

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n = 1 | a_n) = \Phi(a_n), \quad (4.131)$$

where

$$\begin{aligned} \Phi(a) &= \int_{-\infty}^a \mathcal{N}(\theta | 0, 1) d\theta, \\ a_n &= \mathbf{w}^\top \boldsymbol{\phi}_n. \end{aligned} \quad (4.132)$$

(a)

We have

$$p(t_n | \boldsymbol{\phi}_n) = (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \quad (4.133)$$

Then,

$$p(\mathbf{t} | \boldsymbol{\Phi}) = \prod_{n=1}^N (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \quad (4.134)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \boldsymbol{\Phi}). \quad (4.135)$$

The right hand side can be written as

$$-\sum_{n=1}^N (t_n \ln \Phi(a_n) + (1 - t_n) \ln (1 - \Phi(a_n))). \quad (4.136)$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \left( t_n \frac{\mathcal{N}(a_n | 0, 1)}{\Phi(a_n)} - (1 - t_n) \frac{\mathcal{N}(a_n | 0, 1)}{1 - \Phi(a_n)} \right) \boldsymbol{\phi}_n. \quad (4.137)$$

The right hand side can be written as

$$\begin{aligned} & - \sum_{n=1}^N \left( \frac{t_n}{\Phi(a_n)} - \frac{1-t_n}{1-\Phi(a_n)} \right) \mathcal{N}(a_n|0,1) \phi_n \\ & = \sum_{n=1}^N \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)(1-\Phi(a_n))} (\Phi(a_n) - t_n) \phi_n. \end{aligned} \quad (4.138)$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)(1-\Phi(a_n))} (\Phi(a_n) - t_n) \phi_n. \quad (4.139)$$

(b)

We have

$$\begin{aligned} \nabla \nabla E(\mathbf{w}) & = \sum_{n=1}^N \frac{-a_n \mathcal{N}(a_n|0,1)}{\Phi(a_n)(1-\Phi(a_n))} (\Phi(a_n) - t_n) \phi_n \phi_n^\top \\ & \quad - \sum_{n=1}^N \frac{(\mathcal{N}(a_n|0,1))^2}{(\Phi(a_n))^2(1-\Phi(a_n))} (\Phi(a_n) - t_n) \phi_n \phi_n^\top \\ & \quad + \sum_{n=1}^N \frac{(\mathcal{N}(a_n|0,1))^2}{\Phi(a_n)(1-\Phi(a_n))^2} (\Phi(a_n) - t_n) \phi_n \phi_n^\top \\ & \quad + \sum_{n=1}^N \frac{(\mathcal{N}(a_n|0,1))^2}{\Phi(a_n)(1-\Phi(a_n))} \phi_n \phi_n^\top. \end{aligned} \quad (4.140)$$

Therefore,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N b_n \phi_n \phi_n^\top, \quad (4.141)$$

where

$$\begin{aligned} b_n & = \left( \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)(1-\Phi(a_n))} \right)^2 ((\Phi(a_n))^2 - 2t_n \Phi(a_n) + t_n) \\ & \quad - \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)(1-\Phi(a_n))} a_n (\Phi(a_n) - t_n). \end{aligned} \quad (4.142)$$

## 4.20

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (4.143)$$

where

$$\begin{aligned} y_{nk} &= \frac{\exp(a_{nk})}{\sum_{k'=1}^K \exp(a_{nk'})}, \\ a_{nk} &= \mathbf{w}_k^\top \boldsymbol{\phi}_n. \end{aligned} \quad (4.144)$$

Then,

$$p(\mathbf{T} | \mathbf{W}) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (4.145)$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T} | \mathbf{W}). \quad (4.146)$$

By 4.18,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^N (y_{nk} - t_{nk}) \boldsymbol{\phi}_n. \quad (4.147)$$

By 4.17,

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = \sum_{n=1}^N y_{nk} (I_{kk'} - y_{nk'}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\top. \quad (4.148)$$

The right hand side can be written as

$$\mathbf{H}_{kk'} = \boldsymbol{\Phi}^\top \mathbf{R}_{kk'} \boldsymbol{\Phi}, \quad (4.149)$$

where

$$R_{kk'nn'} = y_{nk} (I_{kk'} - y_{nk'}) I_{nn'}. \quad (4.150)$$

Let

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{K1} & \cdots & \mathbf{H}_{KK} \end{bmatrix}, \quad (4.151)$$

and

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_K \end{bmatrix}, \quad (4.152)$$

where  $\mathbf{u}_1, \dots, \mathbf{u}_K$  are vectors in the same dimension as  $\mathbf{w}$ . Then,

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} = \sum_{k=1}^K \sum_{k'=1}^K \mathbf{v}_k^\top \mathbf{R}_{kk'} \mathbf{v}_{k'}, \quad (4.153)$$

where

$$\mathbf{v}_k = \Phi \mathbf{u}_k. \quad (4.154)$$

The right hand side can be written as

$$\begin{aligned} & \sum_{k=1}^K \sum_{k'=1}^K \sum_{n=1}^N \sum_{n'=1}^N v_{kn} y_{nk} (I_{kk'} - y_{nk'}) I_{nn'} v_{k'n'} \\ &= \sum_{k=1}^K \sum_{k'=1}^K \sum_{n=1}^N v_{kn} y_{nk} (I_{kk'} - y_{nk'}) v_{k'n}. \end{aligned} \quad (4.155)$$

The right hand side can be written as

$$\sum_{n=1}^N \sum_{k=1}^K y_{nk} v_{kn} \left( v_{kn} - \sum_{k'=1}^K y_{nk'} v_{k'n} \right). \quad (4.156)$$

By the Jensen's inequality,

$$\begin{aligned} & \sum_{k=1}^K y_{nk} v_{kn} \left( v_{kn} - \sum_{k'=1}^K y_{nk'} v_{k'n} \right) \\ & \geq \left( \sum_{k=1}^K y_{nk} v_{kn} \right) \left( \sum_{k=1}^K y_{nk} v_{kn} - \sum_{k'=1}^K y_{nk'} v_{k'n} \right), \end{aligned} \quad (4.157)$$

where the right hand side is zero. Then,

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} \geq 0. \quad (4.158)$$

Therefore,  $\mathbf{H}$  is positive semidefinite.

## 4.21

Let

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta. \quad (4.159)$$

The right hand side can be written as

$$\int_{-\infty}^0 \mathcal{N}(\theta|0,1)d\theta + \int_0^a \mathcal{N}(\theta|0,1)d\theta = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^a \exp\left(-\frac{\theta^2}{2}\right) d\theta. \quad (4.160)$$

The second term of the right hand side can be written as

$$\frac{1}{\sqrt{2\pi}} \int_0^{\frac{a}{\sqrt{2}}} \exp(-t^2) \sqrt{2} dt = \frac{1}{2} \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right), \quad (4.161)$$

where

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp(-t^2) dt. \quad (4.162)$$

Therefore,

$$\Phi(a) = \frac{1}{2} \left( 1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) \right). \quad (4.163)$$

## 4.22

Let  $\mathcal{D}$  be a set of variables dependent on  $\boldsymbol{\theta}$  in  $M$  dimensions. By marginalisation,

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad (4.164)$$

Let  $\boldsymbol{\theta}_{\text{MAP}}$  be a stationary point of  $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ . Then, we have the Taylor series

$$\begin{aligned} & \ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})) \\ & \simeq \ln(p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}})) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^T \mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}}), \end{aligned} \quad (4.165)$$

where

$$\mathbf{A} = -\nabla \nabla \ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MAP}}}. \quad (4.166)$$

Then,

$$\begin{aligned} & \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ & \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}}) \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^T \mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})\right) d\boldsymbol{\theta}. \end{aligned} \quad (4.167)$$

The integral of the right hand side can be written as

$$(2\pi)^{\frac{M}{2}} (\det \mathbf{A}^{-1})^{\frac{1}{2}} = (2\pi)^{\frac{M}{2}} (\det \mathbf{A})^{-\frac{1}{2}}. \quad (4.168)$$

Therefore,

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) p(\boldsymbol{\theta}_{\text{MAP}}) (2\pi)^{\frac{M}{2}} (\det \mathbf{A})^{-\frac{1}{2}}, \quad (4.169)$$

so that

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \ln p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{A}). \quad (4.170)$$

## 4.23

Let  $\mathcal{D}$  be a set of  $N$  independent and identically distributed variables dependent on  $\boldsymbol{\theta}$  in  $M$  dimensions such that

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0). \quad (4.171)$$

By 4.22,

$$\begin{aligned} \ln p(\mathcal{D}) &\simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \ln p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) \\ &\quad - \frac{1}{2} \ln (\det (\mathbf{H} + \mathbf{V}_0^{-1})), \end{aligned} \quad (4.172)$$

where  $\boldsymbol{\theta}_{\text{MAP}}$  is a stationary point of  $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$  and

$$\mathbf{H} = -\nabla \nabla \ln p(\mathcal{D}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MAP}}}. \quad (4.173)$$

The right hand side of the approximation can be written as

$$\begin{aligned} &\ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{V}_0) \\ &- \frac{1}{2} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m})^\top \mathbf{V}_0^{-1} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}) + \frac{M}{2} \ln(2\pi) \\ &- \frac{1}{2} \ln (\det (\mathbf{H} + \mathbf{V}_0^{-1})) \\ &= \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \frac{1}{2} \ln (\det \mathbf{V}_0^{-1}) - \frac{1}{2} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m})^\top \mathbf{V}_0^{-1} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}) \\ &- \frac{1}{2} \ln (\det (\mathbf{H} + \mathbf{V}_0^{-1})). \end{aligned} \quad (4.174)$$

If  $\mathbf{V}_0^{-1}$  can be neglected, the right hand side can be approximated as

$$\ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{1}{2} \ln(\det \mathbf{H}). \quad (4.175)$$

Since the elements of  $\mathcal{D}$  are independent and identically distributed,

$$\mathbf{H} = N\bar{\mathbf{H}}, \quad (4.176)$$

where

$$\bar{\mathbf{H}} = \frac{1}{N} \sum_{n=1}^N \mathbf{H}_n, \quad (4.177)$$

and  $\mathbf{H}_1, \dots, \mathbf{H}_N$  are respectively the one for each element. Then,

$$\det \mathbf{H} = N^M \det \bar{\mathbf{H}}. \quad (4.178)$$

Therefore,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2} \ln N. \quad (4.179)$$

## 4.24

Let  $t_1, \dots, t_N$  be binary variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= y_n^{t_n} (1 - y_n)^{1-t_n}, \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0), \end{aligned} \quad (4.180)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \boldsymbol{\phi}_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.181)$$

By marginalisation,

$$p(t_{N+1}|\mathbf{t}) = \int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \quad (4.182)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \quad (4.183)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{w}$  can be written as  $-E(\mathbf{w})$  where

$$E(\mathbf{w}) = \sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0). \quad (4.184)$$

Let  $\mathbf{w}_{\text{MAP}}$  be a stationary point of  $E$ . Then, we have the Taylor series

$$E(\mathbf{w}) = E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3), \quad (4.185)$$

where

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \quad (4.186)$$

Then,

$$p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}). \quad (4.187)$$

Then,

$$\begin{aligned} & \int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w} \\ & \simeq \int \sigma(a_{N+1})^{t_{N+1}} (1 - \sigma(a_{N+1}))^{1-t_{N+1}} p(a_{N+1})da_{N+1}, \end{aligned} \quad (4.188)$$

where

$$p(a_{N+1}) = \int \delta(a_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}_{N+1}) \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}) d\mathbf{w}, \quad (4.189)$$

and  $\delta$  is the Dirac delta function. Since  $t_{N+1}$  is a binary variable, the right hand side of the approximation can be written as

$$v_{N+1}^{t_{N+1}} (1 - v_{N+1})^{1-t_{N+1}}, \quad (4.190)$$

where

$$v_{N+1} = \int \sigma(a_{N+1})p(a_{N+1})da_{N+1}. \quad (4.191)$$

We have

$$\mathbb{E} a_{N+1} = \int a_{N+1}p(a_{N+1})da_{N+1}. \quad (4.192)$$

The right hand side can be written as

$$\int \mathbf{w}^\top \boldsymbol{\phi}_{N+1} \mathcal{N}(\mathbf{w} | \mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}) d\mathbf{w} = \mathbf{w}_{\text{MAP}}^\top \boldsymbol{\phi}_{N+1}. \quad (4.193)$$

Then,

$$\text{var } a_{N+1} = \int (a_{N+1} - \mathbf{w}_{\text{MAP}}^\top \boldsymbol{\phi}_{N+1})^2 p(a_{N+1}) da_{N+1}. \quad (4.194)$$

The right hand side can be written as

$$\int ((\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \boldsymbol{\phi}_{N+1})^2 p(a_{N+1}) da_{N+1} = \boldsymbol{\phi}_{N+1}^\top \mathbf{A}^{-1} \boldsymbol{\phi}_{N+1}. \quad (4.195)$$

Then,

$$p(a_{N+1}) = \mathcal{N}(a_{N+1} | \mathbf{w}_{\text{MAP}}^\top \boldsymbol{\phi}_{N+1}, \boldsymbol{\phi}_{N+1}^\top \mathbf{A}^{-1} \boldsymbol{\phi}_{N+1}). \quad (4.196)$$

Therefore,

$$p(t_{N+1} | \mathbf{t}) = v_{N+1}^{t_{N+1}} (1 - v_{N+1})^{1-t_{N+1}}, \quad (4.197)$$

where

$$v_{N+1} = \int \sigma(a) \mathcal{N}(a | \mathbf{w}_{\text{MAP}}^\top \boldsymbol{\phi}_{N+1}, \boldsymbol{\phi}_{N+1}^\top \mathbf{A}^{-1} \boldsymbol{\phi}_{N+1}) da. \quad (4.198)$$

## 4.25

Let  $\lambda$  be a constant such that

$$\frac{d\sigma(a)}{da} \Big|_{a=0} = \frac{d\Phi(\lambda a)}{da} \Big|_{a=0}, \quad (4.199)$$

where

$$\begin{aligned} \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ \Phi(a) &= \int_{-\infty}^a \mathcal{N}(\theta | 0, 1) d\theta. \end{aligned} \quad (4.200)$$

By 4.12,

$$\frac{d\sigma(a)}{da} = \sigma(a) (1 - \sigma(a)). \quad (4.201)$$

We have

$$\frac{d\Phi(\lambda a)}{da} = \lambda \mathcal{N}(a|0, 1). \quad (4.202)$$

Then,

$$\frac{1}{4} = \lambda(2\pi)^{-\frac{1}{2}}. \quad (4.203)$$

Therefore,

$$\lambda^2 = \frac{\pi}{8}. \quad (4.204)$$

## 4.26

Let

$$I(\mu) = \int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da, \quad (4.205)$$

where

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta. \quad (4.206)$$

By the transformation

$$z = \frac{\theta - \mu}{\sigma}, \quad (4.207)$$

the right hand side can be written as

$$\begin{aligned} & \int \Phi(\lambda(\mu + \sigma z)) \mathcal{N}(\mu + \sigma z|\mu, \sigma^2) \sigma dz \\ &= \int \Phi(\lambda(\mu + \sigma z)) \mathcal{N}(z|0, 1) dz. \end{aligned} \quad (4.208)$$

Then,

$$\frac{\partial}{\partial \mu} I(\mu) = \lambda \int \mathcal{N}(\lambda(\mu + \sigma z)|0, 1) \mathcal{N}(z|0, 1) dz. \quad (4.209)$$

The logarithm of the integrand of the right hand side can be written as

$$\begin{aligned} & -\frac{1}{2} \ln(2\pi) - \frac{\lambda^2(\mu + \sigma z)^2}{2} - \frac{1}{2} \ln(2\pi) - \frac{z^2}{2} \\ &= -\ln(2\pi) - \frac{1 + \sigma^2 \lambda^2}{2} \left( z + \frac{\mu \sigma \lambda^2}{1 + \sigma^2 \lambda^2} \right)^2 + \frac{\mu^2 \sigma^2 \lambda^4}{2(1 + \sigma^2 \lambda^2)} - \frac{\mu^2 \lambda^2}{2}. \end{aligned} \quad (4.210)$$

The right hand side can be written as

$$\begin{aligned}
& -\ln(2\pi) - \frac{1+\sigma^2\lambda^2}{2} \left( z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2} \right)^2 - \frac{\mu^2\lambda^2}{2(1+\sigma^2\lambda^2)} \\
& = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(1+\sigma^2\lambda^2)^{-1} - \frac{1+\sigma^2\lambda^2}{2} \left( z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2} \right)^2 \\
& \quad - \ln\lambda - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\lambda^{-2}+\sigma^2) - \frac{\mu^2}{2(\lambda^{-2}+\sigma^2)}. \tag{4.211}
\end{aligned}$$

Then, the integral can be written as

$$\begin{aligned}
& \frac{1}{\lambda} \mathcal{N}(\mu|0, \lambda^{-2}+\sigma^2) \int \mathcal{N}\left(z|-\frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}, (1+\sigma^2\lambda^2)^{-1}\right) dz \\
& = \frac{1}{\lambda} \mathcal{N}(\mu|0, \lambda^{-2}+\sigma^2). \tag{4.212}
\end{aligned}$$

Then,

$$\frac{\partial}{\partial\mu} I(\mu) = \mathcal{N}(\mu|0, \lambda^{-2}+\sigma^2), \tag{4.213}$$

so that

$$I(\mu) = \int_{-\infty}^{\mu} \mathcal{N}(m|0, \lambda^{-2}+\sigma^2) dm. \tag{4.214}$$

By the transformation

$$m' = (\lambda^{-2}+\sigma^2)^{-\frac{1}{2}} m, \tag{4.215}$$

the right hand side can be written as

$$\begin{aligned}
& \int_{-\infty}^{(\lambda^{-2}+\sigma^2)^{-\frac{1}{2}}\mu} (\lambda^{-2}+\sigma^2)^{-\frac{1}{2}} \mathcal{N}(m'|0, 1) (\lambda^{-2}+\sigma^2)^{\frac{1}{2}} dm' \\
& = \int_{-\infty}^{(\lambda^{-2}+\sigma^2)^{-\frac{1}{2}}\mu} \mathcal{N}(m'|0, 1) dm'. \tag{4.216}
\end{aligned}$$

Therefore,

$$I(\mu) = \Phi\left((\lambda^{-2}+\sigma^2)^{-\frac{1}{2}}\mu\right). \tag{4.217}$$

## 5 Neural Networks

### 5.1

Let

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left( \sum_{m=1}^M w_{km}^{(2)} \sigma \left( \sum_{d=1}^D w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) + w_{k0}^{(2)} \right), \quad (5.1)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (5.2)$$

We have

$$\sigma(a) = \frac{\exp\left(\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)}. \quad (5.3)$$

The right hand side can be written as

$$\frac{\exp\left(\frac{a}{2}\right) - \exp\left(-\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)} + \frac{\exp\left(-\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)} = \tanh\left(\frac{a}{2}\right) + 1 - \sigma(a). \quad (5.4)$$

Then,

$$\sigma(a) = \frac{1}{2} \left( 1 + \tanh\left(\frac{a}{2}\right) \right). \quad (5.5)$$

Then, the argument of the expression of  $y_k(\mathbf{x}, \mathbf{w})$  can be written as

$$\begin{aligned} & \sum_{m=1}^M w_{km}^{(2)} \left( \frac{1}{2} \left( 1 + \tanh\left(\frac{1}{2} \left( \sum_{d=1}^D w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) \right) \right) \right) + w_{k0}^{(2)} \\ &= \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} \tanh\left(\frac{1}{2} \sum_{d=1}^D w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)}\right) + \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} + w_{k0}^{(2)}. \end{aligned} \quad (5.6)$$

Therefore,

$$\begin{aligned} & y_k(\mathbf{x}, \mathbf{w}) \\ &= \sigma \left( \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} \tanh\left(\frac{1}{2} \sum_{d=1}^D w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)}\right) + \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} + w_{k0}^{(2)} \right). \end{aligned} \quad (5.7)$$

## 5.2

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables such that

$$p(\mathbf{t}_n | \mathbf{w}) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}_n, \beta^{-1} \mathbf{I}), \quad (5.8)$$

where

$$\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n, \mathbf{w}). \quad (5.9)$$

Then, the logarithm of  $p(\mathbf{T} | \mathbf{w})$  except the terms independent of  $\mathbf{w}$  can be written as

$$-\frac{\beta}{2} \sum_{n=1}^N \|\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\|^2. \quad (5.10)$$

Therefore, maximising  $p(\mathbf{T} | \mathbf{w})$  is equivalent to minimising  $E(\mathbf{w})$ , where

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \quad (5.11)$$

## 5.3

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables such that

$$p(\mathbf{t}_n | \mathbf{w}) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}_n, \Sigma), \quad (5.12)$$

where

$$\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n, \mathbf{w}). \quad (5.13)$$

(a)

The logarithm of  $p(\mathbf{T} | \mathbf{w})$  except the terms independent of  $\mathbf{w}$  and  $\Sigma$  can be written as

$$-\frac{1}{2} \ln(\det \Sigma) - \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}_n)^\top \Sigma^{-1} (\mathbf{t}_n - \mathbf{y}_n). \quad (5.14)$$

Setting the derivatives with respect to  $\mathbf{w}$  and  $\Sigma$  to zero gives

$$\begin{aligned} \mathbf{0} &= - \sum_{n=1}^N \frac{\partial \mathbf{y}_n}{\partial \mathbf{w}} \Sigma^{-1} (\mathbf{t}_n - \mathbf{y}_n), \\ \mathbf{O} &= -\frac{1}{2} \Sigma^{-1} + \frac{1}{2} (\Sigma^{-1})^2 \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}_n) (\mathbf{t}_n - \mathbf{y}_n)^\top. \end{aligned} \quad (5.15)$$

Therefore, the maximum likelihood solution for  $\Sigma$  is given by

$$\Sigma = \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^\top. \quad (5.16)$$

(b)

If  $\Sigma$  is fixed and known, then the maximum likelihood solution for  $\mathbf{w}$  is given by minimising  $E(\mathbf{w})$  where

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}_n)^\top \Sigma^{-1} (\mathbf{t}_n - \mathbf{y}_n). \quad (5.17)$$

## 5.4

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n = 1 | \mathbf{w}) = (1 - \epsilon)y_n + \epsilon(1 - y_n), \quad (5.18)$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \quad (5.19)$$

Then,

$$p(t_n | \mathbf{w}) = ((1 - \epsilon)y_n + \epsilon(1 - y_n))^{t_n} (\epsilon y_n + (1 - \epsilon)(1 - y_n))^{1-t_n}. \quad (5.20)$$

Therefore,

$$\begin{aligned} -\ln p(\mathbf{t} | \mathbf{w}) &= - \sum_{n=1}^N t_n \ln ((1 - \epsilon)y_n + \epsilon(1 - y_n)) \\ &\quad - \sum_{n=1}^N (1 - t_n) \ln (\epsilon y_n + (1 - \epsilon)(1 - y_n)). \end{aligned} \quad (5.21)$$

## 5.5

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n | \mathbf{w}) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (5.22)$$

where

$$y_{nk} = y_k(\mathbf{x}_n, \mathbf{w}). \quad (5.23)$$

Therefore,

$$\ln p(\mathbf{T}|\mathbf{w}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (5.24)$$

## 5.6

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n} (1 - y_n)^{1-t_n}, \quad (5.25)$$

where

$$\begin{aligned} y_n &= \sigma(a_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ a_n &= a(\mathbf{x}_n, \mathbf{w}). \end{aligned} \quad (5.26)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \quad (5.27)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)). \quad (5.28)$$

By 4.12,

$$\frac{\partial E(\mathbf{w})}{\partial a_n} = -y_n(1 - y_n) \left( \frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right). \quad (5.29)$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_n} = y_n - t_n. \quad (5.30)$$

## 5.7

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n|\mathbf{w}) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (5.31)$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{a}_n), \\ y_k(\mathbf{a}) &= \frac{\exp(a_k)}{\sum_{k'=1}^K \exp(a_{k'})}, \\ \mathbf{a}_n &= \mathbf{a}(\mathbf{x}_n, \mathbf{w}). \end{aligned} \tag{5.32}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{T}|\mathbf{w}). \tag{5.33}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \tag{5.34}$$

Then,

$$\frac{\partial E(\mathbf{w})}{\partial a_{nk'}} = -\sum_{k=1}^K \frac{t_{nk}}{y_{nk}} \frac{\partial y_{nk}}{\partial a_{nk'}}. \tag{5.35}$$

By 4.17,

$$\frac{\partial y_{nk}}{\partial a_{nk'}} = y_{nk}(I_{kk'} - y_{nk'}). \tag{5.36}$$

Then,

$$\frac{\partial E(\mathbf{w})}{\partial a_{nk'}} = \sum_{k=1}^K t_{nk}(y_{nk'} - I_{kk'}). \tag{5.37}$$

The right hand side can be written as

$$y_{nk'} \sum_{k=1}^K t_{nk} - t_{nk'}. \tag{5.38}$$

Since  $\mathbf{t}_1, \dots, \mathbf{t}_N$  are from the standard basis in  $K$  dimensions,

$$\sum_{k=1}^K t_{nk} = 1. \tag{5.39}$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_{nk}} = y_{nk} - t_{nk}. \tag{5.40}$$

## 5.8

We have

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \quad (5.41)$$

Then,

$$\frac{d}{da} \tanh a = 1 - \left( \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \right)^2. \quad (5.42)$$

Therefore,

$$\frac{d}{da} \tanh a = 1 - (\tanh a)^2. \quad (5.43)$$

## 5.9

Let  $t_1, \dots, t_N$  be variables from  $\{-1, 1\}$  such that

$$p(t_n | \mathbf{w}) = \left( \frac{1 + y_n}{2} \right)^{\frac{1+t_n}{2}} \left( \frac{1 - y_n}{2} \right)^{\frac{1-t_n}{2}}, \quad (5.44)$$

where

$$\begin{aligned} y_n &= y(\mathbf{x}_n, \mathbf{w}), \\ -1 &\leq y(\mathbf{x}_n, \mathbf{w}) \leq 1. \end{aligned} \quad (5.45)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}). \quad (5.46)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N \left( \frac{1 + t_n}{2} \ln \frac{1 + y_n}{2} + \frac{1 - t_n}{2} \ln \frac{1 - y_n}{2} \right). \quad (5.47)$$

The appropriate choice of  $y$  is  $\tanh$ .

## 5.10

Let

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}), \quad (5.48)$$

where  $E$  is a real function of real vectors. Then,  $\mathbf{H}$  is a real symmetric matrix. Therefore, by 2.20,  $\mathbf{H}$  is positive definite if and only if all of its eigenvalues are positive.

## 5.11

Let  $\mathbf{w}$  be a real vector in  $M$  dimensions. We have the Taylor series

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*) + O(\|\mathbf{w} - \mathbf{w}^*\|^3), \quad (5.49)$$

where

$$\begin{aligned} \nabla E(\mathbf{w}^*) &= \mathbf{0}, \\ \mathbf{H} &= \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*}. \end{aligned} \quad (5.50)$$

Let  $\mathbf{u}_1, \dots, \mathbf{u}_M$  be eigenvectors of  $\mathbf{H}$  such that

$$\mathbf{H}\mathbf{u}_m = \lambda_m \mathbf{u}_m. \quad (5.51)$$

Then, there exists  $\alpha_1, \dots, \alpha_M$  such that

$$\mathbf{w} - \mathbf{w}^* = \sum_{m=1}^M \alpha_m \mathbf{u}_m. \quad (5.52)$$

By 2.19, since  $\mathbf{H}$  is a real symmetric matrix,

$$\mathbf{u}_m^\top \mathbf{u}_{m'} = I_{mm'}. \quad (5.53)$$

Then,

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} \sum_{m=1}^M \lambda_m \alpha_m^2. \quad (5.54)$$

Therefore, the contours of constant  $E$  are ellipses whose axes are aligned with  $\mathbf{u}_1, \dots, \mathbf{u}_M$  with lengths which are proportional to  $\lambda_1^{-\frac{1}{2}}, \dots, \lambda_M^{-\frac{1}{2}}$ .

## 5.12

Let  $\mathbf{w}$  be a real vector. We have the Taylor series

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*) + O(\|\mathbf{w} - \mathbf{w}^*\|^3), \quad (5.55)$$

where

$$\begin{aligned} \nabla E(\mathbf{w}^*) &= \mathbf{0}, \\ \mathbf{H} &= \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*}. \end{aligned} \quad (5.56)$$

If  $\mathbf{H}$  is positive definite, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \quad (5.57)$$

Then,  $\mathbf{w}^*$  is a local minimum of  $E$ . On the other hand, if  $\mathbf{w}^*$  is a local minimum of  $E$ , then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \quad (5.58)$$

Then,  $\mathbf{H}$  is positive definite. Therefore, the necessary and sufficient condition for  $\mathbf{w}^*$  to be a local minimum is that  $\mathbf{H}$  is positive definite.

### 5.13

Let  $\mathbf{w}$  be a vector in  $M$  dimensions. We have the Taylor series

$$\begin{aligned} E(\mathbf{w}) &= E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{b} + \frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}}) \\ &\quad + O(\|\mathbf{w} - \hat{\mathbf{w}}\|^3), \end{aligned} \quad (5.59)$$

where

$$\begin{aligned} \mathbf{b} &= \nabla E(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}. \\ \mathbf{H} &= \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}. \end{aligned} \quad (5.60)$$

Since  $\mathbf{b}$  is a vector in  $M$  dimensions and  $\mathbf{H}$  is a  $M \times M$  symmetric matrix, the number of independent elements of the right hand side is

$$M + \frac{M(M+1)}{2} = \frac{M(M+3)}{2}. \quad (5.61)$$

### 5.14

Let  $w$  be a variable. We have the Taylor series

$$\begin{aligned} E_n(w_{mm'} + \epsilon) &= E_n(w_{mm'}) + \left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} \epsilon + O(\epsilon^2), \\ E_n(w_{mm'} - \epsilon) &= E_n(w_{mm'}) - \left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} \epsilon + O(\epsilon^2). \end{aligned} \quad (5.62)$$

Therefore,

$$\left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} = \frac{E_n(w_{mm'} + \epsilon) - E_n(w_{mm'} - \epsilon)}{2\epsilon} + O(\epsilon^2). \quad (5.63)$$

## 5.15 (Incomplete)

Let

$$J_{km} = \frac{\partial y_k}{\partial x_m}. \quad (5.64)$$

The right hand side can be written as

$$\sum_{m'=1}^M W_{mm'} \frac{\partial y_k}{\partial a_{m'}}, \quad (5.65)$$

where

$$W_{mm'} = \frac{\partial a_{m'}}{\partial x_m}. \quad (5.66)$$

We have

$$\frac{\partial y_k}{\partial a_{m'}} = \sum_{m''=1}^M \frac{\partial y_k}{\partial a_{m''}} \frac{\partial a_{m''}}{\partial a_{m'}}. \quad (5.67)$$

## 5.16

Let  $\mathbf{y}_1, \dots, \mathbf{y}_N$  be vectors dependent on  $\mathbf{w}$ . Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \|\mathbf{y}_n - \mathbf{t}_n\|^2. \quad (5.68)$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (\nabla \mathbf{y}_n)^\top (\mathbf{y}_n - \mathbf{t}_n). \quad (5.69)$$

Therefore,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N (\nabla \text{vec}((\nabla \mathbf{y}_n)^\top)^\top (\mathbf{y}_n - \mathbf{t}_n) + \sum_{n=1}^N (\nabla \mathbf{y}_n)^\top (\nabla \mathbf{y}_n)). \quad (5.70)$$

## 5.17

Let  $y$  be a scalar dependent on  $\mathbf{x}$  and  $\mathbf{w}$ . Let

$$E(\mathbf{w}) = \frac{1}{2} \int \int (y - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt. \quad (5.71)$$

Then,

$$\nabla E(\mathbf{w}) = \int \int (y - t)p(\mathbf{x}, t) \nabla y d\mathbf{x} dt. \quad (5.72)$$

The right hand side can be written as

$$\begin{aligned} & \int y \nabla y \left( \int p(\mathbf{x}, t) dt \right) d\mathbf{x} - \int \nabla y \left( \int t p(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x} \\ &= \int y \nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla y \mathbb{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (5.73)$$

Then,

$$\begin{aligned} \nabla \nabla E(\mathbf{w}) &= \int \nabla y (\nabla y)^T p(\mathbf{x}) d\mathbf{x} + \int y \nabla \nabla y p(\mathbf{x}) d\mathbf{x} \\ &\quad - \int \nabla \nabla y \mathbb{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (5.74)$$

The second and the third terms of the right hand side can be written as

$$\mathbb{E}(y \nabla \nabla y) - \mathbb{E}(\nabla \nabla y \mathbb{E}(t|\mathbf{x})) = \mathbb{E}((y - \mathbb{E}(t|\mathbf{x})) \nabla \nabla y). \quad (5.75)$$

Therefore, if

$$y = \mathbb{E}(t|\mathbf{x}), \quad (5.76)$$

then

$$\nabla \nabla E(\mathbf{w}) = \int \nabla y (\nabla y)^T p(\mathbf{x}) d\mathbf{x}. \quad (5.77)$$

## 5.18

Let  $y_1, \dots, y_N$  be scalars such that

$$\begin{aligned} y_n &= \mathbf{w}_n^{(2)\top} \mathbf{z} + \mathbf{w}_n^{(0)\top} \mathbf{x}, \\ z_m &= \tanh(\mathbf{w}_m^{(1)\top} \mathbf{x}). \end{aligned} \quad (5.78)$$

Let

$$E = \frac{1}{2} \sum_{n=1}^N (y_n - t_n)^2. \quad (5.79)$$

Then,

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}}, \\ \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{w}_m^{(1)}}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.\end{aligned}\tag{5.80}$$

Therefore,

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= (y_n - t_n) \mathbf{x}, \\ \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= (y_n - t_n) \mathbf{A} \mathbf{w}_n^{(2)}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= (y_n - t_n) \mathbf{z},\end{aligned}\tag{5.81}$$

where

$$A_{mm'} = (1 - z_m^2) x_{m'}. \tag{5.82}$$

## 5.19

Let  $t_1, \dots, t_N$  be binary variables such that

$$p(t_n | \mathbf{w}) = y_n^{t_n} (1 - y_n)^{1-t_n}, \tag{5.83}$$

where

$$\begin{aligned}y_n &= \sigma(a_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ a_n &= a(\mathbf{x}_n, \mathbf{w}).\end{aligned}\tag{5.84}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}). \tag{5.85}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)). \tag{5.86}$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \left( \frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right) \nabla y_n. \tag{5.87}$$

We have

$$\nabla y_n = \frac{\partial y_n}{\partial a_n} \nabla a_n. \quad (5.88)$$

By 4.12,

$$\frac{\partial y_n}{\partial a_n} = y_n(1 - y_n). \quad (5.89)$$

Then,

$$\nabla y_n = y_n(1 - y_n) \nabla a_n. \quad (5.90)$$

Then,

$$\nabla E(\mathbf{w}) = - \sum_{n=1}^N (t_n(1 - y_n) - (1 - t_n)y_n) \nabla a_n, \quad (5.91)$$

so that

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \nabla a_n. \quad (5.92)$$

Then,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N \nabla a_n (\nabla y_n)^\top + \sum_{n=1}^N (y_n - t_n) \nabla \nabla a_n. \quad (5.93)$$

Therefore,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n(1 - y_n) \nabla a_n (\nabla a_n)^\top + \sum_{n=1}^N (y_n - t_n) \nabla \nabla a_n. \quad (5.94)$$

## 5.20

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$p(\mathbf{t}_n | \mathbf{w}) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (5.95)$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{a}_n), \\ y_k(\mathbf{a}) &= \frac{\exp(a_k)}{\sum_{k'=1}^K \exp(a_{k'})}, \\ \mathbf{a}_n &= \mathbf{a}(\mathbf{x}_n, \mathbf{w}) \end{aligned} \quad (5.96)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{T}|\mathbf{w}). \quad (5.97)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (5.98)$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K \frac{t_{nk}}{y_{nk}} \nabla y_{nk}. \quad (5.99)$$

We have

$$\nabla y_{nk} = \sum_{k'=1}^K \frac{\partial y_{nk}}{\partial a_{nk'}} \nabla a_{nk'}. \quad (5.100)$$

By 4.17,

$$\frac{\partial y_{nk}}{\partial a_{nk'}} = y_{nk}(I_{kk'} - y_{nk'}). \quad (5.101)$$

Then,

$$\nabla y_{nk} = y_{nk} \sum_{k'=1}^K (I_{kk'} - y_{nk'}) \nabla a_{nk'}. \quad (5.102)$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \sum_{k'=1}^K (I_{kk'} - y_{nk'}) \nabla a_{nk}, \quad (5.103)$$

so that

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N \sum_{k=1}^K (y_{nk} - t_{nk}) \nabla a_{nk}. \quad (5.104)$$

Then,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N \sum_{k=1}^K \nabla a_{nk} (\nabla y_{nk})^\top + \sum_{n=1}^N \sum_{k=1}^K (y_{nk} - t_{nk}) \nabla \nabla a_{nk}. \quad (5.105)$$

Therefore,

$$\begin{aligned} \nabla \nabla E(\mathbf{w}) &= \sum_{n=1}^N \sum_{k=1}^K y_{nk} \nabla a_{nk} \sum_{k'=1}^K (I_{kk'} - y_{nk'}) (\nabla a_{nk'})^\top \\ &\quad + \sum_{n=1}^N \sum_{k=1}^K (y_{nk} - t_{nk}) \nabla \nabla a_{nk}. \end{aligned} \quad (5.106)$$

## 5.21

Let  $\mathbf{y}_1, \dots, \mathbf{y}_N$  be vectors dependent on  $\mathbf{w}$ . Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \|\mathbf{y}_n - \mathbf{t}_n\|^2. \quad (5.107)$$

By 5.16,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N (\nabla \text{vec}(\nabla \mathbf{y}_n)^\top)^\top (\mathbf{y}_n - \mathbf{t}_n) + \mathbf{H}, \quad (5.108)$$

where

$$\mathbf{H} = \sum_{n=1}^N (\nabla \mathbf{y}_n)^\top (\nabla \mathbf{y}_n). \quad (5.109)$$

Let

$$\mathbf{H}' = \sum_{n=1}^{N+1} (\nabla \mathbf{y}_n)^\top (\nabla \mathbf{y}_n). \quad (5.110)$$

Then,

$$\mathbf{H}' = \mathbf{H} + \mathbf{B}^\top \mathbf{B}, \quad (5.111)$$

where

$$\mathbf{B} = \nabla \mathbf{y}_{N+1}. \quad (5.112)$$

By 2.24,

$$\begin{bmatrix} \mathbf{H} & \mathbf{B}^\top \\ \mathbf{B} & -\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{H}'^{-1} & \mathbf{H}'^{-1} \mathbf{B}^\top \\ \mathbf{B} \mathbf{H}'^{-1} & -\mathbf{I} - \mathbf{B} \mathbf{H}'^{-1} \mathbf{B}^\top \end{bmatrix}, \quad (5.113)$$

and

$$\begin{bmatrix} -\mathbf{I} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{H} \end{bmatrix}^{-1} = \begin{bmatrix} -\mathbf{M} & \mathbf{M}^\top \mathbf{B} \mathbf{H}^{-1} \\ \mathbf{H}^{-1} \mathbf{B}^\top \mathbf{M} & \mathbf{H}^{-1} - \mathbf{H}^{-1} \mathbf{B}^\top \mathbf{M}^\top \mathbf{B} \mathbf{H}^{-1} \end{bmatrix}, \quad (5.114)$$

where

$$\mathbf{M} = (\mathbf{I} + \mathbf{B} \mathbf{H}^{-1} \mathbf{B}^\top)^{-1}. \quad (5.115)$$

Therefore,

$$\mathbf{H}'^{-1} = \mathbf{H}^{-1} - \mathbf{H}^{-1} \mathbf{B}^\top \mathbf{M} \mathbf{B} \mathbf{H}^{-1}. \quad (5.116)$$

## 5.22

Let  $a_1, \dots, a_N$  be variables such that

$$\begin{aligned} a_n &= \mathbf{w}_n^{(2)\top} \mathbf{z}, \\ z_m &= h(b_m), \\ b_m &= \mathbf{w}_m^{(1)\top} \mathbf{x}. \end{aligned} \tag{5.117}$$

Let  $E$  be a function of  $a_1, \dots, a_N$ . Then,

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \sum_{n=1}^N \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial z_m} \frac{\partial z_m}{\partial b_m} \frac{\partial b_m}{\partial \mathbf{w}_m^{(1)}}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial \mathbf{w}_n^{(2)}}. \end{aligned} \tag{5.118}$$

Then,

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= h'(b_m) \mathbf{x} \sum_{n=1}^N w_{nm}^{(2)} \frac{\partial E}{\partial a_n}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial a_n} \mathbf{z}. \end{aligned} \tag{5.119}$$

Therefore,

$$\begin{aligned} \frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= C_{mm'}^{(11)} \mathbf{x} \mathbf{x}^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= C_{mn}^{(12)} \mathbf{x} \mathbf{z}^\top + h'(b_m) \frac{\partial E}{\partial a_n} \mathbf{x} \mathbf{v}^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= C_{nn'}^{(22)} \mathbf{z} \mathbf{z}^\top, \end{aligned} \tag{5.120}$$

where  $\mathbf{v}$  is a vector from the standard basis in  $M$  dimensions such that

$$v_m = 1, \tag{5.121}$$

and

$$\begin{aligned}
C_{mm'}^{(11)} &= h''(b_m) I_{mm'} \sum_{n=1}^N w_{nm}^{(2)} \frac{\partial E}{\partial a_n} \\
&\quad + h'(b_m) h'(b_{m'}) \sum_{n=1}^N w_{nm}^{(2)} \sum_{n'=1}^N w_{n'm'}^{(2)} \frac{\partial^2 E}{\partial y_n \partial a_{n'}}, \\
C_{mn}^{(12)} &= h'(b_m) \sum_{n'=1}^N w_{n'm}^{(2)} \frac{\partial^2 E}{\partial a_n \partial a_{n'}}, \\
C_{nn'}^{(22)} &= \frac{\partial^2 E}{\partial a_n \partial a_{n'}}.
\end{aligned} \tag{5.122}$$

## 5.23

Let  $a_1, \dots, a_N$  be variables such that

$$\begin{aligned}
a_n &= \mathbf{w}_n^{(2)\top} \mathbf{z} + \mathbf{w}_n^{(0)\top} \mathbf{x}, \\
z_m &= h(b_m), \\
b_m &= \mathbf{w}_m^{(1)\top} \mathbf{x}.
\end{aligned} \tag{5.123}$$

Let  $E$  be a function of  $a_1, \dots, a_N$ . Then,

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial \mathbf{w}_n^{(0)}}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \sum_{n=1}^N \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial z_m} \frac{\partial z_m}{\partial b_m} \frac{\partial b_m}{\partial \mathbf{w}_m^{(1)}}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial \mathbf{w}_n^{(2)}}.
\end{aligned} \tag{5.124}$$

Then,

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial a_n} \mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= h'(b_m) \mathbf{x} \sum_{n=1}^N w_{nm}^{(2)} \frac{\partial E}{\partial a_n}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial a_n} \mathbf{z}.
\end{aligned} \tag{5.125}$$

Therefore,

$$\begin{aligned}
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(0)}} &= C_{nn'}^{(22)} \mathbf{x} \mathbf{x}^\top, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_m^{(1)}} &= C_{nm}^{(12)} \mathbf{x} \mathbf{x}^\top, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(2)}} &= C_{nn'}^{(22)} \mathbf{x} \mathbf{z}^\top, \\
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= C_{mm'}^{(11)} \mathbf{x} \mathbf{x}^\top, \\
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= C_{mn}^{(12)} \mathbf{x} \mathbf{z}^\top + h'(b_m) \frac{\partial E}{\partial a_n} \mathbf{x} \mathbf{v}^\top, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= C_{nn'}^{(22)} \mathbf{z} \mathbf{z}^\top,
\end{aligned} \tag{5.126}$$

where  $\mathbf{v}$  is a vector from the standard basis in  $M$  dimensions such that

$$v_m = 1, \tag{5.127}$$

and

$$\begin{aligned}
C_{mm'}^{(11)} &= h''(b_m) I_{mm'} \sum_{n=1}^N w_{nm}^{(2)} \frac{\partial E}{\partial a_n} \\
&\quad + h'(b_m) h'(b_{m'}) \sum_{n=1}^N w_{nm}^{(2)} \sum_{n'=1}^N w_{n'm'}^{(2)} \frac{\partial^2 E}{\partial a_n \partial a_{n'}}, \\
C_{nm}^{(12)} &= h'(b_m) \sum_{n'=1}^N w_{n'm}^{(2)} \frac{\partial^2 E}{\partial a_n \partial a_{n'}}, \\
C_{nn'}^{(22)} &= \frac{\partial^2 E}{\partial a_n \partial a_{n'}}.
\end{aligned} \tag{5.128}$$

## 5.24

Let  $y_1, \dots, y_N$  be variables such that

$$\begin{aligned}
y_n &= \mathbf{w}_n^\top \mathbf{z} + w_{n0}, \\
z_m &= h(\mathbf{w}_m^\top \mathbf{x} + w_{m0}).
\end{aligned} \tag{5.129}$$

(a)

Let

$$\tilde{\mathbf{x}} = a\mathbf{x} + b\mathbf{v}, \quad (5.130)$$

where

$$\mathbf{v} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Then,

$$z_m = h \left( \frac{1}{a} \mathbf{w}_m^\top (\tilde{\mathbf{x}} - b\mathbf{v}) + w_{m0} \right). \quad (5.131)$$

Therefore,

$$\tilde{z}_m = h (\tilde{\mathbf{w}}_m^\top \mathbf{x} + \tilde{w}_{m0}), \quad (5.132)$$

where

$$\begin{aligned} \tilde{\mathbf{w}}_m &= \frac{1}{a} \mathbf{w}_m, \\ \tilde{w}_{m0} &= w_{m0} - \frac{b}{a} \mathbf{w}_m^\top \mathbf{v}. \end{aligned} \quad (5.133)$$

(b)

Let

$$\tilde{y}_n = c y_n + d. \quad (5.134)$$

Then,

$$\frac{\tilde{y}_n - d}{c} = \mathbf{w}_n^\top \mathbf{z} + w_{n0}. \quad (5.135)$$

Therefore,

$$\tilde{y}_n = \tilde{\mathbf{w}}_n^\top \mathbf{z} + \tilde{w}_{n0}, \quad (5.136)$$

where

$$\begin{aligned} \tilde{\mathbf{w}}_n &= c \mathbf{w}_n, \\ \tilde{w}_{n0} &= c w_{n0} + d. \end{aligned} \quad (5.137)$$

## 5.25

Let  $E$  be a quadratic error function such that

$$E(\mathbf{w}) = E_0 + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*), \quad (5.138)$$

where  $\mathbf{w}^*$  represents the minimum and  $\mathbf{H}$  is a positive definite and constant matrix whose eigenvectors and eigenvalues are  $\mathbf{u}_1, \dots, \mathbf{u}_M$  and  $\eta_1, \dots, \eta_M$ . Let

$$\begin{aligned} \mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \rho \nabla E, \\ \mathbf{w}^{(0)} &= \mathbf{0}. \end{aligned} \quad (5.139)$$

(a)

We have

$$\nabla E = \mathbf{H} (\mathbf{w} - \mathbf{w}^*). \quad (5.140)$$

Then,

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho \mathbf{H} (\mathbf{w}^{(\tau-1)} - \mathbf{w}^*). \quad (5.141)$$

so that

$$\mathbf{u}_m^\top \mathbf{w}^{(\tau)} = \mathbf{u}_m^\top \mathbf{w}^{(\tau-1)} - \rho \mathbf{u}_m^\top \mathbf{H} (\mathbf{w}^{(\tau-1)} - \mathbf{w}^*). \quad (5.142)$$

Since  $\mathbf{H}$  is symmetric,

$$\mathbf{u}_m^\top \mathbf{H} \mathbf{w} = \mathbf{w}^\top \mathbf{H} \mathbf{u}_m. \quad (5.143)$$

The right hand side can be written as  $\eta_m w_m$ , where

$$w_m = \mathbf{w}^\top \mathbf{u}_m. \quad (5.144)$$

Then,

$$w_m^{(\tau)} = w_m^{(\tau-1)} - \rho \eta_m (w_m^{(\tau-1)} - w_m^*), \quad (5.145)$$

so that

$$w_m^{(\tau)} - w_m^* = (1 - \rho \eta_m) (w_m^{(\tau-1)} - w_m^*). \quad (5.146)$$

Therefore,

$$w_m^{(\tau)} = w_m^* + (1 - \rho \eta_m)^\tau (w_m^{(0)} - w_m^*), \quad (5.147)$$

so that

$$w_m^{(\tau)} = (1 - (1 - \rho \eta_m)^\tau) w_m^*. \quad (5.148)$$

(b)

By (a), if

$$|1 - \rho\eta_m| < 1, \quad (5.149)$$

then

$$\lim_{\tau \rightarrow \infty} w_m^{(\tau)} = w_m^*. \quad (5.150)$$

(c)

By (a) and the Taylor series

$$(1 - \rho\eta_m)^\tau = 1 - \rho\tau\eta_m + \rho^2\tau^2 O(\eta_m^2), \quad (5.151)$$

we have

$$w_m^{(\tau)} = (\rho\tau\eta_m - \rho^2\tau^2 O(\eta_m^2)) w_m^*, \quad (5.152)$$

so that

$$w_m^{(\tau)} = \rho\tau\eta_m (1 - \rho\tau\eta_m O(\eta_m)) w_m^*. \quad (5.153)$$

Therefore,

$$|w_m^{(\tau)}| \ll |w_m|^*, \quad \text{if } \eta_m \ll (\rho\tau)^{-1}, \quad (5.154)$$

while

$$w_m^{(\tau)} \simeq w_m^*, \quad \text{if } \eta_m \gg (\rho\tau)^{-1}. \quad (5.155)$$

## 5.26 (Incomplete)

Let

$$\tilde{E} = E + \lambda\Omega, \quad (5.156)$$

where

$$\Omega = \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \left( \frac{\partial y_{nk}}{\partial \xi} \right)^2. \quad (5.157)$$

(a)

We have

$$\frac{\partial y_{nk}}{\partial \xi} = \mathcal{G}y_{nk}, \quad (5.158)$$

where

$$\begin{aligned}\mathcal{G} &= \sum_{d=1}^D \tau_d \frac{\partial}{\partial x_d}, \\ \tau_d &= \frac{\partial x_d}{\partial \xi}.\end{aligned}\tag{5.159}$$

Therefore,

$$\Omega = \sum_{n=1}^N \Omega_n,\tag{5.160}$$

where

$$\Omega_n = \frac{1}{2} \sum_{k=1}^K (\mathcal{G} y_{nk})^2.\tag{5.161}$$

(b)

$$\begin{aligned}\alpha_j &= \mathcal{G} z_j = h'(a_j) \beta_j? \\ \beta_j &= \mathcal{G} a_j = \sum_{d=1}^D w_{jd} \alpha_d?\end{aligned}\tag{5.162}$$

## 5.27

Let

$$\tilde{E} = \frac{1}{2} \int \int \int (y(\mathbf{x} + \boldsymbol{\xi}) - t)^2 p(t|\mathbf{x}) p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} dt d\boldsymbol{\xi},\tag{5.163}$$

where

$$\begin{aligned}\mathbb{E} \boldsymbol{\xi} &= 0, \\ \text{var } \boldsymbol{\xi} &= \lambda \mathbf{I}.\end{aligned}\tag{5.164}$$

By the Taylor series

$$y(\mathbf{x} + \boldsymbol{\xi}) = y(\mathbf{x}) + \rho(\mathbf{x}, \boldsymbol{\xi}) + O(\|\boldsymbol{\xi}\|^3),\tag{5.165}$$

where

$$\rho(\mathbf{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \nabla y(\mathbf{x}) + \frac{1}{2} \boldsymbol{\xi}^\top \nabla \nabla y(\mathbf{x}) \boldsymbol{\xi},\tag{5.166}$$

the integrand can be written as

$$\begin{aligned}(y(\mathbf{x}) - t)^2 &+ 2(y(\mathbf{x}) - t)(\rho(\mathbf{x}, \boldsymbol{\xi}) + O(\|\boldsymbol{\xi}\|^3)) \\ &+ (\rho(\mathbf{x}, \boldsymbol{\xi}) + O(\|\boldsymbol{\xi}\|^3))^2.\end{aligned}\tag{5.167}$$

Omitting the terms of  $O(\|\boldsymbol{\xi}\|^3)$ , the integral can be approximated as

$$2E + 2 \iint (y(\mathbf{x}) - t) \rho(\mathbf{x}, \boldsymbol{\xi}) p(t|\mathbf{x}) p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} dt d\boldsymbol{\xi} \\ + \iint \rho(\mathbf{x}, \boldsymbol{\xi})^2 p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} d\boldsymbol{\xi}, \quad (5.168)$$

where

$$E = \frac{1}{2} \iint (y(\mathbf{x}) - t)^2 p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt. \quad (5.169)$$

Omitting the terms of  $O(\|\boldsymbol{\xi}\|^3)$ , the second and third terms can be approximated as

$$2 \left( \int \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi} \right)^\top \iint (y(\mathbf{x}) - t) \nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\ + \int \boldsymbol{\xi}^\top \left( \iint (y(\mathbf{x}) - t) \nabla \nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \right) \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi} \\ + \int \boldsymbol{\xi}^\top \left( \int (\nabla y(\mathbf{x})) (\nabla y(\mathbf{x}))^\top p(\mathbf{x}) d\mathbf{x} \right) \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi}. \quad (5.170)$$

Since

$$\mathbb{E} \boldsymbol{\xi} = \mathbf{0}, \quad (5.171)$$

the first term can be written as  $\mathbf{0}$ . Since

$$\mathbb{E} \boldsymbol{\xi} \boldsymbol{\xi}^\top = \lambda \mathbf{I}, \quad (5.172)$$

The second term can be written as

$$\lambda \operatorname{tr} \left( \iint (y(\mathbf{x}) - t) \nabla \nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \right), \quad (5.173)$$

where the integral can be written as

$$\int \left( y(\mathbf{x}) \int p(t|\mathbf{x}) dt - \int t p(t|\mathbf{x}) dt \right) \nabla \nabla y(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ = \int (y(\mathbf{x}) - \mathbb{E}(t|\mathbf{x})) \nabla \nabla y(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (5.174)$$

Similarly, the third term can be written as  $2\lambda\Omega$ , where

$$\Omega = \frac{1}{2} \operatorname{tr} \left( \int \nabla y(\mathbf{x}) (\nabla y(\mathbf{x}))^\top p(\mathbf{x}) d\mathbf{x} \right). \quad (5.175)$$

Therefore, if

$$y(\mathbf{x}) = \text{E}(t|\mathbf{x}), \quad (5.176)$$

then

$$\tilde{E} \simeq E + \lambda\Omega. \quad (5.177)$$

## 5.28 (Incomplete)

### 5.29

Let  $w_1, \dots, w_M$  be variables such that

$$p(w_m) = \sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2). \quad (5.178)$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda\Omega(\mathbf{w}), \quad (5.179)$$

where

$$\Omega(\mathbf{w}) = - \sum_{m=1}^M \ln p(w_m). \quad (5.180)$$

We have

$$\frac{\partial \Omega}{\partial w_m} = - \frac{\sum_{k=1}^K \pi_k \frac{w_m - \mu_k}{\sigma_k^2} \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(w_m | \mu_{k'}, \sigma_{k'}^2)}. \quad (5.181)$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial w_m} = \frac{\partial E}{\partial w_m} - \lambda \sum_{k=1}^K \gamma_k(w_m) \frac{w_m - \mu_k}{\sigma_k^2}, \quad (5.182)$$

where

$$\gamma_k(w_m) = \frac{\pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(w_m | \mu_{k'}, \sigma_{k'}^2)}. \quad (5.183)$$

## 5.30

Let  $w_1, \dots, w_M$  be variables such that

$$p(w_m) = \sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2). \quad (5.184)$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda\Omega(\mathbf{w}), \quad (5.185)$$

where

$$\Omega(\mathbf{w}) = -\sum_{m=1}^M \ln p(w_m). \quad (5.186)$$

We have

$$\frac{\partial \Omega}{\partial \mu_k} = -\frac{\sum_{k=1}^K \pi_k \frac{w_m - \mu_k}{\sigma_k^2} \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(w_m | \mu_{k'}, \sigma_{k'}^2)}. \quad (5.187)$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial \mu_k} = -\lambda \sum_{k=1}^K \gamma_k(w_m) \frac{w_m - \mu_k}{\sigma_k^2}, \quad (5.188)$$

where

$$\gamma_k(w_m) = \frac{\pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(w_m | \mu_{k'}, \sigma_{k'}^2)}. \quad (5.189)$$

### 5.31

Let  $w_1, \dots, w_M$  be variables such that

$$p(w_m) = \sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2). \quad (5.190)$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda\Omega(\mathbf{w}), \quad (5.191)$$

where

$$\Omega(\mathbf{w}) = -\sum_{m=1}^M \ln p(w_m). \quad (5.192)$$

We have

$$\frac{\partial \Omega}{\partial \sigma_k} = -\frac{\sum_{k=1}^K \pi_k \left( -\frac{1}{\sigma_k} + \frac{(w_m - \mu_k)^2}{\sigma_k^3} \right) \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}. \quad (5.193)$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial \sigma_k} = -\lambda \sum_{k=1}^K \gamma_k(w_m) \left( -\frac{1}{\sigma_k} + \frac{(w_m - \mu_k)^2}{\sigma_k^3} \right), \quad (5.194)$$

where

$$\gamma_k(w_m) = \frac{\pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}. \quad (5.195)$$

### 5.32

Let  $w_1, \dots, w_M$  be variables such that

$$p(w_m) = \sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2), \quad (5.196)$$

where

$$\pi_k = \frac{\exp(\eta_k)}{\sum_{k=1}^K \exp(\eta_k)}. \quad (5.197)$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \quad (5.198)$$

where

$$\Omega(\mathbf{w}) = - \sum_{m=1}^M \ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2) \right). \quad (5.199)$$

(a)

If  $k \neq k'$ , then

$$\frac{\partial \pi_k}{\partial \eta_{k'}} = - \frac{\exp(\eta_{k'}) \exp(\eta_k)}{(\sum_{k=1}^K \exp(\eta_k))^2}. \quad (5.200)$$

We have

$$\frac{\partial \pi_k}{\partial \eta_k} = \frac{\exp(\eta_k)}{\sum_{k=1}^K \exp(\eta_k)} - \left( \frac{\exp(\eta_k)}{\sum_{k=1}^K \exp(\eta_k)} \right)^2. \quad (5.201)$$

Therefore,

$$\frac{\partial \pi_k}{\partial \eta_{k'}} = I_{kk'} \pi_k - \pi_k \pi_{k'}. \quad (5.202)$$

(b)

We have

$$\frac{\partial \Omega}{\partial \eta_k} = \sum_{k'=1}^K \frac{\partial \Omega}{\partial \pi_{k'}} \frac{\partial \pi_{k'}}{\partial \eta_k}. \quad (5.203)$$

We have

$$\frac{\partial \Omega}{\partial \pi_{k'}} = -\frac{\sum_{m=1}^M \mathcal{N}(w_m | \mu_{k'}, \sigma_{k'}^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}. \quad (5.204)$$

By (a),

$$\frac{\partial \Omega}{\partial \eta_k} = -\sum_{k'=1}^K \frac{\sum_{m=1}^M \mathcal{N}(w_m | \mu_{k'}, \sigma_{k'}^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)} (I_{kk'} \pi_k - \pi_k \pi_{k'}). \quad (5.205)$$

The right hand side can be written as

$$\begin{aligned} & -\sum_{m=1}^M \frac{\pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)} + \pi_k \sum_{m=1}^M \frac{\sum_{k'=1}^K \pi_{k'} \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)} \\ & = \sum_{m=1}^M \left( \pi_k - \frac{\pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)} \right). \end{aligned} \quad (5.206)$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial \eta_k} = \lambda \sum_{m=1}^M (\pi_k - \gamma_k(w_m)), \quad (5.207)$$

where

$$\gamma_k(w_m) = \frac{\pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}{\sum_{k=1}^K \pi_k \mathcal{N}(w_m | \mu_k, \sigma_k^2)}. \quad (5.208)$$

### 5.33

Let  $(x_1, x_2)$  be the Cartesian coordinates of the end-effector of a two-link robot arm whose joint angles are  $\theta_1$  and  $\theta_2$  and whose arm lengths are  $L_1$  and  $L_2$ . Let us assume that the origin of the coordinate system is given by the attachment point of the lower arm. We have

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} L_1 \cos \theta_1 \\ L_1 \sin \theta_1 \end{bmatrix} + \begin{bmatrix} L_2 \cos(\theta_2 - (\pi - \theta_1)) \\ L_2 \sin(\theta_2 - (\pi - \theta_1)) \end{bmatrix}. \quad (5.209)$$

The second term of the right hand side can be written as

$$\begin{bmatrix} L_2 \cos(\theta_1 + \theta_2 - \pi) \\ L_2 \sin(\theta_1 + \theta_2 - \pi) \end{bmatrix} = \begin{bmatrix} -L_2 \cos(\theta_1 + \theta_2) \\ -L_2 \sin(\theta_1 + \theta_2) \end{bmatrix}. \quad (5.210)$$

Therefore,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} L_1 \cos \theta_1 - L_2 \cos(\theta_1 + \theta_2) \\ L_1 \sin \theta_1 - L_2 \sin(\theta_1 + \theta_2) \end{bmatrix}. \quad (5.211)$$

### 5.34

Let

$$E_n(\mathbf{w}) = -\ln \left( \sum_{k=1}^K \pi_k \mathcal{N}_{nk} \right), \quad (5.212)$$

where

$$\begin{aligned} \pi_k &= \frac{\exp(a_k)}{\sum_{k=1}^K \exp(a_k)}, \\ \mathcal{N}_{nk} &= \mathcal{N}(\mathbf{t}_n | \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w}) \mathbf{I}). \end{aligned} \quad (5.213)$$

Then,

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{k'=1}^K \frac{\partial E_n}{\partial \pi_{k'}} \frac{\partial \pi_{k'}}{\partial a_k^\pi}. \quad (5.214)$$

We have

$$\frac{\partial E_n}{\partial \pi_{k'}} = -\frac{\mathcal{N}_{nk'}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.215)$$

By 5.32(a),

$$\frac{\partial \pi_{k'}}{\partial a_k^\pi} = I_{kk'} \pi_{k'} - \pi_k \pi_{k'}. \quad (5.216)$$

Then,

$$\frac{\partial E_n}{\partial a_k^\pi} = -\sum_{k'=1}^K \frac{\mathcal{N}_{nk'}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}} (I_{kk'} \pi_{k'} - \pi_k \pi_{k'}). \quad (5.217)$$

The right hand side can be written as

$$\pi_k \frac{\sum_{k'=1}^K \pi_{k'} \mathcal{N}_{nk'}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}} - \frac{\pi_k \mathcal{N}_{nk}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}} = \pi_k - \frac{\pi_k \mathcal{N}_{nk}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.218)$$

Therefore,

$$\frac{\partial E_n}{\partial a_k^\pi} = \pi_k - \gamma_{nk}, \quad (5.219)$$

where

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}_{nk}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.220)$$

### 5.35

Let

$$E_n(\mathbf{w}) = -\ln \left( \sum_{k=1}^K \pi_k \mathcal{N}_{nk} \right), \quad (5.221)$$

where

$$\begin{aligned} \mathcal{N}_{nk} &= \mathcal{N}(\mathbf{t}_n | \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w}) \mathbf{I}), \\ \mu_{km} &= a_{km}^\mu. \end{aligned} \quad (5.222)$$

Then,

$$\frac{\partial E_n}{\partial a_{km}^\mu} = \frac{\partial E_n}{\partial \mathcal{N}_{nk}} \frac{\partial \mathcal{N}_{nk}}{\partial a_{km}^\mu}. \quad (5.223)$$

We have

$$\frac{\partial E_n}{\partial \mathcal{N}_{nk}} = -\frac{\pi_k}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.224)$$

We have

$$\frac{\partial \mathcal{N}_{nk}}{\partial a_{km}^\mu} = \frac{t_{nm} - \mu_{km}}{\sigma_k^2} \mathcal{N}_{nk}. \quad (5.225)$$

Then,

$$\frac{\partial E_n}{\partial a_{km}^\mu} = -\frac{\pi_k}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}} \frac{t_{nm} - \mu_{km}}{\sigma_k^2} \mathcal{N}_{nk}. \quad (5.226)$$

Therefore,

$$\frac{\partial E_n}{\partial a_{km}^\mu} = \gamma_{nk} \frac{\mu_{km} - t_{nm}}{\sigma_k^2}, \quad (5.227)$$

where

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}_{nk}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.228)$$

## 5.36

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables in  $M$  dimensions. Let

$$E_n(\mathbf{w}) = -\ln \left( \sum_{k=1}^K \pi_k \mathcal{N}_{nk} \right), \quad (5.229)$$

where

$$\begin{aligned} \mathcal{N}_{nk} &= \mathcal{N} (\mathbf{t}_n | \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w}) \mathbf{I}), \\ \sigma_k &= \exp(a_k^\sigma). \end{aligned} \quad (5.230)$$

Then,

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial \mathcal{N}_{nk}} \frac{\partial \mathcal{N}_{nk}}{\partial a_k^\sigma}. \quad (5.231)$$

We have

$$\frac{\partial E_n}{\partial \mathcal{N}_{nk}} = -\frac{\pi_k}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.232)$$

We have

$$\frac{\partial \mathcal{N}_{nk}}{\partial a_k^\sigma} = \frac{\partial \mathcal{N}_{nk}}{\partial \sigma_k} \frac{\partial \sigma_k}{\partial a_k^\sigma}. \quad (5.233)$$

The right hand side can be written as

$$\left( -M\sigma_k^{-1} + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3} \right) \mathcal{N}_{nk} \sigma_k = \left( -M + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2} \right) \mathcal{N}_{nk}. \quad (5.234)$$

Then,

$$\frac{\partial E_n}{\partial a_k^\sigma} = -\frac{\pi_k}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}} \left( -M + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2} \right) \mathcal{N}_{nk}. \quad (5.235)$$

Therefore,

$$\frac{\partial E_n}{\partial a_k^\sigma} = \gamma_{nk} \left( M - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2} \right), \quad (5.236)$$

where

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}_{nk}}{\sum_{k=1}^K \pi_k \mathcal{N}_{nk}}. \quad (5.237)$$

### 5.37

Let  $\mathbf{t}$  be a variable such that

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{x}) \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}). \quad (5.238)$$

(a)

We have

$$\mathbb{E}(\mathbf{t}|\mathbf{x}) = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \quad (5.239)$$

The right hand side can be written as

$$\sum_{k=1}^K \pi_k(\mathbf{x}) \int \mathbf{t} \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t} = \sum_{k=1}^K \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}). \quad (5.240)$$

Therefore,

$$\mathbb{E}(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}). \quad (5.241)$$

(b)

We have

$$\text{cov}(\mathbf{t}|\mathbf{x}) = \int (\mathbf{t} - \mathbb{E}(\mathbf{t}|\mathbf{x})) (\mathbf{t} - \mathbb{E}(\mathbf{t}|\mathbf{x}))^\top p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \quad (5.242)$$

The right hand side can be written as

$$\begin{aligned} & \int \mathbf{t} \mathbf{t}^\top p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \mathbb{E}(\mathbf{t}|\mathbf{x}) \left( \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right)^\top - \left( \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) \mathbb{E}(\mathbf{t}|\mathbf{x})^\top \\ & + \mathbb{E}(\mathbf{t}|\mathbf{x}) \mathbb{E}(\mathbf{t}|\mathbf{x})^\top \int p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \\ & = \int \mathbf{t} \mathbf{t}^\top p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \mathbb{E}(\mathbf{t}|\mathbf{x}) \mathbb{E}(\mathbf{t}|\mathbf{x})^\top. \end{aligned} \quad (5.243)$$

The first term of the right hand side can be written as

$$\sum_{k=1}^K \pi_k(\mathbf{x}) \int \mathbf{t} \mathbf{t}^\top \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t}. \quad (5.244)$$

The integral of the right hand side can be written as

$$\begin{aligned}
& \int (\mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x}) + \boldsymbol{\mu}_k(\mathbf{x})) (\mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x}) + \boldsymbol{\mu}_k(\mathbf{x}))^\top \mathcal{N}(\mathbf{t} | \boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t} \\
&= \int (\mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x})) (\mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x}))^\top \mathcal{N}(\mathbf{t} | \boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t} \\
&\quad + \boldsymbol{\mu}_k(\mathbf{x}) \left( \int (\mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x})) \mathcal{N}(\mathbf{t} | \boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t} \right)^\top \\
&\quad + \left( \int (\mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x})) \mathcal{N}(\mathbf{t} | \boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t} \right) \boldsymbol{\mu}_k(\mathbf{x})^\top \\
&\quad + \boldsymbol{\mu}_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x})^\top \int \mathcal{N}(\mathbf{t} | \boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}) d\mathbf{t}.
\end{aligned} \tag{5.245}$$

The right hand side can be written as

$$\sigma_k^2(\mathbf{x})\mathbf{I} + \boldsymbol{\mu}_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})^\top. \tag{5.246}$$

Therefore, by (a),

$$\begin{aligned}
\text{cov}(\mathbf{t} | \mathbf{x}) &= \sum_{k=1}^K \pi_k(\mathbf{x}) (\sigma_k^2(\mathbf{x})\mathbf{I} + \boldsymbol{\mu}_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})^\top) \\
&\quad - \left( \sum_{k=1}^K \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}) \right) \left( \sum_{k=1}^K \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}) \right)^\top.
\end{aligned} \tag{5.247}$$

### 5.38

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned}
p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1}\mathbf{I}).
\end{aligned} \tag{5.248}$$

By marginalisation,

$$p(t_{N+1} | \mathbf{t}) = \int p(t_{N+1} | \mathbf{w}) p(\mathbf{w} | \mathbf{t}) d\mathbf{w}. \tag{5.249}$$

By the Bayes' theorem,

$$p(\mathbf{w} | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}). \tag{5.250}$$

The logarithm of the right hand side except the terms independent of  $\mathbf{w}$  can be written as  $-E(\mathbf{w})$  where

$$E(\mathbf{w}) = \frac{\beta}{2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n, \mathbf{w}))^2 + \frac{\alpha}{2} \|\mathbf{w}\|^2. \quad (5.251)$$

Let  $\mathbf{w}_{\text{MAP}}$  be a stationary point of  $p(\mathbf{w}|\mathbf{t})$ . Then, we have a Taylor series

$$\begin{aligned} E(\mathbf{w}) &= E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2} (\mathbf{w} - \mathbf{w}_{\text{MAP}})^T \mathbf{A} (\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ &\quad + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3), \end{aligned} \quad (5.252)$$

where

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \quad (5.253)$$

Then,

$$p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}). \quad (5.254)$$

By a Taylor series

$$y(\mathbf{x}_{N+1}, \mathbf{w}) = y(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) + \mathbf{g}^T(\mathbf{w} - \mathbf{w}_{\text{MAP}}) + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^2), \quad (5.255)$$

where

$$\mathbf{g} = \nabla_{\mathbf{w}} y(\mathbf{x}_{N+1}, \mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}, \quad (5.256)$$

we have

$$p(t_{N+1}|\mathbf{w}) \simeq \mathcal{N}(t_{N+1}|y(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) + \mathbf{g}^T(\mathbf{w} - \mathbf{w}_{\text{MAP}}), \beta^{-1}). \quad (5.257)$$

Then, the logarithm of the integrand except the terms independent of  $\mathbf{w}$  can be approximated as

$$\begin{aligned} &- \frac{\beta}{2} (t_{N+1} - y(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) - \mathbf{g}^T(\mathbf{w} - \mathbf{w}_{\text{MAP}}))^2 \\ &- \frac{1}{2} (\mathbf{w} - \mathbf{w}_{\text{MAP}})^T \mathbf{A} (\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ &= -\frac{1}{2} \mathbf{v}^T \mathbf{M} \mathbf{v}, \end{aligned} \quad (5.258)$$

where

$$\begin{aligned} \mathbf{v} &= \begin{bmatrix} \mathbf{w} - \mathbf{w}_{\text{MAP}} \\ t_{N+1} - y(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) \end{bmatrix}, \\ \mathbf{M} &= \begin{bmatrix} \mathbf{A} + \beta \mathbf{g} \mathbf{g}^T & -\beta \mathbf{g} \\ -\beta \mathbf{g}^T & \beta \end{bmatrix}. \end{aligned} \quad (5.259)$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{g} \\ \mathbf{g}^\top \mathbf{A}^{-1} & \beta^{-1} + \mathbf{g}^\top \mathbf{A}^{-1}\mathbf{g} \end{bmatrix}. \quad (5.260)$$

Therefore,

$$p(t_{N+1}|\mathbf{t}) \simeq \mathcal{N}(t_{N+1}|y(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}), \beta^{-1} + \mathbf{g}^\top \mathbf{A}^{-1}\mathbf{g}). \quad (5.261)$$

### 5.39

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (5.262)$$

where  $\mathbf{w}$  is a vector in  $M$  dimensions. By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \quad (5.263)$$

The logarithm of the integrand of the right hand side can be written as

$$\begin{aligned} & -\frac{N}{2} \ln(2\pi\beta^{-1}) - \frac{\beta}{2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n, \mathbf{w}))^2 \\ & - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\alpha^{-1}\mathbf{I})) - \frac{\alpha}{2} \|\mathbf{w}\|^2 \\ & = -E(\mathbf{w}) - \frac{N+M}{2} \ln(2\pi) + \frac{N}{2} \ln \beta + \frac{M}{2} \ln \alpha, \end{aligned} \quad (5.264)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \sum_{n=1}^N (t_n - y(\mathbf{x}_n, \mathbf{w}))^2 + \frac{\alpha}{2} \|\mathbf{w}\|^2. \quad (5.265)$$

Let  $\mathbf{w}_{\text{MAP}}$  be a stationary point of  $E$ . Then, we have a Taylor series

$$\begin{aligned} E(\mathbf{w}) &= E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ &\quad + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3), \end{aligned} \quad (5.266)$$

where

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \quad (5.267)$$

Then, the logarithm of the integrand can be approximated as

$$\begin{aligned} & -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^T \mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ & - \frac{N+M}{2} \ln(2\pi) + \frac{N}{2} \ln \beta + \frac{M}{2} \ln \alpha. \end{aligned} \quad (5.268)$$

Therefore,

$$\ln p(\mathbf{t}) \simeq -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2} \ln |\det \mathbf{A}| - \frac{N}{2} \ln(2\pi) + \frac{N}{2} \ln \beta + \frac{M}{2} \ln \alpha. \quad (5.269)$$

## 5.40

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables from the standard basis in  $K$  dimensions such that

$$\begin{aligned} p(\mathbf{t}_n | \mathbf{w}) &= \prod_{k=1}^K y_{nk}^{t_{nk}}, \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (5.270)$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{a}_n), \\ y_k(\mathbf{a}) &= \frac{\exp(a_k)}{\sum_{k'=1}^K \exp(a_{k'})}, \\ \mathbf{a}_n &= \mathbf{a}(\mathbf{x}_n, \mathbf{w}). \end{aligned} \quad (5.271)$$

By marginalisation,

$$p(\mathbf{t}_{N+1} | \mathbf{T}) = \int p(\mathbf{t}_{N+1} | \mathbf{w}) p(\mathbf{w} | \mathbf{T}) d\mathbf{w}. \quad (5.272)$$

By the Bayes' theorem,

$$p(\mathbf{w} | \mathbf{T}) p(\mathbf{T}) = p(\mathbf{T} | \mathbf{w}) p(\mathbf{w}). \quad (5.273)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{w}$  can be written as  $-E(\mathbf{w})$  where

$$E(\mathbf{w}) = - \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} + \frac{\alpha}{2} \|\mathbf{w}\|^2. \quad (5.274)$$

Let  $\mathbf{w}_{\text{MAP}}$  be a stationary point of  $E$ . Then, we have a Taylor series

$$\begin{aligned} E(\mathbf{w}) = & E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\top} \mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ & + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3), \end{aligned} \quad (5.275)$$

where

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \quad (5.276)$$

Then,

$$p(\mathbf{w}|\mathbf{T}) \simeq \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}). \quad (5.277)$$

We have

$$p(\mathbf{t}_{N+1}|\mathbf{w}) = \prod_{k=1}^K y_k(\mathbf{a}_{N+1})^{t_{N+1,k}}. \quad (5.278)$$

We have the Taylor series

$$\begin{aligned} \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}) = & \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) + \mathbf{B}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ & + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^2), \end{aligned} \quad (5.279)$$

where

$$\mathbf{B} = \nabla_{\mathbf{w}} \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \quad (5.280)$$

Then,

$$p(\mathbf{t}_{N+1}|\mathbf{w}) \simeq \int \delta(\mathbf{a}_{N+1} - \mathbf{c}_{N+1}) \left( \prod_{k=1}^K y_k(\mathbf{a}_{N+1})^{t_{N+1,k}} \right) d\mathbf{a}_{N+1}, \quad (5.281)$$

where  $\delta$  is the Dirac delta function and

$$\mathbf{c}_{N+1} = \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) + \mathbf{B}(\mathbf{w} - \mathbf{w}_{\text{MAP}}). \quad (5.282)$$

Then,

$$\int p(\mathbf{t}_{N+1}|\mathbf{w}) p(\mathbf{w}|\mathbf{T}) d\mathbf{w} \simeq \int \left( \prod_{k=1}^K y_k(\mathbf{a}_{N+1})^{t_{N+1,k}} \right) p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1}, \quad (5.283)$$

where

$$p(\mathbf{a}_{N+1}) = \int \delta(\mathbf{a}_{N+1} - \mathbf{c}_{N+1}) \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}) d\mathbf{w}. \quad (5.284)$$

Since  $\mathbf{t}_{N+1}$  is a variable from the standard basis in  $K$  dimensions, the right hand side of the approximation can be written as

$$\prod_{k=1}^K \left( \int y_k(\mathbf{a}_{N+1}) p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1} \right)^{t_{N+1,k}}. \quad (5.285)$$

We have

$$\mathbb{E} \mathbf{a}_{N+1} = \int \mathbf{a}_{N+1} p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1}. \quad (5.286)$$

The right hand side can be written as

$$\int \mathbf{c}_{N+1} \mathcal{N}(\mathbf{w} | \mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}) d\mathbf{w} = \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}). \quad (5.287)$$

Then,

$$\text{var } \mathbf{a}_{N+1} = \int (\mathbf{a}_{N+1} - \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}))^2 p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1}. \quad (5.288)$$

The right hand side can be written as

$$\int (\mathbf{c}_{N+1} - \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}))^2 \mathcal{N}(\mathbf{w} | \mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}) d\mathbf{w} = \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top. \quad (5.289)$$

Then,

$$p(\mathbf{a}_{N+1}) = \mathcal{N}(\mathbf{a}_{N+1} | \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}), \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top). \quad (5.290)$$

Therefore,

$$p(\mathbf{t}_{N+1} | \mathbf{T}) \simeq \prod_{k=1}^K v_{N+1,k}^{t_{N+1,k}}, \quad (5.291)$$

where

$$v_{N+1,k} = \int y_k(\mathbf{a}) \mathcal{N}(\mathbf{a} | \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}), \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top) d\mathbf{a}. \quad (5.292)$$

## 5.41

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables such that

$$\begin{aligned} p(\mathbf{t}_n | \mathbf{w}) &= \prod_{k=1}^K y_{nk}^{t_{nk}}, \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (5.293)$$

where

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ y_{nk} &= y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^K y_{nk} &= 1. \end{aligned} \tag{5.294}$$

By marginalisation,

$$p(\mathbf{T}) = \int p(\mathbf{T}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{5.295}$$

The logarithm of the integrand of the right hand side can be written as

$$\begin{aligned} &\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln (\det(\alpha^{-1}\mathbf{I})) - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} \\ &= -E(\mathbf{w}) - \frac{M}{2} \ln(2\pi) + \frac{M}{2} \ln \alpha, \end{aligned} \tag{5.296}$$

where

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk} + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \tag{5.297}$$

Let  $\mathbf{w}_{\text{MAP}}$  be a stationary point of  $E$ . Then, we have a Taylor series

$$\begin{aligned} E(\mathbf{w}) &= E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\ &\quad + O(\|(\mathbf{w} - \mathbf{w}_{\text{MAP}})\|^3), \end{aligned} \tag{5.298}$$

where

$$\mathbf{A} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \tag{5.299}$$

Then, the logarithm of the integrand can be approximated as

$$-E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) - \frac{M}{2} \ln(2\pi) + \frac{M}{2} \ln \alpha. \tag{5.300}$$

Therefore,

$$\ln p(\mathbf{T}) \simeq -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2} \ln |\det \mathbf{A}| + \frac{M}{2} \ln \alpha. \tag{5.301}$$

## 6 Kernel Methods

### 6.1

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, 1), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \lambda^{-1} \mathbf{I}). \end{aligned} \quad (6.1)$$

By the Bayes' theorem,

$$p(\mathbf{w} | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}). \quad (6.2)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{w}$  can be written as  $-E_{\mathbf{w}}(\mathbf{w})$  where

$$E_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2, \quad (6.3)$$

so that

$$E_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2. \quad (6.4)$$

(a)

Setting the derivative of  $E_{\mathbf{w}}$  with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = -\Phi^\top (\mathbf{t} - \Phi \mathbf{w}) + \lambda \mathbf{w}. \quad (6.5)$$

Let the stationary point of  $E_{\mathbf{w}}$  be  $\mathbf{w}_{\text{MAP}}$ . Then,

$$\mathbf{w}_{\text{MAP}} = \Phi^\top \mathbf{a}_{\text{MAP}}, \quad (6.6)$$

where

$$\mathbf{a}_{\text{MAP}} = \frac{1}{\lambda} (\mathbf{t} - \Phi \mathbf{w}_{\text{MAP}}). \quad (6.7)$$

(b)

Let

$$\mathbf{w} = \Phi^\top \mathbf{a}. \quad (6.8)$$

Then,

$$E_{\mathbf{w}}(\mathbf{w}) = E_{\mathbf{a}}(\mathbf{a}), \quad (6.9)$$

where

$$E_{\mathbf{a}}(\mathbf{a}) = \frac{1}{2} \|\mathbf{t} - \Phi \Phi^T \mathbf{a}\|^2 + \frac{\lambda}{2} \mathbf{a}^T \Phi \Phi^T \mathbf{a}. \quad (6.10)$$

The right hand side can be written as

$$\frac{1}{2} \|\mathbf{t}\|^2 - \mathbf{t}^T \mathbf{K} \mathbf{a} + \frac{1}{2} \mathbf{a}^T \mathbf{K} \mathbf{K} \mathbf{a} + \frac{\lambda}{2} \mathbf{a}^T \mathbf{K} \mathbf{a}, \quad (6.11)$$

where

$$\mathbf{K} = \Phi \Phi^T. \quad (6.12)$$

Setting the derivative of  $E_{\mathbf{a}}$  with respect to  $\mathbf{a}$  to zero gives

$$\mathbf{0} = -\mathbf{K} \mathbf{t} + \mathbf{K} \mathbf{K} \mathbf{a} + \lambda \mathbf{K} \mathbf{a}. \quad (6.13)$$

Then, the stationary point of  $E_{\mathbf{a}}$  is given by

$$\mathbf{a}_{\text{MAP}} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}. \quad (6.14)$$

(c)

By (a) and (b),

$$\frac{1}{\lambda} (\mathbf{t} - \Phi \mathbf{w}_{\text{MAP}}) = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}, \quad (6.15)$$

so that

$$\mathbf{t} - \Phi \mathbf{w}_{\text{MAP}} = \lambda (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}. \quad (6.16)$$

The right hand side can be written as

$$(\mathbf{K} + \lambda \mathbf{I} - \mathbf{K})(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t} = \mathbf{t} - \mathbf{K} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}. \quad (6.17)$$

Then,

$$\Phi \mathbf{w}_{\text{MAP}} = \mathbf{K} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}, \quad (6.18)$$

so that

$$\Phi^T \Phi \mathbf{w}_{\text{MAP}} = \Phi^T \mathbf{K} (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}. \quad (6.19)$$

By (b),

$$\mathbf{w}_{\text{MAP}} = \Phi^T (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}, \quad (6.20)$$

so that

$$\lambda \mathbf{w}_{\text{MAP}} = \lambda \Phi^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}. \quad (6.21)$$

Then,

$$(\Phi^\top \Phi + \lambda \mathbf{I}) \mathbf{w}_{\text{MAP}} = \Phi^\top (\mathbf{K} + \lambda \mathbf{I})(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{t}. \quad (6.22)$$

Therefore,

$$\mathbf{w}_{\text{MAP}} = (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{t}. \quad (6.23)$$

## 6.2 (Incomplete)

Let  $t_1, \dots, t_N$  be variables such that

$$t_n \in \{-1, 1\}, \\ p(t_n | \mathbf{w}) = \left( \frac{1 + y_n}{2} \right)^{\frac{1+t_n}{2}} \left( \frac{1 - y_n}{2} \right)^{\frac{1-t_n}{2}}, \quad (6.24)$$

where

$$y_n = f(\mathbf{w}^\top \phi_n), \\ f(a) = \begin{cases} 1, & a \geq 0, \\ -1, & \text{otherwise.} \end{cases} \quad (6.25)$$

Let

$$E(\mathbf{w}) = - \sum_{n \in \mathcal{M}} \mathbf{w}^\top \phi_n t_n, \quad (6.26)$$

where  $\mathcal{M}$  is the set of all misclassification. Setting the derivative of  $E$  with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = - \sum_{n \in \mathcal{M}} t_n \phi_n. \quad (6.27)$$

## 6.3

We have

$$\|\mathbf{x} - \mathbf{x}_n\|^2 = \mathbf{x}^\top \mathbf{x} - 2\mathbf{x}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{x}_n. \quad (6.28)$$

The right hand side can be written as

$$k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{x}_n) + k(\mathbf{x}_n, \mathbf{x}_n), \quad (6.29)$$

where

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'. \quad (6.30)$$

## 6.4

Let  $\mathbf{A}$  be a  $2 \times 2$  matrix with positive eigenvalues and with at least one negative element. We have

$$\det(\lambda\mathbf{I} - \mathbf{A}) = (\lambda - A_{11})(\lambda - A_{22}) - A_{12}A_{21}. \quad (6.31)$$

The right hand side can be written as

$$\lambda^2 - (A_{11} + A_{22})\lambda + A_{11}A_{22} - A_{12}A_{21}. \quad (6.32)$$

Then,

$$\begin{aligned} (A_{11} + A_{22})^2 - 4(A_{11}A_{22} - A_{12}A_{21}) &> 0, \\ A_{11} + A_{22} &> 0, \\ A_{11}A_{22} - A_{12}A_{21} &> 0. \end{aligned} \quad (6.33)$$

Therefore, an example is

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix}. \quad (6.34)$$

## 6.5

Let

$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^\top \boldsymbol{\phi}_1(\mathbf{x}'). \quad (6.35)$$

(a)

Let

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}'), \quad (6.36)$$

where  $c$  is a positive constant. The right hand side can be written as

$$c\boldsymbol{\phi}_1(\mathbf{x})^\top \boldsymbol{\phi}_1(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}'), \quad (6.37)$$

where

$$\boldsymbol{\phi} = \sqrt{c}\boldsymbol{\phi}_1. \quad (6.38)$$

Therefore,  $k$  is a valid kernel.

(b)

Let

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}'). \quad (6.39)$$

The right hand side can be written as

$$f(\mathbf{x})\phi_1(\mathbf{x})^\top \phi_1(\mathbf{x}')f(\mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}'), \quad (6.40)$$

where

$$\phi = f\phi_1. \quad (6.41)$$

Therefore,  $k$  is a valid kernel.

## 6.6

Let

$$k_1(\mathbf{x}, \mathbf{x}') = \phi_1(\mathbf{x})^\top \phi_1(\mathbf{x}'). \quad (6.42)$$

(a)

Let

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=0}^J c_j q_j(\mathbf{x}, \mathbf{x}'), \quad (6.43)$$

where  $c_0, \dots, c_J$  are nonnegative constants and

$$q_j(\mathbf{x}, \mathbf{x}') = (k_1(\mathbf{x}, \mathbf{x}'))^j. \quad (6.44)$$

We have

$$q_0(\mathbf{x}, \mathbf{x}') = 1. \quad (6.45)$$

Let us assume that  $q_j$  is a valid kernel. Then,

$$q_{j+1}(\mathbf{x}, \mathbf{x}') = q_j(\mathbf{x}, \mathbf{x}')\phi_1(\mathbf{x})^\top \phi_1(\mathbf{x}'). \quad (6.46)$$

The right hand side can be written as

$$q_j(\mathbf{x}, \mathbf{x}') \sum_{m=1}^M \phi_{1m}(\mathbf{x})\phi_{1m}(\mathbf{x}') = \sum_{m=1}^M \phi_{1m}(\mathbf{x})q_j(\mathbf{x}, \mathbf{x}')\phi_{1m}(\mathbf{x}'). \quad (6.47)$$

By 6.5(b) and 6.7(a), the right hand side is a valid kernel. Then,  $q_{j+1}$  is a valid kernel. Then, the assumption is proved by induction on  $j$ . Therefore, by 6.5(a) and 6.7(a),  $k$  is a valid kernel.

(b)

Let

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}')). \quad (6.48)$$

The right hand side can be written as

$$\sum_{j=0}^{\infty} \frac{1}{j!} (k_1(\mathbf{x}, \mathbf{x}'))^j. \quad (6.49)$$

Therefore, by (a),  $k$  is a valid kernel.

## 6.7

Let

$$\begin{aligned} k_1(\mathbf{x}, \mathbf{x}') &= \boldsymbol{\phi}_1(\mathbf{x})^\top \boldsymbol{\phi}_1(\mathbf{x}'), \\ k_2(\mathbf{x}, \mathbf{x}') &= \boldsymbol{\phi}_2(\mathbf{x})^\top \boldsymbol{\phi}_2(\mathbf{x}'). \end{aligned} \quad (6.50)$$

(a)

Let

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}'). \quad (6.51)$$

The right hand side can be written as

$$\boldsymbol{\phi}_1(\mathbf{x})^\top \boldsymbol{\phi}_1(\mathbf{x}') + \boldsymbol{\phi}_2(\mathbf{x})^\top \boldsymbol{\phi}_2(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}'), \quad (6.52)$$

where

$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{bmatrix}. \quad (6.53)$$

Therefore,  $k$  is a valid kernel.

(b)

Let

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}'). \quad (6.54)$$

Then,

$$\ln k(\mathbf{x}, \mathbf{x}') = \ln k_1(\mathbf{x}, \mathbf{x}') + \ln k_2(\mathbf{x}, \mathbf{x}'). \quad (6.55)$$

By (a) and 6.6(b), the right hand side is a valid kernel. Then,  $\ln k$  is a valid kernel. Therefore, by 6.6(b),  $k$  is a valid kernel.

## 6.8

(a)

Let

$$k(\mathbf{x}, \mathbf{x}') = k_1(\phi(\mathbf{x}), \phi(\mathbf{x}')), \quad (6.56)$$

where  $k_1$  is a valid kernel. Then,  $k$  is a valid kernel.

(b)

Let

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A} \mathbf{x}', \quad (6.57)$$

where  $\mathbf{A}$  is a symmetric positive semidefinite matrix. There exists  $\mathbf{M}$  such that

$$\mathbf{A} = \mathbf{M}^\top \mathbf{M}. \quad (6.58)$$

Then,

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}'), \quad (6.59)$$

where

$$\phi(\mathbf{x}) = \mathbf{M}\mathbf{x}. \quad (6.60)$$

Therefore,  $k$  is a valid kernel.

## 6.9

Let

$$\begin{aligned} k_a(\mathbf{x}_a, \mathbf{x}'_a) &= \phi_a(\mathbf{x}_a)^\top \phi_a(\mathbf{x}'_a), \\ k_b(\mathbf{x}_b, \mathbf{x}'_b) &= \phi_b(\mathbf{x}_b)^\top \phi_b(\mathbf{x}'_b). \end{aligned} \quad (6.61)$$

(a)

Let

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b), \quad (6.62)$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}. \quad (6.63)$$

The right hand side can be written as

$$\phi_a(\mathbf{x}_a)^\top \phi_a(\mathbf{x}'_a) + \phi_b(\mathbf{x}_b)^\top \phi_b(\mathbf{x}'_b) = \phi(\mathbf{x})^\top \phi(\mathbf{x}'), \quad (6.64)$$

where

$$\phi(\mathbf{x}) = \begin{bmatrix} \phi_a(\mathbf{x}_a) \\ \phi_b(\mathbf{x}_b) \end{bmatrix}. \quad (6.65)$$

Therefore,  $k$  is a valid kernel.

(b)

Let

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b), \quad (6.66)$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}. \quad (6.67)$$

Then,

$$\ln k(\mathbf{x}, \mathbf{x}') = \ln k_a(\mathbf{x}_a, \mathbf{x}'_a) + \ln k_b(\mathbf{x}_b, \mathbf{x}'_b). \quad (6.68)$$

By (a) and 6.6(b), the right hand side is a valid kernel. Then,  $\ln k$  is a valid kernel. Therefore, by 6.6(b),  $k$  is a valid kernel.

## 6.10 (Incomplete)

Let

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})f(\mathbf{x}'). \quad (6.69)$$

## 6.11

Let

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right). \quad (6.70)$$

The right hand side can be written as

$$\exp\left(-\frac{\mathbf{x}^\top \mathbf{x}}{2\sigma^2}\right) \exp\left(\frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2}\right) \exp\left(-\frac{\mathbf{x}'^\top \mathbf{x}'}{2\sigma^2}\right). \quad (6.71)$$

We have

$$\exp\left(\frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2}\right) = \sum_{j=0}^{\infty} q_j(\mathbf{x}, \mathbf{x}'), \quad (6.72)$$

where

$$q_j(\mathbf{x}, \mathbf{x}') = \frac{1}{j!} \left( \frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2} \right)^j. \quad (6.73)$$

By 6.6(a),  $q_j$  is a valid kernel. Then, there exists  $\psi_j$  such that

$$q_j(\mathbf{x}, \mathbf{x}') = \psi_j(\mathbf{x})^\top \psi_j(\mathbf{x}'). \quad (6.74)$$

Then,

$$\exp \left( \frac{\mathbf{x}^\top \mathbf{x}'}{\sigma^2} \right) = \sum_{j=0}^{\infty} \psi_j(\mathbf{x})^\top \psi_j(\mathbf{x}'). \quad (6.75)$$

The right hand side can be written as  $\phi(\mathbf{x})^\top \phi(\mathbf{x}')$ , where

$$\phi = \begin{bmatrix} \psi_1 \\ \vdots \\ \psi_j \\ \vdots \end{bmatrix}. \quad (6.76)$$

Therefore,  $k$  can be written as the inner product of an infinite-dimensional feature vector.

## 6.12

Let  $A_1, \dots, A_{2^{|D|}}$  are all the subsets of  $D$  ordered in the alphabetical order where  $|.|$  is the number of elements in  $(.)$ . Let

$$k(A_m, A_{m'}) = 2^{|A_m \cap A_{m'}|}. \quad (6.77)$$

Let  $\phi$  be a vector such that

$$\phi_m(A) = \begin{cases} 1, & \text{if } A_m \subseteq A, \\ 0, & \text{otherwise.} \end{cases} \quad (6.78)$$

If

$$|A_m \cap A_{m'}| = 0, \quad (6.79)$$

then

$$k(A_m, A_{m'}) = 1. \quad (6.80)$$

The right hand side can be written as

$$\phi(A_m)^\top \phi(A_{m'}). \quad (6.81)$$

Let us assume that if

$$|A_m \cap A_{m'}| = n, \quad (6.82)$$

then

$$k(A_m, A_{m'}) = \phi(A_m)^\top \phi(A_{m'}). \quad (6.83)$$

Let

$$A_{m''} = A_{m'} \cup \{x\}, \quad (6.84)$$

where  $x$  is an element such that

$$x \in A_m - A_{m'}. \quad (6.85)$$

Then,

$$|A_m \cap A_{m''}| = n + 1, \quad (6.86)$$

so that

$$k(A_m, A_{m''}) = 2^{n+1}. \quad (6.87)$$

We have

$$\phi(A_m)^\top \phi(A_{m''}) = \phi(A_m)^\top \phi(A_{m'}) + \phi(A_m)^\top (\phi(A_{m''}) - \phi(A_{m'})). \quad (6.88)$$

By the assumption, the first term of the right hand side is  $2^n$ . By the assumption,  $\phi(A_m)$  and  $\phi(A_{m'})$  have a set of  $2^n$  1s in common. Then,  $\phi(A_m)$  and  $\phi(A_{m''}) - \phi(A_{m'})$  have another set of  $2^n$  1s in common. Then, the second term of the right hand side is  $2^n$ . Then,

$$\phi(A_m)^\top \phi(A_{m''}) = 2^{n+1}, \quad (6.89)$$

so that

$$k(A_m, A_{m''}) = \phi(A_m)^\top \phi(A_{m''}). \quad (6.90)$$

Therefore, the assumption is proved by induction on  $n$ .

## 6.13

Let

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^\top \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}'), \quad (6.91)$$

where

$$\begin{aligned} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) &= \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x} | \boldsymbol{\theta}), \\ \mathbf{F} &= \mathbb{E}_{\mathbf{x}} (\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^\top). \end{aligned} \quad (6.92)$$

Let

$$k_{\psi}(\mathbf{x}, \mathbf{x}') = \mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x})^\top \mathbf{F}_{\psi}^{-1} \mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x}'), \quad (6.93)$$

where  $\psi$  is invertible and differentiable and

$$\begin{aligned} \mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x}) &= \nabla_{\psi(\boldsymbol{\theta})} \ln p(\mathbf{x} | \boldsymbol{\theta}), \\ \mathbf{F}_{\psi} &= \mathbb{E}_{\mathbf{x}} (\mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x})^\top). \end{aligned} \quad (6.94)$$

We have

$$\mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x}) = \frac{\partial \boldsymbol{\theta}}{\partial \psi(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x} | \boldsymbol{\theta}). \quad (6.95)$$

Then,

$$\mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x}) = \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}). \quad (6.96)$$

We have

$$\mathbf{F}_{\psi} = \int \mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}_{\psi}(\boldsymbol{\theta}, \mathbf{x})^\top p(\mathbf{x}) d\mathbf{x}. \quad (6.97)$$

The right hand side can be written as

$$\begin{aligned} & \int \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^\top \left( \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^\top p(\mathbf{x}) d\mathbf{x} \\ &= \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \left( \int \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^\top p(\mathbf{x}) d\mathbf{x} \right) \left( \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^\top. \end{aligned} \quad (6.98)$$

Then,

$$\mathbf{F}_{\psi} = \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{F} \left( \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^\top. \quad (6.99)$$

Then,

$$\begin{aligned} & k_{\psi}(\mathbf{x}, \mathbf{x}') \\ &= \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^T \left( \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^T \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \mathbf{F}^{-1} \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left( \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}'). \end{aligned} \quad (6.100)$$

The right hand side can be written as

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^T \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'). \quad (6.101)$$

Therefore,

$$k_{\psi}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'). \quad (6.102)$$

## 6.14

Let  $\mathbf{x}$  be a variable such that

$$p(\mathbf{x}|\boldsymbol{\mu}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}). \quad (6.103)$$

Let

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^T \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}'), \quad (6.104)$$

where

$$\begin{aligned} \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) &= \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}|\boldsymbol{\mu}), \\ \mathbf{F} &= \mathbb{E}_{\mathbf{x}} (\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) \mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^T). \end{aligned} \quad (6.105)$$

We have

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = -\mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (6.106)$$

Then,

$$\mathbf{F} = \int \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) d\mathbf{x}. \quad (6.107)$$

The right hand side can be written as

$$\mathbf{S}^{-1} \left( \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) d\mathbf{x} \right) \mathbf{S}^{-1} = \mathbf{S}^{-1}. \quad (6.108)$$

Then,

$$\mathbf{F} = \mathbf{S}^{-1}, \quad (6.109)$$

so that

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^T \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (6.110)$$

Therefore,

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (6.111)$$

## 6.15

Let  $k$  be a positive semidefinite kernel function. Then,

$$k(x, x)k(x', x') - k(x, x')^2 = \det \mathbf{K}, \quad (6.112)$$

where

$$\mathbf{K} = \begin{bmatrix} k(x, x) & k(x, x') \\ k(x, x') & k(x', x') \end{bmatrix}. \quad (6.113)$$

Since  $\mathbf{K}$  is positive semidefinite,

$$\det \mathbf{K} \geq 0. \quad (6.114)$$

Therefore,

$$k(x, x')^2 \leq k(x, x)k(x', x'). \quad (6.115)$$

## 6.16 (Incomplete)

Let

$$J(\mathbf{w}) = f(\mathbf{w}^\top \phi(\mathbf{x}_1), \dots, \mathbf{w}^\top \phi(\mathbf{x}_N)) + g(\mathbf{w}^\top \mathbf{w}), \quad (6.116)$$

where  $g$  is a monotonically increasing function. Setting the derivative with respect to  $\mathbf{w}$  to zero gives

$$\mathbf{0} = \sum_{n=1}^N \frac{\partial f}{\partial \mathbf{w}^\top \phi(\mathbf{x}_n)} \phi(\mathbf{x}_n) + 2g'(\mathbf{w}^\top \mathbf{w}) \mathbf{w}. \quad (6.117)$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} J(\mathbf{w}) = -\frac{1}{2g'(\mathbf{w}^\top \mathbf{w})} \sum_{n=1}^N \frac{\partial f}{\partial \mathbf{w}^\top \phi(\mathbf{x}_n)} \phi(\mathbf{x}_n). \quad (6.118)$$

## 6.17

Let

$$E = \frac{1}{2} \sum_{n=1}^N \int (y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n)^2 p(\boldsymbol{\xi}) d\boldsymbol{\xi}. \quad (6.119)$$

By the transformation

$$\mathbf{z} = \mathbf{x}_n + \boldsymbol{\xi}, \quad (6.120)$$

we have

$$E = \frac{1}{2} \sum_{n=1}^N \int (y(\mathbf{z}) - t_n)^2 p(\mathbf{z} - \mathbf{x}_n) d\mathbf{z}. \quad (6.121)$$

Setting the variation with respect to  $y$  to zero gives

$$0 = \sum_{n=1}^N (y(\mathbf{z}) - t_n) p(\mathbf{z} - \mathbf{x}_n). \quad (6.122)$$

The right hand side can be written as

$$y(\mathbf{z}) \sum_{n=1}^N p(\mathbf{z} - \mathbf{x}_n) - \sum_{n=1}^N t_n p(\mathbf{z} - \mathbf{x}_n). \quad (6.123)$$

Therefore,

$$y(\mathbf{x}) = \sum_{n=1}^N t_n k(\mathbf{x}, \mathbf{x}_n), \quad (6.124)$$

where

$$k(\mathbf{x}, \mathbf{x}_n) = \frac{p(\mathbf{x} - \mathbf{x}_n)}{\sum_{n=1}^N p(\mathbf{x} - \mathbf{x}_n)}. \quad (6.125)$$

## 6.18

Let  $x$  be a variable such that

$$p(x, t) = \frac{1}{N} \sum_{n=1}^N f(x - x_n, t - t_n), \quad (6.126)$$

where

$$f(x, t) = \mathcal{N}(x|0, \sigma^2) \mathcal{N}(t|0, \sigma^2). \quad (6.127)$$

(a)

By the Bayes' theorem,

$$p(t|x) = \frac{p(x, t)}{p(x)}. \quad (6.128)$$

By marginalisation,

$$p(x) = \int p(x, t) dt. \quad (6.129)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} \sum_{n=1}^N \mathcal{N}(x - x_n | 0, \sigma^2) \int \mathcal{N}(t - t_n | 0, \sigma^2) dt \\ &= \frac{1}{N} \sum_{n=1}^N \mathcal{N}(x - x_n | 0, \sigma^2). \end{aligned} \quad (6.130)$$

Then,

$$p(t|x) = \frac{\frac{1}{N} \sum_{n=1}^N \mathcal{N}(x - x_n | 0, \sigma^2) \mathcal{N}(t - t_n | 0, \sigma^2)}{\frac{1}{N} \sum_{n=1}^N \mathcal{N}(x - x_n | 0, \sigma^2)}. \quad (6.131)$$

Therefore,

$$p(t|x) = \sum_{n=1}^N k(x, x_n) \mathcal{N}(t - t_n | 0, \sigma^2), \quad (6.132)$$

where

$$k(x, x_n) = \frac{\mathcal{N}(x - x_n | 0, \sigma^2)}{\sum_{n=1}^N \mathcal{N}(x - x_n | 0, \sigma^2)}. \quad (6.133)$$

**(b)**

We have

$$\mathbb{E}(t|x) = \int t p(t|x) dt. \quad (6.134)$$

By (a), the right hand side can be written as

$$\sum_{n=1}^N k(x, x_n) \int t \mathcal{N}(t - t_n | 0, \sigma^2) dt. \quad (6.135)$$

By the transformation

$$t' = t - t_n, \quad (6.136)$$

the integral can be written as

$$\int (t' + t_n) \mathcal{N}(t' | 0, \sigma^2) dt' = \int t' \mathcal{N}(t' | 0, \sigma^2) dt' + t_n \int \mathcal{N}(t' | 0, \sigma^2) dt'. \quad (6.137)$$

The right hand side can be written as  $t_n$ . Therefore,

$$\mathbb{E}(t|x) = \sum_{n=1}^N k(x, x_n) t_n. \quad (6.138)$$

(c)

We have

$$\text{var}(t|x) = \int (t - \mathbb{E}(t|x))^2 p(t|x) dt. \quad (6.139)$$

By (a) and (b), the right hand side can be written as

$$\sum_{n=1}^N k(x, x_n) \int \left( t - \sum_{n=1}^N k(x, x_n) t_n \right)^2 \mathcal{N}(t - t_n | 0, \sigma^2) dt. \quad (6.140)$$

By the transformation

$$t' = t - t_n, \quad (6.141)$$

the integral can be written as

$$\begin{aligned} & \int \left( t' + t_n - \sum_{n=1}^N k(x, x_n) t_n \right)^2 \mathcal{N}(t'|0, \sigma^2) dt' \\ &= \int t'^2 \mathcal{N}(t'|0, \sigma^2) dt' + 2 \left( t_n - \sum_{n=1}^N k(x, x_n) t_n \right) \int t' \mathcal{N}(t'|0, \sigma^2) dt' \quad (6.142) \\ &+ \left( t_n - \sum_{n=1}^N k(x, x_n) t_n \right)^2 \int \mathcal{N}(t'|0, \sigma^2) dt'. \end{aligned}$$

The right hand side can be written as

$$\sigma^2 + \left( t_n - \sum_{n=1}^N k(x, x_n) t_n \right)^2. \quad (6.143)$$

Then,

$$\text{var}(t|x) = \sum_{n=1}^N k(x, x_n) \left( \sigma^2 + \left( t_n - \sum_{n'=1}^N k(x, x_{n'}) t_{n'} \right)^2 \right). \quad (6.144)$$

Therefore,

$$\text{var}(t|x) = \sigma^2 + \sum_{n=1}^N k(x, x_n) \left( t_n - \sum_{n'=1}^N k(x, x_{n'}) t_{n'} \right)^2. \quad (6.145)$$

## 6.19

Let

$$E = \frac{1}{2} \sum_{n=1}^N \int (y(\mathbf{x}_n - \boldsymbol{\xi}_n) - t_n)^2 p(\boldsymbol{\xi}_n) d\boldsymbol{\xi}_n. \quad (6.146)$$

By the transformation

$$\mathbf{z} = \mathbf{x}_n - \boldsymbol{\xi}_n, \quad (6.147)$$

we have

$$E = \frac{1}{2} \sum_{n=1}^N \int (y(\mathbf{z}) - t_n)^2 p(\mathbf{x}_n - \mathbf{z}) d\mathbf{z}. \quad (6.148)$$

Setting the variation with respect to  $y$  to zero gives

$$0 = \sum_{n=1}^N (y(\mathbf{z}) - t_n) p(\mathbf{x}_n - \mathbf{z}). \quad (6.149)$$

The right hand side can be written as

$$y(\mathbf{z}) \sum_{n=1}^N p(\mathbf{x}_n - \mathbf{z}) - \sum_{n=1}^N t_n p(\mathbf{x}_n - \mathbf{z}). \quad (6.150)$$

Therefore,

$$y(\mathbf{x}) = \sum_{n=1}^N t_n k(\mathbf{x}, \mathbf{x}_n), \quad (6.151)$$

where

$$k(\mathbf{x}, \mathbf{x}_n) = \frac{p(\mathbf{x}_n - \mathbf{x})}{\sum_{n=1}^N p(\mathbf{x}_n - \mathbf{x})}. \quad (6.152)$$

## 6.20

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(\mathbf{t}|\mathbf{y}) &= \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}), \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}), \end{aligned} \quad (6.153)$$

where

$$K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}). \quad (6.154)$$

By the Bayes' theorem,

$$p(t_{N+1}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}'), \quad (6.155)$$

where

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}. \quad (6.156)$$

By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y}. \quad (6.157)$$

The logarithm of the integrand of the right hand side except the terms independent of  $\mathbf{y}_N$  can be written as

$$\begin{aligned} & -\frac{\beta}{2}(\mathbf{t} - \mathbf{y})^\top(\mathbf{t} - \mathbf{y}) - \frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1} \mathbf{y} \\ &= -\frac{1}{2} \begin{bmatrix} \mathbf{y} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I} \\ -\beta\mathbf{I} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{t} \end{bmatrix}. \end{aligned} \quad (6.158)$$

By 2.24,

$$\begin{bmatrix} \beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I} \\ -\beta\mathbf{I} & \beta\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{K} & \mathbf{K} \\ \mathbf{K} & \mathbf{K} + \beta^{-1}\mathbf{I} \end{bmatrix}. \quad (6.159)$$

Then,

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}), \quad (6.160)$$

where

$$\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}. \quad (6.161)$$

Then,

$$p(\mathbf{t}') = \mathcal{N}(\mathbf{t}'|\mathbf{0}, \mathbf{C}'), \quad (6.162)$$

where

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\top & c \end{bmatrix}, \quad (6.163)$$

where

$$\begin{aligned} k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}. \end{aligned} \quad (6.164)$$

By 2.24,

$$\begin{bmatrix} c & \mathbf{k}^\top \\ \mathbf{k} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} s^{-1} & -s^{-1}\mathbf{k}^\top \mathbf{C}^{-1} \\ -s^{-1}\mathbf{C}^{-1}\mathbf{k} & \mathbf{C}^{-1} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\top \mathbf{C}^{-1} \end{bmatrix}, \quad (6.165)$$

where

$$s = c - \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{k}. \quad (6.166)$$

Therefore,

$$p(t_{N+1} | \mathbf{t}) = \mathcal{N}(t_{N+1} | m, s), \quad (6.167)$$

where

$$m = \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{t}. \quad (6.168)$$

## 6.21

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(\mathbf{t} | \mathbf{y}) &= \mathcal{N}(\mathbf{t} | \mathbf{y}, \beta^{-1} \mathbf{I}), \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y} | \mathbf{0}, \boldsymbol{\Phi} \boldsymbol{\Phi}^\top). \end{aligned} \quad (6.169)$$

By 3.10,

$$p(t_{N+1} | \mathbf{t}) = \mathcal{N}(t_{N+1} | m, s), \quad (6.170)$$

where

$$\begin{aligned} m &= \beta \boldsymbol{\phi}_{N+1}^\top (\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{t}, \\ s &= \beta^{-1} + \boldsymbol{\phi}_{N+1}^\top (\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\phi}_{N+1}. \end{aligned} \quad (6.171)$$

By 2.26,

$$(\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} = \mathbf{I} - \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi}, \quad (6.172)$$

where

$$\mathbf{C} = \beta^{-1} \mathbf{I} + \boldsymbol{\Phi} \boldsymbol{\Phi}^\top. \quad (6.173)$$

Then,

$$m = \beta \boldsymbol{\phi}_{N+1}^\top (\mathbf{I} - \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi}) \boldsymbol{\Phi}^\top \mathbf{t}. \quad (6.174)$$

The right hand side can be written as

$$\beta (\boldsymbol{\phi}_{N+1}^\top \boldsymbol{\Phi}^\top - \boldsymbol{\phi}_{N+1}^\top \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi} \boldsymbol{\Phi}^\top) \mathbf{t} = \beta \boldsymbol{\phi}_{N+1}^\top \boldsymbol{\Phi}^\top \mathbf{C}^{-1} (\mathbf{C} - \boldsymbol{\Phi} \boldsymbol{\Phi}^\top) \mathbf{t}. \quad (6.175)$$

The right hand side can be written as

$$\beta \mathbf{k}^\top \mathbf{C}^{-1} \beta^{-1} \mathbf{I} \mathbf{t} = \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{t}, \quad (6.176)$$

where

$$\mathbf{k} = \boldsymbol{\Phi} \boldsymbol{\phi}_{N+1}. \quad (6.177)$$

Similarly,

$$s = \beta^{-1} + \boldsymbol{\phi}_{N+1}^\top (\mathbf{I} - \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi}) \boldsymbol{\phi}_{N+1}. \quad (6.178)$$

The right hand side can be written as

$$\beta^{-1} + \boldsymbol{\phi}_{N+1}^\top \boldsymbol{\phi}_{N+1} - \boldsymbol{\phi}_{N+1}^\top \boldsymbol{\Phi}^\top \mathbf{C}^{-1} \boldsymbol{\Phi} \boldsymbol{\phi}_{N+1} = c - \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{k}, \quad (6.179)$$

where

$$c = \beta^{-1} + \boldsymbol{\phi}_{N+1}^\top \boldsymbol{\phi}_{N+1}. \quad (6.180)$$

Therefore,

$$\begin{aligned} m &= \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{t}, \\ s &= c - \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{k}. \end{aligned} \quad (6.181)$$

## 6.22

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(\mathbf{t}|\mathbf{y}) &= \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}), \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}), \end{aligned} \quad (6.182)$$

where

$$K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}). \quad (6.183)$$

(a)

By the Bayes' theorem,

$$p(t_{N+1}, \dots, t_{N+L}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}'), \quad (6.184)$$

where

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \\ \vdots \\ t_{N+L} \end{bmatrix}. \quad (6.185)$$

By 6.20,

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}), \quad (6.186)$$

where

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{K}. \quad (6.187)$$

Let

$$\begin{aligned} K'_{nl} &= k(\mathbf{x}_n, \mathbf{x}_{N+l}), \\ \Gamma_{ll'} &= \beta^{-1} I_{ll'} + k(\mathbf{x}_{N+l}, \mathbf{x}_{N+l'}). \end{aligned} \quad (6.188)$$

Then,

$$p(\mathbf{t}') = \mathcal{N}(\mathbf{t}' | \mathbf{0}, \mathbf{C}'), \quad (6.189)$$

where

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{K}' \\ \mathbf{K}'^\top & \Gamma \end{bmatrix}. \quad (6.190)$$

By 2.24,

$$\begin{bmatrix} \Gamma & \mathbf{K}'^\top \\ \mathbf{K}' & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{K}'^\top\mathbf{C}^{-1} \\ -\mathbf{C}^{-1}\mathbf{K}'\mathbf{S}^{-1} & \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{K}'\mathbf{S}^{-1}\mathbf{K}'^\top\mathbf{C}^{-1} \end{bmatrix}, \quad (6.191)$$

where

$$\mathbf{S} = \Gamma - \mathbf{K}'^\top\mathbf{C}^{-1}\mathbf{K}'. \quad (6.192)$$

Therefore,

$$p(t_{N+1}, \dots, t_{N+L} | \mathbf{t}_N) = \mathcal{N}(t_{N+1}, \dots, t_{N+L} | \mathbf{m}, \mathbf{S}), \quad (6.193)$$

where

$$\mathbf{m} = \mathbf{K}'^\top\mathbf{C}^{-1}\mathbf{t}. \quad (6.194)$$

(b)

Let

$$\begin{aligned} k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= \beta^{-1} + k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}). \end{aligned} \quad (6.195)$$

By (a),

$$p(t_{N+1} | \mathbf{t}) = \mathcal{N}(t_{N+1} | m, s), \quad (6.196)$$

where

$$\begin{aligned} m &= \mathbf{k}^\top\mathbf{C}^{-1}\mathbf{t}, \\ s &= c - \mathbf{k}^\top\mathbf{C}^{-1}\mathbf{k}. \end{aligned} \quad (6.197)$$

## 6.23

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be variables such that

$$\begin{aligned} p(\mathbf{t}_n | \mathbf{y}_n) &= \mathcal{N}(\mathbf{t}_n | \mathbf{y}_n, \beta^{-1}\mathbf{I}), \\ p(\mathbf{Y}) &= \mathcal{N}(\mathbf{Y} | \mathbf{O}, \mathbf{K}), \end{aligned} \quad (6.198)$$

where

$$K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}). \quad (6.199)$$

By the Bayes' theorem,

$$p(\mathbf{t}_{N+1} | \mathbf{T})p(\mathbf{T}) = p(\mathbf{T}'). \quad (6.200)$$

By marginalisation,

$$p(\mathbf{T}) = \int p(\mathbf{T} | \mathbf{Y})p(\mathbf{Y})d\mathbf{Y}. \quad (6.201)$$

The logarithm of the integrand of the right hand side except the terms independent of  $\mathbf{T}$  and  $\mathbf{Y}$  can be written as

$$-\frac{\beta}{2} \sum_{n=1}^N \|\mathbf{t}_n - \mathbf{y}_n\|^2 - \frac{1}{2} \text{tr}(\mathbf{Y}^\top \mathbf{K}^{-1} \mathbf{Y}) = -\frac{1}{2} \text{tr} \mathbf{M}, \quad (6.202)$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{T} \end{bmatrix}^\top \begin{bmatrix} \beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I} \\ -\beta\mathbf{I} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ \mathbf{T} \end{bmatrix}. \quad (6.203)$$

By 2.24,

$$\begin{bmatrix} \beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I} \\ -\beta\mathbf{I} & \beta\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{K} & \mathbf{K} \\ \mathbf{K} & \beta^{-1}\mathbf{I} + \mathbf{K} \end{bmatrix}. \quad (6.204)$$

Then,

$$p(\mathbf{T}) = \mathcal{N}(\mathbf{T} | \mathbf{O}, \mathbf{C}), \quad (6.205)$$

where

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{K}. \quad (6.206)$$

Then,

$$p(\mathbf{T}') = \mathcal{N}(\mathbf{T}' | \mathbf{O}, \mathbf{C}'), \quad (6.207)$$

where

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\top & c \end{bmatrix}, \quad (6.208)$$

where

$$\begin{aligned} k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= \beta^{-1} + k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}). \end{aligned} \quad (6.209)$$

By 2.24,

$$\begin{bmatrix} c & \mathbf{k}^\top \\ \mathbf{k} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} s^{-1} & -s^{-1}\mathbf{k}^\top\mathbf{C}^{-1} \\ -s^{-1}\mathbf{C}^{-1}\mathbf{k} & \mathbf{C}^{-1} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\top\mathbf{C}^{-1} \end{bmatrix}, \quad (6.210)$$

where

$$s = c - \mathbf{k}^\top\mathbf{C}^{-1}\mathbf{k}. \quad (6.211)$$

Therefore,

$$p(\mathbf{t}_{N+1}|\mathbf{T}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{m}, s\mathbf{I}), \quad (6.212)$$

where

$$\mathbf{m} = \mathbf{T}^\top\mathbf{C}^{-1}\mathbf{k}. \quad (6.213)$$

## 6.24

(a)

Let  $\mathbf{M}$  be a  $D \times D$  diagonal matrix such that  $0 < M_{dd} < 1$ . For any vector  $\mathbf{v}$  in  $D$  dimensions,

$$\mathbf{v}^\top\mathbf{M}\mathbf{v} = \sum_{d=1}^D \sum_{d'=1}^D v_d M_{dd'} v_{d'}. \quad (6.214)$$

The right hand side can be written as

$$\sum_{d=1}^D M_{dd} v_d^2 > 0. \quad (6.215)$$

Therefore,  $\mathbf{M}$  is positive definite.

(b)

Let  $\mathbf{M}_1$  and  $\mathbf{M}_2$  be  $D \times D$  positive definite matrices. Then, for any vector  $\mathbf{v}$  in  $D$  dimensions,

$$\begin{aligned} \mathbf{v}^\top\mathbf{M}_1\mathbf{v} &> 0, \\ \mathbf{v}^\top\mathbf{M}_2\mathbf{v} &> 0. \end{aligned} \quad (6.216)$$

Then,

$$\mathbf{v}^\top(\mathbf{M}_1 + \mathbf{M}_2)\mathbf{v} > 0. \quad (6.217)$$

Therefore,  $\mathbf{M}_1 + \mathbf{M}_2$  is positive definite.

## 6.25

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n|a_n) &= \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n}, \\ p(\mathbf{a}) &= \mathcal{N}(\mathbf{a}|\mathbf{0}, \mathbf{C}), \end{aligned} \quad (6.218)$$

where

$$\begin{aligned} \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ C_{nn'} &= k(\mathbf{x}_n, \mathbf{x}_{n'}) + \nu I_{nn'}. \end{aligned} \quad (6.219)$$

(a)

By the Bayes' theorem,

$$p(\mathbf{a}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{a})p(\mathbf{a}). \quad (6.220)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{t}$  and  $\mathbf{a}$  can be written as

$$\Psi(\mathbf{a}) = \sum_{n=1}^N (t_n \ln \sigma(a_n) + (1 - t_n) \ln (1 - \sigma(a_n))) - \frac{1}{2} \mathbf{a}^\top \mathbf{C}^{-1} \mathbf{a}. \quad (6.221)$$

In order to maximise  $p(\mathbf{a}|\mathbf{t})$  with respect to  $\mathbf{a}$ ,  $\Psi$  needs to be maximised. We have

$$\frac{d}{da} \sigma(a) = \sigma(a) (1 - \sigma(a)). \quad (6.222)$$

Then,

$$(\nabla \Psi(\mathbf{a}))_n = (t_n (1 - \sigma(a_n)) - (1 - t_n) \sigma(a_n)) - (\mathbf{C}^{-1} \mathbf{a})_n, \quad (6.223)$$

so that

$$\nabla \Psi(\mathbf{a}) = \mathbf{t} - \boldsymbol{\sigma} - \mathbf{C}^{-1} \mathbf{a}, \quad (6.224)$$

where

$$\sigma_n = \sigma(a_n). \quad (6.225)$$

Since  $\boldsymbol{\sigma}$  non-linearly depends on  $\mathbf{a}$ , the Taylor series

$$\nabla \Psi(\mathbf{a}^{\text{new}}) = \nabla \Psi(\mathbf{a}) + \nabla \nabla \Psi(\mathbf{a}) (\mathbf{a}^{\text{new}} - \mathbf{a}) + O(\|\mathbf{a} - \mathbf{a}^{\text{new}}\|^2). \quad (6.226)$$

is used to iteratively find  $\mathbf{a}$  which maximises  $\Psi$ . Setting the left hand side to zero and neglecting the second order term of the right hand side gives

$$\mathbf{0} = \nabla\Psi(\mathbf{a}) + \nabla\nabla\Psi(\mathbf{a})(\mathbf{a}^{\text{new}} - \mathbf{a}). \quad (6.227)$$

Then,

$$\mathbf{a}^{\text{new}} = \mathbf{a} - (\nabla\nabla\Psi(\mathbf{a}))^{-1} \nabla\Psi(\mathbf{a}). \quad (6.228)$$

We have

$$\nabla\nabla\Psi(\mathbf{a}) = -\mathbf{W} - \mathbf{C}^{-1}, \quad (6.229)$$

where

$$W_{nn'} = \sigma(a_n)(1 - \sigma(a_n)) I_{nn'}. \quad (6.230)$$

Then, the right hand side can be written as

$$\mathbf{a}^{\text{new}} = \mathbf{a} + (\mathbf{W} + \mathbf{C}^{-1})^{-1} (\mathbf{t} - \boldsymbol{\sigma} - \mathbf{C}^{-1}\mathbf{a}). \quad (6.231)$$

The right hand side can be written as

$$\begin{aligned} & (\mathbf{W} + \mathbf{C}^{-1})^{-1} ((\mathbf{W} + \mathbf{C}^{-1})\mathbf{a} + \mathbf{t} - \boldsymbol{\sigma} - \mathbf{C}^{-1}\mathbf{a}) \\ &= \mathbf{C}(\mathbf{W}\mathbf{C} + \mathbf{I})^{-1}(\mathbf{W}\mathbf{a} + \mathbf{t} - \boldsymbol{\sigma}). \end{aligned} \quad (6.232)$$

Therefore,

$$\mathbf{a}^{\text{new}} = \mathbf{C}(\mathbf{W}\mathbf{C} + \mathbf{I})^{-1}(\mathbf{W}\mathbf{a} + \mathbf{t} - \boldsymbol{\sigma}). \quad (6.233)$$

(b)

Let  $\mathbf{a}_{\text{MAP}}$  be a stationary point of  $p(\mathbf{a}|\mathbf{t})$  which is found by the iterative process given in (a). We have the Taylor series

$$\begin{aligned} \Psi(\mathbf{a}) &= \Psi(\mathbf{a}_{\text{MAP}}) + \nabla\Psi(\mathbf{a}_{\text{MAP}})(\mathbf{a} - \mathbf{a}_{\text{MAP}}) \\ &+ \frac{1}{2}(\mathbf{a} - \mathbf{a}_{\text{MAP}})^T \nabla\nabla\Psi(\mathbf{a}_{\text{MAP}})(\mathbf{a} - \mathbf{a}_{\text{MAP}}) + O(\|\mathbf{a} - \mathbf{a}_{\text{MAP}}\|^3). \end{aligned} \quad (6.234)$$

Then,

$$\Psi(\mathbf{a}) \simeq \Psi(\mathbf{a}_{\text{MAP}}) - \frac{1}{2}(\mathbf{a} - \mathbf{a}_{\text{MAP}})^T \mathbf{H}(\mathbf{a} - \mathbf{a}_{\text{MAP}}), \quad (6.235)$$

where

$$\mathbf{H} = \mathbf{W} |_{\mathbf{a}=\mathbf{a}_{\text{MAP}}} + \mathbf{C}^{-1}. \quad (6.236)$$

Therefore,

$$p(\mathbf{a}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{a}|\mathbf{a}_{\text{MAP}}, \mathbf{H}^{-1}). \quad (6.237)$$

(c)

By (a),

$$\mathbf{a}_{\text{MAP}} = \mathbf{C} (\mathbf{W}\mathbf{C} + \mathbf{I})^{-1} (\mathbf{W}\mathbf{a}_{\text{MAP}} + \mathbf{t} - \boldsymbol{\sigma}). \quad (6.238)$$

Then,

$$(\mathbf{W}\mathbf{C} + \mathbf{I}) \mathbf{C}^{-1} \mathbf{a}_{\text{MAP}} = \mathbf{W}\mathbf{a}_{\text{MAP}} + \mathbf{t} - \boldsymbol{\sigma}. \quad (6.239)$$

Therefore,

$$\mathbf{a}_{\text{MAP}} = \mathbf{C}(\mathbf{t} - \boldsymbol{\sigma}). \quad (6.240)$$

## 6.26

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | a_n) &= \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n}, \\ p(\mathbf{a}) &= \mathcal{N}(\mathbf{a} | \mathbf{0}, \mathbf{C}), \end{aligned} \quad (6.241)$$

where

$$\begin{aligned} \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ C_{nn'} &= k(\mathbf{x}_n, \mathbf{x}_{n'}) + \nu I_{nn'}. \end{aligned} \quad (6.242)$$

By marginalisation,

$$p(a_{N+1} | \mathbf{t}) = \int p(a_{N+1} | \mathbf{a}) p(\mathbf{a} | \mathbf{t}) d\mathbf{a}. \quad (6.243)$$

Let

$$\begin{aligned} \mathbf{a}' &= \begin{bmatrix} \mathbf{a} \\ a_{N+1} \end{bmatrix}, \\ k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \nu. \end{aligned} \quad (6.244)$$

Then,

$$p(\mathbf{a}') = \mathcal{N}(\mathbf{a}' | \mathbf{0}, \mathbf{C}'), \quad (6.245)$$

where

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\top & c \end{bmatrix}. \quad (6.246)$$

By 2.24,

$$\begin{bmatrix} c & \mathbf{k}^\top \\ \mathbf{k} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} s^{-1} & -s^{-1}\mathbf{k}^\top \mathbf{C}^{-1} \\ -s^{-1}\mathbf{C}^{-1}\mathbf{k} & \mathbf{C}^{-1} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\top \mathbf{C}^{-1} \end{bmatrix}, \quad (6.247)$$

where

$$s = c - \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{k}. \quad (6.248)$$

Then,

$$p(a_{N+1}|\mathbf{a}) = \mathcal{N}(a_{N+1}|m, s), \quad (6.249)$$

where

$$m = \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{a}. \quad (6.250)$$

By 6.25(b),

$$p(\mathbf{a}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{a}|\mathbf{a}_{\text{MAP}}, \mathbf{H}^{-1}), \quad (6.251)$$

where  $\mathbf{a}_{\text{MAP}}$  is a stationary point of  $p(\mathbf{a}|\mathbf{t})$  and

$$\begin{aligned} \mathbf{H} &= \mathbf{W} |_{\mathbf{a}=\mathbf{a}_{\text{MAP}}} + \mathbf{C}^{-1}, \\ W_{nn'} &= \sigma(a_n)(1 - \sigma(a_n)) I_{nn'}. \end{aligned} \quad (6.252)$$

Then, the logarithm of the integrand except the terms independent of  $\mathbf{t}$  and  $\mathbf{a}$  can be approximated as

$$\begin{aligned} &- \frac{1}{2s} (a_{N+1} - m)^2 - \frac{1}{2} (\mathbf{a} - \mathbf{a}_{\text{MAP}})^\top \mathbf{H} (\mathbf{a} - \mathbf{a}_{\text{MAP}}) \\ &= -\frac{1}{2} \mathbf{a}'^\top \mathbf{M} \mathbf{a}' + \mathbf{a}'^\top \mathbf{v} - \frac{1}{2} \mathbf{a}_{\text{MAP}}^\top \mathbf{H} \mathbf{a}_{\text{MAP}}. \end{aligned} \quad (6.253)$$

where

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} \mathbf{H} + s^{-1} \mathbf{C}^{-1} \mathbf{k} \mathbf{k}^\top \mathbf{C}^{-1} & -s^{-1} \mathbf{C}^{-1} \mathbf{k} \\ -s^{-1} \mathbf{k}^\top \mathbf{C}^{-1} & s^{-1} \end{bmatrix}, \\ \mathbf{v} &= \begin{bmatrix} \mathbf{H} \mathbf{a}_{\text{MAP}} \\ 0 \end{bmatrix}. \end{aligned} \quad (6.254)$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{H}^{-1} & \mathbf{H}^{-1} \mathbf{C}^{-1} \mathbf{k} \\ \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{H}^{-1} & s + \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{H}^{-1} \mathbf{C}^{-1} \mathbf{k} \end{bmatrix}, \quad (6.255)$$

so that

$$\mathbf{M}^{-1} \mathbf{v} = \begin{bmatrix} \mathbf{a}_{\text{MAP}} \\ \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{a}_{\text{MAP}} \end{bmatrix}. \quad (6.256)$$

Then,

$$p(a_{N+1}|\mathbf{t}) \simeq \mathcal{N}(a_{N+1}|\mu, \sigma^2), \quad (6.257)$$

where

$$\begin{aligned} \mu &= \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{a}_{\text{MAP}}, \\ \sigma^2 &= s + \mathbf{k}^\top \mathbf{C}^{-1} \mathbf{H}^{-1} \mathbf{C}^{-1} \mathbf{k}. \end{aligned} \quad (6.258)$$

By 6.25(c),

$$\mathbf{a}_{\text{MAP}} = \mathbf{C}(\mathbf{t} - \boldsymbol{\sigma}), \quad (6.259)$$

where

$$\sigma_n = \sigma(a_n). \quad (6.260)$$

Therefore,

$$\begin{aligned} \mu &= \mathbf{k}^\top(\mathbf{t} - \boldsymbol{\sigma}), \\ \sigma^2 &= c - \mathbf{k}^\top \mathbf{C}^{-1} (\mathbf{C} - \mathbf{H}^{-1}) \mathbf{C}^{-1} \mathbf{k}. \end{aligned} \quad (6.261)$$

## 6.27 (Incomplete)

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | a_n) &= \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n}, \\ p(\mathbf{a} | \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{a} | \mathbf{0}, \mathbf{C}), \end{aligned} \quad (6.262)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \quad (6.263)$$

and  $\mathbf{C}$  is dependent on  $\boldsymbol{\theta}$  with  $M$  dimensions.

### (a)

By marginalisation,

$$p(\mathbf{t} | \boldsymbol{\theta}) = \int p(\mathbf{t} | \mathbf{a}) p(\mathbf{a} | \boldsymbol{\theta}) d\mathbf{a}. \quad (6.264)$$

Let  $\mathbf{a}_{\text{MAP}}$  be a stationary point of  $p(\mathbf{a} | \mathbf{t})$  and

$$\Psi(\mathbf{a}) = \ln p(\mathbf{t} | \mathbf{a}) + \ln p(\mathbf{a} | \boldsymbol{\theta}). \quad (6.265)$$

By 6.25(b),

$$\Psi(\mathbf{a}) \simeq \Psi(\mathbf{a}_{\text{MAP}}) - \frac{1}{2} (\mathbf{a} - \mathbf{a}_{\text{MAP}})^\top \mathbf{H} (\mathbf{a} - \mathbf{a}_{\text{MAP}}), \quad (6.266)$$

where

$$\begin{aligned} \mathbf{H} &= \mathbf{W} |_{\mathbf{a}=\mathbf{a}_{\text{MAP}}} + \mathbf{C}^{-1}, \\ W_{nn'} &= \sigma(a_n) (1 - \sigma(a_n)) I_{nn'}. \end{aligned} \quad (6.267)$$

Therefore,

$$\ln p(\mathbf{t} | \boldsymbol{\theta}) \simeq \Psi(\mathbf{a}_{\text{MAP}}) + \frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{H}). \quad (6.268)$$

(b)

We have

$$\frac{\partial \Psi(\mathbf{a}_{\text{MAP}})}{\partial \theta_m} = \sum_{n=1}^N \sum_{n'=1}^N \frac{\partial \Psi(\mathbf{a}_{\text{MAP}})}{\partial C_{nn'}} \frac{\partial C_{nn'}}{\partial \theta_m}. \quad (6.269)$$

The right hand side can be written as

$$\frac{1}{2} \text{tr} \left( \mathbf{a}_{\text{MAP}} \mathbf{a}_{\text{MAP}}^\top (\mathbf{C}^{-1})^2 \frac{\partial \mathbf{C}}{\partial \theta_m} \right). \quad (6.270)$$

We have

$$\frac{\partial \ln(\det \mathbf{H})}{\partial \theta_m} = \sum_{n=1}^N \sum_{n'=1}^N \frac{\partial \ln(\det \mathbf{H})}{\partial H_{nn'}} \frac{\partial H_{nn'}}{\partial \theta_m}. \quad (6.271)$$

By 3.21(a), the right hand side can be written as

$$\text{tr} \left( \mathbf{H}^{-1} \frac{\partial \mathbf{C}^{-1}}{\partial \theta_m} \right). \quad (6.272)$$

Therefore,

$$\frac{\partial \ln p(\mathbf{t}|\boldsymbol{\theta})}{\partial \theta_m} \simeq \frac{1}{2} \text{tr} \left( \mathbf{a}_{\text{MAP}} \mathbf{a}_{\text{MAP}}^\top (\mathbf{C}^{-1})^2 \frac{\partial \mathbf{C}}{\partial \theta_m} - \mathbf{H}^{-1} \frac{\partial \mathbf{C}^{-1}}{\partial \theta_m} \right). \quad (6.273)$$

## 7 Sparse Kernel Machines

### 7.1 (Incomplete)

#### 7.2

Let  $t_1, \dots, t_N$  be variables. In order to minimise  $\|\mathbf{w}\|^2$  under the constraint

$$t_n (\mathbf{w}^\top \phi(\mathbf{x}_n) + b) \geq \gamma, \quad (7.1)$$

for  $\gamma > 0$ , let

$$L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N a_n (t_n (\mathbf{w}^\top \phi(\mathbf{x}_n) + b) - \gamma). \quad (7.2)$$

Setting the derivatives of  $L$  with respect to  $\mathbf{w}$  and  $b$  to zero gives

$$\begin{aligned} \mathbf{0} &= \mathbf{w} - \sum_{n=1}^N a_n t_n \phi(\mathbf{x}_n), \\ 0 &= - \sum_{n=1}^N a_n t_n. \end{aligned} \quad (7.3)$$

Therefore,

$$\begin{aligned} \mathbf{w} &= \sum_{n=1}^N a_n t_n \phi(\mathbf{x}_n), \\ 0 &= \sum_{n=1}^N a_n t_n, \end{aligned} \quad (7.4)$$

which is independent of  $\gamma$ .

### 7.3 (Incomplete)

### 7.4 (Incomplete)

### 7.5 (Incomplete)

### 7.6

Let  $t_1, \dots, t_N$  be a variable such that

$$\begin{aligned} t_n &\in \{-1, 1\}, \\ p(t_n|y_n) &= \sigma(y_n)^{\frac{1+t_n}{2}} (1 - \sigma(y_n))^{\frac{1-t_n}{2}}, \end{aligned} \quad (7.5)$$

where

$$\begin{aligned} \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ y_n &= \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n) + b. \end{aligned} \quad (7.6)$$

We have

$$1 - \sigma(a) = \sigma(-a). \quad (7.7)$$

Then,

$$p(t_n|y_n) = \sigma(y_n t_n). \quad (7.8)$$

Then,

$$p(\mathbf{t}|\mathbf{y}) = \prod_{n=1}^N \sigma(y_n t_n), \quad (7.9)$$

so that

$$-\ln p(\mathbf{t}|\mathbf{y}) = -\sum_{n=1}^N \ln \sigma(y_n t_n). \quad (7.10)$$

Therefore,

$$-\ln p(\mathbf{t}|\mathbf{y}) = \sum_{n=1}^N \ln (1 + \exp(-y_n t_n)). \quad (7.11)$$

### 7.7

Let  $t_1, \dots, t_N$  be variables. Let

$$E = C \sum_{n=1}^N E_\epsilon(y_n - t_n) + \frac{1}{2} \|\mathbf{w}\|^2, \quad (7.12)$$

where  $\epsilon > 0$  and

$$E_\epsilon(a) = \begin{cases} 0, & |a| < \epsilon, \\ |a| - \epsilon, & \text{otherwise,} \end{cases} \quad (7.13)$$

$$y_n = \mathbf{w}^\top \phi(\mathbf{x}_n) + b.$$

In order to minimise  $E$  under the constraints

$$\begin{aligned} t_n &\leq y_n + \epsilon + \xi_n, \\ t_n &\geq y_n - \epsilon - \hat{\xi}_n, \end{aligned} \quad (7.14)$$

where  $\xi_n \geq 0$  and  $\hat{\xi}_n \geq 0$ , let us minimise

$$\begin{aligned} L = & C \sum_{n=1}^N (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) \\ & - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n). \end{aligned} \quad (7.15)$$

Setting the derivatives of  $L$  with respect to  $\mathbf{w}$ ,  $b$ ,  $\xi_n$  and  $\hat{\xi}_n$  to zero gives

$$\begin{aligned} \mathbf{0} &= \mathbf{w} - \sum_{n=1}^N a_n \phi(\mathbf{x}_n) + \sum_{n=1}^N \hat{a}_n \phi(\mathbf{x}_n), \\ 0 &= -\sum_{n=1}^N a_n + \sum_{n=1}^N \hat{a}_n, \\ 0 &= C - \mu_n - a_n, \\ 0 &= C - \hat{\mu}_n - \hat{a}_n. \end{aligned} \quad (7.16)$$

Then,  $L$  can be written as

$$\begin{aligned} & \sum_{n=1}^N ((\mu_n + a_n) \xi_n + (\hat{\mu}_n + \hat{a}_n) \hat{\xi}_n) + \frac{1}{2} \left\| \sum_{n=1}^N (\hat{a}_n - a_n) \phi(\mathbf{x}_n) \right\|^2 \\ & - \sum_{n=1}^N (\mu_n \xi_n + \hat{\mu}_n \hat{\xi}_n) - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n) \\ & - \sum_{n=1}^N \hat{a}_n (\epsilon + \hat{\xi}_n - y_n + t_n). \end{aligned} \quad (7.17)$$

Therefore, minimising  $L$  is equivalent to maximising

$$\begin{aligned}\tilde{L} = & -\frac{1}{2} \sum_{n=1}^N \sum_{n'=1}^N (\hat{a}_n - a_n)(\hat{a}_{n'} - a_{n'}) k(\mathbf{x}_n, \mathbf{x}_{n'}) \\ & + \epsilon \sum_{n=1}^N (a_n + \hat{a}_n) + \sum_{n=1}^N (a_n - \hat{a}_n)(y_n - t_n),\end{aligned}\tag{7.18}$$

where

$$k(\mathbf{x}_n, \mathbf{x}_{n'}) = \boldsymbol{\phi}(\mathbf{x}_n)^\top \boldsymbol{\phi}(\mathbf{x}_{n'}).\tag{7.19}$$

## 7.8 (Incomplete)

Let  $t_1, \dots, t_N$  be variables. Let

$$E = C \sum_{n=1}^N E_\epsilon(y_n - t_n) + \frac{1}{2} \|\mathbf{w}\|^2,\tag{7.20}$$

where  $\epsilon > 0$  and

$$\begin{aligned}E_\epsilon(a) &= \begin{cases} 0, & |a| < \epsilon, \\ |a| - \epsilon, & \text{otherwise,} \end{cases} \\ y_n &= \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n) + b.\end{aligned}\tag{7.21}$$

(a)

In order to minimise  $E$  under the constraints

$$t_n \leq y_n + \epsilon + \xi_n,\tag{7.22}$$

where  $\xi_n \geq 0$ , let us minimise

$$L = C \sum_{n=1}^N \xi_n + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N \mu_n \xi_n - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n).\tag{7.23}$$

Setting the derivatives of  $L$  with respect to  $\mathbf{w}$ ,  $b$  and  $\xi_n$  to zero gives

$$\begin{aligned}\mathbf{0} &= \mathbf{w} - \sum_{n=1}^N a_n \boldsymbol{\phi}(\mathbf{x}_n), \\ 0 &= -\sum_{n=1}^N a_n, \\ 0 &= C - \mu_n - a_n.\end{aligned}\tag{7.24}$$

Then,  $L$  can be written as

$$\begin{aligned} & \sum_{n=1}^N (\mu_n + a_n) \xi_n + \frac{1}{2} \left\| \sum_{n=1}^N a_n \boldsymbol{\phi}(\mathbf{x}_n) \right\|^2 - \sum_{n=1}^N \mu_n \xi_n \\ & - \sum_{n=1}^N a_n (\epsilon + \xi_n + y_n - t_n). \end{aligned} \quad (7.25)$$

Therefore, minimising  $L$  is equivalent to maximising

$$\tilde{L} = -\frac{1}{2} \sum_{n=1}^N \sum_{n'=1}^N a_n a_{n'} k(\mathbf{x}_n, \mathbf{x}_{n'}) + \sum_{n=1}^N a_n (y_n - t_n), \quad (7.26)$$

where

$$k(\mathbf{x}_n, \mathbf{x}_{n'}) = \boldsymbol{\phi}(\mathbf{x}_n)^T \boldsymbol{\phi}(\mathbf{x}_{n'}). \quad (7.27)$$

## 7.9

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^T \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{A}^{-1}). \end{aligned} \quad (7.28)$$

By the Bayes' theorem,

$$p(\mathbf{w} | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}). \quad (7.29)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{t}$  and  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{\beta}{2} \sum_{n=1}^N \|t_n - \mathbf{w}^T \boldsymbol{\phi}_n\|^2 - \frac{1}{2} \mathbf{w}^T \mathbf{A} \mathbf{w} \\ & = -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^T \begin{bmatrix} \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \mathbf{A} & -\beta \boldsymbol{\Phi}^T \\ -\beta \boldsymbol{\Phi} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}. \end{aligned} \quad (7.30)$$

Therefore,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}, \mathbf{S}), \quad (7.31)$$

where

$$\begin{aligned} \mathbf{m} &= \beta \mathbf{S} \boldsymbol{\Phi}^T \mathbf{t}, \\ \mathbf{S} &= (\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \mathbf{A})^{-1}. \end{aligned} \quad (7.32)$$

## 7.10

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}_n, \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{A}^{-1}). \end{aligned} \quad (7.33)$$

By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}. \quad (7.34)$$

The logarithm of the integrand of the right hand side except the terms independent of  $\mathbf{t}$  and  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{\beta}{2} \sum_{n=1}^N \|t_n - \mathbf{w}^\top \boldsymbol{\phi}_n\|^2 - \frac{1}{2} \mathbf{w}^\top \mathbf{A} \mathbf{w} \\ &= -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \mathbf{A} & -\beta \boldsymbol{\Phi}^\top \\ -\beta \boldsymbol{\Phi} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}. \end{aligned} \quad (7.35)$$

By 2.24,

$$\begin{bmatrix} \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \mathbf{A} & -\beta \boldsymbol{\Phi}^\top \\ -\beta \boldsymbol{\Phi} & \beta \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1} \boldsymbol{\Phi}^\top \\ \boldsymbol{\Phi} \mathbf{A}^{-1} & \beta^{-1} \mathbf{I} + \boldsymbol{\Phi} \mathbf{A}^{-1} \boldsymbol{\Phi}^\top \end{bmatrix}. \quad (7.36)$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t} | \mathbf{0}, \beta^{-1} \mathbf{I} + \boldsymbol{\Phi} \mathbf{A}^{-1} \boldsymbol{\Phi}^\top). \quad (7.37)$$