

# Solutions Manual to Pattern Recognition and Machine Learning

Hiromichi Inawashiro

November 24, 2024

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Probability Distributions</b>	<b>39</b>
<b>3</b>	<b>Linear Models for Regression</b>	<b>93</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 to zero gives

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

Substituting

$$y(x_n, \mathbf{w}) = \sum_{j=0}^M w_j x_n^j \quad (1.3)$$

gives

$$0 = \sum_{n=1}^N x_n^i \left( \sum_{j=0}^M w_j x_n^j - t_n \right). \quad (1.4)$$

Therefore,

$$\operatorname{argmin}_{w_j} E(\mathbf{w}) = \left\{ w_j \mid \sum_{j=0}^M A_{ij} w_j = T_i \right\}, \quad (1.5)$$

where

$$\begin{aligned} A_{ij} &= \sum_{n=1}^N x_n^{i+j}, \\ T_i &= \sum_{n=1}^N x_n^i t_n. \end{aligned} \quad (1.6)$$

## 1.2

Let

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2. \quad (1.7)$$

Setting the derivative to zero gives

$$\mathbf{0} = \sum_{n=1}^N \frac{\partial y(x_n, \mathbf{w})}{\partial \mathbf{w}} (y(x_n, \mathbf{w}) - t_n) + \lambda \mathbf{w}. \quad (1.8)$$

Substituting

$$y(x_n, \mathbf{w}) = \sum_{j=0}^M w_j x_n^j \quad (1.9)$$

gives

$$0 = \sum_{n=1}^N x_n^i \left( \sum_{j=0}^M w_j x_n^j - t_n \right) + \lambda w_i. \quad (1.10)$$

Therefore,

$$\operatorname{argmin}_{w_j} E(\mathbf{w}) = \left\{ w_j \mid \sum_{j=0}^M \tilde{A}_{ij} w_j = T_i \right\}, \quad (1.11)$$

where

$$\begin{aligned} \tilde{A}_{ij} &= \sum_{n=1}^N x_n^{i+j} + \lambda \delta_{ij}, \\ T_i &= \sum_{n=1}^N x_n^i t_n. \end{aligned} \quad (1.12)$$

### 1.3

Let  $a$ ,  $o$  and  $l$  be the events where an apple, orange and lime are selected respectively. The probability that an apple is selected is given by

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \quad (1.13)$$

Substituting  $p(a|r) = \frac{3}{10}$ ,  $p(r) = \frac{1}{5}$ ,  $p(a|g) = \frac{1}{2}$ ,  $p(r) = \frac{1}{5}$ ,  $p(a|g) = \frac{3}{10}$  and  $p(g) = \frac{3}{5}$  gives

$$p(a) = \frac{17}{50}. \quad (1.14)$$

If an orange is selected, the probability that it came from the green box is given by

$$p(g|o) = \frac{p(g, o)}{p(o)}. \quad (1.15)$$

Here,

$$\begin{aligned} p(g, o) &= p(o|g)p(g), \\ p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g). \end{aligned} \quad (1.16)$$

Substituting  $p(o|r) = \frac{2}{5}$ ,  $p(r) = \frac{1}{5}$ ,  $p(o|b) = \frac{1}{2}$ ,  $p(b) = \frac{1}{5}$ ,  $p(o|g) = \frac{3}{10}$  and  $p(g) = \frac{3}{5}$  gives  $p(g, o) = \frac{9}{50}$  and  $p(o) = \frac{9}{25}$ . Therefore,

$$p(g|o) = \frac{1}{2}. \quad (1.17)$$

## 1.4

Let

$$x = g(y) \quad (1.18)$$

and  $\hat{x}$  and  $\hat{y}$  be the locations of the maximum of  $p_x(x)$  and  $p_y(y)$  respectively. Let us assume that there exists  $\epsilon > 0$  such that  $g'(y) \neq 0$  for  $|y - \hat{y}| < \epsilon$ . Then, Taking the derivative of the transformation

$$p_y(y) = p_x(g(y)) |g'(y)| \quad (1.19)$$

and substituting  $y = \hat{y}$  gives

$$0 = g'(\hat{y})p'_x(g(\hat{y})) + p_x(g(\hat{y}))g''(\hat{y}). \quad (1.20)$$

Therefore, in general,

$$\hat{x} \neq g(\hat{y}). \quad (1.21)$$

Here, let us assume that

$$g(y) = ay + b. \quad (1.22)$$

Then, Taking the derivative of the transformation and substituting  $y = \hat{y}$  gives

$$0 = p'_x(g(\hat{y})). \quad (1.23)$$

Therefore,

$$\hat{x} = g(\hat{y}). \quad (1.24)$$

## 1.5

By the definition,

$$\text{var } f(x) = E(f(x) - E f(x))^2. \quad (1.25)$$

The right hand side can be written as

$$E((f(x))^2 - 2f(x) E f(x) + (E f(x))^2) = E(f(x))^2 - (E f(x))^2. \quad (1.26)$$

Therefore,

$$\text{var } f(x) = E(f(x))^2 - (E f(x))^2. \quad (1.27)$$

## 1.6

By the definition,

$$\text{cov}(x, y) = E((x - E x)(y - E y)). \quad (1.28)$$

The right hand side can be written as

$$E xy - E(x E y) - E(y E x) + E(E x E y) = E xy - E x E y. \quad (1.29)$$

The right hand side can be written as

$$\int xyp(x, y)dxdy - \int xp(x)dx \int yp(y)dy. \quad (1.30)$$

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

$$f(x, y) = f(x)f(y). \quad (1.31)$$

Then,

$$\int xyp(x, y)dxdy = \int p(x)dx \int p(y)dy. \quad (1.32)$$

Therefore,

$$\text{cov}(x, y) = 0. \quad (1.33)$$

## 1.7

Let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx. \quad (1.34)$$

Then

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2 + y^2)\right) dx dy. \quad (1.35)$$

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

$$\int_0^{\infty} \int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} dr d\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr. \quad (1.36)$$

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

$$2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{1}{2}s^2\right) s ds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_0^{\infty}. \quad (1.37)$$

Therefore,

$$I = (2\pi\sigma^2)^{\frac{1}{2}}. \quad (1.38)$$

By the definition,

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.39)$$

Then

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx. \quad (1.40)$$

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

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}t^2\right) dt = (2\pi\sigma^2)^{-\frac{1}{2}} I. \quad (1.41)$$

Therefore,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1. \quad (1.42)$$

## 1.8

Let  $x$  be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.43)$$

Then

$$\mathbb{E} x = \int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.44)$$

By the definition, the right hand side can be written as

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \quad (1.45)$$

By the transformation  $y = x - \mu$ , it can be written as

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (y + \mu) \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy. \quad (1.46)$$

Since

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = 0, \quad (1.47)$$

and

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mu \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = \mu \int_{-\infty}^{\infty} \mathcal{N}(y|\mu, \sigma^2) dy, \quad (1.48)$$

we have

$$\mathbb{E} x = \mu. \quad (1.49)$$

By the definition,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1 \quad (1.50)$$

can be written as

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 1. \quad (1.51)$$

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

$$\begin{aligned} & (2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) (\sigma^2)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ & + (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} (\sigma^2)^{-2} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 0. \end{aligned} \quad (1.52)$$



The left hand side can be written as

$$\begin{aligned} & -\frac{1}{2}(\sigma^2)^{-1} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2}(\sigma^2)^{-2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx \\ & = -\frac{1}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2} \text{var } x. \end{aligned} \quad (1.53)$$

Therefore,

$$\text{var } x = \sigma^2. \quad (1.54)$$

## 1.9

Let

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.55)$$

Setting its derivative with respect to  $x$  to zero gives

$$0 = (2\pi\sigma^2)^{-\frac{1}{2}} \left(-\frac{1}{\sigma^2}(x - \mu)\right) \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.56)$$

Therefore, the mode is given by  $\mu$ .

Similarly, let

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1.57)$$

Setting its derivative with respect to  $\mathbf{x}$  to zero gives

$$\mathbf{0} = -(2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x} - \boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1.58)$$

Therefore, the mode is given by  $\boldsymbol{\mu}$ .

## 1.10

By the definition,

$$\mathbb{E}(x + y) = \int \int (x + y)p(x, y)dx dy. \quad (1.59)$$

The right hand side can be written as

$$\int x \left( \int p(x, y) dy \right) dx + \int y \left( \int p(x, y) dx \right) dy = \int xp(x)dx + \int yp(y)dy. \quad (1.60)$$

By the definition, the right hand side can be written as

$$E x + E y. \quad (1.61)$$

Therefore,

$$E(x + y) = E x + E y. \quad (1.62)$$

Similarly, by the definition,

$$\text{var}(x + y) = E (x + y - E(x + y))^2 \quad (1.63)$$

By the result above and the definition, the right hand side can be written as

$$\begin{aligned} & E (x - E x)^2 + 2 E ((x - E x) (y - E y)) + E (y - E y)^2 \\ &= \text{var } x + 2 \text{cov}(x, y) + \text{var } y. \end{aligned} \quad (1.64)$$

If  $x$  and  $y$  are independent, then

$$\text{cov}(x, y) = 0, \quad (1.65)$$

by 1.6. Therefore,

$$\text{var}(x + y) = \text{var } x + \text{var } y. \quad (1.66)$$

## 1.11

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

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (1.67)$$

Then

$$\ln p(\mathbf{x}) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2. \quad (1.68)$$

To maximise it with respect to  $\mu$  and  $\sigma^2$ , setting the partial derivatives to zero gives

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu), \\ 0 &= -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{n=1}^N (x_n - \mu)^2. \end{aligned} \quad (1.69)$$

Therefore,

$$\begin{aligned}\mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N x_n, \\ \sigma_{\text{ML}}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2.\end{aligned}\tag{1.70}$$

## 1.12

Let  $x_m$  and  $x_n$  be independent variables. Then

$$\mathbb{E} x_m x_n = \mathbb{E} x_m \mathbb{E} x_n.\tag{1.71}$$

If they are samples from the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ , the right hand side is given by  $\mu^2$ . On the other hand, by the definition,

$$\mathbb{E} x_n^2 = \text{var } x_n + (\mathbb{E} x_n)^2.\tag{1.72}$$

If  $x_n$  is a sample from the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ , the right hand side is given by  $\sigma^2 + \mu^2$ . Therefore,

$$\mathbb{E} x_m x_n = \mu^2 + \delta_{mn} \sigma^2.\tag{1.73}$$

Here, since

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n,\tag{1.74}$$

we have

$$\mathbb{E} \mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n.\tag{1.75}$$

Therefore,

$$\mathbb{E} \mu_{\text{ML}} = \mu.\tag{1.76}$$

Similarly, since

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2,\tag{1.77}$$

we have

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n - \mu_{\text{ML}})^2.\tag{1.78}$$

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n^2 - 2\mu_{\text{ML}} x_n + \mu_{\text{ML}}^2) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n^2 - \frac{2}{N} \mathbb{E} \left( \mu_{\text{ML}} \left( \sum_{n=1}^N x_n \right) \right) + \mathbb{E} \mu_{\text{ML}}^2. \quad (1.79)$$

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

$$\frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) = \mu^2 + \sigma^2, \quad (1.80)$$

while the second and third terms can be written as

$$-2 \mathbb{E} \mu_{\text{ML}}^2 + \mathbb{E} \mu_{\text{ML}}^2 = -\mathbb{E} \mu_{\text{ML}}^2. \quad (1.81)$$

Here,

$$\mathbb{E} \mu_{\text{ML}}^2 = \mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N x_n \right)^2. \quad (1.82)$$

The right hand side can be written as

$$\frac{1}{N^2} \sum_{n=1}^N \mathbb{E} x_n^2 + \frac{2}{N^2} \sum_{1 \leq m < n \leq N} \mathbb{E} x_m x_n = \frac{1}{N} (\mu^2 + \sigma^2) + \frac{N-1}{N} \mu^2. \quad (1.83)$$

Therefore,

$$\mathbb{E} \mu_{\text{ML}}^2 = \mu^2 + \frac{1}{N} \sigma^2. \quad (1.84)$$

Thus,

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{N-1}{N} \sigma^2. \quad (1.85)$$

### 1.13

Let  $x_1, \dots, x_N$  be a set of variables whose mean is  $\mu$  and variance is  $\sigma^2$ . Then

$$\mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n - \mu)^2. \quad (1.86)$$

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n^2 - 2\mu x_n + \mu^2) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n^2 - \frac{2\mu}{N} \sum_{n=1}^N \mathbb{E} x_n + \mu^2. \quad (1.87)$$

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

$$\frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) = \mu^2 + \sigma^2, \quad (1.88)$$

while the second term can be written as

$$-\frac{2\mu}{N} \sum_{n=1}^N \mu = -2\mu^2. \quad (1.89)$$

Therefore,

$$\mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \sigma^2. \quad (1.90)$$

## 1.14

Let

$$\begin{aligned} w_{ij}^S &= \frac{1}{2}(w_{ij} + w_{ji}), \\ w_{ij}^A &= \frac{1}{2}(w_{ij} - w_{ji}). \end{aligned} \quad (1.91)$$

Then

$$\begin{aligned} w_{ij} &= w_{ij}^S + w_{ij}^A, \\ w_{ij}^S &= w_{ji}^S, \\ w_{ij}^A &= -w_{ji}^A. \end{aligned} \quad (1.92)$$

Here,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = \frac{1}{2} \sum_{i=1}^D \sum_{j=1}^D (w_{ij} - w_{ji}) x_i x_j. \quad (1.93)$$

The right hand side can be written as

$$\frac{1}{2} \left( \sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j - \sum_{i=1}^D \sum_{j=1}^D w_{ji} x_i x_j \right) = 0. \quad (1.94)$$

Therefore,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = 0. \quad (1.95)$$

Additionally,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = \sum_{i=1}^D \sum_{j=1}^D (w_{ij}^S + w_{ij}^A) x_i x_j. \quad (1.96)$$

The right hand side can be written as

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j + \sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = \sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j, \quad (1.97)$$

where the result above is used. Therefore,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = \sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j. \quad (1.98)$$

Finally, since the matrix  $w_{ij}^S$  is a  $D \times D$  symmetric matrix, its number of independent parameters is  $\frac{D(D+1)}{2}$ .

## 1.15

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

$$n(1, M) = n(1, M - 1) = 1. \quad (1.99)$$

Let us assume that

$$n(D, M) = \sum_{i=1}^D n(i, M - 1). \quad (1.100)$$

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. Therefore,

$$n(D + 1, M) = n(D, M) + n(D + 1, M - 1). \quad (1.101)$$

Thus,

$$n(D + 1, M) = \sum_{i=1}^{D+1} n(i, M - 1). \quad (1.102)$$

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

Additionally,

$$\sum_{i=1}^1 \frac{(i+M-2)!}{(i-1)!(M-1)!} = 1. \quad (1.103)$$

Let us assume that

$$\sum_{i=1}^D \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}. \quad (1.104)$$

Then

$$\sum_{i=1}^{D+1} \frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!}. \quad (1.105)$$

The right hand side can be written as

$$\frac{D(D+M-1)! + M(D+M-1)!}{D!M!} = \frac{(D+M)!}{D!M!}. \quad (1.106)$$

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

Finally, by 1.14,

$$n(D, 2) = \frac{D(D+1)}{2}. \quad (1.107)$$

Let us assume that

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!}. \quad (1.108)$$

Then, by the result above,

$$n(D, M+1) = \sum_{i=1}^D n(i, M). \quad (1.109)$$

By the assumption and result above, the right hand side can be written as

$$\sum_{i=1}^D \frac{(i+M-1)!}{(i-1)!M!} = \frac{(D+M)!}{(D-1)!(M+1)!}. \quad (1.110)$$

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

## 1.16

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. Then, by 1.15,

$$N(D, M) = \sum_{m=0}^M n(D, m), \quad (1.111)$$

where

$$n(D, m) = \frac{(D + m - 1)!}{(D - 1)!m!}. \quad (1.112)$$

Additionally,

$$N(D, 0) = 1. \quad (1.113)$$

Let us assume that

$$\sum_{m=0}^M n(D, m) = \frac{(D + M)!}{D!M!}. \quad (1.114)$$

Then

$$\sum_{m=0}^{M+1} n(D, m) = \frac{(D + M)!}{D!M!} + \frac{(D + M)!}{(D - 1)!(M + 1)!}. \quad (1.115)$$

The right hand side can be written as

$$\frac{(M + 1)(D + M)! + D(D + M)!}{D!(M + 1)!} = \frac{(D + M + 1)!}{D!(M + 1)!}. \quad (1.116)$$

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

$$N(D, M) = \frac{(D + M)!}{D!M!}. \quad (1.117)$$

Additionally, by the approximation

$$n! \simeq n^n \exp(-n), \quad (1.118)$$

the right hand side can be approximated as

$$\frac{(D + M)^{D+M} \exp(-(D + M))}{D^D \exp(-D) M^M \exp(-M)} = \frac{(D + M)^{D+M}}{D^D M^M}. \quad (1.119)$$



The right hand side can be written as

$$D^M \left(1 + \frac{M}{D}\right)^D \left(\frac{1}{M} + \frac{1}{D}\right)^M = M^D \left(1 + \frac{D}{M}\right)^M \left(\frac{1}{D} + \frac{1}{M}\right)^D. \quad (1.120)$$

Therefore,  $N(D, M)$  can be approximated as  $D^M$  for  $D \gg M$  and as  $M^D$  for  $M \gg D$ .

Finally, by the result above,

$$\begin{aligned} N(10, 3) &= 286, \\ N(100, 3) &= 176851, \\ N(1000, 3) &= 167668501. \end{aligned} \quad (1.121)$$

## 1.17

Let

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \quad (1.122)$$

Then

$$\Gamma(x+1) = \int_0^\infty u^x \exp(-u) du. \quad (1.123)$$

The right hand side can be written as

$$[-u^x \exp(-u)]_{u=0}^{u=\infty} + \int_0^\infty x u^{x-1} \exp(-u) du = x \Gamma(x). \quad (1.124)$$

Therefore,

$$\Gamma(x+1) = x \Gamma(x). \quad (1.125)$$

Since

$$\Gamma(1) = \int_0^1 \exp(-u) du, \quad (1.126)$$

and the right hand side can be written as 1,

$$\Gamma(1) = 0!. \quad (1.127)$$

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

$$\Gamma(x) = (x-1)!. \quad (1.128)$$

Then,

$$\Gamma(x+1) = x\Gamma(x), \quad (1.129)$$

where the right hand side can be written as  $x!$ . Therefore,

$$\Gamma(x+1) = x!. \quad (1.130)$$

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

## 1.18

Let us consider the transformation from Cartesian to polar coordinates

$$\prod_{i=1}^D \int_{-\infty}^{\infty} \exp(-x_i^2) dx_i = S_D \int_0^{\infty} \exp(-r^2) r^{D-1} dr, \quad (1.131)$$

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  $s = r^2$ , the right hand side can be written as

$$\frac{S_D}{2} \int_0^{\infty} \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \quad (1.132)$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \quad (1.133)$$

Additionally, the volume of the sphere can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \quad (1.134)$$

The right hand side can be written as

$$S_D \left[ \frac{r^D}{D} \right]_{r=0}^{r=1} = \frac{S_D}{D}. \quad (1.135)$$

Therefore,

$$V_D = \frac{S_D}{D}. \quad (1.136)$$

Finally, the results above reduce to

$$\begin{aligned} S_2 &= \frac{2\pi}{\Gamma(1)}, \\ V_2 &= \frac{S_2}{2}. \end{aligned} \tag{1.137}$$

Therefore,

$$\begin{aligned} S_2 &= 2\pi, \\ V_2 &= \pi. \end{aligned} \tag{1.138}$$

Similarly,

$$\begin{aligned} S_3 &= \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})}, \\ V_3 &= \frac{S_3}{3}. \end{aligned} \tag{1.139}$$

Therefore,

$$\begin{aligned} S_3 &= 4\pi, \\ V_3 &= \frac{4}{3}\pi. \end{aligned} \tag{1.140}$$

## 1.19

The volume of a cube of side 2 in  $D$  dimensions is  $2^D$ . Therefore, the ratio of the volume of the cocentric sphere of radius 1 divided by the volume of the cube is given by

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma(\frac{D}{2})}, \tag{1.141}$$

by 1.18.

Additionally, by Stering's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}} \exp(-x) x^{\frac{x+1}{2}}, \tag{1.142}$$

the ratio can be approximated as

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D2^{D-1}(2\pi)^{\frac{1}{2}} \exp\left(1 - \frac{D}{2}\right) \left(\frac{D}{2} - 1\right)^{\frac{D}{4}}}. \tag{1.143}$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}} \frac{1}{D} \left( \frac{e^2 \pi^2}{8D - 16} \right)^{\frac{D}{4}}. \quad (1.144)$$

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

Finally, 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

$$\frac{\sqrt{\sum_{i=1}^D 1^2}}{1} = \sqrt{D}. \quad (1.145)$$

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

## 1.20

For a vector  $\mathbf{x}$  in  $D$  dimensions, let

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \quad (1.146)$$

Then

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_r^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\phi, \quad (1.147)$$

where  $\phi$  is the vector of the angular components of the polar corrdinate 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

$$\begin{aligned} & (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon} \int J dr' d\phi \\ & = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} d\mathbf{x}. \end{aligned} \quad (1.148)$$

Therefore,

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} \simeq p(r) \epsilon, \quad (1.149)$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad (1.150)$$

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

Additionally, setting the derivative of  $p(r)$  to zero gives

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left( (D-1)r^{D-2} - \frac{r^D}{\sigma^2} \right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \quad (1.151)$$

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

$$\hat{r} = \sqrt{D-1}\sigma. \quad (1.152)$$

Additionally, by the expression of  $p(r)$  above,

$$\frac{p(\hat{r} + \epsilon)}{p(\hat{r})} = \left( \frac{\hat{r} + \epsilon}{\hat{r}} \right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \quad (1.153)$$

Using the expression of  $\hat{r}$  above, the right hand side can be written as

$$\begin{aligned} & \exp\left((D-1)\ln\left(1 + \frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\hat{r}^2}{\sigma^2}\ln\left(1 + \frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \end{aligned} \quad (1.154)$$

By the Taylor series

$$\ln(1+x) = x - \frac{1}{2}x^2 + o(x^3), \quad (1.155)$$

the right hand side can be approximated as

$$\exp\left(\frac{\hat{r}^2}{\sigma^2}\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) = \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \quad (1.156)$$

Therefore,

$$p(\hat{r} + \epsilon) \simeq p(\hat{r}) \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \quad (1.157)$$

Finally, let a vector of length  $\hat{r}$  be  $\hat{\mathbf{r}}$ . Then, by the definition of  $p(\mathbf{x})$ ,

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp\left(\frac{\hat{r}^2}{2\sigma^2}\right). \quad (1.158)$$

Substituting the expression of  $\hat{r}$  above, the right hand side can be written as  $\exp\left(\frac{D-1}{2}\right)$ . Therefore,

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp\left(\frac{D-1}{2}\right). \quad (1.159)$$

### 1.21

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

$$0 \leq a(b - a). \quad (1.160)$$

Therefore,

$$a \leq (ab)^{\frac{1}{2}}. \quad (1.161)$$

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,

$$p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2) \Rightarrow \mathbf{x} \in \mathcal{C}_1, \quad (1.162)$$

and

$$p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1) \Rightarrow \mathbf{x} \in \mathcal{C}_2. \quad (1.163)$$

Then, using the inequality above,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} \leq \int_{\mathcal{R}_1} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}, \quad (1.164)$$

and

$$\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int_{\mathcal{R}_2} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \quad (1.165)$$

Therefore,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \quad (1.166)$$

### 1.22

Let

$$E L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.167)$$

If

$$L_{kj} = 1 - \delta_{kj}, \quad (1.168)$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \quad (1.169)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \quad (1.170)$$

Therefore,

$$\mathbb{E} L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.171)$$

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

### 1.23

Let

$$\mathbb{E} L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.172)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.173)$$

Therefore,

$$\mathbb{E} L = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.174)$$

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

### 1.24 (Incomplete)

Let

$$\mathbb{E} L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} + \lambda \int_{\forall k p(\mathcal{C}_k | \mathbf{x}) < \theta} p(\mathbf{x}) d\mathbf{x}. \quad (1.175)$$

## 1.25

Let

$$E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.176)$$

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

$$\mathbf{0} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \quad (1.177)$$

The integral of the right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = \mathbf{y}(\mathbf{x}) p(\mathbf{x}) - p(\mathbf{x}) \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \quad (1.178)$$

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

$$\mathbf{0} = p(\mathbf{x}) (\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})). \quad (1.179)$$

Thus,

$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \quad (1.180)$$

Finally, for a single target variable  $t$ , it reduces to

$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = E_t(t|\mathbf{x}). \quad (1.181)$$

## 1.26

Let

$$E L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.182)$$

The right hand side can be written as

$$\begin{aligned} & \int \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &= \int \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ & \quad + 2 \int \int (\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top (E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ & \quad + \int \int \|E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \end{aligned} \quad (1.183)$$



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

$$\int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 \left( \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \quad (1.184)$$

The second term can be written as

$$2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top \left( \int (\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.185)$$

Since

$$\begin{aligned} \int \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t}}{p(\mathbf{x})}, \\ \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}), \end{aligned} \quad (1.186)$$

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

$$\int \left( \int \|\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x} = \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.187)$$

Therefore,

$$\mathbf{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.188)$$

Thus,

$$\underset{\mathbf{y}(\mathbf{x})}{\text{argmin}} \mathbf{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \quad (1.189)$$

## 1.27 (Incomplete)

Let

$$\mathbf{E} L_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \quad (1.190)$$

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

$$0 = q p(\mathbf{x}) \int |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt. \quad (1.191)$$

Therefore,

$$\operatorname{argmin}_{y(\mathbf{x})} E L_q = \left\{ y(\mathbf{x}) \mid \int |y(\mathbf{x}) - t|^{q-1} \operatorname{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = 0 \right\}. \quad (1.192)$$

Additionally, if  $q = 1$ , the integral can be written as

$$p(\mathbf{x}) \int \operatorname{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = p(\mathbf{x}) \left( \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x}) dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x}) dt \right). \quad (1.193)$$

Therefore,

$$\operatorname{argmin}_{y(\mathbf{x})} E L_1 = \operatorname{median}(t|\mathbf{x}). \quad (1.194)$$

Finally,

$$\lim_{q \rightarrow 0} \left( \operatorname{argmin}_{y(\mathbf{x})} E L_q \right) = \operatorname{mode}(t|\mathbf{x})? \quad (1.195)$$

## 1.28

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \quad (1.196)$$

Let  $h(p)$  be a function to relate  $h$  and  $p$ . Then

$$h(p^2) = 2h(p). \quad (1.197)$$

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

$$h(p^n) = nh(p). \quad (1.198)$$

Then, by the first assumption,

$$h(p^{n+1}) = h(p^n) + h(p). \quad (1.199)$$

Therefore,

$$h(p^{n+1}) = (n+1)h(p). \quad (1.200)$$

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

Additionally, for positive integers  $m$  and  $n$ ,

$$h(p^n) = h\left(p^{\frac{n}{m}m}\right). \quad (1.201)$$

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

$$h\left(p^{\frac{n}{m}}\right) = \frac{n}{m}h(p). \quad (1.202)$$

Finally, by the continuity, for a positive real number  $a$ ,

$$h(p^a) = ah(p). \quad (1.203)$$

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

$$(p \ln p)h'(p) = h(p). \quad (1.204)$$

Therefore,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + \text{const}. \quad (1.205)$$

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

$$h(p) \propto \ln p. \quad (1.206)$$

## 1.29

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

$$H(x) = - \sum_{i=1}^M p(x_i) \ln p(x_i), \quad (1.207)$$

where

$$\sum_{i=1}^M p(x_i) = 1. \quad (1.208)$$

By Jensen's inequality,

$$\sum_{i=1}^M p(x_i) \ln \frac{1}{p(x_i)} \leq \ln \left( \sum_{i=1}^M 1 \right). \quad (1.209)$$

Therefore,

$$H(x) \leq \ln M. \quad (1.210)$$

### 1.30

Let

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu, \sigma^2), \\ q(x) &= \mathcal{N}(x|m, s^2). \end{aligned} \quad (1.211)$$

By the definition,

$$\text{KL}(p||q) = - \int p(x) \ln \frac{q(x)}{p(x)} dx. \quad (1.212)$$

The right hand side can be written as

$$\begin{aligned} & - \int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi \sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\ &= - \int_{-\infty}^{\infty} p(x) \left( -\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2} \right) dx. \end{aligned} \quad (1.213)$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x) dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x) dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx. \quad (1.214)$$

The first term can be written as  $\ln \frac{s}{\sigma}$ . The second term can be written as

$$\frac{1}{2s^2} \int_{-\infty}^{\infty} (x-\mu + \mu-m)^2 p(x) dx = \frac{\sigma^2 + (\mu-m)^2}{2s^2}. \quad (1.215)$$

The third term can be written as  $-\frac{1}{2}$ . Therefore,

$$\text{KL}(p||q) = \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu-m)^2}{2s^2} - \frac{1}{2}. \quad (1.216)$$

### 1.31

Let  $\mathbf{x}$  and  $\mathbf{y}$  be two variables. Then, by the definition,

$$\begin{aligned} H(\mathbf{x}) &= - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}, \\ H(\mathbf{y}) &= - \int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \\ H(\mathbf{x}, \mathbf{y}) &= - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \end{aligned} \quad (1.217)$$

Note that

$$\begin{aligned} H(\mathbf{x}) &= - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}, \\ H(\mathbf{y}) &= - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} \right) \ln p(\mathbf{y}) d\mathbf{y}. \end{aligned} \quad (1.218)$$

Therefore,

$$H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}. \quad (1.219)$$

Since

$$\int \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = 1, \quad (1.220)$$

Jensen's inequality can be used to write that

$$- \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y} \geq - \ln \left( \int \int p(\mathbf{x})p(\mathbf{y}) d\mathbf{x} d\mathbf{y} \right). \quad (1.221)$$

The right hand side can be written as

$$- \ln \left( \int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{y}) d\mathbf{y} \right) = 0. \quad (1.222)$$

Thus,

$$H(\mathbf{x}, \mathbf{y}) \leq H(\mathbf{x}) + H(\mathbf{y}). \quad (1.223)$$

## 1.32

Let  $\mathbf{x}$  be a vector of continuous variables and

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (1.224)$$

where  $\mathbf{A}$  is a nonsingular matrix. By the definition,

$$H(\mathbf{y}) = - \int p_y(\mathbf{y}) \ln p_y(\mathbf{y}) d\mathbf{y}. \quad (1.225)$$

By the transformation

$$p_y(\mathbf{y}) = p_x(\mathbf{A}\mathbf{x}) |\det \mathbf{A}^{-1}|, \quad (1.226)$$

the right hand side can be written as

$$- \int p_x(\mathbf{Ax}) \ln p_x(\mathbf{Ax}) |\det \mathbf{A}| d\mathbf{x} - \ln |\det \mathbf{A}^{-1}| \int p_y(\mathbf{y}) d\mathbf{y}. \quad (1.227)$$

By the transformation

$$\mathbf{x}' = \mathbf{Ax}, \quad (1.228)$$

the first term can be written as

$$- \int p_x(\mathbf{x}') \ln p_x(\mathbf{x}') d\mathbf{x}' = H(\mathbf{x}), \quad (1.229)$$

and the second term can be written as

$$- \ln |\det \mathbf{A}^{-1}| = \ln |\det \mathbf{A}|. \quad (1.230)$$

Therefore,

$$H(\mathbf{y}) = H(\mathbf{x}) + \ln |\det \mathbf{A}|. \quad (1.231)$$

### 1.33

Let  $x$  and  $y$  be two discrete variables. By the definition,

$$H(y|x) = - \sum_i \sum_j p(x_i, y_j) \ln p(y_j|x_i). \quad (1.232)$$

If  $H(y|x)$  is zero, then

$$0 = - \sum_i p(x_i) \sum_j p(y_j|x_i) \ln p(y_j|x_i). \quad (1.233)$$

Since

$$\begin{aligned} p(x_i) &\geq 0, \\ p(y_j|x_i) \ln p(y_j|x_i) &\leq 0. \end{aligned} \quad (1.234)$$

for all  $i$  and  $j$ , the equation reduces to

$$p(y_j|x_i) \ln p(y_j|x_i) = 0. \quad (1.235)$$

Therefore,  $p(y_j|x_i)$  is zero or one. Thus, since

$$\sum_j p(y_j|x_i) = 1, \quad (1.236)$$

$p(y_j|x_i)$  is one for a unique  $x_i$  and zero for others.

### 1.34

Let

$$L(p(x)) = - \int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1 \left( \int_{-\infty}^{\infty} p(x) dx - 1 \right) + \lambda_2 \left( \int_{-\infty}^{\infty} xp(x) dx - \mu \right) + \lambda_3 \left( \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right). \quad (1.237)$$

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

$$0 = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2. \quad (1.238)$$

Therefore,

$$p(x) = \exp \left( -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 \right). \quad (1.239)$$

Therefore,

$$p(x) = \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} + \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right). \quad (1.240)$$

Substituting it to

$$\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= 1, \\ \int_{-\infty}^{\infty} xp(x) dx &= \mu, \\ \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx &= \sigma^2, \end{aligned} \quad (1.241)$$

gives

$$\begin{aligned} \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= 1, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} x \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= \mu, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} (x - \mu)^2 \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= \sigma^2. \end{aligned} \quad (1.242)$$

By the transformation

$$y = \sqrt{-\lambda_3} \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right), \quad (1.243)$$

they can be written as

$$\begin{aligned} & \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy = 1, \\ & \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \left( (-\lambda_3)^{-\frac{1}{2}} y + \mu - \frac{\lambda_2}{2\lambda_3} \right) \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy = \mu, \\ & \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \left( (-\lambda_3)^{-\frac{1}{2}} y - \frac{\lambda_2}{2\lambda_3} \right)^2 \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy = \sigma^2. \end{aligned} \quad (1.244)$$

Since

$$\begin{aligned} & \int_{-\infty}^{\infty} \exp(-y^2) dy = \Gamma \left( \frac{1}{2} \right), \\ & \int_{-\infty}^{\infty} y \exp(-y^2) dy = 0, \\ & \int_{-\infty}^{\infty} y^2 \exp(-y^2) dy = \Gamma \left( \frac{3}{2} \right), \end{aligned} \quad (1.245)$$

they can be written as

$$\begin{aligned} & \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) (-\lambda_3)^{-\frac{1}{2}} \Gamma \left( \frac{1}{2} \right) = 1, \\ & \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) (-\lambda_3)^{-\frac{1}{2}} \Gamma \left( \frac{1}{2} \right) = \mu, \\ & \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \left( (-\lambda_3)^{-\frac{3}{2}} \Gamma \left( \frac{3}{2} \right) + (-\lambda_3)^{-\frac{1}{2}} \frac{\lambda_2^2}{4\lambda_3^2} \Gamma \left( \frac{1}{2} \right) \right) = \sigma^2. \end{aligned} \quad (1.246)$$

Therefore,

$$\begin{aligned} \lambda_1 &= 1 - \frac{1}{2} \ln(2\pi\sigma^2), \\ \lambda_2 &= 0, \\ \lambda_3 &= -\frac{1}{2\sigma^2}. \end{aligned} \quad (1.247)$$

Thus,

$$p(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right). \quad (1.248)$$



### 1.35

Let  $x$  be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.249)$$

Then, by the definition,

$$H(x) = - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \ln \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.250)$$

The right hand side can be written as

$$\begin{aligned} & - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2 \right) dx \\ &= \frac{1}{2} \ln(2\pi\sigma^2) \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx. \end{aligned} \quad (1.251)$$

Therefore,

$$H(x) = \frac{1}{2} (1 + \ln(2\pi\sigma^2)). \quad (1.252)$$

### 1.36 (Incomplete)

Let  $f$  be a strictly convex function. Then, by the definition,

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b), \quad (1.253)$$

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

$$x = \lambda a + (1 - \lambda)b. \quad (1.254)$$

Then, the inequality can be written as

$$f(x) \leq \frac{b - x}{b - a} f(a) + \frac{x - a}{b - a} f(b). \quad (1.255)$$

Let

$$g(x) = \frac{b - x}{b - a} f(a) + \frac{x - a}{b - a} f(b) - f(x). \quad (1.256)$$

Then,

$$g(x) \geq 0. \quad (1.257)$$

Additionally, for  $x > a$ ,

$$g(x) = (x - a) \left( \frac{f(b) - f(a)}{b - a} - \frac{f(x) - f(a)}{x - a} \right). \quad (1.258)$$

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

$$\begin{aligned} f'(c) &= \frac{f(b) - f(a)}{b - a}, \\ f'(y) &= \frac{f(x) - f(a)}{x - a}. \end{aligned} \quad (1.259)$$

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

$$f'(y) \leq f'(c). \quad (1.260)$$

### 1.37

Let  $\mathbf{x}$  and  $\mathbf{y}$  be two variables. Then, by the definition,

$$H(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \quad (1.261)$$

The right hand side can be written as

$$\begin{aligned} & - \int \int p(\mathbf{x}, \mathbf{y}) (\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})) d\mathbf{x} d\mathbf{y} \\ & = - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (1.262)$$

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

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \quad (1.263)$$

### 1.38

Let  $f$  be a strictly convex function. Then, by the definition,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \quad (1.264)$$

where  $0 \leq \lambda \leq 1$ . Let us assume that

$$f\left(\sum_{i=1}^M \lambda_i x_i\right) \leq \sum_{i=1}^M \lambda_i f(x_i), \quad (1.265)$$

where  $\lambda_i \geq 0$  and

$$\sum_{i=1}^M \lambda_i = 1. \quad (1.266)$$

Here, let  $\lambda_i \geq 0$  and

$$\sum_{i=1}^{M+1} \lambda_i = 1. \quad (1.267)$$

Then, by the definition,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{i=1}^M \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right). \quad (1.268)$$

By the assumption,

$$f\left(\sum_{i=1}^M \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right) \leq \sum_{i=1}^M \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \quad (1.269)$$

Therefore,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{i=1}^M \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \quad (1.270)$$

Thus,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \sum_{i=1}^{M+1} \lambda_i f(x_i). \quad (1.271)$$

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

### 1.39

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

$$\begin{aligned}p(x = 0, y = 0) &= \frac{1}{3}, \\p(x = 0, y = 1) &= \frac{1}{3}, \\p(x = 1, y = 0) &= 0, \\p(x = 1, y = 1) &= \frac{1}{3}.\end{aligned}\tag{1.272}$$

(a)

By the definition,

$$H(x) = - \sum p(x) \ln p(x).\tag{1.273}$$

By the distribution,

$$\begin{aligned}p(x = 0) &= \frac{2}{3}, \\p(x = 1) &= \frac{1}{3}.\end{aligned}\tag{1.274}$$

Therefore,

$$H(x) = \ln 3 - \frac{2}{3} \ln 2.\tag{1.275}$$

(b)

By the definition,

$$H(y) = - \sum p(y) \ln p(y).\tag{1.276}$$

By the distribution,

$$\begin{aligned}p(y = 0) &= \frac{1}{3}, \\p(y = 1) &= \frac{2}{3}.\end{aligned}\tag{1.277}$$

Therefore,

$$H(y) = \ln 3 - \frac{2}{3} \ln 2.\tag{1.278}$$

(c)

By the definition,

$$H(y|x) = - \sum p(x, y) \ln p(y|x). \quad (1.279)$$

By the definition,

$$\begin{aligned} p(y = 0|x = 0) &= \frac{p(x = 0, y = 0)}{p(x = 0)}, \\ p(y = 0|x = 1) &= \frac{p(x = 1, y = 0)}{p(x = 1)}, \\ p(y = 1|x = 0) &= \frac{p(x = 0, y = 1)}{p(x = 0)}, \\ p(y = 1|x = 1) &= \frac{p(x = 1, y = 1)}{p(x = 1)}. \end{aligned} \quad (1.280)$$

Then, by the distribution,

$$\begin{aligned} p(y = 0|x = 0) &= \frac{1}{2}, \\ p(y = 0|x = 1) &= 0, \\ p(y = 1|x = 0) &= \frac{1}{2}, \\ p(y = 1|x = 1) &= 1. \end{aligned} \quad (1.281)$$

Therefore,

$$H(y|x) = \frac{2}{3} \ln 2. \quad (1.282)$$

(d)

By the definition,

$$H(x|y) = - \sum p(x, y) \ln p(x|y). \quad (1.283)$$

By the definition,

$$\begin{aligned}
p(x=0|y=0) &= \frac{p(x=0, y=0)}{p(y=0)}, \\
p(x=0|y=1) &= \frac{p(x=0, y=1)}{p(y=1)}, \\
p(x=1|y=0) &= \frac{p(x=1, y=0)}{p(y=0)}, \\
p(x=1|y=1) &= \frac{p(x=1, y=1)}{p(y=1)}.
\end{aligned} \tag{1.284}$$

Then, by the distribution,

$$\begin{aligned}
p(x=0|y=0) &= 1, \\
p(x=0|y=1) &= \frac{1}{2}, \\
p(x=1|y=0) &= 0, \\
p(x=1|y=1) &= \frac{1}{2}.
\end{aligned} \tag{1.285}$$

Therefore,

$$H(x|y) = \frac{2}{3} \ln 2. \tag{1.286}$$

(e)

By the definition,

$$H(x, y) = - \sum p(x, y) \ln p(x, y). \tag{1.287}$$

Therefore,

$$H(x, y) = \ln 3. \tag{1.288}$$

(f)

By the definition,

$$I(x, y) = - \sum p(x, y) \ln \frac{p(x)p(y)}{p(x, y)}. \tag{1.289}$$

By the distribution, the right hand side can be written as

$$H(x) + H(y) - H(x, y). \quad (1.290)$$

Therefore,

$$I(x, y) = \ln 3 - \frac{4}{3} \ln 2. \quad (1.291)$$

## 1.40

Let  $\{x_i\}$  be a set of points where  $x_i > 0$ , and let  $\{\lambda_i\}$  be a set of coefficients where  $\lambda_i \geq 0$  and

$$\sum_{i=1}^M \lambda_i = 1. \quad (1.292)$$

By Jensen's inequality,

$$\sum_{i=1}^M \lambda_i \ln x_i \leq \ln \left( \sum_{i=1}^M \lambda_i x_i \right). \quad (1.293)$$

Therefore,

$$\prod_{i=1}^M x_i^{\lambda_i} \leq \sum_{i=1}^M \lambda_i x_i. \quad (1.294)$$

Substituting

$$\lambda_i = \frac{1}{M} \quad (1.295)$$

gives

$$\left( \prod_{i=1}^M x_i \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{i=1}^M x_i. \quad (1.296)$$

## 1.41

Let  $\mathbf{x}$  and  $\mathbf{y}$  be continuous variables. Then, by the definition,

$$I(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y}. \quad (1.297)$$

The right hand side can be written as

$$\begin{aligned}
& - \int \int p(\mathbf{x}, \mathbf{y}) \left( \ln p(\mathbf{x}) + \ln \frac{p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \\
& = - \int \left( \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} + \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y}.
\end{aligned} \tag{1.298}$$

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

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}). \tag{1.299}$$

By the definition,

$$I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x}). \tag{1.300}$$

Thus,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \tag{1.301}$$



## 2 Probability Distributions

### 2.1

Let  $x$  be a variable such that

$$p(x|\mu) = \mu^x(1 - \mu)^{1-x}, \quad (2.1)$$

where  $x \in \{0, 1\}$ . Then,

$$\sum_x p(x|\mu) = 1. \quad (2.2)$$

By the definition,

$$\begin{aligned} \mathbb{E} x &= \mu, \\ \mathbb{E} x^2 &= \mu, \end{aligned} \quad (2.3)$$

Since

$$\text{var } x = \mathbb{E} x^2 - (\mathbb{E} x)^2, \quad (2.4)$$

we have

$$\text{var } x = \mu(1 - \mu). \quad (2.5)$$

By the definition,

$$\mathbb{H}(x) = - \sum_x p(x|\mu) \ln p(x|\mu). \quad (2.6)$$

Therefore,

$$\mathbb{H}(x) = -\mu \ln \mu - (1 - \mu) \ln(1 - \mu). \quad (2.7)$$

### 2.2

Let  $x$  be a variable such that

$$p(x|\mu) = \left( \frac{1 - \mu}{2} \right)^{\frac{1-x}{2}} \left( \frac{1 + \mu}{2} \right)^{\frac{1+x}{2}}, \quad (2.8)$$

where  $x \in \{-1, 1\}$ . Then,

$$\sum_x p(x|\mu) = 1. \quad (2.9)$$

By the definition,

$$\begin{aligned} \mathbb{E} x &= \mu, \\ \mathbb{E} x^2 &= 1, \end{aligned} \quad (2.10)$$

Since

$$\text{var } x = \mathbb{E} x^2 - (\mathbb{E} x)^2, \quad (2.11)$$

we have

$$\text{var } x = 1 - \mu^2. \quad (2.12)$$

By the definition,

$$H(x) = - \sum_x p(x|\mu) \ln p(x|\mu). \quad (2.13)$$

Therefore,

$$H(x) = -\frac{1-\mu}{2} \ln \frac{1-\mu}{2} - \frac{1+\mu}{2} \ln \frac{1+\mu}{2}. \quad (2.14)$$

### 2.3

By the definition,

$$\begin{aligned} \binom{N}{m} &= \frac{N!}{m!(N-m)!}, \\ \binom{N}{m-1} &= \frac{N!}{(m-1)!(N-m+1)!} \end{aligned} \quad (2.15)$$

Therefore,

$$\binom{N}{m} + \binom{N}{m-1} = \frac{(N-m+1)N! + mN!}{m!(N-m+1)!}. \quad (2.16)$$

By the definition, the right hand side can be written as

$$\frac{(N+1)!}{m!(N+1-m)!} = \binom{N+1}{m}. \quad (2.17)$$

Thus,

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}. \quad (2.18)$$

Note that

$$1+x = \sum_{m=0}^1 \binom{1}{m} x^m. \quad (2.19)$$

Let us assume that

$$(1+x)^N = \sum_{m=0}^N \binom{N}{m} x^m. \quad (2.20)$$

Then,

$$(1+x)^{N+1} = \sum_{m=0}^N \binom{N}{m} x^m + \sum_{m=0}^N \binom{N}{m} x^{m+1}. \quad (2.21)$$

By the result above, the right hand side can be written as

$$\sum_{m=0}^N \binom{N}{m} x^m + \sum_{m=1}^{N+1} \binom{N}{m-1} x^m = 1 + x^{N+1} + \sum_{m=1}^N \binom{N+1}{m} x^m. \quad (2.22)$$

Therefore,

$$(1+x)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} x^m. \quad (2.23)$$

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

Finally, let  $m$  be a variable such that

$$p(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}. \quad (2.24)$$

Then

$$\sum_{m=0}^N p(m|\mu) = \sum_{m=0}^N \binom{N}{m} \mu^m (1-\mu)^{N-m}. \quad (2.25)$$

By the result above, the right hand side can be written as

$$(1-\mu)^N \sum_{m=0}^N \binom{N}{m} \left( \frac{\mu}{1-\mu} \right)^m = (1-\mu)^N \left( 1 + \frac{\mu}{1-\mu} \right)^N. \quad (2.26)$$

Therefore,

$$\sum_{m=0}^N p(m|\mu) = 1. \quad (2.27)$$

## 2.4

Let  $m$  be a variable such that

$$p(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}. \quad (2.28)$$

Then

$$E m = \sum_{m=0}^N m \binom{N}{m} \mu^m (1-\mu)^{N-m}. \quad (2.29)$$

Taking the derivative of

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

with respect to  $\mu$  gives

$$\sum_{m=0}^N m \binom{N}{m} \mu^{m-1} (1-\mu)^{N-m} - \sum_{m=0}^N (N-m) \binom{N}{m} \mu^m (1-\mu)^{N-m-1} = 0. \quad (2.31)$$

The first term of the left hand side can be written as  $\frac{1}{\mu} E m$ . Since

$$(N-m) \binom{N}{m} = N \binom{N-1}{m}, \quad (2.32)$$

the second term of the left hand side can be written as

$$-N \sum_{m=0}^{N-1} \binom{N-1}{m} \mu^m (1-\mu)^{N-m-1} = -N. \quad (2.33)$$

Therefore,

$$E m = N\mu. \quad (2.34)$$

Taking the second derivative of

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

with respect to  $\mu$  gives

$$\begin{aligned} & \sum_{m=0}^N m(m-1) \binom{N}{m} \mu^{m-2} (1-\mu)^{N-m} \\ & - 2 \sum_{m=0}^N m(N-m) \binom{N}{m} \mu^{m-1} (1-\mu)^{N-m-1} \\ & + \sum_{m=0}^N (N-m)(N-m-1) \binom{N}{m} \mu^m (1-\mu)^{N-m-2} = 0. \end{aligned} \quad (2.36)$$

The first term of the left hand side can be written as  $\frac{1}{\mu^2} E m(m-1)$ . Since

$$\begin{aligned} m(N-m) \binom{N}{m} &= N(N-1) \binom{N-2}{m-1}, \\ (N-m)(N-m-1) \binom{N}{m} &= N(N-1) \binom{N-2}{m}, \end{aligned} \quad (2.37)$$

the second and third term of the left hand side can be written as

$$\begin{aligned} -2N(N-1) \sum_{m=1}^{N-1} \binom{N-2}{m-1} \mu^{m-1} (1-\mu)^{N-m-1} &= -2N(N-1), \\ N(N-1) \sum_{m=0}^N \binom{N-2}{m} \mu^m (1-\mu)^{N-m-2} &= N(N-1). \end{aligned} \quad (2.38)$$

Therefore,

$$E m(m-1) = N(N-1)\mu^2. \quad (2.39)$$

Thus, since

$$\text{var } m = E m(m-1) + E m - (E m)^2, \quad (2.40)$$

we have

$$\text{var } m = N\mu(1-\mu). \quad (2.41)$$

## 2.5

By the definition,

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1} \exp(-x) dx \int_0^\infty y^{b-1} \exp(-y) dy. \quad (2.42)$$

By the transformation  $t = x + y$ , the right hand side can be written as

$$\begin{aligned} &\int_0^\infty x^{a-1} \left( \int_x^\infty (t-x)^{b-1} \exp(-t) dt \right) dx \\ &= \int_0^\infty \left( \int_0^t x^{a-1} (t-x)^{b-1} dx \right) \exp(-t) dt. \end{aligned} \quad (2.43)$$

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

$$\begin{aligned} &\int_0^\infty \left( \int_0^1 (t\mu)^{a-1} t^{b-1} (1-\mu)^{b-1} t d\mu \right) \exp(-t) dt \\ &= \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu \int_0^\infty t^{a+b-1} \exp(-t) dt. \end{aligned} \quad (2.44)$$

By the definition, the second integral of the right hand side can be written as  $\Gamma(a+b)$ . Therefore,

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \quad (2.45)$$

## 2.6

Let  $\mu$  be a variable such that

$$p(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}. \quad (2.46)$$

Then

$$\begin{aligned} E\mu &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a(1-\mu)^{b-1}d\mu, \\ E\mu^2 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1}(1-\mu)^{b-1}d\mu. \end{aligned} \quad (2.47)$$

Since

$$\begin{aligned} \int_0^1 \mu^a(1-\mu)^{b-1}d\mu &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}, \\ \int_0^1 \mu^{a+1}(1-\mu)^{b-1}d\mu &= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}, \end{aligned} \quad (2.48)$$

we have

$$\begin{aligned} E\mu &= \frac{a}{a+b}, \\ E\mu^2 &= \frac{a(a+1)}{(a+b)(a+b+1)}. \end{aligned} \quad (2.49)$$

Since

$$\text{var } \mu = E\mu^2 - (E\mu)^2, \quad (2.50)$$

we have

$$\text{var } \mu = \frac{ab}{(a+b)^2(a+b+1)}. \quad (2.51)$$

Since

$$\frac{\partial}{\partial \mu} p(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \left( \frac{a-1}{\mu} - \frac{b-1}{1-\mu} \right), \quad (2.52)$$

we have

$$\text{mode } \mu = \frac{a-1}{a+b-2}. \quad (2.53)$$

## 2.7

Let  $m$  and  $l$  be a variable such that

$$p(m, l|\mu) = \binom{m+l}{m} \mu^m (1-\mu)^l, \quad (2.54)$$

where

$$p(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \quad (2.55)$$

By 2.6,

$$E(\mu|a, b) = \frac{a}{a+b}. \quad (2.56)$$

Note that

$$\mu_{\text{ML}} = \frac{m}{m+l}. \quad (2.57)$$

Since

$$p(\mu|m, l, a, b) \propto p(m, l|\mu)p(\mu|a, b), \quad (2.58)$$

we have

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1} (1-\mu)^{l+b-1}. \quad (2.59)$$

Therefore, by 2.6,

$$E(\mu|m, l, a, b) = \frac{m+a}{m+l+a+b}. \quad (2.60)$$

Thus,

$$E(\mu|m, l, a, b) = \lambda \mu_{\text{ML}} + (1-\lambda) E(\mu|a, b), \quad (2.61)$$

where

$$\lambda = \frac{m+l}{m+l+a+b}. \quad (2.62)$$

## 2.8

Let  $x$  and  $y$  be variables. Then, by the definition,

$$E x = \int x p(x) dx. \quad (2.63)$$

The right hand side can be written as

$$\int x \left( \int p(x, y) dy \right) dx = \int \left( \int x p(x|y) dx \right) p(y) dy. \quad (2.64)$$

Therefore,

$$\mathbb{E} x = \mathbb{E}_y (\mathbb{E}_x(x|y)) . \quad (2.65)$$

Additionally, by the definition,

$$\text{var } x = \mathbb{E} (x - \mathbb{E} x)^2 . \quad (2.66)$$

By the result above, the right hand side can be written as

$$\begin{aligned} & \mathbb{E}_y (\mathbb{E}_x ((x - \mathbb{E}_x(x|y) + \mathbb{E}_x(x|y) - \mathbb{E} x)^2 | y)) \\ &= \mathbb{E}_y (\mathbb{E}_x ((x - \mathbb{E}_x(x|y))^2 | y)) \\ & \quad + 2 \mathbb{E}_y (((\mathbb{E}_x(x|y) - \mathbb{E} x) \mathbb{E}_x(x - \mathbb{E}_x(x|y)) | y)) \\ & \quad + \mathbb{E}_y ((\mathbb{E}_x(x|y) - \mathbb{E} x)^2 | y) . \end{aligned} \quad (2.67)$$

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

$$2 \mathbb{E}_y ((\mathbb{E}_x(x|y) - \mathbb{E} x) (\mathbb{E}_x(x|y) - \mathbb{E}_x(x|y))) = 0. \quad (2.68)$$

By the result above, the third term can be written as

$$\mathbb{E}_y (\mathbb{E}_x(x|y) - \mathbb{E}_y (\mathbb{E}_x(x|y)))^2 = \text{var}_y (\mathbb{E}_x(x|y)) . \quad (2.69)$$

Therefore,

$$\text{var } x = \mathbb{E}_y (\text{var}_x(x|y)) + \text{var}_y (\mathbb{E}_x(x|y)) . \quad (2.70)$$

## 2.9

For a vector  $\boldsymbol{\mu}$  in 2 dimensions, 2.5 gives

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

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

$$\int_{\substack{\sum_{k=1}^M \mu_k = 1 \\ \mu_k \geq 0}} \prod_{k=1}^M \mu_k^{\alpha_k - 1} d\boldsymbol{\mu} = \frac{\prod_{k=1}^M \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^M \alpha_k)} .$$

Under the constraint

$$\sum_{k=1}^{M+1} \mu_k = 1, \quad (2.71)$$



we have

$$\begin{aligned}
& \int_0^{1-\sum_{k=1}^{M-1} \mu_k} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\mu_{M+1} \\
&= \left( \prod_{k=1}^{M-1} \mu_k^{\alpha_k-1} \right) \int_0^{1-\sum_{k=1}^{M-1} \mu_k} \mu_{M+1}^{\alpha_{M+1}-1} \left( 1 - \sum_{k=1}^{M-1} \mu_k - \mu_{M+1} \right)^{\alpha_M-1} d\mu_{M+1}.
\end{aligned} \tag{2.72}$$

By the transformation

$$\mu'_{M+1} = \frac{\mu_{M+1}}{1 - \sum_{k=1}^{M-1} \mu_k}, \tag{2.73}$$

the integral of the right hand side can be written as

$$\begin{aligned}
& \int_0^1 \left( \left( 1 - \sum_{k=1}^{M-1} \mu_k \right) \mu'_{M+1} \right)^{\alpha_{M+1}-1} \left( \left( 1 - \sum_{k=1}^{M-1} \mu_k \right) (1 - \mu'_{M+1}) \right)^{\alpha_M-1} \left( 1 - \sum_{k=1}^{M-1} \mu_k \right) d\mu'_{M+1} \\
&= \left( 1 - \sum_{k=1}^{M-1} \mu_k \right)^{\alpha_M + \alpha_{M+1} - 1} \int_0^1 \mu'_{M+1}^{\alpha_{M+1}-1} (1 - \mu'_{M+1})^{\alpha_M-1} d\mu'_{M+1}.
\end{aligned} \tag{2.74}$$

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

$$\frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \tag{2.75}$$

Therefore,

$$\int_0^{1-\sum_{k=1}^{M-1} \mu_k} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\mu_{M+1} = \left( \prod_{k=1}^{M-1} \mu_k^{\alpha_k-1} \right) \left( 1 - \sum_{k=1}^{M-1} \mu_k \right)^{\alpha_M + \alpha_{M+1} - 1} \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \tag{2.76}$$

By the assumption,

$$\int_{\substack{\sum_{k=1}^M \mu_k = 1 \\ \mu_k \geq 0}} \left( \prod_{k=1}^{M-1} \mu_k^{\alpha_k-1} \right) \mu_M^{\alpha_M + \alpha_{M+1} - 1} d\boldsymbol{\mu} = \frac{(\prod_{k=1}^{M-1} \Gamma(\alpha_k))\Gamma(\alpha_M + \alpha_{M+1})}{\Gamma(\sum_{k=1}^{M+1} \alpha_k)}.$$

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

$$\int_{\substack{\sum_{k=1}^{M+1} \mu_k = 1 \\ \mu_k \geq 0}} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\boldsymbol{\mu} = \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})} \frac{(\prod_{k=1}^{M-1} \Gamma(\alpha_k))\Gamma(\alpha_M + \alpha_{M+1})}{\Gamma(\sum_{k=1}^{M+1} \alpha_k)}.$$

The right hand side can be written as

$$\frac{\prod_{k=1}^{M+1} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{M+1} \alpha_k)}. \quad (2.77)$$

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

## 2.10

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

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k-1}. \quad (2.78)$$

Then

$$\begin{aligned} \mathbb{E} \mu_j &= \int \mu_j p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \\ \mathbb{E} \mu_j^2 &= \int \mu_j^2 p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \\ \mathbb{E} \mu_j \mu_l &= \int \mu_j \mu_l p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}. \end{aligned} \quad (2.79)$$

If  $j \neq l$ , then the right hand sides can be written as

$$\begin{aligned} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 1)} &= \frac{\alpha_j}{\sum_{k=1}^K \alpha_k}, \\ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+2)}{\Gamma(\alpha_j)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 2)} &= \frac{\alpha_j(\alpha_j + 1)}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}, \\ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+1)\Gamma(\alpha_l+1)}{\Gamma(\alpha_j)\Gamma(\alpha_l)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 2)} &= \frac{\alpha_j \alpha_l}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}. \end{aligned} \quad (2.80)$$

Therefore,

$$\begin{aligned} \mathbb{E} \mu_j &= \frac{\alpha_j}{\sum_{k=1}^K \alpha_k}, \\ \mathbb{E} \mu_j^2 &= \frac{\alpha_j(\alpha_j + 1)}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}, \\ \mathbb{E} \mu_j \mu_l &= \frac{\alpha_j \alpha_l}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}. \end{aligned} \quad (2.81)$$

Since

$$\begin{aligned}\text{var } \mu_j &= \text{E } \mu_j^2 - (\text{E } \mu_j)^2, \\ \text{cov } (\mu_j, \mu_l) &= \text{E } \mu_j \mu_l - \text{E } \mu_j \text{E } \mu_l,\end{aligned}\tag{2.82}$$

we have

$$\begin{aligned}\text{var } \mu_j &= \frac{\alpha_j (\sum_{k=1}^K \alpha_k - \alpha_j)}{\sum_{k=1}^K \alpha_k^2 (\sum_{k=1}^K \alpha_k + 1)}, \\ \text{cov } (\mu_j, \mu_l) &= -\frac{\alpha_j \alpha_l}{(\sum_{k=1}^K \alpha_k)^2 (\sum_{k=1}^K \alpha_k + 1)}.\end{aligned}\tag{2.83}$$

## 2.11

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

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k-1}.\tag{2.84}$$

Then

$$\text{E } \ln \mu_j = \int (\ln \mu_j) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}.\tag{2.85}$$

Since

$$\frac{\partial}{\partial \alpha_j} p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \left( \frac{\Gamma'\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} - \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} + \ln \mu_j \right) p(\boldsymbol{\mu}|\boldsymbol{\alpha}),\tag{2.86}$$

we have

$$\text{E } \ln \mu_j = \frac{\partial}{\partial \alpha_j} \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} + \left( \psi(\alpha_j) - \psi\left(\sum_{k=1}^K \alpha_k\right) \right) \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu},\tag{2.87}$$

where

$$\psi(a) = \frac{d}{da} \ln \Gamma(a).\tag{2.88}$$

Therefore,

$$\text{E } \ln \mu_j = \psi(\alpha_j) - \psi\left(\sum_{k=1}^K \alpha_k\right).\tag{2.89}$$

## 2.12

Let  $x$  be a variable such that

$$p(x|a, b) = \frac{1}{b-a}, \quad (2.90)$$

where  $a < b$ . Then

$$\int_a^b p(x|a, b) dx = 1. \quad (2.91)$$

Note that

$$\begin{aligned} \mathbb{E} x &= \int_a^b x p(x|a, b) dx, \\ \mathbb{E} x^2 &= \int_a^b x^2 p(x|a, b) dx. \end{aligned} \quad (2.92)$$

The right hand sides can be written as

$$\begin{aligned} \frac{1}{b-a} \int_a^b x dx &= \frac{1}{2}(a+b), \\ \frac{1}{b-a} \int_a^b x^2 dx &= \frac{1}{3}(a^2 + ab + b^2). \end{aligned} \quad (2.93)$$

Therefore,

$$\begin{aligned} \mathbb{E} x &= \frac{1}{2}(a+b), \\ \mathbb{E} x^2 &= \frac{1}{3}(a^2 + ab + b^2). \end{aligned} \quad (2.94)$$

Since

$$\text{var } x = \mathbb{E} x^2 - (\mathbb{E} x)^2, \quad (2.95)$$

we have

$$\text{var } x = \frac{1}{12}(b-a)^2. \quad (2.96)$$

## 2.13

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}). \end{aligned} \quad (2.97)$$

Then, by the definition,

$$\text{KL}(p||q) = - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}. \quad (2.98)$$

Since

$$\ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \frac{(2\pi)^{-\frac{D}{2}} (|\det \mathbf{L}|)^{-\frac{1}{2}} \exp \left( -\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) \right)}{(2\pi)^{-\frac{D}{2}} (|\det \boldsymbol{\Sigma}|)^{-\frac{1}{2}} \exp \left( -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right)}, \quad (2.99)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{2} \ln \left| \frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}} \right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & + \frac{1}{2} \int (\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & - \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \end{aligned} \quad (2.100)$$

Let us look at each term. Since

$$\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = 1, \quad (2.101)$$

the first term can be written as  $\frac{1}{2} \ln \left| \frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}} \right|$ . Since

$$(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) = (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m}), \quad (2.102)$$

the second term can be written as

$$\begin{aligned} & \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} \int (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & + \frac{1}{2} (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m}) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \end{aligned} \quad (2.103)$$

Since

$$\begin{aligned} & \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = 1, \\ & \int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\mu}, \\ & \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\Sigma}, \end{aligned} \quad (2.104)$$

it can be written as

$$\frac{1}{2} \text{tr} (\mathbf{L}^{-1} \mathbf{\Sigma}) + \frac{1}{2} (\boldsymbol{\mu} - \mathbf{m})^T \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}). \quad (2.105)$$

Since

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} = \mathbf{\Sigma}, \quad (2.106)$$

the third term can be written as

$$-\frac{1}{2} \text{tr} (\mathbf{\Sigma}^{-1} \mathbf{\Sigma}) = -\frac{D}{2}. \quad (2.107)$$

Therefore,

$$\text{KL}(p||q) = \frac{1}{2} \left( \ln \left| \frac{\det \mathbf{L}}{\det \mathbf{\Sigma}} \right| + \text{tr} (\mathbf{L}^{-1} \mathbf{\Sigma}) + (\boldsymbol{\mu} - \mathbf{m})^T \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}) - D \right). \quad (2.108)$$

## 2.14

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

$$\begin{aligned} L(p(\mathbf{x})) = & - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} + \lambda \left( \int p(\mathbf{x}) d\mathbf{x} - 1 \right) \\ & + \mathbf{l}^T \left( \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\mu} \right) + \mathbf{m}^T \left( \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x} - \mathbf{\Sigma} \right) \mathbf{m}. \end{aligned} \quad (2.109)$$

Then

$$\frac{\delta L(p(\mathbf{x}))}{\delta p(\mathbf{x})} = -\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^T \mathbf{x} + \mathbf{m}^T (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{m}. \quad (2.110)$$

Setting the left hand side to zero gives

$$p(\mathbf{x}) = \exp (-1 + \lambda + \mathbf{l}^T \mathbf{x} + \mathbf{m}^T (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{m}), \quad (2.111)$$

so that

$$p(\mathbf{x}) = \exp (-1 + \lambda - \mathbf{l}^T \mathbf{M} \mathbf{l} + (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M} \mathbf{l})^T \mathbf{M}^{-1} (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M} \mathbf{l})), \quad (2.112)$$

where

$$\mathbf{M} = (\mathbf{m} \mathbf{m}^T)^{-1}. \quad (2.113)$$

Substituting it to

$$\begin{aligned}\int p(\mathbf{x})d\mathbf{x} &= 1, \\ \int \mathbf{x}p(\mathbf{x})d\mathbf{x} &= \boldsymbol{\mu}, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x})d\mathbf{x} &= \boldsymbol{\Sigma},\end{aligned}\tag{2.114}$$

and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l}\tag{2.115}$$

gives

$$\begin{aligned}\exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}) \int \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= 1, \\ \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}) \int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= \boldsymbol{\mu}, \\ \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}) \int (\mathbf{y} + \mathbf{M}\mathbf{l})(\mathbf{y} + \mathbf{M}\mathbf{l})^\top \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= \boldsymbol{\Sigma}.\end{aligned}\tag{2.116}$$

Since

$$\begin{aligned}\int \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \left( \Gamma\left(\frac{1}{2}\right) \right)^D, \\ \int \mathbf{y} \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \mathbf{0}, \\ \int \mathbf{y}\mathbf{y}^\top \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \Gamma\left(\frac{3}{2}\right) \left( \Gamma\left(\frac{1}{2}\right) \right)^{D-1} \mathbf{I},\end{aligned}\tag{2.117}$$

they can be written as

$$\begin{aligned}\exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}) \left( \Gamma\left(\frac{1}{2}\right) \right)^D (\det \mathbf{M})^{\frac{1}{2}} &= 1, \\ \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}) (\boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \left( \Gamma\left(\frac{1}{2}\right) \right)^D (\det \mathbf{M})^{\frac{1}{2}} &= \boldsymbol{\mu}, \\ \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}) \left( \Gamma\left(\frac{3}{2}\right) \left( \Gamma\left(\frac{1}{2}\right) \right)^{D-1} \mathbf{M} + \mathbf{M}\mathbf{l}(\mathbf{M}\mathbf{l})^\top \left( \Gamma\left(\frac{1}{2}\right) \right)^D \right) (\det \mathbf{M})^{\frac{1}{2}} &= \boldsymbol{\Sigma}.\end{aligned}\tag{2.118}$$

Therefore,

$$\begin{aligned}\lambda &= 1 - \frac{D}{2} \ln \pi - \frac{1}{2} \ln(\det \mathbf{M}), \\ \mathbf{l} &= \mathbf{0}, \\ \mathbf{M} &= 2\mathbf{\Sigma}.\end{aligned}\tag{2.119}$$

Thus,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} (\det \mathbf{\Sigma})^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).\tag{2.120}$$

## 2.15

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}).\tag{2.121}$$

Then, by the definition,

$$\mathbf{H}(\mathbf{x}) = - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x}.\tag{2.122}$$

The right hand side can be written as

$$\begin{aligned}& - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) \left( -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln |\det \mathbf{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x} \\ &= \left( \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\det \mathbf{\Sigma}| \right) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} \\ &+ \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x}.\end{aligned}\tag{2.123}$$

Since

$$\begin{aligned}\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} &= 1, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) d\mathbf{x} &= \mathbf{\Sigma},\end{aligned}\tag{2.124}$$

the first and second term of the right hand side can be written as

$$\frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\det \mathbf{\Sigma}|\tag{2.125}$$



and

$$\frac{1}{2} \text{tr} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}) = \frac{D}{2}. \quad (2.126)$$

Therefore,

$$H(\mathbf{x}) = \frac{D}{2} (1 + \ln(2\pi)) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}|. \quad (2.127)$$

## 2.16

Let  $x$  be a variable such that

$$x = x_1 + x_2, \quad (2.128)$$

where

$$\begin{aligned} p(x_1) &= \mathcal{N}(x_1 | \mu_1, \tau_1^{-1}), \\ p(x_2) &= \mathcal{N}(x_2 | \mu_2, \tau_2^{-1}). \end{aligned} \quad (2.129)$$

Then

$$p(x) = \int_{-\infty}^{\infty} p(x|x_2)p(x_2)dx_2. \quad (2.130)$$

The right hand side can be written as

$$\begin{aligned} & \int_{-\infty}^{\infty} \mathcal{N}(x | \mu_1 + x_2, \tau_1^{-1}) \mathcal{N}(x_2 | \mu_2, \tau_2^{-1}) dx_2 \\ &= \int_{-\infty}^{\infty} \left( \frac{\tau_1}{2\pi} \right)^{\frac{1}{2}} \exp \left( -\frac{\tau_1}{2} (x - \mu_1 - x_2)^2 \right) \left( \frac{\tau_2}{2\pi} \right)^{\frac{1}{2}} \exp \left( -\frac{\tau_2}{2} (x_2 - \mu_2)^2 \right) dx_2. \end{aligned} \quad (2.131)$$

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

$$\begin{aligned} & -\frac{\tau_1 + \tau_2}{2} \left( x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1}{2} (x - \mu_1)^2 - \frac{\tau_2}{2} \mu_2^2 \\ & + \frac{\tau_1 + \tau_2}{2} \left( \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 \\ & = -\frac{\tau_1 + \tau_2}{2} \left( x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)} (x - \mu_1 - \mu_2)^2. \end{aligned} \quad (2.132)$$

Therefore,

$$p(x) = \mathcal{N}(x | \mu_1 + \mu_2, \tau_1^{-1} + \tau_2^{-1}). \quad (2.133)$$

Thus, by 1.35,

$$H(x) = \frac{1}{2} \left( 1 + \ln(2\pi) + \ln(\tau_1^{-1} + \tau_2^{-1}) \right). \quad (2.134)$$

## 2.17

Let  $\Sigma$  be a matrix and

$$\begin{aligned} \mathbf{S} &= \frac{1}{2} \left( \Sigma^{-1} + (\Sigma^{-1})^\top \right), \\ \mathbf{A} &= \frac{1}{2} \left( \Sigma^{-1} - (\Sigma^{-1})^\top \right). \end{aligned} \quad (2.135)$$

Then

$$\Sigma^{-1} = \mathbf{S} + \mathbf{A}. \quad (2.136)$$

Therefore,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S} (\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.137)$$

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

$$\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top (\Sigma^{-1})^\top (\mathbf{x} - \boldsymbol{\mu}). \quad (2.138)$$

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

$$-\frac{1}{2} \left( \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)^\top (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.139)$$

Thus,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}) = 0. \quad (2.140)$$

Hence

$$(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.141)$$

## 2.18

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

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.142)$$

where  $i = 1, \dots, D$  and  $\mathbf{u}_i$  are unit vectors. Taking the inner product with  $\bar{\mathbf{u}}_i$  on both sides gives

$$\bar{\mathbf{u}}_i^\top \Sigma \mathbf{u}_i = \lambda_i. \quad (2.143)$$

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

$$\bar{\mathbf{u}}_i^\top \Sigma^\top \mathbf{u}_i = (\bar{\Sigma} \bar{\mathbf{u}}_i)^\top \mathbf{u}_i. \quad (2.144)$$

The right hand side can be writted as

$$\bar{\lambda}_i \bar{\mathbf{u}}_i^\top \mathbf{u}_i = \bar{\lambda}_i. \quad (2.145)$$

Therefore,

$$\lambda_i = \bar{\lambda}_i. \quad (2.146)$$

Additionally, for  $i \neq j$ , taking the inner product with  $\mathbf{u}_j$  on n both sides of the original equation gives

$$\mathbf{u}_j^\top \Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_j^\top \mathbf{u}_i. \quad (2.147)$$

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

$$\mathbf{u}_j^\top \Sigma^\top \mathbf{u}_i = (\Sigma \mathbf{u}_j)^\top \mathbf{u}_i. \quad (2.148)$$

The right hand side can be written as  $\lambda_j \mathbf{u}_j^\top \mathbf{u}_i$ . Therefore,

$$\lambda_i \mathbf{u}_j^\top \mathbf{u}_i = \lambda_j \mathbf{u}_j^\top \mathbf{u}_i. \quad (2.149)$$

Thus, if  $\lambda_i \neq \lambda_j$ , then

$$\mathbf{u}_j^\top \mathbf{u}_i = 0. \quad (2.150)$$

## 2.19

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

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.151)$$

where  $i = 1, \dots, D$  and  $\mathbf{u}_i$  are unit vectors. Let

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.152)$$

Then

$$\mathbf{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}. \quad (2.153)$$

By 2.18,

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}. \quad (2.154)$$

Therefore,

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top, \\ \mathbf{\Sigma}^{-1} &= \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\top, \end{aligned} \quad (2.155)$$

Thus,

$$\begin{aligned} \mathbf{\Sigma} &= \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^\top, \\ \mathbf{\Sigma}^{-1} &= \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\top. \end{aligned} \quad (2.156)$$

## 2.20

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

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.157)$$

where  $i = 1, \dots, D$  and  $\mathbf{u}_i$  are unit vectors. Let

$$\begin{aligned} \mathbf{\Lambda} &= \text{diag}(\lambda_1, \dots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.158)$$

By 2.19,

$$\mathbf{a}^\top \mathbf{\Sigma} \mathbf{a} = \mathbf{b}^\top \mathbf{\Lambda} \mathbf{b}, \quad (2.159)$$

where

$$\mathbf{b} = \mathbf{U}^\top \mathbf{a}. \quad (2.160)$$

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

$$\mathbf{a}^\top \mathbf{\Sigma} \mathbf{a} > 0 \quad (2.161)$$

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

$$\lambda_i > 0. \quad (2.162)$$

### 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

$$\Sigma \Lambda = \mathbf{I}. \quad (2.163)$$

Taking the transpose of the both sides gives

$$\Lambda^\top \Sigma = \mathbf{I}. \quad (2.164)$$

Therefore,

$$\Lambda^\top = \Lambda. \quad (2.165)$$

### 2.23

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

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.166)$$

where  $i = 1, \dots, D$  and  $\mathbf{u}_i$  are unit vectors. Let

$$\begin{aligned} \Lambda' &= \text{diag} \left( \lambda_1^{-\frac{1}{2}}, \dots, \lambda_D^{-\frac{1}{2}} \right), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.167)$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = \Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{U} \Lambda' \Lambda'^\top \mathbf{U}^\top (\mathbf{x}-\boldsymbol{\mu}) = \Delta} d\mathbf{x}. \quad (2.168)$$

By the transformation

$$\mathbf{y} = \Lambda'^\top \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu}) \quad (2.169)$$

and the property

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}, \quad (2.170)$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left| \det \left( \mathbf{U} \mathbf{\Lambda}'^{-1} \right) \right| d\mathbf{y} = |\det \mathbf{\Sigma}|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \quad (2.171)$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det \mathbf{\Sigma}|^{\frac{1}{2}} \Delta^D V_D, \quad (2.172)$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \quad (2.173)$$

## 2.24

Let

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$$

be a partitioned matrix where  $\mathbf{A}$  is a square matrix and  $\mathbf{D}$  is an invertible matrix. By an LDU decomposition, we have

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

Therefore,

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

Thus,

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

## 2.25

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}), \quad (2.174)$$

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

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \quad (2.175)$$

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

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\ = & -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\ & - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{bc}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\ & - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{ca}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{cb}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{cc}(\mathbf{x}_c - \boldsymbol{\mu}_c). \end{aligned} \quad (2.176)$$

Excluding the terms independent of  $\mathbf{x}_a$ , the right hand side can be written as

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\top \boldsymbol{\Sigma}_{a|b,c}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}), \quad (2.177)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c), \\ \boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \quad (2.178)$$

Therefore,

$$p(\mathbf{x}_a | \mathbf{x}_b, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}). \quad (2.179)$$

Multiplying both sides by  $p(\mathbf{x}_c)$  and integrating both sides with respect to  $\mathbf{x}_c$  gives

$$p(\mathbf{x}_a | \mathbf{x}_b) = \int \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}) p(\mathbf{x}_c) d\mathbf{x}_c. \quad (2.180)$$

Thus,

$$p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \quad (2.181)$$

where

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) + \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} \boldsymbol{\mu}_c, \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1}.\end{aligned}\tag{2.182}$$

## 2.26

We have

$$\begin{aligned}& \left( \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{D} \mathbf{A}^{-1} \right) (\mathbf{A} + \mathbf{B} \mathbf{C} \mathbf{D}) \\ &= \mathbf{I} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{D} + \mathbf{A}^{-1} \mathbf{B} \mathbf{C} \mathbf{D} \\ & \quad - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{D} \mathbf{A}^{-1} \mathbf{B} \mathbf{C} \mathbf{D}.\end{aligned}\tag{2.183}$$

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

$$\begin{aligned}& \mathbf{A}^{-1} \mathbf{B} \left( \mathbf{C} - (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} (\mathbf{I} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B} \mathbf{C}) \right) \mathbf{D} \\ &= \mathbf{A}^{-1} \mathbf{B} \left( \mathbf{C} - (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B}) \mathbf{C} \right) \mathbf{D}.\end{aligned}\tag{2.184}$$

The right hand side can be written as

$$\mathbf{A}^{-1} \mathbf{B} (\mathbf{C} - \mathbf{C}) \mathbf{D} = \mathbf{O}.\tag{2.185}$$

Therefore,

$$\left( \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{D} \mathbf{A}^{-1} \right) (\mathbf{A} + \mathbf{B} \mathbf{C} \mathbf{D}) = \mathbf{I}.\tag{2.186}$$

Thus,

$$(\mathbf{A} + \mathbf{B} \mathbf{C} \mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{D} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{D} \mathbf{A}^{-1}.\tag{2.187}$$

## 2.27

Let  $\mathbf{x}$  and  $\mathbf{z}$  be two variables. Then

$$\mathbf{E}(\mathbf{x} + \mathbf{z}) = \int \int (\mathbf{x} + \mathbf{z}) p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}.\tag{2.188}$$

The right hand side can be written as

$$\int \mathbf{x} \left( \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \right) d\mathbf{x} + \int \mathbf{z} \left( \int p(\mathbf{x}, \mathbf{z}) d\mathbf{x} \right) d\mathbf{z} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} + \int \mathbf{z} p(\mathbf{z}) d\mathbf{z}.\tag{2.189}$$



The right hand side can be written as  $E \mathbf{x} + E \mathbf{z}$ . Therefore,

$$E(\mathbf{x} + \mathbf{z}) = E \mathbf{x} + E \mathbf{z}. \quad (2.190)$$

Additionally,

$$\text{cov}(\mathbf{x} + \mathbf{z}) = \int \int (\mathbf{x} + \mathbf{z} - E(\mathbf{x} + \mathbf{z})) (\mathbf{x} + \mathbf{z} - E(\mathbf{x} + \mathbf{z}))^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \quad (2.191)$$

The right hand side can be written as

$$\begin{aligned} & \int \int (\mathbf{x} - E \mathbf{x}) (\mathbf{x} - E \mathbf{x})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} + \int \int (\mathbf{x} - E \mathbf{x}) (\mathbf{z} - E \mathbf{z})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ & + \int \int (\mathbf{z} - E \mathbf{z}) (\mathbf{x} - E \mathbf{x})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} + \int \int (\mathbf{z} - E \mathbf{z}) (\mathbf{z} - E \mathbf{z})^\top p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \end{aligned} \quad (2.192)$$

The first and fourth terms can be written as  $\text{cov } \mathbf{x}$  and  $\text{cov } \mathbf{z}$ . If  $\mathbf{x}$  and  $\mathbf{z}$  are independent, the second and third terms can be written as

$$\begin{aligned} & \int (\mathbf{x} - E \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int (\mathbf{z} - E \mathbf{z})^\top p(\mathbf{z}) d\mathbf{z} = \mathbf{O}, \\ & \int (\mathbf{z} - E \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \int (\mathbf{x} - E \mathbf{x})^\top p(\mathbf{x}) d\mathbf{x} = \mathbf{O}. \end{aligned} \quad (2.193)$$

Therefore,

$$\text{cov}(\mathbf{x} + \mathbf{z}) = \text{cov } \mathbf{x} + \text{cov } \mathbf{z}. \quad (2.194)$$

## 2.28

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

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

where

$$E \mathbf{z} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}$$

and

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

Then, by 2.29,

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

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

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ & + \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ = & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu}) + \\ & -\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}))^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})) \\ & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L} \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}). \end{aligned} \tag{2.195}$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{2.196}$$

Therefore,

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \mathbf{\Lambda}^{-1}), \\ p(\mathbf{y} | \mathbf{x}) &= \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}). \end{aligned} \tag{2.197}$$

## 2.29

Let

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

Then, by 2.24,

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

## 2.30

Let

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

Then

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

## 2.31

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

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \quad (2.198)$$

where

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}), \\ p(\mathbf{z}) &= \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}). \end{aligned} \quad (2.199)$$

By the definition,

$$p(\mathbf{y}) = \int p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (2.200)$$

The right hand side can be written as

$$\int \mathcal{N}(\mathbf{y} | \mathbf{x} + \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}) \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}) d\mathbf{x}. \quad (2.201)$$

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

$$-\frac{1}{2}(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}})^\top \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \quad (2.202)$$

The first and second order terms can be written as

$$-\mathbf{x}^\top (\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} - \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}}) + \mathbf{y}^\top \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} = \mathbf{u}^\top \mathbf{v} \quad (2.203)$$

and

$$-\frac{1}{2} \mathbf{x}^\top (\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}) \mathbf{x} + \frac{1}{2} \mathbf{x}^\top \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \mathbf{y} + \frac{1}{2} \mathbf{y}^\top \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{y}^\top \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \mathbf{y} = -\frac{1}{2} \mathbf{u}^\top \mathbf{R} \mathbf{u}, \quad (2.204)$$

respectively, where

$$\mathbf{u} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \mathbf{v} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} \\ \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} & -\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \\ -\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} & \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \end{bmatrix}.$$

Therefore, the logarithm of the integrand except the terms independent of  $\mathbf{u}$  can be written as

$$-\frac{1}{2} (\mathbf{u} - \mathbf{R}^{-1}\mathbf{v})^\top \mathbf{R} (\mathbf{u} - \mathbf{R}^{-1}\mathbf{v}), \quad (2.205)$$

where

$$\mathbf{R}^{-1} = \begin{bmatrix} \Sigma_{\mathbf{x}} & \Sigma_{\mathbf{x}} \\ \Sigma_{\mathbf{x}} & \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{z}} \end{bmatrix}, \mathbf{R}^{-1}\mathbf{v} = \begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{x}} + \mu_{\mathbf{z}} \end{bmatrix}.$$

by 2.29 and 2.30. Thus,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mu_{\mathbf{x}} + \mu_{\mathbf{z}}, \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{z}}). \quad (2.206)$$

## 2.32

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\ p(\mathbf{y} | \mathbf{x}) &= \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}). \end{aligned} \quad (2.207)$$

By the definition,

$$p(\mathbf{y} | \mathbf{x})p(\mathbf{x}) = p(\mathbf{x} | \mathbf{y})p(\mathbf{y}). \quad (2.208)$$

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

$$-\frac{1}{2} (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.209)$$

Since the first term can be written as

$$\begin{aligned} & -\frac{1}{2} (\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ &= -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L} \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ & \quad -\frac{1}{2} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}), \end{aligned} \quad (2.210)$$

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

$$\begin{aligned}
& -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\top (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) + \frac{1}{2} \mathbf{z}^\top (\mathbf{A}^\top \mathbf{L} \mathbf{A} + \boldsymbol{\Lambda}) \mathbf{z} \\
& -\frac{1}{2}(\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}) \\
& = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\top (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A}) (\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) \\
& -\frac{1}{2}(\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b})^\top \mathbf{M} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}),
\end{aligned} \tag{2.211}$$

where

$$\begin{aligned}
\mathbf{z} &= (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}), \\
\mathbf{M} &= \mathbf{L} - \mathbf{L} \mathbf{A} (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L}.
\end{aligned} \tag{2.212}$$

we have

$$\boldsymbol{\mu} + \mathbf{z} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu}). \tag{2.213}$$

By 2.26,

$$(\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} = \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \boldsymbol{\Lambda}^{-1}. \tag{2.214}$$

Therefore,

$$\mathbf{M} = (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1}. \tag{2.215}$$

Thus,

$$\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x} | (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu}), (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}), \\
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y} | \mathbf{A} \boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top).
\end{aligned} \tag{2.216}$$

## 2.33

Refer to 2.32.

## 2.34

Let  $\mathbf{X}$  be a set of  $N$  variables such that

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.217}$$

To maximise it with respect to  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ , setting the partial derivatives to zero gives

$$\begin{aligned}\mathbf{0} &= \sum_{n=1}^N (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x}_n - \boldsymbol{\mu}), \\ \mathbf{O} &= -\frac{N}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top.\end{aligned}\tag{2.218}$$

Therefore,

$$\begin{aligned}\boldsymbol{\mu}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top.\end{aligned}\tag{2.219}$$

## 2.35

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}).\tag{2.220}$$

Then

$$\mathbb{E} \mathbf{x} \mathbf{x}^\top = \int \mathbf{x} \mathbf{x}^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.\tag{2.221}$$

The right hand side can be written as

$$\begin{aligned}& \int (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &= \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} + \boldsymbol{\mu} \int (\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &+ \left( \int (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \right) \boldsymbol{\mu}^\top + \boldsymbol{\mu} \boldsymbol{\mu}^\top \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.\end{aligned}\tag{2.222}$$

Since

$$\begin{aligned}\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} &= 1, \\ \int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} &= \boldsymbol{\mu}, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} &= \boldsymbol{\Sigma},\end{aligned}\tag{2.223}$$

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

$$\mathbb{E} \mathbf{x}\mathbf{x}^\top = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.224)$$

Additionally, let  $\mathbf{x}_n$  and  $\mathbf{x}_m$  be variables such that

$$\begin{aligned} p(\mathbf{x}_n) &= \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \Sigma), \\ p(\mathbf{x}_m) &= \mathcal{N}(\mathbf{x}_m | \boldsymbol{\mu}, \Sigma). \end{aligned} \quad (2.225)$$

If  $n \neq m$ , then

$$\mathbb{E} \mathbf{x}_n \mathbf{x}_m^\top = \mathbb{E} \mathbf{x}_n \mathbb{E} \mathbf{x}_m^\top. \quad (2.226)$$

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

$$\mathbb{E} \mathbf{x}_n \mathbf{x}_m^\top = \delta_{nm} \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.227)$$

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

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \Sigma). \quad (2.228)$$

By 2.34,

$$\begin{aligned} \boldsymbol{\mu}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ \Sigma_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \end{aligned} \quad (2.229)$$

Then

$$\mathbb{E} \Sigma_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbb{E} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \quad (2.230)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} \sum_{n=1}^N \mathbb{E} \mathbf{x}_n \mathbf{x}_n^\top - \frac{1}{N^2} \sum_{n=1}^N \mathbb{E} \left( \sum_{n=1}^N \mathbf{x}_n \right) \mathbf{x}_n^\top - \frac{1}{N^2} \sum_{n=1}^N \mathbb{E} \mathbf{x}_n \left( \sum_{n=1}^N \mathbf{x}_n \right)^\top \\ & + \frac{1}{N^3} \sum_{n=1}^N \mathbb{E} \left( \sum_{n=1}^N \mathbf{x}_n \right) \left( \sum_{n=1}^N \mathbf{x}_n \right)^\top. \end{aligned} \quad (2.231)$$

The first term can be written as  $\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top$ . The second and third terms can be written as

$$-\frac{1}{N} ((\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top) + (N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\top) = -\frac{1}{N} \Sigma - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.232)$$

The fourth term can be written as

$$\frac{1}{N^2} (N (\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top) + N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\top) = \frac{1}{N}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.233)$$

Therefore,

$$\mathbb{E} \boldsymbol{\Sigma}_{\text{ML}} = \frac{N-1}{N}\boldsymbol{\Sigma}. \quad (2.234)$$

## 2.36

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

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (2.235)$$

Let us assume that  $\mu$  is known. Then, by 2.34,

$$\sigma_{\text{ML}}^{2(N)} = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2. \quad (2.236)$$

The right hand side can be written as

$$\frac{1}{N}(x_N - \mu)^2 + \frac{1}{N} \sum_{n=1}^{N-1} (x_n - \mu)^2 = \frac{1}{N}(x_N - \mu)^2 + \frac{N-1}{N} \sigma_{\text{ML}}^{2(N-1)}. \quad (2.237)$$

Therefore,

$$\sigma_{\text{ML}}^{2(N)} = \sigma_{\text{ML}}^{2(N-1)} + \frac{1}{N} \left( (x_N - \mu)^2 - \sigma_{\text{ML}}^{2(N-1)} \right). \quad (2.238)$$

Since

$$\frac{\partial}{\partial \sigma^2} (-\ln p(x_n | \sigma^2)) = \frac{1}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} (x_n - \mu)^2, \quad (2.239)$$

we have

$$\sigma_{\text{ML}}^{2(N)} = \sigma_{\text{ML}}^{2(N-1)} - \frac{\sigma_{\text{ML}}^{2(N-1)}}{N} \frac{\partial}{\partial \sigma_{\text{ML}}^{2(N-1)}} \left( -\ln p(x_N | \sigma_{\text{ML}}^{2(N-1)}) \right). \quad (2.240)$$



## 2.37

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

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.241)$$

Let us assume that  $\boldsymbol{\mu}$  is known. Then, by 2.34,

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \quad (2.242)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top + \frac{1}{N} \sum_{n=1}^{N-1} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top \\ &= \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top + \frac{N-1}{N} \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}. \end{aligned} \quad (2.243)$$

Therefore,

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} + \frac{1}{N} \left( (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top - \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} \right). \quad (2.244)$$

Since

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} (-\ln p(\mathbf{x}_N | \boldsymbol{\Sigma})) = -\frac{1}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top, \quad (2.245)$$

we have

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} - \frac{\boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}}{N} \frac{\partial}{\partial \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}} \left( -\ln p(\mathbf{x}_N | \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}) \right). \quad (2.246)$$

## 2.38

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

$$\begin{aligned} p(x_n | \mu) &= \mathcal{N}(x_n | \mu, \sigma^2), \\ p(\mu) &= \mathcal{N}(\mu | \mu_0, \sigma_0^2). \end{aligned} \quad (2.247)$$

By the definition,

$$p(\mu | \mathbf{x}) p(\mathbf{x}) = p(\mathbf{x} | \mu) p(\mu). \quad (2.248)$$

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

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2. \quad (2.249)$$

The first term can be written as

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}} + \mu_{\text{ML}} - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{N}{2\sigma^2} (\mu_{\text{ML}} - \mu)^2. \quad (2.250)$$

where

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n, \quad (2.251)$$

as derived in 2.34. Therefore, the logarithm except the terms independent of  $\mathbf{x}$  and  $\mu$  can be written as

$$\begin{aligned} & -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{N}{2\sigma^2} (\mu_{\text{ML}} - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \\ & = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{1}{2\sigma_N^2} (\mu - \mu_N)^2 + \frac{\mu_N^2}{2\sigma_N^2}, \end{aligned} \quad (2.252)$$

where

$$\begin{aligned} \mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\ \sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}. \end{aligned} \quad (2.253)$$

Therefore,

$$p(\mu|\mathbf{x}) = \mathcal{N} \left( \mu \mid \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2} \right). \quad (2.254)$$

### 2.39 (Incomplete)

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

$$\begin{aligned} p(x_n|\mu) &= \mathcal{N}(x_n|\mu, \sigma^2), \\ p(\mu) &= \mathcal{N}(\mu|\mu_0, \sigma_0^2). \end{aligned} \quad (2.255)$$

Then, by 2.38,

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2), \quad (2.256)$$

where

$$\begin{aligned} \mu_N &= \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{n=1}^N x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\ \sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}. \end{aligned} \quad (2.257)$$

Then

$$\begin{aligned} \mu_N &= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2} \mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} x_N + \frac{\sigma^2 - \sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\ \sigma_N^2 &= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2} \sigma_{N-1}^2. \end{aligned} \quad (2.258)$$

Additionally, we have

$$\begin{aligned} p(\mu|x_1, \dots, x_{N-1}) &= \mathcal{N}(\mu|\mu_{N-1}, \sigma_{N-1}^2), \\ p(x_N|\mu) &= \mathcal{N}(x_N|\mu, \sigma^2). \end{aligned} \quad (2.259)$$

Then,  $\ln p(\mu|x_1, \dots, x_{N-1}) + \ln p(x_N|\mu)$  except the terms independent of  $\mu$  or  $x_N$  can be written as

$$\begin{aligned} & -\frac{1}{2\sigma_{N-1}^2}(\mu - \mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N - \mu)^2 \\ &= -\frac{1}{2\frac{\sigma_{N-1}^2 \sigma^2}{\sigma_{N-1}^2 + \sigma^2}} \left( \mu - \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2} \mu_{N-1} - \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2} x_N \right)^2 \\ &+ \frac{1}{2\frac{\sigma_{N-1}^2 \sigma^2}{\sigma_{N-1}^2 + \sigma^2}} \left( \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2} \mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2} x_N \right) - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}. \end{aligned} \quad (2.260)$$

Therefoere,

$$\begin{aligned} \mu_N &= \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2} \mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2} x_N, \\ \sigma_N^2 &= \frac{\sigma_{N-1}^2 \sigma^2}{\sigma_{N-1}^2 + \sigma^2}. \end{aligned} \quad (2.261)$$

## 2.40

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

$$\begin{aligned} p(\mathbf{x}_n|\boldsymbol{\mu}) &= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ p(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}). \end{aligned} \quad (2.262)$$

By the definition,

$$p(\boldsymbol{\mu}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}). \quad (2.263)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{X}$  and  $\boldsymbol{\mu}$  can be written as

$$-\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0). \quad (2.264)$$

The first term can be written as

$$\begin{aligned} & -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}) \\ &= -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{N}{2} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}). \end{aligned} \quad (2.265)$$

where

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \quad (2.266)$$

as derived in 2.34. Therefore, the logarithm except the terms independent of  $\mathbf{X}$  and  $\boldsymbol{\mu}$  can be written as

$$\begin{aligned} & -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{N}{2} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}) \\ & -\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\ &= -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\top \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) \\ & \quad + \frac{1}{2} \boldsymbol{\mu}_N^\top \boldsymbol{\Sigma}_N^{-1} \boldsymbol{\mu}_N, \end{aligned} \quad (2.267)$$

where

$$\begin{aligned}\boldsymbol{\mu}_N &= (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} (N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0), \\ \boldsymbol{\Sigma}_N &= (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}.\end{aligned}\tag{2.268}$$

Therefore,

$$p(\boldsymbol{\mu}|\mathbf{X}) = \mathcal{N}\left(\boldsymbol{\mu} \mid (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} (N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0), (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\right).\tag{2.269}$$

## 2.41

By the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda).\tag{2.270}$$

Then

$$\int_0^\infty \text{Gam}(\lambda|a, b) d\lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a-1} \exp(-b\lambda) d\lambda.\tag{2.271}$$

By the transformation

$$\lambda' = b\lambda,\tag{2.272}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a-1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{\Gamma(a)} \int_0^\infty \lambda'^{a-1} \exp(-\lambda') d\lambda'.\tag{2.273}$$

The right hand side can be written as

$$\frac{1}{\Gamma(a)} \Gamma(a) = 1.\tag{2.274}$$

Therefore,

$$\int_0^\infty \text{Gam}(\lambda|a, b) d\lambda = 1.\tag{2.275}$$

## 2.42

Let  $\lambda$  be a variable such that

$$p(\lambda) = \text{Gam}(\lambda|a, b).\tag{2.276}$$

By the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.277)$$

Then

$$\text{E } \lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right) d\lambda. \quad (2.278)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.279)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b\Gamma(a)} \int_0^\infty \lambda'^a \exp(-\lambda') d\lambda'. \quad (2.280)$$

The right hand side can be written as

$$\frac{1}{b\Gamma(a)} \Gamma(a+1) = \frac{a}{b}. \quad (2.281)$$

Therefore,

$$\text{E } \lambda = \frac{a}{b}. \quad (2.282)$$

Additionally,

$$\text{E } \lambda^2 = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1} \exp\left(-\frac{\lambda}{b}\right) d\lambda. \quad (2.283)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.284)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b^2\Gamma(a)} \int_0^\infty \lambda'^{a+1} \exp(-\lambda') d\lambda'. \quad (2.285)$$

The right hand side can be written as

$$\frac{1}{b^2\Gamma(a)} \Gamma(a+2) = \frac{a(a+1)}{b^2}. \quad (2.286)$$

Therefore,

$$E \lambda^2 = \frac{a(a+1)}{b^2}. \quad (2.287)$$

By the definition,

$$\text{var } \lambda = E \lambda^2 - (E \lambda)^2. \quad (2.288)$$

Therefore,

$$\text{var } \lambda = \frac{a}{b^2}. \quad (2.289)$$

Finally, setting the derivative of  $\text{Gam}(\lambda|a, b)$  with respect to  $\lambda$  to zero gives

$$0 = \frac{b^a}{\Gamma(a)} \left( \frac{a-1}{\lambda} - b \right) \lambda^{a-1} \exp \left( -\frac{\lambda}{b} \right). \quad (2.290)$$

Therefore,

$$\text{mode } \lambda = \frac{a-1}{b}. \quad (2.291)$$

## 2.43

Let

$$p(x|\sigma^2, q) = \frac{q}{2\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \exp \left( -\frac{|x|^q}{2\sigma^2} \right). \quad (2.292)$$

Then

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = \frac{q}{\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \int_0^{\infty} \exp \left( -\frac{x^q}{2\sigma^2} \right) dx. \quad (2.293)$$

By the transformation

$$x' = \frac{x^q}{2\sigma^2}, \quad (2.294)$$

the right hand side can be written as

$$\begin{aligned} & \frac{q}{\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \int_0^{\infty} \exp(-x') (2\sigma^2)^{\frac{1}{q}} \frac{1}{q} x'^{\frac{1}{q}-1} dx' \\ &= \frac{1}{\Gamma(\frac{1}{q})} \int_0^{\infty} x'^{\frac{1}{q}-1} \exp(-x') dx'. \end{aligned} \quad (2.295)$$

The right hand side can be written as

$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma \left( \frac{1}{q} \right) = 1. \quad (2.296)$$

Therefore,

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = 1. \quad (2.297)$$

Additionally,

$$p(x|\sigma^2, 2) = \frac{1}{\Gamma(\frac{1}{2})} (2\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \quad (2.298)$$

Therefore,

$$p(x|\sigma^2, 2) = \mathcal{N}(x|0, \sigma^2). \quad (2.299)$$

Finally, let  $\mathbf{t} = (t_1, \dots, t_N)^\top$  and  $\mathbf{X} = \{x_1, \dots, x_N\}$  such that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \quad (2.300)$$

where

$$p(\epsilon_n) = p(\epsilon_n|\sigma^2, q). \quad (2.301)$$

Therefore, the logarithm of  $p(\epsilon_n)$  except the terms independent of  $\mathbf{w}$  and  $\sigma^2$  can be written as

$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q} \ln(2\sigma^2). \quad (2.302)$$

Thus, 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

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln(2\sigma^2). \quad (2.303)$$

## 2.44

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

$$\begin{aligned} p(x_n|\mu, \tau) &= \mathcal{N}(x_n|\mu, \tau^{-1}), \\ p(\mu, \tau) &= \mathcal{N}(\mu|\mu_0, (\beta\tau)^{-1}) \text{Gam}(\tau|a, b). \end{aligned} \quad (2.304)$$

By the definition,

$$p(\mu, \tau|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu, \tau)p(\mu, \tau). \quad (2.305)$$



The logarithm of the right hand side except the terms independent of  $\mathbf{x}$ ,  $\mu$  and  $\tau$  can be written as

$$\begin{aligned} & \frac{N}{2} \ln \tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \mu)^2 + \frac{1}{2} \ln \tau - \frac{\beta \tau}{2} (\mu - \mu_0)^2 + (a - 1) \ln \tau - b\tau \\ &= \left( a + \frac{N-1}{2} \right) \ln \tau - \frac{N\tau}{2} (\bar{x} - \mu)^2 - \frac{\beta \tau}{2} (\mu - \mu_0)^2 - b\tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \bar{x})^2, \end{aligned} \quad (2.306)$$

where

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.307)$$

Since

$$-\frac{N\tau}{2} (\bar{x} - \mu)^2 - \frac{\beta \tau}{2} (\mu - \mu_0)^2 = -\frac{(N + \beta)\tau}{2} \left( \mu - \frac{N\bar{x} + \beta\mu_0}{N + \beta} \right)^2 - \frac{N\beta\tau(\bar{x} - \mu_0)^2}{2(N + \beta)}, \quad (2.308)$$

the right hand side can be written as

$$-\frac{(N + \beta)\tau}{2} \left( \mu - \frac{N\bar{x} + \beta\mu_0}{N + \beta} \right)^2 + \left( a + \frac{N-1}{2} \right) \ln \tau - \left( b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N + \beta)} + \frac{1}{2} \sum_{n=1}^N (x_n - \bar{x})^2 \right) \tau. \quad (2.309)$$

Therefore,

$$\begin{aligned} p(\mu, \tau | \mathbf{x}) &= \mathcal{N} \left( \mu \mid \frac{N\bar{x} + \beta\mu_0}{N + \beta}, ((N + \beta)\tau)^{-1} \right) \\ &\quad \text{Gam} \left( \tau \mid a + \frac{N+1}{2}, b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N + \beta)} + \frac{1}{2} \sum_{n=1}^N (x_n - \bar{x})^2 \right). \end{aligned} \quad (2.310)$$

## 2.45

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.311)$$

Then

$$p(\mathbf{X} | \boldsymbol{\Lambda}) = \prod_{n=1}^N \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.312)$$

The right hand side except the terms independent of  $\mathbf{\Lambda}$  can be written as

$$(\det \mathbf{\Lambda})^{\frac{N}{2}} \exp \left( -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x}_n - \boldsymbol{\mu}) \right) = (\det \mathbf{\Lambda})^{\frac{N}{2}} \exp \left( -\frac{1}{2} \text{tr} (\mathbf{S} \mathbf{\Lambda}) \right), \quad (2.313)$$

where

$$\mathbf{S} = \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \quad (2.314)$$

Therefore,

$$p(\mathbf{X}|\mathbf{\Lambda}) \propto (\det \mathbf{\Lambda})^{\frac{N}{2}} \exp \left( -\frac{1}{2} \text{tr} (\mathbf{S} \mathbf{\Lambda}) \right). \quad (2.315)$$

Let us assume that a prior distribution of  $\mathbf{\Lambda}$  is given by

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu) (\det \mathbf{\Lambda})^{\frac{\nu-D-1}{2}} \exp \left( -\frac{1}{2} \text{tr} (\mathbf{W}^{-1} \mathbf{\Lambda}) \right). \quad (2.316)$$

Then, by the definition,

$$p(\mathbf{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) \propto p(\mathbf{X}|\mathbf{\Lambda}) \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu), \quad (2.317)$$

where the right hand side except the terms independent of  $\mathbf{\Lambda}$  can be written as

$$(\det \mathbf{\Lambda})^{\frac{\nu+N-D-1}{2}} \exp \left( -\frac{1}{2} \text{tr} ((\mathbf{W}^{-1} + \mathbf{S}) \mathbf{\Lambda}) \right). \quad (2.318)$$

Therefore,

$$p(\mathbf{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) = \mathcal{W} \left( \mathbf{\Lambda} \mid (\mathbf{W}^{-1} + \mathbf{S})^{-1}, \nu + N \right). \quad (2.319)$$

Thus,  $\mathcal{W}$  is a conjugate prior distribution of  $\mathbf{\Lambda}$ .

## 2.46

Let  $x$  be a variable such that

$$p(x|\mu, \tau, a, b) = \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b). \quad (2.320)$$

Then

$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau. \quad (2.321)$$

The right hand side can be written as

$$\begin{aligned} & \int_0^\infty \left( \frac{\tau}{2\pi} \right)^{\frac{1}{2}} \exp \left( -\frac{\tau}{2} (x - \mu)^2 \right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) d\tau \\ &= (2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp \left( -\left( b + \frac{(x - \mu)^2}{2} \right) \tau \right) d\tau. \end{aligned} \quad (2.322)$$

By the transformation

$$\tau' = \left( b + \frac{(x - \mu)^2}{2} \right) \tau, \quad (2.323)$$

the integral of the right hand side can be written as

$$\int_0^\infty \left( \frac{\tau'}{b + \frac{(x - \mu)^2}{2}} \right)^{a-\frac{1}{2}} \exp(-\tau') \frac{d\tau'}{b + \frac{(x - \mu)^2}{2}} = \Gamma \left( a + \frac{1}{2} \right) \left( b + \frac{(x - \mu)^2}{2} \right)^{-a-\frac{1}{2}}. \quad (2.324)$$

Therefore,

$$p(x|\mu, \tau, a, b) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left( b + \frac{(x - \mu)^2}{2} \right)^{-a-\frac{1}{2}}. \quad (2.325)$$

Let

$$\begin{aligned} \nu &= 2a, \\ \lambda &= \frac{a}{b}. \end{aligned} \quad (2.326)$$

Then

$$p(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( \frac{\lambda}{\pi\nu} \right)^{\frac{1}{2}} \left( 1 + \frac{\lambda(x - \mu)^2}{\nu} \right)^{-\frac{\nu+1}{2}}. \quad (2.327)$$

## 2.47

By the definition,

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left( \frac{\lambda}{\pi\nu} \right)^{\frac{1}{2}} \left( 1 + \frac{\lambda(x - \mu)^2}{\nu} \right)^{-\frac{\nu+1}{2}}. \quad (2.328)$$

By the transformation

$$y = \frac{\lambda(x - \mu)^2}{\nu}, \quad (2.329)$$

the right hand side except the terms independent of  $x$  can be written as

$$(1+y)^{-\frac{\lambda(x-\mu)^2}{2y}-\frac{1}{2}}. \quad (2.330)$$

In the limit  $y \rightarrow \infty$ , it becomes

$$\exp\left(-\frac{\lambda}{2}(x-\mu)^2\right). \quad (2.331)$$

Therefore, in the limit  $\nu \rightarrow \infty$ ,  $\text{St}(x|\mu, \lambda, \nu)$  becomes  $\mathcal{N}(x|\mu, \lambda^{-1})$ .

## 2.48

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

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right). \quad (2.332)$$

Then

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.333)$$

The right hand side can be written as

$$\begin{aligned} & \int_0^\infty (2\pi)^{-\frac{D}{2}} (\det(\eta\boldsymbol{\Lambda}))^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \eta\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right) \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \eta^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu}{2}\eta\right) d\eta \\ &= (2\pi)^{-\frac{D}{2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} (\det \boldsymbol{\Lambda})^{\frac{1}{2}} \int_0^\infty \eta^{\frac{D+\nu}{2}-1} \exp\left(-\frac{1}{2}(\nu + (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}))\eta\right) d\eta. \end{aligned} \quad (2.334)$$

By the transformation

$$\eta' = \frac{1}{2}(\nu + (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}))\eta, \quad (2.335)$$

the integral of the right hand side can be written as

$$\begin{aligned} & \int_0^\infty \left(\frac{2\eta'}{\nu + (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}-1} \exp(-\eta') \frac{2}{\nu + (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})} d\eta' \\ &= \left(\frac{2}{\nu + (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}} \Gamma\left(\frac{D+\nu}{2}\right). \end{aligned} \quad (2.336)$$

Therefore,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.337)$$

## 2.49

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

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu). \quad (2.338)$$

By the definition,

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.339)$$

First,

$$\mathbb{E} \mathbf{x} = \int \mathbf{x} \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \quad (2.340)$$

The right hand side can be written as

$$\begin{aligned} & \int \mathbf{x} \left( \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left( \int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \quad (2.341)$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \quad (2.342)$$

Therefore,

$$\mathbb{E} \mathbf{x} = \boldsymbol{\mu}. \quad (2.343)$$

Additionally,

$$\text{cov} \mathbf{x} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \quad (2.344)$$

The right hand side can be written as

$$\begin{aligned} & \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \left( \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left( \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \quad (2.345)$$

The right hand side can be written as

$$\int (\eta \mathbf{\Lambda})^{-1} \text{Gam} \left( \eta \mid \frac{\nu}{2}, \frac{\nu}{2} \right) d\eta = \mathbf{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \int \eta^{\frac{\nu}{2}-2} \exp \left( -\frac{\nu}{2} \eta \right) d\eta. \quad (2.346)$$

By the transformation

$$\eta' = \frac{\nu}{2} \eta, \quad (2.347)$$

the integral of the right hand side can be written as

$$\int \left( \frac{2}{\nu} \eta' \right)^{-\frac{\nu}{2}-2} \exp(-\eta') \frac{2}{\nu} d\eta' = \left( \frac{2}{\nu} \right)^{\frac{\nu}{2}-1} \Gamma \left( \frac{\nu}{2} - 1 \right). \quad (2.348)$$

Therefore, the right hand side can be written as

$$\mathbf{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left( \frac{2}{\nu} \right)^{\frac{\nu}{2}-1} \Gamma \left( \frac{\nu}{2} - 1 \right) = \frac{\frac{\nu}{2}}{\frac{\nu}{2} - 1} \mathbf{\Lambda}^{-1}. \quad (2.349)$$

Thus,

$$\text{cov } \mathbf{x} = \frac{\nu}{\nu - 2} \mathbf{\Lambda}^{-1}. \quad (2.350)$$

Finally, setting the derivative of  $\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)$  with respect to  $\mathbf{x}$  to zero gives

$$\mathbf{0} = -\frac{1}{2} (\mathbf{\Lambda} + \mathbf{\Lambda}^\top) (\mathbf{x} - \boldsymbol{\mu}) \int \eta \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta \mathbf{\Lambda})^{-1}) \text{Gam} \left( \eta \mid \frac{\nu}{2}, \frac{\nu}{2} \right) d\eta. \quad (2.351)$$

Therefore,

$$\text{mode } \mathbf{x} = \boldsymbol{\mu}. \quad (2.352)$$

## 2.50

By the definition,

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \mathbf{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left( 1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu} \right)^{-\frac{D+\nu}{2}}. \quad (2.353)$$

By the transformation

$$y = \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu}, \quad (2.354)$$

the right hand side except the terms independent of  $\mathbf{x}$  can be written as

$$(1+y)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2y} - \frac{D}{2}}. \quad (2.355)$$

In the limit  $y \rightarrow \infty$ , it becomes

$$\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right). \quad (2.356)$$

Therefore, in the limit  $\nu \rightarrow \infty$ ,  $\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$  becomes  $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ .

## 2.51

We have

$$\exp(iA) \exp(-iA) = 1. \quad (2.357)$$

The left hand side can be written as

$$(\cos A + i \sin A)(\cos A - i \sin A) = \cos^2 A + \sin^2 A. \quad (2.358)$$

Therefore,

$$\cos^2 A + \sin^2 A = 1. \quad (2.359)$$

Additionally,

$$\cos(A - B) = \text{Re}(\exp(i(A - B))). \quad (2.360)$$

The right hand side can be written as

$$\text{Re}(\exp(iA) \exp(-iB)) = \text{Re}((\cos A + i \sin A)(\cos B - i \sin B)). \quad (2.361)$$

The right hand side can be written as  $\cos A \cos B + \sin A \sin B$ . Therefore,

$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \quad (2.362)$$

Finally,

$$\sin(A - B) = \text{Im}(\exp(i(A - B))). \quad (2.363)$$

The right hand side can be written as

$$\text{Im}(\exp(iA) \exp(-iB)) = \text{Im}((\cos A + i \sin A)(\cos B - i \sin B)). \quad (2.364)$$

The right hand side can be written as  $\sin A \cos B - \cos A \sin B$ . Therefore,

$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \quad (2.365)$$

## 2.52 (Incomplete)

Let  $\theta$  be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.366)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \quad (2.367)$$

By the Taylor series

$$\cos \alpha = 1 - \frac{1}{2}\alpha^2 + O(\alpha^4) \quad (2.368)$$

and the transformation

$$\xi = m^{\frac{1}{2}}(\theta - \theta_0), \quad (2.369)$$

we have

$$\exp(m \cos(\theta - \theta_0)) = \exp\left(m \left(1 - \frac{1}{2}(\theta - \theta_0)^2 + O((\theta - \theta_0)^4)\right)\right). \quad (2.370)$$

## 2.53

Let  $\theta_0$  be a parameter such that

$$\sum_{n=1}^N \sin(\theta_n - \theta_0) = 0. \quad (2.371)$$

The left hand side can be written as

$$\sum_{n=1}^N (\sin \theta_n \cos \theta_0 - \cos \theta_n \sin \theta_0) = \cos \theta_0 \sum_{n=1}^N \sin \theta_n - \sin \theta_0 \sum_{n=1}^N \cos \theta_n. \quad (2.372)$$

Therefore,

$$\theta_0 = \arctan \left( \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right). \quad (2.373)$$



## 2.54

Let  $\theta$  be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.374)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \quad (2.375)$$

Setting the first and second derivatives with respect to  $\theta$  to zero gives

$$\begin{aligned} 0 &= -m \sin(\theta - \theta_0) p(\theta|\theta_0, m), \\ 0 &= (m^2 \sin^2(\theta - \theta_0) - m \cos(\theta - \theta_0)) p(\theta|\theta_0, m). \end{aligned} \quad (2.376)$$

Therefore,

$$\begin{aligned} \operatorname{argmax}_{\theta} p(\theta|\theta_0, m) &= \theta_0, \\ \operatorname{argmin}_{\theta} p(\theta|\theta_0, m) &= \theta_0 - \pi \operatorname{sgn}(\theta_0 - \pi). \end{aligned} \quad (2.377)$$

## 2.55

Let

$$\theta_0^{\text{ML}} = \arctan \left( \frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right). \quad (2.378)$$

Let

$$\begin{aligned} \bar{r} \cos \bar{\theta} &= \frac{1}{N} \sum_{n=1}^N \cos \theta_n, \\ \bar{r} \sin \bar{\theta} &= \frac{1}{N} \sum_{n=1}^N \sin \theta_n. \end{aligned} \quad (2.379)$$

Then

$$\theta_0^{\text{ML}} = \bar{\theta}. \quad (2.380)$$

Here,

$$\frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}}) = \left( \frac{1}{N} \sum_{n=1}^N \cos \theta_n \right) \cos \theta_0^{\text{ML}} + \left( \frac{1}{N} \sum_{n=1}^N \sin \theta_n \right) \sin \theta_0^{\text{ML}}. \quad (2.381)$$

By the result above, the right hand side can be written as

$$\bar{r} \cos^2 \bar{\theta} + \bar{r} \sin^2 \bar{\theta} = \bar{r}. \quad (2.382)$$

Therefore,

$$\frac{1}{N} \sum_{n=1}^N \cos(\theta_n - \theta_0^{\text{ML}}) = \bar{r}. \quad (2.383)$$

## 2.56

By the definition,

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \quad (2.384)$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp((a-1) \ln \mu + (b-1) \ln(1-\mu)) \quad (2.385)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}.$$

Additionally, by the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.386)$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \exp((a-1) \ln \lambda - b\lambda). \quad (2.387)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ -b \end{bmatrix}.$$

Finally, for

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.388)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta, \quad (2.389)$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m \cos \theta_0 \cos \theta + m \sin \theta_0 \sin \theta). \quad (2.390)$$

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,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right). \quad (2.391)$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})), \quad (2.392)$$

where

$$\begin{aligned} h(\mathbf{x}) &= (2\pi)^{-\frac{D}{2}}, \\ g(\boldsymbol{\eta}) &= (\det(-2\boldsymbol{\eta}_2))^{-\frac{1}{2}} \exp \left( \frac{1}{4} \boldsymbol{\eta}_1^\top \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1 \right), \\ \boldsymbol{\eta} &= \begin{bmatrix} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ -\frac{1}{2} \boldsymbol{\Sigma}^{-1} \end{bmatrix}, \\ \mathbf{u}(\mathbf{x}) &= \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \mathbf{x}^\top \end{bmatrix}. \end{aligned}$$

## 2.58

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

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})). \quad (2.393)$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} = 1 \quad (2.394)$$

with respect to  $\boldsymbol{\eta}$  gives

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x}) h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = \mathbf{0}. \quad (2.395)$$

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \int \mathbf{u}(\mathbf{x}) p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathbb{E} \mathbf{u}(\mathbf{x}). \quad (2.396)$$

Therefore,

$$\mathbb{E} \mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}. \quad (2.397)$$

Thus,

$$\mathbb{E} \mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}). \quad (2.398)$$

Taking the second derivative with respect to  $\boldsymbol{\eta}$  gives

$$\begin{aligned} \nabla \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} + 2\nabla g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} \\ + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = \mathbf{0}. \end{aligned} \quad (2.399)$$

The left hand side can be written as

$$\begin{aligned} \frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \frac{2\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int \mathbf{u}(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \int \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} \\ = \frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2 \mathbb{E} \mathbf{u}(\mathbf{x}) \mathbb{E} \mathbf{u}(\mathbf{x})^\top + \mathbb{E} (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top). \end{aligned} \quad (2.400)$$

Therefore,

$$\mathbb{E} (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top) = -\frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\top}{g^2(\boldsymbol{\eta})}. \quad (2.401)$$

By the definition,

$$\text{cov} \mathbf{u}(\mathbf{x}) = \mathbb{E} (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top) - \mathbb{E} \mathbf{u}(\mathbf{x}) \mathbb{E} \mathbf{u}(\mathbf{x})^\top. \quad (2.402)$$

Thus,

$$\text{cov} \mathbf{u}(\mathbf{x}) = -\frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\top}{g^2(\boldsymbol{\eta})}. \quad (2.403)$$

Hence,

$$\text{cov} \mathbf{u}(\mathbf{x}) = -\nabla \nabla \ln g(\boldsymbol{\eta}). \quad (2.404)$$

## 2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right). \quad (2.405)$$

Then

$$\int p(x|\sigma) dx = \frac{1}{\sigma} \int f\left(\frac{x}{\sigma}\right) dx. \quad (2.406)$$

By the transformation

$$x' = \frac{x}{\sigma}, \quad (2.407)$$

the right hand side can be written as

$$\frac{1}{\sigma} \int f(x') \sigma dx' = \int f(x') dx'. \quad (2.408)$$

Therefore,  $p(x|\sigma)$  will be normalised if  $f(x)$  is normalised.

## 2.60

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

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \quad (2.409)$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \quad (2.410)$$

Since

$$\int p(\mathbf{x}) d\mathbf{x} = 1, \quad (2.411)$$

we have

$$\sum_i h_i \Delta_i = 1. \quad (2.412)$$

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

$$\ln \left( \prod_i h_i^{n_i} \right) = \sum_i n_i \ln h_i, \quad (2.413)$$

where

$$\sum_i n_i = N. \quad (2.414)$$

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda \left( \sum_i h_i \Delta_i - 1 \right) \quad (2.415)$$

with respect to  $h_i$  and  $\lambda$  to zero gives

$$\begin{aligned} \frac{n_i}{h_i} + \lambda \Delta_i &= 0, \\ \sum_i h_i \Delta_i - 1 &= 0. \end{aligned} \quad (2.416)$$

Therefore,

$$\begin{aligned} \lambda &= -N, \\ h_i &= \frac{n_i}{N \Delta_i}. \end{aligned} \quad (2.417)$$

Thus, 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

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})}, \quad (2.418)$$

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{x}_{(K)} - \mathbf{x}\|} d\mathbf{x}', \quad (2.419)$$

$K$  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

$$\frac{1 - \exp(-2a)}{1 + \exp(-2a)} = \frac{2}{1 + \exp(-2a)} - 1. \quad (3.2)$$

By the definition

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \quad (3.3)$$

the right hand side can be written as

$$\tanh a = 2\sigma(2a) - 1. \quad (3.4)$$

Let

$$y(x_n, \mathbf{w}) = w_0 + \sum_{j=1}^M w_j \sigma\left(\frac{x - \mu_j}{s}\right). \quad (3.5)$$

By the result above, the right hand side can be written as

$$w_0 + \sum_{j=1}^M w_j \frac{\tanh\left(\frac{x - \mu_j}{2s}\right) + 1}{2} = w_0 + \frac{1}{2} \sum_{j=1}^M w_j + \frac{1}{2} \sum_{j=1}^M w_j \tanh\left(\frac{x - \mu_j}{2s}\right). \quad (3.6)$$

Therefore,  $y(x_n, \mathbf{w})$  is equivalent to

$$y(x_n, \mathbf{u}) = u_0 + \sum_{j=1}^M u_j \tanh\left(\frac{x - \mu_j}{2s}\right), \quad (3.7)$$

where

$$\begin{aligned} u_0 &= w_0 + \frac{1}{2} \sum_{j=1}^M w_j, \\ u_j &= \frac{w_j}{2}, \quad j = 1, \dots, M. \end{aligned} \quad (3.8)$$

### 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

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N r_n (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2. \quad (3.11)$$

Setting the derivative to zero gives

$$\mathbf{0} = - \sum_{n=1}^N r_n \phi(\mathbf{x}_n) (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)). \quad (3.12)$$

The right hand side can be written as

$$- \sum_{n=1}^N r_n t_n \phi(\mathbf{x}_n) + \left( \sum_{n=1}^N r_n \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^\top \right) \mathbf{w}. \quad (3.13)$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} E_D(\mathbf{w}) = (\Phi'^\top \Phi')^{-1} \Phi'^\top \mathbf{t}', \quad (3.14)$$

where

$$\Phi' = \begin{bmatrix} \sqrt{r_1} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \sqrt{r_N} \phi(\mathbf{x}_N)^\top \end{bmatrix}, \mathbf{t}' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \vdots \\ \sqrt{r_N} t_N \end{bmatrix}.$$



### 3.4 (Incomplete)

Let

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(\mathbf{x}_n, \mathbf{w}) - t_n)^2, \quad (3.15)$$

where

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{i=1}^D w_i (x_i + \epsilon_i), \\ p(\epsilon_i) &= \mathcal{N}(\epsilon_i | 0, \sigma^2). \end{aligned} \quad (3.16)$$

Setting the derivative of  $E_D(\mathbf{w})$  to zero gives

$$\mathbf{0} = \sum_{n=1}^N \begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n).$$

The right hand side can be written as

$$\sum_{n=1}^N \begin{bmatrix} 1 \\ \mathbf{x}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n) + \sum_{n=1}^N \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n).$$

### 3.5

Let

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2. \quad (3.17)$$

Then, the minimisation of  $E_D(\mathbf{w})$  under the constraint

$$\sum_{j=1}^M |w_j|^q \leq \eta \quad (3.18)$$

reduces to the minimisation of

$$E_D(\mathbf{w}) + \lambda \left( \sum_{j=1}^M |w_j|^q - \eta \right) \quad (3.19)$$

with respect to  $\mathbf{w}$  and  $\lambda$ . Then,

$$\eta = \sum_{j=1}^M |w_j^*(\lambda)|^q, \quad (3.20)$$

where

$$\mathbf{w}^*(\lambda) = \underset{\mathbf{w}}{\operatorname{argmin}} \left( E_D(\mathbf{w}) + \lambda \left( \sum_{j=1}^M |w_j|^q - \eta \right) \right). \quad (3.21)$$

### 3.6

Let  $\mathbf{t}_1, \dots, \mathbf{t}_N$  be a variable in  $D$  dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}, \Sigma) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{W}), \Sigma), \quad (3.22)$$

where

$$\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^\top \phi(\mathbf{x}_n). \quad (3.23)$$

Then

$$\begin{aligned} & \ln \left( \prod_{n=1}^N p(\mathbf{t}_n | \mathbf{W}, \Sigma) \right) \\ &= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln(\det \Sigma) - \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \phi(\mathbf{x}_n))^\top \Sigma^{-1} (\mathbf{t}_n - \mathbf{W}^\top \phi(\mathbf{x}_n)). \end{aligned} \quad (3.24)$$

Setting the derivatives with respect to  $\mathbf{W}$  and  $\Sigma$  to zero gives

$$\begin{aligned} \mathbf{O} &= -\frac{1}{2} (\Sigma^{-1} + (\Sigma^{-1})^\top) \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \phi(\mathbf{x}_n)) (\phi(\mathbf{x}_n))^\top, \\ \mathbf{O} &= -\frac{N}{2} (\Sigma^{-1})^\top + \frac{1}{2} (\Sigma^{-1})^2 \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \phi(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}^\top \phi(\mathbf{x}_n))^\top. \end{aligned} \quad (3.25)$$

Therefore,

$$\begin{aligned} \mathbf{W}_{\text{ML}} &= (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{t}, \\ \Sigma_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}_{\text{ML}}^\top \phi(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}_{\text{ML}}^\top \phi(\mathbf{x}_n))^\top. \end{aligned} \quad (3.26)$$

where

$$\Phi = \begin{bmatrix} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \phi(\mathbf{x}_N)^\top \end{bmatrix}.$$

### 3.7

Let  $t_1, \dots, t_N$  be a variable such that

$$\begin{aligned} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.27)$$

By the definition,

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \beta)p(\mathbf{t}) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}). \quad (3.28)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{t}$  and  $\mathbf{w}$  can be written as

$$\begin{aligned} & -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\ & = -\frac{\beta}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^\top (\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0), \end{aligned} \quad (3.29)$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\top \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\top \end{bmatrix}.$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + \frac{1}{2}\mathbf{m}_N^\top \mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\mathbf{t}^\top \mathbf{t} - \frac{1}{2}\mathbf{m}_0^\top \mathbf{S}_0^{-1}\mathbf{m}_0 \quad (3.30)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.31)$$

Therefore,

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N). \quad (3.32)$$

### 3.8

Let

$$\begin{aligned} p(\mathbf{w}|\mathbf{X}_N, \mathbf{t}_N, \beta) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \\ p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}, \beta) &= \mathcal{N}(t_{N+1}|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_{N+1}), \beta^{-1}), \end{aligned} \quad (3.33)$$

where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi_N^\top \mathbf{t}_N), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \Phi_N^\top \Phi_N.\end{aligned}\tag{3.34}$$

By the definition,

$$p(\mathbf{w}|\mathbf{X}_{N+1}, \mathbf{t}_{N+1}, \beta) \propto p(\mathbf{w}|\mathbf{X}_N, \mathbf{t}_N, \beta) p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}, \beta).\tag{3.35}$$

The logarithm of the right hand side except the terms independent of  $\mathbf{X}_{N+1}$  and  $\mathbf{w}$  can be written as

$$\begin{aligned}& -\frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) - \frac{\beta}{2} (t_{N+1} - \mathbf{w}^\top \phi(\mathbf{x}_{N+1}))^2 \\ &= -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu}_{N+1})^\top \boldsymbol{\Lambda}_{N+1} (\mathbf{w} - \boldsymbol{\mu}_{N+1}) + \frac{1}{2} \boldsymbol{\mu}_{N+1}^\top \boldsymbol{\Lambda}_{N+1} \boldsymbol{\mu}_{N+1} \\ & \quad - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} t_{N+1}^2,\end{aligned}\tag{3.36}$$

where

$$\begin{aligned}\boldsymbol{\mu}_{N+1} &= \boldsymbol{\Lambda}_{N+1}^{-1} (\mathbf{S}_N^{-1} \mathbf{m}_N + \beta t_{N+1} \phi(\mathbf{x}_{N+1})), \\ \boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_N^{-1} + \beta \phi(\mathbf{x}_{N+1}) \phi(\mathbf{x}_{N+1})^\top.\end{aligned}\tag{3.37}$$

Therefore,

$$\begin{aligned}\boldsymbol{\mu}_{N+1} &= \mathbf{m}_{N+1}, \\ \boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_{N+1}^{-1}.\end{aligned}\tag{3.38}$$

Thus,

$$p(\mathbf{w}|\mathbf{X}_{N+1}, \mathbf{t}_{N+1}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}).\tag{3.39}$$

### 3.9 (Incomplete)

Let

$$\begin{aligned}p(\mathbf{w}|\mathbf{X}_N, \mathbf{t}_N, \beta) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \\ p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}, \beta) &= \mathcal{N}(t_{N+1}|\mathbf{w}^\top \phi(\mathbf{x}_{N+1}), \beta^{-1}),\end{aligned}\tag{3.40}$$

where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi_N^\top \mathbf{t}_N), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \Phi_N^\top \Phi_N.\end{aligned}\tag{3.41}$$

By the definition,

$$p(\mathbf{w}|\mathbf{X}_{N+1}, \mathbf{t}_{N+1}, \beta) \propto p(\mathbf{w}|\mathbf{X}_N, \mathbf{t}_N, \beta) p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}, \beta).\tag{3.42}$$

The logarithm of the right hand side except the terms independent of  $\mathbf{X}_{N+1}$  and  $\mathbf{w}$  can be written as

$$-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) - \frac{\beta}{2}(t_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_{N+1}))^2. \quad (3.43)$$

It can be written as

$$-\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \boldsymbol{\phi}(\mathbf{x}_{N+1})^\top & -\beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \\ -\beta \boldsymbol{\phi}(\mathbf{x}_{N+1})^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_N^{-1} \mathbf{m}_N \\ 0 \end{bmatrix} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N.$$

### 3.10

Let  $t$  be a variable such that

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{w}) &= \mathcal{N}(t | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}), \\ p(\mathbf{w}|\mathbf{t}) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N). \end{aligned} \quad (3.44)$$

By the definition, we have

$$p(t|\mathbf{x}, \mathbf{t}) = \int p(t|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{t}) d\mathbf{w}. \quad (3.45)$$

The logarithm of the integrand of the right hand side except the terms independent of  $t$  and  $\mathbf{w}$  can be written as

$$-\frac{\beta}{2}(t - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}))^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \quad (3.46)$$

It can be written as

$$-\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x})^\top & -\beta \boldsymbol{\phi}(\mathbf{x}) \\ -\beta \boldsymbol{\phi}(\mathbf{x})^\top & \beta \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_N^{-1} \mathbf{m}_N \\ 0 \end{bmatrix} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x})^\top & -\beta \boldsymbol{\phi}(\mathbf{x}) \\ -\beta \boldsymbol{\phi}(\mathbf{x})^\top & \beta \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}_N & \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}) \\ \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N & \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}.$$

Therefore,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x})^\top & -\beta \boldsymbol{\phi}(\mathbf{x}) \\ \beta \boldsymbol{\phi}(\mathbf{x})^\top & \beta \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_N^{-1} \mathbf{m}_N \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{m}_N \\ \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}.$$

Thus,

$$p(t|\mathbf{x}, \mathbf{t}) = \mathcal{N}(t | \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})), \quad (3.47)$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \quad (3.48)$$

### 3.11

Let  $t$  be a variable such that

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{w}) &= \mathcal{N}(t|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}), \\ p(\mathbf{w}|\mathbf{t}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \end{aligned} \quad (3.49)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}_N^\top \mathbf{t}_N), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N. \end{aligned} \quad (3.50)$$

Then, by 3.10,

$$p(t|\mathbf{x}, \mathbf{t}) = \mathcal{N}(t | \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})), \quad (3.51)$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \quad (3.52)$$

Then

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top (\mathbf{S}_N - \mathbf{S}_{N+1}) \boldsymbol{\phi}(\mathbf{x}). \quad (3.53)$$

By the expression of  $\mathbf{S}_N$  above,

$$\mathbf{S}_{N+1} = (\mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \boldsymbol{\phi}(\mathbf{x}_{N+1})^\top)^{-1}. \quad (3.54)$$

By the identity

$$(\mathbf{M} + \mathbf{v} \mathbf{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1} \mathbf{v}) (\mathbf{v}^\top \mathbf{M}^{-1})}{1 + \mathbf{v}^\top \mathbf{M}^{-1} \mathbf{v}}, \quad (3.55)$$

the right hand side can be written as

$$\mathbf{S}_N - \frac{\beta (\mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1})) (\boldsymbol{\phi}(\mathbf{x}_{N+1})^\top \mathbf{S}_N)}{1 + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1})}. \quad (3.56)$$

Therefore,

$$\boldsymbol{\phi}(\mathbf{x})^\top (\mathbf{S}_N - \mathbf{S}_{N+1}) \boldsymbol{\phi}(\mathbf{x}) = \frac{\beta (\boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1}))^2}{1 + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_{N+1})}. \quad (3.57)$$

Thus,

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}). \quad (3.58)$$

### 3.12

Let  $t_1, \dots, t_N$  be variables such that

$$\begin{aligned} p(t_n | \mathbf{x}_n, \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0), \end{aligned} \quad (3.59)$$

where  $\mathbf{w}$  and  $\boldsymbol{\phi}$  are vectors in  $M$  dimensions. By the definition,

$$p(\mathbf{w}, \beta | \mathbf{t}) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}, \beta). \quad (3.60)$$

The logarithm of the right hand side except the terms independent of  $\mathbf{t}$ ,  $\mathbf{w}$  and  $\beta$  can be written as

$$\begin{aligned} & -\frac{N}{2} \ln \beta - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{M}{2} \ln \beta - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ & - (a_0 - 1) \ln \beta - b_0 \beta \\ = & - \left( a_0 + \frac{N+M}{2} - 1 \right) \ln \beta - \frac{\beta}{2} \mathbf{w}^\top (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \mathbf{w} + \beta \mathbf{w}^\top (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}) \\ & - \frac{\beta}{2} \mathbf{t}^\top \mathbf{t} - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 - b_0 \beta. \end{aligned} \quad (3.61)$$

The right hand side can be written as

$$-\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) - (a_N - 1) \ln \beta - b_N \beta, \quad (3.62)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S}_N &= (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \\ a_N &= a_0 + \frac{N+M}{2}, \\ b_N &= b_0 + \frac{1}{2} \mathbf{t}^\top \mathbf{t} + \frac{1}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N. \end{aligned} \quad (3.63)$$

Therefore,

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \text{Gam}(\beta | a_N, b_N). \quad (3.64)$$

### 3.13 (Incomplete)