

Solutions Manual to Pattern Recognition and Machine Learning

Hiromichi Inawashiro

May 20, 2025

Contents

1	Introduction	1
2	Probability Distributions	42
3	Linear Models for Regression	101
4	Linear Models for Classification	126
5	Neural Networks	152

1 Introduction

1.1

Let

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

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

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

If

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

then

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

Therefore,

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

Thus,

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

where

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

If

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

then

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

1.2

Let

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

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

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

If

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

then

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

Therefore,

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

Thus,

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

where

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

If

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

then

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

1.3

Let a , o and l be the events where an apple, orange and lime are selected respectively.

(a)

The probability that an apple is selected is given by

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \tag{1.17}$$

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

$$p(a) = \frac{17}{50}. \tag{1.18}$$

(b)

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

$$p(g|o) = \frac{p(g,o)}{p(o)}. \tag{1.19}$$

Here,

$$\begin{aligned} p(g,o) &= p(o|g)p(g), \\ p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g). \end{aligned} \tag{1.20}$$

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

$$\begin{aligned} p(g, o) &= \frac{9}{50}, \\ p(o) &= \frac{9}{25} \end{aligned} \tag{1.21}$$

Therefore,

$$p(g|o) = \frac{1}{2}. \tag{1.22}$$

1.4

Let

$$x = g(y) \tag{1.23}$$

and \hat{x} and \hat{y} be the locations of the maximum of $p_x(x)$ and $p_y(y)$ respectively.

(a)

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

$$p_y(y) = p_x(g(y)) |g'(y)| \tag{1.24}$$

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

$$0 = g'(\hat{y}) p'_x(g(\hat{y})) + p_x(g(\hat{y})) g''(\hat{y}). \tag{1.25}$$

Therefore, in general,

$$\hat{x} \neq g(\hat{y}). \tag{1.26}$$

(b)

Let us assume that

$$g(y) = ay + b. \tag{1.27}$$

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

$$0 = p'_x(g(\hat{y})). \tag{1.28}$$

Therefore,

$$\hat{x} = g(\hat{y}). \tag{1.29}$$

1.5

By the definition,

$$\text{var } f(x) = E(f(x) - E f(x))^2. \quad (1.30)$$

The right hand side can be written as

$$E((f(x))^2 - 2f(x) E f(x) + (E f(x))^2) = E(f(x))^2 - (E f(x))^2. \quad (1.31)$$

Therefore,

$$\text{var } f(x) = E(f(x))^2 - (E f(x))^2. \quad (1.32)$$

1.6

By the definition,

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

The right hand side can be written as

$$E xy - E(x E y) - E(y E x) + E(E x E y) = E xy - E x E y. \quad (1.34)$$

The right hand side can be written as

$$\int xyp(x, y)dxdy - \int xp(x)dx \int yp(y)dy. \quad (1.35)$$

If x and y are independent, by the definition,

$$f(x, y) = f(x)f(y). \quad (1.36)$$

Then,

$$\int xyp(x, y)dxdy = \int p(x)dx \int p(y)dy. \quad (1.37)$$

Therefore,

$$\text{cov}(x, y) = 0. \quad (1.38)$$

1.7

(a)

Let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx. \quad (1.39)$$

Then,

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2 + y^2)\right) dx dy. \quad (1.40)$$

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

$$\int_0^{\infty} \int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} dr d\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr. \quad (1.41)$$

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

$$2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{1}{2}s^2\right) s ds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_0^{\infty}. \quad (1.42)$$

Therefore,

$$I = (2\pi\sigma^2)^{\frac{1}{2}}. \quad (1.43)$$

(b)

By the definition,

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.44)$$

Then,

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

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

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}t^2\right) dt = (2\pi\sigma^2)^{-\frac{1}{2}} I. \quad (1.46)$$

Therefore,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1. \quad (1.47)$$

1.8

(a)

Let x be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.48)$$

Then

$$\mathbb{E} x = \int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.49)$$

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

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

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

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (y + \mu) \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy. \quad (1.51)$$

Since

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = 0, \quad (1.52)$$

and

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mu \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = \mu \int_{-\infty}^{\infty} \mathcal{N}(y|\mu, \sigma^2) dy, \quad (1.53)$$

we have

$$\mathbb{E} x = \mu. \quad (1.54)$$

(b)

The property

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1 \quad (1.55)$$

can be written as

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

Taking the derivative with respect to σ^2 gives

$$\begin{aligned} & (2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) (\sigma^2)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ & + (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} (\sigma^2)^{-2} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 0. \end{aligned} \quad (1.57)$$

The left hand side can be written as

$$\begin{aligned} & -\frac{1}{2} (\sigma^2)^{-1} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2} (\sigma^2)^{-2} \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx \\ & = -\frac{1}{2} (\sigma^2)^{-1} + \frac{1}{2} (\sigma^2)^{-2} \text{var } x. \end{aligned} \quad (1.58)$$

Therefore,

$$\text{var } x = \sigma^2. \quad (1.59)$$

1.9

(a)

Let x be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.60)$$

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

$$0 = (2\pi\sigma^2)^{-\frac{1}{2}} \left(-\frac{1}{\sigma^2}(x-\mu)\right) \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \quad (1.61)$$

Therefore,

$$\text{mode } x = \mu. \quad (1.62)$$

(b)

Let \mathbf{x} be a variable such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (1.63)$$

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

$$\mathbf{0} = -(2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x} - \boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \quad (1.64)$$

Therefore,

$$\text{mode } \mathbf{x} = \boldsymbol{\mu}. \quad (1.65)$$

1.10

(a)

By the definition,

$$\mathbb{E}(x + y) = \int \int (x + y)p(x, y)dxdy. \quad (1.66)$$

The right hand side can be written as

$$\int x \left(\int p(x, y)dy \right) dx + \int y \left(\int p(x, y)dx \right) dy = \int xp(x)dx + \int yp(y)dy. \quad (1.67)$$

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

$$\mathbb{E} x + \mathbb{E} y. \quad (1.68)$$

Therefore,

$$\mathbb{E}(x + y) = \mathbb{E} x + \mathbb{E} y. \quad (1.69)$$

(b)

By the definition,

$$\text{var}(x + y) = \mathbb{E} (x + y - \mathbb{E}(x + y))^2 \quad (1.70)$$

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

$$\begin{aligned} & \mathbb{E} (x - \mathbb{E} x)^2 + 2 \mathbb{E} ((x - \mathbb{E} x)(y - \mathbb{E} y)) + \mathbb{E} (y - \mathbb{E} y)^2 \\ &= \text{var } x + 2 \text{cov}(x, y) + \text{var } y. \end{aligned} \quad (1.71)$$

By 1.6, if x and y are independent, then

$$\text{cov}(x, y) = 0. \quad (1.72)$$

Therefore,

$$\text{var}(x + y) = \text{var } x + \text{var } y. \quad (1.73)$$

1.11

Let x_1, \dots, x_N be variables such that

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (1.74)$$

Then,

$$\ln p(\mathbf{x}) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2. \quad (1.75)$$

Setting the derivatives with respect to μ and σ^2 to zero gives

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu), \\ 0 &= -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{n=1}^N (x_n - \mu)^2. \end{aligned} \quad (1.76)$$

Therefore, the maximum likelihood solutions for μ and σ^2 are given by

$$\begin{aligned} \mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N x_n, \\ \sigma_{\text{ML}}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2. \end{aligned} \quad (1.77)$$

1.12

(a)

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

$$\begin{aligned} p(x_n) &= \mathcal{N}(x_n | \mu, \sigma^2), \\ p(x_{n'}) &= \mathcal{N}(x_{n'} | \mu, \sigma^2). \end{aligned} \quad (1.78)$$

Then,

$$\mathbb{E} x_n x_{n'} = \mu^2. \quad (1.79)$$

By the property

$$\mathbb{E} x_n^2 = \text{var } x_n + (\mathbb{E} x_n)^2, \quad (1.80)$$

we have

$$\mathbb{E} x_n^2 = \sigma^2 + \mu^2. \quad (1.81)$$

Therefore,

$$\mathbb{E} x_n x_{n'} = \mu^2 + I_{nn'} \sigma^2. \quad (1.82)$$

(b)

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

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (1.83)$$

Then, by 1.11, the maximum likelihood solution for μ is given by

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (1.84)$$

Then,

$$\mathbb{E} \mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n. \quad (1.85)$$

Therefore,

$$\mathbb{E} \mu_{\text{ML}} = \mu. \quad (1.86)$$

Similarly, by 1.11, the maximum likelihood solution for σ^2 is given by

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2. \quad (1.87)$$

Then,

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n - \mu_{\text{ML}})^2. \quad (1.88)$$

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n^2 - 2\mu_{\text{ML}} x_n + \mu_{\text{ML}}^2) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n^2 - \frac{2}{N} \mathbb{E} \left(\mu_{\text{ML}} \left(\sum_{n=1}^N x_n \right) \right) + \mathbb{E} \mu_{\text{ML}}^2. \quad (1.89)$$

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

$$\frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) = \mu^2 + \sigma^2, \quad (1.90)$$

while the second and third terms can be written as

$$-2 \mathbb{E} \mu_{\text{ML}}^2 + \mathbb{E} \mu_{\text{ML}}^2 = -\mathbb{E} \mu_{\text{ML}}^2. \quad (1.91)$$

Here,

$$\mathbb{E} \mu_{\text{ML}}^2 = \mathbb{E} \left(\frac{1}{N} \sum_{n=1}^N x_n \right)^2. \quad (1.92)$$

The right hand side can be written as

$$\frac{1}{N^2} \sum_{n=1}^N \mathbb{E} x_n^2 + \frac{2}{N^2} \sum_{1 \leq n < n' \leq N} \mathbb{E} x_n x_{n'} = \frac{1}{N} (\mu^2 + \sigma^2) + \frac{N-1}{N} \mu^2. \quad (1.93)$$

Therefore,

$$\mathbb{E} \mu_{\text{ML}}^2 = \mu^2 + \frac{1}{N} \sigma^2. \quad (1.94)$$

Thus,

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{N-1}{N} \sigma^2. \quad (1.95)$$

1.13

Let x_1, \dots, x_N be variables such that

$$\begin{aligned} \mathbb{E} x_n &= \mu, \\ \text{var } x_n &= \sigma^2. \end{aligned} \quad (1.96)$$

Here,

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

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \text{var } x_n = \sigma^2. \quad (1.98)$$

Therefore,

$$\mathbb{E} \left(\frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \sigma^2. \quad (1.99)$$

1.14

(a)

Let

$$\begin{aligned} w_{dd'}^S &= \frac{1}{2}(w_{dd'} + w_{d'd}), \\ w_{dd'}^A &= \frac{1}{2}(w_{dd'} - w_{d'd}). \end{aligned} \quad (1.100)$$

Then,

$$\begin{aligned} w_{dd'} &= w_{dd'}^S + w_{dd'}^A, \\ w_{dd'}^S &= w_{d'd}^S, \\ w_{dd'}^A &= -w_{d'd}^A. \end{aligned} \quad (1.101)$$

(b)

Here,

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^A x_d x_{d'} = \frac{1}{2} \sum_{d=1}^D \sum_{d'=1}^D (w_{dd'} - w_{d'd}) x_d x_{d'}. \quad (1.102)$$

The right hand side can be written as

$$\frac{1}{2} \left(\sum_{d=1}^D \sum_{d'=1}^D w_{dd'} x_d x_{d'} - \sum_{d=1}^D \sum_{d'=1}^D w_{d'd} x_d x_{d'} \right) = 0. \quad (1.103)$$

Therefore,

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^A x_d x_{d'} = 0. \quad (1.104)$$

(c)

We have

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'} x_d x_{d'} = \sum_{d=1}^D \sum_{d'=1}^D (w_{dd'}^S + w_{dd'}^A) x_d x_{d'}. \quad (1.105)$$

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

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^S x_d x_{d'} + \sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^A x_d x_{d'} = \sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^S x_d x_{d'}, \quad (1.106)$$

Therefore,

$$\sum_{d=1}^D \sum_{d'=1}^D w_{dd'} x_d x_{d'} = \sum_{d=1}^D \sum_{d'=1}^D w_{dd'}^S x_d x_{d'}. \quad (1.107)$$

(d)

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

1.15

(a)

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

$$n(1, M) = n(1, M - 1) = 1. \quad (1.108)$$

Let us assume that

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

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,

$$n(D + 1, M) = n(D, M) + n(D + 1, M - 1). \quad (1.110)$$

Thus,

$$n(D + 1, M) = \sum_{d=1}^{D+1} n(d, M - 1). \quad (1.111)$$

Hence, the assumption is proved by induction on D .

(b)

Note that

$$\sum_{d=1}^1 \frac{(d + M - 2)!}{(d - 1)!(M - 1)!} = 1. \quad (1.112)$$

Let us assume that

$$\sum_{d=1}^D \frac{(d+M-2)!}{(d-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!}. \quad (1.113)$$

Then

$$\sum_{d=1}^{D+1} \frac{(d+M-2)!}{(d-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} + \frac{(D+M-1)!}{D!(M-1)!}. \quad (1.114)$$

The right hand side can be written as

$$\frac{D(D+M-1)! + M(D+M-1)!}{D!M!} = \frac{(D+M)!}{D!M!}. \quad (1.115)$$

Therefore, the assumption is proved by induction on D .

(c)

By 1.14(d),

$$n(D, 2) = \frac{D(D+1)}{2}. \quad (1.116)$$

Let us assume that

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!}. \quad (1.117)$$

Then, by (a),

$$n(D, M+1) = \sum_{d=1}^D n(d, M). \quad (1.118)$$

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

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

Therefore, the assumption is proved by induction on M .

1.16

(a)

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

$$N(D, M) = \sum_{m=0}^M n(D, m), \quad (1.120)$$

where

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

(b)

By (a),

$$N(D, 0) = 1. \quad (1.122)$$

Let us assume that

$$\sum_{m=0}^M n(D, m) = \frac{(D + M)!}{D!M!}. \quad (1.123)$$

Then,

$$\sum_{m=0}^{M+1} n(D, m) = \frac{(D + M)!}{D!M!} + \frac{(D + M)!}{(D - 1)!(M + 1)!}. \quad (1.124)$$

The right hand side can be written as

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

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

$$N(D, M) = \frac{(D + M)!}{D!M!}. \quad (1.126)$$

(c)

By the approximation

$$n! \simeq n^n \exp(-n), \quad (1.127)$$

we have

$$\frac{(D+M)!}{D!M!} \simeq \frac{(D+M)^{D+M}}{D^D M^M}. \quad (1.128)$$

The right hand side can be written as

$$D^M \left(1 + \frac{M}{D}\right)^D \left(\frac{1}{M} + \frac{1}{D}\right)^M = M^D \left(1 + \frac{D}{M}\right)^M \left(\frac{1}{D} + \frac{1}{M}\right)^D. \quad (1.129)$$

Therefore,

$$N(D, M) \simeq \begin{cases} D^M, & D \gg M, \\ M^D, & M \gg D. \end{cases} \quad (1.130)$$

(d)

By (b),

$$\begin{aligned} N(10, 3) &= 286, \\ N(100, 3) &= 176851, \\ N(1000, 3) &= 167668501. \end{aligned} \quad (1.131)$$

1.17

(a)

Let

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \quad (1.132)$$

Then

$$\Gamma(x+1) = \int_0^\infty u^x \exp(-u) du. \quad (1.133)$$

The right hand side can be written as

$$[-u^x \exp(-u)]_{u=0}^{u=\infty} + \int_0^\infty x u^{x-1} \exp(-u) du = x \Gamma(x). \quad (1.134)$$

Therefore,

$$\Gamma(x+1) = x \Gamma(x). \quad (1.135)$$

(b)

Since

$$\Gamma(1) = \int_0^\infty \exp(-u) du, \quad (1.136)$$

we have

$$\Gamma(1) = 0!. \quad (1.137)$$

For a positive integer x , let us assume that

$$\Gamma(x) = (x-1)!. \quad (1.138)$$

Then,

$$\Gamma(x+1) = x\Gamma(x). \quad (1.139)$$

Therefore,

$$\Gamma(x+1) = x!. \quad (1.140)$$

Thus, the assumption is proved by induction on x .

1.18

(a)

Let us consider the transformation from Cartesian to polar coordinates

$$\prod_{d=1}^D \int_{-\infty}^{\infty} \exp(-x_d^2) dx_d = S_D \int_0^\infty \exp(-r^2) r^{D-1} dr, \quad (1.141)$$

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

$$s = r^2, \quad (1.142)$$

the right hand side can be written as

$$\frac{S_D}{2} \int_0^\infty \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \quad (1.143)$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \quad (1.144)$$

(b)

The volume of the sphere can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \quad (1.145)$$

Therefore,

$$V_D = \frac{S_D}{D}. \quad (1.146)$$

(c)

By (a) and (b),

$$\begin{aligned} S_2 &= 2\pi, \\ V_2 &= \pi. \end{aligned} \quad (1.147)$$

Similarly,

$$\begin{aligned} S_3 &= 4\pi, \\ V_3 &= \frac{4}{3}\pi. \end{aligned} \quad (1.148)$$

1.19

(a)

The volume of a cube of side 2 in D dimensions is 2^D . Therefore, by 1.18, the ratio of the volume of the cocentric sphere of radius 1 divided by the volume of the cube is given by

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} \Gamma\left(\frac{D}{2}\right)}. \quad (1.149)$$

(b)

By the Sterling's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}} \exp(-x) x^{\frac{x+1}{2}}, \quad (1.150)$$

we have

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} (2\pi)^{\frac{1}{2}} \exp\left(1 - \frac{D}{2}\right) \left(\frac{D}{2} - 1\right)^{\frac{D}{4}}}. \quad (1.151)$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}} \frac{1}{D} \left(\frac{e^2\pi^2}{8D-16} \right)^{\frac{D}{4}}. \quad (1.152)$$

Therefore,

$$\lim_{D \rightarrow \infty} \frac{V_D}{2^D} = 0. \quad (1.153)$$

(c)

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

$$\frac{\sqrt{\sum_{i=1}^D 1^2}}{1} = \sqrt{D}. \quad (1.154)$$

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

1.20

(a)

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

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \quad (1.155)$$

Then

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_r^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\phi, \quad (1.156)$$

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

$$\begin{aligned} & (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon} \int J dr' d\phi \\ & = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} d\mathbf{x}. \end{aligned} \quad (1.157)$$

Therefore,

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} \simeq p(r)\epsilon, \quad (1.158)$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad (1.159)$$

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

(b)

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

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left((D-1)r^{D-2} - \frac{r^D}{\sigma^2} \right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \quad (1.160)$$

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

$$\hat{r} = \sqrt{D-1}\sigma. \quad (1.161)$$

(c)

By the expression of $p(r)$ above,

$$\frac{p(\hat{r} + \epsilon)}{p(\hat{r})} = \left(\frac{\hat{r} + \epsilon}{\hat{r}} \right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \quad (1.162)$$

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

$$\begin{aligned} & \exp\left((D-1)\ln\left(1 + \frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\hat{r}^2}{\sigma^2}\ln\left(1 + \frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \end{aligned} \quad (1.163)$$

By the Taylor series

$$\ln(1+x) = x - \frac{1}{2}x^2 + o(x^3), \quad (1.164)$$

the right hand side can be approximated as

$$\exp\left(\frac{\hat{r}^2}{\sigma^2}\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) = \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \quad (1.165)$$

Therefore,

$$p(\hat{r} + \epsilon) \simeq p(\hat{r}) \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \quad (1.166)$$

(d)

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

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp\left(\frac{\hat{r}^2}{2\sigma^2}\right). \quad (1.167)$$

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

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp\left(\frac{D-1}{2}\right). \quad (1.168)$$

1.21

(a)

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

$$0 \leq a(b-a). \quad (1.169)$$

Therefore,

$$a \leq (ab)^{\frac{1}{2}}. \quad (1.170)$$

(b)

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

$$\begin{aligned} p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2) &\Rightarrow \mathbf{x} \in \mathcal{C}_1, \\ p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1) &\Rightarrow \mathbf{x} \in \mathcal{C}_2. \end{aligned} \quad (1.171)$$

Then, by (a),

$$\begin{aligned} \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} &\leq \int_{\mathcal{R}_1} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}, \\ \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} &\leq \int_{\mathcal{R}_2} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \end{aligned} \quad (1.172)$$

Therefore,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \quad (1.173)$$

1.22

Let

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

If

$$L_{kj} = 1 - I_{kj}, \quad (1.175)$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \quad (1.176)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \quad (1.177)$$

Therefore,

$$E L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.178)$$

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

1.23

Let

$$E L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.179)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.180)$$

Therefore,

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

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

1.24 (Incomplete)

Let

$$\mathbb{E} L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} + \lambda \int_{\forall k p(\mathcal{C}_k|\mathbf{x}) < \theta} p(\mathbf{x}) d\mathbf{x}. \quad (1.182)$$

1.25

Let

$$\mathbb{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.183)$$

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

$$\mathbf{0} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \quad (1.184)$$

The integral of the right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = \mathbf{y}(\mathbf{x}) p(\mathbf{x}) - p(\mathbf{x}) \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \quad (1.185)$$

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

$$\mathbf{0} = p(\mathbf{x}) (\mathbf{y}(\mathbf{x}) - \mathbb{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})). \quad (1.186)$$

Thus,

$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} \mathbb{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathbb{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \quad (1.187)$$

For a single target variable t , it reduces to

$$\underset{\mathbf{y}(\mathbf{x})}{\operatorname{argmin}} \mathbb{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathbb{E}_t(t|\mathbf{x}). \quad (1.188)$$

1.26

Let

$$\mathbb{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.189)$$

The right hand side can be written as

$$\begin{aligned}
& \int \int \| \mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t} \|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
&= \int \int \| \mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
&\quad + 2 \int \int (\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top (\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
&\quad + \int \int \| \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t} \|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}.
\end{aligned} \tag{1.190}$$

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

$$\int \| \mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \|^2 \left(\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \| \mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \|^2 p(\mathbf{x}) d\mathbf{x}. \tag{1.191}$$

The integral of the second term can be written as

$$\int (\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top \left(\int (\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.192}$$

Since

$$\begin{aligned}
\int \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t}}{p(\mathbf{x})}, \\
\int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}),
\end{aligned} \tag{1.193}$$

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

$$\int \left(\int \| \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t} \|^2 p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x} = \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.194}$$

Therefore,

$$\mathbb{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \| \mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \|^2 p(\mathbf{x}) d\mathbf{x} + \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.195}$$

Thus,

$$\underset{\mathbf{y}(\mathbf{x})}{\text{argmin}} \mathbb{E} L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.196}$$

1.27 (Incomplete)

(a)

Let

$$E L_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \quad (1.197)$$

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

$$0 = q p(\mathbf{x}) \int |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt. \quad (1.198)$$

Therefore,

$$\underset{y(\mathbf{x})}{\text{argmin}} E L_q = \left\{ y(\mathbf{x}) \mid \int |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = 0 \right\}. \quad (1.199)$$

(b)

We have

$$E L_1 = \int \left(\int \text{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.200)$$

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

$$\int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x}) dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x}) dt. \quad (1.201)$$

Therefore,

$$\underset{y(\mathbf{x})}{\text{argmin}} E L_1 = \text{median}(t|\mathbf{x}). \quad (1.202)$$

(c)

We have

$$\lim_{q \rightarrow 0} \left(\underset{y(\mathbf{x})}{\text{argmin}} E L_q \right) = \text{mode}(t|\mathbf{x})? \quad (1.203)$$

1.28

(a)

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \quad (1.204)$$

Then,

$$h(p^2) = 2h(p). \quad (1.205)$$

Let us assume that, for a positive integer n ,

$$h(p^n) = nh(p). \quad (1.206)$$

Then, by the first assumption,

$$h(p^{n+1}) = h(p^n) + h(p), \quad (1.207)$$

so that

$$h(p^{n+1}) = (n+1)h(p). \quad (1.208)$$

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

(b)

For positive integers m and n ,

$$h(p^n) = h(p^{\frac{n}{m}m}). \quad (1.209)$$

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

$$h(p^{\frac{n}{m}}) = \frac{n}{m}h(p). \quad (1.210)$$

(c)

By the continuity, for a positive real number a ,

$$h(p^a) = ah(p). \quad (1.211)$$

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

$$(p \ln p)h'(p) = h(p). \quad (1.212)$$

Then,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + \text{const} . \quad (1.213)$$

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

$$h(p) \propto \ln p. \quad (1.214)$$

1.29

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

$$H(x) = - \sum_{m=1}^M p(x_m) \ln p(x_m), \quad (1.215)$$

where

$$\sum_{m=1}^M p(x_m) = 1. \quad (1.216)$$

By the Jensen's inequality,

$$\sum_{m=1}^M p(x_i) \ln \frac{1}{p(x_m)} \leq \ln \left(\sum_{m=1}^M 1 \right). \quad (1.217)$$

Therefore,

$$H(x) \leq \ln M. \quad (1.218)$$

1.30

Let

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu, \sigma^2), \\ q(x) &= \mathcal{N}(x|m, s^2). \end{aligned} \quad (1.219)$$

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

$$\text{KL}(p||q) = - \int p(x) \ln \frac{q(x)}{p(x)} dx. \quad (1.220)$$

The right hand side can be written as

$$\begin{aligned}
& - \int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\
& = - \int_{-\infty}^{\infty} p(x) \left(-\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2} \right) dx.
\end{aligned} \tag{1.221}$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x) dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x) dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx. \tag{1.222}$$

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

$$\frac{1}{2s^2} \int_{-\infty}^{\infty} (x-\mu + \mu-m)^2 p(x) dx = \frac{\sigma^2 + (\mu-m)^2}{2s^2}. \tag{1.223}$$

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

$$\text{KL}(p||q) = \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu-m)^2}{2s^2} - \frac{1}{2}. \tag{1.224}$$

1.31

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

$$\begin{aligned}
H(\mathbf{x}) &= - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}, \\
H(\mathbf{y}) &= - \int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \\
H(\mathbf{x}, \mathbf{y}) &= - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}.
\end{aligned} \tag{1.225}$$

Note that

$$\begin{aligned}
H(\mathbf{x}) &= - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}, \\
H(\mathbf{y}) &= - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} \right) \ln p(\mathbf{y}) d\mathbf{y}.
\end{aligned} \tag{1.226}$$

Therefore,

$$H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y}. \quad (1.227)$$

Since

$$\int \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x}d\mathbf{y} = 1, \quad (1.228)$$

The Jensen's inequality can be used to write that

$$- \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y} \geq - \ln \left(\int \int p(\mathbf{x})p(\mathbf{y}) d\mathbf{x}d\mathbf{y} \right). \quad (1.229)$$

The right hand side can be written as

$$- \ln \left(\int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{y}) d\mathbf{y} \right) = 0. \quad (1.230)$$

Therefore,

$$H(\mathbf{x}, \mathbf{y}) \leq H(\mathbf{x}) + H(\mathbf{y}). \quad (1.231)$$

1.32

Let \mathbf{x} be a vector and let

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (1.232)$$

where \mathbf{A} is a nonsingular matrix. Since

$$1 = \int p_x(\mathbf{x}) d\mathbf{x} = \int p_x(\mathbf{A}^{-1}\mathbf{y}) |\det \mathbf{A}^{-1}| d\mathbf{y}, \quad (1.233)$$

we have

$$p_y(\mathbf{y}) = p_x(\mathbf{A}^{-1}\mathbf{y}) |\det \mathbf{A}^{-1}|, \quad (1.234)$$

so that

$$\ln p_y(\mathbf{y}) = \ln p_x(\mathbf{A}^{-1}\mathbf{y}) - \ln |\det \mathbf{A}|. \quad (1.235)$$

By the definition, the entropy is given by

$$H(\mathbf{y}) = - \int p_y(\mathbf{y}) \ln p_y(\mathbf{y}) d\mathbf{y}. \quad (1.236)$$

Then, the right hand side can be written as

$$- \int p_y(\mathbf{y}) \ln p_x(\mathbf{A}^{-1}\mathbf{y}) d\mathbf{y} + \ln |\det \mathbf{A}| \int p_y(\mathbf{y}) d\mathbf{y}. \quad (1.237)$$

By the transformation

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}, \quad (1.238)$$

the first term can be written as

$$- \int p_y(\mathbf{A}\mathbf{x}) \ln p_x(\mathbf{x}) |\det \mathbf{A}| d\mathbf{x}. \quad (1.239)$$

Since

$$1 = \int p_y(\mathbf{y}) d\mathbf{y} = \int p_y(\mathbf{A}\mathbf{x}) |\det \mathbf{A}| d\mathbf{x}, \quad (1.240)$$

we have

$$p_x(\mathbf{x}) = p_y(\mathbf{A}\mathbf{x}) |\det \mathbf{A}|. \quad (1.241)$$

Then, the first term can be written as

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

Therefore,

$$H(\mathbf{y}) = H(\mathbf{x}) + \ln |\det \mathbf{A}|. \quad (1.243)$$

1.33

Let x and y be two discrete variables with K and L states. By the definition, the conditional entropy is given by

$$H(y|x) = - \sum_{k=1}^K \sum_{l=1}^L p(x_k, y_l) \ln p(y_l|x_k). \quad (1.244)$$

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

$$0 = - \sum_{k=1}^K p(x_k) \sum_{l=1}^L p(y_l|x_k) \ln p(y_l|x_k). \quad (1.245)$$

Since

$$\begin{aligned} p(x_k) &\geq 0, \\ p(y_l|x_k) \ln p(y_l|x_k) &\leq 0, \end{aligned} \quad (1.246)$$

the equation reduces to

$$p(y_l|x_k) \ln p(y_l|x_k) = 0. \quad (1.247)$$

Then, $p(y_l|x_k)$ is zero or one. Therefore, since

$$\sum_{l=1}^L p(y_l|x_k) = 1, \quad (1.248)$$

we have

$$p(y_l|x_k) = \begin{cases} 1, & \text{if } K = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (1.249)$$

1.34

Let x be a variable. By the definition, the entropy is given by

$$H(x) = - \int_{-\infty}^{\infty} p(x) \ln p(x) dx. \quad (1.250)$$

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

$$\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= 1, \\ \int_{-\infty}^{\infty} xp(x) dx &= \mu, \\ \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx &= \sigma^2, \end{aligned} \quad (1.251)$$

let

$$\begin{aligned} L(p) = & H(x) + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) dx - 1 \right) + \lambda_2 \left(\int_{-\infty}^{\infty} xp(x) dx - \mu \right) \\ & + \lambda_3 \left(\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right). \end{aligned} \quad (1.252)$$

Setting the variation with respect to p to zero gives

$$0 = -\ln p - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2. \quad (1.253)$$

Therefore,

$$p(x) = \exp(-1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2), \quad (1.254)$$

so that

$$p(x) = c \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right), \quad (1.255)$$

where

$$c = \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right). \quad (1.256)$$

Substituting it to the constraints gives

$$\begin{aligned} c \int_{-\infty}^{\infty} \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx &= 1, \\ c \int_{-\infty}^{\infty} x \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx &= \mu, \\ c \int_{-\infty}^{\infty} (x - \mu)^2 \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx &= \sigma^2. \end{aligned} \quad (1.257)$$

By the transformation

$$y = \sqrt{-\lambda_3} \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right), \quad (1.258)$$

they can be written as

$$\begin{aligned} c \int_{-\infty}^{\infty} \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= 1, \\ c \int_{-\infty}^{\infty} \left((- \lambda_3)^{-\frac{1}{2}} y + \mu - \frac{\lambda_2}{2\lambda_3}\right) \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= \mu, \\ c \int_{-\infty}^{\infty} \left((- \lambda_3)^{-\frac{1}{2}} y - \frac{\lambda_2}{2\lambda_3}\right)^2 \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= \sigma^2. \end{aligned} \quad (1.259)$$

Since

$$\begin{aligned} \int_{-\infty}^{\infty} \exp(-y^2) dy &= \Gamma\left(\frac{1}{2}\right), \\ \int_{-\infty}^{\infty} y \exp(-y^2) dy &= 0, \\ \int_{-\infty}^{\infty} y^2 \exp(-y^2) dy &= \Gamma\left(\frac{3}{2}\right), \end{aligned} \quad (1.260)$$

they can be written as

$$\begin{aligned} c(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right) &= 1, \\ c\left(\mu - \frac{\lambda_2}{2\lambda_3}\right)(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right) &= \mu, \\ c\left((-\lambda_3)^{-\frac{3}{2}}\Gamma\left(\frac{3}{2}\right) + (-\lambda_3)^{-\frac{1}{2}}\frac{\lambda_2^2}{4\lambda_3^2}\Gamma\left(\frac{1}{2}\right)\right) &= \sigma^2. \end{aligned} \quad (1.261)$$

Then,

$$\begin{aligned} \lambda_1 &= 1 - \frac{1}{2}\ln(2\pi\sigma^2), \\ \lambda_2 &= 0, \\ \lambda_3 &= -\frac{1}{2\sigma^2}. \end{aligned} \quad (1.262)$$

Therefore,

$$p(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.263)$$

1.35

Let x be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \quad (1.264)$$

Then, by the definition, the entropy is given by

$$H(x) = - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \ln \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.265)$$

The right hand side can be written as

$$\begin{aligned} & - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2 \right) dx \\ &= \frac{1}{2}\ln(2\pi\sigma^2) \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx. \end{aligned} \quad (1.266)$$

Therefore,

$$H(x) = \frac{1}{2} (1 + \ln(2\pi\sigma^2)). \quad (1.267)$$

1.36 (Incomplete)

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

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b), \quad (1.268)$$

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

$$x = \lambda a + (1 - \lambda)b. \quad (1.269)$$

Then, the inequality can be written as

$$f(x) \leq \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b). \quad (1.270)$$

Let

$$g(x) = \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b) - f(x). \quad (1.271)$$

Then,

$$g(x) \geq 0. \quad (1.272)$$

Additionally, for $x > a$,

$$g(x) = (x-a) \left(\frac{f(b)-f(a)}{b-a} - \frac{f(x)-f(a)}{x-a} \right). \quad (1.273)$$

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

$$\begin{aligned} f'(c) &= \frac{f(b)-f(a)}{b-a}, \\ f'(y) &= \frac{f(x)-f(a)}{x-a}. \end{aligned} \quad (1.274)$$

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

$$f'(y) \leq f'(c). \quad (1.275)$$

1.37

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

$$H(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \quad (1.276)$$

The right hand side can be written as

$$\begin{aligned}
& - \int \int p(\mathbf{x}, \mathbf{y}) (\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})) d\mathbf{x}d\mathbf{y} \\
& = - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x}d\mathbf{y} - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}.
\end{aligned} \tag{1.277}$$

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

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \tag{1.278}$$

1.38

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

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \tag{1.279}$$

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

$$f\left(\sum_{m=1}^M \lambda_m x_m\right) \leq \sum_{m=1}^M \lambda_m f(x_m), \tag{1.280}$$

where $\lambda_m \geq 0$ and

$$\sum_{m=1}^M \lambda_m = 1. \tag{1.281}$$

Here, let $\lambda_m \geq 0$ and

$$\sum_{m=1}^{M+1} \lambda_m = 1. \tag{1.282}$$

Then, by the definition,

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right). \tag{1.283}$$

By the assumption,

$$f\left(\sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right) \leq \sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m). \tag{1.284}$$

Then,

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{m=1}^M \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m). \quad (1.285)$$

Then,

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \sum_{m=1}^{M+1} \lambda_m f(x_m). \quad (1.286)$$

Therefore, the assumption is proved by induction on M .

1.39

Let x and y be two binary variables where

$$\begin{aligned} p(x=0, y=0) &= \frac{1}{3}, \\ p(x=0, y=1) &= \frac{1}{3}, \\ p(x=1, y=0) &= 0, \\ p(x=1, y=1) &= \frac{1}{3}. \end{aligned} \quad (1.287)$$

(a)

By the definition, the entropy is given by

$$H(x) = - \sum p(x) \ln p(x). \quad (1.288)$$

By the distribution,

$$\begin{aligned} p(x=0) &= \frac{2}{3}, \\ p(x=1) &= \frac{1}{3}. \end{aligned} \quad (1.289)$$

Therefore,

$$H(x) = \ln 3 - \frac{2}{3} \ln 2. \quad (1.290)$$

(b)

By the definition, the entropy is given by

$$H(y) = - \sum p(y) \ln p(y). \quad (1.291)$$

By the distribution,

$$\begin{aligned} p(y = 0) &= \frac{1}{3}, \\ p(y = 1) &= \frac{2}{3}. \end{aligned} \quad (1.292)$$

Therefore,

$$H(y) = \ln 3 - \frac{2}{3} \ln 2. \quad (1.293)$$

(c)

By the definition, the conditional entropy is given by

$$H(y|x) = - \sum p(x, y) \ln p(y|x). \quad (1.294)$$

By the definition,

$$\begin{aligned} p(y = 0|x = 0) &= \frac{p(x = 0, y = 0)}{p(x = 0)}, \\ p(y = 0|x = 1) &= \frac{p(x = 1, y = 0)}{p(x = 1)}, \\ p(y = 1|x = 0) &= \frac{p(x = 0, y = 1)}{p(x = 0)}, \\ p(y = 1|x = 1) &= \frac{p(x = 1, y = 1)}{p(x = 1)}. \end{aligned} \quad (1.295)$$

Then, by the distribution,

$$\begin{aligned} p(y = 0|x = 0) &= \frac{1}{2}, \\ p(y = 0|x = 1) &= 0, \\ p(y = 1|x = 0) &= \frac{1}{2}, \\ p(y = 1|x = 1) &= 1. \end{aligned} \quad (1.296)$$

Therefore,

$$H(y|x) = \frac{2}{3} \ln 2. \quad (1.297)$$

(d)

By the definition, the conditional entropy is given by

$$H(x|y) = - \sum p(x, y) \ln p(x|y). \quad (1.298)$$

By the definition,

$$\begin{aligned} p(x = 0|y = 0) &= \frac{p(x = 0, y = 0)}{p(y = 0)}, \\ p(x = 0|y = 1) &= \frac{p(x = 0, y = 1)}{p(y = 1)}, \\ p(x = 1|y = 0) &= \frac{p(x = 1, y = 0)}{p(y = 0)}, \\ p(x = 1|y = 1) &= \frac{p(x = 1, y = 1)}{p(y = 1)}. \end{aligned} \quad (1.299)$$

Then, by the distribution,

$$\begin{aligned} p(x = 0|y = 0) &= 1, \\ p(x = 0|y = 1) &= \frac{1}{2}, \\ p(x = 1|y = 0) &= 0, \\ p(x = 1|y = 1) &= \frac{1}{2}. \end{aligned} \quad (1.300)$$

Therefore,

$$H(x|y) = \frac{2}{3} \ln 2. \quad (1.301)$$

(e)

By the definition, the entropy is given by

$$H(x, y) = - \sum p(x, y) \ln p(x, y). \quad (1.302)$$

Therefore,

$$H(x, y) = \ln 3. \quad (1.303)$$

(f)

By the definition, the mutual information is given by

$$I(x, y) = - \sum p(x, y) \ln \frac{p(x)p(y)}{p(x, y)}. \quad (1.304)$$

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

$$H(x) + H(y) - H(x, y). \quad (1.305)$$

Therefore,

$$I(x, y) = \ln 3 - \frac{4}{3} \ln 2. \quad (1.306)$$

1.40

Let x_1, \dots, x_M be numbers where $x_m > 0$, and let $\lambda_1, \dots, \lambda_M$ be numbers where $\lambda_m \geq 0$ and

$$\sum_{m=1}^M \lambda_m = 1. \quad (1.307)$$

By the Jensen's inequality,

$$\sum_{m=1}^M \lambda_m \ln x_m \leq \ln \left(\sum_{m=1}^M \lambda_m x_m \right), \quad (1.308)$$

so that

$$\prod_{m=1}^M x_m^{\lambda_m} \leq \sum_{m=1}^M \lambda_m x_m. \quad (1.309)$$

Substituting

$$\lambda_m = \frac{1}{M} \quad (1.310)$$

to the inequality gives

$$\left(\prod_{m=1}^M x_m \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{m=1}^M x_m. \quad (1.311)$$

1.41

Let \mathbf{x} and \mathbf{y} be continuous variables. Then, by the definition, the mutual information is given by

$$I(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x}d\mathbf{y}. \quad (1.312)$$

The right hand side can be written as

$$\begin{aligned} & - \int \int p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{x}) + \ln \frac{p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x}d\mathbf{y} \\ & = - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} + \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x}d\mathbf{y}. \end{aligned} \quad (1.313)$$

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

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}). \quad (1.314)$$

By the definition,

$$I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x}). \quad (1.315)$$

Therefore,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \quad (1.316)$$

2 Probability Distributions

2.1

Let x be a variable such that

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

(a)

We have

$$\sum_{x \in \{0,1\}} p(x) = 1 - \mu + \mu.\tag{2.2}$$

Therefore,

$$\sum_{x \in \{0,1\}} p(x) = 1.\tag{2.3}$$

(b)

By the definition,

$$\begin{aligned}\mathbb{E} x &= \sum_{x \in \{0,1\}} xp(x), \\ \mathbb{E} x^2 &= \sum_{x \in \{0,1\}} x^2p(x),\end{aligned}\tag{2.4}$$

Therefore,

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

Since

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

we have

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

(c)

By the definition,

$$H(x) = - \sum_{x \in \{0,1\}} p(x) \ln p(x). \quad (2.8)$$

Therefore,

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

2.2

Let x be a variable such that

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

(a)

We have

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

Therefore,

$$\sum_{x \in \{-1,1\}} p(x) = 1. \quad (2.12)$$

(b)

By the definition,

$$\begin{aligned} E x &= \sum_{x \in \{-1,1\}} x p(x), \\ E x^2 &= \sum_{x \in \{-1,1\}} x^2 p(x). \end{aligned} \quad (2.13)$$

The right hand sides can be written as

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

Therefore,

$$\begin{aligned} \mathbb{E} x &= \mu, \\ \mathbb{E} x^2 &= 1. \end{aligned} \tag{2.15}$$

Since

$$\text{var } x = \mathbb{E} x^2 - (\mathbb{E} x)^2, \tag{2.16}$$

we have

$$\text{var } x = 1 - \mu^2. \tag{2.17}$$

(c)

By the definition,

$$H(x) = - \sum_{x \in \{-1, 1\}} p(x|\mu) \ln p(x|\mu). \tag{2.18}$$

Therefore,

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

2.3

(a)

By the definition,

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

Then,

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

The right hand side can be written as

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

Therefore,

$$\binom{N}{n} + \binom{N}{n-1} = \binom{N+1}{n}. \tag{2.23}$$

(b)

Note that

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

Let us assume that

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

Then,

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

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

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

Then,

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

Therefore, the assumption is proved by induction on N .

(c)

Let n be a variable such that

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

Then,

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

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

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

Therefore,

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

2.4

Let n be a variable such that

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

(a)

We have

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

By 2.3(c),

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

Taking the derivative with respect to μ gives

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

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

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

Since

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

the second term can be written as

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

Therefore,

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

(b)

By 2.3(c),

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

Taking the second derivative with respect to μ gives

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

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

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

Since

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

the second and third terms can be written as

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

Then,

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

Therefore, since

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

we have

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

2.5

By the definition,

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

By the transformation

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

the right hand side can be written as

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

By the transformation

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

the right hand side can be written as

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

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

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

2.6

Let μ be a variable such that

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

Then, by the definition,

$$\begin{aligned} E\mu &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu, \\ E\mu^2 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu. \end{aligned} \quad (2.56)$$

By 2.5,

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

Therefore,

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

Since

$$\text{var } \mu = E\mu^2 - (E\mu)^2, \quad (2.59)$$

we have

$$\text{var } \mu = \frac{ab}{(a+b)^2(a+b+1)}. \quad (2.60)$$

Since

$$\frac{\partial}{\partial \mu} p(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \left(\frac{a-1}{\mu} - \frac{b-1}{1-\mu} \right), \quad (2.61)$$

we have

$$\text{mode } \mu = \frac{a-1}{a+b-2}. \quad (2.62)$$

2.7

Let m and l be variables such that

$$\begin{aligned} p(m, l|\mu) &= \binom{m+l}{m} \mu^m (1-\mu)^l, \\ p(\mu) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \end{aligned} \quad (2.63)$$

By 2.6,

$$E \mu = \frac{a}{a+b}. \quad (2.64)$$

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

$$0 = \binom{m+l}{m} \mu^m (1-\mu)^l \left(\frac{m}{\mu} + \frac{l}{1-\mu} \right). \quad (2.65)$$

Then, the maximum likelihood solution for μ is given by

$$\mu_{\text{ML}} = \frac{m}{m+l}. \quad (2.66)$$

By the Bayes' theorem,

$$p(\mu|m, l)p(m, l) = p(m, l|\mu)p(\mu). \quad (2.67)$$

Then, by 2.5,

$$p(\mu|m, l) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1} (1-\mu)^{l+b-1}. \quad (2.68)$$

The, by 2.6,

$$E(\mu|m, l) = \frac{m+a}{m+l+a+b}. \quad (2.69)$$

Therefore,

$$E(\mu|m, l) = \lambda \mu_{\text{ML}} + (1-\lambda) E \mu, \quad (2.70)$$

where

$$\lambda = \frac{m+l}{m+l+a+b}. \quad (2.71)$$

2.8

Let x and y be variables.

(a)

By the definition,

$$E x = \int x p(x) dx. \quad (2.72)$$

The right hand side can be written as

$$\int x \left(\int p(x, y) dy \right) dx = \int \left(\int x p(x|y) dx \right) p(y) dy. \quad (2.73)$$

Therefore,

$$E x = E_y (E_x(x|y)). \quad (2.74)$$

(b)

By the definition,

$$\text{var } x = \text{E} (x - \text{E} x)^2. \quad (2.75)$$

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

$$\text{E}_y \left(\text{E}_x \left((x - \text{E} x)^2 | y \right) \right) = \text{E}_y \left(\text{E}_x \left((x - \text{E}_x(x|y) + \text{E}_x(x|y) - \text{E} x)^2 | y \right) \right). \quad (2.76)$$

The right hand side can be written as

$$\begin{aligned} & \text{E}_y \left(\text{E}_x \left((x - \text{E}_x(x|y))^2 | y \right) \right) \\ & + 2 \text{E}_y \left(((\text{E}_x(x|y) - \text{E} x) \text{E}_x(x - \text{E}_x(x|y)) | y) \right) \\ & + \text{E}_y \left((\text{E}_x(x|y) - \text{E} x)^2 | y \right). \end{aligned} \quad (2.77)$$

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

$$2 \text{E}_y ((\text{E}_x(x|y) - \text{E} x) (\text{E}_x(x|y) - \text{E}_x(x|y))) = 0. \quad (2.78)$$

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

$$\text{E}_y (\text{E}_x(x|y) - \text{E}_y (\text{E}_x(x|y)))^2 = \text{var}_y (\text{E}_x(x|y)). \quad (2.79)$$

Therefore,

$$\text{var } x = \text{E}_y (\text{var}_x(x|y)) + \text{var}_y (\text{E}_x(x|y)). \quad (2.80)$$

2.9 (Incomplete)

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

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

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

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

Under the constraint

$$\sum_{m=1}^{M+1} \mu_m = 1, \quad (2.81)$$

we have

$$\int_0^c \prod_{m=1}^{M+1} \mu_m^{\alpha_m-1} d\mu_{M+1} = \left(\prod_{m=1}^{M-1} \mu_m^{\alpha_m-1} \right) \int_0^c \mu_{M+1}^{\alpha_{M+1}-1} (c - \mu_{M+1})^{\alpha_M-1} d\mu_{M+1}, \quad (2.82)$$

where

$$c = 1 - \sum_{m=1}^{M-1} \mu_m. \quad (2.83)$$

By the transformation

$$\mu'_{M+1} = \frac{\mu_{M+1}}{c}, \quad (2.84)$$

the integral of the right hand side can be written as

$$\begin{aligned} & \int_0^1 (c\mu'_{M+1})^{\alpha_{M+1}-1} (c(1 - \mu'_{M+1}))^{\alpha_M-1} c d\mu'_{M+1} \\ &= c^{\alpha_M+\alpha_{M+1}-1} \int_0^1 \mu'_{M+1}^{\alpha_{M+1}-1} (1 - \mu'_{M+1})^{\alpha_M-1} d\mu'_{M+1}. \end{aligned} \quad (2.85)$$

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

$$\frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \quad (2.86)$$

Then,

$$\int_0^c \prod_{m=1}^{M+1} \mu_m^{\alpha_m-1} d\mu_{M+1} = \left(\prod_{m=1}^{M-1} \mu_m^{\alpha_m-1} \right) c^{\alpha_M+\alpha_{M+1}-1} \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M + \alpha_{M+1})}. \quad (2.87)$$

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

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

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

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

The right hand side can be written as

$$\frac{\prod_{m=1}^{M+1} \Gamma(\alpha_m)}{\Gamma(\sum_{m=1}^{M+1} \alpha_m)}. \quad (2.88)$$

Therefore, the assumption is proved by induction on M .

2.10

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

$$p(\boldsymbol{\mu}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}. \quad (2.89)$$

Then, by the definition,

$$\begin{aligned} \mathbb{E} \mu_k &= \int \mu_k p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \\ \mathbb{E} \mu_k^2 &= \int \mu_k^2 p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \\ \mathbb{E} \mu_k \mu_{k'} &= \int \mu_k \mu_{k'} p(\boldsymbol{\mu}) d\boldsymbol{\mu}. \end{aligned} \quad (2.90)$$

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

$$\begin{aligned} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_k+1)}{\Gamma(\alpha_k)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 1)} &= \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}, \\ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_k+2)}{\Gamma(\alpha_k)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 2)} &= \frac{\alpha_k(\alpha_k + 1)}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}, \\ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_k+1)\Gamma(\alpha_{k'}+1)}{\Gamma(\alpha_k)\Gamma(\alpha_{k'})} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 2)} &= \frac{\alpha_k \alpha_{k'}}{\sum_{k=1}^K \alpha_k (\sum_{k=1}^K \alpha_k + 1)}. \end{aligned} \quad (2.91)$$

Then,

$$\begin{aligned} \mathbb{E} \mu_k &= \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}. \\ \mathbb{E} \mu_k^2 &= \frac{\alpha_k(\alpha_k + 1)}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1 \right)}, \\ \mathbb{E} \mu_k \mu_{k'} &= \frac{\alpha_k \alpha_{k'}}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1 \right)}. \end{aligned} \quad (2.92)$$

Since

$$\begin{aligned} \text{var} \mu_k &= \mathbb{E} \mu_k^2 - (\mathbb{E} \mu_k)^2, \\ \text{cov}(\mu_k, \mu_{k'}) &= \mathbb{E} \mu_k \mu_{k'} - \mathbb{E} \mu_k \mathbb{E} \mu_{k'}, \end{aligned} \quad (2.93)$$

we have

$$\begin{aligned} \text{var} \mu_k &= \frac{\alpha_k \left(\left(\sum_{k=1}^K \alpha_k \right) - \alpha_k \right)}{\left(\sum_{k=1}^K \alpha_k \right)^2 \left(\sum_{k=1}^K \alpha_k + 1 \right)}, \\ \text{cov}(\mu_k, \mu_{k'}) &= - \frac{\alpha_k \alpha_{k'}}{\left(\sum_{k=1}^K \alpha_k \right)^2 \left(\sum_{k=1}^K \alpha_k + 1 \right)}. \end{aligned} \quad (2.94)$$

2.11

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

$$p(\boldsymbol{\mu}) = \frac{\Gamma \left(\sum_{k=1}^K \alpha_k \right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}. \quad (2.95)$$

Then, by the definition,

$$\mathbb{E} \ln \mu_k = \int (\ln \mu_k) p(\boldsymbol{\mu}) d\boldsymbol{\mu}. \quad (2.96)$$

Since

$$\frac{\partial}{\partial \alpha_k} p(\boldsymbol{\mu}) = \left(\frac{\Gamma' \left(\sum_{k=1}^K \alpha_k \right)}{\Gamma \left(\sum_{k=1}^K \alpha_k \right)} - \frac{\Gamma'(\alpha_k)}{\Gamma(\alpha_k)} + \ln \mu_k \right) p(\boldsymbol{\mu}), \quad (2.97)$$

we have

$$\mathbb{E} \ln \mu_k = \frac{\partial}{\partial \alpha_k} \int p(\boldsymbol{\mu}) d\boldsymbol{\mu} + \left(\psi(\alpha_k) - \psi \left(\sum_{k=1}^K \alpha_k \right) \right) \int p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \quad (2.98)$$

where

$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \quad (2.99)$$

Therefore,

$$\mathbb{E} \ln \mu_k = \psi(\alpha_k) - \psi\left(\sum_{k=1}^K \alpha_k\right). \quad (2.100)$$

2.12

Let x be a variable such that

$$p(x) = \frac{1}{b-a}, \quad (2.101)$$

where $a < b$. Then

$$\int_a^b p(x) dx = 1. \quad (2.102)$$

Then, by the definition,

$$\begin{aligned} \mathbb{E} x &= \frac{1}{b-a} \int_a^b x dx, \\ \mathbb{E} x^2 &= \frac{1}{b-a} \int_a^b x^2 dx. \end{aligned} \quad (2.103)$$

Then,

$$\begin{aligned} \mathbb{E} x &= \frac{1}{2}(a+b), \\ \mathbb{E} x^2 &= \frac{1}{3}(a^2 + ab + b^2). \end{aligned} \quad (2.104)$$

Since

$$\text{var } x = \mathbb{E} x^2 - (\mathbb{E} x)^2, \quad (2.105)$$

we have

$$\text{var } x = \frac{1}{12}(b-a)^2. \quad (2.106)$$

2.13

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}). \end{aligned} \quad (2.107)$$

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

$$\text{KL}(p||q) = - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}. \quad (2.108)$$

Note that

$$\ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \frac{(2\pi)^{-\frac{D}{2}} |\det \mathbf{L}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) \right)}{(2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right)}. \quad (2.109)$$

The right hand side can be written as

$$\frac{1}{2} \ln \left| \frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}} \right| + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}). \quad (2.110)$$

Then, the integral can be written as

$$\begin{aligned} & \frac{1}{2} \ln \left| \frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}} \right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & + \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & - \frac{1}{2} \int (\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \end{aligned} \quad (2.111)$$

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

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\Sigma}, \quad (2.112)$$

we have

$$\int (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \text{tr } \boldsymbol{\Sigma}. \quad (2.113)$$

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

$$\text{tr} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}) = D. \quad (2.114)$$

Since

$$(\mathbf{x} - \mathbf{m})^\top \mathbf{L}^{-1} (\mathbf{x} - \mathbf{m}) = (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m}), \quad (2.115)$$

the integral of the third term can be written as

$$\begin{aligned} & \int (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{L}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & + 2(\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} \int (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}) \int \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ & = \text{tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}). \end{aligned} \quad (2.116)$$

Therefore,

$$\text{KL}(p||q) = \frac{1}{2} \left(\ln \left| \frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}} \right| - D + \text{tr}(\mathbf{L}^{-1} \boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}) \right). \quad (2.117)$$

2.14

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

$$\text{H}(\mathbf{x}) = - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \quad (2.118)$$

In order to maximise $\text{H}(x)$ with the constraints

$$\begin{aligned} \int p(\mathbf{x}) d\mathbf{x} &= 1, \\ \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} &= \boldsymbol{\mu}, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x}) d\mathbf{x} &= \boldsymbol{\Sigma}, \end{aligned} \quad (2.119)$$

let

$$\begin{aligned} L(p) = & \text{H}(\mathbf{x}) + \lambda \left(\int p(\mathbf{x}) d\mathbf{x} - 1 \right) + \mathbf{l}^\top \left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\mu} \right) \\ & + \mathbf{m}^\top \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\Sigma} \right) \mathbf{m}. \end{aligned} \quad (2.120)$$

Setting the variation with respect to p to zero gives

$$0 = -\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^\top \mathbf{x} + \mathbf{m}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{m}. \quad (2.121)$$

Then,

$$p(\mathbf{x}) = \exp(-1 + \lambda + \mathbf{l}^\top \mathbf{x} + \mathbf{m}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{m}), \quad (2.122)$$

so that

$$p(\mathbf{x}) = c \exp\left(-(\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})^\top \mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})\right), \quad (2.123)$$

where

$$\begin{aligned} c &= \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M}\mathbf{l}), \\ \mathbf{M} &= -(\mathbf{m}\mathbf{m}^\top)^{-1}. \end{aligned} \quad (2.124)$$

Substituting it to the constraints and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l} \quad (2.125)$$

gives

$$\begin{aligned} c \int \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= 1, \\ c \int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= \boldsymbol{\mu}, \\ c \int (\mathbf{y} + \mathbf{M}\mathbf{l})(\mathbf{y} + \mathbf{M}\mathbf{l})^\top \exp(-\mathbf{y}^\top \mathbf{M}^{-1} \mathbf{y}) d\mathbf{y} &= \boldsymbol{\Sigma}. \end{aligned} \quad (2.126)$$

Since

$$\begin{aligned} \int \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \left(\Gamma\left(\frac{1}{2}\right)\right)^D, \\ \int \mathbf{y} \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \mathbf{0}, \\ \int \mathbf{y}\mathbf{y}^\top \exp(-\mathbf{y}^\top \mathbf{y}) d\mathbf{y} &= \Gamma\left(\frac{3}{2}\right) \left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1} \mathbf{I}, \end{aligned} \quad (2.127)$$

they can be written as

$$\begin{aligned}
c \left(\Gamma \left(\frac{1}{2} \right) \right)^D |\det \mathbf{M}|^{\frac{1}{2}} &= 1, \\
c(\boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \left(\Gamma \left(\frac{1}{2} \right) \right)^D |\det \mathbf{M}|^{\frac{1}{2}} &= \boldsymbol{\mu}, \\
c \left(\Gamma \left(\frac{3}{2} \right) \left(\Gamma \left(\frac{1}{2} \right) \right)^{D-1} \mathbf{M} + \mathbf{M}\mathbf{l}(\mathbf{M}\mathbf{l})^\top \left(\Gamma \left(\frac{1}{2} \right) \right)^D \right) |\det \mathbf{M}|^{\frac{1}{2}} &= \boldsymbol{\Sigma}.
\end{aligned} \tag{2.128}$$

Then,

$$\begin{aligned}
\lambda &= 1 - \frac{D}{2} \ln \pi - \frac{1}{2} \ln |\det \mathbf{M}|, \\
\mathbf{l} &= \mathbf{0}, \\
\mathbf{M} &= 2\boldsymbol{\Sigma}.
\end{aligned} \tag{2.129}$$

Therefore,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right). \tag{2.130}$$

2.15

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.131}$$

Then, by the definition, the entropy is given by

$$\mathbf{H}(\mathbf{x}) = - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \tag{2.132}$$

The right hand side can be written as

$$\begin{aligned}
& - \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \left(-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln |\det \boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x} \\
&= \left(\frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}| \right) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\
&+ \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.
\end{aligned} \tag{2.133}$$

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

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

we have

$$\int (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \text{tr } \boldsymbol{\Sigma}. \quad (2.135)$$

Then, the second integral can be written as

$$\text{tr } (\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}) = D. \quad (2.136)$$

Therefore,

$$H(\mathbf{x}) = \frac{D}{2} (1 + \ln(2\pi)) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}|. \quad (2.137)$$

2.16

Let x be a variable such that

$$x = x_1 + x_2, \quad (2.138)$$

where

$$\begin{aligned} p(x_1) &= \mathcal{N}(x_1|\mu_1, \tau_1^{-1}), \\ p(x_2) &= \mathcal{N}(x_2|\mu_2, \tau_2^{-1}). \end{aligned} \quad (2.139)$$

By marginalisation,

$$p(x) = \int_{-\infty}^{\infty} p(x|x_2)p(x_2)dx_2. \quad (2.140)$$

The right hand side can be written as

$$\begin{aligned} & \int_{-\infty}^{\infty} \mathcal{N}(x|\mu_1 + x_2, \tau_1^{-1}) \mathcal{N}(x_2|\mu_2, \tau_2^{-1}) dx_2 \\ &= \int_{-\infty}^{\infty} \left(\frac{\tau_1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_1}{2}(x - \mu_1 - x_2)^2\right) \left(\frac{\tau_2}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_2}{2}(x_2 - \mu_2)^2\right) dx_2. \end{aligned} \quad (2.141)$$

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

$$\begin{aligned}
& -\frac{\tau_1 + \tau_2}{2} \left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 \\
& + \frac{\tau_1 + \tau_2}{2} \left(\frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 \\
& = -\frac{\tau_1 + \tau_2}{2} \left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2} \right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2.
\end{aligned} \tag{2.142}$$

Then,

$$p(x) = \mathcal{N}(x | \mu_1 + \mu_2, \tau_1^{-1} + \tau_2^{-1}). \tag{2.143}$$

Therefore, by 1.35,

$$H(x) = \frac{1}{2} (1 + \ln(2\pi) + \ln(\tau_1^{-1} + \tau_2^{-1})). \tag{2.144}$$

2.17

Let Σ be a matrix and

$$\begin{aligned}
\mathbf{S} &= \frac{1}{2} (\Sigma^{-1} + (\Sigma^{-1})^\top), \\
\mathbf{A} &= \frac{1}{2} (\Sigma^{-1} - (\Sigma^{-1})^\top).
\end{aligned} \tag{2.145}$$

Then,

$$\Sigma^{-1} = \mathbf{S} + \mathbf{A}, \tag{2.146}$$

so that

$$(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{S} (\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A} (\mathbf{x} - \boldsymbol{\mu}). \tag{2.147}$$

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

$$\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top (\Sigma^{-1})^\top (\mathbf{x} - \boldsymbol{\mu}). \tag{2.148}$$

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

$$-\frac{1}{2} (\Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}))^\top (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}). \tag{2.149}$$

Then,

$$(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = 0. \quad (2.150)$$

Therefore,

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

2.18

(a)

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

$$\boldsymbol{\Sigma} \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.152)$$

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

$$\overline{\mathbf{u}_d}^\top \boldsymbol{\Sigma} \mathbf{u}_d = \lambda_d, \quad (2.153)$$

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

$$\overline{\mathbf{u}_d}^\top \boldsymbol{\Sigma}^\top \mathbf{u}_d = (\overline{\boldsymbol{\Sigma} \mathbf{u}_d})^\top \mathbf{u}_d. \quad (2.154)$$

The right hand side can be writtet as

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

Therefore,

$$\lambda_d = \overline{\lambda_d}. \quad (2.156)$$

(b)

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

$$\boldsymbol{\Sigma} \mathbf{u}_d = \lambda_d \mathbf{u}_d \quad (2.157)$$

gives

$$\mathbf{u}_{d'}^\top \boldsymbol{\Sigma} \mathbf{u}_d = \lambda_d \mathbf{u}_{d'}^\top \mathbf{u}_d. \quad (2.158)$$

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

$$\mathbf{u}_{d'}^\top \boldsymbol{\Sigma}^\top \mathbf{u}_d = (\boldsymbol{\Sigma} \mathbf{u}_{d'})^\top \mathbf{u}_d. \quad (2.159)$$

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

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

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

$$\mathbf{u}_{d'}^\top \mathbf{u}_d = 0. \quad (2.161)$$

2.19

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

$$\Sigma \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.162)$$

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

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.163)$$

Then

$$\Sigma \mathbf{U} = \mathbf{U} \Lambda. \quad (2.164)$$

By 2.18,

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}. \quad (2.165)$$

Then,

$$\begin{aligned} \Sigma &= \mathbf{U} \Lambda \mathbf{U}^\top, \\ \Sigma^{-1} &= \mathbf{U} \Lambda^{-1} \mathbf{U}^\top, \end{aligned} \quad (2.166)$$

Therefore,

$$\begin{aligned} \Sigma &= \sum_{d=1}^D \lambda_d \mathbf{u}_d \mathbf{u}_d^\top, \\ \Sigma^{-1} &= \sum_{d=1}^D \frac{1}{\lambda_d} \mathbf{u}_d \mathbf{u}_d^\top. \end{aligned} \quad (2.167)$$

2.20

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

$$\Sigma \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.168)$$

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

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.169)$$

By 2.19,

$$\mathbf{a}^\top \Sigma \mathbf{a} = \mathbf{b}^\top \Lambda \mathbf{b}, \quad (2.170)$$

where

$$\mathbf{b} = \mathbf{U}^\top \mathbf{a}. \quad (2.171)$$

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

$$\mathbf{a}^\top \mathbf{\Sigma} \mathbf{a} > 0 \quad (2.172)$$

for any real vector \mathbf{a} is

$$\lambda_d > 0. \quad (2.173)$$

2.21

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

2.22

Let $\mathbf{\Sigma}$ be a $D \times D$ symmetric matrix and

$$\mathbf{\Sigma} \mathbf{\Lambda} = \mathbf{I}. \quad (2.174)$$

Taking the transpose of the both sides gives

$$\mathbf{\Lambda}^\top \mathbf{\Sigma} = \mathbf{I}. \quad (2.175)$$

Therefore,

$$\mathbf{\Lambda}^\top = \mathbf{\Lambda}. \quad (2.176)$$

2.23

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

$$\mathbf{\Sigma} \mathbf{u}_d = \lambda_d \mathbf{u}_d, \quad (2.177)$$

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

$$\begin{aligned} \mathbf{\Lambda}' &= \text{diag} \left(\lambda_1^{-\frac{1}{2}}, \dots, \lambda_D^{-\frac{1}{2}} \right), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \quad (2.178)$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{U}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}'^\top \mathbf{U}^\top (\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \quad (2.179)$$

By the transformation

$$\mathbf{y} = \boldsymbol{\Lambda}'^\top \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu}) \quad (2.180)$$

and the property

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}, \quad (2.181)$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left| \det \left(\mathbf{U}\boldsymbol{\Lambda}'^{-1} \right) \right| d\mathbf{y} = |\det \boldsymbol{\Sigma}|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \quad (2.182)$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det \boldsymbol{\Sigma}|^{\frac{1}{2}} \Delta^D V_D, \quad (2.183)$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \quad (2.184)$$

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

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

Then,

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

Therefore,

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

2.25

Let \mathbf{x} be a variable such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2.185)$$

where

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

Let

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \quad (2.186)$$

where

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

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

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ & -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\top \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) \\ & -\frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{ca}(\mathbf{x}_a - \boldsymbol{\mu}_a). \end{aligned} \quad (2.187)$$

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

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\top \boldsymbol{\Sigma}_{a|b,c}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}), \quad (2.188)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c), \\ \boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \quad (2.189)$$

Then,

$$p(\mathbf{x}_a|\mathbf{x}_b, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}). \quad (2.190)$$

By marginalisation,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \int p(\mathbf{x}_a|\mathbf{x}_b, \mathbf{x}_c)p(\mathbf{x}_c)d\mathbf{x}_c. \quad (2.191)$$

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

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\top \boldsymbol{\Sigma}_{a|b,c}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_{cc}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\ & = -\frac{1}{2}\mathbf{v}^\top \mathbf{M}\mathbf{v}, \end{aligned} \quad (2.192)$$

where

$$\begin{aligned} \mathbf{v} &= \begin{bmatrix} \mathbf{x}_c - \boldsymbol{\mu}_c \\ \mathbf{x}_a - \boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{bmatrix}, \\ \mathbf{M} &= \begin{bmatrix} \boldsymbol{\Lambda}_{cc} + \boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1} \boldsymbol{\Lambda}_{ac} & \boldsymbol{\Lambda}_{ac}^\top \\ \boldsymbol{\Lambda}_{ac} & \boldsymbol{\Lambda}_{aa} \end{bmatrix}. \end{aligned} \quad (2.193)$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{cc}^{-1} & -\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1} \\ -\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1} & \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^\top \boldsymbol{\Lambda}_{aa}^{-1} \end{bmatrix}. \quad (2.194)$$

Therefore,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \quad (2.195)$$

where

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

2.26

We have

$$\begin{aligned} & \left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1} \right) (\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}) \\ & = \mathbf{I} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D} \\ & \quad + \mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D}. \end{aligned} \quad (2.197)$$

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

$$\begin{aligned} & \mathbf{A}^{-1}\mathbf{B} \left(\mathbf{C} - (\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}(\mathbf{I} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}) \right) \mathbf{D} \\ & = \mathbf{A}^{-1}\mathbf{B} \left(\mathbf{C} - (\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})\mathbf{C} \right) \mathbf{D}. \end{aligned} \quad (2.198)$$

The right hand side can be written as

$$\mathbf{A}^{-1}\mathbf{B}(\mathbf{C} - \mathbf{C})\mathbf{D} = \mathbf{O}. \quad (2.199)$$

Then,

$$\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}\right)(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}) = \mathbf{I}. \quad (2.200)$$

Therefore,

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}. \quad (2.201)$$

2.27

(a)

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

$$\mathbf{E}(\mathbf{x} + \mathbf{z}) = \int \int (\mathbf{x} + \mathbf{z})p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \quad (2.202)$$

The right hand side can be written as

$$\int \mathbf{x} \left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{z} \right) d\mathbf{x} + \int \mathbf{z} \left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{x} \right) d\mathbf{z} = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} + \int \mathbf{z}p(\mathbf{z})d\mathbf{z}. \quad (2.203)$$

Therefore,

$$\mathbf{E}(\mathbf{x} + \mathbf{z}) = \mathbf{E}\mathbf{x} + \mathbf{E}\mathbf{z}. \quad (2.204)$$

(b)

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

$$\text{cov}(\mathbf{x} + \mathbf{z}) = \int \int (\mathbf{x} + \mathbf{z} - \mathbf{E}(\mathbf{x} + \mathbf{z}))(\mathbf{x} + \mathbf{z} - \mathbf{E}(\mathbf{x} + \mathbf{z}))^T p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \quad (2.205)$$

The right hand side can be written as

$$\begin{aligned} & \int \int (\mathbf{x} - \mathbf{E}\mathbf{x})(\mathbf{x} - \mathbf{E}\mathbf{x})^T p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} + \int \int (\mathbf{x} - \mathbf{E}\mathbf{x})(\mathbf{z} - \mathbf{E}\mathbf{z})^T p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} \\ & + \int \int (\mathbf{z} - \mathbf{E}\mathbf{z})(\mathbf{x} - \mathbf{E}\mathbf{x})^T p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} + \int \int (\mathbf{z} - \mathbf{E}\mathbf{z})(\mathbf{z} - \mathbf{E}\mathbf{z})^T p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \end{aligned} \quad (2.206)$$

Each term can be written as

$$\begin{aligned}
\int (\mathbf{x} - \mathbf{E} \mathbf{x}) (\mathbf{x} - \mathbf{E} \mathbf{x})^\top \left(\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \right) d\mathbf{x} &= \int (\mathbf{x} - \mathbf{E} \mathbf{x}) (\mathbf{x} - \mathbf{E} \mathbf{x})^\top p(\mathbf{x}) d\mathbf{x}, \\
\int (\mathbf{x} - \mathbf{E} \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int (\mathbf{z} - \mathbf{E} \mathbf{z})^\top p(\mathbf{z}) d\mathbf{z} &= (\mathbf{E} \mathbf{x} - \mathbf{E} \mathbf{x})(\mathbf{E} \mathbf{z} - \mathbf{E} \mathbf{z})^\top, \\
\int (\mathbf{z} - \mathbf{E} \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \int (\mathbf{x} - \mathbf{E} \mathbf{x})^\top p(\mathbf{x}) d\mathbf{x} &= (\mathbf{E} \mathbf{z} - \mathbf{E} \mathbf{z})(\mathbf{E} \mathbf{x} - \mathbf{E} \mathbf{x})^\top, \\
\int (\mathbf{z} - \mathbf{E} \mathbf{z}) (\mathbf{z} - \mathbf{E} \mathbf{z})^\top \left(\int p(\mathbf{x}, \mathbf{z}) d\mathbf{x} \right) d\mathbf{z} &= \int (\mathbf{z} - \mathbf{E} \mathbf{z}) (\mathbf{z} - \mathbf{E} \mathbf{z})^\top p(\mathbf{z}) d\mathbf{z}.
\end{aligned} \tag{2.207}$$

Therefore,

$$\text{cov}(\mathbf{x} + \mathbf{z}) = \text{cov} \mathbf{x} + \text{cov} \mathbf{z}. \tag{2.208}$$

2.28

Let \mathbf{z} be a variable such that

$$\begin{aligned}
\mathbf{z} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \\
\mathbf{E} \mathbf{z} &= \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}, \\
\text{cov} \mathbf{z} &= \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \\ \mathbf{A} \boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \end{bmatrix},
\end{aligned} \tag{2.209}$$

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

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

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

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

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}). \tag{2.211}$$

Therefore,

$$\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\
p(\mathbf{y} | \mathbf{x}) &= \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}).
\end{aligned} \tag{2.212}$$

2.29

Let

$$\mathbf{R} = \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}. \tag{2.213}$$

By 2.24,

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \\ \mathbf{A} \boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \end{bmatrix}. \tag{2.214}$$

2.30

Let

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \\ \mathbf{A} \boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \end{bmatrix}. \tag{2.215}$$

Then,

$$\mathbf{R}^{-1} \begin{bmatrix} \boldsymbol{\Lambda} \boldsymbol{\mu} - \mathbf{A}^\top \mathbf{L} \mathbf{b} \\ \mathbf{L} \mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A} \boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \tag{2.216}$$

2.31

Let \mathbf{y} be a variable such that

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \quad (2.217)$$

where

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}), \\ p(\mathbf{z}) &= \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}). \end{aligned} \quad (2.218)$$

By marginalisation,

$$p(\mathbf{y}) = \int p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (2.219)$$

The right hand side can be written as

$$\int \mathcal{N}(\mathbf{y} | \mathbf{x} + \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}) \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}) d\mathbf{x}. \quad (2.220)$$

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

$$-\frac{1}{2}(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}})^\top \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \quad (2.221)$$

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

$$-\frac{1}{2}\mathbf{u}^\top \mathbf{R} \mathbf{u} + \mathbf{u}^\top \mathbf{v} = -\frac{1}{2}(\mathbf{u} - \mathbf{R}^{-1}\mathbf{v})^\top \mathbf{R}(\mathbf{u} - \mathbf{R}^{-1}\mathbf{v}) + \frac{1}{2}\mathbf{v}^\top \mathbf{R}^{-1}\mathbf{v}, \quad (2.222)$$

where

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \\ \mathbf{R} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} & -\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \\ -\boldsymbol{\Sigma}_{\mathbf{z}}^{-1} & \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \end{bmatrix}, \\ \mathbf{v} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} \\ \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \boldsymbol{\mu}_{\mathbf{z}} \end{bmatrix}, \end{aligned} \quad (2.223)$$

By 2.29 and 2.30,

$$\begin{aligned} \mathbf{R}^{-1} &= \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x}} \\ \boldsymbol{\Sigma}_{\mathbf{x}} & \boldsymbol{\Sigma}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{z}} \end{bmatrix}, \\ \mathbf{R}^{-1}\mathbf{v} &= \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}} \end{bmatrix}. \end{aligned} \quad (2.224)$$

Therefore,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{z}}). \quad (2.225)$$

2.32

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}). \end{aligned} \quad (2.226)$$

By the Bayes' theorem,

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}). \quad (2.227)$$

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

$$\begin{aligned} & -\frac{1}{2}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})^\top \mathbf{L}(\mathbf{y} - \mathbf{Ax} - \mathbf{b}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\ &= -\frac{1}{2}(\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L}(\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ & \quad - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}). \end{aligned} \quad (2.228)$$

The right hand side can be written as

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{LA})(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ & -\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\top (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{LA})(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) + \frac{1}{2}\mathbf{z}^\top (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{LA})\mathbf{z} \\ & \quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}), \end{aligned} \quad (2.229)$$

where

$$\mathbf{z} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{LA})^{-1} \mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}). \quad (2.230)$$

The right hand side can be written as

$$\begin{aligned} & -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\top (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{LA})(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) \\ & -\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\top \mathbf{M}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}), \end{aligned} \quad (2.231)$$

where

$$\mathbf{M} = \mathbf{L} - \mathbf{LA}(\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{LA})^{-1} \mathbf{A}^\top \mathbf{L}. \quad (2.232)$$

We have

$$\boldsymbol{\mu} + \mathbf{z} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} ((\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A}) \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b})). \quad (2.233)$$

Then,

$$\boldsymbol{\mu} + \mathbf{z} = (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu}). \quad (2.234)$$

By 2.26,

$$(\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} = \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \boldsymbol{\Lambda}^{-1}. \quad (2.235)$$

Then,

$$\mathbf{M} = \mathbf{L} - \mathbf{L} \mathbf{A} \left(\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \boldsymbol{\Lambda}^{-1} \right) \mathbf{A}^\top \mathbf{L}. \quad (2.236)$$

The right hand side can be written as

$$\begin{aligned} & \mathbf{L} - \mathbf{L} \left(\mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top - \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1} \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top \right) \mathbf{L} \\ &= \mathbf{L} - \mathbf{L} \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1} (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top - \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top) \mathbf{L}. \end{aligned} \quad (2.237)$$

The right hand side can be written as

$$\begin{aligned} & \mathbf{L} - \mathbf{L} \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1} \\ &= \mathbf{L} (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top - \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top) (\mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top)^{-1}. \end{aligned} \quad (2.238)$$

Then,

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

Therefore,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x} | (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\mathbf{A}^\top \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu}), (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}), \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y} | \mathbf{A} \boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A} \boldsymbol{\Lambda}^{-1} \mathbf{A}^\top). \end{aligned} \quad (2.240)$$

2.33

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

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \\ p(\mathbf{y} | \mathbf{x}) &= \mathcal{N}(\mathbf{y} | \mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}). \end{aligned} \quad (2.241)$$

By the Bayes' theorem,

$$p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) = p(\mathbf{x} | \mathbf{y}) p(\mathbf{y}). \quad (2.242)$$

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

$$-\frac{1}{2} (\mathbf{y} - \mathbf{Ax} - \mathbf{b})^\top \mathbf{L} (\mathbf{y} - \mathbf{Ax} - \mathbf{b}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x} - \boldsymbol{\mu}). \quad (2.243)$$

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

$$-\frac{1}{2} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} + \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} -\mathbf{A}^\top \mathbf{L} \mathbf{b} + \boldsymbol{\Lambda} \boldsymbol{\mu} \\ \mathbf{L} \mathbf{b} \end{bmatrix}. \quad (2.244)$$

By 2.24,

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

so that

$$\begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A} & -\mathbf{A}^\top \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{bmatrix}^{-1} \begin{bmatrix} -\mathbf{A}^\top \mathbf{L} \mathbf{b} + \boldsymbol{\Lambda} \boldsymbol{\mu} \\ \mathbf{L} \mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A} \boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \quad (2.246)$$

Then,

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

By 2.25,

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu} + (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}), (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}). \quad (2.248)$$

We have

$$\begin{aligned} & \boldsymbol{\mu} + (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{A} \boldsymbol{\mu} - \mathbf{b}) \\ &= (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\boldsymbol{\Lambda} \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b})). \end{aligned} \quad (2.249)$$

Therefore,

$$p(\mathbf{x} | \mathbf{y}) = \mathcal{N}(\mathbf{x} | (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1} (\boldsymbol{\Lambda} \boldsymbol{\mu} + \mathbf{A}^\top \mathbf{L} (\mathbf{y} - \mathbf{b})), (\boldsymbol{\Lambda} + \mathbf{A}^\top \mathbf{L} \mathbf{A})^{-1}). \quad (2.250)$$

2.34

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.251)$$

Then,

$$\ln \left(\prod_{n=1}^N p(\mathbf{x}_n) \right) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\det \boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}). \quad (2.252)$$

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

$$\begin{aligned} \mathbf{0} &= \sum_{n=1}^N (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x}_n - \boldsymbol{\mu}), \\ \mathbf{O} &= -\frac{N}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \end{aligned} \quad (2.253)$$

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

$$\begin{aligned} \boldsymbol{\mu}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \end{aligned} \quad (2.254)$$

2.35

(a)

Let \mathbf{x} be a variable such that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.255)$$

Then,

$$\mathbf{E} \mathbf{x} \mathbf{x}^\top = \mathbf{E}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^\top. \quad (2.256)$$

The right hand side can be written as

$$\mathbf{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top + \boldsymbol{\mu} \mathbf{E}(\mathbf{x} - \boldsymbol{\mu})^\top + \mathbf{E}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\mu}^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \quad (2.257)$$

Since

$$\begin{aligned} \mathbf{E} \mathbf{x} &= \boldsymbol{\mu}, \\ \text{cov } \mathbf{x} &= \boldsymbol{\Sigma}, \end{aligned} \tag{2.258}$$

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

$$\mathbf{E} \mathbf{x} \mathbf{x}^\top = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{2.259}$$

(b)

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

$$\begin{aligned} p(\mathbf{x}_n) &= \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ p(\mathbf{x}_m) &= \mathcal{N}(\mathbf{x}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned} \tag{2.260}$$

If $n \neq m$, then

$$\mathbf{E} \mathbf{x}_n \mathbf{x}_m^\top = \mathbf{E} \mathbf{x}_n \mathbf{E} \mathbf{x}_m^\top. \tag{2.261}$$

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

$$\mathbf{E} \mathbf{x}_n \mathbf{x}_m^\top = I_{nm} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{2.262}$$

(c)

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

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.263}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \end{aligned} \tag{2.264}$$

Then,

$$\mathbf{E} \boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{E}(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top. \tag{2.265}$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} \sum_{n=1}^N \mathbb{E} \mathbf{x}_n \mathbf{x}_n^\top - \frac{1}{N^2} \sum_{n=1}^N \mathbb{E} \left(\sum_{n=1}^N \mathbf{x}_n \right) \mathbf{x}_n^\top - \frac{1}{N^2} \sum_{n=1}^N \mathbb{E} \mathbf{x}_n \left(\sum_{n=1}^N \mathbf{x}_n \right)^\top \\ & + \frac{1}{N^3} \sum_{n=1}^N \mathbb{E} \left(\sum_{n=1}^N \mathbf{x}_n \right) \left(\sum_{n=1}^N \mathbf{x}_n \right)^\top. \end{aligned} \quad (2.266)$$

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

$$-\frac{1}{N} ((\Sigma + \mu\mu^\top) + (N-1)\mu\mu^\top) = -\frac{1}{N} \Sigma - \mu\mu^\top. \quad (2.267)$$

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

$$\frac{1}{N^2} (N(\Sigma + \mu\mu^\top) + N(N-1)\mu\mu^\top) = \frac{1}{N} \Sigma + \mu\mu^\top. \quad (2.268)$$

Then,

$$\mathbb{E} \Sigma_{\text{ML}} = (\Sigma + \mu\mu^\top) + 2 \left(-\frac{1}{N} \Sigma - \mu\mu^\top \right) + \frac{1}{N} \Sigma + \mu\mu^\top. \quad (2.269)$$

Therefore,

$$\mathbb{E} \Sigma_{\text{ML}} = \frac{N-1}{N} \Sigma. \quad (2.270)$$

2.36

Let x_1, \dots, x_N be variables such that

$$p(x_n) = \mathcal{N}(x_n | \mu, \sigma^2). \quad (2.271)$$

Let us assume that μ is known. By 2.34, the maximum likelihood solution for σ^2 is given by

$$\sigma_{\text{ML}}^2{}^{(N)} = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2. \quad (2.272)$$

The right hand side can be written as

$$\frac{1}{N} (x_N - \mu)^2 + \frac{1}{N} \sum_{n=1}^{N-1} (x_n - \mu)^2 = \frac{1}{N} (x_N - \mu)^2 + \frac{N-1}{N} \sigma_{\text{ML}}^2{}^{(N-1)}. \quad (2.273)$$

Then,

$$\sigma_{\text{ML}}^{2(N)} = \sigma_{\text{ML}}^{2(N-1)} + \frac{1}{N} \left((x_N - \mu)^2 - \sigma_{\text{ML}}^{2(N-1)} \right). \quad (2.274)$$

We have

$$\frac{\partial}{\partial \sigma^2} (-\ln p(x_n)) = \frac{1}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} (x_n - \mu)^2. \quad (2.275)$$

Therefore,

$$\sigma_{\text{ML}}^{2(N)} = \sigma_{\text{ML}}^{2(N-1)} - \frac{2 \left(\sigma_{\text{ML}}^{2(N-1)} \right)^2}{N} \left(\frac{\partial}{\partial \sigma^2} (-\ln p(x_N)) \right) \Big|_{\sigma^2 = \sigma_{\text{ML}}^{2(N-1)}}. \quad (2.276)$$

2.37

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

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (2.277)$$

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

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \quad (2.278)$$

The right hand side can be written as

$$\begin{aligned} & \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top + \frac{1}{N} \sum_{n=1}^{N-1} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top \\ &= \frac{1}{N} (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top + \frac{N-1}{N} \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}. \end{aligned} \quad (2.279)$$

Then,

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} + \frac{1}{N} \left((\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top - \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} \right). \quad (2.280)$$

By 3.21(a), we have

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} (-\ln p(x_n)) = -\frac{1}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^\top. \quad (2.281)$$

Therefore,

$$\boldsymbol{\Sigma}_{\text{ML}}^{(N)} = \boldsymbol{\Sigma}_{\text{ML}}^{(N-1)} - \frac{\left(\boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}\right)^2}{N} \left(\frac{\partial}{\partial \boldsymbol{\Sigma}} (-\ln p(\mathbf{x}_N)) \right) \Big|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{ML}}^{(N-1)}}. \quad (2.282)$$

2.38

Let x_1, \dots, x_N be variables such that

$$\begin{aligned} p(x_n|\mu) &= \mathcal{N}(x_n|\mu, \sigma^2), \\ p(\mu) &= \mathcal{N}(\mu|\mu_0, \sigma_0^2). \end{aligned} \quad (2.283)$$

By the Bayes' theorem,

$$p(\mu|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu)p(\mu). \quad (2.284)$$

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

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2. \quad (2.285)$$

By 2.34, the maximum likelihood solution for μ is given by

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.286)$$

Then, the first term can be written as

$$\begin{aligned} & -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}} + \mu_{\text{ML}} - \mu)^2 \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{\mu_{\text{ML}} - \mu}{\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}}) - \frac{N}{2\sigma^2} (\mu_{\text{ML}} - \mu)^2. \end{aligned} \quad (2.287)$$

Since the second term of the right hand side is zero, the logarithm except the terms independent of \mathbf{x} and μ can be written as

$$\begin{aligned} & -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{N}{2\sigma^2} (\mu_{\text{ML}} - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 - \frac{1}{2\sigma_N^2} (\mu - \mu_N)^2 + \frac{\mu_N^2}{2\sigma_N^2} - \frac{N\mu_{\text{ML}}^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}, \end{aligned} \quad (2.288)$$

where

$$\begin{aligned}\mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\ \sigma_N^2 &= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}.\end{aligned}\tag{2.289}$$

Therefore,

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2).\tag{2.290}$$

2.39 (Incomplete)

Let x_1, \dots, x_N be variables such that

$$\begin{aligned}p(x_n|\mu) &= \mathcal{N}(x_n|\mu, \sigma^2), \\ p(\mu) &= \mathcal{N}(\mu|\mu_0, \sigma_0^2).\end{aligned}\tag{2.291}$$

Then, by 2.38,

$$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2),\tag{2.292}$$

where

$$\begin{aligned}\mu_N &= \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{n=1}^N x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\ \sigma_N^2 &= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}.\end{aligned}\tag{2.293}$$

Then

$$\begin{aligned}\mu_N &= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2}x_N + \frac{\sigma^2 - \sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\ \sigma_N^2 &= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2}\sigma_{N-1}^2.\end{aligned}\tag{2.294}$$

Additionally, we have

$$\begin{aligned}p(\mu|x_1, \dots, x_{N-1}) &= \mathcal{N}(\mu|\mu_{N-1}, \sigma_{N-1}^2), \\ p(x_N|\mu) &= \mathcal{N}(x_N|\mu, \sigma^2).\end{aligned}\tag{2.295}$$

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

$$\begin{aligned}
& -\frac{1}{2\sigma_{N-1}^2}(\mu - \mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N - \mu)^2 \\
& = -\frac{1}{2\frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}} \left(\mu - \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} - \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N \right)^2 \\
& \quad + \frac{1}{2\frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}} \left(\frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N \right) - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}.
\end{aligned} \tag{2.296}$$

Therefoere,

$$\begin{aligned}
\mu_N &= \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N, \\
\sigma_N^2 &= \frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}.
\end{aligned} \tag{2.297}$$

2.40

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

$$\begin{aligned}
p(\mathbf{x}_n|\boldsymbol{\mu}) &= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\
p(\boldsymbol{\mu}) &= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}).
\end{aligned} \tag{2.298}$$

By the Bayes' theorem,

$$p(\boldsymbol{\mu}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}). \tag{2.299}$$

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

$$-\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0). \tag{2.300}$$

The first term can be written as

$$\begin{aligned}
& -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}) \\
& = -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{N}{2} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}).
\end{aligned} \tag{2.301}$$

where

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \quad (2.302)$$

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

$$\begin{aligned} & -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{N}{2} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu}) \\ & - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\ & = -\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\top \boldsymbol{\Sigma}_N^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) \\ & + \frac{1}{2} \boldsymbol{\mu}_N^\top \boldsymbol{\Sigma}_N^{-1} \boldsymbol{\mu}_N, \end{aligned} \quad (2.303)$$

where

$$\begin{aligned} \boldsymbol{\mu}_N &= (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} (N\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0), \\ \boldsymbol{\Sigma}_N &= (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}. \end{aligned} \quad (2.304)$$

Therefore,

$$p(\boldsymbol{\mu}|\mathbf{X}) = \mathcal{N} \left(\boldsymbol{\mu} \mid (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} (N\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_{\text{ML}} + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0), (N\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \right). \quad (2.305)$$

2.41

By the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.306)$$

Then

$$\int_0^\infty \text{Gam}(\lambda|a, b) d\lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a-1} \exp(-b\lambda) d\lambda. \quad (2.307)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.308)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a-1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{\Gamma(a)} \int_0^\infty \lambda'^{a-1} \exp(-\lambda') d\lambda'. \quad (2.309)$$

The right hand side can be written as

$$\frac{1}{\Gamma(a)} \Gamma(a) = 1. \quad (2.310)$$

Therefore,

$$\int_0^\infty \text{Gam}(\lambda|a, b) d\lambda = 1. \quad (2.311)$$

2.42

Let λ be a variable such that

$$p(\lambda) = \text{Gam}(\lambda|a, b). \quad (2.312)$$

By the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.313)$$

Then

$$\text{E } \lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right) d\lambda. \quad (2.314)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.315)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b\Gamma(a)} \int_0^\infty \lambda'^a \exp(-\lambda') d\lambda'. \quad (2.316)$$

The right hand side can be written as

$$\frac{1}{b\Gamma(a)} \Gamma(a+1) = \frac{a}{b}. \quad (2.317)$$

Therefore,

$$\text{E } \lambda = \frac{a}{b}. \quad (2.318)$$

Additionally,

$$\mathbb{E} \lambda^2 = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1} \exp\left(-\frac{\lambda}{b}\right) d\lambda. \quad (2.319)$$

By the transformation

$$\lambda' = b\lambda, \quad (2.320)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b^2 \Gamma(a)} \int_0^\infty \lambda'^{a+1} \exp(-\lambda') d\lambda'. \quad (2.321)$$

The right hand side can be written as

$$\frac{1}{b^2 \Gamma(a)} \Gamma(a+2) = \frac{a(a+1)}{b^2}. \quad (2.322)$$

Therefore,

$$\mathbb{E} \lambda^2 = \frac{a(a+1)}{b^2}. \quad (2.323)$$

By the definition,

$$\text{var } \lambda = \mathbb{E} \lambda^2 - (\mathbb{E} \lambda)^2. \quad (2.324)$$

Therefore,

$$\text{var } \lambda = \frac{a}{b^2}. \quad (2.325)$$

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

$$0 = \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{\lambda} - b \right) \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right). \quad (2.326)$$

Therefore,

$$\text{mode } \lambda = \frac{a-1}{b}. \quad (2.327)$$

2.43

Let

$$p(x|\sigma^2, q) = \frac{q}{2\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \exp\left(-\frac{|x|^q}{2\sigma^2}\right). \quad (2.328)$$

Then

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = \frac{q}{\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \int_0^{\infty} \exp\left(-\frac{x^q}{2\sigma^2}\right) dx. \quad (2.329)$$

By the transformation

$$x' = \frac{x^q}{2\sigma^2}, \quad (2.330)$$

the right hand side can be written as

$$\begin{aligned} & \frac{q}{\Gamma(\frac{1}{q})} (2\sigma^2)^{-\frac{1}{q}} \int_0^{\infty} \exp(-x') (2\sigma^2)^{\frac{1}{q}} \frac{1}{q} x^{\frac{1}{q}-1} dx' \\ &= \frac{1}{\Gamma(\frac{1}{q})} \int_0^{\infty} x^{\frac{1}{q}-1} \exp(-x') dx'. \end{aligned} \quad (2.331)$$

The right hand side can be written as

$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma\left(\frac{1}{q}\right) = 1. \quad (2.332)$$

Therefore,

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) dx = 1. \quad (2.333)$$

Additionally,

$$p(x|\sigma^2, 2) = \frac{1}{\Gamma(\frac{1}{2})} (2\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \quad (2.334)$$

Therefore,

$$p(x|\sigma^2, 2) = \mathcal{N}(x|0, \sigma^2). \quad (2.335)$$

Finally, let $\mathbf{t} = (t_1, \dots, t_N)^\top$ and $\mathbf{X} = \{x_1, \dots, x_N\}$ such that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \quad (2.336)$$

where

$$p(\epsilon_n) = p(\epsilon_n|\sigma^2, q). \quad (2.337)$$

Therefore, the logarithm of $p(\epsilon_n)$ except the terms independent of \mathbf{w} and σ^2 can be written as

$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q} \ln(2\sigma^2). \quad (2.338)$$

Thus, the logarithm of $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$ except the terms independent of \mathbf{w} and σ^2 can be written as

$$-\frac{1}{2\sigma^2} \sum_{n=1}^N |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln(2\sigma^2). \quad (2.339)$$

2.44

Let x_1, \dots, x_N be variables such that

$$\begin{aligned} p(x_n|\mu, \tau) &= \mathcal{N}(x_n|\mu, \tau^{-1}), \\ p(\mu, \tau) &= \mathcal{N}(\mu|\mu_0, (\beta\tau)^{-1}) \text{Gam}(\tau|a, b). \end{aligned} \quad (2.340)$$

By the Bayes' theorem,

$$p(\mu, \tau|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu, \tau)p(\mu, \tau). \quad (2.341)$$

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

$$\begin{aligned} & \frac{N}{2} \ln \tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \mu)^2 + \frac{1}{2} \ln \tau - \frac{\beta\tau}{2} (\mu - \mu_0)^2 + (a-1) \ln \tau - b\tau \\ &= \left(a + \frac{N-1}{2}\right) \ln \tau - \frac{N\tau}{2} (\bar{x} - \mu)^2 - \frac{\beta\tau}{2} (\mu - \mu_0)^2 - b\tau - \frac{\tau}{2} \sum_{n=1}^N (x_n - \bar{x})^2, \end{aligned} \quad (2.342)$$

where

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n. \quad (2.343)$$

Since

$$-\frac{N\tau}{2} (\bar{x} - \mu)^2 - \frac{\beta\tau}{2} (\mu - \mu_0)^2 = -\frac{(N+\beta)\tau}{2} \left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2 - \frac{N\beta\tau(\bar{x} - \mu_0)^2}{2(N+\beta)}, \quad (2.344)$$

the right hand side can be written as

$$\begin{aligned} & -\frac{(N+\beta)\tau}{2} \left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2 \\ & + \left(a + \frac{N-1}{2}\right) \ln \tau - \left(b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N+\beta)} + \frac{1}{2} \sum_{n=1}^N (x_n - \bar{x})^2\right) \tau. \end{aligned} \quad (2.345)$$

Therefore,

$$p(\mu, \tau | \mathbf{x}) = \mathcal{N} \left(\mu \mid \frac{N\bar{x} + \beta\mu_0}{N + \beta}, ((N + \beta)\tau)^{-1} \right) \\ \text{Gam} \left(\tau \mid a + \frac{N + 1}{2}, b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N + \beta)} + \frac{1}{2} \sum_{n=1}^N (x_n - \bar{x})^2 \right). \quad (2.346)$$

2.45

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.347)$$

Then

$$p(\mathbf{X} | \boldsymbol{\Lambda}) = \prod_{n=1}^N \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.348)$$

The right hand side except the terms independent of $\boldsymbol{\Lambda}$ can be written as

$$(\det \boldsymbol{\Lambda})^{\frac{N}{2}} \exp \left(-\frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Lambda} (\mathbf{x}_n - \boldsymbol{\mu}) \right) = (\det \boldsymbol{\Lambda})^{\frac{N}{2}} \exp \left(-\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Lambda}) \right), \quad (2.349)$$

where

$$\mathbf{S} = \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\top. \quad (2.350)$$

Therefore,

$$p(\mathbf{X} | \boldsymbol{\Lambda}) \propto (\det \boldsymbol{\Lambda})^{\frac{N}{2}} \exp \left(-\frac{1}{2} \text{tr}(\mathbf{S}\boldsymbol{\Lambda}) \right). \quad (2.351)$$

Let us assume that a prior distribution of $\boldsymbol{\Lambda}$ is given by

$$\mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) = B(\mathbf{W}, \nu) (\det \boldsymbol{\Lambda})^{\frac{\nu-D-1}{2}} \exp \left(-\frac{1}{2} \text{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda}) \right). \quad (2.352)$$

Then, by the definition,

$$p(\boldsymbol{\Lambda} | \mathbf{X}, \mathbf{W}, \nu) \propto p(\mathbf{X} | \boldsymbol{\Lambda}) \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu), \quad (2.353)$$

where the right hand side except the terms independent of $\mathbf{\Lambda}$ can be written as

$$(\det \mathbf{\Lambda})^{\frac{\nu+N-D-1}{2}} \exp \left(-\frac{1}{2} \text{tr} ((\mathbf{W}^{-1} + \mathbf{S}) \mathbf{\Lambda}) \right). \quad (2.354)$$

Therefore,

$$p(\mathbf{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) = \mathcal{W} \left(\mathbf{\Lambda} \mid (\mathbf{W}^{-1} + \mathbf{S})^{-1}, \nu + N \right). \quad (2.355)$$

Thus, \mathcal{W} is a conjugate prior distribution of $\mathbf{\Lambda}$.

2.46

Let x be a variable such that

$$p(x|\mu, \tau, a, b) = \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b). \quad (2.356)$$

Then

$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau. \quad (2.357)$$

The right hand side can be written as

$$\begin{aligned} & \int_0^\infty \left(\frac{\tau}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\tau}{2} (x - \mu)^2 \right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) d\tau \\ &= (2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp \left(-\left(b + \frac{(x - \mu)^2}{2} \right) \tau \right) d\tau. \end{aligned} \quad (2.358)$$

By the transformation

$$\tau' = \left(b + \frac{(x - \mu)^2}{2} \right) \tau, \quad (2.359)$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{\tau'}{b + \frac{(x - \mu)^2}{2}} \right)^{a-\frac{1}{2}} \exp(-\tau') \frac{d\tau'}{b + \frac{(x - \mu)^2}{2}} = \Gamma \left(a + \frac{1}{2} \right) \left(b + \frac{(x - \mu)^2}{2} \right)^{-a-\frac{1}{2}}. \quad (2.360)$$

Therefore,

$$p(x|\mu, \tau, a, b) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left(b + \frac{(x - \mu)^2}{2} \right)^{-a-\frac{1}{2}}. \quad (2.361)$$

Let

$$\begin{aligned}\nu &= 2a, \\ \lambda &= \frac{a}{b}.\end{aligned}\tag{2.362}$$

Then

$$p(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}.\tag{2.363}$$

2.47

By the definition,

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}.\tag{2.364}$$

By the transformation

$$y = \frac{\lambda(x-\mu)^2}{\nu},\tag{2.365}$$

the right hand side except the terms independent of x can be written as

$$(1+y)^{-\frac{\lambda(x-\mu)^2}{2y}-\frac{1}{2}}.\tag{2.366}$$

In the limit $y \rightarrow \infty$, it becomes

$$\exp\left(-\frac{\lambda}{2}(x-\mu)^2\right).\tag{2.367}$$

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

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \eta, \nu) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right).\tag{2.368}$$

Then

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta.\tag{2.369}$$

The right hand side can be written as

$$\begin{aligned} & \int_0^\infty (2\pi)^{-\frac{D}{2}} (\det(\eta\mathbf{\Lambda}))^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \eta \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right) \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \eta^{\frac{\nu}{2}-1} \exp\left(-\frac{\nu}{2}\eta\right) d\eta \\ &= (2\pi)^{-\frac{D}{2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} (\det \mathbf{\Lambda})^{\frac{1}{2}} \int_0^\infty \eta^{\frac{D+\nu}{2}-1} \exp\left(-\frac{1}{2}(\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu}))\eta\right) d\eta. \end{aligned} \quad (2.370)$$

By the transformation

$$\eta' = \frac{1}{2} (\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})) \eta, \quad (2.371)$$

the integral of the right hand side can be written as

$$\begin{aligned} & \int_0^\infty \left(\frac{2\eta'}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})} \right)^{\frac{D+\nu}{2}-1} \exp(-\eta') \frac{2}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})} d\eta' \\ &= \left(\frac{2}{\nu + (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})} \right)^{\frac{D+\nu}{2}} \Gamma\left(\frac{D+\nu}{2}\right). \end{aligned} \quad (2.372)$$

Therefore,

$$p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \mathbf{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.373)$$

2.49

Let \mathbf{x} be a variable such that

$$p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu). \quad (2.374)$$

By the definition,

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.375)$$

First,

$$\mathbb{E} \mathbf{x} = \int \mathbf{x} \text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) d\mathbf{x}. \quad (2.376)$$

The right hand side can be written as

$$\begin{aligned} & \int \mathbf{x} \left(\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left(\int \mathbf{x} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \quad (2.377)$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \quad (2.378)$$

Therefore,

$$\mathbb{E} \mathbf{x} = \boldsymbol{\mu}. \quad (2.379)$$

Additionally,

$$\text{cov } \mathbf{x} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \quad (2.380)$$

The right hand side can be written as

$$\begin{aligned} & \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \left(\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}) d\mathbf{x} \right) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \quad (2.381)$$

The right hand side can be written as

$$\int (\eta\boldsymbol{\Lambda})^{-1} \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \int \eta^{\frac{\nu}{2}-2} \exp\left(-\frac{\nu}{2}\eta\right) d\eta. \quad (2.382)$$

By the transformation

$$\eta' = \frac{\nu}{2}\eta, \quad (2.383)$$

the integral of the right hand side can be written as

$$\int \left(\frac{2}{\nu}\eta'\right)^{-\frac{\nu}{2}-2} \exp(-\eta') \frac{2}{\nu} d\eta' = \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2} - 1\right). \quad (2.384)$$

Therefore, the right hand side can be written as

$$\boldsymbol{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2} - 1\right) = \frac{\frac{\nu}{2}}{\frac{\nu}{2} - 1} \boldsymbol{\Lambda}^{-1}. \quad (2.385)$$

Thus,

$$\text{cov } \mathbf{x} = \frac{\nu}{\nu - 2} \mathbf{\Lambda}^{-1}. \quad (2.386)$$

Finally, setting the derivative of $\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)$ with respect to \mathbf{x} to zero gives

$$\mathbf{0} = -\frac{1}{2} (\mathbf{\Lambda} + \mathbf{\Lambda}^\top) (\mathbf{x} - \boldsymbol{\mu}) \int \eta \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta \mathbf{\Lambda})^{-1}) \text{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \quad (2.387)$$

Therefore,

$$\text{mode } \mathbf{x} = \boldsymbol{\mu}. \quad (2.388)$$

2.50

By the definition,

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2}) (\det \mathbf{\Lambda})^{\frac{1}{2}}}{\Gamma(\frac{\nu}{2}) (\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.389)$$

By the transformation

$$y = \frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{\nu}, \quad (2.390)$$

the right hand side except the terms independent of \mathbf{x} can be written as

$$(1 + y)^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})}{2y} - \frac{D}{2}}. \quad (2.391)$$

In the limit $y \rightarrow \infty$, it becomes

$$\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Lambda} (\mathbf{x} - \boldsymbol{\mu})\right). \quad (2.392)$$

Therefore, in the limit $\nu \rightarrow \infty$, $\text{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)$ becomes $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1})$.

2.51

We have

$$\exp(iA) \exp(-iA) = 1. \quad (2.393)$$

The left hand side can be written as

$$(\cos A + i \sin A)(\cos A - i \sin A) = \cos^2 A + \sin^2 A. \quad (2.394)$$

Therefore,

$$\cos^2 A + \sin^2 A = 1. \quad (2.395)$$

Additionally,

$$\cos(A - B) = \operatorname{Re}(\exp(i(A - B))). \quad (2.396)$$

The right hand side can be written as

$$\operatorname{Re}(\exp(iA) \exp(-iB)) = \operatorname{Re}((\cos A + i \sin A)(\cos B - i \sin B)). \quad (2.397)$$

The right hand side can be written as $\cos A \cos B + \sin A \sin B$. Therefore,

$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \quad (2.398)$$

Finally,

$$\sin(A - B) = \operatorname{Im}(\exp(i(A - B))). \quad (2.399)$$

The right hand side can be written as

$$\operatorname{Im}(\exp(iA) \exp(-iB)) = ((\cos A + i \sin A)(\cos B - i \sin B)). \quad (2.400)$$

The right hand side can be written as $\sin A \cos B - \cos A \sin B$. Therefore,

$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \quad (2.401)$$

2.52 (Incomplete)

Let θ be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.402)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \quad (2.403)$$

By the Taylor series

$$\cos \alpha = 1 - \frac{1}{2}\alpha^2 + O(\alpha^4) \quad (2.404)$$

and the transformation

$$\xi = m^{\frac{1}{2}}(\theta - \theta_0), \quad (2.405)$$

we have

$$\exp(m \cos(\theta - \theta_0)) = \exp\left(m \left(1 - \frac{1}{2}(\theta - \theta_0)^2 + O((\theta - \theta_0)^4)\right)\right). \quad (2.406)$$

2.53

Let θ_0 be a parameter such that

$$\sum_{n=1}^N \sin(\theta_n - \theta_0) = 0. \quad (2.407)$$

The left hand side can be written as

$$\sum_{n=1}^N (\sin \theta_n \cos \theta_0 - \cos \theta_n \sin \theta_0) = \cos \theta_0 \sum_{n=1}^N \sin \theta_n - \sin \theta_0 \sum_{n=1}^N \cos \theta_n. \quad (2.408)$$

Therefore,

$$\theta_0 = \arctan \left(\frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right). \quad (2.409)$$

2.54

Let θ be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.410)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \quad (2.411)$$

Setting the first and second derivatives with respect to θ to zero gives

$$\begin{aligned} 0 &= -m \sin(\theta - \theta_0) p(\theta|\theta_0, m), \\ 0 &= (m^2 \sin^2(\theta - \theta_0) - m \cos(\theta - \theta_0)) p(\theta|\theta_0, m). \end{aligned} \quad (2.412)$$

Therefore,

$$\begin{aligned} \operatorname{argmax}_{\theta} p(\theta|\theta_0, m) &= \theta_0, \\ \operatorname{argmin}_{\theta} p(\theta|\theta_0, m) &= \theta_0 - \pi \operatorname{sgn}(\theta_0 - \pi). \end{aligned} \quad (2.413)$$

2.55

Let

$$\theta_0^{\text{ML}} = \arctan \left(\frac{\sum_{n=1}^N \sin \theta_n}{\sum_{n=1}^N \cos \theta_n} \right). \quad (2.414)$$

Let

$$\begin{aligned} \bar{r} \cos \bar{\theta} &= \frac{1}{N} \sum_{n=1}^N \cos \theta_n, \\ \bar{r} \sin \bar{\theta} &= \frac{1}{N} \sum_{n=1}^N \sin \theta_n. \end{aligned} \quad (2.415)$$

Then

$$\theta_0^{\text{ML}} = \bar{\theta}. \quad (2.416)$$

Here,

$$\frac{1}{N} \sum_{n=1}^N \cos (\theta_n - \theta_0^{\text{ML}}) = \left(\frac{1}{N} \sum_{n=1}^N \cos \theta_n \right) \cos \theta_0^{\text{ML}} + \left(\frac{1}{N} \sum_{n=1}^N \sin \theta_n \right) \sin \theta_0^{\text{ML}}. \quad (2.417)$$

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

$$\bar{r} \cos^2 \bar{\theta} + \bar{r} \sin^2 \bar{\theta} = \bar{r}. \quad (2.418)$$

Therefore,

$$\frac{1}{N} \sum_{n=1}^N \cos (\theta_n - \theta_0^{\text{ML}}) = \bar{r}. \quad (2.419)$$

2.56

By the definition,

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \quad (2.420)$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp((a-1) \ln \mu + (b-1) \ln(1-\mu)) \quad (2.421)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a - 1 \\ b - 1 \end{bmatrix}.$$

Additionally, by the definition,

$$\text{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \quad (2.422)$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \exp((a - 1) \ln \lambda - b\lambda). \quad (2.423)$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a - 1 \\ -b \end{bmatrix}.$$

Finally, for

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0)), \quad (2.424)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta, \quad (2.425)$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m \cos \theta_0 \cos \theta + m \sin \theta_0 \sin \theta). \quad (2.426)$$

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,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right). \quad (2.427)$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})), \quad (2.428)$$

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

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})). \quad (2.429)$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} = 1 \quad (2.430)$$

with respect to $\boldsymbol{\eta}$ gives

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x}) h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = \mathbf{0}. \quad (2.431)$$

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \int \mathbf{u}(\mathbf{x}) p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathbb{E} \mathbf{u}(\mathbf{x}). \quad (2.432)$$

Therefore,

$$\mathbb{E} \mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}. \quad (2.433)$$

Thus,

$$\mathbf{E} \mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}). \quad (2.434)$$

Taking the second derivative with respect to $\boldsymbol{\eta}$ gives

$$\begin{aligned} \nabla \nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} + 2\nabla g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} \\ + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = \mathbf{O}. \end{aligned} \quad (2.435)$$

The left hand side can be written as

$$\begin{aligned} \frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \frac{2\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int \mathbf{u}(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} + \int \mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top p(\mathbf{x}|\boldsymbol{\eta}) d\mathbf{x} \\ = \frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2 \mathbf{E} \mathbf{u}(\mathbf{x}) \mathbf{E} \mathbf{u}(\mathbf{x})^\top + \mathbf{E} (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top). \end{aligned} \quad (2.436)$$

Therefore,

$$\mathbf{E} (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top) = -\frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\top}{g^2(\boldsymbol{\eta})}. \quad (2.437)$$

By the definition,

$$\text{cov} \mathbf{u}(\mathbf{x}) = \mathbf{E} (\mathbf{u}(\mathbf{x}) \mathbf{u}(\mathbf{x})^\top) - \mathbf{E} \mathbf{u}(\mathbf{x}) \mathbf{E} \mathbf{u}(\mathbf{x})^\top. \quad (2.438)$$

Thus,

$$\text{cov} \mathbf{u}(\mathbf{x}) = -\frac{\nabla \nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\top}{g^2(\boldsymbol{\eta})}. \quad (2.439)$$

Hence,

$$\text{cov} \mathbf{u}(\mathbf{x}) = -\nabla \nabla \ln g(\boldsymbol{\eta}). \quad (2.440)$$

2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right). \quad (2.441)$$

Then

$$\int p(x|\sigma) dx = \frac{1}{\sigma} \int f\left(\frac{x}{\sigma}\right) dx. \quad (2.442)$$

By the transformation

$$x' = \frac{x}{\sigma}, \quad (2.443)$$

the right hand side can be written as

$$\frac{1}{\sigma} \int f(x') \sigma dx' = \int f(x') dx'. \quad (2.444)$$

Therefore, $p(x|\sigma)$ will be normalised if $f(x)$ is normalised.

2.60

Let \mathbf{x} be a variable such that

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \quad (2.445)$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \quad (2.446)$$

Since

$$\int p(\mathbf{x}) d\mathbf{x} = 1, \quad (2.447)$$

we have

$$\sum_i h_i \Delta_i = 1. \quad (2.448)$$

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

$$\ln \left(\prod_i h_i^{n_i} \right) = \sum_i n_i \ln h_i, \quad (2.449)$$

where

$$\sum_i n_i = N. \quad (2.450)$$

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda \left(\sum_i h_i \Delta_i - 1 \right) \quad (2.451)$$

with respect to h_i and λ to zero gives

$$\begin{aligned}\frac{n_i}{h_i} + \lambda \Delta_i &= 0, \\ \sum_i h_i \Delta_i - 1 &= 0.\end{aligned}\tag{2.452}$$

Then,

$$\begin{aligned}\lambda &= -N, \\ h_i &= \frac{n_i}{N \Delta_i}.\end{aligned}\tag{2.453}$$

Therefore, the maximum likelihood estimator for the $\{h_i\}$ is $\frac{n_i}{N \Delta_i}$.

2.61 (Incomplete)

Let \mathbf{x} be a variable and $\mathbf{x}_1, \dots, \mathbf{x}_N$ be observations. Let

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})},\tag{2.454}$$

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}' - \mathbf{x}\| \leq \|\mathbf{x}_{(K)} - \mathbf{x}\|} d\mathbf{x}',\tag{2.455}$$

K is a constant and $\mathbf{x}_{(K)}$ is the K th nearest observation from the point \mathbf{x} .

3 Linear Models for Regression

3.1

By the definition,

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \quad (3.1)$$

The right hand side can be written as

$$\frac{1 - \exp(-2a)}{1 + \exp(-2a)} = \frac{2}{1 + \exp(-2a)} - 1. \quad (3.2)$$

Therefore,

$$\tanh a = 2\sigma(2a) - 1, \quad (3.3)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (3.4)$$

Let

$$y(x_n, \mathbf{w}) = w_0 + \sum_{m=1}^M w_m \sigma\left(\frac{x - \mu_m}{s}\right). \quad (3.5)$$

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

$$w_0 + \sum_{m=1}^M w_m \frac{1 + \tanh\left(\frac{x - \mu_m}{2s}\right)}{2} = w_0 + \frac{1}{2} \sum_{m=1}^M w_m + \frac{1}{2} \sum_{m=1}^M w_m \tanh\left(\frac{x - \mu_m}{2s}\right). \quad (3.6)$$

Therefore, $y(x_n, \mathbf{w})$ is equivalent to

$$y(x_n, \mathbf{u}) = u_0 + \sum_{m=1}^M u_m \tanh\left(\frac{x - \mu_m}{2s}\right), \quad (3.7)$$

where

$$\begin{aligned} u_0 &= w_0 + \frac{1}{2} \sum_{m=1}^M w_m, \\ u_m &= \frac{1}{2} w_m. \end{aligned} \quad (3.8)$$

3.2 (Incomplete)

Let Φ 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 Φ ?

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 Φ ?

3.3

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N r_n (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2. \quad (3.11)$$

The right hand side can be written as

$$\frac{1}{2} \|\mathbf{t}' - \Phi' \mathbf{w}\|^2, \quad (3.12)$$

where

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \vdots \\ \sqrt{r_N} t_N \end{bmatrix}, \Phi' = \begin{bmatrix} \sqrt{r_1} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \sqrt{r_N} \phi(\mathbf{x}_N)^\top \end{bmatrix}.$$

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

$$\mathbf{0} = -\Phi'^\top (\mathbf{t}' - \Phi' \mathbf{w}). \quad (3.13)$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} E(\mathbf{w}) = (\Phi'^\top \Phi')^{-1} \Phi'^\top \mathbf{t}'. \quad (3.14)$$

3.4 (Incomplete)

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (y(\mathbf{x}_n, \mathbf{w}) - t_n)^2, \quad (3.15)$$

where

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{m=1}^M w_m (x_m + \epsilon_m), \\ p(\epsilon_m) &= \mathcal{N}(\epsilon_m | 0, \sigma^2). \end{aligned} \quad (3.16)$$

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

$$\mathbf{0} = \sum_{n=1}^N \begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix} (y(\mathbf{x}_n, \mathbf{w}) - t_n).$$

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

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2. \quad (3.17)$$

Then, the minimisation of $E(\mathbf{w})$ under the constraint

$$\sum_{m=1}^M |w_m|^q \leq \eta \quad (3.18)$$

reduces to the minimisation of

$$E(\mathbf{w}) + \lambda \left(\sum_{m=1}^M |w_m|^q - \eta \right) \quad (3.19)$$

with respect to \mathbf{w} and λ . Then,

$$\eta = \sum_{m=1}^M |w_m^*(\lambda)|^q, \quad (3.20)$$

where

$$\mathbf{w}^*(\lambda) = \underset{\mathbf{w}}{\operatorname{argmin}} \left(E(\mathbf{w}) + \lambda \left(\sum_{m=1}^M |w_m|^q - \eta \right) \right). \quad (3.21)$$

3.6

Let $\mathbf{t}_1, \dots, \mathbf{t}_N$ be variables in D dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{W}), \boldsymbol{\Sigma}), \quad (3.22)$$

where

$$\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n). \quad (3.23)$$

Then,

$$\begin{aligned} & \ln \left(\prod_{n=1}^N p(\mathbf{t}_n | \mathbf{W}, \boldsymbol{\Sigma}) \right) \\ &= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n))^\top \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)). \end{aligned} \quad (3.24)$$

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

$$\begin{aligned} \mathbf{O} &= -\frac{1}{2} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)) (\boldsymbol{\phi}(\mathbf{x}_n))^\top, \\ \mathbf{O} &= -\frac{N}{2} (\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n))^\top. \end{aligned} \quad (3.25)$$

Therefore,

$$\begin{aligned} \mathbf{W}_{\text{ML}} &= (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{t}, \\ \boldsymbol{\Sigma}_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{W}_{\text{ML}}^\top \boldsymbol{\phi}(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}_{\text{ML}}^\top \boldsymbol{\phi}(\mathbf{x}_n))^\top, \end{aligned} \quad (3.26)$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\top \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\top \end{bmatrix}.$$

3.7

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.27)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \quad (3.28)$$

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,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N). \quad (3.32)$$

3.8

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.33)$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.34)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi_N^\top \mathbf{t}_N), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \Phi_N^\top \Phi_N. \end{aligned} \quad (3.35)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \quad (3.36)$$

The right hand side can be written as

$$p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \quad (3.37)$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N). \quad (3.38)$$

The logarithm of the right hand side except the terms independent of \mathbf{w} can be written as

$$\begin{aligned} & -\frac{\beta}{2} (t_{N+1} - \mathbf{w}^\top \phi(\mathbf{x}_{N+1}))^2 - \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) \\ &= -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu}_{N+1})^\top \boldsymbol{\Lambda}_{N+1} (\mathbf{w} - \boldsymbol{\mu}_{N+1}) + \frac{1}{2} \boldsymbol{\mu}_{N+1}^\top \boldsymbol{\Lambda}_{N+1} \boldsymbol{\mu}_{N+1} \\ & \quad - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} t_{N+1}^2, \end{aligned} \quad (3.39)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{N+1} &= \boldsymbol{\Lambda}_{N+1}^{-1} (\mathbf{S}_N^{-1} \mathbf{m}_N + \beta t_{N+1} \phi(\mathbf{x}_{N+1})), \\ \boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_N^{-1} + \beta \phi(\mathbf{x}_{N+1}) \phi(\mathbf{x}_{N+1})^\top. \end{aligned} \quad (3.40)$$

Therefore,

$$\begin{aligned} \boldsymbol{\mu}_{N+1} &= \mathbf{m}_{N+1}, \\ \boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_{N+1}^{-1}. \end{aligned} \quad (3.41)$$

Thus,

$$p(\mathbf{w}|\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}). \quad (3.42)$$

3.9 (Incomplete)

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.43)$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.44)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}_N^\top \mathbf{t}_N), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}_N^\top \boldsymbol{\Phi}_N. \end{aligned} \quad (3.45)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \quad (3.46)$$

The right hand side can be written as

$$p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \quad (3.47)$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N). \quad (3.48)$$

The logarithm of the right hand side except the terms independent of \mathbf{w} can be written as

$$-\frac{\beta}{2} (t_{N+1} - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_{N+1}))^2 - \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N). \quad (3.49)$$

3.10

Let t be a variable such that

$$\begin{aligned} p(t|\mathbf{w}) &= \mathcal{N}(t|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.50)$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.51)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.52)$$

By marginalisation,

$$p(t|\mathbf{t}) = \int p(t|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \quad (3.53)$$

The logarithm of the integrand of the right hand side except the terms independent of t and \mathbf{w} can be written as

$$-\frac{\beta}{2}(t - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}))^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \quad (3.54)$$

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,

$$p(t|\mathbf{t}) = \mathcal{N}(t|\mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})), \quad (3.55)$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \quad (3.56)$$

3.11

Let t be a variable such that

$$\begin{aligned} p(t|\mathbf{w}) &= \mathcal{N}(t|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \quad (3.57)$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.58)$$

where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi_N^\top \mathbf{t}_N), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \Phi_N^\top \Phi_N.\end{aligned}\tag{3.59}$$

Then, by 3.10,

$$p(t|\mathbf{t}) = \mathcal{N}(t \mid \mathbf{m}_N^\top \phi(\mathbf{x}), \sigma_N^2(\mathbf{x})),\tag{3.60}$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^\top \mathbf{S}_N \phi(\mathbf{x}).\tag{3.61}$$

Then,

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \phi(\mathbf{x})^\top (\mathbf{S}_N - \mathbf{S}_{N+1}) \phi(\mathbf{x}).\tag{3.62}$$

By the expression of \mathbf{S}_N above,

$$\mathbf{S}_{N+1} = (\mathbf{S}_N^{-1} + \beta \phi(\mathbf{x}_{N+1}) \phi(\mathbf{x}_{N+1})^\top)^{-1}.\tag{3.63}$$

By the identity

$$(\mathbf{M} + \mathbf{v} \mathbf{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1} \mathbf{v}) (\mathbf{v}^\top \mathbf{M}^{-1})}{1 + \mathbf{v}^\top \mathbf{M}^{-1} \mathbf{v}},\tag{3.64}$$

the right hand side can be written as

$$\mathbf{S}_N - \frac{\beta (\mathbf{S}_N \phi(\mathbf{x}_{N+1})) (\phi(\mathbf{x}_{N+1})^\top \mathbf{S}_N)}{1 + \beta \phi(\mathbf{x}_{N+1})^\top \mathbf{S}_N \phi(\mathbf{x}_{N+1})}.\tag{3.65}$$

Therefore,

$$\phi(\mathbf{x})^\top (\mathbf{S}_N - \mathbf{S}_{N+1}) \phi(\mathbf{x}) = \frac{\beta (\phi(\mathbf{x})^\top \mathbf{S}_N \phi(\mathbf{x}_{N+1}))^2}{1 + \beta \phi(\mathbf{x}_{N+1})^\top \mathbf{S}_N \phi(\mathbf{x}_{N+1})}.\tag{3.66}$$

Thus,

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}).\tag{3.67}$$

3.12

Let t_1, \dots, t_N be variables such that

$$\begin{aligned}p(t_n | \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0),\end{aligned}\tag{3.68}$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{w}, \beta | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta). \quad (3.69)$$

The logarithm of the right hand side except the terms independent of \mathbf{t} , \mathbf{w} and β can be written as

$$\begin{aligned} & -\frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{M}{2} \ln \beta^{-1} - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ & + (a_0 - 1) \ln \beta - b_0 \beta \\ = & -\frac{M}{2} \ln \beta - \frac{\beta}{2} \mathbf{w}^\top (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \mathbf{w} + \beta \mathbf{w}^\top (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}) - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 \\ & + \left(a_0 + \frac{N}{2} - 1 \right) \ln \beta - b_0 \beta. \end{aligned} \quad (3.70)$$

The right hand side can be written as

$$-\frac{M}{2} \ln \beta - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) + (a_N - 1) \ln \beta - b_N \beta, \quad (3.71)$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi}, \\ a_N &= a_0 + \frac{N}{2}, \\ b_N &= b_0 + \frac{1}{2} \|\mathbf{t}\|^2 + \frac{1}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N. \end{aligned} \quad (3.72)$$

Therefore,

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \text{Gam}(\beta | a_N, b_N). \quad (3.73)$$

Substituting it to the result of the Bayes' theorem above, we have

$$p(\mathbf{t}) = \frac{\mathcal{N}(\mathbf{t} | \boldsymbol{\Phi} \mathbf{w}, \beta^{-1} \mathbf{I}) \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0)}{\mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \text{Gam}(\beta | a_N, b_N)}. \quad (3.74)$$

The logarithm of the right hand side can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^\top (\mathbf{t} - \Phi \mathbf{w}) \\
& -\frac{M}{2} \ln(2\pi) - \frac{M}{2} \ln \beta^{-1} - \frac{1}{2} \det \mathbf{S}_0 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\
& + a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1) \ln \beta - b_0 \beta \\
& + \frac{M}{2} \ln(2\pi) + \frac{M}{2} \ln \beta^{-1} + \frac{1}{2} \det \mathbf{S}_N + \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) \\
& - a_N \ln b_N + \ln \Gamma(a_N) - (a_N - 1) \ln \beta + b_N \beta \\
& = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \det \mathbf{S}_0 + a_0 \ln b_0 - \ln \Gamma(a_0) + \frac{1}{2} \det \mathbf{S}_N - a_N \ln b_N + \ln \Gamma(a_N).
\end{aligned} \tag{3.75}$$

Therefore,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left(\frac{\det \mathbf{S}_N}{\det \mathbf{S}_0} \right)^{\frac{1}{2}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_N^{a_N}}. \tag{3.76}$$

3.13

Let t_1, \dots, t_N be variables such that

$$\begin{aligned}
p(t_n | \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}), \\
p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0),
\end{aligned} \tag{3.77}$$

where \mathbf{w} and ϕ are vectors in M dimensions. Then, by 3.12,

$$p(\mathbf{w}, \beta | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \beta^{-1} \mathbf{S}_N) \text{Gam}(\beta | a_N, b_N), \tag{3.78}$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \Phi^\top \mathbf{t}), \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \Phi^\top \Phi, \\
a_N &= a_0 + \frac{N}{2}, \\
b_N &= b_0 + \frac{1}{2} \|\mathbf{t}\|^2 + \frac{1}{2} \mathbf{m}_0^\top \mathbf{S}_0 \mathbf{m}_0 - \frac{1}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N.
\end{aligned} \tag{3.79}$$

By marginalisation,

$$p(t | \mathbf{t}) = \int \int p(t | \mathbf{w}, \beta) p(\mathbf{w}, \beta | \mathbf{t}) d\mathbf{w} d\beta. \tag{3.80}$$

The right hand side can be written as

$$\int \left(\int \mathcal{N}(t|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}) \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1} \mathbf{S}_N) d\mathbf{w} \right) \text{Gam}(\beta|a_N, b_N) d\beta. \quad (3.81)$$

The logarithm of the integrand with respect to \mathbf{w} except the terms independent of \mathbf{w} can be written as

$$-\frac{\beta}{2} (t - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}))^2 - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N). \quad (3.82)$$

It can be written as

$$-\frac{\beta}{2} \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top & -\boldsymbol{\phi}(\mathbf{x}) \\ -\boldsymbol{\phi}(\mathbf{x})^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix} + \beta \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix}^\top \begin{bmatrix} \mathbf{S}_N^{-1} \mathbf{m}_N \\ 0 \end{bmatrix} - \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top & -\boldsymbol{\phi}(\mathbf{x}) \\ -\boldsymbol{\phi}(\mathbf{x})^\top & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}_N & \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}) \\ \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N & 1 + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}.$$

Then,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\top & -\boldsymbol{\phi}(\mathbf{x}) \\ -\boldsymbol{\phi}(\mathbf{x})^\top & 1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_N^{-1} \mathbf{m}_N \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{m}_N \\ \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}.$$

Therefore, the integral with respect to \mathbf{w} can be written as

$$\mathcal{N}(t|\mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1} (1 + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}))). \quad (3.83)$$

Then, the logarithm of the integrand with respect to β except the terms independent of β can be written as

$$\begin{aligned} & -\frac{1}{2} \ln \beta^{-1} - \frac{\beta}{2(1 + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}))} (t - \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}))^2 + (a_N - 1) \ln \beta - b_N \beta \\ & = \left(a_N + \frac{1}{2} - 1 \right) \ln \beta - \left(b_N + \frac{(t - \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}))^2}{2(1 + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}))} \right) \beta. \end{aligned} \quad (3.84)$$

Therefore, the integral with respect to β except the terms independent of t can be written as

$$\left(b_N + \frac{(t - \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}))^2}{2(1 + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}))} \right)^{-a_N - \frac{1}{2}}. \quad (3.85)$$

Thus,

$$p(t|\mathbf{x}, \mathbf{t}) = \text{St}(t|\mu, \lambda, \nu), \quad (3.86)$$

where

$$\begin{aligned} \mu &= \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}), \\ \lambda &= \frac{a_N}{b_N} (1 + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}))^{-1}, \\ \nu &= 2a_N. \end{aligned} \quad (3.87)$$

3.14 (Incomplete)

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.88)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.89)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.90)$$

Let

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}). \quad (3.91)$$

Then,

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n, \quad (3.92)$$

where

$$k(\mathbf{x}, \mathbf{x}') = \beta \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}'). \quad (3.93)$$

Let us suppose that $\boldsymbol{\phi}_j(\mathbf{x})$ are linearly independent, $N > M$ and

$$\phi_0(\mathbf{x}) = 1. \quad (3.94)$$

Then, we can construct a new basis set $\psi_j(\mathbf{x})$ such that

$$\boldsymbol{\Psi}^\top \boldsymbol{\Psi} = \mathbf{I}? \quad (3.95)$$

$$\sum_{n=1}^N \psi_j(\mathbf{x}_n) \psi_k(\mathbf{x}_n) = I_{jk} \quad (3.96)$$

where

$$\mathbf{\Psi} = \begin{bmatrix} \boldsymbol{\psi}(\mathbf{x}_1)^\top \\ \vdots \\ \boldsymbol{\psi}(\mathbf{x}_N)^\top \end{bmatrix}$$

and

$$\psi_0(\mathbf{x}) = 1. \quad (3.97)$$

Under the basis set, if $\alpha = 0$, then

$$\mathbf{S}_N^{-1} = \beta \mathbf{I}, \quad (3.98)$$

so that

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}'). \quad (3.99)$$

Then,

$$\sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^N \sum_{j=0}^{M-1} \psi_j(\mathbf{x}) \psi_j(\mathbf{x}_n) = 1? \quad (3.100)$$

3.15

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.101)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N), \quad (3.102)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.103)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \quad (3.104)$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N. \quad (3.105)$$

By 3.22, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α and β to zero gives

$$\begin{aligned} \alpha &= \frac{\gamma}{\mathbf{m}_N^\top \mathbf{m}_N}, \\ \beta &= \frac{N - \gamma}{\|\mathbf{t} - \Phi \mathbf{m}_N\|^2}, \end{aligned} \quad (3.106)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m} \quad (3.107)$$

and $\lambda_1, \dots, \lambda_M$ are the eigenvalues of $\beta \Phi^\top \Phi$. If α and β are set as above, then

$$E(\mathbf{m}_N) = \frac{N}{2}. \quad (3.108)$$

3.16

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.109)$$

where \mathbf{w} and ϕ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}). \quad (3.110)$$

Integrating both sides with respect to \mathbf{w} gives

$$p(\mathbf{t}) = \int p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}. \quad (3.111)$$

The logarithm of the integrand of the right hand side except the terms independent of \mathbf{w} can be written as

$$-\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} = -\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}^\top \begin{bmatrix} \alpha \mathbf{I} + \beta \Phi^\top \Phi & -\beta \Phi^\top \\ -\beta \Phi & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}.$$

By 2.24,

$$\begin{bmatrix} \alpha \mathbf{I} + \beta \Phi^\top \Phi & -\beta \Phi^\top \\ -\beta \Phi & \beta \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \alpha^{-1} \mathbf{I} & \alpha^{-1} \Phi^\top \\ \alpha^{-1} \Phi & \alpha^{-1} \Phi \Phi^\top + \beta^{-1} \mathbf{I} \end{bmatrix}.$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t} | \mathbf{0}, \alpha^{-1} \Phi \Phi^\top + \beta^{-1} \mathbf{I}). \quad (3.112)$$

3.17

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.113)$$

where \mathbf{w} and ϕ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}). \quad (3.114)$$

Then,

$$p(\mathbf{t}) = \int p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}) d\mathbf{w}. \quad (3.115)$$

The logarithm of the integrand of the right hand side can be written as

$$-\frac{N}{2} \ln(2\pi\beta^{-1}) - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\alpha^{-1} \mathbf{I})) - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.116)$$

Therefore,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp(-E(\mathbf{w})) d\mathbf{w}, \quad (3.117)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.118)$$

3.18

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.119)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.120)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.121)$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi} \right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi} \right)^{\frac{M}{2}} \int \exp(-E(\mathbf{w})) d\mathbf{w}, \quad (3.122)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.123)$$

The first term of the definition of $E(\mathbf{w})$ can be written as

$$\begin{aligned} & \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N - \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N)\|^2 \\ &= \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2 - \beta(\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N)^\top \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N). \end{aligned} \quad (3.124)$$

Similarly, the second term can be written as

$$\begin{aligned} & \frac{\alpha}{2} (\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N)^\top (\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N) \\ &= \frac{\alpha}{2} (\mathbf{w} - \mathbf{m}_N)^\top (\mathbf{w} - \mathbf{m}_N) + \alpha \mathbf{m}_N^\top (\mathbf{w} - \mathbf{m}_N) + \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N. \end{aligned} \quad (3.125)$$

Here,

$$\begin{aligned} & -\beta(\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N)^\top \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \alpha \mathbf{m}_N^\top (\mathbf{w} - \mathbf{m}_N) \\ &= (-\beta \boldsymbol{\Phi}^\top \mathbf{t} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{m}_N + \alpha \mathbf{m}_N)^\top (\mathbf{w} - \mathbf{m}_N). \end{aligned} \quad (3.126)$$

By the definitions of \mathbf{m}_N and \mathbf{S}_N above, the right hand can be written as

$$(-\beta \boldsymbol{\Phi}^\top \mathbf{t} + \mathbf{S}_N^{-1} \mathbf{m}_N)^\top (\mathbf{w} - \mathbf{m}_N) = 0. \quad (3.127)$$

Therefore,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N). \quad (3.128)$$

3.19

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.129)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.130)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.131)$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp(-E(\mathbf{w})) d\mathbf{w}, \quad (3.132)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{w}\|^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}. \quad (3.133)$$

By 3.18,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N). \quad (3.134)$$

Therefore, the integral in the expression above of $p(\mathbf{t})$ can be written as

$$\begin{aligned} & \exp(-E(\mathbf{m}_N)) \int \exp\left(-\frac{1}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N)\right) d\mathbf{w} \\ &= (2\pi)^{\frac{M}{2}} (\det \mathbf{S}_N)^{\frac{1}{2}} \exp(-E(\mathbf{m}_N)). \end{aligned} \quad (3.135)$$

Thus,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N). \quad (3.136)$$

3.20

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}(t_n|\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \quad (3.137)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \quad (3.138)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.139)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \quad (3.140)$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N. \quad (3.141)$$

Let $\mathbf{u}_1, \dots, \mathbf{u}_M$ be eigenvectors of $\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ such that

$$\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{u}_m = \lambda_m \mathbf{u}_m. \quad (3.142)$$

Then,

$$\mathbf{S}_N^{-1} \mathbf{u}_m = (\alpha + \lambda_m) \mathbf{u}_m, \quad (3.143)$$

so that

$$\det \mathbf{S}_N = \prod_{m=1}^M \frac{1}{\alpha + \lambda_m}. \quad (3.144)$$

Therefore, setting the derivative of $\ln p(\mathbf{t}|\alpha, \beta)$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \sum_{m=1}^M \frac{1}{\alpha + \lambda_m} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N. \quad (3.145)$$

Multiplying both sides by 2α gives

$$\alpha \mathbf{m}_N^\top \mathbf{m}_N = M - \sum_{m=1}^M \frac{\alpha}{\alpha + \lambda_m}. \quad (3.146)$$

The right hand side can be written as

$$\sum_{m=1}^M \left(1 - \frac{\alpha}{\alpha + \lambda_m} \right) = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.147)$$

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N}, \quad (3.148)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_i}{\alpha + \lambda_m}. \quad (3.149)$$

3.21

(a)

Let Σ be a $M \times M$ real symmetric matrix such that

$$\Sigma \mathbf{u}_m = \lambda_m \mathbf{u}_m, \quad (3.150)$$

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

$$\begin{aligned} \Lambda &= \text{diag}(\lambda_1, \dots, \lambda_M), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_M]. \end{aligned} \quad (3.151)$$

By 2.19,

$$\begin{aligned} \Sigma &= \mathbf{U} \Lambda \mathbf{U}^T, \\ \mathbf{U}^T \mathbf{U} &= \mathbf{I}. \end{aligned} \quad (3.152)$$

Therefore,

$$\det \Sigma = \prod_{m=1}^M \lambda_m, \quad (3.153)$$

so that

$$\ln(\det \Sigma) = \sum_{m=1}^M \ln \lambda_i. \quad (3.154)$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \Sigma) = \sum_{m=1}^M \frac{\partial \lambda_m}{\partial \alpha} \frac{1}{\lambda_m}. \quad (3.155)$$

Therefore,

$$\frac{\partial}{\partial \alpha} \ln(\det \Sigma) = \text{tr} \left(\Lambda^{-1} \frac{\partial \Lambda}{\partial \alpha} \right). \quad (3.156)$$

The right hand side can be written as

$$\text{tr} \left(\mathbf{U} \mathbf{\Lambda}^{-1} \mathbf{U}^\top \frac{\partial \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top}{\partial \alpha} \right) = \text{tr} \left(\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha} \right). \quad (3.157)$$

Therefore,

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{\Sigma}) = \text{tr} \left(\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha} \right). \quad (3.158)$$

(b)

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.159)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N), \quad (3.160)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned} \quad (3.161)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \quad (3.162)$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N. \quad (3.163)$$

By 3.21(a),

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{S}_N^{-1}) = \text{tr}(\mathbf{S}_N). \quad (3.164)$$

The right hand side can be written as

$$\sum_{m=1}^M \frac{1}{\alpha + \lambda_m}, \quad (3.165)$$

where $\lambda_1, \dots, \lambda_M$ are eigenvalues of $\beta \Phi^\top \Phi$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \sum_{m=1}^M \frac{1}{\alpha + \lambda_m} - \frac{1}{2} \mathbf{m}_N^\top \mathbf{m}_N, \quad (3.166)$$

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^\top \mathbf{m}_N}, \quad (3.167)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.168)$$

3.22

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I}), \end{aligned} \quad (3.169)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N), \quad (3.170)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \Phi^\top \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \Phi^\top \Phi. \end{aligned} \quad (3.171)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \quad (3.172)$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2} \|\mathbf{t} - \Phi \mathbf{m}_N\|^2 + \frac{\alpha}{2} \mathbf{m}_N^\top \mathbf{m}_N. \quad (3.173)$$

By 3.21(a),

$$\frac{\partial}{\partial \beta} \ln(\det \mathbf{S}_N^{-1}) = \text{tr}(\mathbf{S}_N \Phi^\top \Phi). \quad (3.174)$$

Since

$$\mathbf{S}_N \boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \frac{1}{\beta} (\mathbf{I} - \alpha \mathbf{S}_N), \quad (3.175)$$

the right hand side can be written as

$$\frac{1}{\beta} \left(M - \alpha \sum_{m=1}^M \frac{1}{\alpha + \lambda_m} \right) = \frac{1}{\beta} \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}, \quad (3.176)$$

where $\lambda_1, \dots, \lambda_M$ are eigenvalues of $\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to β to zero gives

$$0 = \frac{N}{2\beta} - \frac{1}{2\beta} \sum_{m=1}^M \frac{\lambda_i}{\alpha + \lambda_m} - \frac{1}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2. \quad (3.177)$$

Thus,

$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2}, \quad (3.178)$$

where

$$\gamma = \sum_{m=1}^M \frac{\lambda_m}{\alpha + \lambda_m}. \quad (3.179)$$

3.23

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}, \beta) &= \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}), \\ p(\mathbf{w}, \beta) &= \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) \text{Gam}(\beta | a_0, b_0), \end{aligned} \quad (3.180)$$

where \mathbf{w} and $\boldsymbol{\phi}$ are vectors in M dimensions. By marginalisation,

$$p(\mathbf{t}) = \int \int p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta) d\mathbf{w} d\beta. \quad (3.181)$$

The right hand side can be written as

$$\int \left(\int \left(\prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \right) \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0) d\mathbf{w} \right) \text{Gam}(\beta | a_0, b_0) d\beta. \quad (3.182)$$

The logarithm of the integrand with respect to \mathbf{w} can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \beta^{-1} - \frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n))^2 \\
& - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln \det(\beta^{-1} \mathbf{S}_0) - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\
& = -\frac{N+M}{2} \ln(2\pi) + \frac{N+M}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) \\
& - \frac{\beta}{2} \mathbf{w}^\top (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \mathbf{w} + \beta \mathbf{w}^\top (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}) - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0.
\end{aligned} \tag{3.183}$$

The right hand side can be written as

$$\begin{aligned}
& -\frac{N+M}{2} \ln(2\pi) + \frac{N+M}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) \\
& - \frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0,
\end{aligned} \tag{3.184}$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{t}), \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi}.
\end{aligned} \tag{3.185}$$

Therefore, the logarithm of the integral with respect to \mathbf{w} can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) + \frac{N}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}_N) \\
& + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0.
\end{aligned} \tag{3.186}$$

Then, the logarithm of the integrand with respect to β can be written as

$$\begin{aligned}
& -\frac{N}{2} \ln(2\pi) + \frac{N}{2} \ln \beta - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}_N) \\
& + \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} \|\mathbf{t}\|^2 - \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 \\
& - \ln \Gamma(a_0) + a_0 \ln b_0 + (a_0 - 1) \ln \beta - b_0 \beta \\
& = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}_N) \\
& - \ln \Gamma(a_0) + a_0 \ln b_0 + (a_N - 1) \ln \beta - b_N \beta,
\end{aligned} \tag{3.187}$$

where

$$\begin{aligned} a_N &= a_0 + \frac{N}{2}, \\ b_N &= b_0 + \frac{\beta}{2} \|\mathbf{t}\|^2 + \frac{\beta}{2} \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 - \frac{\beta}{2} \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N. \end{aligned} \quad (3.188)$$

Therefore, the logarithm of the integral with respect to β can be written as

$$-\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{S}_0) + \frac{1}{2} \ln(\det \mathbf{S}_N) - \ln \Gamma(a_0) + a_0 \ln b_0 + \ln \Gamma(a_N) - a_N \ln b_N. \quad (3.189)$$

Thus,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left(\frac{\det \mathbf{S}_N}{\det \mathbf{S}_0} \right)^{\frac{1}{2}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{b_0^{a_0}}{b_N^{a_N}}. \quad (3.190)$$

3.24

Refer to 3.12.

4 Linear Models for Classification

4.1

Let x_1, \dots, x_M and y_1, \dots, y_N be two sets of data points. Then, the corresponding convex hulls are defined as the sets of all points \mathbf{x} and \mathbf{y} such that

$$\begin{aligned}\mathbf{x} &= \sum_{m=1}^M \alpha_m \mathbf{x}_m, \\ \mathbf{y} &= \sum_{n=1}^N \beta_n \mathbf{y}_n,\end{aligned}\tag{4.1}$$

where

$$\begin{aligned}\sum_{m=1}^M \alpha_m &= \sum_{n=1}^N \beta_n = 1, \\ \alpha_m &\geq 0, \beta_n \geq 0.\end{aligned}\tag{4.2}$$

Let us assume that $\alpha_1, \dots, \alpha_M$ and β_1, \dots, β_N below are subject to the constraints above.

If the convex hulls intersect, then there exist $\alpha_1, \dots, \alpha_M$ and β_1, \dots, β_N such that

$$\sum_{m=1}^M \alpha_m \mathbf{x}_m = \sum_{n=1}^N \beta_n \mathbf{y}_n.\tag{4.3}$$

Then,

$$\sum_{m=1}^M \alpha_m (\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0) = \hat{\mathbf{w}}^\top \sum_{m=1}^M \alpha_m \mathbf{x}_m + w_0 \sum_m \alpha_m,\tag{4.4}$$

for any $\hat{\mathbf{w}}$ and w_0 . The right hand side can be written as

$$\hat{\mathbf{w}}^\top \sum_{n=1}^N \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^N \beta_n = \sum_{n=1}^N \beta_n (\hat{\mathbf{w}}^\top \mathbf{y}_n + w_0).\tag{4.5}$$

Therefore, there do not exist $\hat{\mathbf{w}}$ and w_0 such that

$$\begin{aligned}\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0 &> 0, \\ \hat{\mathbf{w}}^\top \mathbf{y}_n + w_0 &< 0.\end{aligned}\tag{4.6}$$

Conversely, if there exist $\hat{\mathbf{w}}$ and w_0 such that

$$\begin{aligned}\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0 &> 0, \\ \hat{\mathbf{w}}^\top \mathbf{y}_n + w_0 &< 0,\end{aligned}\tag{4.7}$$

then

$$\begin{aligned}\sum_{m=1}^M \alpha_m (\hat{\mathbf{w}}^\top \mathbf{x}_m + w_0) &> 0, \\ \sum_{n=1}^N \beta_n (\hat{\mathbf{w}}^\top \mathbf{y}_n + w_0) &< 0.\end{aligned}\tag{4.8}$$

The left hand sides can be written as

$$\begin{aligned}\hat{\mathbf{w}}^\top \sum_{m=1}^M \alpha_m \mathbf{x}_m + w_0 \sum_{m=1}^M \alpha_m &= \hat{\mathbf{w}}^\top \sum_{m=1}^M \alpha_m \mathbf{x}_m + w_0, \\ \hat{\mathbf{w}}^\top \sum_{n=1}^N \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^N \beta_n &= \hat{\mathbf{w}}^\top \sum_{n=1}^N \beta_n \mathbf{y}_n + w_0.\end{aligned}\tag{4.9}$$

Therefore, there do not exist $\alpha_1, \dots, \alpha_M$ and β_1, \dots, β_N such that

$$\sum_{m=1}^M \alpha_m \mathbf{x}_m = \sum_{n=1}^N \beta_n \mathbf{y}_n.\tag{4.10}$$

Thus, the convex hulls do not intersect.

4.2 (Incomplete)

Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{w}_1, \dots, \mathbf{w}_K$ are variables in M dimensions and $\mathbf{t}_1, \dots, \mathbf{t}_N$ are ones in K dimensions. Let

$$E(\tilde{\mathbf{W}}) = \frac{1}{2} \text{tr} \left((\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^\top (\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}) \right),\tag{4.11}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\top \end{bmatrix},$$

$$\tilde{\mathbf{W}} = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^\top \\ \vdots \\ \mathbf{t}_N^\top \end{bmatrix}.$$

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

$$\mathbf{0} = \tilde{\mathbf{X}}^\top (\tilde{\mathbf{X}} \tilde{\mathbf{W}} - \mathbf{T}). \quad (4.12)$$

Therefore,

$$\underset{\tilde{\mathbf{W}}}{\operatorname{argmin}} E(\tilde{\mathbf{W}}) = \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^\top \mathbf{T}. \quad (4.13)$$

Let $\tilde{\mathbf{W}}^*$ denote the least-square solution above. Then,

$$(\tilde{\mathbf{W}}^*)^\top \tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{T}^\top \tilde{\mathbf{X}} \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{x}} - \mathbf{t}_n, \quad (4.14)$$

where $\tilde{\mathbf{x}}$ is a vector in $M + 1$ dimensions whose first element is 1. The right hand side can be written as

$$\mathbf{T}^\top \left(\tilde{\mathbf{X}} \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{x}} - \mathbf{v}_n \right) = \mathbf{0}, \quad (4.15)$$

where \mathbf{v}_n is a vector in N dimensions whose n th element is 1 and other elements are zero. Therefore,

$$(\tilde{\mathbf{W}}^*)^\top \tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{0}. \quad (4.16)$$

Thus, if

$$\mathbf{a}^\top \mathbf{t}_n + b = 0, \quad (4.17)$$

then

$$\mathbf{a}^\top (\tilde{\mathbf{W}}^*)^\top \tilde{\mathbf{x}} + b = 0. \quad (4.18)$$

4.3 (Incomplete)

4.4

Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \quad (4.19)$$

where N_k is the number of \mathbf{x}_n such that n is in \mathcal{C}_k . Setting the derivatives of

$$\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1) + \lambda (\|\mathbf{w}\|^2 - 1) \quad (4.20)$$

with respect to \mathbf{w} and λ to zero gives

$$\begin{aligned} \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} &= \mathbf{0}, \\ \|\mathbf{w}\|^2 - 1 &= 0. \end{aligned} \quad (4.21)$$

Therefore, $\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1)$ under the constraint

$$\|\mathbf{w}\|^2 = 1 \quad (4.22)$$

is maximised if

$$\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1. \quad (4.23)$$

4.5

Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \quad (4.24)$$

where N_k is the number of \mathbf{x}_n such that n is in \mathcal{C}_k . Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \quad (4.25)$$

where

$$\begin{aligned} s_k^2 &= \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2, \\ y_n &= \mathbf{w}^\top \mathbf{x}_n, \\ m_k &= \mathbf{w}^\top \mathbf{m}_k. \end{aligned} \quad (4.26)$$

Then, $J(\mathbf{w})$ can be written as

$$\frac{(\mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1))^2}{\sum_{n \in \mathcal{C}_1} (\mathbf{w}^\top(\mathbf{x}_n - \mathbf{m}_1))^2 + \sum_{n \in \mathcal{C}_2} (\mathbf{w}^\top(\mathbf{x}_n - \mathbf{m}_2))^2} = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}, \quad (4.27)$$

where

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top, \\ \mathbf{S}_W &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top. \end{aligned} \quad (4.28)$$

4.6

Let $\mathbf{x}_1, \dots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \quad (4.29)$$

where N_k is the number of \mathbf{x}_n such that n is in \mathcal{C}_k . Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \quad (4.30)$$

where

$$\begin{aligned} s_k^2 &= \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2, \\ y_n &= \mathbf{w}^\top \mathbf{x}_n, \\ m_k &= \mathbf{w}^\top \mathbf{m}_k. \end{aligned} \quad (4.31)$$

Then, by 4.5,

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}}, \quad (4.32)$$

where

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top, \\ \mathbf{S}_W &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top. \end{aligned} \quad (4.33)$$

Let

$$E = \frac{1}{2} \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n)^2, \quad (4.34)$$

where

$$t_n = \begin{cases} \frac{N}{N_1}, & n \in \mathcal{C}_1, \\ -\frac{N}{N_2}, & n \in \mathcal{C}_2. \end{cases} \quad (4.35)$$

Setting the derivative with respect to \mathbf{w} and w_0 gives

$$\begin{aligned} 0 &= \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n), \\ 0 &= \sum_{n=1}^N (\mathbf{w}^\top \mathbf{x}_n + w_0 - t_n) \mathbf{x}_n. \end{aligned} \quad (4.36)$$

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

$$\mathbf{w}^\top \sum_{n=1}^N \mathbf{x}_n + Nw_0 - \sum_{n=1}^N t_n = N(\mathbf{w}^\top \mathbf{m} + w_0), \quad (4.37)$$

where

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n. \quad (4.38)$$

Therefore,

$$w_0 = -\mathbf{w}^\top \mathbf{m}. \quad (4.39)$$

Then, the right hand side of the second equation above can be written as

$$\begin{aligned} & \sum_{n=1}^N (\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}) - t_n) \mathbf{x}_n \\ &= \sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}) - \frac{N}{N_1} \right) \mathbf{x}_n + \sum_{n \in \mathcal{C}_2} \left(\mathbf{w}^\top (\mathbf{x}_n - \mathbf{m}) + \frac{N}{N_2} \right) \mathbf{x}_n. \end{aligned} \quad (4.40)$$

Since

$$\begin{aligned} \mathbf{m} &= \frac{N_1}{N} \mathbf{m}_1 + \frac{N_2}{N} \mathbf{m}_2, \\ \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1) &= \mathbf{0}, \end{aligned} \quad (4.41)$$

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

$$\begin{aligned} & \sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^\top \left(\mathbf{x}_n - \mathbf{m}_1 + \frac{N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \right) - \frac{N}{N_1} \right) (\mathbf{x}_n - \mathbf{m}_1 + \mathbf{m}_1) \\ &= \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^\top \mathbf{w} - N \mathbf{m}_1. \end{aligned} \quad (4.42)$$

Similarly, the second term can be written as

$$\left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^\top \mathbf{w} - N \mathbf{m}_2. \quad (4.43)$$

Therefore,

$$\begin{aligned} \mathbf{0} = & \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^\top \mathbf{w} - N \mathbf{m}_1 \\ & + \left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^\top \mathbf{w} - N \mathbf{m}_2. \end{aligned} \quad (4.44)$$

Thus,

$$\left(\mathbf{S}_W + \frac{N_1 N_2}{N} \mathbf{S}_B \right) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2). \quad (4.45)$$

4.7

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (4.46)$$

Then,

$$\sigma(-a) = \frac{1}{1 + \exp(a)}. \quad (4.47)$$

The right hand side can be written as

$$1 - \frac{\exp(a)}{1 + \exp(a)} = 1 - \frac{1}{1 + \exp(-a)}. \quad (4.48)$$

Therefore,

$$\sigma(-a) = 1 - \sigma(a). \quad (4.49)$$

Additionally,

$$\exp(-a) = \frac{1}{\sigma(a)} - 1. \quad (4.50)$$

Then,

$$a = -\ln \left(\frac{1}{\sigma(a)} - 1 \right). \quad (4.51)$$

Therefore,

$$\sigma^{-1}(y) = \ln \left(\frac{y}{1 - y} \right). \quad (4.52)$$

4.8

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

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad (4.53)$$

where

$$p(\mathcal{C}_1) + p(\mathcal{C}_2) = 1. \quad (4.54)$$

By the Bayes' theorem,

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}. \quad (4.55)$$

The right hand side can be written as

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \quad (4.56)$$

where

$$a = \ln \left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \right). \quad (4.57)$$

Substituting the expressions above of $p(\mathbf{x}|\mathcal{C}_k)$, we have

$$\begin{aligned} a = & -\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \ln p(\mathcal{C}_1) \\ & + \frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \boldsymbol{\Sigma}) + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) - \ln p(\mathcal{C}_2). \end{aligned} \quad (4.58)$$

Therefore,

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + w_0), \quad (4.59)$$

where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ w_0 &= -\frac{1}{2} \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2). \end{aligned} \quad (4.60)$$

4.9

Let $\mathbf{t}_1, \dots, \mathbf{t}_N$ be variables in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^K (p(\boldsymbol{\phi}_n, \mathcal{C}_k))^{t_{nk}}, \quad (4.61)$$

where

$$\sum_{k=1}^K p(\mathcal{C}_k) = 1. \quad (4.62)$$

Then,

$$p(\mathbf{T}, \mathbf{\Phi}) = \prod_{n=1}^N \prod_{k=1}^K (p(\phi_n, \mathcal{C}_k))^{t_{nk}}. \quad (4.63)$$

If

$$p(\mathcal{C}_k) = \pi_k, \quad (4.64)$$

then, by the Bayes' theorem,

$$\ln p(\mathbf{T}, \mathbf{\Phi}) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \pi_k + \ln p(\phi_n | \mathcal{C}_k)). \quad (4.65)$$

Setting the derivatives of

$$\ln p(\mathbf{T}, \mathbf{\Phi}) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \quad (4.66)$$

with respect to π_k and λ to zero gives

$$\begin{aligned} 0 &= \frac{1}{\pi_k} \sum_{n=1}^N t_{nk} + \lambda, \\ 0 &= \sum_{k=1}^K \pi_k - 1. \end{aligned} \quad (4.67)$$

Then,

$$\lambda = - \sum_{k=1}^K \sum_{n=1}^N t_{nk}. \quad (4.68)$$

The right hand side can be written as $-N$. Therefore, the maximum likelihood solution for π_k is given by

$$\pi_k = \frac{N_k}{N}, \quad (4.69)$$

where

$$N_k = \sum_{n=1}^N t_{nk}. \quad (4.70)$$

4.10

Let $\mathbf{t}_1, \dots, \mathbf{t}_N$ be variables in K dimensions such that

$$p(\mathbf{t}_n, \phi_n) = \prod_{k=1}^K (p(\phi_n, \mathcal{C}_k))^{t_{nk}}, \quad (4.71)$$

where

$$\sum_{k=1}^K p(\mathcal{C}_k) = 1. \quad (4.72)$$

Then,

$$p(\mathbf{T}, \Phi) = \prod_{n=1}^N \prod_{k=1}^K (p(\phi_n, \mathcal{C}_k))^{t_{nk}}. \quad (4.73)$$

If

$$p(\phi_n | \mathcal{C}_k) = \mathcal{N}(\phi_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad (4.74)$$

then, by the Bayes' theorem,

$$\ln p(\mathbf{T}, \Phi) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} (\ln \mathcal{N}(\phi_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) + \ln p(\mathcal{C}_k)). \quad (4.75)$$

The right hand side can be written as

$$\sum_{n=1}^N \sum_{k=1}^K t_{nk} \left(-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} (\phi_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\phi_n - \boldsymbol{\mu}_k) + \ln p(\mathcal{C}_k) \right). \quad (4.76)$$

By 3.21(a), setting the derivatives of $\ln p(\mathbf{T}, \Phi)$ with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ to zero gives

$$\begin{aligned} \mathbf{0} &= \frac{1}{2} \sum_{n=1}^N t_{nk} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\phi_n - \boldsymbol{\mu}_k), \\ \mathbf{O} &= -\frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K t_{nk} \left((\boldsymbol{\Sigma}^{-1})^\top - (\boldsymbol{\Sigma}^{-1})^2 (\phi_n - \boldsymbol{\mu}_k)(\phi_n - \boldsymbol{\mu}_k)^\top \right). \end{aligned} \quad (4.77)$$

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

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^N t_{nk} \boldsymbol{\phi}_n, \\ \boldsymbol{\Sigma} &= \frac{1}{N} \sum_{k=1}^K N_k \mathbf{S}_k,\end{aligned}\tag{4.78}$$

where

$$\begin{aligned}N_k &= \sum_{n=1}^N t_{nk}, \\ \mathbf{S}_k &= \frac{1}{N_k} \sum_{n=1}^N t_{nk} (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\top.\end{aligned}\tag{4.79}$$

4.11

Let $\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_M$ be variables such that

$$p(\boldsymbol{\phi}_m | \mathcal{C}_k) = \prod_{l=1}^L \mu_{kml}^{\phi_{ml}},\tag{4.80}$$

where

$$\sum_{k=1}^K p(\mathcal{C}_k) = 1.\tag{4.81}$$

Then,

$$p(\boldsymbol{\Phi} | \mathcal{C}_k) = \prod_{m=1}^M \prod_{l=1}^L \mu_{kml}^{\phi_{ml}}.\tag{4.82}$$

By the Bayes' theorem,

$$p(\mathcal{C}_k | \boldsymbol{\Phi}) = \frac{p(\boldsymbol{\Phi} | \mathcal{C}_k) p(\mathcal{C}_k)}{\sum_{k=1}^K p(\boldsymbol{\Phi} | \mathcal{C}_k) p(\mathcal{C}_k)}.\tag{4.83}$$

Therefore,

$$p(\mathcal{C}_k | \boldsymbol{\Phi}) = \frac{\exp(a_k(\boldsymbol{\Phi}))}{\sum_{k=1}^K \exp(a_k(\boldsymbol{\Phi}))},\tag{4.84}$$

where

$$a_k(\boldsymbol{\Phi}) = \left(\sum_{m=1}^M \sum_{l=1}^L \phi_{ml} \ln \mu_{kml} \right) + \ln p(\mathcal{C}_k).\tag{4.85}$$

4.12

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (4.86)$$

Then,

$$\frac{d\sigma(a)}{da} = \frac{\exp(-a)}{(1 + \exp(-a))^2}. \quad (4.87)$$

The right hand side can be written as

$$\frac{1}{1 + \exp(-a)} - \frac{1}{(1 + \exp(-a))^2} = \sigma(a) - (\sigma(a))^2. \quad (4.88)$$

Therefore,

$$\frac{d\sigma(a)}{da} = \sigma(a) (1 - \sigma(a)). \quad (4.89)$$

4.13

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n | \mathbf{w}) &= y_n^{t_n} (1 - y_n)^{1-t_n}, \end{aligned} \quad (4.90)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \phi_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.91)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \mathbf{w}). \quad (4.92)$$

The right hand side can be written as

$$-\ln \left(\prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n} \right) = -\sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)). \quad (4.93)$$

Then, by 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \left(\frac{t_n}{y_n} y_n (1 - y_n) \phi_n - \frac{1 - t_n}{1 - y_n} y_n (1 - y_n) \phi_n \right). \quad (4.94)$$

The right hand side can be written as

$$-\sum_{n=1}^N (t_n(1 - y_n)\phi_n - (1 - t_n)y_n\phi_n) = \sum_{n=1}^N (y_n - t_n)\phi_n. \quad (4.95)$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n)\phi_n. \quad (4.96)$$

4.14

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n}(1 - y_n)^{1-t_n}, \end{aligned} \quad (4.97)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \phi_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.98)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \quad (4.99)$$

By 4.13, setting the derivative with respect to \mathbf{w} to zero gives

$$\mathbf{0} = \sum_{n=1}^N (y_n - t_n) \phi_n. \quad (4.100)$$

If ϕ_1, \dots, ϕ_N are linearly independent, then

$$y_n = t_n. \quad (4.101)$$

Then,

$$\sigma(\mathbf{w}^\top \phi_n) = \begin{cases} 1, & t_n = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (4.102)$$

Therefore,

$$\mathbf{w}^\top \phi_n = \begin{cases} \infty, & t_n = 1, \\ -\infty, & \text{otherwise.} \end{cases} \quad (4.103)$$

4.15

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n} (1 - y_n)^{1-t_n}, \end{aligned} \quad (4.104)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \phi_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.105)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \quad (4.106)$$

By 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \phi_n. \quad (4.107)$$

Then, by 4.12,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n (1 - y_n) \phi_n \phi_n^\top. \quad (4.108)$$

The right hand side can be written as

$$\mathbf{H} = \Phi^\top \mathbf{R} \Phi, \quad (4.109)$$

where

$$R_{nn'} = \begin{cases} y_n(1 - y_n), & n = n', \\ 0, & \text{otherwise.} \end{cases} \quad (4.110)$$

Then,

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} = (\Phi \mathbf{u})^\top \mathbf{R} (\Phi \mathbf{u}). \quad (4.111)$$

Since

$$y_n(1 - y_n) > 0, \quad (4.112)$$

we have

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} > 0. \quad (4.113)$$

Therefore, \mathbf{H} is positive definite. Thus, E is a convex function of \mathbf{w} and it has a unique minimum.

4.16

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1 | \phi_n) &= \pi_n. \end{aligned} \quad (4.114)$$

Then,

$$p(t_n | \phi_n) = \pi_n^{t_n} (1 - \pi_n)^{1-t_n}. \quad (4.115)$$

Therefore,

$$p(\mathbf{t} | \Phi) = \prod_{n=1}^N \pi_n^{t_n} (1 - \pi_n)^{1-t_n}. \quad (4.116)$$

Thus,

$$-\ln p(\mathbf{t} | \Phi) = -\sum_{n=1}^N (t_n \ln \pi_n + (1 - t_n) \ln(1 - \pi_n)). \quad (4.117)$$

4.17

Let

$$y_k = \frac{\exp(a_k)}{\sum_{k=1}^K \exp(a_k)}. \quad (4.118)$$

Then,

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k)}{\sum_{k=1}^K \exp(a_k)} - \frac{\exp(2a_k)}{\left(\sum_{k=1}^K \exp(a_k)\right)^2}. \quad (4.119)$$

The right hand side can be written as $y_k(1 - y_k)$. If $k \neq k'$, then

$$\frac{\partial y_k}{\partial a_{k'}} = -\frac{\exp(a_k + a_{k'})}{\left(\sum_{k=1}^K \exp(a_k)\right)^2}. \quad (4.120)$$

The right hand side can be written as $-y_k y_{k'}$. Therefore,

$$\frac{\partial y_k}{\partial a_{k'}} = y_k(I_{kk'} - y_{k'}). \quad (4.121)$$

4.18

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

$$t_{nk} \in \{0, 1\},$$

$$p(\mathbf{t}_n | \mathbf{W}) = \prod_{k=1}^K y_{nk}^{t_{nk}}, \quad (4.122)$$

where

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_{k=1}^K \exp(a_{nk})}, \quad (4.123)$$

$$a_{nk} = \mathbf{w}_k^\top \phi_n.$$

Then,

$$p(\mathbf{T} | \mathbf{W}) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (4.124)$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T} | \mathbf{W}). \quad (4.125)$$

The right hand side can be written as

$$-\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (4.126)$$

Then, by 4.17,

$$\nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = -\sum_{n=1}^N \sum_{k=1}^K y_{nk} (I_{kk'} - y_{nk'}) \frac{t_{nk}}{y_{nk}} \phi_n. \quad (4.127)$$

The right hand side can be written as

$$-\sum_{n=1}^N \left(\sum_{k=1}^K (I_{kk'} - y_{nk'}) t_{nk} \right) \phi_n = -\sum_{n=1}^N (t_{nk'} - y_{nk'}) \phi_n. \quad (4.128)$$

Therefore,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^N (y_{nk} - t_{nk}) \phi_n. \quad (4.129)$$

4.19

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1 | a_n) &= \Phi(a_n), \end{aligned} \tag{4.130}$$

where

$$\begin{aligned} \Phi(a) &= \int_{-\infty}^a \mathcal{N}(\theta | 0, 1) d\theta, \\ a_n &= \mathbf{w}^\top \phi_n. \end{aligned} \tag{4.131}$$

Then,

$$p(t_n | \phi_n) = (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \tag{4.132}$$

Therefore,

$$p(\mathbf{t} | \Phi) = \prod_{n=1}^N (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \tag{4.133}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \phi). \tag{4.134}$$

The right hand side can be written as

$$-\sum_{n=1}^N (t_n \ln \Phi(a_n) + (1 - t_n) \ln (1 - \Phi(a_n))). \tag{4.135}$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \left(t_n \frac{\mathcal{N}(a_n | 0, 1)}{\Phi(a_n)} - (1 - t_n) \frac{\mathcal{N}(a_n | 0, 1)}{1 - \Phi(a_n)} \right) \phi_n. \tag{4.136}$$

The right hand side can be written as

$$\begin{aligned} &-\sum_{n=1}^N \left(\frac{t_n}{\Phi(a_n)} - \frac{1 - t_n}{1 - \Phi(a_n)} \right) \mathcal{N}(a_n | 0, 1) \phi_n \\ &= \sum_{n=1}^N \frac{\mathcal{N}(a_n | 0, 1)}{\Phi(a_n) (1 - \Phi(a_n))} (\Phi(a_n) - t_n) \phi_n. \end{aligned} \tag{4.137}$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N \frac{\mathcal{N}(a_n|0, 1)}{\Phi(a_n)(1 - \Phi(a_n))} (\Phi(a_n) - t_n) \phi_n. \quad (4.138)$$

Then,

$$\begin{aligned} \nabla \nabla E(\mathbf{w}) &= \sum_{n=1}^N \frac{-a_n \mathcal{N}(a_n|0, 1)}{\Phi(a_n)(1 - \Phi(a_n))} (\Phi(a_n) - t_n) \phi_n \phi_n^\top \\ &\quad - \sum_{n=1}^N \frac{(\mathcal{N}(a_n|0, 1))^2}{(\Phi(a_n))^2 (1 - \Phi(a_n))} (\Phi(a_n) - t_n) \phi_n \phi_n^\top \\ &\quad + \sum_{n=1}^N \frac{(\mathcal{N}(a_n|0, 1))^2}{\Phi(a_n)(1 - \Phi(a_n))^2} (\Phi(a_n) - t_n) \phi_n \phi_n^\top \\ &\quad + \sum_{n=1}^N \frac{(\mathcal{N}(a_n|0, 1))^2}{\Phi(a_n)(1 - \Phi(a_n))} \phi_n \phi_n^\top. \end{aligned} \quad (4.139)$$

Therefore,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N b_n \phi_n \phi_n^\top, \quad (4.140)$$

where

$$\begin{aligned} b_n &= \left(\frac{\mathcal{N}(a_n|0, 1)}{\Phi(a_n)(1 - \Phi(a_n))} \right)^2 ((\Phi(a_n))^2 - 2t_n \Phi(a_n) + t_n) \\ &\quad - \frac{\mathcal{N}(a_n|0, 1)}{\Phi(a_n)(1 - \Phi(a_n))} a_n (\Phi(a_n) - t_n). \end{aligned} \quad (4.141)$$

4.20 (Incomplete)

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

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(\mathbf{t}_n | \mathbf{W}) &= \prod_{k=1}^K y_{nk}^{t_{nk}}, \end{aligned} \quad (4.142)$$

where

$$\begin{aligned} y_{nk} &= \frac{\exp(a_{nk})}{\sum_{k=1}^K \exp(a_{nk})}, \\ a_{nk} &= \mathbf{w}_k^\top \boldsymbol{\phi}_n. \end{aligned} \quad (4.143)$$

Then,

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (4.144)$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{W}). \quad (4.145)$$

By 4.18,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^N (y_{nk} - t_{nk}) \boldsymbol{\phi}_n. \quad (4.146)$$

Additionally, by 4.17,

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = \sum_{n=1}^N y_{nk} (I_{kk'} - y_{nk'}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\top. \quad (4.147)$$

The right hand side can be written as

$$\mathbf{H}_{kk'} = \boldsymbol{\Phi}^\top \mathbf{R}_{kk'} \boldsymbol{\Phi}, \quad (4.148)$$

where

$$R_{kk'nn'} = \begin{cases} y_{nk} (I_{kk'} - y_{nk'}), & n = n', \\ 0, & \text{otherwise.} \end{cases} \quad (4.149)$$

Let

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{K1} & \cdots & \mathbf{H}_{KK} \end{bmatrix},$$

and

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_K \end{bmatrix},$$

where $\mathbf{u}_1, \dots, \mathbf{u}_K$ are vectors in the same dimension as \mathbf{w} . Then,

$$\mathbf{u}^\top \mathbf{H} \mathbf{u} = \sum_{k=1}^K \sum_{k'=1}^K \mathbf{u}_k^\top \mathbf{H}_{kk'} \mathbf{u}_{k'}, \quad (4.150)$$

Then, the right hand side can be written as

$$\sum_{k=1}^K \sum_{k'=1}^K (\Phi \mathbf{u}_k)^\top \mathbf{R}_{kk'} (\Phi \mathbf{u}_{k'}). \quad (4.151)$$

4.21

Let

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta. \quad (4.152)$$

The right hand side can be written as

$$\int_{-\infty}^0 \mathcal{N}(\theta|0, 1) d\theta + \int_0^a \mathcal{N}(\theta|0, 1) d\theta = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^a \exp\left(-\frac{\theta^2}{2}\right) d\theta. \quad (4.153)$$

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

$$\frac{1}{\sqrt{2\pi}} \int_0^{\frac{a}{\sqrt{2}}} \exp(-t^2) \sqrt{2} dt = \frac{1}{2} \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right), \quad (4.154)$$

where

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp(-t^2) dt. \quad (4.155)$$

Therefore,

$$\Phi(a) = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) \right). \quad (4.156)$$

4.22

Let $\boldsymbol{\theta}$ be a variable in M dimensions. By a Taylor expansion,

$$\ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})) \simeq \ln(p(\mathcal{D}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)) + \mathbf{v}(\boldsymbol{\theta}_0)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \quad (4.157)$$

where

$$\begin{aligned}\mathbf{v}(\boldsymbol{\theta}) &= \nabla \ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})), \\ \mathbf{A}(\boldsymbol{\theta}) &= -\nabla \nabla \ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})).\end{aligned}\tag{4.158}$$

Let $\boldsymbol{\theta}_{\text{MAP}}$ be a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then,

$$\ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})) \simeq \ln(p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}})) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^\top \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}}),\tag{4.159}$$

so that

$$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^\top \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})\right).\tag{4.160}$$

By marginalisation, integrating both sides with respect to $\boldsymbol{\theta}$ gives

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}}) \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^\top \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})\right) d\boldsymbol{\theta}.\tag{4.161}$$

The integral of the right hand side can be written as

$$(2\pi)^{\frac{M}{2}} (\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})^{-1})^{\frac{1}{2}} = (2\pi)^{\frac{M}{2}} (\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}}))^{-\frac{1}{2}}.\tag{4.162}$$

Therefore,

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}})(2\pi)^{\frac{M}{2}} (\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}}))^{-\frac{1}{2}},\tag{4.163}$$

so that

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \ln p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})).\tag{4.164}$$

4.23

Let $\boldsymbol{\theta}$ be a variable in M dimensions. By 4.22,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \ln p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})),\tag{4.165}$$

where $\boldsymbol{\theta}_{\text{MAP}}$ is a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and

$$\mathbf{A}(\boldsymbol{\theta}) = -\nabla \nabla \ln(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})).\tag{4.166}$$

If

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0), \quad (4.167)$$

then

$$\nabla \nabla \ln p(\boldsymbol{\theta}) = -\mathbf{V}_0^{-1}, \quad (4.168)$$

so that

$$\mathbf{A}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta}) + \mathbf{V}_0^{-1}, \quad (4.169)$$

where

$$\mathbf{H}(\boldsymbol{\theta}) = -\nabla \nabla \ln p(\mathcal{D}|\boldsymbol{\theta}). \quad (4.170)$$

Then, the right hand side of the approximation above can be written as

$$\begin{aligned} & \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det \mathbf{V}_0) - \frac{1}{2} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0)^\top \mathbf{V}_0^{-1} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0) \\ & + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\mathbf{H}(\boldsymbol{\theta}_{\text{MAP}}) + \mathbf{V}_0^{-1})) \\ = & \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \frac{1}{2} \ln(\det \mathbf{V}_0^{-1}) - \frac{1}{2} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0)^\top \mathbf{V}_0^{-1} (\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0) \\ & - \frac{1}{2} \ln(\det(\mathbf{H}(\boldsymbol{\theta}_{\text{MAP}}) + \mathbf{V}_0^{-1})). \end{aligned} \quad (4.171)$$

If \mathbf{V}_0^{-1} can be neglected, the right hand side can be written as

$$\ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{1}{2} \ln(\det \mathbf{H}(\boldsymbol{\theta}_{\text{MAP}})). \quad (4.172)$$

If each data point is independent and identically distributed, then

$$\mathbf{H}(\boldsymbol{\theta}) = N \bar{\mathbf{H}}(\boldsymbol{\theta}), \quad (4.173)$$

where

$$\bar{\mathbf{H}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^N \mathbf{H}_n(\boldsymbol{\theta}), \quad (4.174)$$

and $\mathbf{H}_n(\boldsymbol{\theta})$ is the one for each data point. Then,

$$\det \mathbf{H}(\boldsymbol{\theta}_{\text{MAP}}) = N^M \det \bar{\mathbf{H}}(\boldsymbol{\theta}_{\text{MAP}}). \quad (4.175)$$

Therefore,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2} \ln N. \quad (4.176)$$

4.24 (Incomplete)

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n} (1 - y_n)^{1-t_n}, \end{aligned} \quad (4.177)$$

where

$$\begin{aligned} y_n &= \sigma(\mathbf{w}^\top \phi_n), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \quad (4.178)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \quad (4.179)$$

If

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0), \quad (4.180)$$

then the logarithm of the right hand side except the terms independent of \mathbf{w} and \mathbf{t} can be written as

$$\sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0). \quad (4.181)$$

Then, by 4.22,

$$p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{S}_N)? \quad (4.182)$$

where \mathbf{w}_{MAP} is the maximum likelihood solution for $p(\mathbf{w})$ and

$$\mathbf{S}_N = -\nabla \nabla \ln p(\mathbf{w}|\mathbf{t}). \quad (4.183)$$

By marginalisation,

$$p(\mathcal{C}_1|\mathbf{t}) = \int p(\mathcal{C}_1|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \quad (4.184)$$

The logarithm of the integrand of the right hand side except the terms independent of \mathbf{w} can be approximated as

$$-\ln(1 + \exp(-\mathbf{w}^\top \phi)) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) = \quad (4.185)$$

4.25

Let

$$\begin{aligned}\sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ \Phi(a) &= \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta.\end{aligned}\tag{4.186}$$

By 4.12,

$$\frac{d\sigma(a)}{da} = \sigma(a) (1 - \sigma(a)).\tag{4.187}$$

On the other hand, the right hand side can be written as

$$\frac{d\Phi(\lambda a)}{da} = \lambda \mathcal{N}(\lambda a|0, 1).\tag{4.188}$$

Let us assume that

$$\left. \frac{d\sigma(a)}{da} \right|_{a=0} = \left. \frac{d\Phi(\lambda a)}{da} \right|_{a=0}.\tag{4.189}$$

Then,

$$\frac{1}{4} = \lambda (2\pi)^{-\frac{1}{2}}.\tag{4.190}$$

Therefore,

$$\lambda^2 = \frac{\pi}{8}.\tag{4.191}$$

4.26

Let

$$I(\mu) = \int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da,\tag{4.192}$$

where

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0, 1) d\theta.\tag{4.193}$$

By the transformation

$$z = \frac{a - \mu}{\sigma},\tag{4.194}$$

the right hand side can be written as

$$\int \Phi(\lambda(\mu + \sigma z)) \mathcal{N}(\mu + \sigma z|\mu, \sigma^2) \sigma dz = \int \Phi(\lambda(\mu + \sigma z)) \mathcal{N}(z|0, 1) dz.\tag{4.195}$$

Then,

$$\frac{\partial}{\partial \mu} I(\mu) = \lambda \int \mathcal{N}(\lambda(\mu + \sigma z)|0, 1) \mathcal{N}(z|0, 1) dz. \quad (4.196)$$

The logarithm of the integrand of the right hand side can be written as

$$\begin{aligned} & -\frac{1}{2} \ln(2\pi) - \frac{\lambda^2(\mu + \sigma z)^2}{2} - \frac{1}{2} \ln(2\pi) - \frac{z^2}{2} \\ & = -\ln(2\pi) - \frac{1 + \sigma^2 \lambda^2}{2} \left(z + \frac{\mu \sigma \lambda^2}{1 + \sigma^2 \lambda^2} \right)^2 + \frac{\mu^2 \sigma^2 \lambda^4}{2(1 + \sigma^2 \lambda^2)} - \frac{\mu^2 \lambda^2}{2}. \end{aligned} \quad (4.197)$$

The right hand side can be written as

$$\begin{aligned} & -\ln(2\pi) - \frac{1 + \sigma^2 \lambda^2}{2} \left(z + \frac{\mu \sigma \lambda^2}{1 + \sigma^2 \lambda^2} \right)^2 - \frac{\mu^2 \lambda^2}{2(1 + \sigma^2 \lambda^2)} \\ & = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(1 + \sigma^2 \lambda^2)^{-1} - \frac{1 + \sigma^2 \lambda^2}{2} \left(z + \frac{\mu \sigma \lambda^2}{1 + \sigma^2 \lambda^2} \right)^2 \\ & \quad - \ln \lambda - \frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\lambda^{-2} + \sigma^2) - \frac{\mu^2}{2(\lambda^{-2} + \sigma^2)}. \end{aligned} \quad (4.198)$$

Then, the integral can be written as

$$\begin{aligned} & \int \mathcal{N} \left(z \middle| -\frac{\mu \sigma \lambda^2}{1 + \sigma^2 \lambda^2}, (1 + \sigma^2 \lambda^2)^{-1} \right) \frac{1}{\lambda} \mathcal{N}(\mu|0, \lambda^{-2} + \sigma^2) dz \\ & = \frac{1}{\lambda} \mathcal{N}(\mu|0, \lambda^{-2} + \sigma^2). \end{aligned} \quad (4.199)$$

Therefore,

$$\frac{\partial}{\partial \mu} I(\mu) = \mathcal{N}(\mu|0, \lambda^{-2} + \sigma^2). \quad (4.200)$$

Integrating both sides with respect to μ gives

$$I(\mu) = \int_{-\infty}^{\mu} \mathcal{N}(m|0, \lambda^{-2} + \sigma^2) dm. \quad (4.201)$$

By the transformation

$$m' = \frac{m}{(\lambda^{-2} + \sigma^2)^{\frac{1}{2}}}, \quad (4.202)$$

the right hand side can be written as

$$\int_{-\infty}^{\frac{\mu}{(\lambda^{-2} + \sigma^2)^{\frac{1}{2}}}} (\lambda^{-2} + \sigma^2)^{-\frac{1}{2}} \mathcal{N}(m'|0, 1) (\lambda^{-2} + \sigma^2)^{\frac{1}{2}} dm' = \Phi \left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{\frac{1}{2}}} \right). \quad (4.203)$$

Therefore,

$$I(\mu) = \Phi \left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{\frac{1}{2}}} \right). \quad (4.204)$$

5 Neural Networks

5.1

Let

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\sum_{m=1}^M w_{km}^{(2)} \sigma \left(\sum_{d=1}^D w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) + w_{k0}^{(2)} \right), \quad (5.1)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \quad (5.2)$$

Here,

$$\sigma(a) = \frac{\exp\left(\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)}. \quad (5.3)$$

The right hand side can be written as

$$\tanh\left(\frac{a}{2}\right) + \sigma(-a) = \tanh\left(\frac{a}{2}\right) + 1 - \sigma(a). \quad (5.4)$$

Therefore,

$$\sigma(a) = \frac{1}{2} \left(1 + \tanh\left(\frac{a}{2}\right) \right). \quad (5.5)$$

Then, the argument of the right hand side can be written as

$$\begin{aligned} & \sum_{m=1}^M w_{km}^{(2)} \left(\frac{1}{2} \left(1 + \tanh \left(\frac{1}{2} \left(\sum_{d=1}^D w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) \right) \right) \right) + w_{k0}^{(2)} \\ &= \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} \tanh \left(\frac{1}{2} \sum_{d=1}^D w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)} \right) + \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} + w_{k0}^{(2)}. \end{aligned} \quad (5.6)$$

Therefore,

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} \tanh \left(\frac{1}{2} \sum_{d=1}^D w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)} \right) + \frac{1}{2} \sum_{m=1}^M w_{km}^{(2)} + w_{k0}^{(2)} \right). \quad (5.7)$$

5.2

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

$$p(\mathbf{t}_n | \mathbf{x}_n, \mathbf{w}) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1} \mathbf{I}). \quad (5.8)$$

Then, the logarithm of the likelihood except the terms independent of \mathbf{w} can be written as

$$-\frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^\top (\beta^{-1} \mathbf{I})^{-1} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})) = -\frac{\beta}{2} \sum_{n=1}^N \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \quad (5.9)$$

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

$$\mathbf{0} = -\beta \sum_{n=1}^N \frac{\partial \mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial \mathbf{w}} (\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n). \quad (5.10)$$

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \quad (5.11)$$

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

$$\mathbf{0} = \sum_{n=1}^N \frac{\partial \mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial \mathbf{w}} (\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n). \quad (5.12)$$

Therefore, maximising the likelihood is equivalent to minimising $E(\mathbf{w})$.

5.3

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

$$p(\mathbf{t}_n | \mathbf{x}_n, \mathbf{w}) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \Sigma). \quad (5.13)$$

Then, the logarithm of the likelihood except the terms independent of \mathbf{w} and Σ can be written as

$$-\frac{1}{2} \ln(\det \Sigma) - \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^\top \Sigma^{-1} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})). \quad (5.14)$$

Setting the derivatives with respect to \mathbf{w} and Σ to zero gives

$$\begin{aligned} \mathbf{0} &= - \sum_{n=1}^N \frac{\partial \mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial \mathbf{w}} \Sigma^{-1} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})), \\ \mathbf{O} &= -\frac{1}{2} \Sigma^{-1} + \frac{1}{2} (\Sigma^{-1})^2 \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})) (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^\top. \end{aligned} \quad (5.15)$$

Therefore, the maximum likelihood solution for Σ is given by

$$\Sigma = \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})) (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^\top. \quad (5.16)$$

On the other hand, if Σ is fixed and known, then the maximum likelihood solution for \mathbf{w} is given by minimising

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^\top \Sigma^{-1} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})). \quad (5.17)$$

5.4

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1 | \mathbf{x}_n) &= (1 - \epsilon)y_n + \epsilon(1 - y_n), \end{aligned} \quad (5.18)$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \quad (5.19)$$

Then,

$$p(t_n | \mathbf{x}_n) = ((1 - \epsilon)y_n + \epsilon(1 - y_n))^{t_n} (\epsilon y_n + (1 - \epsilon)(1 - y_n))^{1-t_n}. \quad (5.20)$$

Therefore,

$$\begin{aligned} & - \ln \left(\prod_{n=1}^N p(t_n | \mathbf{x}_n) \right) \\ &= - \sum_{n=1}^N (t_n \ln((1 - \epsilon)y_n + \epsilon(1 - y_n)) + (1 - t_n) \ln(\epsilon y_n + (1 - \epsilon)(1 - y_n))). \end{aligned} \quad (5.21)$$

5.5

Let $\mathbf{t}_1, \dots, \mathbf{t}_N$ be variables in K dimensions such that

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1 | \mathbf{x}_n) &= y_{nk}, \end{aligned} \tag{5.22}$$

where

$$y_{nk} = y_k(\mathbf{x}_n, \mathbf{w}). \tag{5.23}$$

Then,

$$p(t_{nk} | \mathbf{x}_n) = y_{nk}^{t_{nk}}, \tag{5.24}$$

so that

$$p(\mathbf{t}_n | \mathbf{x}_n) = \prod_{k=1}^K y_{nk}^{t_{nk}}. \tag{5.25}$$

Therefore,

$$\ln \left(\prod_{n=1}^N p(\mathbf{t}_n | \mathbf{x}_n) \right) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \tag{5.26}$$

5.6

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1 | \mathbf{x}_n) &= y_n, \end{aligned} \tag{5.27}$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \tag{5.28}$$

Then,

$$p(t_n | \mathbf{x}_n) = y_n^{t_n} (1 - y_n)^{1-t_n}. \tag{5.29}$$

Let

$$E(\mathbf{w}) = - \ln \left(\prod_{n=1}^N p(t_n | \mathbf{x}_n) \right). \tag{5.30}$$

Then,

$$E(\mathbf{w}) = - \sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)). \tag{5.31}$$

If

$$y_n = \sigma(a_n), \quad (5.32)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \quad (5.33)$$

then, by 4.12,

$$\frac{\partial E(\mathbf{w})}{\partial a_n} = -y_n(1 - y_n) \left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right). \quad (5.34)$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_n} = y_n - t_n. \quad (5.35)$$

5.7

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

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1 | \mathbf{x}_n) &= y_{nk}, \end{aligned} \quad (5.36)$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^K y_{nk} &= 1. \end{aligned} \quad (5.37)$$

Then,

$$p(\mathbf{t}_n | \mathbf{x}_n) = \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (5.38)$$

Let

$$E(\mathbf{w}) = -\ln \left(\prod_{n=1}^N p(\mathbf{t}_n | \mathbf{x}_n) \right). \quad (5.39)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (5.40)$$

If

$$y_{nk} = \frac{\exp(a_k(\mathbf{x}_n, \mathbf{w}))}{\sum_{k=1}^K \exp(a_k(\mathbf{x}_n, \mathbf{w}))}, \quad (5.41)$$

then, by 4.17,

$$\frac{\partial E(\mathbf{w})}{\partial a_{k'}} = - \sum_{n=1}^N \sum_{k=1}^K t_{nk} y_{nk} (I_{kk'} - y_{nk}) \frac{1}{y_{nk}}. \quad (5.42)$$

The right hand side can be written as

$$- \sum_{n=1}^N \sum_{k=1}^K t_{nk} (I_{kk'} - y_{nk}) = - \sum_{n=1}^N \left(\sum_{k=1}^K t_{nk} y_{nk} - t_{nk'} \right). \quad (5.43)$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_k} = \sum_{n=1}^N (y_{nk} - t_{nk}). \quad (5.44)$$

5.8

Setting the derivative of

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \quad (5.45)$$

gives

$$\frac{d}{da} \tanh a = 1 - \left(\frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \right)^2. \quad (5.46)$$

Therefore,

$$\frac{d}{da} \tanh a = 1 - (\tanh a)^2. \quad (5.47)$$

5.9

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{-1, 1\}, \\ p(t_n = 1 | \mathbf{x}_n) &= \frac{1 + y_n}{2}, \end{aligned} \quad (5.48)$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \quad (5.49)$$

Then,

$$p(t_n|\mathbf{x}_n) = \left(\frac{1+y_n}{2}\right)^{\frac{1+t_n}{2}} \left(\frac{1-y_n}{2}\right)^{\frac{1-t_n}{2}}. \quad (5.50)$$

Let

$$E(\mathbf{w}) = -\ln \left(\prod_{n=1}^N p(t_n|\mathbf{x}_n) \right). \quad (5.51)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N \left(\frac{1+t_n}{2} \ln \frac{1+y_n}{2} + \frac{1-t_n}{2} \ln \frac{1-y_n}{2} \right). \quad (5.52)$$

The appropriate choice of y is \tanh .

5.10

Let

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}), \quad (5.53)$$

where E is a real function of real vectors. Then, \mathbf{H} is a real symmetric matrix. Therefore, by 2.20, \mathbf{H} is positive if and only if its eigenvalues are positive.

5.11

Let \mathbf{w} be a real vector in M dimensions. Let E be a real function of \mathbf{w} . Let \mathbf{w}^* be a vector such that

$$\nabla E(\mathbf{w}^*) = \mathbf{0}. \quad (5.54)$$

Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*), \quad (5.55)$$

where

$$\mathbf{H} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*}. \quad (5.56)$$

Let $\mathbf{u}_1, \dots, \mathbf{u}_M$ be eigenvectors such that

$$\mathbf{H}\mathbf{u}_m = \lambda_m \mathbf{u}_m. \quad (5.57)$$

Note that \mathbf{H} is a real symmetric matrix. Then, by ??, we have

$$\mathbf{u}_m^\top \mathbf{u}_{m'} = I_{mm'}. \quad (5.58)$$

Therefore, there exists $\alpha_1, \dots, \alpha_M$ such that

$$\mathbf{w} - \mathbf{w}^* = \sum_{m=1}^M \alpha_m \mathbf{u}_m. \quad (5.59)$$

Then, the approximation can be written as

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} \sum_{m=1}^M \lambda_m \alpha_m^2. \quad (5.60)$$

Therefore, the contours of constant E are ellipses whose axes are aligned with $\mathbf{u}_1, \dots, \mathbf{u}_M$ with lengths which are proportional to $\lambda_1^{-\frac{1}{2}}, \dots, \lambda_M^{-\frac{1}{2}}$.

5.12

Let \mathbf{w} be a real vector. Let E be a real function of \mathbf{w} . Let \mathbf{w}^* be a vector such that

$$\nabla E(\mathbf{w}^*) = \mathbf{0}. \quad (5.61)$$

Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*), \quad (5.62)$$

where

$$\mathbf{H} = \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*}. \quad (5.63)$$

If \mathbf{H} is positive definite, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \quad (5.64)$$

Therefore, \mathbf{w}^* is a local minimum of the right hand side. On the other hand, if \mathbf{w}^* is a local minimum of the right hand side, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \quad (5.65)$$

Therefore, \mathbf{H} is positive definite. Thus, the necessary and sufficient condition for \mathbf{w}^* to be a local minimum is that \mathbf{H} be positive definite.

5.13

Let \mathbf{w} be a vector in M dimensions. Let E be a function of \mathbf{w} . Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{b} + \frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^\top \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}}), \quad (5.66)$$

where

$$\begin{aligned} \mathbf{b} &= \nabla E(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}} \\ \mathbf{H} &= \nabla \nabla E(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}} \end{aligned} \quad (5.67)$$

Since \mathbf{b} is a vector in M dimensions and \mathbf{H} is a $M \times M$ symmetric matrix, the number of independent elements of the right hand side is

$$M + \frac{M(M+1)}{2} = \frac{M(M+3)}{2}. \quad (5.68)$$

5.14

Let w be a variable. Let E_n be a function of w . Then, by a Taylor expansion,

$$\begin{aligned} E_n(w_{mm'} + \epsilon) &= E_n(w_{mm'}) + \left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} \epsilon + O(\epsilon^2), \\ E_n(w_{mm'} - \epsilon) &= E_n(w_{mm'}) - \left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} \epsilon + O(\epsilon^2). \end{aligned} \quad (5.69)$$

Therefore,

$$\left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} = \frac{E_n(w_{mm'} + \epsilon) - E_n(w_{mm'} - \epsilon)}{2\epsilon} + O(\epsilon^2). \quad (5.70)$$

5.15 (Incomplete)

5.16

Let $\mathbf{t}_1, \dots, \mathbf{t}_N$ be vectors. Let $\mathbf{y}_1, \dots, \mathbf{y}_N$ be vectors which are dependent on a vector \mathbf{w} . Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \|\mathbf{y}_n - \mathbf{t}_n\|^2. \quad (5.71)$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (\nabla \mathbf{y}_n)^\top (\mathbf{y}_n - \mathbf{t}_n), \quad (5.72)$$

so that

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N (\nabla \text{vec} (\nabla \mathbf{y}_n)^\top)^\top (\mathbf{y}_n - \mathbf{t}_n) + \sum_{n=1}^N (\nabla \mathbf{y}_n)^\top (\nabla \mathbf{y}_n). \quad (5.73)$$

5.17

Let t be a variable. Let y be a function of a vector \mathbf{x} and a vector \mathbf{w} . Let

$$E(\mathbf{w}) = \frac{1}{2} \int \int (y - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt. \quad (5.74)$$

Then,

$$\nabla E(\mathbf{w}) = \int \int (y - t) p(\mathbf{x}, t) \nabla y d\mathbf{x} dt. \quad (5.75)$$

The right hand side can be written as

$$\begin{aligned} & \int y \nabla y \left(\int p(\mathbf{x}, t) dt \right) d\mathbf{x} - \int \nabla y \left(\int t p(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x} \\ &= \int y \nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla y E(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \end{aligned} \quad (5.76)$$

Then,

$$\nabla \nabla E(\mathbf{w}) = \int \nabla y (\nabla y)^\top p(\mathbf{x}) d\mathbf{x} + \int y \nabla \nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla \nabla y E(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (5.77)$$

The second and the third terms of the right hand side can be written as

$$E(y \nabla \nabla y) - E(\nabla \nabla y E(t|\mathbf{x})) = E((y - E(t|\mathbf{x})) \nabla \nabla y). \quad (5.78)$$

Therefore, if

$$y = E(t|\mathbf{x}), \quad (5.79)$$

then

$$\nabla \nabla E(\mathbf{w}) = \int \nabla y (\nabla y)^\top p(\mathbf{x}) d\mathbf{x}. \quad (5.80)$$

5.18

Let t_1, \dots, t_N be variables. Let y_1, \dots, y_N be variables such that

$$\begin{aligned} y_n &= \mathbf{w}_n^{(2)\top} \mathbf{z} + \mathbf{w}_n^{(0)\top} \mathbf{x}, \\ z_m &= \tanh \left(\mathbf{w}_m^{(1)\top} \mathbf{x} \right). \end{aligned} \quad (5.81)$$

Let

$$E = \frac{1}{2} \sum_{n=1}^N (y_n - t_n)^2. \quad (5.82)$$

Then,

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}}, \\ \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{w}_m^{(1)}}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}. \end{aligned} \quad (5.83)$$

Therefore,

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= (y_n - t_n) \mathbf{x}, \\ \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= (y_n - t_n) \mathbf{A} \mathbf{w}_n^{(2)}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= (y_n - t_n) \mathbf{z}, \end{aligned} \quad (5.84)$$

where

$$A_{mm'} = (1 - z_m^2) x_{m'}. \quad (5.85)$$

5.19

Let t_1, \dots, t_N be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n | \mathbf{w}) &= y_n^{t_n} (1 - y_n)^{1-t_n}, \end{aligned} \quad (5.86)$$

where

$$\begin{aligned} y_n &= \sigma(a_n(\mathbf{w})), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}, \end{aligned} \quad (5.87)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \quad (5.88)$$

The right hand side can be written as

$$-\ln \left(\prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n} \right) = -\sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)). \quad (5.89)$$

Then, by 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^N \left(\frac{t_n}{y_n} y_n (1 - y_n) \nabla a_n - \frac{1 - t_n}{1 - y_n} y_n (1 - y_n) \nabla a_n \right). \quad (5.90)$$

The right hand side can be written as

$$-\sum_{n=1}^N (t_n (1 - y_n) \nabla a_n - (1 - t_n) y_n \nabla a_n) = \sum_{n=1}^N (y_n - t_n) \nabla a_n. \quad (5.91)$$

Then, by 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N (y_n - t_n) \nabla a_n. \quad (5.92)$$

Therefore, by 4.12,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N y_n (1 - y_n) \nabla a_n (\nabla a_n)^\top + \sum_{n=1}^N (y_n - t_n) \nabla \nabla a_n. \quad (5.93)$$

5.20

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

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1 | \mathbf{x}_n) &= y_{nk}, \end{aligned} \quad (5.94)$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^K y_{nk} &= 1. \end{aligned} \quad (5.95)$$

Then,

$$p(\mathbf{t}_n|\mathbf{x}_n) = \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (5.96)$$

Let

$$E(\mathbf{w}) = -\ln \left(\prod_{n=1}^N p(\mathbf{t}_n|\mathbf{x}_n) \right). \quad (5.97)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (5.98)$$

If

$$y_{nk} = \frac{\exp(a_k(\mathbf{x}_n, \mathbf{w}))}{\sum_{k=1}^K \exp(a_k(\mathbf{x}_n, \mathbf{w}))}, \quad (5.99)$$

then, by 5.7,

$$\frac{\partial E(\mathbf{w})}{\partial a_k} = \sum_{n=1}^N (y_{nk} - t_{nk}). \quad (5.100)$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^N \sum_{k=1}^K (y_{nk} - t_{nk}) \nabla a_k. \quad (5.101)$$

Then,

$$\nabla \nabla E(\mathbf{w}) = \sum_{n=1}^N \sum_{k=1}^K y_{nk}(1 - y_{nk}) \nabla a_k (\nabla a_k)^\top + \sum_{n=1}^N \sum_{k=1}^K (y_{nk} - t_{nk}) \nabla \nabla a_k. \quad (5.102)$$

5.21 (Incomplete)

5.22

Let y_1, \dots, y_N be variables such that

$$\begin{aligned} y_n &= \mathbf{w}_n^{(2)\top} \mathbf{z}, \\ z_m &= h(a_m), \\ a_m &= \mathbf{w}_m^{(1)\top} \mathbf{x}. \end{aligned} \quad (5.103)$$

Let E be a function of y_1, \dots, y_N . Then,

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial z_m} \frac{\partial z_m}{\partial a_m} \frac{\partial a_m}{\partial \mathbf{w}_m^{(1)}}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.\end{aligned}\tag{5.104}$$

Therefore,

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} w_{nm}^{(2)} h'(a_m) \mathbf{x}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \mathbf{z}.\end{aligned}\tag{5.105}$$

Thus,

$$\begin{aligned}\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= w_{nm}^{(2)} \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm'}^{(2)} h'(a_{m'}) h'(a_m) \mathbf{x} + \frac{\partial E}{\partial y_n} h''(a_m) I_{mm'} \mathbf{x} \right)^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= h'(a_m) \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} \mathbf{z} + \frac{\partial E}{\partial y_n} \mathbf{v} \right)^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{z} \mathbf{z}^\top,\end{aligned}\tag{5.106}$$

where

$$v_m = \begin{cases} 1, & m = n, \\ 0 & \text{otherwise.} \end{cases}\tag{5.107}$$

5.23

Let y_1, \dots, y_N be variables such that

$$\begin{aligned}y_n &= \mathbf{w}_n^{(2)\top} \mathbf{z} + \mathbf{w}_n^{(0)\top} \mathbf{x}, \\ z_m &= h(a_m), \\ a_m &= \mathbf{w}_m^{(1)\top} \mathbf{x}.\end{aligned}\tag{5.108}$$

Let E be a function of y_1, \dots, y_N . Then,

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}}, \\ \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial z_m} \frac{\partial z_m}{\partial a_m} \frac{\partial a_m}{\partial \mathbf{w}_m^{(1)}}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.\end{aligned}\tag{5.109}$$

Therefore,

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \mathbf{x}, \\ \frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} w_{nm}^{(2)} h'(a_m) \mathbf{x}, \\ \frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \mathbf{z}.\end{aligned}\tag{5.110}$$

Thus,

$$\begin{aligned}\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(0)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{x} \mathbf{x}^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_m^{(1)}} &= \frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} h'(a_m) \mathbf{x} \mathbf{x}^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{x} \mathbf{z}^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= w_{nm}^{(2)} \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm'}^{(2)} h'(a_{m'}) h'(a_m) \mathbf{x} + \frac{\partial E}{\partial y_n} h''(a_m) I_{mm'} \mathbf{x} \right)^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= h'(a_m) \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} \mathbf{z} + \frac{\partial E}{\partial y_n} \mathbf{v} \right)^\top, \\ \frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{z} \mathbf{z}^\top,\end{aligned}\tag{5.111}$$

where

$$v_m = \begin{cases} 1, & m = n, \\ 0 & \text{otherwise.} \end{cases}\tag{5.112}$$

5.24

Let y_1, \dots, y_n be variables such that

$$\begin{aligned} y_n &= \mathbf{w}_n^\top \mathbf{z} + w_{n0}, \\ z_m &= h(\mathbf{w}_m^\top \mathbf{x} + w_{m0}). \end{aligned} \quad (5.113)$$

(a)

Let

$$\tilde{\mathbf{x}} = a\mathbf{x} + b\mathbf{v}, \quad (5.114)$$

where

$$\mathbf{v} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Then,

$$z_m = h\left(\frac{1}{a}\mathbf{w}_m^\top (\tilde{\mathbf{x}} - b\mathbf{v}) + w_{m0}\right). \quad (5.115)$$

Therefore,

$$\tilde{z}_m = h(\tilde{\mathbf{w}}_m^\top \mathbf{x} + \tilde{w}_{m0}). \quad (5.116)$$

where

$$\begin{aligned} \tilde{\mathbf{w}}_m &= \frac{1}{a}\mathbf{w}_m, \\ \tilde{w}_{m0} &= w_{m0} - \frac{b}{a}\mathbf{w}_m^\top \mathbf{v}. \end{aligned} \quad (5.117)$$

(b)

Let

$$\tilde{y}_n = cy_n + d. \quad (5.118)$$

Then,

$$\frac{\tilde{y}_n - d}{c} = \mathbf{w}_n^\top \mathbf{z} + w_{n0}. \quad (5.119)$$

Therefore,

$$\tilde{y}_n = \tilde{\mathbf{w}}_n^\top \mathbf{z} + \tilde{w}_{n0}, \quad (5.120)$$

where

$$\begin{aligned} \tilde{\mathbf{w}}_n &= c\mathbf{w}_n, \\ \tilde{w}_{n0} &= cw_{n0} + d. \end{aligned} \quad (5.121)$$