Solutions Manual to
Pattern Recognition and Machine Learning

Hiromichi Inawashiro

June 13, 2025

Contents

1   Introduction

1.1

Let

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

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

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

If

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

then

$$\mathbf{0} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \left(\mathbf{w}^\intercal \boldsymbol{\phi}(x_n) - t_n \right). \tag{1.4}$$

Then,

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

Therefore,

$$\operatorname*{argmin}_{\mathbf{w}} E(\mathbf{w}) = \mathbf{A}^{-1} \mathbf{v}, \tag{1.6}$$

where

$$\mathbf{A} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\intercal,$$

$$\mathbf{v} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n). \tag{1.7}$$

If

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

then

$$A_{mm'} = \sum_{n=1}^{N} x_n^{m+m'},$$

$$v_m = \sum_{n=1}^{N} t_n x_n^m. \tag{1.8}$$

1.2

Let

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

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

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

If

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

then

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

Then,

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

Therefore,

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

where

$$\mathbf{A} = \sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^\mathsf{T} + \lambda \mathbf{I},$$

$$\mathbf{v} = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n). \tag{1.15}$$

If

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

then

$$A_{mm'} = \sum_{n=1}^{N} x_n^{m+m'} + \lambda I_{mm'},$$

$$v_m = \sum_{n=1}^{N} t_n x_n^m. \tag{1.16}$$

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,

$$p(g, o) = p(o|g)p(g),$$

$$p(o) = p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g). \tag{1.20}$$

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

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

Therefore,

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

1.4

Let x and y be variables such that

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

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

$$|y - \hat{y}| < \epsilon, \tag{1.24}$$

then

$$g'(y) \neq 0. \tag{1.25}$$

(a)

Since

$$\int p_x(x)dx = \int p_x\left(g(y)\right)|g'(y)|dy, \tag{1.26}$$

we have

$$p_y(y) = p_x\left(g(y)\right)|g'(y)|. \tag{1.27}$$

Taking the derivative and substituting

$$y = \hat{y} \tag{1.28}$$

gives

$$0 = g'(\hat{y})p'_x\left(g(\hat{y})\right) + p_x\left(g(\hat{y})\right)g''(\hat{y}). \tag{1.29}$$

Therefore, in general,

$$\hat{x} \neq g(\hat{y}). \tag{1.30}$$

(b)

By (a), if
$$g(y) = ay + b, \tag{1.31}$$

then
$$0 = ap'_x \left(g \left(\hat{y} \right) \right). \tag{1.32}$$

Therefore,
$$\hat{x} = g \left(\hat{y} \right). \tag{1.33}$$

1.5

We have
$$\operatorname{var} f(x) = \mathrm{E} \left(f(x) - \mathrm{E} f(x) \right)^2. \tag{1.34}$$

The right hand side can be written as
$$\mathrm{E} \left((f(x))^2 - 2f(x) \mathrm{E} f(x) + (\mathrm{E} f(x))^2 \right) = \mathrm{E} \left(f(x) \right)^2 - (\mathrm{E} f(x))^2. \tag{1.35}$$

Therefore,
$$\operatorname{var} f(x) = \mathrm{E} \left(f(x) \right)^2 - (\mathrm{E} f(x))^2. \tag{1.36}$$

1.6

We have
$$\operatorname{cov}(x, y) = \mathrm{E} \left((x - \mathrm{E} x) (y - \mathrm{E} y) \right). \tag{1.37}$$

The right hand side can be written as
$$\mathrm{E} \, xy - \mathrm{E} \left(x \, \mathrm{E} \, y \right) - \mathrm{E} \left(y \, \mathrm{E} \, x \right) + \mathrm{E} \left(\mathrm{E} \, x \, \mathrm{E} \, y \right) = \mathrm{E} \, xy - \mathrm{E} \, x \, \mathrm{E} \, y. \tag{1.38}$$

The right hand side can be written as
$$\int xyp(x, y) dx dy - \int xp(x) dx \int yp(y) dy. \tag{1.39}$$

If x and y are independent, by the definition,
$$f(x, y) = f(x)f(y). \tag{1.40}$$

Then,
$$\int xyp(x, y) dx dy = \int p(x) dx \int p(y) dy. \tag{1.41}$$

Therefore,
$$\operatorname{cov}(x, y) = 0. \tag{1.42}$$

1.7

(a)

Let

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx. \tag{1.43}$$

Then,

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\left(x^2 + y^2\right)\right) dxdy. \tag{1.44}$$

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

$$\int_0^{\infty}\int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} drd\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdr. \tag{1.45}$$

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

$$2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{1}{2}s^2\right) sds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_0^{\infty}. \tag{1.46}$$

Therefore,

$$I = \left(2\pi\sigma^2\right)^{\frac{1}{2}}. \tag{1.47}$$

(b)

By the definition,

$$\mathcal{N}\left(x|\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \tag{1.48}$$

Then,

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

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

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}t^2\right) dt = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} I. \tag{1.50}$$

Therefore,

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) dx = 1. \tag{1.51}$$

1.8

(a)

Let x be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \tag{1.52}$$

Then,

$$\mathrm{E}\,x = \int_{-\infty}^{\infty} x\mathcal{N}\left(x|\mu, \sigma^2\right) dx. \tag{1.53}$$

The right hand side can be written as

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} x\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \tag{1.54}$$

By the transformation

$$y = x - \mu, \tag{1.55}$$

the integral can be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty} (y+\mu)\exp\left(-\frac{1}{2\sigma^2}y^2\right) dy \\
&= \int_{-\infty}^{\infty} y\exp\left(-\frac{1}{2\sigma^2}y^2\right) dy + \mu\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}y^2\right) dy.
\end{aligned} \tag{1.56}$$

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

$$\mu\left(2\pi\sigma^2\right)^{\frac{1}{2}}. \tag{1.57}$$

Therefore,

$$\mathrm{E}\,x = \mu. \tag{1.58}$$

(b)

By 1.7(b),

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) dx = 1, \tag{1.59}$$

so that

$$\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 1. \tag{1.60}$$

Taking the derivative with respect to σ^2 gives

$$(2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) \left(\sigma^2\right)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$
$$+ \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} \left(\sigma^2\right)^{-2} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 0. \tag{1.61}$$

The left hand side can be written as

$$-\frac{1}{2} \left(\sigma^2\right)^{-1} \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) dx + \frac{1}{2} \left(\sigma^2\right)^{-2} \int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}\left(x|\mu,\sigma^2\right) dx$$
$$= -\frac{1}{2} \left(\sigma^2\right)^{-1} + \frac{1}{2} \left(\sigma^2\right)^{-2} \operatorname{var} x. \tag{1.62}$$

Therefore,

$$\operatorname{var} x = \sigma^2. \tag{1.63}$$

1.9

(a)

Let x be a variable such that

$$p(x) = \mathcal{N}\left(x|\mu,\sigma^2\right). \tag{1.64}$$

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

$$0 = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \left(-\frac{1}{\sigma^2}(x-\mu)\right) \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.65}$$

Therefore,

$$\operatorname{mode} x = \mu. \tag{1.66}$$

(b)

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right). \tag{1.67}$$

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

$$\mathbf{0} = -\left(2\pi\right)^{-\frac{D}{2}} \left(\det\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right) (\mathbf{x}-\boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{1.68}$$

Therefore,
$$\text{mode}\,\mathbf{x} = \boldsymbol{\mu}. \tag{1.69}$$

1.10

(a)

We have
$$\mathrm{E}(x + y) = \iint (x + y)p(x, y)dxdy. \tag{1.70}$$

The right hand side can be written as

$$\int x \left(\int p(x, y)dy \right) dx + \int y \left(\int p(x, y)dx \right) dy = \int xp(x)dx + \int yp(y)dy. \tag{1.71}$$

The right hand side can be written as

$$\mathrm{E}\,x + \mathrm{E}\,y. \tag{1.72}$$

Therefore,
$$\mathrm{E}(x + y) = \mathrm{E}\,x + \mathrm{E}\,y. \tag{1.73}$$

(b)

We have
$$\mathrm{var}(x + y) = \mathrm{E}\left(x + y - \mathrm{E}(x + y)\right)^2 \tag{1.74}$$

The right hand side can be written as

$$\begin{aligned}
\mathrm{E}\left(x - \mathrm{E}\,x\right)^2 + 2\,\mathrm{E}\left((x - \mathrm{E}\,x)(y - \mathrm{E}\,y)\right) + \mathrm{E}\left(y - \mathrm{E}\,y\right)^2 \\
= \mathrm{var}\,x + 2\,\mathrm{cov}(x, y) + \mathrm{var}\,y.
\end{aligned} \tag{1.75}$$

By 1.6, if x and y are independent, then

$$\mathrm{cov}(x, y) = 0. \tag{1.76}$$

Therefore,
$$\mathrm{var}(x + y) = \mathrm{var}\,x + \mathrm{var}\,y. \tag{1.77}$$

9

1.11

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

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right). \tag{1.78}$$

Then,

$$\ln\left(\prod_{n=1}^{N} p(x_n)\right) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{1.79}$$

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

$$0 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu),$$

$$0 = -\frac{N}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{1.80}$$

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

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

$$\sigma_{\text{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2. \tag{1.81}$$

1.12

(a)

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

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right),$$
$$p(x_{n'}) = \mathcal{N}\left(x_{n'} | \mu, \sigma^2\right). \tag{1.82}$$

Then,

$$\text{E}\, x_n x_{n'} = \mu^2. \tag{1.83}$$

By the property

$$\text{E}\, x_n^2 = \text{var}\, x_n + \left(\text{E}\, x_n\right)^2, \tag{1.84}$$

we have

$$\mathrm{E}\,x_n^2 = \sigma^2 + \mu^2. \tag{1.85}$$

Therefore,

$$\mathrm{E}\,x_n x_{n'} = \mu^2 + I_{nn'}\sigma^2. \tag{1.86}$$

(b)

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

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right). \tag{1.87}$$

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

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{1.88}$$

Then,

$$\mathrm{E}\,\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\,x_n. \tag{1.89}$$

Therefore,

$$\mathrm{E}\,\mu_{\mathrm{ML}} = \mu. \tag{1.90}$$

(c)

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

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right). \tag{1.91}$$

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

$$\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \mu_{\mathrm{ML}}\right)^2. \tag{1.92}$$

Then,

$$\mathrm{E}\,\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left(x_n - \mu_{\mathrm{ML}}\right)^2. \tag{1.93}$$

The right hand side can be writen as

$$\frac{1}{N} \sum_{n=1}^{N} \mathrm{E} \left(x_n^2 - 2\mu_{\mathrm{ML}} x_n + \mu_{\mathrm{ML}}^2 \right)$$
$$= \frac{1}{N} \sum_{n=1}^{N} \mathrm{E}\, x_n^2 - \frac{2}{N} \mathrm{E} \left(\mu_{\mathrm{ML}} \left(\sum_{n=1}^{N} x_n \right) \right) + \mathrm{E}\, \mu_{\mathrm{ML}}^2. \tag{1.94}$$

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

$$\frac{1}{N} \sum_{n=1}^{N} \left(\mu^2 + \sigma^2 \right) = \mu^2 + \sigma^2, \tag{1.95}$$

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

$$-\frac{2}{N} \mathrm{E} \left(N \left(\frac{1}{N} \sum_{n=1}^{N} x_n \right)^2 \right) + \mathrm{E} \left(\frac{1}{N} \sum_{n=1}^{N} x_n \right)^2 = -\mathrm{E} \left(\frac{1}{N} \sum_{n=1}^{N} x_n \right)^2. \tag{1.96}$$

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

$$-\frac{1}{N^2} \sum_{n=1}^{N} \mathrm{E}\, x_n^2 - \frac{2}{N^2} \sum_{1 \le n < n' \le N} \mathrm{E}\, x_n x_{n'}$$
$$= -\frac{1}{N^2} N \left(\mu^2 + \sigma^2 \right) - \frac{2}{N^2} \frac{N(N-1)}{2} \mu^2. \tag{1.97}$$

The right hand side can be written as

$$-\frac{1}{N} \left(\mu^2 + \sigma^2 \right) - \frac{N-1}{N} \mu^2 = -\mu^2 - \frac{1}{N} \sigma^2. \tag{1.98}$$

Then,

$$\mathrm{E}\, \sigma_{\mathrm{ML}}^2 = \mu^2 + \sigma^2 - \mu^2 - \frac{1}{N} \sigma^2. \tag{1.99}$$

Therefore,

$$\mathrm{E}\, \sigma_{\mathrm{ML}}^2 = \frac{N-1}{N} \sigma^2. \tag{1.100}$$

1.13

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

$$
\begin{aligned}
\mathrm{E}\, x_n &= \mu, \\
\mathrm{var}\, x_n &= \sigma^2.
\end{aligned} \tag{1.101}
$$

We have

$$
\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right) = \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\left(x_n - \mu\right)^2. \tag{1.102}
$$

The right hand side can be writen as

$$
\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{var}\, x_n = \frac{\sigma^2}{N}. \tag{1.103}
$$

Therefore,

$$
\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right) = \frac{\sigma^2}{N}. \tag{1.104}
$$

1.14

Let

$$
\begin{aligned}
w_{dd'}^{\mathrm{S}} &= \frac{1}{2}(w_{dd'} + w_{d'd}), \\
w_{dd'}^{\mathrm{A}} &= \frac{1}{2}(w_{dd'} - w_{d'd}).
\end{aligned} \tag{1.105}
$$

(a)

We have

$$
\begin{aligned}
w_{dd'} &= w_{dd'}^{\mathrm{S}} + w_{dd'}^{\mathrm{A}}, \\
w_{dd'}^{\mathrm{S}} &= w_{d'd}^{\mathrm{S}}, \\
w_{dd'}^{\mathrm{A}} &= -w_{d'd}^{\mathrm{A}}.
\end{aligned} \tag{1.106}
$$

(b)

We have

$$
\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{A}} x_d x_{d'} = \frac{1}{2}\sum_{d=1}^{D}\sum_{d'=1}^{D}(w_{dd'} - w_{d'd}) x_d x_{d'}. \tag{1.107}
$$

The right hand side can be written as

$$\frac{1}{2}\left(\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'} x_d x_{d'} - \sum_{d=1}^{D}\sum_{d'=1}^{D} w_{d'd} x_d x_{d'}\right) = 0. \tag{1.108}$$

Therefore,

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{A}} x_d x_{d'} = 0. \tag{1.109}$$

(c)

We have

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'} x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} \left(w_{dd'}^{\mathrm{S}} + w_{dd'}^{\mathrm{A}}\right) x_d x_{d'}. \tag{1.110}$$

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

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{S}} x_d x_{d'} + \sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{A}} x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{S}} x_d x_{d'}, \tag{1.111}$$

Therefore,

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'} x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{S}} x_d x_{d'}. \tag{1.112}$$

(d)

Since $\mathbf{W}^{\mathrm{S}}$ 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

$$n(1, M) = n(1, M - 1) = 1. \tag{1.113}$$

Let us assume that

$$n(D, M) = \sum_{d=1}^{D} n(d, M - 1). \tag{1.114}$$

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

$$n(D + 1, M) = n(D, M) + n(D + 1, M - 1), \tag{1.115}$$

so that

$$n(D + 1, M) = \sum_{d=1}^{D+1} n(d, M - 1). \tag{1.116}$$

Therefore, the assumption is proved by induction on D.

(b)

We have

$$\sum_{d=1}^{1} \frac{(d + M - 2)!}{(d - 1)!(M - 1)!} = 1. \tag{1.117}$$

Let us assume that

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

Then,

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

The right hand side can be written as

$$\frac{D(D + M - 1)! + M(D + M - 1)!}{D!M!} = \frac{(D + M)!}{D!M!}. \tag{1.120}$$

Therefore, the assumption is proved by induction on D.

(c)

By 1.14(d),

$$n(D, 2) = \frac{D(D + 1)}{2}. \tag{1.121}$$

Let us assume that

$$n(D, M) = \frac{(D + M - 1)!}{(D - 1)! M!}. \tag{1.122}$$

By (a),

$$n(D, M + 1) = \sum_{d=1}^{D} n(d, M). \tag{1.123}$$

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

$$\sum_{d=1}^{D} \frac{(d + M - 1)!}{(d - 1)! M!} = \frac{(D + M)!}{(D - 1)!(M + 1)!}. \tag{1.124}$$

Therefore, the assumption is proved by induction on M.

1.16

(a)

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

$$N(D, M) = \sum_{m=0}^{M} n(D, m), \tag{1.125}$$

where

$$n(D, m) = \frac{(D + m - 1)!}{(D - 1)! m!}. \tag{1.126}$$

(b)

By (a),

$$N(D, 0) = 1. \tag{1.127}$$

Let us assume that

$$\sum_{m=0}^{M} n(D, m) = \frac{(D + M)!}{D! M!}. \tag{1.128}$$

16

Then,
$$\sum_{m=0}^{M+1} n(D,m) = \frac{(D+M)!}{D!M!} + \frac{(D+M)!}{(D-1)!(M+1)!}. \tag{1.129}$$
The right hand side can be written as
$$\frac{(M+1)(D+M)! + D(D+M)!}{D!(M+1)!} = \frac{(D+M+1)!}{D!(M+1)!}. \tag{1.130}$$
Then, the assumption is proved by induction on M. Therefore,
$$N(D,M) = \frac{(D+M)!}{D!M!}. \tag{1.131}$$

(c)

By the approximation
$$n! \simeq n^n \exp(-n), \tag{1.132}$$
we have
$$\frac{(D+M)!}{D!M!} \simeq \frac{(D+M)^{D+M}}{D^D M^M}. \tag{1.133}$$
The right hand side can be written as
$$D^M \left(1 + \frac{M}{D}\right)^D \left(\frac{1}{M} + \frac{1}{D}\right)^M = M^D \left(1 + \frac{D}{M}\right)^M \left(\frac{1}{D} + \frac{1}{M}\right)^D. \tag{1.134}$$
Therefore,
$$N(D,M) \simeq \begin{cases} D^M, & D \gg M, \\ M^D, & M \gg D. \end{cases} \tag{1.135}$$

(d)

By (b),
$$N(10,3) = 286,$$
$$N(100,3) = 176851, \tag{1.136}$$
$$N(1000,3) = 167668501.$$

1.17

Let
$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u)du. \tag{1.137}$$

17

(a)

We have

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

The right hand side can be written as

$$[-u^x \exp(-u)]_{u=0}^{u=\infty} + \int_0^\infty x u^{x-1} \exp(-u) du = x\Gamma(x). \tag{1.139}$$

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.140}$$

(b)

We have

$$\Gamma(1) = \int_0^\infty \exp(-u) du, \tag{1.141}$$

so that

$$\Gamma(1) = 0!. \tag{1.142}$$

For a positive integer x, let us assume that

$$\Gamma(x) = (x-1)!. \tag{1.143}$$

By (a),

$$\Gamma(x+1) = x\Gamma(x), \tag{1.144}$$

so that

$$\Gamma(x+1) = x!. \tag{1.145}$$

Therefore, the assumption is proved by induction on x.

1.18

(a)

Let

$$\prod_{d=1}^D \int_{-\infty}^\infty \exp\left(-x_d^2\right) dx_i = S_D \int_0^\infty \exp\left(-r^2\right) r^{D-1} dr, \tag{1.146}$$

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

$$s = r^2,\tag{1.147}$$

the right hand side can be written as

$$\frac{S_D}{2}\int_0^\infty \exp(-s)s^{\frac{D-1}{2}}s^{-\frac{1}{2}}ds = \frac{S_D}{2}\Gamma\left(\frac{D}{2}\right).\tag{1.148}$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}.\tag{1.149}$$

(b)

The volume of the sphere can can be written as

$$V_D = S_D \int_0^1 r^{D-1}dr.\tag{1.150}$$

Therefore,

$$V_D = \frac{S_D}{D}.\tag{1.151}$$

(c)

By (a) and (b),

$$S_2 = 2\pi,$$
$$V_2 = \pi.\tag{1.152}$$

Similarly,

$$S_3 = 4\pi,$$
$$V_3 = \frac{4}{3}\pi.\tag{1.153}$$

1.19

(a)

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

cube is given by

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma\left(\frac{D}{2}\right)}. \tag{1.154}$$

(b)

By (a) and the Stering's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}}\exp(-x)x^{\frac{x+1}{2}}, \tag{1.155}$$

we have

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D2^{D-1}(2\pi)^{\frac{1}{2}}\exp\left(1-\frac{D}{2}\right)\left(\frac{D}{2}-1\right)^{\frac{D}{4}}}. \tag{1.156}$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}}\frac{1}{D}\left(\frac{e^2\pi^2}{8D-16}\right)^{\frac{D}{4}}. \tag{1.157}$$

Therefore,

$$\lim_{D\to\infty}\frac{V_D}{2^D} = 0. \tag{1.158}$$

(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

$$\frac{\sqrt{\sum_{i=1}^{D}1^2}}{1} = \sqrt{D}. \tag{1.159}$$

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

1.20

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

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}}\exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \tag{1.160}$$

(a)

We have

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x})d\mathbf{x} = \int_r^{r+\epsilon}\int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\boldsymbol{\phi}, \qquad (1.161)$$

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

$$(2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon}\int J dr' d\boldsymbol{\phi}$$

$$=(2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} d\mathbf{x}. \qquad (1.162)$$

Therefore,

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x})d\mathbf{x} \simeq p(r)\epsilon, \qquad (1.163)$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \qquad (1.164)$$

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

(b)

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

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left((D-1)r^{D-2} - \frac{r^D}{\sigma^2}\right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \qquad (1.165)$$

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

$$\hat{r} = \sqrt{D-1}\sigma. \qquad (1.166)$$

(c)

We have

$$\frac{p\left(\hat{r} + \epsilon\right)}{p(\hat{r})} = \left(\frac{\hat{r}+\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \qquad (1.167)$$

21

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

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

By the Taylor series

$$
\ln(1+x) = x - \frac{1}{2}x^2 + o\left(x^3\right), \tag{1.169}
$$

the right hand side can be approximated as

$$
\exp\left(\frac{\hat{r}^2}{\sigma^2}\left(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) = \exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.170}
$$

Therefore,

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

(d)

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

$$
\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)} = \exp\left(\frac{\hat{r}^2}{2\sigma^2}\right). \tag{1.172}
$$

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

$$
\exp\left(\frac{D-1}{2}\right). \tag{1.173}
$$

Therefore,

$$
\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)} = \exp\left(\frac{D-1}{2}\right). \tag{1.174}
$$

1.21

(a)

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

$$
0 \le a(b-a). \tag{1.175}
$$

Therefore,

$$
a \le (ab)^{\frac{1}{2}}. \tag{1.176}
$$

(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,

$$p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2) \Rightarrow \mathbf{x} \in \mathcal{C}_1,$$
$$p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1) \Rightarrow \mathbf{x} \in \mathcal{C}_2. \tag{1.177}$$

By (a),

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} \leq \int_{\mathcal{R}_1} (p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x},$$
$$\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int_{\mathcal{R}_2} (p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \tag{1.178}$$

Therefore,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \leq \int (p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}. \tag{1.179}$$

1.22

Let

$$\mathrm{E}\, L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.180}$$

If

$$L_{kj} = 1 - I_{kj}, \tag{1.181}$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)) \, d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \tag{1.182}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)) \, d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \tag{1.183}$$

Then,

$$\mathrm{E}\, L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.184}$$

23

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

1.23

Let

$$\mathrm{E}\,L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.185}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left(\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.186}$$

Then,

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

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

1.24   (Incomplete)

Let

$$\mathrm{E}\,L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} + \lambda \int_{\forall k p(\mathcal{C}_k|\mathbf{x}) < \theta} p(\mathbf{x}) d\mathbf{x}. \tag{1.188}$$

1.25

Let

$$\mathrm{E}\,L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.189}$$

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

$$\mathbf{0} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) \, p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \tag{1.190}$$

The integral of the right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = \mathbf{y}(\mathbf{x}) p(\mathbf{x}) - p(\mathbf{x}) \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \tag{1.191}$$

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

$$\mathbf{0} = p(\mathbf{x}) \left(\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \right). \tag{1.192}$$

Therefore,

$$\operatorname*{argmin}_{\mathbf{y}(\mathbf{x})} E\, L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.193}$$

For a single target variable t, it reduces to

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

1.26

Let

$$E\, L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.195}$$

The right hand side can be written as

$$
\begin{aligned}
\iint & \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
= & \iint \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
& + 2 \iint (\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^{\mathsf{T}} \left(E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t} \right) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
& + \iint \|E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}.
\end{aligned}
\tag{1.196}
$$

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

$$\int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 \left(\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \tag{1.197}$$

The integral of the second term can be written as

$$\int (\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^{\mathsf{T}} \left(\int \left(E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t} \right) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.198}$$

25

Since

$$\int E_t(t|x)p(t|x)dt = E_t(t|x)\frac{\int p(x,t)dt}{p(x)},$$

$$\int t p(t|x)dt = E_t(t|x),$$

(1.199)

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

$$\int \left(\int \|E_t(t|x) - t\|^2 p(t|x)dt \right) p(x)dx = \int \text{var}_t(t|x)p(x)dx.$$

(1.200)

Then,

$$E\, L(t, y(x)) = \int \|y(x) - E_t(t|x)\|^2 p(x)dx + \int \text{var}_t(t|x)p(x)dx.$$

(1.201)

Therefore,

$$\underset{y(x)}{\text{argmin}}\, E\, L(t, y(x)) = E_t(t|x).$$

(1.202)

1.27 (Incomplete)

(a)

Let

$$E\, L_q = \int \int |y(x) - t|^q p(x, t)dx dt.$$

(1.203)

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

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

(1.204)

Therefore,

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

(1.205)

(b)

We have

$$\mathrm{E}\,L_1 = \int \left(\int \mathrm{sign}(y(\mathbf{x}) - t)p(t|\mathbf{x})dt \right) p(\mathbf{x})d\mathbf{x}. \qquad (1.206)$$

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

$$\int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})dt. \qquad (1.207)$$

Therefore,

$$\underset{y(\mathbf{x})}{\mathrm{argmin}}\,\mathrm{E}\,L_1 = \mathrm{median}(t|\mathbf{x}). \qquad (1.208)$$

(c)

We have

$$\lim_{q \to 0} \left(\underset{y(\mathbf{x})}{\mathrm{argmin}}\,\mathrm{E}\,L_q \right) = \mathrm{mode}(t|\mathbf{x})? \qquad (1.209)$$

1.28

(a)

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \qquad (1.210)$$

Then,

$$h\left(p^2\right) = 2h(p). \qquad (1.211)$$

Let us assume that, for a positive integer n,

$$h\left(p^n\right) = nh(p). \qquad (1.212)$$

Then, by the first assumption,

$$h\left(p^{n+1}\right) = h\left(p^n\right) + h(p), \qquad (1.213)$$

so that

$$h\left(p^{n+1}\right) = (n+1)h(p). \qquad (1.214)$$

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

27

(b)

For positive integers m and n,

$$h\left(p^{n}\right) = h\left(p^{\frac{n}{m}m}\right). \tag{1.215}$$

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 $mh\left(p^{\frac{n}{m}}\right)$. Therefore,

$$h\left(p^{\frac{n}{m}}\right) = \frac{n}{m}h(p). \tag{1.216}$$

(c)

By the continuity, for a positive real number a,

$$h\left(p^{a}\right) = ah(p). \tag{1.217}$$

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

$$(p\ln p)h'(p) = h(p). \tag{1.218}$$

Then,

$$\int \frac{h'(p)}{h(p)}dp = \int \frac{1}{p\ln p}dp + \text{const}. \tag{1.219}$$

Ignorting 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,

$$h(p) \propto \ln p. \tag{1.220}$$

1.29

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

$$\mathrm{H}(x) = -\sum_{m=1}^{M} p(x_m)\ln p(x_m), \tag{1.221}$$

where

$$\sum_{m=1}^{M} p(x_m) = 1. \tag{1.222}$$

28

By the Jensen's inequality,

$$\sum_{m=1}^{M} p(x_i) \ln \frac{1}{p(x_m)} \leq \ln \left(\sum_{m=1}^{M} 1 \right). \tag{1.223}$$

Therefore,

$$\mathrm{H}(x) \leq \ln M. \tag{1.224}$$

1.30

Let

$$\begin{aligned}
p(x) &= \mathcal{N}\left(x|\mu, \sigma^2\right), \\
q(x) &= \mathcal{N}\left(x|m, s^2\right).
\end{aligned} \tag{1.225}$$

Then, the Kullback-Leibler divergence is given by

$$\mathrm{KL}\left(p||q\right) = -\int p(x) \ln \frac{q(x)}{p(x)} dx. \tag{1.226}$$

The right hand side can be written as

$$\begin{aligned}
&- \int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\
&= - \int_{-\infty}^{\infty} p(x) \left(-\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2} \right) dx.
\end{aligned} \tag{1.227}$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x)dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x)dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx. \tag{1.228}$$

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

$$\begin{aligned}
&\frac{1}{2s^2} \int_{-\infty}^{\infty} (x - \mu + \mu - m)^2 p(x)dx \\
&= \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx + \frac{\mu - m}{s^2} \int_{-\infty}^{\infty} (x-\mu)p(x)dx \\
&\quad + \frac{(\mu - m)^2}{2s^2} \int_{-\infty}^{\infty} p(x)dx,
\end{aligned} \tag{1.229}$$

where the right hand side can be written as
$$\frac{\sigma^2 + (\mu - m)^2}{2s^2}.$$ (1.230)

The third term can be written as $-\frac{1}{2}$. Therefore,
$$\mathrm{KL}\,(p||q) = \ln\frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{2s^2} - \frac{1}{2}.$$ (1.231)

1.31

Let $\mathbf{x}$ and $\mathbf{y}$ be two variables. We have
$$\mathrm{H}(\mathbf{x}) = -\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x},$$
$$\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y})\ln p(\mathbf{y})d\mathbf{y},$$ (1.232)
$$\mathrm{H}(\mathbf{x},\mathbf{y}) = -\int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}.$$

Since
$$\mathrm{H}(\mathbf{x}) = -\int\left(\int p(\mathbf{x},\mathbf{y})d\mathbf{y}\right)\ln p(\mathbf{x})d\mathbf{x},$$
$$\mathrm{H}(\mathbf{y}) = -\int\left(\int p(\mathbf{x},\mathbf{y})d\mathbf{x}\right)\ln p(\mathbf{y})d\mathbf{y},$$ (1.233)

we have
$$\mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{x},\mathbf{y}) = -\int\int p(\mathbf{x},\mathbf{y})\ln\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}d\mathbf{x}d\mathbf{y}.$$ (1.234)

Since
$$\int\int p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} = 1,$$ (1.235)

the Jensen's inequality can be used to have
$$-\int\int p(\mathbf{x},\mathbf{y})\ln\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}d\mathbf{x}d\mathbf{y} \geq -\ln\left(\int\int p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y}\right).$$ (1.236)

The right hand side can be written as
$$-\ln\left(\int p(\mathbf{x})d\mathbf{x}\int p(\mathbf{y})d\mathbf{y}\right) = 0.$$ (1.237)

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

30

1.32

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad (1.239)$$

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

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

Then,

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

We have

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

The right hand side can be written as

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

The first term can be written as

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

By the transformation

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

it can be written as

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

Therefore,

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

1.33

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

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

If

$$\mathrm{H}(y|x) = 0, \tag{1.249}$$

then

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

Since

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

the equation reduces to

$$p(y_l|x_k)\ln p(y_l|x_k) = 0. \tag{1.252}$$

Then,

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

Since

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

we have

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

1.34

Let x be a variable. We have

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

32

In order to maximise H(x) with the constratints

$$\int_{-\infty}^{\infty} p(x)dx = 1,$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu, \tag{1.257}$$

$$\int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx = \sigma^2,$$

let

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

Setting the variation with respect to p to zero gives

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

Then,

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

so that

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

where

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

Substituting it to the constraints gives

$$c \int_{-\infty}^{\infty} \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx = 1,$$

$$c \int_{-\infty}^{\infty} x \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx = \mu, \tag{1.263}$$

$$c \int_{-\infty}^{\infty} (x - \mu)^2 \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx = \sigma^2.$$

By the transformation

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

they can be written as

$$c \int_{-\infty}^{\infty} \exp\left(-y^2\right) (-\lambda_3)^{-\frac{1}{2}} dy = 1,$$

$$c \int_{-\infty}^{\infty} \left((-\lambda_3)^{-\frac{1}{2}} y + \mu - \frac{\lambda_2}{2\lambda_3} \right) \exp\left(-y^2\right) (-\lambda_3)^{-\frac{1}{2}} dy = \mu, \tag{1.265}$$

$$c \int_{-\infty}^{\infty} \left((-\lambda_3)^{-\frac{1}{2}} y - \frac{\lambda_2}{2\lambda_3} \right)^2 \exp\left(-y^2\right) (-\lambda_3)^{-\frac{1}{2}} dy = \sigma^2.$$

Since

$$\int_{-\infty}^{\infty} \exp\left(-y^2\right) dy = \Gamma\left(\frac{1}{2}\right),$$

$$\int_{-\infty}^{\infty} y \exp\left(-y^2\right) dy = 0, \tag{1.266}$$

$$\int_{-\infty}^{\infty} y^2 \exp\left(-y^2\right) dy = \Gamma\left(\frac{3}{2}\right),$$

they can be written as

$$c(-\lambda_3)^{-\frac{1}{2}} \Gamma\left(\frac{1}{2}\right) = 1,$$

$$c \left(\mu - \frac{\lambda_2}{2\lambda_3} \right) (-\lambda_3)^{-\frac{1}{2}} \Gamma\left(\frac{1}{2}\right) = \mu, \tag{1.267}$$

$$c \left((-\lambda_3)^{-\frac{3}{2}} \Gamma\left(\frac{3}{2}\right) + (-\lambda_3)^{-\frac{1}{2}} \frac{\lambda_2^2}{4\lambda_3^2} \Gamma\left(\frac{1}{2}\right) \right) = \sigma^2.$$

Then,

$$\lambda_1 = 1 - \frac{1}{2} \ln\left(2\pi\sigma^2\right),$$

$$\lambda_2 = 0, \tag{1.268}$$

$$\lambda_3 = -\frac{1}{2\sigma^2}.$$

Therefore,

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

34

1.35

Let x be a variable such that

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

Then,

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

The right hand side can be written as

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

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

Therefore,

$$\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln\left(2\pi\sigma^2\right)\right). \tag{1.273}$$

1.36   (Incomplete)

Let f be a strictly convex function. Then,

$$f\left(\lambda a + (1-\lambda)b\right) \le \lambda f(a) + (1-\lambda)f(b), \tag{1.274}$$

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

$$x = \lambda a + (1-\lambda)b. \tag{1.275}$$

Then, the inequality can be written as

$$f(x) \le \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b). \tag{1.276}$$

Let

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

Then,

$$g(x) \ge 0. \tag{1.278}$$

35

Additionally, for $x > a$,

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

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

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

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

$$f'(y) \leq f'(c). \tag{1.281}$$

1.37

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

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

The right hand side can be written as

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

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

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

1.38

Let f be a strictly convex function. Then,

$$f \left(\lambda x_1 + (1 - \lambda) x_2 \right) \leq \lambda f(x_1) + (1 - \lambda) f(x_2), \tag{1.285}$$

where
$$0 \le \lambda \le 1. \tag{1.286}$$

Let us assume that
$$f\left(\sum_{m=1}^{M} \lambda_m x_m\right) \le \sum_{m=1}^{M} \lambda_m f(x_m), \tag{1.287}$$

where
$$\sum_{m=1}^{M} \lambda_m = 1, \tag{1.288}$$
$$\lambda_m \ge 0.$$

Since f is strictly convex,
$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \le \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{m=1}^{M} \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right), \tag{1.289}$$

where
$$\sum_{m=1}^{M+1} \lambda_m = 1, \tag{1.290}$$
$$\lambda_m \ge 0.$$

By the assumption,
$$f\left(\sum_{m=1}^{M} \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right) \le \sum_{m=1}^{M} \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m). \tag{1.291}$$

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

so that
$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \le \sum_{m=1}^{M+1} \lambda_m f(x_m). \tag{1.293}$$

Therefore, the assumption is proved by induction on M.

1.39

Let x and y be two binary variables where

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

(a)

We have
$$
H(x) = -\sum p(x) \ln p(x).
\tag{1.295}
$$

By the distribution,
$$
\begin{aligned}
p(x = 0) &= \frac{2}{3}, \\
p(x = 1) &= \frac{1}{3}.
\end{aligned}
\tag{1.296}
$$

Therefore,
$$
H(x) = \ln 3 - \frac{2}{3} \ln 2.
\tag{1.297}
$$

(b)

We have
$$
H(y) = -\sum p(y) \ln p(y).
\tag{1.298}
$$

By the distribution,
$$
\begin{aligned}
p(y = 0) &= \frac{1}{3}, \\
p(y = 1) &= \frac{2}{3}.
\end{aligned}
\tag{1.299}
$$

Therefore,
$$
H(y) = \ln 3 - \frac{2}{3} \ln 2.
\tag{1.300}
$$

(c)

We have

$$H(y|x) = -\sum p(x,y)\ln p(y|x). \tag{1.301}$$

By the Bayes' theorem,

$$
\begin{aligned}
p(y=0|x=0) &= \frac{p(x=0, y=0)}{p(x=0)}, \\
p(y=0|x=1) &= \frac{p(x=1, y=0)}{p(x=1)}, \\
p(y=1|x=0) &= \frac{p(x=0, y=1)}{p(x=0)}, \\
p(y=1|x=1) &= \frac{p(x=1, y=1)}{p(x=1)}.
\end{aligned}
\tag{1.302}
$$

By the distribution,

$$
\begin{aligned}
p(y=0|x=0) &= \frac{1}{2}, \\
p(y=0|x=1) &= 0, \\
p(y=1|x=0) &= \frac{1}{2}, \\
p(y=1|x=1) &= 1.
\end{aligned}
\tag{1.303}
$$

Therefore,

$$H(y|x) = \frac{2}{3}\ln 2. \tag{1.304}$$

(d)

We have

$$H(x|y) = -\sum p(x,y)\ln p(x|y). \tag{1.305}$$

By the Bayes' theorem,

$$p(x = 0|y = 0) = \frac{p(x = 0, y = 0)}{p(y = 0)},$$

$$p(x = 0|y = 1) = \frac{p(x = 0, y = 1)}{p(y = 1)},$$

$$p(x = 1|y = 0) = \frac{p(x = 1, y = 0)}{p(y = 0)},$$

$$p(x = 1|y = 1) = \frac{p(x = 1, y = 1)}{p(y = 1)}.$$

$$(1.306)$$

By the distribution,

$$p(x = 0|y = 0) = 1,$$

$$p(x = 0|y = 1) = \frac{1}{2},$$

$$p(x = 1|y = 0) = 0,$$

$$p(x = 1|y = 1) = \frac{1}{2}.$$

$$(1.307)$$

Therefore,

$$\mathrm{H}(x|y) = \frac{2}{3}\ln 2. \tag{1.308}$$

(e)

We have

$$\mathrm{H}(x, y) = -\sum p(x, y)\ln p(x, y). \tag{1.309}$$

Therefore,

$$\mathrm{H}(x, y) = \ln 3. \tag{1.310}$$

(f)

We have

$$\mathrm{I}(x, y) = -\sum p(x, y)\ln \frac{p(x)p(y)}{p(x, y)}. \tag{1.311}$$

The right hand side can be written as

$$\mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x, y). \tag{1.312}$$

Therefore,

$$\mathrm{I}(x, y) = \ln 3 - \frac{4}{3}\ln 2. \tag{1.313}$$

40

1.40

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

$$\sum_{m=1}^{M} \lambda_m = 1,$$
$$\lambda_m \geq 0,$$
$$x_m > 0. \qquad (1.314)$$

By the Jensen's inequality,

$$\sum_{m=1}^{M} \lambda_m \ln x_m \leq \ln \left(\sum_{m=1}^{M} \lambda_m x_m \right), \qquad (1.315)$$

so that

$$\prod_{m=1}^{M} x_m^{\lambda_m} \leq \sum_{m=1}^{M} \lambda_m x_m. \qquad (1.316)$$

Substituting

$$\lambda_m = \frac{1}{M} \qquad (1.317)$$

to the inequality gives

$$\left(\prod_{m=1}^{M} x_m \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{m=1}^{M} x_m. \qquad (1.318)$$

1.41

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

(a)

We have

$$I(\mathbf{x}, \mathbf{y}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}. \qquad (1.319)$$

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

$$- \iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{x}) + \ln \frac{p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y}$$
$$= - \iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{x}) - \ln p(\mathbf{x}|\mathbf{y}) \right) d\mathbf{x} d\mathbf{y}. \qquad (1.320)$$

41

The right hand side can be written as

$$- \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} + \int\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y}. \qquad (1.321)$$

Therefore,

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{H}(\mathbf{x}) - \mathrm{H}(\mathbf{x}|\mathbf{y}). \qquad (1.322)$$

(b)

We have

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{I}(\mathbf{y}, \mathbf{x}). \qquad (1.323)$$

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

$$\mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}|\mathbf{x}). \qquad (1.324)$$

Therefore,

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}|\mathbf{x}). \qquad (1.325)$$

2 Probability Distributions

2.1

Let x be a variable such that

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

(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,

$$\mathrm{E}\, x = \sum_{x \in \{0,1\}} x p(x),$$
$$\mathrm{E}\, x^2 = \sum_{x \in \{0,1\}} x^2 p(x), \tag{2.4}$$

Therefore,

$$\mathrm{E}\, x = \mu,$$
$$\mathrm{E}\, x^2 = \mu. \tag{2.5}$$

Since

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

we have

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

(c)

By the definition,

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

Therefore,

$$H(x) = -\mu \ln \mu - (1 - \mu) \ln(1 - \mu). \tag{2.9}$$

2.2

Let x be a variable such that

$$x \in \{-1, 1\},$$

$$p(x) = \left(\frac{1 - \mu}{2}\right)^{\frac{1-x}{2}} \left(\frac{1 + \mu}{2}\right)^{\frac{1+x}{2}}. \tag{2.10}$$

(a)

We have

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

Therefore,

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

(b)

By the definition,

$$\mathrm{E}\, x = \sum_{x \in \{-1,1\}} x p(x),$$
$$\mathrm{E}\, x^2 = \sum_{x \in \{-1,1\}} x^2 p(x). \tag{2.13}$$

The right hand sides can be written as

$$-\frac{1 - \mu}{2} + \frac{1 + \mu}{2} = \mu,$$
$$\frac{1 - \mu}{2} + \frac{1 + \mu}{2} = 1. \tag{2.14}$$

44

Therefore,
$$\mathrm{E}\,x = \mu,$$
$$\mathrm{E}\,x^2 = 1. \tag{2.15}$$

Since
$$\operatorname{var} x = \mathrm{E}\,x^2 - (\mathrm{E}\,x)^2, \tag{2.16}$$

we have
$$\operatorname{var} x = 1 - \mu^2. \tag{2.17}$$

(c)

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

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

2.3

(a)

By the definition,
$$\binom{N}{n} = \frac{N!}{n!(N-n)!},$$
$$\binom{N}{n-1} = \frac{N!}{(n-1)!(N-n+1)!} \tag{2.20}$$

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

45

(b)

Note that

$$1 + x = \sum_{n=0}^{1} \binom{1}{n} x^n. \tag{2.24}$$

Let us assume that

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

Then,

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

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

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

Then,

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

Therefore, the assumption is proved by induction on N.

(c)

Let n be a variable such that

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

Then,

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

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

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

Therefore,

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

2.4

Let n be a variable such that

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

(a)

We have

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

By 2.3(c),

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

Taking the derivative with respect to μ gives

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

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

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

Since

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

the second term can be written as

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

Therefore,

$$\mathrm{E}\,n = N\mu. \tag{2.40}$$

(b)

By 2.3(c),

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

Taking the second derivative with respect to μ gives

$$\sum_{n=0}^{N} n(n-1)\binom{N}{n}\mu^{n-2}(1-\mu)^{N-n}$$

$$-2\sum_{n=0}^{N} n(N-n)\binom{N}{n}\mu^{n-1}(1-\mu)^{N-n-1} \tag{2.42}$$

$$+\sum_{n=0}^{N}(N-n)(N-n-1)\binom{N}{n}\mu^{n}(1-\mu)^{N-n-2} = 0.$$

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

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

Since

$$n(N-n)\binom{N}{n} = N(N-1)\binom{N-2}{n-1},$$
$$(N-n)(N-n-1)\binom{N}{n} = N(N-1)\binom{N-2}{n}, \tag{2.44}$$

the second and third terms can be written as

$$-2N(N-1)\sum_{n=1}^{N-1}\binom{N-2}{n-1}\mu^{n-1}(1-\mu)^{N-n-1} = -2N(N-1),$$

$$N(N-1)\sum_{n=0}^{N}\binom{N-2}{n}\mu^{n}(1-\mu)^{N-n-2} = N(N-1). \tag{2.45}$$

Then,

$$\mathrm{E}\,n(n-1) = N(N-1)\mu^2. \tag{2.46}$$

Therefore, since

$$\mathrm{var}\,n = \mathrm{E}\,n(n-1) + \mathrm{E}\,n - (\mathrm{E}\,n)^2, \tag{2.47}$$

we have

$$\mathrm{var}\,n = N\mu(1-\mu). \tag{2.48}$$

2.5

By the definition,

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

By the transformation

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

the right hand side can be written as

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

By the transformation

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

the right hand side can be written as

$$\int_0^\infty \left(\int_0^1 (t\mu)^{a-1}t^{b-1}(1-\mu)^{b-1}td\mu\right)\exp(-t)dt$$
$$= \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu \int_0^\infty t^{a+b-1}\exp(-t)dt. \qquad (2.53)$$

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

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

2.6

Let μ be a variable such that

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

Then, by the definition,

$$
\mathrm{E}\,\mu = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu,
$$
$$
\mathrm{E}\,\mu^2 = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu.
$$

(2.56)

By 2.5,

$$
\int_0^1 \mu^a (1-\mu)^{b-1} d\mu = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)},
$$
$$
\int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu = \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}.
$$

(2.57)

Therefore,

$$
\mathrm{E}\,\mu = \frac{a}{a+b},
$$
$$
\mathrm{E}\,\mu^2 = \frac{a(a+1)}{(a+b)(a+b+1)}.
$$

(2.58)

Since

$$
\mathrm{var}\,\mu = \mathrm{E}\,\mu^2 - (\mathrm{E}\mu)^2,
$$

(2.59)

we have

$$
\mathrm{var}\,\mu = \frac{ab}{(a+b)^2(a+b+1)}.
$$

(2.60)

Since

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

(2.61)

we have

$$
\mathrm{mode}\,\mu = \frac{a-1}{a+b-2}.
$$

(2.62)

2.7

Let m and l be variables such that

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

(2.63)

By 2.6,
$$\mathrm{E}\,\mu = \frac{a}{a+b}. \tag{2.64}$$
Setting the derivative of $p(m, l|\mu)$ with respect to μ to zero gives
$$0 = \binom{m+l}{m}\mu^m(1-\mu)^l\left(\frac{m}{\mu} + \frac{l}{1-\mu}\right). \tag{2.65}$$
Then, the maximum likelihood solution for μ is given by
$$\mu_{\mathrm{ML}} = \frac{m}{m+l}. \tag{2.66}$$
By the Bayes' thereorem,
$$p(\mu|m, l)p(m, l) = p(m, l|\mu)p(\mu). \tag{2.67}$$
Then, by 2.5,
$$p(\mu|m, l) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}. \tag{2.68}$$
The, by 2.6,
$$\mathrm{E}(\mu|m, l) = \frac{m+a}{m+l+a+b}. \tag{2.69}$$
Therefore,
$$\mathrm{E}(\mu|m, l) = \lambda\mu_{\mathrm{ML}} + (1-\lambda)\,\mathrm{E}\,\mu, \tag{2.70}$$
where
$$\lambda = \frac{m+l}{m+l+a+b}. \tag{2.71}$$

2.8

Let x and y be variables.

(a)

By the definition,
$$\mathrm{E}\,x = \int xp(x)dx. \tag{2.72}$$
The right hand side can be written as
$$\int x\left(\int p(x, y)dy\right)dx = \int\left(\int xp(x|y)dx\right)p(y)dy. \tag{2.73}$$
Therefore,
$$\mathrm{E}\,x = \mathrm{E}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.74}$$

(b)

By the definition,
$$\operatorname{var} x = \operatorname{E} \left(x - \operatorname{E} x \right)^2. \tag{2.75}$$

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

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

The right hand side can be written as

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

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

$$2 \operatorname{E}_y \left(\left(\operatorname{E}_x(x|y) - \operatorname{E} x \right) \left(\operatorname{E}_x(x|y) - \operatorname{E}_x(x|y) \right) \right) = 0. \tag{2.78}$$

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

$$\operatorname{E}_y \left(\operatorname{E}_x(x|y) - \operatorname{E}_y \left(\operatorname{E}_x(x|y) \right) \right)^2 = \operatorname{var}_y \left(\operatorname{E}_x(x|y) \right). \tag{2.79}$$

Therefore,
$$\operatorname{var} x = \operatorname{E}_y \left(\operatorname{var}_x(x|y) \right) + \operatorname{var}_y \left(\operatorname{E}_x(x|y) \right). \tag{2.80}$$

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

$$\sum_{m=1}^{M+1} \mu_m = 1, \tag{2.81}$$

we have

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

where

$$c = 1 - \sum_{m=1}^{M-1} \mu_m. \tag{2.83}$$

By the transformation

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

the integral of the right hand side can be written as

$$\int_0^1 (c\mu'_{M+1})^{\alpha_{M+1} - 1} \left(c(1 - \mu'_{M+1}) \right)^{\alpha_M - 1} c d\mu'_{M+1}$$
$$= c^{\alpha_M + \alpha_{M+1} - 1} \int_0^1 \mu'_{M+1}{}^{\alpha_{M+1} - 1} (1 - \mu'_{M+1})^{\alpha_M - 1} d\mu'_{M+1}. \tag{2.85}$$

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

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

Then,

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

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

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

53

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

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

Therefore, the assumption is proved by induction on M.

2.10

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

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

Then, by the definition,

$$\mathrm{E}\,\mu_k=\int\mu_k p(\boldsymbol{\mu})d\boldsymbol{\mu},$$
$$\mathrm{E}\,\mu_k^2=\int\mu_k^2 p(\boldsymbol{\mu})d\boldsymbol{\mu}, \tag{2.90}$$
$$\mathrm{E}\,\mu_k\mu_{k'}=\int\mu_k\mu_{k'} p(\boldsymbol{\mu})d\boldsymbol{\mu}.$$

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

$$\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\frac{\Gamma(\alpha_k+1)}{\Gamma(\alpha_k)}\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k+1)}=\frac{\alpha_k}{\sum_{k=1}^{K}\alpha_k},$$

$$\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\frac{\Gamma(\alpha_k+2)}{\Gamma(\alpha_k)}\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k+2\right)}=\frac{\alpha_k(\alpha_k+1)}{\sum_{k=1}^{K}\alpha_k(\sum_{k=1}^{K}\alpha_k+1)}, \tag{2.91}$$

$$\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\frac{\Gamma(\alpha_k+1)\Gamma(\alpha_{k'}+1)}{\Gamma(\alpha_k)\Gamma(\alpha_{k'})}\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}\alpha_k+2\right)}=\frac{\alpha_k\alpha_{k'}}{\sum_{k=1}^{K}\alpha_k(\sum_{k=1}^{K}\alpha_k+1)}.$$

Then,

$$\mathrm{E}\,\mu_k = \frac{\alpha_k}{\sum_{k=1}^{K}\alpha_k}.$$

$$\mathrm{E}\,\mu_k^2 = \frac{\alpha_k(\alpha_k+1)}{\sum_{k=1}^{K}\alpha_k\left(\sum_{k=1}^{K}\alpha_k+1\right)}, \tag{2.92}$$

$$\mathrm{E}\,\mu_k\mu_{k'} = \frac{\alpha_k\alpha_{k'}}{\sum_{k=1}^{K}\alpha_k\left(\sum_{k=1}^{K}\alpha_k+1\right)}.$$

Since

$$\mathrm{var}\,\mu_k = \mathrm{E}\,\mu_k^2 - (\mathrm{E}\,\mu_k)^2,$$
$$\mathrm{cov}\,(\mu_k,\mu_{k'}) = \mathrm{E}\,\mu_k\mu_{k'} - \mathrm{E}\,\mu_k\,\mathrm{E}\,\mu_{k'}, \tag{2.93}$$

we have

$$\mathrm{var}\,\mu_k = \frac{\alpha_k\left(\left(\sum_{k=1}^{K}\alpha_k\right)-\alpha_k\right)}{(\sum_{k=1}^{K}\alpha_k)^2(\sum_{k=1}^{K}\alpha_k+1)},$$

$$\mathrm{cov}(\mu_k,\mu_{k'}) = -\frac{\alpha_k\alpha_{k'}}{(\sum_{k=1}^{K}\alpha_k)^2(\sum_{k=1}^{K}\alpha_k+1)}. \tag{2.94}$$

2.11

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

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

Then, by the definition,

$$\mathrm{E}\ln\mu_k = \int (\ln\mu_k)\,p(\boldsymbol{\mu})d\boldsymbol{\mu}. \tag{2.96}$$

Since

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

we have

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

55

where

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

Therefore,

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

2.12

Let x be a variable such that

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

where $a < b$. Then

$$\int_a^b p(x)dx = 1. \tag{2.102}$$

Then, by the definition,

$$\mathrm{E}\,x = \frac{1}{b-a} \int_a^b x dx,$$
$$\mathrm{E}\,x^2 = \frac{1}{b-a} \int_a^b x^2 dx. \tag{2.103}$$

Then,

$$\mathrm{E}\,x = \frac{1}{2}(a+b),$$
$$\mathrm{E}\,x^2 = \frac{1}{3}\left(a^2 + ab + b^2\right). \tag{2.104}$$

Since

$$\mathrm{var}\,x = \mathrm{E}\,x^2 - (\mathrm{E}\,x)^2, \tag{2.105}$$

we have

$$\mathrm{var}\,x = \frac{1}{12}(b-a)^2. \tag{2.106}$$

2.13

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$
$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}). \qquad (2.107)$$

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

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

Note that

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

The right hand side can be written as

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

Then, the integral can be written as

$$\frac{1}{2} \ln \left| \frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}} \right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$$
$$+ \frac{1}{2} \int (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \qquad (2.111)$$
$$- \frac{1}{2} \int (\mathbf{x}-\mathbf{m})^{\mathsf{T}}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}.$$

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

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

we have

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

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

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

Since

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

the integral of the third term can be written as

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

Therefore,

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

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 constratints

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

let

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

Setting the variation with respect to p to zero gives

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

Then,

$$p(\mathbf{x}) = \exp\left(-1 + \lambda + \mathbf{l}^\mathsf{T}\mathbf{x} + \mathbf{m}^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{m}\right), \tag{2.122}$$

so that

$$p(\mathbf{x}) = c\exp\left(-(\mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml})^\mathsf{T}\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml})\right), \tag{2.123}$$

where

$$\begin{aligned}
c &= \exp\left(-1 + \lambda - \mathbf{l}^\mathsf{T}\mathbf{Ml}\right), \\
\mathbf{M} &= -\left(\mathbf{mm}^\mathsf{T}\right)^{-1}.
\end{aligned} \tag{2.124}$$

Substituting it to the constraints and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml} \tag{2.125}$$

gives

$$\begin{aligned}
c\int \exp\left(-\mathbf{y}^\mathsf{T}\mathbf{M}^{-1}\mathbf{y}\right) d\mathbf{y} &= 1, \\
c\int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{Ml}) \exp\left(-\mathbf{y}^\mathsf{T}\mathbf{M}^{-1}\mathbf{y}\right) d\mathbf{y} &= \boldsymbol{\mu}, \\
c\int (\mathbf{y} + \mathbf{Ml})(\mathbf{y} + \mathbf{Ml})^\mathsf{T} \exp\left(-\mathbf{y}^\mathsf{T}\mathbf{M}^{-1}\mathbf{y}\right) d\mathbf{y} &= \boldsymbol{\Sigma}.
\end{aligned} \tag{2.126}$$

Since

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

they can be written as

$$c \left(\Gamma \left(\frac{1}{2} \right) \right)^D |\det \mathbf{M}|^{\frac{1}{2}} = 1,$$

$$c(\boldsymbol{\mu} + \mathbf{Ml}) \left(\Gamma \left(\frac{1}{2} \right) \right)^D |\det \mathbf{M}|^{\frac{1}{2}} = \boldsymbol{\mu},$$

$$c \left(\Gamma \left(\frac{3}{2} \right) \left(\Gamma \left(\frac{1}{2} \right) \right)^{D-1} \mathbf{M} + \mathbf{Ml}(\mathbf{Ml})^\intercal \left(\Gamma \left(\frac{1}{2} \right) \right)^D \right) |\det \mathbf{M}|^{\frac{1}{2}} = \boldsymbol{\Sigma}. \tag{2.128}$$

Then,

$$\lambda = 1 - \frac{D}{2} \ln \pi - \frac{1}{2} \ln |\det \mathbf{M}|,$$

$$\mathbf{l} = \mathbf{0}, \tag{2.129}$$

$$\mathbf{M} = 2\boldsymbol{\Sigma}.$$

Therefore,

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

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

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

The right hand side can be written as

$$-\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \left(-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln |\det \boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x}$$

$$= \left(\frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln |\det \boldsymbol{\Sigma}| \right) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$$

$$+ \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \tag{2.133}$$

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

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

we have

$$\int (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \operatorname{tr} \boldsymbol{\Sigma}. \tag{2.135}$$

Then, the second integral can be written as

$$\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) = D. \tag{2.136}$$

Therefore,

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

2.16

Let x be a variable such that

$$x = x_1 + x_2, \tag{2.138}$$

where

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

By marginalisation,

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

The right hand side can be written as

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

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

$$-\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1(x-\mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2$$

$$+ \frac{\tau_1 + \tau_2}{2}\left(\frac{\tau_1(x-\mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2$$

$$= -\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1(x-\mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2.$$

$$(2.142)$$

Then,

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

Therefore, by 1.35,

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

2.17

Let $\boldsymbol{\Sigma}$ be a matrix and

$$\mathbf{S} = \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right),$$
$$\mathbf{A} = \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} - \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right). \tag{2.145}$$

Then,

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

so that

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

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

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.148}$$

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

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

Then,
$$(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = 0. \tag{2.150}$$

Therefore,
$$(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.151}$$

2.18

(a)

Let $\boldsymbol{\Sigma}$ be a $D \times D$ real symmetric matrix such that
$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d \mathbf{u}_d, \tag{2.152}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_D$ are unit vectors. Then,
$$\overline{\mathbf{u}_d}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d, \tag{2.153}$$

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
$$\overline{\mathbf{u}_d}^\mathsf{T}\overline{\boldsymbol{\Sigma}}^\mathsf{T}\mathbf{u}_d = \left(\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{u}_d}\right)^\mathsf{T}\mathbf{u}_d. \tag{2.154}$$

The right hand side can be writtet as
$$\overline{\lambda_d}\,\overline{\mathbf{u}_d}^\mathsf{T}\mathbf{u}_d = \overline{\lambda_d}. \tag{2.155}$$

Therefore,
$$\lambda_d = \overline{\lambda_d}. \tag{2.156}$$

(b)

For $d \neq d'$, taking the inner product with $\mathbf{u}_d'$ on both sides of
$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d \mathbf{u}_d \tag{2.157}$$

gives
$$\mathbf{u}_{d'}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d \mathbf{u}_{d'}^\mathsf{T}\mathbf{u}_d. \tag{2.158}$$

Since $\boldsymbol{\Sigma}$ is symmetric, the left hand side can be written as
$$\mathbf{u}_{d'}^\mathsf{T}\boldsymbol{\Sigma}^\mathsf{T}\mathbf{u}_d = (\boldsymbol{\Sigma}\mathbf{u}_{d'})^\mathsf{T}\mathbf{u}_d. \tag{2.159}$$

The right hand side can be written as $\lambda_{d'}\mathbf{u}_{d'}^\mathsf{T}\mathbf{u}_d$. Then,
$$\lambda_d \mathbf{u}_{d'}^\mathsf{T}\mathbf{u}_d = \lambda_{d'}\mathbf{u}_{d'}^\mathsf{T}\mathbf{u}_d. \tag{2.160}$$

Therefore, if $\lambda_d \neq \lambda_{d'}$, then
$$\mathbf{u}_{d'}^\mathsf{T}\mathbf{u}_d = 0. \tag{2.161}$$

2.19

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

$$\mathbf{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.162}$$

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

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

Then

$$\mathbf{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}. \tag{2.164}$$

By 2.18,

$$\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}. \tag{2.165}$$

Then,

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}, \\ \mathbf{\Sigma}^{-1} &= \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\mathsf{T}, \end{aligned} \tag{2.166}$$

Therefore,

$$\begin{aligned} \mathbf{\Sigma} &= \sum_{d=1}^{D} \lambda_d\mathbf{u}_d\mathbf{u}_d^\mathsf{T}, \\ \mathbf{\Sigma}^{-1} &= \sum_{d=1}^{D} \frac{1}{\lambda_d}\mathbf{u}_d\mathbf{u}_d^\mathsf{T}. \end{aligned} \tag{2.167}$$

2.20

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

$$\mathbf{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.168}$$

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

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

By 2.19,

$$\mathbf{a}^\mathsf{T}\mathbf{\Sigma}\mathbf{a} = \mathbf{b}^\mathsf{T}\mathbf{\Lambda}\mathbf{b}, \tag{2.170}$$

where
$$\mathbf{b} = \mathbf{U}^\mathsf{T}\mathbf{a}. \tag{2.171}$$

The right hand side can be written as $\sum_{d=1}^{D} \lambda_d b_d^2$. Therefore, the necessary and sufficient condition for
$$\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{a} > 0 \tag{2.172}$$

for any real vector $\mathbf{a}$ is
$$\lambda_d > 0. \tag{2.173}$$

2.21

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

2.22

Let $\boldsymbol{\Sigma}$ be a $D \times D$ symmetric matrix and
$$\boldsymbol{\Sigma}\boldsymbol{\Lambda} = \mathbf{I}. \tag{2.174}$$

Taking the transpose of the both sides gives
$$\boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{\Sigma} = \mathbf{I}. \tag{2.175}$$

Therefore,
$$\boldsymbol{\Lambda}^\mathsf{T} = \boldsymbol{\Lambda}. \tag{2.176}$$

2.23

Let $\boldsymbol{\Sigma}$ be a $D \times D$ real symmetric matrix such that
$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.177}$$

where $u_1, \cdots, u_D$ are unit vectors. Let
$$\begin{aligned}
\boldsymbol{\Lambda}' &= \operatorname{diag}\left(\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_D^{-\frac{1}{2}}\right), \\
\mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].
\end{aligned} \tag{2.178}$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{U}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}'^\mathsf{T}\mathbf{U}^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \tag{2.179}$$

By the transformation

$$\mathbf{y} = \boldsymbol{\Lambda}'^\mathsf{T}\mathbf{U}^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu}) \tag{2.180}$$

and the property

$$\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}, \tag{2.181}$$

the right hand side can be written as

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

Therefore,

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

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \tag{2.184}$$

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

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

where

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

Let

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \tag{2.186}$$

where

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

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

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$
$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T}\boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\mathsf{T}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) \tag{2.187}$$
$$-\frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\mathsf{T}\boldsymbol{\Lambda}_{ca}(\mathbf{x}_a - \boldsymbol{\mu}_a).$$

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

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

where

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

Then,

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

By marginalisation,

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

67

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

$$
-\frac{1}{2}\left(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_{a|b,c}^{-1}\left(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}\right) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Lambda}_{cc}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\
= -\frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v},
\tag{2.192}
$$

where

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

By 2.24,

$$
\mathbf{M}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{cc}^{-1} & -\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^{\mathsf{T}}\boldsymbol{\Lambda}_{aa}^{-1} \\ -\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1} & \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^{\mathsf{T}}\boldsymbol{\Lambda}_{aa}^{-1} \end{bmatrix}.
\tag{2.194}
$$

Therefore,

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

where

$$
\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right), \\
\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^{\mathsf{T}}\boldsymbol{\Lambda}_{aa}^{-1}.
\tag{2.196}
$$

2.26

We have

$$
\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\right)(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}) \\
= \mathbf{I} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D} \\
+ \mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D}.
\tag{2.197}
$$

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

$$
\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{I} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\right)\right)\mathbf{D} \\
= \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C} - \left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)\mathbf{C}\right)\mathbf{D}.
\tag{2.198}
$$

The right hand side can be written as

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

Then,

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

Therefore,

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

2.27

(a)

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

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

The right hand side can be written as

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

Therefore,

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \mathrm{E}\,\mathbf{x} + \mathrm{E}\,\mathbf{z}. \tag{2.204}$$

(b)

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

$$\mathrm{cov}(\mathbf{x} + \mathbf{z}) = \int\int \left(\mathbf{x} + \mathbf{z} - \mathrm{E}(\mathbf{x} + \mathbf{z})\right)\left(\mathbf{x} + \mathbf{z} - \mathrm{E}(\mathbf{x} + \mathbf{z})\right)^{\mathsf{T}} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \tag{2.205}$$

The right hand side can be written as

$$\int\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathbf{x} - \mathrm{E}\,\mathbf{x})^{\mathsf{T}} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} + \int\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathbf{z} - \mathrm{E}\,\mathbf{z})^{\mathsf{T}} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}$$
$$+ \int\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathbf{x} - \mathrm{E}\,\mathbf{x})^{\mathsf{T}} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z} + \int\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathbf{z} - \mathrm{E}\,\mathbf{z})^{\mathsf{T}} p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \tag{2.206}$$

Each term can be written as

$$\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})\,(\mathbf{x} - \mathrm{E}\,\mathbf{x})^{\mathsf{T}}\left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{z}\right)d\mathbf{x} = \int (\mathbf{x} - \mathrm{E}\,\mathbf{x})\,(\mathbf{x} - \mathrm{E}\,\mathbf{x})^{\mathsf{T}}\,p(\mathbf{x})d\mathbf{x},$$

$$\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})p(\mathbf{x})d\mathbf{x}\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})^{\mathsf{T}}p(\mathbf{z})d\mathbf{z} = (\mathrm{E}\,\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathrm{E}\,\mathbf{z} - \mathrm{E}\,\mathbf{z})^{\mathsf{T}},$$

$$\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})p(\mathbf{z})d\mathbf{z}\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})^{\mathsf{T}}p(\mathbf{x})d\mathbf{x} = (\mathrm{E}\,\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathrm{E}\,\mathbf{x} - \mathrm{E}\,\mathbf{x})^{\mathsf{T}},$$

$$\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})\,(\mathbf{z} - \mathrm{E}\,\mathbf{z})^{\mathsf{T}}\left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{x}\right)d\mathbf{z} = \int (\mathbf{z} - \mathrm{E}\,\mathbf{z})\,(\mathbf{z} - \mathrm{E}\,\mathbf{z})^{\mathsf{T}}\,p(\mathbf{z})d\mathbf{z}.$$

$$(2.207)$$

Therefore,

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

2.28

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

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

$$\mathrm{E}\,\mathbf{z} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}, \tag{2.209}$$

$$\mathrm{cov}\,\mathbf{z} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}} \end{bmatrix},$$

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

$$(\mathrm{cov}\,\mathbf{z})^{-1} = \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathsf{T}}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix}.$$

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

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})$$

$$+\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{L}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})$$

$$=-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)(\mathbf{x} - \boldsymbol{\mu})+$$

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})\right)^\mathsf{T}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})\right)$$

$$+\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}).$$

$$(2.210)$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}\left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}\right)^\mathsf{T}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}\right). \qquad (2.211)$$

Therefore,

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right). \qquad (2.212)$$

2.29

Let

$$\mathbf{R} = \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} & -\mathbf{A}^\mathsf{T}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix}. \qquad (2.213)$$

By 2.24,

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \end{bmatrix}. \qquad (2.214)$$

2.30

Let

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T} \end{bmatrix}. \qquad (2.215)$$

Then,

$$\mathbf{R}^{-1}\begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \qquad (2.216)$$

2.31

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

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \tag{2.217}$$

where

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

By marginalisation,

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

The right hand side can be written as

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

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

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}}\right)^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{z}}^{-1} \left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_{\mathbf{z}}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \tag{2.221}$$

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

$$-\frac{1}{2}\mathbf{u}^{\mathsf{T}}\mathbf{R}\mathbf{u} + \mathbf{u}^{\mathsf{T}}\mathbf{v} = -\frac{1}{2}\left(\mathbf{u} - \mathbf{R}^{-1}\mathbf{v}\right)^{\mathsf{T}}\mathbf{R}\left(\mathbf{u} - \mathbf{R}^{-1}\mathbf{v}\right) + \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{v}, \tag{2.222}$$

where

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

By 2.29 and 2.30,

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

Therefore,

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

2.32

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{Ax}+\mathbf{b}, \mathbf{L}^{-1}\right). \tag{2.226}$$

By the Bayes' theorem,

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

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

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

The right hand side can be written as

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

where

$$\mathbf{z} = \left(\boldsymbol{\Lambda}+\mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}). \tag{2.230}$$

The right hand side can be written as

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})^{\mathsf{T}}\left(\boldsymbol{\Lambda}+\mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z}) \\
&\quad - \frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\mathsf{T}}\mathbf{M}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}),
\end{aligned} \tag{2.231}
$$

where

$$\mathbf{M} = \mathbf{L} - \mathbf{L}\mathbf{A}\left(\boldsymbol{\Lambda}+\mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{L}. \tag{2.232}$$

We have

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

Then,

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

By 2.26,

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

Then,

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

The right hand side can be written as

$$\mathbf{L} - \mathbf{L} \left(\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right)^{-1} \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right) \mathbf{L}$$
$$= \mathbf{L} - \mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right)^{-1} \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right) \mathbf{L}.$$
$$(2.237)$$

The right hand side can be written as

$$\mathbf{L} - \mathbf{L}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right)^{-1}$$
$$= \mathbf{L} \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal - \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right) \left(\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right)^{-1}. \quad (2.238)$$

Then,

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

Therefore,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x} \,\middle|\, \left(\boldsymbol{\Lambda} + \mathbf{A}^\intercal \mathbf{L}\mathbf{A}\right)^{-1} \left(\mathbf{A}^\intercal \mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right), \left(\boldsymbol{\Lambda} + \mathbf{A}^\intercal \mathbf{L}\mathbf{A}\right)^{-1}\right),$$
$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right).$$
$$(2.240)$$

2.33

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

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}\right). \tag{2.241}$$

By the Bayes' theorem,

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

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

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{Ax} - \mathbf{b}\right)^\mathsf{T} \mathbf{L}\left(\mathbf{y} - \mathbf{Ax} - \mathbf{b}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.243}$$

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

$$-\frac{1}{2}\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} & -\mathbf{A}^\mathsf{T}\mathbf{L}\\-\mathbf{L}\mathbf{A} & \mathbf{L}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix} + \begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}^\mathsf{T}\begin{bmatrix}-\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} + \boldsymbol{\Lambda}\boldsymbol{\mu}\\\mathbf{L}\mathbf{b}\end{bmatrix}. \tag{2.244}$$

By 2.24,

$$\begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} & -\mathbf{A}^\mathsf{T}\mathbf{L}\\-\mathbf{L}\mathbf{A} & \mathbf{L}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\\\mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\end{bmatrix}, \tag{2.245}$$

so that

$$\begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A} & -\mathbf{A}^\mathsf{T}\mathbf{L}\\-\mathbf{L}\mathbf{A} & \mathbf{L}\end{bmatrix}^{-1}\begin{bmatrix}-\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} + \boldsymbol{\Lambda}\boldsymbol{\mu}\\\mathbf{L}\mathbf{b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}\\\mathbf{A}\boldsymbol{\mu} + \mathbf{b}\end{bmatrix}. \tag{2.246}$$

Then,

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right). \tag{2.247}$$

By 2.25,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu} + \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}\right), \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\right). \tag{2.248}$$

We have

$$\boldsymbol{\mu} + \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}\right)$$
$$= \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b})\right). \tag{2.249}$$

Therefore,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b})\right), \left(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\right). \tag{2.250}$$

2.34

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

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

Then,

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

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

$$\mathbf{0} = \sum_{n=1}^{N}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right)(\mathbf{x}_n - \boldsymbol{\mu}),$$

$$\mathbf{O} = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2 \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.253}$$

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

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

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^{\mathsf{T}}. \tag{2.254}$$

2.35

(a)

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

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

Then,

$$\mathrm{E}\,\mathbf{x}\mathbf{x}^{\mathsf{T}} = \mathrm{E}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.256}$$

The right hand side can be written as

$$\mathrm{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} + \boldsymbol{\mu}\,\mathrm{E}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} + \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\mu}^{\mathsf{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.257}$$

Since
$$\mathrm{E}\,\mathbf{x} = \boldsymbol{\mu},$$
$$\mathrm{cov}\,\mathbf{x} = \boldsymbol{\Sigma}, \tag{2.258}$$

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

$$\mathrm{E}\,\mathbf{x}\mathbf{x}^{\mathsf{T}} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.259}$$

(b)

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

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(\mathbf{x}_m) = \mathcal{N}\left(\mathbf{x_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.260}$$

If $n \neq m$, then
$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^{\mathsf{T}} = \mathrm{E}\,\mathbf{x}_n\,\mathrm{E}\,\mathbf{x}_m^{\mathsf{T}}. \tag{2.261}$$

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

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

(c)

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

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

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

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n,$$
$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}. \tag{2.264}$$

Then,

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

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^{N} \mathrm{E}\, \mathbf{x}_n \mathbf{x}_n^\mathsf{T} - \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{E} \left(\sum_{n=1}^{N} \mathbf{x_n} \right) \mathbf{x}_n^\mathsf{T} - \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{E}\, \mathbf{x}_n \left(\sum_{n=1}^{N} \mathbf{x_n} \right)^\mathsf{T}$$
$$+ \frac{1}{N^3} \sum_{n=1}^{N} \mathrm{E} \left(\sum_{n=1}^{N} \mathbf{x_n} \right) \left(\sum_{n=1}^{N} \mathbf{x_n} \right)^\mathsf{T}.$$

(2.266)

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

$$-\frac{1}{N} \left((\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}) + (N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} \right) = -\frac{1}{N}\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}.$$

(2.267)

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

$$\frac{1}{N^2} \left(N\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} \right) + N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} \right) = \frac{1}{N}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}.$$

(2.268)

Then,

$$\mathrm{E}\, \boldsymbol{\Sigma}_{\mathrm{ML}} = (\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}) + 2\left(-\frac{1}{N}\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} \right) + \frac{1}{N}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}.$$

(2.269)

Therefore,

$$\mathrm{E}\, \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

(2.270)

2.36

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

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

(2.271)

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

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

(2.272)

The right hand side can be written as

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

(2.273)

Then,

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

We have

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

Therefore,

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

2.37

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

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

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

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.278}$$

The right hand side can be written as

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

Then,

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}\left((\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} - \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right). \tag{2.280}$$

By 3.21(a), we have

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\left(-\ln p(x_n)\right) = -\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.281}$$

Therefore,

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

2.38

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

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

By the Bayes' theorem,

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

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

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

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

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n. \tag{2.286}$$

Then, the first term can be written as

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

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{\mu_{\mathrm{ML}} - \mu}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}}) - \frac{N}{2\sigma^2}(\mu_{\mathrm{ML}} - \mu)^2. \tag{2.287}$$

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

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2}(\mu_{\mathrm{ML}} - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{1}{2\sigma_N^2}(\mu - \mu_N)^2 + \frac{\mu_N^2}{2\sigma_N^2} - \frac{N\mu_{\mathrm{ML}}^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}, \tag{2.288}$$

where
$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0,$$
$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}.$$
$$(2.289)$$

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

2.39

Let $x_1, \cdots, x_N$ be variables such that
$$p(x_n|\mu) = \mathcal{N}\left(x_n|\mu, \sigma^2\right),$$
$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$
$$(2.291)$$

(a)

By 2.38,
$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right), \tag{2.292}$$

where
$$\mu_N = \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2}\sum_{n=1}^{N}x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0,$$
$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}.$$
$$(2.293)$$

Then,
$$\frac{1}{\sigma_N^2} = \frac{(N-1)\sigma_0^2 + \sigma^2}{\sigma^2\sigma_0^2} + \frac{1}{\sigma^2}. \tag{2.294}$$

Therefore,
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \tag{2.295}$$

We have
$$\frac{\mu_N}{\sigma_N^2} = \frac{1}{\sigma^2}\sum_{n=1}^{N}x_n + \frac{\mu_0}{\sigma_0^2}, \tag{2.296}$$

so that
$$\frac{\mu_{N-1}}{\sigma_{N-1}^2} = \frac{1}{\sigma^2}\sum_{n=1}^{N-1}x_n + \frac{\mu_0}{\sigma_0^2}. \tag{2.297}$$

Therefore,

$$\frac{\mu_N}{\sigma_N^2} = \frac{\mu_{N-1}}{\sigma_{N-1}^2} + \frac{x_N}{\sigma^2}. \tag{2.298}$$

(b)

By the Bayes' theorem,

$$p(\mu|\mathbf{x}_N)p(\mathbf{x}_N) = p(\mathbf{x}_N|\mu)p(\mu). \tag{2.299}$$

Since x_N and $\mathbf{x}_{N-1}$ are independent, it can be written as

$$p(\mu|\mathbf{x}_N)p(x_N)p(\mathbf{x}_{N-1}) = p(x_N|\mu)p(\mathbf{x}_{N-1}|\mu)p(\mu). \tag{2.300}$$

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

$$p(x_N|\mu)p(\mu|\mathbf{x}_{N-1})p(\mathbf{x}_{N-1}). \tag{2.301}$$

Then,

$$p(\mu|\mathbf{x}_N)p(x_N) = p(\mu|\mathbf{x}_{N-1})p(x_N|\mu). \tag{2.302}$$

The logarithm of the integrand of the right hand side except the terms independent of μ or x_N can be written as

$$-\frac{1}{2\sigma_{N-1}^2}(\mu - \mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N - \mu)^2$$

$$= -\frac{1}{2}\left(\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}\right)\left(\mu - \frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}\mu_{N-1} - \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}x_N\right)^2$$

$$+ \frac{1}{2}\left(\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}\right)\left(\frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}\mu_{N-1} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}x_N\right)^2 - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}. \tag{2.303}$$

Then,

$$\mu_N = \frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}\mu_{N-1} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}x_N,$$

$$\sigma_N^2 = \frac{1}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}. \tag{2.304}$$

Therefore,

$$\frac{\mu_N}{\sigma_N^2} = \frac{\mu_{N-1}}{\sigma_{N-1}^2} + \frac{x_N}{\sigma^2},$$
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \tag{2.305}$$

2.40

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

$$p(\mathbf{x}_n|\boldsymbol{\mu}) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(\boldsymbol{\mu}) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right). \tag{2.306}$$

By 2.34, the maximum likelihood solution for $\boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \tag{2.307}$$

By the Bayes' theorem,

$$p(\boldsymbol{\mu}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}). \tag{2.308}$$

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

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathsf{T}}. \tag{2.309}$$

The first term can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})$$
$$= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - (\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})$$
$$- \frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}). \tag{2.310}$$

The second term of the right hand side is zero. Then, the logarithm except the terms independent of $\mathbf{X}$ and $\boldsymbol{\mu}$ can be written as

$$
\begin{aligned}
& -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}) \\
& \quad -\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\mathsf{T} \\
& = -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_N) \\
& \quad +\frac{1}{2}\boldsymbol{\mu}_N^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N - \frac{N}{2}\boldsymbol{\mu}_{\mathrm{ML}}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}},
\end{aligned}
\tag{2.311}
$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_N &= \left(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right), \\
\boldsymbol{\Sigma}_N &= \left(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}.
\end{aligned}
\tag{2.312}
$$

Therefore,

$$
p(\boldsymbol{\mu}|\mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N).
\tag{2.313}
$$

2.41

By the definition,

$$
\mathrm{Gam}(\lambda|a, b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda).
\tag{2.314}
$$

Then,

$$
\int_0^\infty \mathrm{Gam}(\lambda|a, b)d\lambda = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a-1}\exp(-b\lambda)d\lambda.
\tag{2.315}
$$

By the transformation

$$
\lambda' = b\lambda,
\tag{2.316}
$$

the right hand side can be written as

$$
\frac{b^a}{\Gamma(a)}\int_0^\infty \left(\frac{\lambda'}{b}\right)^{a-1}\exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{\Gamma(a)}\int_0^\infty \lambda'^{a-1}\exp(-\lambda')d\lambda'.
\tag{2.317}
$$

The right hand side can be written as

$$\frac{1}{\Gamma(a)}\Gamma(a) = 1. \qquad (2.318)$$

Therefore,

$$\int_0^\infty \text{Gam}(\lambda|a,b)d\lambda = 1. \qquad (2.319)$$

2.42

Let λ be a variable such that

$$p(\lambda) = \text{Gam}(\lambda|a,b). \qquad (2.320)$$

By the definition,

$$\text{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \qquad (2.321)$$

(a)

We have

$$\text{E}\,\lambda = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right)d\lambda. \qquad (2.322)$$

By the transformation

$$\lambda' = b\lambda, \qquad (2.323)$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b\Gamma(a)}\int_0^\infty \lambda'^a \exp(-\lambda')d\lambda'. \qquad (2.324)$$

The right hand side can be written as

$$\frac{1}{b\Gamma(a)}\Gamma(a+1) = \frac{a}{b}. \qquad (2.325)$$

Therefore,

$$\text{E}\,\lambda = \frac{a}{b}. \qquad (2.326)$$

(b)

We have

$$\mathrm{E}\,\lambda^2 = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1} \exp\left(-\frac{\lambda}{b}\right) d\lambda. \tag{2.327}$$

By the transformation

$$\lambda' = b\lambda, \tag{2.328}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b^2\Gamma(a)} \int_0^\infty \lambda'^{a+1} \exp(-\lambda')d\lambda'. \tag{2.329}$$

The right hand side can be written as

$$\frac{1}{b^2\Gamma(a)}\Gamma(a+2) = \frac{a(a+1)}{b^2}. \tag{2.330}$$

Then,

$$\mathrm{E}\,\lambda^2 = \frac{a(a+1)}{b^2}. \tag{2.331}$$

We have

$$\mathrm{var}\,\lambda = \mathrm{E}\,\lambda^2 - (\mathrm{E}\,\lambda)^2. \tag{2.332}$$

Therefore,

$$\mathrm{var}\,\lambda = \frac{a}{b^2}. \tag{2.333}$$

(c)

Setting the derivative of $\mathrm{Gam}(\lambda|a,b)$ with respect to λ to zero gives

$$0 = \frac{b^a}{\Gamma(a)}\left(\frac{a-1}{\lambda} - b\right)\lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right). \tag{2.334}$$

Therefore,

$$\mathrm{mode}\,\lambda = \frac{a-1}{b}. \tag{2.335}$$

2.43

Let

$$p\left(x|\sigma^2, q\right) = \frac{q}{2\Gamma(\frac{1}{q})}\left(2\sigma^2\right)^{-\frac{1}{q}} \exp\left(-\frac{|x|^q}{2\sigma^2}\right). \tag{2.336}$$

(a)

We have

$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = \frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^{\infty} \exp\left(-\frac{x^q}{2\sigma^2}\right) dx. \tag{2.337}$$

By the transformation

$$x' = \frac{x^q}{2\sigma^2}, \tag{2.338}$$

the right hand side can be written as

$$\frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^{\infty} \exp(-x') \left(2\sigma^2\right)^{\frac{1}{q}} \frac{1}{q} x'^{\frac{1}{q}-1} dx'$$

$$= \frac{1}{\Gamma(\frac{1}{q})} \int_0^{\infty} x'^{\frac{1}{q}-1} \exp(-x') dx'. \tag{2.339}$$

The right hand side can be written as

$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma\left(\frac{1}{q}\right) = 1. \tag{2.340}$$

Therefore,

$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = 1. \tag{2.341}$$

(b)

We have

$$p\left(x|\sigma^2, 2\right) = \frac{1}{\Gamma(\frac{1}{2})} \left(2\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{2.342}$$

Therefore,

$$p\left(x|\sigma^2, 2\right) = \mathcal{N}\left(x|0, \sigma^2\right). \tag{2.343}$$

(c)

Let $\mathbf{t} = (t_1, \cdots, t_N)^{\mathsf{T}}$ and $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ such that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \tag{2.344}$$

where
$$p(\epsilon_n) = p\left(\epsilon_n | \sigma^2, q\right). \tag{2.345}$$

Then, the logarithm of $p(\epsilon_n)$ except the terms independent of $\mathbf{w}$ and σ^2 can be written as
$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q}\ln\left(2\sigma^2\right). \tag{2.346}$$

Therefore, the logarithm of $p\left(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2\right)$ except the terms independent of $\mathbf{w}$ and σ^2 can be written as
$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q}\ln\left(2\sigma^2\right). \tag{2.347}$$

2.44

Let $x_1, \cdots, x_N$ be variables such that
$$\begin{aligned}
p(x_n|\mu, \tau) &= \mathcal{N}\left(x_n|\mu, \tau^{-1}\right), \\
p(\mu, \tau) &= \mathcal{N}\left(\mu|\mu_0, (\beta\tau)^{-1}\right)\mathrm{Gam}(\tau|a, b).
\end{aligned} \tag{2.348}$$

By the Bayes' theorem,
$$p(\mu, \tau|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu, \tau)p(\mu, \tau). \tag{2.349}$$

The logarithm of the right hand side except the terms independent of $\mathbf{x}$, μ and τ can be written as
$$\begin{aligned}
&\frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 + \frac{1}{2}\ln\tau - \frac{\beta\tau}{2}(\mu - \mu_0)^2 + (a - 1)\ln\tau - b\tau \\
&= \left(a + \frac{N-1}{2}\right)\ln\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{\beta\tau}{2}(\mu - \mu_0)^2 - b\tau.
\end{aligned} \tag{2.350}$$

Let
$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N}x_n. \tag{2.351}$$

Then,
$$\sum_{n=1}^{N}(x_n - \mu)^2 = \sum_{n=1}^{N}(x_n - \bar{x} + \bar{x} - \mu)^2. \tag{2.352}$$

The right hand side can be written as

$$\sum_{n=1}^{N}(x_n - \bar{x})^2 + 2(\bar{x} - \mu)\sum_{n=1}^{N}(x_n - \bar{x}) + N(\bar{x} - \mu)^2$$
$$= \sum_{n=1}^{N}(x_n - \bar{x})^2 + N(\bar{x} - \mu)^2. \tag{2.353}$$

Then, the logarithm except the terms independent of $\mathbf{x}$, μ and τ can be written as

$$\left(a + \frac{N-1}{2}\right)\ln\tau - \frac{N\tau}{2}(\bar{x} - \mu)^2 - \frac{\beta\tau}{2}(\mu - \mu_0)^2 - b\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2. \tag{2.354}$$

The second and third terms can be written as

$$-\frac{N\tau}{2}(\bar{x} - \mu)^2 - \frac{\beta\tau}{2}(\mu - \mu_0)^2$$
$$= -\frac{(N+\beta)\tau}{2}\left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2 + \frac{(N\bar{x} + \beta\mu_0)^2\tau}{2(N+\beta)} - \frac{N\tau}{2}\bar{x}^2 - \frac{\beta\tau}{2}\mu_0^2. \tag{2.355}$$

The second, third and forth terms on the right hand side can be written as

$$\frac{(N\bar{x} + \beta\mu_0)^2\tau - (N+\beta)N\tau\bar{x}^2 - (N+\beta)\beta\tau\mu_0^2}{2(N+\beta)} = -\frac{N\beta\tau(\bar{x} - \mu_0)^2}{2(N+\beta)}. \tag{2.356}$$

Then, the logrithm except the terms independent of $\mathbf{x}$, μ and τ can be written as

$$-\frac{(N+\beta)\tau}{2}\left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2$$
$$+ \left(a + \frac{N-1}{2}\right)\ln\tau - \left(b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N+\beta)} + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2\right)\tau. \tag{2.357}$$

Therefore,

$$p(\mu, \tau|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \tau_N^{-1}\right)\mathrm{Gam}\left(\tau|a_N, b_N\right), \tag{2.358}$$

where

$$\mu_N = \frac{N\bar{x} + \beta\mu_0}{N + \beta},$$
$$\tau_N = (N + \beta)\tau,$$
$$a_N = a + \frac{N + 1}{2},$$
$$b_N = b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N + \beta)} + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2. \tag{2.359}$$

2.45

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables in D dimensions such that

$$p(\mathbf{x}_n|\mathbf{\Lambda}) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}\right),$$
$$p(\mathbf{\Lambda}) = \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu), \tag{2.360}$$

where

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu)|\det \mathbf{\Lambda}|^{\frac{\nu - D - 1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{-1}\mathbf{\Lambda}\right)\right). \tag{2.361}$$

By the Bayes' theorem,

$$p(\mathbf{\Lambda}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\mathbf{\Lambda})p(\mathbf{\Lambda}). \tag{2.362}$$

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

$$-\frac{N}{2}\ln\left|\det\left(\mathbf{\Lambda}^{-1}\right)\right| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}\mathbf{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})$$
$$+\frac{\nu - D - 1}{2}\ln|\det\mathbf{\Lambda}| - \frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{-1}\mathbf{\Lambda}\right) \tag{2.363}$$
$$=\frac{\nu + N - D - 1}{2}\ln|\det\mathbf{\Lambda}| - \frac{1}{2}\operatorname{tr}\left(\left(\mathbf{W}^{-1} + \mathbf{S}\right)\mathbf{\Lambda}\right),$$

where

$$\mathbf{S} = \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}. \tag{2.364}$$

Then,

$$p(\mathbf{\Lambda}|\mathbf{X}) = \mathcal{W}\left(\mathbf{\Lambda}|\left(\mathbf{W}^{-1} + \mathbf{S}\right)^{-1}, \nu + N\right). \tag{2.365}$$

Therefore, $\mathcal{W}$ is a conjugate prior distribution of $\mathbf{\Lambda}$.

2.46

Let x be a variable such that

$$p(x|\tau) = \mathcal{N}\left(x|\mu, \tau^{-1}\right),$$
$$p(\tau) = \text{Gam}(\tau|a, b). \tag{2.366}$$

By marginalisation,

$$p(x) = \int_0^\infty p(x|\tau)p(\tau)d\tau. \tag{2.367}$$

The right hand side can be written as

$$\int_0^\infty \left(2\pi\tau^{-1}\right)^{-\frac{1}{2}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau)d\tau$$
$$= (2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp\left(-\left(b + \frac{(x-\mu)^2}{2}\right)\tau\right)d\tau. \tag{2.368}$$

By the transformation

$$\tau' = \left(b + \frac{(x-\mu)^2}{2}\right)\tau, \tag{2.369}$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{\tau'}{b + \frac{(x-\mu)^2}{2}}\right)^{a-\frac{1}{2}} \exp(-\tau') \frac{1}{b + \frac{(x-\mu)^2}{2}} d\tau'$$
$$= \Gamma\left(a + \frac{1}{2}\right)\left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.370}$$

Then,

$$p(x) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.371}$$

By the transformation

$$\nu = 2a,$$
$$\lambda = \frac{a}{b}, \tag{2.372}$$

91

the right hand side can be written as

$$(2\pi)^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\nu}{2\lambda}\right)^{-\frac{1}{2}} \left(1 + \frac{(x-\mu)^2}{\frac{\nu}{\lambda}}\right)^{-\frac{\nu+1}{2}}$$
$$= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda}{\nu}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}. \tag{2.373}$$

Therefore,

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda}{\nu}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}. \tag{2.374}$$

2.47

Let

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda}{\nu}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}. \tag{2.375}$$

By the transformation

$$\frac{1}{y} = \frac{\lambda}{\nu}(x-\mu)^2, \tag{2.376}$$

the right hand side except the terms independent of x can be written as

$$\left(1 + \frac{1}{y}\right)^{-\frac{\lambda(x-\mu)^2}{2}y - \frac{1}{2}}. \tag{2.377}$$

By the property

$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = e, \tag{2.378}$$

we have

$$\lim_{y \to \infty} \left(1 + \frac{1}{y}\right)^{-\frac{\lambda(x-\mu)^2}{2}y - \frac{1}{2}} = \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right). \tag{2.379}$$

Therefore,

$$\lim_{\nu \to \infty} \text{St}(x|\mu, \lambda, \nu) = \mathcal{N}(x|\mu, \lambda^{-1}). \tag{2.380}$$

2.48

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

$$p(\mathbf{x}|\eta) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right),$$
$$p(\eta) = \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right). \tag{2.381}$$

By marginalisation,

$$p(\mathbf{x}) = \int_0^\infty p(\mathbf{x}|\eta)p(\eta)d\eta. \tag{2.382}$$

The right hand side can be written as

$$\int_0^\infty (2\pi)^{-\frac{D}{2}}\left|\det(\eta\boldsymbol{\Lambda})^{-1}\right|^{-\frac{1}{2}}\exp\left(-\frac{\eta}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\eta^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\eta\right)d\eta$$

$$=(2\pi)^{-\frac{D}{2}}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\int_0^\infty \eta^{\frac{D+\nu}{2}-1}\exp\left(-\frac{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2}\eta\right)d\eta. \tag{2.383}$$

By the transformation

$$\eta' = \frac{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2}\eta, \tag{2.384}$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{2\eta'}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}-1}\exp(-\eta')\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}d\eta'$$

$$=\left(\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}}\int_0^\infty \eta'^{\frac{D+\nu}{2}-1}\exp(-\eta')d\eta'. \tag{2.385}$$

Then,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\left(\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}}\Gamma\left(\frac{D+\nu}{2}\right). \tag{2.386}$$

The right hand side can be written as

$$(2\pi)^{-\frac{D}{2}}\frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\left(\frac{\nu}{2}\right)^{-\frac{D}{2}}\left(\frac{\nu}{2}\right)^{\frac{D+\nu}{2}}\left(\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}}$$

$$=(2\pi)^{-\frac{D}{2}}\frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\left(\frac{\nu}{2}\right)^{-\frac{D}{2}}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \tag{2.387}$$

Therefore,

$$p(\mathbf{x}) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\det \mathbf{\Lambda}|^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \tag{2.388}$$

2.49

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

$$p(\mathbf{x}) = \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu), \tag{2.389}$$

where

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.390}$$

(a)

We have

$$\mathrm{E}\,\mathbf{x} = \int \mathbf{x}\,\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) d\mathbf{x}. \tag{2.391}$$

The right hand side can be written as

$$\begin{aligned}
&\int \mathbf{x}\left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta\right) d\mathbf{x} \\
&= \int \left(\int \mathbf{x}\,\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right) d\mathbf{x}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta.
\end{aligned} \tag{2.392}$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \tag{2.393}$$

Therefore,

$$\mathrm{E}\,\mathbf{x} = \boldsymbol{\mu}. \tag{2.394}$$

(b)

By (a), we have

$$\mathrm{cov}\,\mathbf{x} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) d\mathbf{x}. \tag{2.395}$$

94

The right hand side can be written as

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \left(\int \mathcal{N} \left(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1} \right) \mathrm{Gam} \left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2} \right) d\eta \right) d\mathbf{x}$$
$$= \int \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N} \left(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1} \right) d\mathbf{x} \right) \mathrm{Gam} \left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2} \right) d\eta. \tag{2.396}$$

The right hand side can be written as

$$\int (\eta \boldsymbol{\Lambda})^{-1} \mathrm{Gam} \left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2} \right) d\eta = \frac{\left(\frac{\nu}{2} \right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\int \eta^{\frac{\nu}{2}-2} \exp \left(-\frac{\nu}{2} \eta \right) d\eta \right) \boldsymbol{\Lambda}^{-1}. \tag{2.397}$$

By the transformation

$$\eta' = \frac{\nu}{2} \eta, \tag{2.398}$$

the integral of the right hand side can be written as

$$\int \left(\frac{2}{\nu} \eta' \right)^{\frac{\nu}{2}-2} \exp(-\eta') \frac{2}{\nu} d\eta' = \left(\frac{2}{\nu} \right)^{\frac{\nu}{2}-1} \Gamma \left(\frac{\nu}{2} - 1 \right). \tag{2.399}$$

Then,

$$\mathrm{cov}\, \mathbf{x} = \frac{\left(\frac{\nu}{2} \right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\frac{2}{\nu} \right)^{\frac{\nu}{2}-1} \Gamma \left(\frac{\nu}{2} - 1 \right) \boldsymbol{\Lambda}^{-1}. \tag{2.400}$$

Therefore,

$$\mathrm{cov}\, \mathbf{x} = \frac{\nu}{\nu - 2} \boldsymbol{\Lambda}^{-1}. \tag{2.401}$$

(c)

Setting the derivative of $p(\mathbf{x})$ to zero gives

$$\mathbf{0} = -\frac{1}{2} \left(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^{\mathsf{T}} \right) (\mathbf{x} - \boldsymbol{\mu}) \int \eta \mathcal{N} \left(\mathbf{x} | \boldsymbol{\mu}, (\eta \boldsymbol{\Lambda})^{-1} \right) \mathrm{Gam} \left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2} \right) d\eta. \tag{2.402}$$

Therefore,

$$\mathrm{mode}\, \mathbf{x} = \boldsymbol{\mu}. \tag{2.403}$$

2.50

Let

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.404)$$

By the transformation

$$y = \frac{\nu}{(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}, \quad (2.405)$$

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

$$\left(1 + \frac{1}{y}\right)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2}y - \frac{D}{2}}. \quad (2.406)$$

By the property

$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^{x} = e, \quad (2.407)$$

we have

$$\lim_{y \to \infty} \left(1 + \frac{1}{y}\right)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2}y - \frac{D}{2}} = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (2.408)$$

Therefore,

$$\lim_{\nu \to \infty} \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.409)$$

2.51

(a)

We have

$$\exp(iA)\exp(-iA) = 1. \quad (2.410)$$

The left hand side can be written as

$$(\cos A + i\sin A)(\cos A - i\sin A) = \cos^2 A + \sin^2 A. \quad (2.411)$$

Therefore,

$$\cos^2 A + \sin^2 A = 1. \quad (2.412)$$

96

(b)

We have

$$\cos(A - B) = \text{Re}\left(\exp\left(i(A - B)\right)\right). \tag{2.413}$$

The right hand side can be written as

$$\text{Re}\left(\exp(iA)\exp(-iB)\right) = \text{Re}\left((\cos A + i\sin A)(\cos B - i\sin B)\right). \tag{2.414}$$

Therefore,

$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \tag{2.415}$$

(c)

We have

$$\sin(A - B) = \text{Im}\left(\exp\left(i(A - B)\right)\right). \tag{2.416}$$

The right hand side can be written as

$$\text{Im}\left(\exp(iA)\exp(-iB)\right) = \left((\cos A + i\sin A)(\cos B - i\sin B)\right). \tag{2.417}$$

Therefore,

$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \tag{2.418}$$

2.52

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right), \tag{2.419}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta. \tag{2.420}$$

By the Taylor series

$$\cos\alpha = 1 - \frac{1}{2}\alpha^2 + O\left(\alpha^4\right), \tag{2.421}$$

the right hand side can be written as

$$\frac{\exp\left(m\left(1 - \frac{1}{2}(\theta - \theta_0)^2 + O\left((\theta - \theta_0)^4\right)\right)\right)}{\int_0^{2\pi}\exp\left(m\left(1 - \frac{1}{2}\theta^2 + O(\theta^4)\right)\right)d\theta}$$

$$= \exp\left(-\frac{m}{2}(\theta - \theta_0)^2\right)\frac{\exp\left(mO\left((\theta - \theta_0)^4\right)\right)}{\int_0^{2\pi}\exp\left(m\left(-\frac{1}{2}\theta^2 + O(\theta^4)\right)\right)d\theta}. \tag{2.422}$$

Therefore,

$$\lim_{m\to\infty} f(\theta|\theta_0, m) = \mathcal{N}\left(\theta|\theta_0, m^{-1}\right). \tag{2.423}$$

2.53

Let

$$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0. \tag{2.424}$$

The left hand side can be written as

$$\sum_{n=1}^{N} (\sin\theta_n \cos\theta_0 - \cos\theta_n \sin\theta_0) = \cos\theta_0 \sum_{n=1}^{N} \sin\theta_n - \sin\theta_0 \sum_{n=1}^{N} \cos\theta_n. \tag{2.425}$$

Therefore,

$$\theta_0 = \arctan\left(\frac{\sum_{n=1}^{N} \sin\theta_n}{\sum_{n=1}^{N} \cos\theta_n}\right). \tag{2.426}$$

2.54

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right), \tag{2.427}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta. \tag{2.428}$$

Setting the first and seconnd derivatives with respect to θ to zero gives

$$
\begin{aligned}
0 &= -m\sin(\theta - \theta_0)f(\theta|\theta_0, m), \\
0 &= \left(m^2\sin^2(\theta - \theta_0) - m\cos(\theta - \theta_0)\right)f(\theta|\theta_0, m).
\end{aligned} \tag{2.429}
$$

Therefore,

$$
\begin{aligned}
\operatorname*{argmax}_{\theta} f(\theta|\theta_0, m) &= \theta_0, \\
\operatorname*{argmin}_{\theta} f(\theta|\theta_0, m) &= \theta_0 - \pi\operatorname{sgn}(\theta_0 - \pi).
\end{aligned} \tag{2.430}
$$

2.55

(a)

Let $\theta_1, \cdots, \theta_N$ be variables such that

$$p(\theta_n) = f(\theta_n|\theta_0, m), \tag{2.431}$$

98

where

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right),$$
$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta. \tag{2.432}$$

Then,

$$\ln\left(\prod_{n=1}^N p(\theta_n|\theta_0, m)\right) = -\frac{N}{2}\ln\left(2\pi I_0(m)\right) + m\sum_{n=1}^N \cos(\theta_n - \theta_0). \tag{2.433}$$

Setting the derivative with respect to θ_0 to zero gives

$$0 = m\sum_{n=1}^N \sin(\theta_n - \theta_0). \tag{2.434}$$

Therefore, by 2.53, the maximum likelihood solution for θ_0 is given by

$$\theta_0^{\text{ML}} = \arctan\left(\frac{\sum_{n=1}^N \sin\theta_n}{\sum_{n=1}^N \cos\theta_n}\right). \tag{2.435}$$

(b)

Let

$$\bar{r}\cos\bar{\theta} = \frac{1}{N}\sum_{n=1}^N \cos\theta_n,$$
$$\bar{r}\sin\bar{\theta} = \frac{1}{N}\sum_{n=1}^N \sin\theta_n. \tag{2.436}$$

By (a),

$$\bar{\theta} = \theta_0^{\text{ML}}. \tag{2.437}$$

We have

$$\frac{1}{N}\sum_{n=1}^N \cos\left(\theta_n - \theta_0^{\text{ML}}\right) = \left(\frac{1}{N}\sum_{n=1}^N \cos\theta_n\right)\cos\theta_0^{\text{ML}} + \left(\frac{1}{N}\sum_{n=1}^N \sin\theta_n\right)\sin\theta_0^{\text{ML}}. \tag{2.438}$$

The right hand side can be written as

$$\bar{r}\cos^2\bar{\theta} + \bar{r}\sin^2\bar{\theta} = \bar{r}. \tag{2.439}$$

Therefore,

$$\frac{1}{N} \sum_{n=1}^{N} \cos\left(\theta_n - \theta_0^{\text{ML}}\right) = \bar{r}. \tag{2.440}$$

2.56

(a)

Let

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}. \tag{2.441}$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \exp\left((a-1)\ln\mu + (b-1)\ln(1-\mu)\right) \tag{2.442}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}. $$

(b)

Let

$$\text{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda). \tag{2.443}$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \exp\left((a-1)\ln\lambda - b\lambda\right). \tag{2.444}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ -b \end{bmatrix}. $$

(c)

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right), \qquad (2.445)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta)d\theta, \qquad (2.446)$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m\cos\theta_0\cos\theta + m\sin\theta_0\sin\theta). \qquad (2.447)$$

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,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \qquad (2.448)$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right), \qquad (2.449)$$

where

$$h(\mathbf{x}) = (2\pi)^{-\frac{D}{2}},$$

$$g(\boldsymbol{\eta}) = (\det(-2\boldsymbol{\eta}_2))^{-\frac{1}{2}} \exp\left(\frac{1}{4}\boldsymbol{\eta}_1^{\mathsf{T}}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1\right),$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix},$$

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}^{\mathsf{T}} \end{bmatrix}.$$

2.58

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

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right). \tag{2.450}$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = 1 \tag{2.451}$$

with respect to $\boldsymbol{\eta}$ gives

$$\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right)d\mathbf{x} + g(\boldsymbol{\eta})\int \mathbf{u}(\mathbf{x})h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right)d\mathbf{x} = \mathbf{0}. \tag{2.452}$$

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathrm{E}\,\mathbf{u}(\mathbf{x}). \tag{2.453}$$

Therefore,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}. \tag{2.454}$$

Thus,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}). \tag{2.455}$$

Taking the second derivative with respect to $\boldsymbol{\eta}$ gives

$$\nabla\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right)d\mathbf{x} + 2\nabla g(\boldsymbol{\eta})\int \mathbf{u}(\mathbf{x})^\mathsf{T}h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right)d\mathbf{x}$$
$$+ g(\boldsymbol{\eta})\int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\mathsf{T}h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\right)d\mathbf{x} = \mathbf{O}. \tag{2.456}$$

The left hand side can be written as

$$\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \frac{2\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\int \mathbf{u}(\mathbf{x})^\mathsf{T}p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\mathsf{T}p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x}$$
$$= \frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2\,\mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^\mathsf{T} + \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\mathsf{T}\right). \tag{2.457}$$

Therefore,

$$E\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\mathsf{T}\right) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\mathsf{T}}{g^2(\boldsymbol{\eta})}. \tag{2.458}$$

By the definition,

$$\text{cov}\,\mathbf{u}(\mathbf{x}) = E\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\mathsf{T}\right) - E\,\mathbf{u}(\mathbf{x})\,E\,\mathbf{u}(\mathbf{x})^\mathsf{T}. \tag{2.459}$$

Thus,

$$\text{cov}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^\mathsf{T}}{g^2(\boldsymbol{\eta})}. \tag{2.460}$$

Hence,

$$\text{cov}\,\mathbf{u}(\mathbf{x}) = -\nabla\nabla \ln g(\boldsymbol{\eta}). \tag{2.461}$$

2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right). \tag{2.462}$$

Then

$$\int p(x|\sigma)dx = \frac{1}{\sigma}\int f\left(\frac{x}{\sigma}\right)dx. \tag{2.463}$$

By the transformation

$$x' = \frac{x}{\sigma}, \tag{2.464}$$

the right hand side can be written as

$$\frac{1}{\sigma}\int f(x')\sigma dx' = \int f(x')dx'. \tag{2.465}$$

Therefore, $p(x|\sigma)$ will be normalised if $f(x)$ is normalised.

2.60

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

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \tag{2.466}$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \tag{2.467}$$

Since

$$\int p(\mathbf{x})d\mathbf{x} = 1,$$ (2.468)

we have

$$\sum_i h_i \Delta_i = 1.$$ (2.469)

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

$$\ln\left(\prod_i h_i^{n_i}\right) = \sum_i n_i \ln h_i,$$ (2.470)

where

$$\sum_i n_i = N.$$ (2.471)

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda\left(\sum_i h_i \Delta_i - 1\right)$$ (2.472)

with respect to h_i and λ to zero gives

$$\frac{n_i}{h_i} + \lambda \Delta_i = 0,$$
$$\sum_i h_i \Delta_i - 1 = 0.$$ (2.473)

Then,

$$\lambda = -N,$$
$$h_i = \frac{n_i}{N\Delta_i}.$$ (2.474)

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 $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be observations. Let

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})},$$ (2.475)

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}'-\mathbf{x}\| \leq \|\mathbf{x}_{(K)}-\mathbf{x}\|} d\mathbf{x}', \qquad (2.476)$$

K is a constant and $\mathbf{x}_{(K)}$ is the Kth nearest observation from the point $\mathbf{x}$.

3   Linear Models for Regression

3.1

By the definition,

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \tag{3.1}$$

The right hand side can be written as

$$\frac{1 - \exp(-2a)}{1 + \exp(-2a)} = \frac{2}{1 + \exp(-2a)} - 1. \tag{3.2}$$

Therefore,

$$\tanh a = 2\sigma(2a) - 1, \tag{3.3}$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{3.4}$$

Let

$$y(x_n, \mathbf{w}) = w_0 + \sum_{m=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right). \tag{3.5}$$

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

$$w_0 + \sum_{m=1}^{M} w_m \frac{1 + \tanh\left(\frac{x - \mu_m}{2s}\right)}{2} = w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m + \frac{1}{2}\sum_{m=1}^{M} w_m \tanh\left(\frac{x - \mu_m}{2s}\right). \tag{3.6}$$

Therefore, $y(x_n, \mathbf{w})$ is equivalent to

$$y(x_n, \mathbf{u}) = u_0 + \sum_{m=1}^{M} u_m \tanh\left(\frac{x - \mu_m}{2s}\right), \tag{3.7}$$

where

$$u_0 = w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m,$$
$$u_m = \frac{1}{2}w_m. \tag{3.8}$$

106

3.2   (Incomplete)

Let $\boldsymbol{\Phi}$ be an $N \times M$ matarix. Then, for any vector $\mathbf{v}$ in N dimensions,

$$\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{v} \tag{3.9}$$

is a projection of $\mathbf{v}$ onto the space spanned by the columns of $\boldsymbol{\Phi}$?
    Additionally, for a vector $\mathbf{t}$ in N dimensions,

$$\left(\boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{t} \tag{3.10}$$

is an orthogonal projection of $\mathbf{t}$ onto the space spanned by the columns of $\boldsymbol{\Phi}$?

3.3

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left(t_n - \mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n) \right)^2 . \tag{3.11}$$

The right hand side can be written as

$$\frac{1}{2} \| \mathbf{t}' - \boldsymbol{\Phi}' \mathbf{w} \|^2 , \tag{3.12}$$

where

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \vdots \\ \sqrt{r_N} t_N \end{bmatrix}, \boldsymbol{\Phi}' = \begin{bmatrix} \sqrt{r_1} \phi(\mathbf{x}_1)^{\mathsf{T}} \\ \vdots \\ \sqrt{r_N} \phi(\mathbf{x}_N)^{\mathsf{T}} \end{bmatrix}.$$

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

$$\mathbf{0} = -\boldsymbol{\Phi}'^{\mathsf{T}} (\mathbf{t}' - \boldsymbol{\Phi}' \mathbf{w}). \tag{3.13}$$

Therefore,

$$\operatorname*{argmin}_{\mathbf{w}} E(\mathbf{w}) = \left(\boldsymbol{\Phi}'^{\mathsf{T}} \boldsymbol{\Phi}' \right)^{-1} \boldsymbol{\Phi}'^{\mathsf{T}} \mathbf{t}'. \tag{3.14}$$

3.4 (Incomplete)

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right)^2, \tag{3.15}$$

where

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{m=1}^{M} w_m (x_m + \epsilon_m), \tag{3.16}$$

$$p(\epsilon_m) = \mathcal{N} \left(\epsilon_m | 0, \sigma^2 \right).$$

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

$$\mathbf{0} = \sum_{n=1}^{N} \begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right).$$

The right hand side can be written as

$$\sum_{n=1}^{N} \begin{bmatrix} 1 \\ \mathbf{x}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right) + \sum_{n=1}^{N} \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right).$$

3.5

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n) \right)^2. \tag{3.17}$$

Then, the minimisation of $E(\mathbf{w})$ under the constraint

$$\sum_{m=1}^{M} |w_m|^q \le \eta \tag{3.18}$$

reduces to the minimisation of

$$E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \tag{3.19}$$

with respect to $\mathbf{w}$ and λ. Then,

$$\eta = \sum_{m=1}^{M} |w_m^*(\lambda)|^q, \tag{3.20}$$

108

where

$$\mathbf{w}^*(\lambda) = \underset{\mathbf{w}}{\operatorname{argmin}} \left(E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \right). \tag{3.21}$$

3.6

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in D dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{W}), \boldsymbol{\Sigma}\right), \tag{3.22}$$

where

$$\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n). \tag{3.23}$$

Then,

$$\ln \left(\prod_{n=1}^{N} p(\mathbf{t}_n | \mathbf{W}, \boldsymbol{\Sigma}) \right)$$

$$= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n)). \tag{3.24}$$

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

$$\mathbf{O} = -\frac{1}{2} \left(\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\mathsf{T} \right) \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n)) (\boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T},$$

$$\mathbf{O} = -\frac{N}{2} (\boldsymbol{\Sigma}^{-1})^\mathsf{T} + \frac{1}{2} (\boldsymbol{\Sigma}^{-1})^2 \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T}. \tag{3.25}$$

Therefore,

$$\mathbf{W}_{\mathrm{ML}} = (\boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\mathsf{T} \mathbf{t},$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n)) (\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n))^\mathsf{T}, \tag{3.26}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\mathsf{T} \end{bmatrix}.$$

3.7

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.27}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.28}$$

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

$$-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)$$
$$= -\frac{\beta}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^\mathsf{T}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0), \tag{3.29}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\mathsf{T} \end{bmatrix}.$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + \frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\mathbf{t}^\mathsf{T}\mathbf{t} - \frac{1}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0, \tag{3.30}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \tag{3.31}$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N). \tag{3.32}$$

3.8

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.33}$$

Then, by 3.7,
$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.34}$$
where
$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^\mathsf{T}\mathbf{t}_N \right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^\mathsf{T}\boldsymbol{\Phi}_N. \end{aligned} \tag{3.35}$$
By the Bayes' theorem,
$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \tag{3.36}$$
Since t_{N+1} and $\mathbf{t}_N$ are independent, it can be written as
$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1})p(\mathbf{t}_N) = p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}). \tag{3.37}$$
By the Bayes' theorem, the right hand side can be written as
$$p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \tag{3.38}$$
Then,
$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(\mathbf{w}|\mathbf{t}_N)p(t_{N+1}|\mathbf{w}). \tag{3.39}$$
The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as
$$\begin{aligned} &-\frac{1}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\mathsf{T} \mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right) - \frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2 \\ =&-\frac{1}{2}\left(\mathbf{w} - \boldsymbol{\mu}_{N+1}\right)^\mathsf{T}\boldsymbol{\Lambda}_{N+1}\left(\mathbf{w} - \boldsymbol{\mu}_{N+1}\right) + \frac{1}{2}\boldsymbol{\mu}_{N+1}^\mathsf{T}\boldsymbol{\Lambda}_{N+1}\boldsymbol{\mu}_{N+1} \\ &-\frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}t_{N+1}^2, \end{aligned} \tag{3.40}$$
where
$$\begin{aligned} \boldsymbol{\mu}_{N+1} &= \boldsymbol{\Lambda}_{N+1}^{-1}\left(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta t_{N+1}\boldsymbol{\phi}(\mathbf{x}_{N+1})\right), \\ \boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^\mathsf{T}. \end{aligned} \tag{3.41}$$
Therefore,
$$\begin{aligned} \boldsymbol{\mu}_{N+1} &= \mathbf{m}_{N+1}, \\ \boldsymbol{\Lambda}_{N+1} &= \mathbf{S}_{N+1}^{-1}. \end{aligned} \tag{3.42}$$
Thus,
$$p(\mathbf{w}|\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}). \tag{3.43}$$

111

3.9   (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.44}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.45}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^\mathsf{T}\mathbf{t}_N\right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^\mathsf{T}\boldsymbol{\Phi}_N. \tag{3.46}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \tag{3.47}$$

Since t_{N+1} and $\mathbf{t}_N$ are independent, it can be written as

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1})p(\mathbf{t}_N) = p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}). \tag{3.48}$$

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

$$p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \tag{3.49}$$

Then,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(\mathbf{w}|\mathbf{t}_N)p(t_{N+1}|\mathbf{w}). \tag{3.50}$$

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

$$-\frac{1}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\mathsf{T}\mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right) - \frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2. \tag{3.51}$$

3.10

Let t be a variable such that

$$p(t|\mathbf{w}) = \mathcal{N}\left(t|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.52}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.53}$$

where

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \mathbf{\Phi}^\intercal \mathbf{t} \right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}^\intercal \mathbf{\Phi}. \qquad (3.54)$$

By marginalisation,

$$p(t|\mathbf{t}) = \int p(t|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \qquad (3.55)$$

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

$$-\frac{\beta}{2} \left(t - \mathbf{w}^\intercal \phi(\mathbf{x}) \right)^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \qquad (3.56)$$

It can be written as

$$-\frac{1}{2} \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix}^\intercal \begin{bmatrix} \mathbf{S}_N^{-1} + \beta\phi(\mathbf{x})\phi(\mathbf{x})^\intercal & -\beta\phi(\mathbf{x}) \\ -\beta\phi(\mathbf{x})^\intercal & \beta \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix} + \begin{bmatrix} \mathbf{w} \\ t \end{bmatrix}^\intercal \begin{bmatrix} \mathbf{S}_N^{-1}\mathbf{m}_N \\ 0 \end{bmatrix} - \frac{1}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \beta\phi(\mathbf{x})\phi(\mathbf{x})^\intercal & -\beta\phi(\mathbf{x}) \\ -\beta\phi(\mathbf{x})^\intercal & \beta \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}_N & \mathbf{S}_N\phi(\mathbf{x}) \\ \phi(\mathbf{x})^\intercal\mathbf{S}_N & \frac{1}{\beta} + \phi(\mathbf{x})^\intercal\mathbf{S}_N\phi(\mathbf{x}) \end{bmatrix}.$$

Therefore,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \beta\phi(\mathbf{x})\phi(\mathbf{x})^\intercal & \beta\phi(\mathbf{x}) \\ \beta\phi(\mathbf{x})^\intercal & \beta \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_N^{-1}\mathbf{m}_N \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{m}_N \\ \mathbf{m}_N^\intercal\phi(\mathbf{x}) \end{bmatrix}.$$

Thus,

$$p(t|\mathbf{t}) = \mathcal{N}\left(t|\mathbf{m}_N^\intercal\phi(\mathbf{x}), \sigma_N^2(\mathbf{x}) \right), \qquad (3.57)$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^\intercal\mathbf{S}_N\phi(\mathbf{x}). \qquad (3.58)$$

3.11

Let t be a variable such that

$$p(t|\mathbf{w}) = \mathcal{N}\left(t|\mathbf{w}^\intercal\phi(\mathbf{x}), \beta^{-1} \right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \qquad (3.59)$$

Then, by 3.7,
$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.60}$$

where
$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}_N^{\mathsf{T}}\mathbf{t}_N\right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}_N^{\mathsf{T}}\mathbf{\Phi}_N.\end{aligned} \tag{3.61}$$

Then, by 3.10,
$$p(t|\mathbf{t}) = \mathcal{N}\left(t \mid \mathbf{m}_N^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right), \tag{3.62}$$

where
$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}). \tag{3.63}$$

Then,
$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}}\left(\mathbf{S}_N - \mathbf{S}_{N+1}\right)\boldsymbol{\phi}(\mathbf{x}). \tag{3.64}$$

By the expression of $\mathbf{S}_N$ above,
$$\mathbf{S}_{N+1} = \left(\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathsf{T}}\right)^{-1}. \tag{3.65}$$

By the identity
$$\left(\mathbf{M} + \mathbf{v}\mathbf{v}^{\mathsf{T}}\right)^{-1} = \mathbf{M}^{-1} - \frac{\left(\mathbf{M}^{-1}\mathbf{v}\right)\left(\mathbf{v}^{\mathsf{T}}\mathbf{M}^{-1}\right)}{1 + \mathbf{v}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{v}}, \tag{3.66}$$

the right hand side can be written as
$$\mathbf{S}_N - \frac{\beta\left(\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)\left(\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathsf{T}}\mathbf{S}_N\right)}{1 + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathsf{T}}}. \tag{3.67}$$

Therefore,
$$\boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}}(\mathbf{S}_N - \mathbf{S}_{N+1})\boldsymbol{\phi}(\mathbf{x}) = \frac{\beta\left(\boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2}{1 + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})^{\mathsf{T}}}. \tag{3.68}$$

Thus,
$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}). \tag{3.69}$$

3.12

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}, \beta) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}, \beta) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0), \tag{3.70}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{w}, \beta|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}, \beta). \tag{3.71}$$

The logarithm of the right hand side except the terms independent of $\mathbf{t}$, $\mathbf{w}$ and β can be written as

$$-\frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{M}{2}\ln\beta^{-1} - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)$$

$$+ (a_0 - 1)\ln\beta - b_0\beta$$

$$= -\frac{M}{2}\ln\beta - \frac{\beta}{2}\mathbf{w}^\mathsf{T}\left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^\mathsf{T}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right) - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0$$

$$+ \left(a_0 + \frac{N}{2} - 1\right)\ln\beta - b_0\beta. \tag{3.72}$$

The right hand side can be written as

$$-\frac{M}{2}\ln\beta - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + (a_N - 1)\ln\beta - b_N\beta, \tag{3.73}$$

where

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right),$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi},$$
$$a_N = a_0 + \frac{N}{2}, \tag{3.74}$$
$$b_N = b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0\mathbf{m}_0 - \frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N.$$

Therefore,

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N, b_N). \tag{3.75}$$

Substituting it to the result of the Bayes' theorem above, we have

$$p(\mathbf{t}) = \frac{\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0)}{\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N, b_N)}. \tag{3.76}$$

The logarithm of the right hand side can be written as

$$
\begin{aligned}
& -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^{\mathsf{T}}(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) \\
& -\frac{M}{2}\ln(2\pi) - \frac{M}{2}\ln\beta^{-1} - \frac{1}{2}\det\mathbf{S}_0 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathsf{T}}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\
& + a_0\ln b_0 - \ln\Gamma(a_0) + (a_0 - 1)\ln\beta - b_0\beta \\
& + \frac{M}{2}\ln(2\pi) + \frac{M}{2}\ln\beta^{-1} + \frac{1}{2}\det\mathbf{S}_N + \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^{\mathsf{T}}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_0) \\
& - a_N\ln b_N + \ln\Gamma(a_N) - (a_N - 1)\ln\beta + b_N\beta \\
= & -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\det\mathbf{S}_0 + a_0\ln b_0 - \ln\Gamma(a_0) + \frac{1}{2}\det\mathbf{S}_N - a_N\ln b_N + \ln\Gamma(a_N).
\end{aligned}
$$
(3.77)

Therefore,

$$
p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}}\left(\frac{\det\mathbf{S}_N}{\det\mathbf{S}_0}\right)^{\frac{1}{2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}.
\tag{3.78}
$$

3.13

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}, \beta) &= \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\phi(\mathbf{x}_n), \beta^{-1}\right), \\
p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0),
\end{aligned}
\tag{3.79}
$$

where $\mathbf{w}$ and ϕ are vectors in M dimensions. Then, by 3.12,

$$
p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N, b_N),
\tag{3.80}
$$

where

$$
\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^{\mathsf{T}}\mathbf{t}\right), \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}, \\
a_N &= a_0 + \frac{N}{2}, \\
b_N &= b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0\mathbf{m}_0 - \frac{1}{2}\mathbf{m}_N^{\mathsf{T}}\mathbf{S}_N^{-1}\mathbf{m}_N.
\end{aligned}
\tag{3.81}
$$

By marginalisation,

$$
p(t|\mathbf{t}) = \int\int p(t|\mathbf{w}, \beta)p(\mathbf{w}, \beta|\mathbf{t})d\mathbf{w}d\beta.
\tag{3.82}
$$

The right hand side can be written as

$$\int \left(\int \mathcal{N}\left(t|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right) d\mathbf{w} \right) \mathrm{Gam}(\beta|a_N, b_N)d\beta.$$
(3.83)

The logarithm of the integrand with respect to $\mathbf{w}$ except the terms indepndent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\left(t - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N).$$
(3.84)

It can be written as

$$-\frac{\beta}{2}\begin{bmatrix}\mathbf{w} \\ t\end{bmatrix}^\intercal \begin{bmatrix} \mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x}) \\ -\boldsymbol{\phi}(\mathbf{x})^\intercal & 1 \end{bmatrix}\begin{bmatrix}\mathbf{w} \\ t\end{bmatrix} + \beta\begin{bmatrix}\mathbf{w} \\ t\end{bmatrix}^\intercal \begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N \\ 0\end{bmatrix} - \frac{\beta}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x}) \\ -\boldsymbol{\phi}(\mathbf{x})^\intercal & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}_N & \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) \\ \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N & 1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}.$$

Then,

$$\begin{bmatrix} \mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x}) \\ -\boldsymbol{\phi}(\mathbf{x})^\intercal & 1 \end{bmatrix}^{-1} \begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N \\ 0\end{bmatrix} = \begin{bmatrix} \mathbf{m}_N \\ \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x}) \end{bmatrix}.$$

Therefore, the integral with respect to $\mathbf{w}$ can be written as

$$\mathcal{N}\left(t|\mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)\right).$$
(3.85)

Then, the logarithm of the integrand with respect to β except the terms independent of β can be written as

$$-\frac{1}{2}\ln \beta^{-1} - \frac{\beta}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}\left(t - \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2 + (a_N - 1)\ln \beta - b_N\beta$$

$$= \left(a_N + \frac{1}{2} - 1\right)\ln \beta - \left(b_N + \frac{\left(t - \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}\right)\beta.$$
(3.86)

Therefore, the integral with respect to β except the terms independent of t can be written as

$$\left(b_N + \frac{\left(t - \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}\right)^{-a_N - \frac{1}{2}}.$$
(3.87)

Thus,
$$p(t|\mathbf{x}, \mathbf{t}) = \mathrm{St}(t|\mu, \lambda, \nu), \tag{3.88}$$

where
$$\begin{aligned}
\mu &= \mathbf{m}_N^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x}), \\
\lambda &= \frac{a_N}{b_N} \left(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)^{-1}, \\
\nu &= 2a_N.
\end{aligned} \tag{3.89}$$

3.14   (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that
$$\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}),
\end{aligned} \tag{3.90}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. Then, by 3.7,
$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.91}$$

where
$$\begin{aligned}
\mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{t}, \\
\mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{\Phi}.
\end{aligned} \tag{3.92}$$

Let
$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x}). \tag{3.93}$$

Then,
$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n, \tag{3.94}$$

where
$$k(\mathbf{x}, \mathbf{x}') = \beta \boldsymbol{\phi}(\mathbf{x})^{\mathsf{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}'). \tag{3.95}$$

Let us suppose that $\phi_j(\mathbf{x})$ are linearly independent, $N > M$ and
$$\phi_0(\mathbf{x}) = 1. \tag{3.96}$$

Then, we can construct a new basis set $\psi_j(\mathbf{x})$ such that
$$\boldsymbol{\Psi}^{\mathsf{T}} \boldsymbol{\Psi} = \mathbf{I}? \tag{3.97}$$

$$\sum_{n=1}^{N} \psi_j(\mathbf{x}_n)\psi_k(\mathbf{x}_n) = I_{jk}? \tag{3.98}$$

where

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\psi}(\mathbf{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\psi}(\mathbf{x}_N)^\mathsf{T} \end{bmatrix}$$

and

$$\psi_0(\mathbf{x}) = 1. \tag{3.99}$$

Under the basis set, if $\alpha = 0$, then

$$\mathbf{S}_N^{-1} = \beta\mathbf{I}, \tag{3.100}$$

so that

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\mathsf{T}\boldsymbol{\psi}(\mathbf{x}'). \tag{3.101}$$

Then,

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^{N}\sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) = 1? \tag{3.102}$$

3.15

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.103}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.104}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \end{aligned} \tag{3.105}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.106}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \tag{3.107}$$

By 3.22, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α and β to zero gives

$$\alpha = \frac{\gamma}{\mathbf{m}_N^\mathsf{T}\mathbf{m}_N},$$
$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2}, \tag{3.108}$$

where

$$\gamma = \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m} \tag{3.109}$$

and $\lambda_1, \cdots, \lambda_M$ are the eigenvalues of $\beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}$. If α and β are set as above, then

$$E(\mathbf{m}_N) = \frac{N}{2}. \tag{3.110}$$

3.16

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{3.111}$$

where $\mathbf{w}$ and ϕ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.112}$$

Integrating both sides with respect to $\mathbf{w}$ gives

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{3.113}$$

The logarithm of the integrand of the right hand side except the terms independent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} = -\frac{1}{2}\begin{bmatrix}\mathbf{w}\\\mathbf{t}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\alpha\mathbf{I} + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi} & -\beta\mathbf{\Phi}^\mathsf{T}\\-\beta\mathbf{\Phi} & \beta\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{w}\\\mathbf{t}\end{bmatrix}.$$

By 2.24,

$$\begin{bmatrix} \alpha\mathbf{I} + \beta\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi} & -\beta\mathbf{\Phi}^\mathsf{T} \\ -\beta\mathbf{\Phi} & \beta\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\mathbf{\Phi}^\mathsf{T} \\ \alpha^{-1}\mathbf{\Phi} & \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^\mathsf{T} + \beta^{-1}\mathbf{I} \end{bmatrix}.$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^\mathsf{T} + \beta^{-1}\mathbf{I}\right). \tag{3.114}$$

3.17

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.115}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.116}$$

Then,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{3.117}$$

The logarithm of the integrand of the right hand side can be written as

$$-\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left(\alpha^{-1}\mathbf{I}\right)\right) - \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}. \tag{3.118}$$

Therefore,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int \exp\left(-E(\mathbf{w})\right)d\mathbf{w}, \tag{3.119}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}. \tag{3.120}$$

3.18

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.121}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.122}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\mathsf{T} \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi}. \end{aligned} \tag{3.123}$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left(-E(\mathbf{w})\right) d\mathbf{w}, \tag{3.124}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}. \tag{3.125}$$

The first term of the definition of $E(\mathbf{w})$ can be written as

$$\begin{aligned} &\frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N - \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N)\|^2 \\ =&\frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 - \beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^\mathsf{T}\boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N). \end{aligned} \tag{3.126}$$

Similarly, the second term can be written as

$$\begin{aligned} &\frac{\alpha}{2}(\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N) \\ =&\frac{\alpha}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) + \alpha\mathbf{m}_N^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \end{aligned} \tag{3.127}$$

Here,

$$\begin{aligned} &- \beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^\mathsf{T}\boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \alpha\mathbf{m}_N^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) \\ =&(-\beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\mathbf{m}_N + \alpha\mathbf{m}_N)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N). \end{aligned} \tag{3.128}$$

By the definitions of $\mathbf{m}_N$ and $\mathbf{S}_N$ above, the right hand can be written as

$$\left(-\beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t} + \mathbf{S}_N^{-1}\mathbf{m}_N\right)^\mathsf{T}(\mathbf{w} - \mathbf{m}_N) = 0. \tag{3.129}$$

Therefore,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.130}$$

3.19

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{3.131}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.132}$$

where

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\ \mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \tag{3.133}$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left(-E(\mathbf{w})\right) d\mathbf{w}, \tag{3.134}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}. \tag{3.135}$$

By 3.18,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.136}$$

Therefore, the integral in the expression above of $p(\mathbf{t})$ can be written as

$$\exp\left(-E(\mathbf{m}_N)\right) \int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N)\right) d\mathbf{w} \\ = (2\pi)^{\frac{M}{2}} \left(\det \mathbf{S}_N\right)^{\frac{1}{2}} \exp\left(-E(\mathbf{m}_N)\right). \tag{3.137}$$

Thus,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathbf{S}_N) - E(\mathbf{m}_N). \tag{3.138}$$

3.20

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{3.139}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \qquad (3.140)$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \end{aligned} \qquad (3.141)$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \quad (3.142)$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \qquad (3.143)$$

Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ be eigenvectors of $\beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ such that

$$\beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\mathbf{u}_m = \lambda_m\mathbf{u}_m. \qquad (3.144)$$

Then,

$$\mathbf{S}_N^{-1}\mathbf{u}_m = (\alpha + \lambda_m)\mathbf{u}_m, \qquad (3.145)$$

so that

$$\det\mathbf{S}_N = \prod_{m=1}^{M} \frac{1}{\alpha + \lambda_m}. \qquad (3.146)$$

Therefore, setting the derivative of $\ln p(\mathbf{t}|\alpha, \beta)$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\sum_{m=1}^{M}\frac{1}{\alpha + \lambda_m} - \frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \qquad (3.147)$$

Multiplying both sides by 2α gives

$$\alpha\mathbf{m}_N^\mathsf{T}\mathbf{m}_N = M - \sum_{m=1}^{M}\frac{\alpha}{\alpha + \lambda_m}. \qquad (3.148)$$

The right hand side can be written as

$$\sum_{m=1}^{M}\left(1 - \frac{\alpha}{\alpha + \lambda_m}\right) = \sum_{m=1}^{M}\frac{\lambda_i}{\alpha + \lambda_m}. \qquad (3.149)$$

124

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^\mathsf{T}\mathbf{m}_N}, \tag{3.150}$$

where

$$\gamma = \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m}. \tag{3.151}$$

3.21

(a)

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

$$\boldsymbol{\Sigma}\mathbf{u}_m = \lambda_m \mathbf{u}_m, \tag{3.152}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_M$ are unit vectors. Let

$$\begin{aligned}\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \cdots, \lambda_M), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_M].\end{aligned} \tag{3.153}$$

By 2.19,

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\mathsf{T}, \\ \mathbf{U}^\mathsf{T}\mathbf{U} &= \mathbf{I}.\end{aligned} \tag{3.154}$$

Therefore,

$$\det \boldsymbol{\Sigma} = \prod_{m=1}^{M} \lambda_m, \tag{3.155}$$

so that

$$\ln(\det \boldsymbol{\Sigma}) = \sum_{m=1}^{M} \ln \lambda_i. \tag{3.156}$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \boldsymbol{\Sigma}) = \sum_{m=1}^{M} \frac{\partial \lambda_m}{\partial \alpha} \frac{1}{\lambda_m}. \tag{3.157}$$

Therefore,

$$\frac{\partial}{\partial \alpha} \ln(\det \boldsymbol{\Sigma}) = \mathrm{tr}\left(\boldsymbol{\Lambda}^{-1}\frac{\partial \boldsymbol{\Lambda}}{\partial \alpha}\right). \tag{3.158}$$

The right hand side can be written as

$$\mathrm{tr}\left(\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\mathsf{T}}\frac{\partial \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}}{\partial \alpha}\right) = \mathrm{tr}\left(\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \alpha}\right). \tag{3.159}$$

Therefore,

$$\frac{\partial}{\partial \alpha}\ln(\det \mathbf{\Sigma}) = \mathrm{tr}\left(\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \alpha}\right). \tag{3.160}$$

(b)

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.161}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.162}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathsf{T}}\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}. \end{aligned} \tag{3.163}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln \alpha + \frac{N}{2}\ln \beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.164}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathsf{T}}\mathbf{m}_N. \tag{3.165}$$

By 3.21(a),

$$\frac{\partial}{\partial \alpha}\ln\left(\det \mathbf{S}_N^{-1}\right) = \mathrm{tr}\left(\mathbf{S}_N\right). \tag{3.166}$$

The right hand side can be written as

$$\sum_{m=1}^{M}\frac{1}{\alpha + \lambda_m}, \tag{3.167}$$

where $\lambda_1, \cdots, \lambda_M$ are eigenvalues of $\beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\sum_{m=1}^{M}\frac{1}{\alpha + \lambda_m} - \frac{1}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N, \tag{3.168}$$

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^\mathsf{T}\mathbf{m}_N}, \tag{3.169}$$

where

$$\gamma = \sum_{m=1}^{M}\frac{\lambda_m}{\alpha + \lambda_m}. \tag{3.170}$$

3.22

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{3.171}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.172}$$

where

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\mathbf{t},$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}. \tag{3.173}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.174}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\mathsf{T}\mathbf{m}_N. \tag{3.175}$$

By 3.21(a),

$$\frac{\partial}{\partial\beta}\ln\left(\det\mathbf{S}_N^{-1}\right) = \mathrm{tr}\left(\mathbf{S}_N\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right). \tag{3.176}$$

127

Since
$$\mathbf{S}_N \mathbf{\Phi}^\mathsf{T} \mathbf{\Phi} = \frac{1}{\beta} \left(\mathbf{I} - \alpha \mathbf{S}_N \right), \tag{3.177}$$

the right hand side can be written as

$$\frac{1}{\beta} \left(M - \alpha \sum_{m=1}^{M} \frac{1}{\alpha + \lambda_m} \right) = \frac{1}{\beta} \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m}, \tag{3.178}$$

where $\lambda_1, \cdots, \lambda_M$ are eigenvalues of $\beta \mathbf{\Phi}^\mathsf{T} \mathbf{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to β to zero gives

$$0 = \frac{N}{2\beta} - \frac{1}{2\beta} \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m} - \frac{1}{2} \|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2. \tag{3.179}$$

Thus,

$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2}, \tag{3.180}$$

where

$$\gamma = \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m}. \tag{3.181}$$

3.23

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned}
p(t_n | \mathbf{w}, \beta) &= \mathcal{N} \left(t_n | \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1} \right), \\
p(\mathbf{w}, \beta) &= \mathcal{N} \left(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0 \right) \mathrm{Gam}(\beta | a_0, b_0),
\end{aligned} \tag{3.182}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By marginalisation,

$$p(\mathbf{t}) = \int \int p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta) d\mathbf{w} d\beta. \tag{3.183}$$

The right hand side can be written as

$$\int \left(\int \left(\prod_{n=1}^{N} \mathcal{N} \left(t_n | \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1} \right) \right) \mathcal{N} \left(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0 \right) d\mathbf{w} \right) \mathrm{Gam}(\beta | a_0, b_0) d\beta. \tag{3.184}$$

The logarithm of the integrand with respect to $\mathbf{w}$ can be written as

$$
-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n))^2
$$

$$
-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\beta^{-1}\mathbf{S}_0) - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^\intercal\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)
$$

$$
= -\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0)
$$

$$
-\frac{\beta}{2}\mathbf{w}^\intercal\left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^\intercal\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\intercal\mathbf{t}\right) - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0.
$$

$$(3.185)$$

The right hand side can be written as

$$
-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0)
$$

$$
-\frac{\beta}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\intercal\mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right) + \frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0,
$$

$$(3.186)$$

where

$$
\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\intercal\mathbf{t}\right),
$$
$$
\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}.
$$

$$(3.187)$$

Therefore, the logarithm of the integral with respect to $\mathbf{w}$ can be written as

$$
-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}_N)
$$

$$
+\frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0.
$$

$$(3.188)$$

Then, the logarithm of the integrand with respect to β can be written as

$$
-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}_N)
$$

$$
+\frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0
$$

$$
-\ln\Gamma(a_0) + a_0\ln b_0 + (a_0 - 1)\ln\beta - b_0\beta
$$

$$
= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}_N)
$$

$$
-\ln\Gamma(a_0) + a_0\ln b_0 + (a_N - 1)\ln\beta - b_N\beta,
$$

$$(3.189)$$

where

$$a_N = a_0 + \frac{N}{2},$$

$$b_N = b_0 + \frac{\beta}{2}\|\mathbf{t}\|^2 + \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{\beta}{2}\mathbf{m}_N^\mathsf{T}\mathbf{S}_N^{-1}\mathbf{m}_N. \tag{3.190}$$

Therefore, the logarithm of the integral with respect to β can be written as

$$-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}_N) - \ln\Gamma(a_0) + a_0\ln b_0 + \ln\Gamma(a_N) - a_N\ln b_N. \tag{3.191}$$

Thus,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}}\left(\frac{\det \mathbf{S}_N}{\det \mathbf{S}_0}\right)^{\frac{1}{2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0{}^{a_0}}{b_N{}^{a_N}}. \tag{3.192}$$

3.24

Refer to 3.12.

4 Linear Models for Classification

4.1

Let $x_1, \cdots, x_M$ and $y_1, \cdots, y_N$ 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

$$\mathbf{x} = \sum_{m=1}^{M} \alpha_m \mathbf{x}_m,$$
$$\mathbf{y} = \sum_{n=1}^{N} \beta_n \mathbf{y}_n, \tag{4.1}$$

where

$$\sum_{m=1}^{M} \alpha_m = \sum_{n=1}^{N} \beta_n = 1,$$
$$\alpha_m \geq 0, \beta_n \geq 0. \tag{4.2}$$

Let us assume that $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ below are subject to the constraints above.

If the convex hulls intersect, then there exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$\sum_{m=1}^{M} \alpha_m \mathbf{x}_m = \sum_{n=1}^{N} \beta_n \mathbf{y}_n. \tag{4.3}$$

Then,

$$\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^\mathsf{T} \mathbf{x}_m + w_0 \right) = \hat{\mathbf{w}}^\mathsf{T} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0 \sum_{m} \alpha_m, \tag{4.4}$$

for any $\hat{\mathbf{w}}$ and w_0. The right hand side can be written as

$$\hat{\mathbf{w}}^\mathsf{T} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n = \sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^\mathsf{T} \mathbf{y}_n + w_0 \right). \tag{4.5}$$

Therefore, there do not exist $\hat{\mathbf{w}}$ and w_0 such that

$$\hat{\mathbf{w}}^\mathsf{T} \mathbf{x}_m + w_0 > 0,$$
$$\hat{\mathbf{w}}^\mathsf{T} \mathbf{y}_n + w_0 < 0. \tag{4.6}$$

Conversely, if there exist $\hat{\mathbf{w}}$ and w_0 such that

$$\begin{aligned}
\hat{\mathbf{w}}^\mathsf{T}\mathbf{x}_m + w_0 &> 0, \\
\hat{\mathbf{w}}^\mathsf{T}\mathbf{y}_n + w_0 &< 0,
\end{aligned} \tag{4.7}$$

then

$$\begin{aligned}
\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^\mathsf{T}\mathbf{x}_m + w_0\right) &> 0, \\
\sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^\mathsf{T}\mathbf{y}_n + w_0\right) &< 0.
\end{aligned} \tag{4.8}$$

The left hand sides can be written as

$$\begin{aligned}
\hat{\mathbf{w}}^\mathsf{T}\sum_{m=1}^{M} \alpha_m\mathbf{x}_m + w_0 \sum_{m=1}^{M} \alpha_m &= \hat{\mathbf{w}}^\mathsf{T}\sum_{m=1}^{M} \alpha_m\mathbf{x}_m + w_0, \\
\hat{\mathbf{w}}^\mathsf{T}\sum_{n=1}^{N} \beta_n\mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n &= \hat{\mathbf{w}}^\mathsf{T}\sum_{n=1}^{N} \beta_n\mathbf{y}_n + w_0.
\end{aligned} \tag{4.9}$$

Therefore, there do not exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$\sum_{m=1}^{M} \alpha_m\mathbf{x}_m = \sum_{n=1}^{N} \beta_n\mathbf{y}_n. \tag{4.10}$$

Thus, the convex hulls do not intersect.

4.2   (Incomplete)

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ and $\mathbf{w}_1, \cdots, \mathbf{w}_K$ are variables in M dimensions and $\mathbf{t}_1, \cdots, \mathbf{t}_N$ are ones in K dimensions. Let

$$E(\tilde{\mathbf{W}}) = \frac{1}{2} \operatorname{tr}\left((\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^\mathsf{T}(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})\right), \tag{4.11}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^\mathsf{T} \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\mathsf{T} \end{bmatrix},$$

$$\tilde{\mathbf{W}} = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^\intercal \\ \vdots \\ \mathbf{t}_N^\intercal \end{bmatrix}.$$

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

$$\mathbf{O} = \tilde{\mathbf{X}}^\intercal(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}). \tag{4.12}$$

Therefore,

$$\operatorname*{argmin}_{\tilde{\mathbf{W}}} E(\tilde{\mathbf{W}}) = \left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\intercal\mathbf{T}. \tag{4.13}$$

Let $\tilde{\mathbf{W}}^*$ denote the least-square solution above. Then,

$$(\tilde{\mathbf{W}}^*)^\intercal\tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{T}^\intercal\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}} - \mathbf{t}_n, \tag{4.14}$$

where $\tilde{\mathbf{x}}$ is a vector in $M + 1$ dimensions whose first element is 1. The right hand side can be written as

$$\mathbf{T}^\intercal\left(\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}} - \mathbf{v}_n\right) = \mathbf{0}? \tag{4.15}$$

where $\mathbf{v}_n$ is a vector in N dimensions whose n th element is 1 and other elements are zero. Therefore,

$$(\tilde{\mathbf{W}}^*)^\intercal\tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{0}. \tag{4.16}$$

Thus, if

$$\mathbf{a}^\intercal\mathbf{t}_n + b = 0, \tag{4.17}$$

then

$$\mathbf{a}^\intercal(\tilde{\mathbf{W}}^*)^\intercal\tilde{\mathbf{x}} + b = 0. \tag{4.18}$$

4.3   (Incomplete)

4.4

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n, \tag{4.19}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Setting the derivatives of

$$\mathbf{w}^{\mathsf{T}}(\mathbf{m}_2 - \mathbf{m}_1) + \lambda \left(\|\mathbf{w}\|^2 - 1 \right) \tag{4.20}$$

with respect to $\mathbf{w}$ and λ to zero gives

$$\mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} = \mathbf{0},$$
$$\|\mathbf{w}\|^2 - 1 = 0. \tag{4.21}$$

Therefore, $\mathbf{w}^{\mathsf{T}}(\mathbf{m}_2 - \mathbf{m}_1)$ under the constratint

$$\|\mathbf{w}\|^2 = 1 \tag{4.22}$$

is maximised if

$$\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1. \tag{4.23}$$

4.5

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \tag{4.24}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{4.25}$$

where

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2,$$
$$y_n = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n, \tag{4.26}$$
$$m_k = \mathbf{w}^{\mathsf{T}}\mathbf{m}_k.$$

Then, $J(\mathbf{w})$ can be written as

$$\frac{\left(\mathbf{w}^{\mathsf{T}}(\mathbf{m}_2 - \mathbf{m}_1)\right)^2}{\sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^{\mathsf{T}}(\mathbf{x}_n - \mathbf{m}_1)\right)^2 + \sum_{n \in \mathcal{C}_2} \left(\mathbf{w}^{\mathsf{T}}(\mathbf{x}_n - \mathbf{m}_2)\right)^2} = \frac{\mathbf{w}^{\mathsf{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathsf{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}, \tag{4.27}$$

where

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathsf{T}},$$
$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathsf{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathsf{T}}. \tag{4.28}$$

4.6

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \tag{4.29}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{4.30}$$

where

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2,$$

$$y_n = \mathbf{w}^\mathsf{T} \mathbf{x}_n, \tag{4.31}$$

$$m_k = \mathbf{w}^\mathsf{T} \mathbf{m}_k.$$

Then, by 4.5,

$$J(\mathbf{w}) = \frac{\mathbf{w}^\mathsf{T} \mathbf{S}_\mathrm{B} \mathbf{w}}{\mathbf{w}^\mathsf{T} \mathbf{S}_\mathrm{W} \mathbf{w}}, \tag{4.32}$$

where

$$\mathbf{S}_\mathrm{B} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\mathsf{T},$$

$$\mathbf{S}_\mathrm{W} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathsf{T} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathsf{T}. \tag{4.33}$$

Let

$$E = \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n + w_0 - t_n \right)^2, \tag{4.34}$$

where

$$t_n = \begin{cases} \frac{N}{N_1}, & n \in \mathcal{C}_1, \\ -\frac{N}{N_2}, & n \in \mathcal{C}_2. \end{cases} \tag{4.35}$$

Setting the derivative with respect to $\mathbf{w}$ and w_0 gives

$$0 = \sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n + w_0 - t_n \right),$$

$$\mathbf{0} = \sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n + w_0 - t_n \right) \mathbf{x}_n. \tag{4.36}$$

135

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

$$\mathbf{w}^\mathsf{T} \sum_{n=1}^{N} \mathbf{x}_n + N w_0 - \sum_{n=1}^{N} t_n = N \left(\mathbf{w}^\mathsf{T} \mathbf{m} + w_0 \right), \tag{4.37}$$

where

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n. \tag{4.38}$$

Therefore,

$$w_0 = -\mathbf{w}^\mathsf{T} \mathbf{m}. \tag{4.39}$$

Then, the right hand side of the second equation above can be written as

$$\begin{aligned}
&\sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \left(\mathbf{x}_n - \mathbf{m} \right) - t_n \right) \mathbf{x}_n \\
&= \sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^\mathsf{T} (\mathbf{x}_n - \mathbf{m}) - \frac{N}{N_1} \right) \mathbf{x}_n + \sum_{n \in \mathcal{C}_2} \left(\mathbf{w}^\mathsf{T} (\mathbf{x}_n - \mathbf{m}) + \frac{N}{N_2} \right) \mathbf{x}_n.
\end{aligned} \tag{4.40}$$

Since

$$\mathbf{m} = \frac{N_1}{N} \mathbf{m}_1 + \frac{N_2}{N} \mathbf{m}_2,$$

$$\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1) = \mathbf{0}, \tag{4.41}$$

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

$$\begin{aligned}
&\sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^\mathsf{T} \left(\mathbf{x}_n - \mathbf{m}_1 + \frac{N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \right) - \frac{N}{N_1} \right) (\mathbf{x}_n - \mathbf{m}_1 + \mathbf{m}_1) \\
&= \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathsf{T} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^\mathsf{T} \mathbf{w} - N \mathbf{m}_1.
\end{aligned} \tag{4.42}$$

Similarly, the second term can be written as

$$\left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathsf{T} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^\mathsf{T} \mathbf{w} - N \mathbf{m}_2. \tag{4.43}$$

136

Therefore,

$$
\mathbf{0} = \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathsf{T}} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2)\mathbf{m}_1^{\mathsf{T}}\mathbf{w} - N\mathbf{m}_1
$$
$$
+ \left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathsf{T}} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1)\mathbf{m}_2^{\mathsf{T}}\mathbf{w} - N\mathbf{m}_2.
$$

$$(4.44)$$

Thus,

$$
\left(\mathbf{S}_{\mathrm{W}} + \frac{N_1 N_2}{N}\mathbf{S}_{\mathrm{B}} \right) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2). \tag{4.45}
$$

4.7

Let

$$
\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{4.46}
$$

Then,

$$
\sigma(-a) = \frac{1}{1 + \exp(a)}. \tag{4.47}
$$

The right hand side can be written as

$$
1 - \frac{\exp(a)}{1 + \exp(a)} = 1 - \frac{1}{1 + \exp(-a)}. \tag{4.48}
$$

Therefore,

$$
\sigma(-a) = 1 - \sigma(a). \tag{4.49}
$$

Additionally,

$$
\exp(-a) = \frac{1}{\sigma(a)} - 1. \tag{4.50}
$$

Then,

$$
a = -\ln\left(\frac{1}{\sigma(a)} - 1 \right). \tag{4.51}
$$

Therefore,

$$
\sigma^{-1}(y) = \ln\left(\frac{y}{1 - y} \right). \tag{4.52}
$$

4.8

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

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right), \tag{4.53}$$

where

$$p(\mathcal{C}_1) + p(\mathcal{C}_2) = 1. \tag{4.54}$$

By the Bayes' theorem,

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}. \tag{4.55}$$

The right hand side can be written as

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{4.56}$$

where

$$a = \ln\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right). \tag{4.57}$$

Substituting the expressions above of $p(\mathbf{x}|\mathcal{C}_k)$, we have

$$\begin{aligned}
a = &-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \ln p(\mathcal{C}_1) \\
&+ \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \ln p(\mathcal{C}_2).
\end{aligned} \tag{4.58}$$

Therefore,

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma\left(\mathbf{w}^\mathsf{T}\mathbf{x} + w_0\right), \tag{4.59}$$

where

$$\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2).
\end{aligned} \tag{4.60}$$

4.9

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^{K}\left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}, \tag{4.61}$$

where
$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1. \tag{4.62}$$

Then,
$$p(\mathbf{T}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k) \right)^{t_{nk}}. \tag{4.63}$$

If
$$p(\mathcal{C}_k) = \pi_k, \tag{4.64}$$

then, by the Bayes' theorem,
$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \pi_k + \ln p(\boldsymbol{\phi}_n | \mathcal{C}_k) \right). \tag{4.65}$$

Setting the derivatives of
$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{4.66}$$

with respect to π_k and λ to zero gives
$$\begin{aligned} 0 &= \frac{1}{\pi_k} \sum_{n=1}^{N} t_{nk} + \lambda, \\ 0 &= \sum_{k=1}^{K} \pi_k - 1. \end{aligned} \tag{4.67}$$

Then,
$$\lambda = - \sum_{k=1}^{K} \sum_{n=1}^{N} t_{nk}. \tag{4.68}$$

The right hand side can be written as $-N$. Therefore, the maximum likelihood solution for π_k is given by
$$\pi_k = \frac{N_k}{N}, \tag{4.69}$$

where
$$N_k = \sum_{n=1}^{N} t_{nk}. \tag{4.70}$$

4.10

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}, \tag{4.71}$$

where

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1. \tag{4.72}$$

Then,

$$p(\mathbf{T}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}. \tag{4.73}$$

If

$$p(\boldsymbol{\phi}_n | \mathcal{C}_k) = \mathcal{N}(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \tag{4.74}$$

then, by the Bayes' theorem,

$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \mathcal{N}(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) + \ln p(\mathcal{C}_k)\right). \tag{4.75}$$

The right hand side can be written as

$$\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) + \ln p(\mathcal{C}_k)\right). \tag{4.76}$$

By 3.21(a), setting the derivatives of $\ln p(\mathbf{T}, \boldsymbol{\Phi})$ with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ to zero gives

$$\mathbf{0} = \frac{1}{2}\sum_{n=1}^{N} t_{nk} \left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k),$$

$$\mathbf{O} = -\frac{1}{2}\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}} - \left(\boldsymbol{\Sigma}^{-1}\right)^2 (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}\right). \tag{4.77}$$

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

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \boldsymbol{\phi}_n,$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{k=1}^{K} N_k \mathbf{S}_k,$$

(4.78)

where

$$N_k = \sum_{n=1}^{N} t_{nk},$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathsf{T}}.$$

(4.79)

4.11

Let $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_M$ be variables such that

$$p(\boldsymbol{\phi}_m | \mathcal{C}_k) = \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}},$$

(4.80)

where

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1.$$

(4.81)

Then,

$$p(\boldsymbol{\Phi} | \mathcal{C}_k) = \prod_{m=1}^{M} \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}}.$$

(4.82)

By the Bayes' theorem,

$$p(\mathcal{C}_k | \boldsymbol{\Phi}) = \frac{p(\boldsymbol{\Phi} | \mathcal{C}_k) p(\mathcal{C}_k)}{\sum_{k=1}^{K} p(\boldsymbol{\Phi} | \mathcal{C}_k) p(\mathcal{C}_k)}.$$

(4.83)

Therefore,

$$p(\mathcal{C}_k | \boldsymbol{\Phi}) = \frac{\exp\left(a_k(\boldsymbol{\Phi})\right)}{\sum_{k=1}^{K} \exp\left(a_k(\boldsymbol{\Phi})\right)},$$

(4.84)

where

$$a_k(\boldsymbol{\Phi}) = \left(\sum_{m=1}^{M} \sum_{l=1}^{L} \phi_{ml} \ln \mu_{kml}\right) + \ln p(\mathcal{C}_k).$$

(4.85)

4.12

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{4.86}$$

Then,

$$\frac{d\sigma(a)}{da} = \frac{\exp(-a)}{(1 + \exp(-a))^2}. \tag{4.87}$$

The right hand side can be written as

$$\frac{1}{1 + \exp(-a)} - \frac{1}{(1 + \exp(-a))^2} = \sigma(a) - (\sigma(a))^2. \tag{4.88}$$

Therefore,

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a)). \tag{4.89}$$

4.13

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n}(1 - y_n)^{1-t_n}, \end{aligned} \tag{4.90}$$

where

$$\begin{aligned} y_n &= \sigma\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n\right), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \tag{4.91}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{4.92}$$

The right hand side can be written as

$$-\ln\left(\prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}\right) = -\sum_{n=1}^{N}(t_n \ln y_n + (1 - t_n)\ln(1 - y_n)). \tag{4.93}$$

Then, by 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{t_n}{y_n}y_n(1 - y_n)\boldsymbol{\phi}_n - \frac{1 - t_n}{1 - y_n}y_n(1 - y_n)\boldsymbol{\phi}_n\right). \tag{4.94}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} (t_n(1 - y_n)\boldsymbol{\phi}_n - (1 - t_n)y_n\boldsymbol{\phi}_n) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \tag{4.95}$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \tag{4.96}$$

4.14

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n}(1 - y_n)^{1-t_n}, \end{aligned} \tag{4.97}$$

where

$$\begin{aligned} y_n &= \sigma\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n\right), \\ \sigma(a) &= \frac{1}{1 + \exp(-a)}. \end{aligned} \tag{4.98}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{4.99}$$

By 4.13, setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \tag{4.100}$$

If $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N$ are linearly independent, then

$$y_n = t_n. \tag{4.101}$$

Then,

$$\sigma\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n\right) = \begin{cases} 1, & t_n = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.102}$$

Threrefore,

$$\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n = \begin{cases} \infty, & t_n = 1, \\ -\infty, & \text{otherwise.} \end{cases} \tag{4.103}$$

4.15

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
t_n &\in \{0, 1\}, \\
p(t_n|\mathbf{w}) &= y_n^{t_n}(1 - y_n)^{1-t_n},
\end{aligned}
\tag{4.104}
$$

where

$$
\begin{aligned}
y_n &= \sigma\left(\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\right), \\
\sigma(a) &= \frac{1}{1 + \exp(-a)}.
\end{aligned}
\tag{4.105}
$$

Let

$$
E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}).
\tag{4.106}
$$

By 4.13,

$$
\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n.
\tag{4.107}
$$

Then, by 4.12,

$$
\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\mathsf{T}.
\tag{4.108}
$$

The right hand side can be written as

$$
\mathbf{H} = \boldsymbol{\Phi}^\mathsf{T}\mathbf{R}\boldsymbol{\Phi},
\tag{4.109}
$$

where

$$
R_{nn'} = \begin{cases} y_n(1 - y_n), & n = n', \\ 0, & \text{otherwise.} \end{cases}
\tag{4.110}
$$

Then,

$$
\mathbf{u}^\mathsf{T}\mathbf{H}\mathbf{u} = (\boldsymbol{\Phi}\mathbf{u})^\mathsf{T}\mathbf{R}\,(\boldsymbol{\Phi}\mathbf{u}).
\tag{4.111}
$$

Since

$$
y_n(1 - y_n) > 0,
\tag{4.112}
$$

we have

$$
\mathbf{u}^\mathsf{T}\mathbf{H}\mathbf{u} > 0.
\tag{4.113}
$$

Therefore, $\mathbf{H}$ is positive definite. Thus, E is a convex function of $\mathbf{w}$ and it has a unique minimum.

144

4.16

Let $t_1, \cdots, t_N$ be variables such that

$$t_n \in \{0, 1\},$$
$$p(t_n = 1 | \boldsymbol{\phi}_n) = \pi_n. \tag{4.114}$$

Then,

$$p(t_n | \boldsymbol{\phi}_n) = \pi_n^{t_n} (1 - \pi_n)^{1-t_n}. \tag{4.115}$$

Therefore,

$$p(\mathbf{t} | \boldsymbol{\Phi}) = \prod_{n=1}^{N} \pi_n^{t_n} (1 - \pi_n)^{1-t_n}. \tag{4.116}$$

Thus,

$$- \ln p(\mathbf{t} | \boldsymbol{\Phi}) = -\sum_{n=1}^{N} \left(t_n \ln \pi_n + (1 - t_n) \ln(1 - \pi_n) \right). \tag{4.117}$$

4.17

Let

$$y_k = \frac{\exp(a_k)}{\sum_{k=1}^{K} \exp(a_k)}. \tag{4.118}$$

Then,

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k)}{\sum_{k=1}^{K} \exp(a_k)} - \frac{\exp(2a_k)}{\left(\sum_{k=1}^{K} \exp(a_k) \right)^2}. \tag{4.119}$$

The right hand side can be written as $y_k(1 - y_k)$. If $k \neq k'$, then

$$\frac{\partial y_k}{\partial a_{k'}} = -\frac{\exp(a_k + a_{k'})}{\left(\sum_{k=1}^{K} \exp(a_k) \right)^2}. \tag{4.120}$$

The right hand side can be written as $-y_k y_{k'}$. Therefore,

$$\frac{\partial y_k}{\partial a_{k'}} = y_k(I_{kk'} - y_{k'}). \tag{4.121}$$

145

4.18

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables such that

$$t_{nk} \in \{0, 1\},$$

$$p(\mathbf{t}_n|\mathbf{W}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{4.122}$$

where

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_{k=1}^{K} \exp(a_{nk})}, \tag{4.123}$$

$$a_{nk} = \mathbf{w}_k^\intercal \boldsymbol{\phi}_n.$$

Then,

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.124}$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{W}). \tag{4.125}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{4.126}$$

Then, by 4.17,

$$\nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk}(I_{kk'} - y_{nk'})\frac{t_{nk}}{y_{nk}}\boldsymbol{\phi}_n. \tag{4.127}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} \left(\sum_{k=1}^{K} (I_{kk'} - y_{nk'})t_{nk} \right) \boldsymbol{\phi}_n = -\sum_{n=1}^{N} (t_{nk'} - y_{nk'})\boldsymbol{\phi}_n. \tag{4.128}$$

Therefore,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^{N} (y_{nk} - t_{nk}) \boldsymbol{\phi}_n. \tag{4.129}$$

146

4.19

Let $t_1, \cdots, t_N$ 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

$$
\begin{aligned}
\Phi(a) &= \int_{-\infty}^{a} \mathcal{N}(\theta | 0, 1) d\theta, \\
a_n &= \mathbf{w}^{\mathsf{T}} \boldsymbol{\phi}_n.
\end{aligned}
\tag{4.131}
$$

Then,

$$
p(t_n | \boldsymbol{\phi}_n) = (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1 - t_n}.
\tag{4.132}
$$

Therefore,

$$
p(\mathbf{t} | \boldsymbol{\Phi}) = \prod_{n=1}^{N} (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1 - t_n}.
\tag{4.133}
$$

Let

$$
E(\mathbf{w}) = -\ln p(\mathbf{t} | \boldsymbol{\phi}).
\tag{4.134}
$$

The right hand side can be written as

$$
-\sum_{n=1}^{N} \left(t_n \ln \Phi(a_n) + (1 - t_n) \ln (1 - \Phi(a_n)) \right).
\tag{4.135}
$$

Then,

$$
\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \frac{\mathcal{N}(a_n | 0, 1)}{\Phi(a_n)} - (1 - t_n) \frac{\mathcal{N}(a_n | 0, 1)}{1 - \Phi(a_n)} \right) \boldsymbol{\phi}_n.
\tag{4.136}
$$

The right hand side can be written as

$$
\begin{aligned}
&-\sum_{n=1}^{N} \left(\frac{t_n}{\Phi(a_n)} - \frac{1 - t_n}{1 - \Phi(a_n)} \right) \mathcal{N}(a_n | 0, 1) \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} \frac{\mathcal{N}(a_n | 0, 1)}{\Phi(a_n) (1 - \Phi(a_n))} (\Phi(a_n) - t_n) \boldsymbol{\phi}_n.
\end{aligned}
\tag{4.137}
$$

147

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n. \qquad (4.138)$$

Then,

$$
\begin{aligned}
\nabla\nabla E(\mathbf{w}) = & \sum_{n=1}^{N} \frac{-a_n\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\intercal} \\
& - \sum_{n=1}^{N} \frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\left(\Phi(a_n)\right)^2\left(1-\Phi(a_n)\right)} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\intercal} \\
& + \sum_{n=1}^{N} \frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\Phi(a_n)\left(1-\Phi(a_n)\right)^2} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\intercal} \\
& + \sum_{n=1}^{N} \frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\Phi(a_n)\left(1-\Phi(a_n)\right)} \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\intercal}.
\end{aligned} \qquad (4.139)
$$

Therefore,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} b_n \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\intercal}, \qquad (4.140)$$

where

$$
\begin{aligned}
b_n = & \left(\frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\right)^2 \left(\left(\Phi(a_n)\right)^2 - 2t_n\Phi(a_n) + t_n\right) \\
& - \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)} a_n \left(\Phi(a_n) - t_n\right).
\end{aligned} \qquad (4.141)
$$

4.20   (Incomplete)

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables such that

$$
\begin{aligned}
t_{nk} &\in \{0,1\}, \\
p(\mathbf{t}_n|\mathbf{W}) &= \prod_{k=1}^{K} y_{nk}^{t_{nk}},
\end{aligned} \qquad (4.142)
$$

148

where

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_{k=1}^{K} \exp(a_{nk})}, \tag{4.143}$$
$$a_{nk} = \mathbf{w}_k^\intercal \boldsymbol{\phi}_n.$$

Then,

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.144}$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{W}). \tag{4.145}$$

By 4.18,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^{N} (y_{nk} - t_{nk}) \, \boldsymbol{\phi}_n. \tag{4.146}$$

Additionally, by 4.17,

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = \sum_{n=1}^{N} y_{nk}(I_{kk'} - y_{nk'}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\intercal. \tag{4.147}$$

The right hand side can be written as

$$\mathbf{H}_{kk'} = \boldsymbol{\Phi}^\intercal \mathbf{R}_{kk'} \boldsymbol{\Phi}, \tag{4.148}$$

where

$$R_{kk'nn'} = \begin{cases} y_{nk}(I_{kk'} - y_{nk'}), & n = n', \\ 0, & \text{otherwise.} \end{cases} \tag{4.149}$$

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 $\mathbf{u}_1, \cdots, \mathbf{u}_K$ are vectors in the same dimension as $\mathbf{w}$. Then,

$$\mathbf{u}^\mathsf{T}\mathbf{H}\mathbf{u} = \sum_{k=1}^K \sum_{k'=1}^K \mathbf{u}_k^\mathsf{T}\mathbf{H}_{kk'}\mathbf{u}_{k'}, \tag{4.150}$$

Then, the right hand side can be written as

$$\sum_{k=1}^K \sum_{k'=1}^K (\mathbf{\Phi}\mathbf{u}_k)^\mathsf{T} \mathbf{R}_{kk'} (\mathbf{\Phi}\mathbf{u}_{k'}). \tag{4.151}$$

4.21

Let

$$\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0,1)d\theta. \tag{4.152}$$

The right hand side can be written as

$$\int_{-\infty}^0 \mathcal{N}(\theta|0,1)d\theta + \int_0^a \mathcal{N}(\theta|0,1)d\theta = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^a \exp\left(-\frac{\theta^2}{2}\right) d\theta. \tag{4.153}$$

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

$$\frac{1}{\sqrt{2\pi}} \int_0^{\frac{a}{\sqrt{2}}} \exp\left(-t^2\right) \sqrt{2}dt = \frac{1}{2}\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right), \tag{4.154}$$

where

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp\left(-t^2\right) dt. \tag{4.155}$$

Therefore,

$$\Phi(a) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right). \tag{4.156}$$

4.22

Let $\boldsymbol{\theta}$ be a variable in M dimensions. By a Taylor expansion,

$$\ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) \simeq \ln\left(p(\mathcal{D}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)\right) + \mathbf{v}(\boldsymbol{\theta}_0)^\mathsf{T}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{A}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0), \tag{4.157}$$

where
$$\mathbf{v}(\boldsymbol{\theta}) = \nabla \ln \left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \right),$$
$$\mathbf{A}(\boldsymbol{\theta}) = -\nabla\nabla \ln \left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \right). \tag{4.158}$$

Let $\boldsymbol{\theta}_{\text{MAP}}$ be a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then,

$$\ln \left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \right) \simeq \ln \left(p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}}) \right) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^{\intercal}\mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}}), \tag{4.159}$$

so that

$$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^{\intercal}\mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}}) \right). \tag{4.160}$$

By marginalisation, integrating both sides with respect to $\boldsymbol{\theta}$ gives

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}}) \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}})^{\intercal}\mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{MAP}}) \right) d\boldsymbol{\theta}. \tag{4.161}$$

The integral of the right hand side can be written as

$$(2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}})^{-1} \right)^{\frac{1}{2}} = (2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}}) \right)^{-\frac{1}{2}}. \tag{4.162}$$

Therefore,

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}})p(\boldsymbol{\theta}_{\text{MAP}})(2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}}) \right)^{-\frac{1}{2}}, \tag{4.163}$$

so that

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \ln p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}}) \right). \tag{4.164}$$

4.23

Let $\boldsymbol{\theta}$ be a variable in M dimensions. By 4.22,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \ln p(\boldsymbol{\theta}_{\text{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det \mathbf{A}(\boldsymbol{\theta}_{\text{MAP}}) \right), \tag{4.165}$$

where $\boldsymbol{\theta}_{\text{MAP}}$ is a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and

$$\mathbf{A}(\boldsymbol{\theta}) = -\nabla\nabla \ln \left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \right). \tag{4.166}$$

If

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0), \tag{4.167}$$

then

$$\nabla\nabla \ln p(\theta) = -\mathbf{V}_0^{-1}, \tag{4.168}$$

so that

$$\mathbf{A}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta}) + \mathbf{V}_0^{-1}, \tag{4.169}$$

where

$$\mathbf{H}(\boldsymbol{\theta}) = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}). \tag{4.170}$$

Then, the right hand side of the approximation above can be written as

$$
\begin{aligned}
&\ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{V}_0) - \frac{1}{2}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0)^{\mathsf{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0) \\
&+ \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left(\mathbf{H}(\boldsymbol{\theta}_{\text{MAP}}) + \mathbf{V}_0^{-1}\right)\right) \\
&= \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) + \frac{1}{2}\ln\left(\det \mathbf{V}_0^{-1}\right) - \frac{1}{2}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0)^{\mathsf{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\text{MAP}} - \mathbf{m}_0) \\
&- \frac{1}{2}\ln\left(\det\left(\mathbf{H}(\boldsymbol{\theta}_{\text{MAP}}) + \mathbf{V}_0^{-1}\right)\right).
\end{aligned}
\tag{4.171}
$$

If $\mathbf{V}_0^{-1}$ can be neglected, the right hand side can be written as

$$\ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{1}{2}\ln\left(\det \mathbf{H}(\boldsymbol{\theta}_{\text{MAP}})\right). \tag{4.172}$$

If each data point is independent and identically distributed, then

$$\mathbf{H}(\boldsymbol{\theta}) = N\bar{\mathbf{H}}(\boldsymbol{\theta}), \tag{4.173}$$

where

$$\bar{\mathbf{H}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n(\boldsymbol{\theta}), \tag{4.174}$$

and $\mathbf{H}_n(\boldsymbol{\theta})$ is the one for each data point. Then,

$$\det \mathbf{H}(\boldsymbol{\theta}_{\text{MAP}}) = N^M \det \bar{\mathbf{H}}(\boldsymbol{\theta}_{\text{MAP}}). \tag{4.175}$$

Threrefore,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\text{MAP}}) - \frac{M}{2}\ln N. \tag{4.176}$$

4.24 (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$t_n \in \{0, 1\},$$
$$p(t_n|\mathbf{w}) = y_n^{t_n}(1 - y_n)^{1-t_n}, \qquad (4.177)$$

where

$$y_n = \sigma\left(\mathbf{w}^\intercal \boldsymbol{\phi}_n\right),$$
$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \qquad (4.178)$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \qquad (4.179)$$

If

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0), \qquad (4.180)$$

then the lograrithm of the right hand side except the terms independent of $\mathbf{w}$ and $\mathbf{t}$ can be written as

$$\sum_{n=1}^{N} (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\intercal \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0). \qquad (4.181)$$

Then, by 4.22,

$$p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)? \qquad (4.182)$$

where $\mathbf{w}_{\mathrm{MAP}}$ is the maximum likelihood solution for $p(\mathbf{w})$ and

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}). \qquad (4.183)$$

By marginalisation,

$$p(\mathcal{C}_1|\mathbf{t}) = \int p(\mathcal{C}_1|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \qquad (4.184)$$

The logarithm of the integrand of the right hand side except the terms independent of $\mathbf{w}$ can be approximated as

$$-\ln\left(1 + \exp\left(-\mathbf{w}^\intercal \boldsymbol{\phi}\right)\right) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^\intercal \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}) = \qquad (4.185)$$

4.25

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)},$$

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1)d\theta. \tag{4.186}$$

By 4.12,

$$\frac{d\sigma(a)}{da} = \sigma(a)\left(1 - \sigma(a)\right). \tag{4.187}$$

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

$$\int \Phi\left(\lambda(\mu + \sigma z)\right)\mathcal{N}(\mu + \sigma z|\mu, \sigma^2)\sigma dz = \int \Phi\left(\lambda(\mu + \sigma z)\right)\mathcal{N}(z|0, 1)dz. \tag{4.195}$$

154

Then,

$$\frac{\partial}{\partial \mu} I(\mu) = \lambda \int \mathcal{N}\left(\lambda(\mu + \sigma z)|0, 1\right) \mathcal{N}(z|0, 1) dz. \tag{4.196}$$

The logarithm of the integrand of the right hand side can be written as

$$
\begin{aligned}
&-\frac{1}{2}\ln(2\pi) - \frac{\lambda^2(\mu + \sigma z)^2}{2} - \frac{1}{2}\ln(2\pi) - \frac{z^2}{2} \\
&= -\ln(2\pi) - \frac{1+\sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}\right)^2 + \frac{\mu^2\sigma^2\lambda^4}{2\left(1+\sigma^2\lambda^2\right)} - \frac{\mu^2\lambda^2}{2}.
\end{aligned}
\tag{4.197}
$$

The right hand side can be written as

$$
\begin{aligned}
&-\ln(2\pi) - \frac{1+\sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}\right)^2 - \frac{\mu^2\lambda^2}{2\left(1+\sigma^2\lambda^2\right)} \\
&= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(1+\sigma^2\lambda^2\right)^{-1} - \frac{1+\sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}\right)^2 \\
&\quad - \ln\lambda - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\lambda^{-2} + \sigma^2\right) - \frac{\mu^2}{2\left(\lambda^{-2} + \sigma^2\right)}.
\end{aligned}
\tag{4.198}
$$

Then, the integral can be written as

$$
\begin{aligned}
&\int \mathcal{N}\left(z \Big| -\frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}, \left(1+\sigma^2\lambda^2\right)^{-1}\right) \frac{1}{\lambda}\mathcal{N}\left(\mu|0, \lambda^{-2} + \sigma^2\right) dz \\
&= \frac{1}{\lambda}\mathcal{N}\left(\mu|0, \lambda^{-2} + \sigma^2\right).
\end{aligned}
\tag{4.199}
$$

Therefore,

$$\frac{\partial}{\partial \mu} I(\mu) = \mathcal{N}\left(\mu|0, \lambda^{-2} + \sigma^2\right). \tag{4.200}$$

Integrating both sides with respect to μ gives

$$I(\mu) = \int_{-\infty}^{\mu} \mathcal{N}\left(m|0, \lambda^{-2} + \sigma^2\right) dm. \tag{4.201}$$

By the transformation

$$m' = \frac{m}{\left(\lambda^{-2} + \sigma^2\right)^{\frac{1}{2}}}, \tag{4.202}$$

the right hand side can be written as

$$\int_{-\infty}^{\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}} \left(\lambda^{-2}+\sigma^2\right)^{-\frac{1}{2}} \mathcal{N}\left(m'|0,1\right)\left(\lambda^{-2}+\sigma^2\right)^{\frac{1}{2}} dm' = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}\right). \tag{4.203}$$

Therefore,

$$I(\mu) = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}\right). \tag{4.204}$$

156

5 Neural Networks

5.1

Let

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\sum_{m=1}^{M} w_{km}^{(2)} \sigma \left(\sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) + w_{k0}^{(2)} \right), \tag{5.1}$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{5.2}$$

Here,

$$\sigma(a) = \frac{\exp\left(\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)}. \tag{5.3}$$

The right hand side can be written as

$$\tanh\left(\frac{a}{2}\right) + \sigma(-a) = \tanh\left(\frac{a}{2}\right) + 1 - \sigma(a). \tag{5.4}$$

Therefore,

$$\sigma(a) = \frac{1}{2}\left(1 + \tanh\left(\frac{a}{2}\right)\right). \tag{5.5}$$

Then, the argument of the right hand side can be written as

$$\sum_{m=1}^{M} w_{km}^{(2)} \left(\frac{1}{2} \left(1 + \tanh\left(\frac{1}{2} \left(\sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) \right) \right) \right) + w_{k0}^{(2)}$$
$$= \frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} \tanh\left(\frac{1}{2} \sum_{d=1}^{D} w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)} \right) + \frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} + w_{k0}^{(2)}. \tag{5.6}$$

Therefore,

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} \tanh\left(\frac{1}{2} \sum_{d=1}^{D} w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)} \right) + \frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} + w_{k0}^{(2)} \right). \tag{5.7}$$

5.2

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables such that

$$p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}\right). \tag{5.8}$$

Then, the logarithm of the likelihood except the terms independent of $\mathbf{w}$ can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^{\mathsf{T}}\left(\beta^{-1}\mathbf{I}\right)^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})) = -\frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{5.9}$$

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

$$\mathbf{0} = -\beta\sum_{n=1}^{N}\frac{\partial\mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial\mathbf{w}}(\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n). \tag{5.10}$$

Let

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{5.11}$$

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

$$\mathbf{0} = \sum_{n=1}^{N}\frac{\partial\mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial\mathbf{w}}(\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n). \tag{5.12}$$

Therefore, maximising the likelihood is equivalent to minimising $E(\mathbf{w})$.

5.3

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables such that

$$p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \boldsymbol{\Sigma}\right). \tag{5.13}$$

Then, the logarithm of the likelihood except the terms independent of $\mathbf{w}$ and $\boldsymbol{\Sigma}$ can be written as

$$-\frac{1}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w}))^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})). \tag{5.14}$$

Setting the derivatives with respect to $\mathbf{w}$ and $\mathbf{\Sigma}$ to zero gives

$$\mathbf{0} = -\sum_{n=1}^{N} \frac{\partial \mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial \mathbf{w}} \mathbf{\Sigma}^{-1} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right),$$

$$\mathbf{O} = -\frac{1}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\left(\mathbf{\Sigma}^{-1}\right)^2 \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^\mathsf{T}.$$

(5.15)

Therefore, the maximum likelihood solution for $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^\mathsf{T}.$$

(5.16)

On the other hand, if $\mathbf{\Sigma}$ is fixed and known, then the maximum likelihood solution for $\mathbf{w}$ is given by minimising

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^\mathsf{T} \mathbf{\Sigma}^{-1} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right).$$

(5.17)

5.4

Let $t_1, \cdots, t_N$ be variables such that

$$t_n \in \{0, 1\},$$
$$p(t_n = 1|\mathbf{x}_n) = (1 - \epsilon)y_n + \epsilon\left(1 - y_n\right),$$

(5.18)

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}).$$

(5.19)

Then,

$$p(t_n|\mathbf{x}_n) = \left((1 - \epsilon)y_n + \epsilon\left(1 - y_n\right)\right)^{t_n} \left(\epsilon y_n + (1 - \epsilon)\left(1 - y_n\right)\right)^{1-t_n}.$$

(5.20)

Therefore,

$$-\ln\left(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n)\right)$$

$$= -\sum_{n=1}^{N} \left(t_n \ln\left((1 - \epsilon)y_n + \epsilon(1 - y_n)\right) + (1 - t_n)\ln\left(\epsilon y_n + (1 - \epsilon)(1 - y_n)\right)\right).$$

(5.21)

5.5

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in K dimensions such that

$$t_{nk} \in \{0, 1\},$$
$$p(t_{nk} = 1 | \mathbf{x}_n) = y_{nk}, \tag{5.22}$$

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 $t_1, \cdots, t_N$ be variables such that

$$t_n \in \{0, 1\},$$
$$p(t_n = 1 | \mathbf{x}_n) = y_n, \tag{5.27}$$

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,

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right). \tag{5.31}$$

If
$$y_n = \sigma(a_n), \tag{5.32}$$

where
$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{5.33}$$

then, by 4.12,
$$\frac{\partial E(\mathbf{w})}{\partial a_n} = -y_n(1 - y_n) \left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right). \tag{5.34}$$

Therefore,
$$\frac{\partial E(\mathbf{w})}{\partial a_n} = y_n - t_n. \tag{5.35}$$

5.7

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables such that
$$t_{nk} \in \{0, 1\}, \\ p(t_{nk} = 1 | \mathbf{x}_n) = y_{nk}, \tag{5.36}$$

where
$$y_{nk} = y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^{K} y_{nk} = 1. \tag{5.37}$$

Then,
$$p(\mathbf{t}_n | \mathbf{x}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{5.38}$$

Let
$$E(\mathbf{w}) = -\ln \left(\prod_{n=1}^{N} p(\mathbf{t}_n | \mathbf{x}_n) \right). \tag{5.39}$$

Then,
$$E(\mathbf{w}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.40}$$

If
$$y_{nk} = \frac{\exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}{\sum_{k=1}^{K} \exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}, \tag{5.41}$$

161

then, by 4.17,

$$\frac{\partial E(\mathbf{w})}{\partial a_{k'}} = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} y_{nk} (I_{kk'} - y_{nk})\frac{1}{y_{nk}}. \tag{5.42}$$

The right hand side can be written as

$$-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(I_{kk'} - y_{nk}) = -\sum_{n=1}^{N}\left(\sum_{k=1}^{K} t_{nk} y_{nk} - t_{nk'}\right). \tag{5.43}$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_{k}} = \sum_{n=1}^{N}(y_{nk} - t_{nk}). \tag{5.44}$$

5.8

Setting the derivative of

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \tag{5.45}$$

gives

$$\frac{d}{da}\tanh a = 1 - \left(\frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}\right)^2. \tag{5.46}$$

Therefore,

$$\frac{d}{da}\tanh a = 1 - (\tanh a)^2. \tag{5.47}$$

5.9

Let $t_1, \cdots, t_N$ be variables such that

$$t_n \in \{-1, 1\},$$
$$p(t_n = 1|\mathbf{x}_n) = \frac{1 + y_n}{2}, \tag{5.48}$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \tag{5.49}$$

Then,

$$p(t_n|\mathbf{x}_n) = \left(\frac{1 + y_n}{2}\right)^{\frac{1+t_n}{2}} \left(\frac{1 - y_n}{2}\right)^{\frac{1-t_n}{2}}. \tag{5.50}$$

Let

$$E(\mathbf{w}) = -\ln\left(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n)\right). \tag{5.51}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{1 + t_n}{2}\ln\frac{1 + y_n}{2} + \frac{1 - t_n}{2}\ln\frac{1 - y_n}{2}\right). \tag{5.52}$$

The appropriate choice of y is tanh.

5.10

Let

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}), \tag{5.53}$$

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

$$\nabla E\left(\mathbf{w}^*\right) = \mathbf{0}. \tag{5.54}$$

Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E\left(\mathbf{w}^*\right) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^*\right)^\mathsf{T}\mathbf{H}\left(\mathbf{w} - \mathbf{w}^*\right), \tag{5.55}$$

where

$$\mathbf{H} = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}^*}. \tag{5.56}$$

Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ be eigenvectors such that

$$\mathbf{H}\mathbf{u}_m = \lambda_m\mathbf{u}_m. \tag{5.57}$$

Note that $\mathbf{H}$ is a real symmetric matrix. Then, by **??**, we have

$$\mathbf{u}_m^\mathsf{T}\mathbf{u}_{m'} = I_{mm'}. \tag{5.58}$$

Therefore, there exists $\alpha_1, \cdots, \alpha_M$ such that

$$\mathbf{w} - \mathbf{w}^* = \sum_{m=1}^{M} \alpha_m \mathbf{u}_m. \tag{5.59}$$

Then, the approximation can be written as

$$E(\mathbf{w}) \simeq E\left(\mathbf{w}^*\right) + \frac{1}{2}\sum_{m=1}^{M} \lambda_m \alpha_m^2. \tag{5.60}$$

Therefore, the contours of constant E are ellipses whose axes are aligned with $\mathbf{u}_1, \cdots, \mathbf{u}_M$ with lengths which are proportional to $\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_M^{-\frac{1}{2}}$.

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

$$\nabla E\left(\mathbf{w}^*\right) = \mathbf{0}. \tag{5.61}$$

Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E\left(\mathbf{w}^*\right) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^*\right)^\mathsf{T}\mathbf{H}\left(\mathbf{w} - \mathbf{w}^*\right), \tag{5.62}$$

where

$$\mathbf{H} = \left.\nabla\nabla E(\mathbf{w})\right|_{\mathbf{w}=\mathbf{w}^*}. \tag{5.63}$$

If $\mathbf{H}$ is positive definite, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \tag{5.64}$$

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

$$\mathbf{w} = \mathbf{w}^*. \tag{5.65}$$

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,

$$E(\mathbf{w}) \simeq E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^{\mathsf{T}} \mathbf{b} + \frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^{\mathsf{T}} \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}}), \tag{5.66}$$

where

$$\begin{aligned} \mathbf{b} &= \left. \nabla E(\mathbf{w}) \right|_{\mathbf{w}=\hat{\mathbf{w}}}. \\ \mathbf{H} &= \left. \nabla\nabla E(\mathbf{w}) \right|_{\mathbf{w}=\hat{\mathbf{w}}}. \end{aligned} \tag{5.67}$$

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

$$M + \frac{M(M+1)}{2} = \frac{M(M+3)}{2}. \tag{5.68}$$

5.14

Let w be a variable. Let E_n be a function of w. Then, by a Taylor expansion,

$$\begin{aligned} E_n(w_{mm'} + \epsilon) &= E_n(w_{mm'}) + \left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} \epsilon + O\left(\epsilon^2\right), \\ E_n(w_{mm'} - \epsilon) &= E_n(w_{mm'}) - \left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} \epsilon + O\left(\epsilon^2\right). \end{aligned} \tag{5.69}$$

Therefore,

$$\left. \frac{\partial E_n}{\partial w} \right|_{w=w_{mm'}} = \frac{E_n(w_{mm'} + \epsilon) - E_n(w_{mm'} - \epsilon)}{2\epsilon} + O\left(\epsilon^2\right). \tag{5.70}$$

5.15 (Incomplete)

5.16

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be vectors. Let $\mathbf{y}_1, \cdots, \mathbf{y}_N$ be vectors which are dependent on a vector $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{t}_n\|^2. \tag{5.71}$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (\nabla \mathbf{y}_n)^\mathsf{T} (\mathbf{y}_n - \mathbf{t}_n), \tag{5.72}$$

so that

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\nabla \operatorname{vec} (\nabla \mathbf{y}_n)^\mathsf{T}\right)^\mathsf{T} (\mathbf{y}_n - \mathbf{t}_n) + \sum_{n=1}^{N} (\nabla \mathbf{y}_n)^\mathsf{T} (\nabla \mathbf{y}_n). \tag{5.73}$$

5.17

Let t be a variable. Let y be a function of a vector $\mathbf{x}$ and a vector $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \int \int (y - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt. \tag{5.74}$$

Then,

$$\nabla E(\mathbf{w}) = \int \int (y - t) p(\mathbf{x}, t) \nabla y \, d\mathbf{x} dt. \tag{5.75}$$

The right hand side can be written as

$$\int y \nabla y \left(\int p(\mathbf{x}, t) dt\right) d\mathbf{x} - \int \nabla y \left(\int t p(t|\mathbf{x}) dt\right) p(\mathbf{x}) d\mathbf{x}$$
$$= \int y \nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla y \, \mathrm{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5.76}$$

Then,

$$\nabla\nabla E(\mathbf{w}) = \int \nabla y (\nabla y)^\mathsf{T} p(\mathbf{x}) d\mathbf{x} + \int y \nabla\nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla\nabla y \, \mathrm{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5.77}$$

The second and the third terms of the right hand side can be written as

$$\mathrm{E}(y \nabla\nabla y) - \mathrm{E}(\nabla\nabla y \, \mathrm{E}(t|\mathbf{x})) = \mathrm{E}((y - \mathrm{E}(t|\mathbf{x})) \nabla\nabla y). \tag{5.78}$$

Therefore, if

$$y = \mathrm{E}(t|\mathbf{x}), \tag{5.79}$$

then

$$\nabla\nabla E(\mathbf{w}) = \int \nabla y (\nabla y)^\mathsf{T} p(\mathbf{x}) d\mathbf{x}. \tag{5.80}$$

5.18

Let $t_1, \cdots, t_N$ be variables. Let $y_1, \cdots, y_N$ be variablees such that

$$
\begin{aligned}
y_n &= \mathbf{w}_n^{(2)\mathsf{T}} \mathbf{z} + \mathbf{w}_n^{(0)\mathsf{T}} \mathbf{x}, \\
z_m &= \tanh\left(\mathbf{w}_m^{(1)\mathsf{T}} \mathbf{x}\right).
\end{aligned}
\tag{5.81}
$$

Let

$$
E = \frac{1}{2} \sum_{n=1}^{N} (y_n - t_n)^2.
\tag{5.82}
$$

Then,

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{w}_m^{(1)}}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.
\end{aligned}
\tag{5.83}
$$

Therefore,

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= (y_n - t_n)\mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= (y_n - t_n)\mathbf{A}\mathbf{w}_n^{(2)}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= (y_n - t_n)\mathbf{z},
\end{aligned}
\tag{5.84}
$$

where

$$
A_{mm'} = \left(1 - z_m^2\right) x_{m'}.
\tag{5.85}
$$

5.19

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
t_n &\in \{0, 1\}, \\
p(t_n|\mathbf{w}) &= y_n^{t_n}(1 - y_n)^{1-t_n},
\end{aligned}
\tag{5.86}
$$

where

$$
\begin{aligned}
y_n &= \sigma\left(a_n(\mathbf{w})\right), \\
\sigma(a) &= \frac{1}{1 + \exp(-a)},
\end{aligned}
\tag{5.87}
$$

Let
$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{5.88}$$

The right hand side can be written as

$$-\ln\left(\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\right) = -\sum_{n=1}^{N}\left(t_n \ln y_n + (1-t_n)\ln(1-y_n)\right). \tag{5.89}$$

Then, by 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{t_n}{y_n}y_n(1-y_n)\nabla a_n - \frac{1-t_n}{1-y_n}y_n(1-y_n)\nabla a_n\right). \tag{5.90}$$

The right hand side can be written as

$$-\sum_{n=1}^{N}\left(t_n(1-y_n)\nabla a_n - (1-t_n)y_n\nabla a_n\right) = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n. \tag{5.91}$$

Then, by 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n. \tag{5.92}$$

Therefore, by 4.12,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\nabla a_n(\nabla a_n)^{\mathsf{T}} + \sum_{n=1}^{N}(y_n - t_n)\nabla\nabla a_n. \tag{5.93}$$

5.20

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables such that

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1|\mathbf{x}_n) &= y_{nk}, \end{aligned} \tag{5.94}$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^{K} y_{nk} &= 1. \end{aligned} \tag{5.95}$$

168

Then,

$$p(\mathbf{t}_n|\mathbf{x}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{5.96}$$

Let

$$E(\mathbf{w}) = -\ln\left(\prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n)\right). \tag{5.97}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.98}$$

If

$$y_{nk} = \frac{\exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}{\sum_{k=1}^{K} \exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}, \tag{5.99}$$

then, by 5.7,

$$\frac{\partial E(\mathbf{w})}{\partial a_k} = \sum_{n=1}^{N}(y_{nk} - t_{nk}). \tag{5.100}$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla a_k. \tag{5.101}$$

Then,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}(1 - y_{nk})\nabla a_k \left(\nabla a_k\right)^{\mathsf{T}} + \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla\nabla a_k. \tag{5.102}$$

5.21   (Incomplete)

5.22

Let $y_1, \cdots, y_N$ be variables such that

$$\begin{aligned} y_n &= \mathbf{w}_n^{(2)\mathsf{T}}\mathbf{z}, \\ z_m &= h(a_m), \\ a_m &= \mathbf{w}_m^{(1)\mathsf{T}}\mathbf{x}. \end{aligned} \tag{5.103}$$

169

Let E be a function of $y_1, \cdots, y_N$. 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,

$$
\begin{aligned}
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= w_{nm}^{(2)} \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm'}^{(2)} h'(a_{m'}) h'(a_m) \mathbf{x} + \frac{\partial E}{\partial y_n} h''(a_m) I_{mm'} \mathbf{x} \right)^{\mathsf{T}}, \\
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= h'(a_m) \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} \mathbf{z} + \frac{\partial E}{\partial y_n} \mathbf{v} \right)^{\mathsf{T}}, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{z} \mathbf{z}^{\mathsf{T}},
\end{aligned}
\tag{5.106}
$$

where

$$
v_m = \begin{cases} 1, & m = n, \\ 0 & \text{otherwise.} \end{cases}
\tag{5.107}
$$

5.23

Let $y_1, \cdots, y_N$ be variables such that

$$
\begin{aligned}
y_n &= \mathbf{w}_n^{(2)\mathsf{T}} \mathbf{z} + \mathbf{w}_n^{(0)\mathsf{T}} \mathbf{x}, \\
z_m &= h(a_m), \\
a_m &= \mathbf{w}_m^{(1)\mathsf{T}} \mathbf{x}.
\end{aligned}
\tag{5.108}
$$

Let E be a function of $y_1, \cdots, y_N$. Then,

$$\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} = \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}},$$
$$\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} = \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial z_m} \frac{\partial z_m}{\partial a_m} \frac{\partial a_m}{\partial \mathbf{w}_m^{(1)}}, \qquad (5.109)$$
$$\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} = \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.$$

Therefore,

$$\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} = \frac{\partial E}{\partial y_n} \mathbf{x},$$
$$\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} = \frac{\partial E}{\partial y_n} w_{nm}^{(2)} h'(a_m) \mathbf{x}, \qquad (5.110)$$
$$\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} = \frac{\partial E}{\partial y_n} \mathbf{z}.$$

Thus,

$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(0)}} = \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{x}\mathbf{x}^\mathsf{T},$$
$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_m^{(1)}} = \frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} h'(a_m) \mathbf{x}\mathbf{x}^\mathsf{T},$$
$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(2)}} = \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{x}\mathbf{z}^\mathsf{T},$$
$$\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} = w_{nm}^{(2)} \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm'}^{(2)} h'(a_{m'}) h'(a_m) \mathbf{x} + \frac{\partial E}{\partial y_n} h''(a_m) I_{mm'} \mathbf{x} \right)^\mathsf{T},$$
$$\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} = h'(a_m) \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} \mathbf{z} + \frac{\partial E}{\partial y_n} \mathbf{v} \right)^\mathsf{T},$$
$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} = \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{z}\mathbf{z}^\mathsf{T},$$

$$(5.111)$$

where

$$v_m = \begin{cases} 1, & m = n, \\ 0 & \text{otherwise.} \end{cases} \qquad (5.112)$$

5.24

Let $y_1, \cdots, y_n$ be variables such that

$$y_n = \mathbf{w}_n^\mathsf{T}\mathbf{z} + w_{n0},$$
$$z_m = h\left(\mathbf{w}_m^\mathsf{T}\mathbf{x} + w_{m0}\right). \tag{5.113}$$

(a)

Let

$$\tilde{\mathbf{x}} = a\mathbf{x} + b\mathbf{v}, \tag{5.114}$$

where

$$\mathbf{v} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Then,

$$z_m = h\left(\frac{1}{a}\mathbf{w}_m^\mathsf{T}\left(\tilde{\mathbf{x}} - b\mathbf{v}\right) + w_{m0}\right). \tag{5.115}$$

Therefore,

$$\tilde{z}_m = h\left(\tilde{\mathbf{w}}_m^\mathsf{T}\mathbf{x} + \tilde{w}_{m0}\right). \tag{5.116}$$

where

$$\tilde{\mathbf{w}}_m = \frac{1}{a}\mathbf{w}_m,$$
$$\tilde{w}_{m0} = w_{m0} - \frac{b}{a}\mathbf{w}_m^\mathsf{T}\mathbf{v}. \tag{5.117}$$

(b)

Let

$$\tilde{y}_n = cy_n + d. \tag{5.118}$$

Then,

$$\frac{\tilde{y}_n - d}{c} = \mathbf{w}_n^\mathsf{T}\mathbf{z} + w_{n0}. \tag{5.119}$$

Therefore,

$$\tilde{y}_n = \tilde{\mathbf{w}}_n^\mathsf{T}\mathbf{z} + \tilde{w}_{n0}, \tag{5.120}$$

where

$$\tilde{\mathbf{w}}_n = c\mathbf{w}_n,$$
$$\tilde{w}_{n0} = cw_{n0} + d. \tag{5.121}$$

5.25   (Incomplete)

Let E be a quadratic error function such that

$$E(\mathbf{w}) = E_0 + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^*\right)^\mathsf{T}\mathbf{H}\left(\mathbf{w} - \mathbf{w}^*\right), \tag{5.122}$$

where $\mathbf{w}^*$ represents the minimum and $\mathbf{H}$ is a positive definite and constant matrix whose eignevectors and eigenvalues are $\mathbf{u}_1, \cdots, \mathbf{u}_M$ and $\eta_1, \cdots, \eta_M$. Let

$$\begin{aligned}
\mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \rho\nabla E, \\
\mathbf{w}^{(0)} &= \mathbf{0}.
\end{aligned} \tag{5.123}$$

(a)

We have

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\mathbf{H}\left(\mathbf{w}^{(\tau-1)} - \mathbf{w}^*\right), \tag{5.124}$$

so that

$$\mathbf{w}^{(\tau)} = \left(\mathbf{I} - \rho\mathbf{H}\right)\mathbf{w}^{(\tau-1)} + \rho\mathbf{H}\mathbf{w}^*. \tag{5.125}$$

Then,

$$\mathbf{u}_m^\mathsf{T}\mathbf{w}^{(\tau)} = \mathbf{u}_m^\mathsf{T}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_m^\mathsf{T}\mathbf{H}\mathbf{w}^{(\tau-1)} + \rho\mathbf{u}_m^\mathsf{T}\mathbf{H}\mathbf{w}^*. \tag{5.126}$$

Since $\mathbf{H}$ is symmetric,

$$\mathbf{u}_m^\mathsf{T}\mathbf{H}\mathbf{w} = \mathbf{w}^\mathsf{T}\mathbf{H}\mathbf{u}_m. \tag{5.127}$$

The right hand side can be written as

$$\eta_m\mathbf{w}^\mathsf{T}\mathbf{u}_m = \eta_m w_m, \tag{5.128}$$

where

$$w_m = \mathbf{w}^\mathsf{T}\mathbf{u}_m. \tag{5.129}$$

Then,

$$w_m^{(\tau)} = (1 - \rho\eta_m)w_m^{(\tau-1)} + \rho\eta_m w_m^*, \tag{5.130}$$

so that

$$w_m^{(\tau)} - w_m^* = (1 - \rho\eta_m)\left(w_m^{(\tau-1)} - w_m^*\right). \tag{5.131}$$

Therefore,

$$w_m^{(\tau)} = w_m^* - (1 - \rho\eta_m)^\tau\left(w_m^{(0)} - w_m^*\right), \tag{5.132}$$

so that

$$w_m^{(\tau)} = \left(1 - (1 - \rho\eta_m)^\tau\right)w_m^*. \tag{5.133}$$

(b)

By (a), if
$$|1 - \rho\eta_m| < 1, \tag{5.134}$$

then
$$\lim_{\tau \to \infty} w_m^{(\tau)} = w_m^*. \tag{5.135}$$

(c)

By (a) and a Taylor series
$$(1 - \rho\eta_m)^\tau = 1 - \rho\tau\eta_m + \rho^2\tau^2 O\left(\eta_m^2\right), \tag{5.136}$$

we have
$$w_m^{(\tau)} = \left(\rho\tau\eta_m - \rho^2\tau^2 O\left(\eta_m^2\right)\right) w_m^*, \tag{5.137}$$

so that
$$w_m^{(\tau)} = \rho\tau\eta_m \left(1 - \rho\tau\eta_m O\left(\eta_m\right)\right) w_m^*. \tag{5.138}$$

Therefore,
$$\begin{aligned} w_m^{(\tau)} &\simeq w_m^*, \quad \text{if} \quad \eta_m \gg (\rho\tau)^{-1}? \\ \left|w_m^{(\tau)}\right| &\ll w_m^*, \quad \text{if} \quad \eta_m \ll (\rho\tau)^{-1}. \end{aligned} \tag{5.139}$$

5.26   (Incomplete)

Let
$$\tilde{E} = E + \lambda\Omega, \tag{5.140}$$

where
$$\Omega = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\frac{\partial y_{nk}}{\partial \xi}\right)^2. \tag{5.141}$$

(a)

We have
$$\frac{\partial y_{nk}}{\partial \xi} = \mathcal{G} y_{nk}, \tag{5.142}$$

where

$$\mathcal{G} = \sum_{d=1}^{D} \tau_d \frac{\partial}{\partial x_d}, \tag{5.143}$$

$$\tau_d = \frac{\partial x_d}{\partial \xi}.$$

Therefore,

$$\Omega = \sum_{n=1}^{N} \Omega_n, \tag{5.144}$$

where

$$\Omega_n = \frac{1}{2} \sum_{k=1}^{K} (\mathcal{G} y_{nk})^2. \tag{5.145}$$

(b)

$$\alpha_j = \mathcal{G} z_j = h'(a_j) \beta_j?$$

$$\beta_j = \mathcal{G} a_j = \sum_{d=1}^{D} w_{jd} \alpha_d? \tag{5.146}$$

5.27

Let

$$\tilde{E} = \frac{1}{2} \iiint (y(\mathbf{x} + \boldsymbol{\xi}) - t)^2 \, p(t|\mathbf{x}) p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} dt d\boldsymbol{\xi}, \tag{5.147}$$

where

$$p(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\xi}|\mathbf{0}, \mathbf{I}). \tag{5.148}$$

By a Taylor series

$$y(\mathbf{x} + \boldsymbol{\xi}) = y(\mathbf{x}) + \boldsymbol{\xi}^\mathsf{T} \nabla y(\mathbf{x}) + \frac{1}{2} \boldsymbol{\xi}^\mathsf{T} \nabla \nabla y(\mathbf{x}) \boldsymbol{\xi} + O\left(\|\boldsymbol{\xi}\|^3\right), \tag{5.149}$$

the integrand can be written as

$$(y(\mathbf{x}) - t)^2 + 2(y(\mathbf{x}) - t) \left(\boldsymbol{\xi}^\mathsf{T} \nabla y(\mathbf{x}) + \frac{1}{2} \boldsymbol{\xi}^\mathsf{T} \nabla \nabla y(\mathbf{x}) \boldsymbol{\xi} + O\left(\|\boldsymbol{\xi}\|^3\right) \right)$$
$$+ \left(\boldsymbol{\xi}^\mathsf{T} \nabla y(\mathbf{x}) + \frac{1}{2} \boldsymbol{\xi}^\mathsf{T} \nabla \nabla y(\mathbf{x}) \boldsymbol{\xi} + O\left(\|\boldsymbol{\xi}\|^3\right) \right)^2. \tag{5.150}$$

Omitting the terms of $O(\|\boldsymbol{\xi}\|^3)$, the integral can be approximated as

$$\iiint (y(\mathbf{x}) - t)^2 p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi}$$

$$+ 2\iiint (y(\mathbf{x}) - t)\left(\boldsymbol{\xi}^\mathsf{T}\nabla y(\mathbf{x}) + \frac{1}{2}\boldsymbol{\xi}^\mathsf{T}\nabla\nabla y(\mathbf{x})\boldsymbol{\xi}\right) p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi}$$

$$+ \iiint \left(\boldsymbol{\xi}^\mathsf{T}\nabla y(\mathbf{x}) + \frac{1}{2}\boldsymbol{\xi}^\mathsf{T}\nabla\nabla y(\mathbf{x})\boldsymbol{\xi}\right)^2 p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi}.$$

(5.151)

Omitting the terms of $O(\|\boldsymbol{\xi}\|^3)$, the second and third terms can be approximated as

$$2\iiint (y(\mathbf{x}) - t)\,\boldsymbol{\xi}^\mathsf{T}\nabla y(\mathbf{x})p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi}$$

$$+ \iiint (y(\mathbf{x}) - t)\,\boldsymbol{\xi}^\mathsf{T}\nabla\nabla y(\mathbf{x})\boldsymbol{\xi}p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi} \qquad (5.152)$$

$$+ \iiint \boldsymbol{\xi}^\mathsf{T}\left(\nabla y(\mathbf{x})\right)\left(\nabla y(\mathbf{x})\right)^\mathsf{T}\boldsymbol{\xi}p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi}.$$

Since

$$\mathrm{E}\,\boldsymbol{\xi} = \mathbf{0}, \qquad (5.153)$$

the integral of the first term can be written as

$$\left(\int \boldsymbol{\xi}p(\boldsymbol{\xi})d\boldsymbol{\xi}\right)^\mathsf{T}\iint (y(\mathbf{x}) - t)\,\nabla y(\mathbf{x})p(t|\mathbf{x})p(\mathbf{x})d\mathbf{x}dt = \mathbf{0}. \qquad (5.154)$$

Since

$$\mathrm{cov}\,\boldsymbol{\xi} = \mathbf{I}, \qquad (5.155)$$

The integrals of the second term can be written as

$$\int \boldsymbol{\xi}^\mathsf{T}\left(\iint (y(\mathbf{x}) - t)\,\nabla\nabla y(\mathbf{x})p(t|\mathbf{x})p(\mathbf{x})d\mathbf{x}dt\right)\boldsymbol{\xi}p(\boldsymbol{\xi})d\boldsymbol{\xi}$$

$$= \mathrm{tr}\left(\iint (y(\mathbf{x}) - t)\,\nabla\nabla y(\mathbf{x})p(t|\mathbf{x})p(\mathbf{x})d\mathbf{x}dt\right). \qquad (5.156)$$

The integral of the right hand side can be written as

$$\int \left(y(\mathbf{x}) - \int tp(t|\mathbf{x})dt\right)\nabla\nabla y(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

$$= \int (y(\mathbf{x}) - \mathrm{E}(t|\mathbf{x}))\,\nabla\nabla y(\mathbf{x})p(\mathbf{x})d\mathbf{x}. \qquad (5.157)$$

Similarly, the integral of the third term can be written as

$$\int \boldsymbol{\xi}^{\mathsf{T}} \left(\int\int (\nabla y(\mathbf{x}))(\nabla y(\mathbf{x}))^{\mathsf{T}} p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dt \right) \boldsymbol{\xi}d\boldsymbol{\xi}$$
$$= \mathrm{tr} \left(\int\int (\nabla y(\mathbf{x}))(\nabla y(\mathbf{x}))^{\mathsf{T}} p(t|\mathbf{x})p(\mathbf{x})d\mathbf{x}dt \right). \tag{5.158}$$

The integral of the right hand side can be written as

$$\int \left(\int p(t|\mathbf{x})dt \right) (\nabla y(\mathbf{x}))(\nabla y(\mathbf{x}))^{\mathsf{T}} p(\mathbf{x})d\mathbf{x}$$
$$= \int (\nabla y(\mathbf{x}))(\nabla y(\mathbf{x}))^{\mathsf{T}} p(\mathbf{x})d\mathbf{x}. \tag{5.159}$$

If

$$y(\mathbf{x}) = \mathrm{E}(t|\mathbf{x}), \tag{5.160}$$

then

$$\tilde{E} \simeq E + \Omega, \tag{5.161}$$

where

$$E = \frac{1}{2} \int\int\int (y(\mathbf{x}) - t)^2 p(t|\mathbf{x})p(\mathbf{x})p(\boldsymbol{\xi})d\mathbf{x}dtd\boldsymbol{\xi},$$
$$\Omega = \frac{1}{2} \mathrm{tr} \left(\int \nabla y(\mathbf{x}) (\nabla y(\mathbf{x}))^{\mathsf{T}} p(\mathbf{x})d\mathbf{x}. \right). \tag{5.162}$$

5.28 (Incomplete)

5.29 (Incomplete)

Let $w_1, \cdots, w_M$ be variables such that

$$p(w_m) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right). \tag{5.163}$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda\Omega(\mathbf{w}), \tag{5.164}$$

where

$$\Omega(\mathbf{w}) = -\sum_{m=1}^{M} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right) \right). \tag{5.165}$$

177

We have

$$\frac{\partial \Omega}{\partial w_m} = -\sum_{k=1}^{K} \pi_k \frac{w_m - \mu_k}{\sigma_k^2}. \tag{5.166}$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial w_m} = \frac{\partial E}{\partial w_m} - \lambda \sum_{k=1}^{K} \pi_k \frac{w_m - \mu_k}{\sigma_k^2}. \tag{5.167}$$