Solutions Manual to
Pattern Recognition and Machine Learning

Hiromichi Inawashiro

February 26, 2026

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

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

then

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

Then,

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

Therefore,

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

where

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

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

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

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

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

If

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

then

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

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,

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

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

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

Let

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

(a)

By the Bayes' theorem,

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

Therefore,

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

(b)

By the Bayes' thorem,

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

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

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

Therefore,

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

1.4

Let x and y be variables such that

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

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

$$|y - \hat{y}| < \epsilon, \tag{1.25}$$

then

$$g'(y) \neq 0. \tag{1.26}$$

(a)

Since

$$\int p_x(x)dx = \int p_x\left(g(y)\right)|g'(y)|dy, \tag{1.27}$$

we have

$$p_y(y) = p_x\left(g(y)\right)|g'(y)|. \tag{1.28}$$

Taking the derivative and substituting

$$y = \hat{y} \tag{1.29}$$

gives

$$0 = g'(\hat{y})p_x'\left(g\left(\hat{y}\right)\right) + p_x\left(g\left(\hat{y}\right)\right)g''\left(\hat{y}\right). \tag{1.30}$$

Therefore, in general,

$$\hat{x} \neq g\left(\hat{y}\right). \tag{1.31}$$

4

(b)

By (a), if
$$g(y) = ay + b, \tag{1.32}$$
then
$$0 = ap'_x \left(g\left(\hat{y} \right) \right). \tag{1.33}$$
Therefore,
$$\hat{x} = g\left(\hat{y} \right). \tag{1.34}$$

1.5

We have
$$\text{var } f(x) = \text{E} \left(f(x) - \text{E} f(x) \right)^2. \tag{1.35}$$
The right hand side can be written as
$$\text{E} \left((f(x))^2 - 2f(x) \, \text{E} \, f(x) + (\text{E} \, f(x))^2 \right) = \text{E} \left(f(x) \right)^2 - (\text{E} \, f(x))^2. \tag{1.36}$$
Therefore,
$$\text{var } f(x) = \text{E} \left(f(x) \right)^2 - (\text{E} \, f(x))^2. \tag{1.37}$$

1.6

We have
$$\text{cov}(x, y) = \text{E} \left((x - \text{E} \, x)(y - \text{E} \, y) \right). \tag{1.38}$$
The right hand side can be written as
$$\text{E} \, xy - \text{E} \left(x \, \text{E} \, y \right) - \text{E} \left(y \, \text{E} \, x \right) + \text{E} \left(\text{E} \, x \, \text{E} \, y \right) = \text{E} \, xy - \text{E} \, x \, \text{E} \, y. \tag{1.39}$$
The right hand side can be written as
$$\int xyp(x, y)dxdy - \int xp(x)dx \int yp(y)dy. \tag{1.40}$$
If x and y are independent, then
$$f(x, y) = f(x)f(y). \tag{1.41}$$
Then,
$$\int xyp(x, y)dxdy = \int p(x)dx \int p(y)dy. \tag{1.42}$$
Therefore,
$$\text{cov}(x, y) = 0. \tag{1.43}$$

5

1.7

(a)

Let

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

Then,

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

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

$$\int_{0}^{\infty} \int_{0}^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} drd\theta$$
$$=2\pi \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdr. \tag{1.46}$$

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

$$2\pi\sigma^2 \int_{0}^{\infty} \exp\left(-\frac{1}{2}s^2\right) sds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_{0}^{\infty}. \tag{1.47}$$

Therefore,

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

(b)

We have

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

Then,

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

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

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

By 1.7(a), it can be written as $(2\pi\sigma^2)^{\frac{1}{2}}$. Therefore,

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

1.8

(a)

Let x be a variable such that

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

Then,

$$\mathrm{E}\, x = \int_{-\infty}^{\infty} x\mathcal{N}\left(x|\mu, \sigma^2\right) dx. \tag{1.54}$$

The right hand side can be written as

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

By the transformation

$$y = x - \mu, \tag{1.56}$$

the integral can be written as

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

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

$$\mu \left(2\pi\sigma^2\right)^{\frac{1}{2}}. \tag{1.58}$$

Therefore,

$$\mathrm{E}\, x = \mu. \tag{1.59}$$

7

(b)

By 1.7(b),

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

so that

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

Taking the derivative with respect to σ^2 gives

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

The left hand side can be written as

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

Therefore,

$$\operatorname{var} x = \sigma^2. \tag{1.64}$$

1.9

(a)

Let x be a variable such that

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

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

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

Therefore,

$$\operatorname{mode} x = \mu. \tag{1.67}$$

(b)

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

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

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

$$\begin{aligned}
\mathbf{0} = & - (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right)(\mathbf{x} - \boldsymbol{\mu}) \\
& \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).
\end{aligned} \tag{1.69}$$

Therefore,

$$\text{mode}\,\mathbf{x} = \boldsymbol{\mu}. \tag{1.70}$$

1.10

(a)

We have

$$\mathrm{E}(x + y) = \iint (x + y)p(x, y)dxdy. \tag{1.71}$$

The right hand side can be written as

$$\begin{aligned}
& \int x \left(\int p(x, y)dy\right) dx + \int y \left(\int p(x, y)dx\right) dy \\
& = \int xp(x)dx + \int yp(y)dy.
\end{aligned} \tag{1.72}$$

The right hand side can be written as

$$\mathrm{E}\,x + \mathrm{E}\,y. \tag{1.73}$$

Therefore,

$$\mathrm{E}(x + y) = \mathrm{E}\,x + \mathrm{E}\,y. \tag{1.74}$$

(b)

We have

$$\mathrm{var}(x + y) = \mathrm{E}\left(x + y - \mathrm{E}(x + y)\right)^2 \tag{1.75}$$

The right hand side can be written as

$$\mathrm{E}\,(x - \mathrm{E}\,x)^2 + 2\,\mathrm{E}\,((x - \mathrm{E}\,x)\,(y - \mathrm{E}\,y)) + \mathrm{E}\,(y - \mathrm{E}\,y)^2$$
$$= \mathrm{var}\,x + 2\,\mathrm{cov}(x, y) + \mathrm{var}\,y. \tag{1.76}$$

By 1.6, if x and y are independent, then

$$\mathrm{cov}(x, y) = 0. \tag{1.77}$$

Therefore,

$$\mathrm{var}(x + y) = \mathrm{var}\,x + \mathrm{var}\,y. \tag{1.78}$$

1.11

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

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

Then,

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

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

$$0 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu),$$
$$0 = -\frac{N}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{1.81}$$

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

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n,$$
$$\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2. \tag{1.82}$$

1.12

(a)

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

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right),$$
$$p(x_{n'}) = \mathcal{N}\left(x_{n'} | \mu, \sigma^2\right). \tag{1.83}$$

Then,

$$\mathrm{E}\, x_n x_{n'} = \mu^2. \tag{1.84}$$

By the property

$$\mathrm{E}\, x_n^2 = \mathrm{var}\, x_n + \left(\mathrm{E}\, x_n\right)^2, \tag{1.85}$$

we have

$$\mathrm{E}\, x_n^2 = \sigma^2 + \mu^2. \tag{1.86}$$

Therefore,

$$\mathrm{E}\, x_n x_{n'} = \mu^2 + I_{nn'}\sigma^2. \tag{1.87}$$

(b)

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

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

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

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

Then,

$$\mathrm{E}\, \mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{E}\, x_n. \tag{1.90}$$

Therefore,

$$\mathrm{E}\, \mu_{\mathrm{ML}} = \mu. \tag{1.91}$$

(c)

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

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

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

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

Then,

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

The right hand side can be writen as

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

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

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

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

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

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

$$\begin{aligned}
&-\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\,x_n^2 - \frac{2}{N^2}\sum_{1\le n<n'\le N}\mathrm{E}\,x_n x_{n'}\\
&= -\frac{1}{N^2}N\left(\mu^2 + \sigma^2\right) - \frac{2}{N^2}\frac{N(N-1)}{2}\mu^2.
\end{aligned}\tag{1.98}$$

The right hand side can be written as

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

Then,

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

Therefore,

$$\mathrm{E}\,\sigma_{\mathrm{ML}}^2 = \frac{N-1}{N}\sigma^2. \tag{1.101}$$

1.13

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

$$\begin{aligned}
\mathrm{E}\,x_n &= \mu, \\
\mathrm{var}\,x_n &= \sigma^2.
\end{aligned} \tag{1.102}$$

We have

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

The right hand side can be writen as

$$\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{var}\,x_n = \frac{\sigma^2}{N}. \tag{1.104}$$

Therefore,

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

1.14

Let

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

(a)

We have

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

(b)

We have

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

The right hand side can be written as

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

Therefore,

$$
\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{A}} x_d x_{d'} = 0.
\tag{1.110}
$$

(c)

We have

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

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

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

Therefore,

$$
\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'} x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}^{\mathrm{S}} x_d x_{d'}.
\tag{1.113}
$$

(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

$$n(1, M) = n(1, M - 1) = 1. \tag{1.114}$$

Let us assume that

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

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

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

so that

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

Therefore, the assumption is proved by induction on D.

(b)

We have

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

Let us assume that

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

15

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

The right hand side can be written as
$$\frac{D(D+M-1)! + M(D+M-1)!}{D!M!} = \frac{(D+M)!}{D!M!}. \tag{1.121}$$

Therefore, the assumption is proved by induction on D.

(c)

By 1.14(d),
$$n(D,2) = \frac{D(D+1)}{2}. \tag{1.122}$$

Let us assume that
$$n(D,M) = \frac{(D+M-1)!}{(D-1)!M!}. \tag{1.123}$$

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

By the assumption and (b), the right hand side can be written as
$$\sum_{d=1}^{D} \frac{(d+M-1)!}{(d-1)!M!} = \frac{(D+M)!}{(D-1)!(M+1)!}. \tag{1.125}$$

Therefore, the assumption is proved by induction on M.

1.16

(a)

Let $N(D,M)$ be the number of independent parameters in all of the terms up to and including the ones of D dimensions and M orders. By 1.15,
$$N(D,M) = \sum_{m=0}^{M} n(D,m), \tag{1.126}$$

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

(b)

By (a),
$$N(D, 0) = 1. \tag{1.128}$$

Let us assume that
$$\sum_{m=0}^{M} n(D, m) = \frac{(D + M)!}{D!M!}. \tag{1.129}$$

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

The right hand side can be written as
$$\frac{(M + 1)(D + M)! + D(D + M)!}{D!(M + 1)!} = \frac{(D + M + 1)!}{D!(M + 1)!}. \tag{1.131}$$

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

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

(c)

By the approximation
$$n! \simeq n^n \exp(-n), \tag{1.133}$$

we have
$$\frac{(D + M)!}{D!M!} \simeq \frac{(D + M)^{D+M}}{D^D M^M}. \tag{1.134}$$

The right hand side can be written as

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

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

(d)

By (b),

$$N(10, 3) = 286,$$
$$N(100, 3) = 176851, \tag{1.137}$$
$$N(1000, 3) = 167668501.$$

1.17

Let

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

(a)

We have

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

The right hand side can be written as

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

Therefore,

$$\Gamma(x + 1) = x\Gamma(x). \tag{1.141}$$

(b)

We have

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

so that

$$\Gamma(1) = 0!. \tag{1.143}$$

For a positive integer x, let us assume that

$$\Gamma(x) = (x - 1)!. \tag{1.144}$$

By (a),

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

18

so that

$$\Gamma(x+1) = x!. \tag{1.146}$$

Therefore, the assumption is proved by induction on x.

1.18

(a)

Let

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

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

$$s = r^2, \tag{1.148}$$

the right hand side can be written as

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

Therefore,

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

(b)

The volume of the sphere can can be written as

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

Therefore,

$$V_D = \frac{S_D}{D}. \tag{1.152}$$

(c)

By (a) and (b),

$$S_2 = 2\pi,$$
$$V_2 = \pi.$$
$$(1.153)$$

Similarly,

$$S_3 = 4\pi,$$
$$V_3 = \frac{4}{3}\pi.$$
$$(1.154)$$

1.19

(a)

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

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

(b)

By (a) and the Stering's formula

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

we have

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

The right hand side can be written as

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

Therefore,

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

(c)

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

$$\frac{\sqrt{\sum_{i=1}^{D} 1^2}}{1} = \sqrt{D}. \tag{1.160}$$

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

1.20

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

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

(a)

We have

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

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

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

Therefore,

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

where

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

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

21

(b)

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

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

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

$$\hat{r} = \sqrt{D-1}\sigma. \tag{1.167}$$

(c)

We have

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

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

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

By the Taylor series

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

the right hand side can be approximated as

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

Therefore,

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

(d)

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

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

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

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

Therefore,

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

1.21

(a)

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

$$0 \leq a(b-a). \tag{1.176}$$

Therefore,

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

(b)

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

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

By (a),

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

Therefore,

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

1.22

Let

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

where

$$L_{kj} = 1 - I_{kj}. \tag{1.182}$$

The right hand side can be written as

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

The right hand side can be written as

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

Then,

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

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

1.23

Let

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

The right hand side can be written as

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

Then,

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

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

1.24 (Incomplete)

Let

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

1.25

Let

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

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

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

The integral of the right hand side can be written as

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

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

$$\mathbf{0} = p(\mathbf{x}) \left(\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \right). \tag{1.193}$$

Therefore,

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

For a single target variable t, it reduces to

$$\operatorname*{argmin}_{\mathbf{y}(\mathbf{x})} \mathrm{E}\,L(t, \mathbf{y}(\mathbf{x})) = \mathrm{E}_t(t|\mathbf{x}). \tag{1.195}$$

1.26

Let

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

The right hand side can be written as

$$\iint \|\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x}) + \mathrm{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} = I_1 + 2I_2 + I_3, \quad (1.197)$$

where

$$I_1 = \iint \|\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t},$$

$$I_2 = \iint (\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x}))^\mathsf{T} (\mathrm{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) \, p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}, \quad (1.198)$$

$$I_3 = \iint \|\mathrm{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}.$$

The expression of I_1 can be written as

$$\int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x})\|^2 \left(\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x}$$
$$= \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \quad (1.199)$$

The expression of I_2 can be written as

$$\int (\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x}))^\mathsf{T} \left(\int (\mathrm{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) \, p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.200)$$

We have

$$\int (\mathrm{E}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) \, p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \mathrm{E}(\mathbf{t}|\mathbf{x}) \int p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \quad (1.201)$$

The right hand side can be written as

$$\mathrm{E}(\mathbf{t}|\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x}) = 0. \quad (1.202)$$

The expression of I_3 can be written as

$$\int \left(\int \|\mathbf{t} - \mathrm{E}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x} = \int \mathrm{var}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.203)$$

Then,

$$\mathrm{E} \, L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \int \mathrm{var}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.204)$$

Therefore,

$$\operatorname*{argmin}_{\mathbf{y}(\mathbf{x})} \mathrm{E} \, L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathrm{E}(\mathbf{t}|\mathbf{x}). \quad (1.205)$$

26

1.27   (Incomplete)

(a)

Let

$$\mathrm{E}\,L_q = \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \tag{1.206}$$

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

$$0 = qp(\mathbf{x}) \int |y(\mathbf{x}) - t|^{q-1} \operatorname{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt. \tag{1.207}$$

Therefore,

$$\begin{aligned}
&\underset{y(\mathbf{x})}{\operatorname{argmin}}\,\mathrm{E}\,L_q \\
&= \left\{ y(\mathbf{x}) \mid \int |y(\mathbf{x}) - t|^{q-1} \operatorname{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = 0 \right\}.
\end{aligned} \tag{1.208}$$

(b)

We have

$$\mathrm{E}\,L_1 = \int \left(\int \operatorname{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.209}$$

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

$$\int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x}) dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x}) dt. \tag{1.210}$$

Therefore,

$$\underset{y(\mathbf{x})}{\operatorname{argmin}}\,\mathrm{E}\,L_1 = \operatorname{median}(t|\mathbf{x}). \tag{1.211}$$

(c)

We have

$$\lim_{q \to 0} \left(\underset{y(\mathbf{x})}{\operatorname{argmin}}\,\mathrm{E}\,L_q \right) = \operatorname{mode}(t|\mathbf{x})? \tag{1.212}$$

27

1.28

(a)

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \tag{1.213}$$

Then,

$$h\left(p^2\right) = 2h(p). \tag{1.214}$$

Let us assume that, for a positive integer n,

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

Then, by the first assumption,

$$h\left(p^{n+1}\right) = h\left(p^n\right) + h(p), \tag{1.216}$$

so that

$$h\left(p^{n+1}\right) = (n+1)h(p). \tag{1.217}$$

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

(b)

For positive integers m and n,

$$h\left(p^n\right) = h\left(p^{\frac{n}{m}m}\right). \tag{1.218}$$

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,

$$h\left(p^{\frac{n}{m}}\right) = \frac{n}{m}h(p). \tag{1.219}$$

(c)

By the continuity, for a positive real number a,

$$h\left(p^a\right) = ah(p). \tag{1.220}$$

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

$$(p \ln p)h'(p) = h(p). \tag{1.221}$$

Then,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + \text{const} .\qquad(1.222)$$

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,

$$h(p) \propto \ln p.\qquad(1.223)$$

1.29

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

$$\mathrm{H}(x) = - \sum_{m=1}^{M} p(x_m) \ln p(x_m),\qquad(1.224)$$

where

$$\sum_{m=1}^{M} p(x_m) = 1.\qquad(1.225)$$

By the Jensen's inequality,

$$\sum_{m=1}^{M} p(x_i) \ln \frac{1}{p(x_m)} \leq \ln \left(\sum_{m=1}^{M} 1 \right).\qquad(1.226)$$

Therefore,

$$\mathrm{H}(x) \leq \ln M.\qquad(1.227)$$

1.30

Let

$$\begin{aligned}p(x) &= \mathcal{N}\left(x|\mu, \sigma^2\right), \\ q(x) &= \mathcal{N}\left(x|m, s^2\right).\end{aligned}\qquad(1.228)$$

Then, the Kullback-Leibler divergence is given by

$$\mathrm{KL}\left(p||q\right) = - \int p(x) \ln \frac{q(x)}{p(x)} dx.\qquad(1.229)$$

The right hand side can be written as

$$-\int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx$$

$$= -\int_{-\infty}^{\infty} p(x) \left(-\frac{1}{2}\ln\frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2}\right) dx. \qquad (1.230)$$

The right hand side can be written as

$$\ln\frac{s}{\sigma}\int_{-\infty}^{\infty} p(x)dx + \frac{1}{2s^2}\int_{-\infty}^{\infty}(x-m)^2 p(x)dx - \frac{1}{2\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx. \quad (1.231)$$

The integral of the second term can be written as

$$\int_{-\infty}^{\infty} (x - \mu + \mu - m)^2 p(x)dx$$

$$= \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx + 2(\mu-m)\int_{-\infty}^{\infty}(x-\mu)p(x)dx \qquad (1.232)$$

$$+ (\mu-m)^2 \int_{-\infty}^{\infty} p(x)dx.$$

Therefore,

$$\mathrm{KL}\,(p\|q) = \ln\frac{s}{\sigma} + \frac{\sigma^2 + (\mu-m)^2}{2s^2} - \frac{1}{2}. \qquad (1.233)$$

1.31

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

$$\mathrm{H}(\mathbf{x}) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x},$$

$$\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \qquad (1.234)$$

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

The expressions of $\mathrm{H}(\mathbf{x})$ and $\mathrm{H}(\mathbf{y})$ can be written as

$$-\int \left(\int p(\mathbf{x}, \mathbf{y})d\mathbf{y}\right) \ln p(\mathbf{x})d\mathbf{x},$$

$$-\int \left(\int p(\mathbf{x}, \mathbf{y})d\mathbf{x}\right) \ln p(\mathbf{y})d\mathbf{y}. \qquad (1.235)$$

Then,

$$\mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{x}, \mathbf{y}) = -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y}. \qquad (1.236)$$

We have

$$\iint p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = 1. \qquad (1.237)$$

Then, by the Jensen's inequality,

$$-\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \geq - \ln \left(\iint p(\mathbf{x})p(\mathbf{y}) d\mathbf{x} d\mathbf{y} \right). \qquad (1.238)$$

The right hand side can be written as

$$-\ln \left(\int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{y}) d\mathbf{y} \right) = 0. \qquad (1.239)$$

Therefore,

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) \leq \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}). \qquad (1.240)$$

1.32

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad (1.241)$$

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

$$\int p_x(\mathbf{x}) d\mathbf{x} = \int p_x \left(\mathbf{A}^{-1} \mathbf{y} \right) \left| \det \mathbf{A}^{-1} \right| d\mathbf{y}. \qquad (1.242)$$

Then,

$$p_y(\mathbf{y}) = p_x \left(\mathbf{A}^{-1} \mathbf{y} \right) \left| \det \mathbf{A}^{-1} \right|. \qquad (1.243)$$

We have

$$\mathrm{H}(\mathbf{y}) = -\int p_y(\mathbf{y}) \ln p_y(\mathbf{y}) d\mathbf{y}. \qquad (1.244)$$

The right hand side can be written as

$$\begin{aligned}
&-\int p_y(\mathbf{y}) \ln \left(p_x \left(\mathbf{A}^{-1} \mathbf{y} \right) \left| \det \mathbf{A}^{-1} \right| \right) d\mathbf{y} \\
&= -\int p_y(\mathbf{y}) \ln p_x \left(\mathbf{A}^{-1} \mathbf{y} \right) d\mathbf{y} + \ln |\det \mathbf{A}| \int p_y(\mathbf{y}) d\mathbf{y}.
\end{aligned} \qquad (1.245)$$

31

The first term can be written as

$$-\left|\det \mathbf{A}^{-1}\right| \int p_x\left(\mathbf{A}^{-1}\mathbf{y}\right) \ln p_x\left(\mathbf{A}^{-1}\mathbf{y}\right) d\mathbf{y}. \qquad (1.246)$$

By the transformation

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}, \qquad (1.247)$$

it can be written as

$$-\int p_x(\mathbf{x}) \ln p_x(\mathbf{x}) d\mathbf{x} = \mathrm{H}(\mathbf{x}). \qquad (1.248)$$

Therefore,

$$\mathrm{H}(\mathbf{y}) = \mathrm{H}(\mathbf{x}) + \ln |\det \mathbf{A}|. \qquad (1.249)$$

1.33

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

$$\mathrm{H}(y|x) = -\sum_{k=1}^{K}\sum_{l=1}^{L} p(x_k, y_l) \ln p(y_l|x_k). \qquad (1.250)$$

If

$$\mathrm{H}(y|x) = 0, \qquad (1.251)$$

then

$$0 = -\sum_{k=1}^{K} p(x_k)\left(\sum_{l=1}^{L} p(y_l|x_k) \ln p(y_l|x_k)\right). \qquad (1.252)$$

We have

$$p(x_k) \geq 0,$$
$$p(y_l|x_k) \ln p(y_l|x_k) \leq 0. \qquad (1.253)$$

Then, the equation reduces to

$$p(y_l|x_k) \ln p(y_l|x_k) = 0. \qquad (1.254)$$

Then,

$$p(y_l|x_k) = \begin{cases} 1, \\ 0. \end{cases} \qquad (1.255)$$

32

We have

$$\sum_{l=1}^{L} p(y_l|x_k) = 1. \tag{1.256}$$

Therefore,

$$p(y_l|x_k) = \begin{cases} 1, & \text{if } K = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1.257}$$

1.34

Let x be a variable. We have

$$H(x) = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx. \tag{1.258}$$

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

$$\int_{-\infty}^{\infty} p(x) dx = 1,$$
$$\int_{-\infty}^{\infty} x p(x) dx = \mu, \tag{1.259}$$
$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2,$$

let

$$L(p) = H(x) + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) dx - 1 \right) + \lambda_2 \left(\int_{-\infty}^{\infty} x p(x) dx - \mu \right) \\ + \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx - \sigma^2 \right). \tag{1.260}$$

Setting the variation with respect to p to zero gives

$$0 = -\ln p - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2. \tag{1.261}$$

Then,

$$p(x) = \exp\left(-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 \right), \tag{1.262}$$

so that

$$p(x) = c \exp\left(-q(x) \right), \tag{1.263}$$

where

$$c = \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right),$$
$$q(x) = -\lambda_3\left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2. \tag{1.264}$$

Substituting it to the constraints gives

$$c\int_{-\infty}^{\infty} \exp\left(-q(x)\right) dx = 1,$$
$$c\int_{-\infty}^{\infty} x\exp\left(-q(x)\right) dx = \mu, \tag{1.265}$$
$$c\int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(-q(x)\right) dx = \sigma^2.$$

By the transformation

$$y^2 = q(x), \tag{1.266}$$

they can be written as

$$c\int_{-\infty}^{\infty} \exp\left(-y^2\right)(-\lambda_3)^{-\frac{1}{2}} dy = 1,$$
$$c\int_{-\infty}^{\infty} \left((-\lambda_3)^{-\frac{1}{2}}y + \mu - \frac{\lambda_2}{2\lambda_3}\right)\exp\left(-y^2\right)(-\lambda_3)^{-\frac{1}{2}} dy = \mu, \tag{1.267}$$
$$c\int_{-\infty}^{\infty} \left((-\lambda_3)^{-\frac{1}{2}}y - \frac{\lambda_2}{2\lambda_3}\right)^2\exp\left(-y^2\right)(-\lambda_3)^{-\frac{1}{2}} dy = \sigma^2.$$

We have

$$\int_{-\infty}^{\infty} \exp\left(-y^2\right) dy = \Gamma\left(\frac{1}{2}\right),$$
$$\int_{-\infty}^{\infty} y\exp\left(-y^2\right) dy = 0, \tag{1.268}$$
$$\int_{-\infty}^{\infty} y^2\exp\left(-y^2\right) dy = \Gamma\left(\frac{3}{2}\right),$$

Then,

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

$$c\left(\mu - \frac{\lambda_2}{2\lambda_3}\right)(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right) = \mu, \tag{1.269}$$

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

Then,

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

$$\lambda_2 = 0, \tag{1.270}$$

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

Therefore,

$$p(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.271}$$

1.35

Let x be a variable such that

$$p(x) = \mathcal{N}\left(x|\mu,\sigma^2\right). \tag{1.272}$$

Then,

$$\mathrm{H}(x) = -\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\ln\mathcal{N}\left(x|\mu,\sigma^2\right)dx. \tag{1.273}$$

The right hand side can be written as

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

$$=\frac{1}{2}\ln\left(2\pi\sigma^2\right)\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)dx \tag{1.274}$$

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

Therefore,

$$\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln\left(2\pi\sigma^2\right)\right). \tag{1.275}$$

1.36 (Incomplete)

Let f be a strictly convex function. Then,

$$f\left(\lambda a + (1-\lambda)b\right) \le \lambda f(a) + (1-\lambda)f(b), \tag{1.276}$$

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

$$x = \lambda a + (1-\lambda)b. \tag{1.277}$$

Then, the inequality can be written as

$$f(x) \le \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b). \tag{1.278}$$

Let

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

Then,

$$g(x) \ge 0. \tag{1.280}$$

Additionally, for $x > a$,

$$g(x) = (x-a)\left(\frac{f(b)-f(a)}{b-a} - \frac{f(x)-f(a)}{x-a}\right). \tag{1.281}$$

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

$$\begin{aligned} f'(c) &= \frac{f(b)-f(a)}{b-a}, \\ f'(y) &= \frac{f(x)-f(a)}{x-a}. \end{aligned} \tag{1.282}$$

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

$$f'(y) \le f'(c). \tag{1.283}$$

1.37

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

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) = -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \tag{1.284}$$

The right hand side can be written as

$$- \iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x}) \right) d\mathbf{x} d\mathbf{y}$$
$$= - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x}. \tag{1.285}$$

We have

$$H(\mathbf{y}|\mathbf{x}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y},$$
$$H(\mathbf{x}) = - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \tag{1.286}$$

Therefore,

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \tag{1.287}$$

1.38

Let f be a strictly convex function. Then,

$$f\left(\lambda x_1 + (1 - \lambda)x_2\right) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \tag{1.288}$$

where

$$0 \leq \lambda \leq 1. \tag{1.289}$$

Let us assume that

$$f\left(\sum_{m=1}^{M} \lambda_m x_m\right) \leq \sum_{m=1}^{M} \lambda_m f(x_m), \tag{1.290}$$

where

$$\sum_{m=1}^{M} \lambda_m = 1,$$
$$\lambda_m \geq 0. \tag{1.291}$$

We have

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

37

where

$$\sum_{m=1}^{M+1} \lambda_m = 1,$$ (1.293)

$$\lambda_m \geq 0.$$

By the assumption,

$$f\left(\sum_{m=1}^{M} \frac{\lambda_m}{1 - \lambda_{M+1}} x_m\right) \leq \sum_{m=1}^{M} \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m).$$ (1.294)

Then,

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{m=1}^{M} \frac{\lambda_m}{1 - \lambda_{M+1}} f(x_m),$$ (1.295)

so that

$$f\left(\sum_{m=1}^{M+1} \lambda_m x_m\right) \leq \sum_{m=1}^{M+1} \lambda_m f(x_m).$$ (1.296)

Therefore, the assumption is proved by induction on M.

1.39

Let x and y be two binary variables where

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

(a)

We have

$$H(x) = -\sum_x p(x) \ln p(x).$$ (1.298)

38

We have
$$p(x = 0) = p(x = 0, y = 0) + p(x = 0, y = 1),$$
$$p(x = 1) = p(x = 1, y = 0) + p(x = 1, y = 1). \tag{1.299}$$

Then,
$$p(x = 0) = \frac{2}{3},$$
$$p(x = 1) = \frac{1}{3}. \tag{1.300}$$

Therefore,
$$H(x) = \ln 3 - \frac{2}{3} \ln 2. \tag{1.301}$$

(b)

We have
$$H(y) = -\sum_y p(y) \ln p(y). \tag{1.302}$$

We have
$$p(y = 0) = p(x = 0, y = 0) + p(x = 1, y = 0),$$
$$p(y = 1) = p(x = 0, y = 1) + p(x = 1, y = 1). \tag{1.303}$$

Then,
$$p(y = 0) = \frac{1}{3},$$
$$p(y = 1) = \frac{2}{3}. \tag{1.304}$$

Therefore,
$$H(y) = \ln 3 - \frac{2}{3} \ln 2. \tag{1.305}$$

(c)

We have
$$H(y|x) = -\sum_{x,y} p(x, y) \ln p(y|x). \tag{1.306}$$

39

By the Bayes' theorem,

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

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

$$p(y = 1 | x = 0) = \frac{p(x = 0, y = 1)}{p(x = 0)},$$ (1.307)

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

Then,

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

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

$$p(y = 1 | x = 0) = \frac{1}{2},$$ (1.308)

$$p(y = 1 | x = 1) = 1.$$

Therefore,

$$\mathrm{H}(y | x) = \frac{2}{3} \ln 2.$$ (1.309)

(d)

We have

$$\mathrm{H}(x | y) = -\sum_{x,y} p(x, y) \ln p(x | y).$$ (1.310)

By the Bayes' theorem,

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

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

$$p(x = 1 | y = 0) = \frac{p(x = 1, y = 0)}{p(y = 0)},$$ (1.311)

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

40

Then,

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

(1.312)

Therefore,

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

(1.313)

(e)

We have

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

(1.314)

Therefore,

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

(1.315)

(f)

We have

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

(1.316)

The right hand side can be written as

$$H(x) + H(y) - H(x, y).$$

(1.317)

Therefore,

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

(1.318)

1.40

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

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

(1.319)

By the Jensen's inequality,

$$\sum_{m=1}^{M} \lambda_m \ln x_m \leq \ln \left(\sum_{m=1}^{M} \lambda_m x_m \right), \tag{1.320}$$

so that

$$\prod_{m=1}^{M} x_m^{\lambda_m} \leq \sum_{m=1}^{M} \lambda_m x_m. \tag{1.321}$$

Substituting

$$\lambda_m = \frac{1}{M} \tag{1.322}$$

to the inequality gives

$$\left(\prod_{m=1}^{M} x_m \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{m=1}^{M} x_m. \tag{1.323}$$

1.41

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

(a)

We have

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = - \iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y}. \tag{1.324}$$

The right hand side can be written as

$$- \iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{x}) + \ln \left(\frac{p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \right) d\mathbf{x} d\mathbf{y}$$
$$= - \iint p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{x}) - \ln p(\mathbf{x}|\mathbf{y}) \right) d\mathbf{x} d\mathbf{y}. \tag{1.325}$$

The right hand side can be written as

$$- \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) \ln p(\mathbf{x}) d\mathbf{x} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y}. \tag{1.326}$$

Therefore,

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

(b)

We have

$$I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x}). \tag{1.328}$$

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

$$H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \tag{1.329}$$

Therefore,

$$I(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}). \tag{1.330}$$

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

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

Therefore,

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

(b)

We have

$$\mathrm{E}\,x = \sum_x x p(x),$$
$$\mathrm{E}\,x^2 = \sum_x x^2 p(x), \tag{2.4}$$

Then,

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

We have

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

Therefore,

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

(c)

We have

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

44

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

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

Therefore,

$$\sum_x p(x) = 1. \tag{2.12}$$

(b)

We have

$$\mathrm{E}\, x = \sum_x x p(x),$$

$$\mathrm{E}\, x^2 = \sum_x x^2 p(x). \tag{2.13}$$

The right hand sides can be written as

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

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

Then,

$$\mathrm{E}\, x = \mu,$$

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

We have

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

45

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

(c)

We have
$$\mathrm{H}(x) = -\sum_x p(x|\mu) \ln p(x|\mu). \tag{2.18}$$

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)

We have
$$\binom{N}{n} = \frac{N!}{n!(N-n)!},$$
$$\binom{N}{n-1} = \frac{N!}{(n-1)!(N-n+1)!} \tag{2.20}$$

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

(b)

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

46

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,

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

47

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

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

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,

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

We have

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

Therefore,

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

49

2.5

We have

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

By the transformation

$$t = x + y, \tag{2.50}$$

the right hand side can be written as

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

By the transformation

$$x = t\mu, \tag{2.52}$$

the right hand side can be written as

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

Then,

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

Therefore,

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

2.6

Let μ be a variable such that

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

(a)

We have

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

$$\mathrm{E}\,\mu^2 = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu. \tag{2.57}$$

By 2.5,

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

$$\int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu = \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}. \tag{2.58}$$

Therefore,

$$\mathrm{E}\,\mu = \frac{a}{a+b},$$

$$\mathrm{E}\,\mu^2 = \frac{a(a+1)}{(a+b)(a+b+1)}. \tag{2.59}$$

(b)

We have

$$\mathrm{var}\,\mu = \mathrm{E}\,\mu^2 - (\mathrm{E}\mu)^2. \tag{2.60}$$

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

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

Therefore,

$$\mathrm{var}\,\mu = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.62}$$

(c)

We have

$$\ln p(\mu) = \ln \Gamma(a+b) - \ln \Gamma(a) - \ln \Gamma(b) + (a-1)\ln \mu + (b-1)\ln(1-\mu). \tag{2.63}$$

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

$$0 = \ln\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)\left(\frac{a-1}{\mu} - \frac{b-1}{1-\mu}\right). \tag{2.64}$$

Therefore,
$$\text{mode}\,\mu = \begin{cases} \frac{a-1}{a+b-2}, & \text{if } a \geq 1 \text{and } b \geq 1, \\ \text{none}, & \text{otherwise.} \end{cases} \tag{2.65}$$

2.7

Let m and l be variables such that
$$p(m, l | \mu) = \binom{m+l}{m} \mu^m (1 - \mu)^l,$$
$$p(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}. \tag{2.66}$$

By 2.6,
$$\mathrm{E}\,\mu = \frac{a}{a+b}. \tag{2.67}$$

Setting the derivative of $p(m, l | \mu)$ with respect to μ to zero gives
$$0 = \binom{m+l}{m} \mu^m (1-\mu)^l \left(\frac{m}{\mu} + \frac{l}{1-\mu} \right). \tag{2.68}$$

Then, the maximum likelihood solution for μ is given by
$$\mu_{\mathrm{ML}} = \frac{m}{m+l}. \tag{2.69}$$

By the Bayes' thereorem,
$$p(\mu|m, l) = \frac{p(m, l|\mu)p(\mu)}{p(m, l)}. \tag{2.70}$$

By 2.5, the right hand side can be written as
$$p(\mu|m, l) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1} (1-\mu)^{l+b-1}. \tag{2.71}$$

By 2.6,
$$\mathrm{E}(\mu|m, l) = \frac{m+a}{m+l+a+b}. \tag{2.72}$$

Therefore,
$$\mathrm{E}(\mu|m, l) = \lambda \mu_{\mathrm{ML}} + (1 - \lambda)\,\mathrm{E}\,\mu, \tag{2.73}$$

where
$$\lambda = \frac{m+l}{m+l+a+b}. \tag{2.74}$$

2.8

Let x and y be variables.

(a)

We have

$$\mathrm{E}\,x = \int xp(x)dx. \tag{2.75}$$

The right hand side can be written as

$$\int x \left(\int p(x, y)dy \right) dx = \int \left(\int xp(x|y)dx \right) p(y)dy. \tag{2.76}$$

Therefore,

$$\mathrm{E}\,x = \mathrm{E}\left(\mathrm{E}(x|y) \right). \tag{2.77}$$

(b)

We have

$$\mathrm{var}\,x = \mathrm{E}\left(x - \mathrm{E}\,x \right)^2. \tag{2.78}$$

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

$$\begin{aligned} &\mathrm{E}\left(\mathrm{E}\left((x - \mathrm{E}\,x)^2 | y \right) \right) \\ &= \mathrm{E}\left(\mathrm{E}\left((f(x, y) + g(y))^2 | y \right) \right), \end{aligned} \tag{2.79}$$

where

$$\begin{aligned} f(x, y) &= x - \mathrm{E}(x|y), \\ g(y) &= \mathrm{E}(x|y) - \mathrm{E}\,x. \end{aligned} \tag{2.80}$$

The right hand side can be written as $I_1 + 2I_2 + I_3$, where

$$\begin{aligned} I_1 &= \mathrm{E}\left(\mathrm{E}\left(f^2(x, y)|y \right) \right), \\ I_2 &= \mathrm{E}\left(\mathrm{E}\left(f(x, y)g(y)|y \right) \right), \\ I_3 &= \mathrm{E}\left(g^2(y) \right). \end{aligned} \tag{2.81}$$

We have

$$\mathrm{var}(x|y) = \mathrm{E}\left(f^2(x, y)|y \right). \tag{2.82}$$

Then, the expression of I_1 can be written as

$$\mathrm{E}\left(\mathrm{var}(x|y) \right). \tag{2.83}$$

We have
$$\mathrm{E}\left(f(x,y)|y\right) = 0. \tag{2.84}$$
Then, the expression of I_2 can be written as
$$\mathrm{E}\left(\mathrm{E}\left(f(x,y)|y\right)g(y)\right) = 0. \tag{2.85}$$
By (a),
$$\mathrm{E}\left(\mathrm{E}(x|y)\right) = \mathrm{E}\,x. \tag{2.86}$$
Then, the expression of I_3 can be written as
$$\mathrm{var}\left(\mathrm{E}(x|y)\right). \tag{2.87}$$
Therefore,
$$\mathrm{var}\,x = \mathrm{E}\left(\mathrm{var}(x|y)\right) + \mathrm{var}\left(\mathrm{E}(x|y)\right). \tag{2.88}$$

2.9

For a vector $\boldsymbol{\mu}$ in 2 dimensions, by 2.5,
$$\int_{\substack{\mu_1+\mu_2=1 \\ \mu_1 \geq 0, \mu_2 \geq 0}} \mu_1^{\alpha_1-1}\mu_2^{\alpha_2-1}d\boldsymbol{\mu} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}. \tag{2.89}$$

For a vector $\boldsymbol{\mu}$ in K dimensions, let us assume that
$$\int_{\substack{\sum_{k=1}^{K}\mu_k=1 \\ \mu_k \geq 0}} \prod_{k=1}^{K} \mu_k^{\alpha_k-1}d\boldsymbol{\mu} = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}. \tag{2.90}$$

For a vector $\boldsymbol{\mu}$ in $K+1$ dimensions, let
$$\boldsymbol{\mu}' = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{K-1} \\ \mu_K + \mu_{K+1} \end{bmatrix}. \tag{2.91}$$

Then,
$$\begin{aligned} &\int_{\substack{\sum_{k=1}^{K+1}\mu_k=1 \\ \mu_k \geq 0}} \prod_{k=1}^{K+1} \mu_k^{\alpha_k-1}d\boldsymbol{\mu} \\ &= \int_{\substack{\sum_{k=1}^{K}\mu'_k=1 \\ \mu'_k \geq 0}} \left(\prod_{k=1}^{K-1} \mu_k'^{\alpha_k-1}\right) I(\mu'_K)d\boldsymbol{\mu}', \end{aligned} \tag{2.92}$$

54

where
$$I(\mu'_K) = \int_0^{\mu'_K} (\mu'_K - \mu_{K+1})^{\alpha_K - 1} \mu_{K+1}^{\alpha_{K+1} - 1} d\mu_{K+1}. \tag{2.93}$$

By the transformation
$$x = \frac{\mu_{K+1}}{\mu'_K}, \tag{2.94}$$

the expression of $I(\mu'_K)$ can be written as

$$\int_0^1 (\mu'_K(1-x))^{\alpha_K - 1} (\mu'_K x)^{\alpha_{K+1} - 1} \mu'_K dx$$
$$= \mu'^{\alpha_K + \alpha_{K+1} - 1}_K \int_0^1 x^{\alpha_{K+1} - 1} (1-x)^{\alpha_K - 1} dx. \tag{2.95}$$

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

$$\frac{\Gamma(\alpha_K)\Gamma(\alpha_{K+1})}{\Gamma(\alpha_K + \alpha_{K+1})}. \tag{2.96}$$

Then,

$$\int_{\substack{\sum_{k=1}^{K+1} \mu_k = 1 \\ \mu_k \geq 0}} \prod_{k=1}^{K+1} \mu_k^{\alpha_k - 1} d\boldsymbol{\mu}$$
$$= \frac{\Gamma(\alpha_K)\Gamma(\alpha_{K+1})}{\Gamma(\alpha_K + \alpha_{K+1})} \int_{\substack{\sum_{k=1}^{K} \mu'_k = 1 \\ \mu'_k \geq 0}} \left(\prod_{k=1}^{K-1} \mu'^{\alpha_k - 1}_k \right) \mu'^{\alpha_K + \alpha_{K+1} - 1}_K d\boldsymbol{\mu}'. \tag{2.97}$$

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

$$\frac{(\prod_{k=1}^{K-1} \Gamma(\alpha_k))\Gamma(\alpha_K + \alpha_{K+1})}{\Gamma(\sum_{k=1}^{K+1} \alpha_k)}. \tag{2.98}$$

Then,

$$\int_{\substack{\sum_{k=1}^{K+1} \mu_k = 1 \\ \mu_k \geq 0}} \prod_{k=1}^{K+1} \mu_k^{\alpha_k - 1} d\boldsymbol{\mu} = \frac{\prod_{k=1}^{K+1} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K+1} \alpha_k)}. \tag{2.99}$$

Therefore, the assumption is proved by induction on K.

2.10

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

$$p(\boldsymbol{\mu}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}. \tag{2.100}$$

We have

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

$$\mathrm{E}\,\mu_k^2 = \int \mu_k^2 p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \tag{2.101}$$

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

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

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

$$\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_k+2)}{\Gamma(\alpha_k)} \prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k + 2\right)} = \frac{\alpha_k(\alpha_k + 1)}{\sum_{k=1}^{K} \alpha_k(\sum_{k=1}^{K} \alpha_k + 1)}, \tag{2.102}$$

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

Then,

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

$$\mathrm{E}\,\mu_k^2 = \frac{\alpha_k(\alpha_k + 1)}{\sum_{k=1}^{K} \alpha_k \left(\sum_{k=1}^{K} \alpha_k + 1\right)}, \tag{2.103}$$

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

We have

$$\mathrm{var}\,\mu_k = \mathrm{E}\,\mu_k^2 - \left(\mathrm{E}\,\mu_k\right)^2,$$

$$\mathrm{cov}\,(\mu_k, \mu_{k'}) = \mathrm{E}\,\mu_k \mu_{k'} - \mathrm{E}\,\mu_k \,\mathrm{E}\,\mu_{k'}. \tag{2.104}$$

Therefore,

$$\operatorname{var} \mu_k = \frac{\alpha_k \left(\left(\sum_{k=1}^{K} \alpha_k \right) - \alpha_k \right)}{(\sum_{k=1}^{K} \alpha_k)^2 (\sum_{k=1}^{K} \alpha_k + 1)}, \qquad (2.105)$$
$$\operatorname{cov}(\mu_k, \mu_{k'}) = -\frac{\alpha_k \alpha_{k'}}{(\sum_{k=1}^{K} \alpha_k)^2 (\sum_{k=1}^{K} \alpha_k + 1)}.$$

2.11

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

$$p(\boldsymbol{\mu}) = \frac{\Gamma \left(\sum_{k=1}^{K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}. \qquad (2.106)$$

We have

$$\mathrm{E} \ln \mu_k = \int (\ln \mu_k) \, p(\boldsymbol{\mu}) d\boldsymbol{\mu}. \qquad (2.107)$$

We have

$$\frac{\partial}{\partial \alpha_k} p(\boldsymbol{\mu}) = \left(\frac{\Gamma' \left(\sum_{k=1}^{K} \alpha_k \right)}{\Gamma \left(\sum_{k=1}^{K} \alpha_k \right)} - \frac{\Gamma'(\alpha_k)}{\Gamma(\alpha_k)} + \ln \mu_k \right) p(\boldsymbol{\mu}). \qquad (2.108)$$

Then,

$$\mathrm{E} \ln \mu_k = \frac{\partial}{\partial \alpha_k} \int p(\boldsymbol{\mu}) d\boldsymbol{\mu} + \left(\psi(\alpha_k) - \psi \left(\sum_{k=1}^{K} \alpha_k \right) \right) \int p(\boldsymbol{\mu}) d\boldsymbol{\mu}, \qquad (2.109)$$

where

$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \qquad (2.110)$$

Therefore,

$$\mathrm{E} \ln \mu_k = \psi(\alpha_k) - \psi \left(\sum_{k=1}^{K} \alpha_k \right). \qquad (2.111)$$

2.12

Let x be a variable such that

$$p(x) = \frac{1}{b-a}, \tag{2.112}$$

where $a < b$. Then

$$\int_a^b p(x)dx = 1. \tag{2.113}$$

We have

$$\operatorname{E} x = \frac{1}{b-a} \int_a^b x dx,$$
$$\operatorname{E} x^2 = \frac{1}{b-a} \int_a^b x^2 dx. \tag{2.114}$$

Then,

$$\operatorname{E} x = \frac{1}{2}(a+b),$$
$$\operatorname{E} x^2 = \frac{1}{3}\left(a^2 + ab + b^2\right). \tag{2.115}$$

We have

$$\operatorname{var} x = \operatorname{E} x^2 - (\operatorname{E} x)^2. \tag{2.116}$$

Therefore,

$$\operatorname{var} x = \frac{1}{12}(b-a)^2. \tag{2.117}$$

2.13

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$
$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}). \tag{2.118}$$

We have

$$\operatorname{KL}(p||q) = -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}. \tag{2.119}$$

The logarithm in the right hand side can be written as

$$\ln \frac{(2\pi)^{-\frac{D}{2}} |\det \mathbf{L}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right)}{(2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$
$$= \frac{1}{2}\ln\left|\frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}}\right| + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}). \tag{2.120}$$

Then, the integral can be written as

$$\frac{1}{2}\ln\left|\frac{\det \boldsymbol{\Sigma}}{\det \mathbf{L}}\right| + \frac{1}{2}I_1 - \frac{1}{2}I_2, \tag{2.121}$$

where

$$\begin{aligned}
I_1 &= \int (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x}, \\
I_2 &= \int (\mathbf{x} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\mathbf{x} - \mathbf{m})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x}.
\end{aligned} \tag{2.122}$$

We have

$$(\mathbf{x} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\mathbf{x} - \mathbf{m}) = (\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \mathbf{m}). \tag{2.123}$$

Then, the expression of I_2 can be written as

$$\begin{aligned}
&\int (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{L}^{-1}(\mathbf{x} - \boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} \\
&+ 2(\boldsymbol{\mu} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}\int (\mathbf{x} - \boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} \\
&+ (\boldsymbol{\mu} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m})\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} \\
=& I_2' + (\boldsymbol{\mu} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m}).
\end{aligned} \tag{2.124}$$

We have

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} = \boldsymbol{\Sigma}, \tag{2.125}$$

so that

$$\int (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} = \operatorname{tr}\boldsymbol{\Sigma}. \tag{2.126}$$

Then,

$$\begin{aligned}
I_1 &= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right), \\
I_2' &= \operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right).
\end{aligned} \tag{2.127}$$

Therefore,

$$\begin{aligned}
&\mathrm{KL}(p||q) \\
=& \frac{1}{2}\left(\ln\left|\frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}}\right| - D + \operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right) + (\boldsymbol{\mu} - \mathbf{m})^\mathsf{T}\mathbf{L}^{-1}(\boldsymbol{\mu} - \mathbf{m})\right).
\end{aligned} \tag{2.128}$$

59

2.14

Let $\mathbf{x}$ be a variable in D dimensions. We have

$$\mathrm{H}(\mathbf{x}) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \tag{2.129}$$

In order to maximise $\mathrm{H}(x)$ with the constratints

$$\int p(\mathbf{x}) d\mathbf{x} = 1,$$
$$\int \mathbf{x} p(\mathbf{x}) d\mathbf{x} = \boldsymbol{\mu}, \tag{2.130}$$
$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x} = \boldsymbol{\Sigma},$$

let

$$\begin{aligned} L(p) =&\mathrm{H}(\mathbf{x}) + \lambda\left(\int p(\mathbf{x}) d\mathbf{x} - 1\right) + \mathbf{l}^{\mathsf{T}}\left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\mu}\right) \\ &+ \mathbf{m}^{\mathsf{T}}\left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\Sigma}\right)\mathbf{m}. \end{aligned} \tag{2.131}$$

Setting the variation with respect to p to zero gives

$$0 = -\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^{\mathsf{T}}\mathbf{x} + \mathbf{m}^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{m}. \tag{2.132}$$

Then,

$$p(\mathbf{x}) = \exp\left(-1 + \lambda + \mathbf{l}^{\mathsf{T}}\mathbf{x} + \mathbf{m}^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{m}\right), \tag{2.133}$$

so that

$$p(\mathbf{x}) = c \exp\left(-(\mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml})^{\mathsf{T}}\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml})\right), \tag{2.134}$$

where

$$\begin{aligned} c &= \exp\left(-1 + \lambda - \mathbf{l}^{\mathsf{T}}\mathbf{Ml}\right), \\ \mathbf{M} &= -\left(\mathbf{mm}^{\mathsf{T}}\right)^{-1}. \end{aligned} \tag{2.135}$$

Substituting it to the constraints and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{Ml} \tag{2.136}$$

gives

$$c \int \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} = 1,$$

$$c \int \left(\mathbf{y} + \boldsymbol{\mu} + \mathbf{Ml}\right) \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} = \boldsymbol{\mu}, \qquad (2.137)$$

$$c \int \left(\mathbf{y} + \mathbf{Ml}\right) \left(\mathbf{y} + \mathbf{Ml}\right)^\mathsf{T} \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} = \boldsymbol{\Sigma}.$$

We have

$$\int \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{y}\right) d\mathbf{y} = \left(\Gamma\left(\frac{1}{2}\right)\right)^D,$$

$$\int \mathbf{y} \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{y}\right) d\mathbf{y} = \mathbf{0}, \qquad (2.138)$$

$$\int \mathbf{y}\mathbf{y}^\mathsf{T} \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{y}\right) d\mathbf{y} = \Gamma\left(\frac{3}{2}\right) \left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1} \mathbf{I},$$

so that

$$\int \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} = \pi^{\frac{D}{2}} |\det \mathbf{M}|^{\frac{1}{2}},$$

$$\int \mathbf{y} \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} = \mathbf{0}, \qquad (2.139)$$

$$\int \mathbf{y}\mathbf{y}^\mathsf{T} \exp\left(-\mathbf{y}^\mathsf{T} \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} = \frac{1}{2} \pi^{\frac{D}{2}} (\det \mathbf{M})^{\frac{1}{2}} \mathbf{M},$$

Then,

$$c\pi^{\frac{D}{2}} |\det \mathbf{M}|^{\frac{1}{2}} = 1,$$

$$c\pi^{\frac{D}{2}} |\det \mathbf{M}|^{\frac{1}{2}} (\boldsymbol{\mu} + \mathbf{Ml}) = \boldsymbol{\mu}, \qquad (2.140)$$

$$c\pi^{\frac{D}{2}} |\det \mathbf{M}|^{\frac{1}{2}} \left(\frac{1}{2}\mathbf{M} + \mathbf{Ml}(\mathbf{Ml})^\mathsf{T}\right) = \boldsymbol{\Sigma}.$$

Then,

$$c = (2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}},$$

$$\mathbf{l} = \mathbf{0}, \qquad (2.141)$$

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

Therefore,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}} |\det \boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \qquad (2.142)$$

2.15

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.143}$$

Then, by the definition, the entropy is given by

$$\mathrm{H}(\mathbf{x}) = - \int \mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) d\mathbf{x}. \tag{2.144}$$

The right hand side can be written as

$$- \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \left(-\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln|\det \boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right) d\mathbf{x}$$
$$= \left(\frac{D}{2} \ln(2\pi) + \frac{1}{2} \ln|\det \boldsymbol{\Sigma}| \right) \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$$
$$+ \frac{1}{2} \int (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \tag{2.145}$$

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

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \boldsymbol{\Sigma}, \tag{2.146}$$

we have

$$\int (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = \mathrm{tr}\, \boldsymbol{\Sigma}. \tag{2.147}$$

Then, the second integral can be written as

$$\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \right) = D. \tag{2.148}$$

Therefore,

$$\mathrm{H}(\mathbf{x}) = \frac{D}{2} \left(1 + \ln(2\pi) \right) + \frac{1}{2} \ln|\det \boldsymbol{\Sigma}|. \tag{2.149}$$

2.16

Let x be a variable such that

$$x = x_1 + x_2, \tag{2.150}$$

where
$$p(x_1) = \mathcal{N}\left(x_1|\mu_1, \tau_1^{-1}\right),$$
$$p(x_2) = \mathcal{N}\left(x_2|\mu_2, \tau_2^{-1}\right). \tag{2.151}$$

By marginalisation,
$$p(x) = \int_{-\infty}^{\infty} p(x|x_2)p(x_2)dx_2. \tag{2.152}$$

The right hand side can be written as
$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu_1 + x_2, \tau_1^{-1}\right)\mathcal{N}\left(x_2|\mu_2, \tau_2^{-1}\right)dx_2$$
$$= \int_{-\infty}^{\infty} \left(\frac{\tau_1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_1}{2}(x-\mu_1-x_2)^2\right)\left(\frac{\tau_2}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_2}{2}(x_2-\mu_2)^2\right)dx_2. \tag{2.153}$$

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

$$-\frac{\tau_1+\tau_2}{2}\left(x_2 - \frac{\tau_1(x-\mu_1)+\tau_2\mu_2}{\tau_1+\tau_2}\right)^2 - \frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2$$
$$+ \frac{\tau_1+\tau_2}{2}\left(\frac{\tau_1(x-\mu_1)+\tau_2\mu_2}{\tau_1+\tau_2}\right)^2$$
$$= -\frac{\tau_1+\tau_2}{2}\left(x_2 - \frac{\tau_1(x-\mu_1)+\tau_2\mu_2}{\tau_1+\tau_2}\right)^2 - \frac{\tau_1\tau_2}{2(\tau_1+\tau_2)}(x-\mu_1-\mu_2)^2. \tag{2.154}$$

Then,
$$p(x) = \mathcal{N}\left(x|\mu_1+\mu_2, \tau_1^{-1}+\tau_2^{-1}\right). \tag{2.155}$$

Therefore, by 1.35,
$$\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln(2\pi) + \ln\left(\tau_1^{-1}+\tau_2^{-1}\right)\right). \tag{2.156}$$

2.17

Let $\boldsymbol{\Sigma}$ be a matrix and
$$\mathbf{S} = \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right),$$
$$\mathbf{A} = \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} - \left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}\right). \tag{2.157}$$

Then,
$$\boldsymbol{\Sigma}^{-1} = \mathbf{S} + \mathbf{A}, \tag{2.158}$$

so that
$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.159}$$

The second term of the right hand side can be written as
$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.160}$$

The second term of the right hand side can be written as
$$-\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.161}$$

Then,
$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = 0. \tag{2.162}$$

Therefore,
$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.163}$$

2.18

(a)

Let $\boldsymbol{\Sigma}$ be a $D \times D$ real symmetric matrix such that
$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.164}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_D$ are unit vectors. Then,
$$\overline{\mathbf{u}_d}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d, \tag{2.165}$$

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
$$\overline{\mathbf{u}_d}^{\mathsf{T}}\overline{\boldsymbol{\Sigma}}^{\mathsf{T}}\mathbf{u}_d = \left(\overline{\boldsymbol{\Sigma}\mathbf{u}_d}\right)^{\mathsf{T}}\mathbf{u}_d. \tag{2.166}$$

The right hand side can be writtet as
$$\overline{\lambda_d}\overline{\mathbf{u}_d}^{\mathsf{T}}\mathbf{u}_d = \overline{\lambda_d}. \tag{2.167}$$

Therefore,
$$\lambda_d = \overline{\lambda_d}. \tag{2.168}$$

(b)

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d \tag{2.169}$$

gives

$$\mathbf{u}_{d'}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_{d'}^{\mathsf{T}}\mathbf{u}_d. \tag{2.170}$$

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

$$\mathbf{u}_{d'}^{\mathsf{T}}\boldsymbol{\Sigma}^{\mathsf{T}}\mathbf{u}_d = (\boldsymbol{\Sigma}\mathbf{u}_{d'})^{\mathsf{T}}\mathbf{u}_d. \tag{2.171}$$

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

$$\lambda_d\mathbf{u}_{d'}^{\mathsf{T}}\mathbf{u}_d = \lambda_{d'}\mathbf{u}_{d'}^{\mathsf{T}}\mathbf{u}_d. \tag{2.172}$$

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

$$\mathbf{u}_{d'}^{\mathsf{T}}\mathbf{u}_d = 0. \tag{2.173}$$

2.19

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.174}$$

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

$$\begin{aligned}
\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \cdots, \lambda_D), \\
\mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].
\end{aligned} \tag{2.175}$$

Then,

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}. \tag{2.176}$$

By 2.18,

$$\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}. \tag{2.177}$$

Then,

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}, \\
\boldsymbol{\Sigma}^{-1} &= \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathsf{T}},
\end{aligned} \tag{2.178}$$

Therefore,

$$\begin{aligned}
\boldsymbol{\Sigma} &= \sum_{d=1}^{D} \lambda_d\mathbf{u}_d\mathbf{u}_d^{\mathsf{T}}, \\
\boldsymbol{\Sigma}^{-1} &= \sum_{d=1}^{D} \frac{1}{\lambda_d}\mathbf{u}_d\mathbf{u}_d^{\mathsf{T}}.
\end{aligned} \tag{2.179}$$

2.20

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d \mathbf{u}_d, \tag{2.180}$$

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

$$\begin{aligned}
\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \cdots, \lambda_D), \\
\mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].
\end{aligned} \tag{2.181}$$

By 2.19,

$$\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{a} = \mathbf{b}^\mathsf{T}\boldsymbol{\Lambda}\mathbf{b}, \tag{2.182}$$

where

$$\mathbf{b} = \mathbf{U}^\mathsf{T}\mathbf{a}. \tag{2.183}$$

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

$$\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{a} > 0 \tag{2.184}$$

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

$$\lambda_d > 0. \tag{2.185}$$

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

$$\boldsymbol{\Sigma}\boldsymbol{\Lambda} = \mathbf{I}. \tag{2.186}$$

Taking the transpose of the both sides gives

$$\boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{\Sigma} = \mathbf{I}. \tag{2.187}$$

Therefore,

$$\boldsymbol{\Lambda}^\mathsf{T} = \boldsymbol{\Lambda}. \tag{2.188}$$

2.23

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d \mathbf{u}_d, \tag{2.189}$$

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

$$\begin{aligned}
\boldsymbol{\Lambda}' &= \text{diag}\left(\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_D^{-\frac{1}{2}}\right), \\
\mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].
\end{aligned} \tag{2.190}$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{U}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}'^\mathsf{T}\mathbf{U}^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \tag{2.191}$$

By the transformation

$$\mathbf{y} = \boldsymbol{\Lambda}'^\mathsf{T}\mathbf{U}^\mathsf{T}(\mathbf{x} - \boldsymbol{\mu}) \tag{2.192}$$

and the property

$$\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}, \tag{2.193}$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left|\det\left(\mathbf{U}\boldsymbol{\Lambda}'^{-1}\right)\right| d\mathbf{y} = |\det\boldsymbol{\Sigma}|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \tag{2.194}$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det\boldsymbol{\Sigma}|^{\frac{1}{2}}\Delta^D V_D, \tag{2.195}$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \tag{2.196}$$

2.24

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

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{O} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}. \tag{2.197}$$

The right hand side can be written as

$$
\begin{bmatrix} \mathbf{I} & \mathbf{BD}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A} - \mathbf{BD}^{-1}\mathbf{C} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}. \tag{2.198}
$$

Then,

$$
\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{M} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{BD}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}, \tag{2.199}
$$

where

$$
\mathbf{M} = \left(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C} \right)^{-1}. \tag{2.200}
$$

Therefore,

$$
\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & -\mathbf{BD}^{-1}\mathbf{M} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{bmatrix}. \tag{2.201}
$$

2.25

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

$$
p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.202}
$$

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

$$
\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \tag{2.203}
$$

where

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

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

$$
\begin{aligned}
&- \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T} \boldsymbol{\Lambda}_{aa} (\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T} \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&- \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\mathsf{T} \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\mathsf{T} \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) \\
&- \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\mathsf{T} \boldsymbol{\Lambda}_{ca}(\mathbf{x}_a - \boldsymbol{\mu}_a).
\end{aligned} \tag{2.204}
$$

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

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\intercal \boldsymbol{\Sigma}_{a|b,c}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}), \tag{2.205}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c), \\
\boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}.
\end{aligned} \tag{2.206}$$

Then,

$$p(\mathbf{x}_a|\mathbf{x}_b, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}). \tag{2.207}$$

By marginalisation,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \int p(\mathbf{x}_a|\mathbf{x}_b, \mathbf{x}_c)p(\mathbf{x}_c)d\mathbf{x}_c. \tag{2.208}$$

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

$$\begin{aligned}
&-\frac{1}{2}\left(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}\right)^\intercal \boldsymbol{\Sigma}_{a|b,c}^{-1}\left(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}\right) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Lambda}_{cc}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\
&= -\frac{1}{2}\mathbf{v}^\intercal \mathbf{M}\mathbf{v},
\end{aligned} \tag{2.209}$$

where

$$\begin{aligned}
\mathbf{v} &= \begin{bmatrix} \mathbf{x}_c - \boldsymbol{\mu}_c \\ \mathbf{x}_a - \boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) \end{bmatrix}, \\
\mathbf{M} &= \begin{bmatrix} \boldsymbol{\Lambda}_{cc} + \boldsymbol{\Lambda}_{ac}^\intercal \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac} & \boldsymbol{\Lambda}_{ac}^\intercal \\ \boldsymbol{\Lambda}_{ac} & \boldsymbol{\Lambda}_{aa} \end{bmatrix}.
\end{aligned} \tag{2.210}$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{cc}^{-1} & -\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^\intercal \boldsymbol{\Lambda}_{aa}^{-1} \\ -\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1} & \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^\intercal \boldsymbol{\Lambda}_{aa}^{-1} \end{bmatrix}. \tag{2.211}$$

Therefore,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \tag{2.212}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right), \\
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\Lambda}_{cc}^{-1}\boldsymbol{\Lambda}_{ac}^\intercal \boldsymbol{\Lambda}_{aa}^{-1}.
\end{aligned} \tag{2.213}$$

69

2.26

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

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{D} & -\mathbf{C}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & \mathbf{MBC} \\ \mathbf{CDM} & -\mathbf{C} + \mathbf{CDMBC} \end{bmatrix}, \qquad (2.214)$$

where

$$\mathbf{M} = (\mathbf{A} + \mathbf{BCD})^{-1}. \qquad (2.215)$$

By 2.24,

$$\begin{bmatrix} -\mathbf{C}^{-1} & \mathbf{D} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}^{-1} = \begin{bmatrix} -\mathbf{N} & \mathbf{NDA}^{-1} \\ \mathbf{A}^{-1}\mathbf{BN} & \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BNDA}^{-1} \end{bmatrix}, \qquad (2.216)$$

where

$$\mathbf{N} = \left(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}\right)^{-1}. \qquad (2.217)$$

Therefore,

$$\mathbf{M} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BNDA}^{-1}. \qquad (2.218)$$

2.27

(a)

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

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \int\int (\mathbf{x} + \mathbf{z})p(\mathbf{x}, \mathbf{z})d\mathbf{x}d\mathbf{z}. \qquad (2.219)$$

The right hand side can be written as

$$\int \mathbf{x}\left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{z}\right)d\mathbf{x} + \int \mathbf{z}\left(\int p(\mathbf{x}, \mathbf{z})d\mathbf{x}\right)d\mathbf{z} = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} + \int \mathbf{z}p(\mathbf{z})d\mathbf{z}. \qquad (2.220)$$

Therefore,

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \mathrm{E}\,\mathbf{x} + \mathrm{E}\,\mathbf{z}. \qquad (2.221)$$

(b)

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

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

$$(2.222)$$

The right hand side can be written as

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

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

$$(2.223)$$

Each term can be written as

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

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

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

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

$$(2.224)$$

Therefore,

$$\mathrm{cov}(\mathbf{x}+\mathbf{z}) = \mathrm{cov}\,\mathbf{x} + \mathrm{cov}\,\mathbf{z}. \qquad\qquad (2.225)$$

2.28

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

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

$$\mathrm{E}\,\mathbf{z} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu}+\mathbf{b} \end{bmatrix}, \qquad\qquad (2.226)$$

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

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

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

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

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

(2.227)

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}\left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}\right)^{\mathsf{T}}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}\right). \qquad (2.228)$$

Therefore,

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right), \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right).
\end{aligned}
\qquad (2.229)
$$

2.29

Let

$$\mathbf{R} = \begin{bmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathsf{T}}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix}. \qquad (2.230)$$

By 2.24,

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}} \end{bmatrix}. \qquad (2.231)$$

2.30

Let

$$\mathbf{R}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}} \\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}} \end{bmatrix}. \qquad (2.232)$$

Then,

$$\mathbf{R}^{-1} \begin{bmatrix} \mathbf{\Lambda}\boldsymbol{\mu} - \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{bmatrix}. \tag{2.233}$$

2.31

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

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \tag{2.234}$$

where

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}\right), \\ p(\mathbf{z}) &= \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z}, \mathbf{\Sigma}_\mathbf{z}\right). \end{aligned} \tag{2.235}$$

By marginalisation,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{2.236}$$

The right hand side can be written as

$$\int \mathcal{N}\left(\mathbf{y}|\mathbf{x} + \boldsymbol{\mu}_\mathbf{z}, \mathbf{\Sigma}_\mathbf{z}\right) \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}\right) d\mathbf{x}. \tag{2.237}$$

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

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_\mathbf{z}\right)^\mathsf{T} \mathbf{\Sigma}_\mathbf{z}^{-1} \left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_\mathbf{z}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^\mathsf{T}\mathbf{\Sigma}_\mathbf{x}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}). \tag{2.238}$$

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

$$-\frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{R}\mathbf{u} + \mathbf{u}^\mathsf{T}\mathbf{v} = -\frac{1}{2}\left(\mathbf{u} - \mathbf{R}^{-1}\mathbf{v}\right)^\mathsf{T}\mathbf{R}\left(\mathbf{u} - \mathbf{R}^{-1}\mathbf{v}\right) + \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{R}^{-1}\mathbf{v}, \tag{2.239}$$

where

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \\ \mathbf{R} &= \begin{bmatrix} \mathbf{\Sigma}_\mathbf{x}^{-1} + \mathbf{\Sigma}_\mathbf{z}^{-1} & -\mathbf{\Sigma}_\mathbf{z}^{-1} \\ -\mathbf{\Sigma}_\mathbf{z}^{-1} & \mathbf{\Sigma}_\mathbf{z}^{-1} \end{bmatrix}, \\ \mathbf{v} &= \begin{bmatrix} \mathbf{\Sigma}_\mathbf{x}^{-1}\boldsymbol{\mu}_\mathbf{x} - \mathbf{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{z} \\ \mathbf{\Sigma}_\mathbf{z}^{-1}\boldsymbol{\mu}_\mathbf{z} \end{bmatrix}, \end{aligned} \tag{2.240}$$

By 2.29 and 2.30,

$$\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{\Sigma_x} & \mathbf{\Sigma_x} \\ \mathbf{\Sigma_x} & \mathbf{\Sigma_x} + \mathbf{\Sigma_z} \end{bmatrix},$$

$$\mathbf{R}^{-1}\mathbf{v} = \begin{bmatrix} \boldsymbol{\mu_x} \\ \boldsymbol{\mu_x} + \boldsymbol{\mu_z} \end{bmatrix}. \tag{2.241}$$

Therefore,

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu_x} + \boldsymbol{\mu_z}, \mathbf{\Sigma_x} + \mathbf{\Sigma_z}\right). \tag{2.242}$$

2.32

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}\right),$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}\right). \tag{2.243}$$

By the Bayes' theorem,

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}). \tag{2.244}$$

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

$$
\begin{aligned}
&-\frac{1}{2}\left(\mathbf{y} - \mathbf{Ax} - \mathbf{b}\right)^\mathsf{T} \mathbf{L}\left(\mathbf{y} - \mathbf{Ax} - \mathbf{b}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\
&= -\frac{1}{2}\left(\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}\right)^\mathsf{T} \mathbf{L}\left(\mathbf{y} - \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}\right) \\
&\quad - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu}).
\end{aligned} \tag{2.245}
$$

The right hand side can be written as

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA}\right)(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\
&\quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}) \\
&= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\mathsf{T}\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA}\right)(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) + \frac{1}{2}\mathbf{z}^\mathsf{T}\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{LA}\right)\mathbf{z} \\
&\quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}),
\end{aligned} \tag{2.246}
$$

where
$$\mathbf{z} = \left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}). \tag{2.247}$$

The right hand side can be written as
$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z})^\mathsf{T}\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)(\mathbf{x} - \boldsymbol{\mu} - \mathbf{z}) \\
&-\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})^\mathsf{T}\mathbf{M}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}),
\end{aligned} \tag{2.248}$$

where
$$\mathbf{M} = \mathbf{L} - \mathbf{L}\mathbf{A}\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}. \tag{2.249}$$

We have
$$\boldsymbol{\mu} + \mathbf{z} = \left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)\boldsymbol{\mu} + \mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})\right). \tag{2.250}$$

Then,
$$\boldsymbol{\mu} + \mathbf{z} = \left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\right). \tag{2.251}$$

By 2.26,
$$\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}\mathbf{A}\mathbf{\Lambda}^{-1}. \tag{2.252}$$

Then,
$$\mathbf{M} = \mathbf{L} - \mathbf{L}\mathbf{A}\left(\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}\mathbf{A}\mathbf{\Lambda}^{-1}\right)\mathbf{A}^\mathsf{T}\mathbf{L}. \tag{2.253}$$

The right hand side can be written as
$$\begin{aligned}
&\mathbf{L} - \mathbf{L}\left(\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)\mathbf{L} \\
=&\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)\mathbf{L}.
\end{aligned} \tag{2.254}$$

The right hand side can be written as
$$\begin{aligned}
&\mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1} \\
=&\mathbf{L}\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T} - \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)\left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}.
\end{aligned} \tag{2.255}$$

Then,
$$\mathbf{M} = \left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right)^{-1}. \tag{2.256}$$

Therefore,
$$\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{x}|\left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\right), \left(\mathbf{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A}\right)^{-1}\right), \\
p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\mathsf{T}\right).
\end{aligned} \tag{2.257}$$

2.33

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

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right),$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right). \tag{2.258}$$

By the Bayes' theorem,

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}). \tag{2.259}$$

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

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}\right)^{\mathsf{T}}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.260}$$

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

$$-\frac{1}{2}\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathsf{T}}\mathbf{L}\\-\mathbf{L}\mathbf{A} & \mathbf{L}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix} + \begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}-\mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{b} + \boldsymbol{\Lambda}\boldsymbol{\mu}\\\mathbf{L}\mathbf{b}\end{bmatrix}. \tag{2.261}$$

By 2.24,

$$\begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathsf{T}}\mathbf{L}\\-\mathbf{L}\mathbf{A} & \mathbf{L}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}}\\\mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}}\end{bmatrix}, \tag{2.262}$$

so that

$$\begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathsf{T}}\mathbf{L}\\-\mathbf{L}\mathbf{A} & \mathbf{L}\end{bmatrix}^{-1}\begin{bmatrix}-\mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{b} + \boldsymbol{\Lambda}\boldsymbol{\mu}\\\mathbf{L}\mathbf{b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}\\\mathbf{A}\boldsymbol{\mu} + \mathbf{b}\end{bmatrix}. \tag{2.263}$$

Then,

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}}\right). \tag{2.264}$$

By 2.25,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu} + \left(\boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{L}\left(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b}\right), \left(\boldsymbol{\Lambda} + \mathbf{A}^{\mathsf{T}}\mathbf{L}\mathbf{A}\right)^{-1}\right). \tag{2.265}$$

We have
$$\boldsymbol{\mu} + (\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{L}\,(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})$$
$$= (\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b}))\,. \tag{2.266}$$

Therefore,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}\,|\,(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1}(\boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y} - \mathbf{b})),(\boldsymbol{\Lambda} + \mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1}\right). \tag{2.267}$$

2.34

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{2.268}$$

Then,

$$\ln\left(\prod_{n=1}^{N} p(\mathbf{x}_n)\right) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\det\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.269}$$

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

$$\mathbf{0} = \sum_{n=1}^{N}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^\mathsf{T}\right)(\mathbf{x}_n - \boldsymbol{\mu}),$$
$$\mathbf{O} = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^\mathsf{T} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}. \tag{2.270}$$

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

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n,$$
$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}. \tag{2.271}$$

2.35

(a)

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right). \tag{2.272}$$

Then,
$$\mathrm{E}\,\mathbf{x}\mathbf{x}^{\mathsf{T}} = \mathrm{E}(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.273}$$

The right hand side can be written as

$$\mathrm{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} + \boldsymbol{\mu}\,\mathrm{E}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} + \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\mu}^{\mathsf{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.274}$$

Since

$$\begin{aligned} \mathrm{E}\,\mathbf{x} &= \boldsymbol{\mu}, \\ \mathrm{cov}\,\mathbf{x} &= \boldsymbol{\Sigma}, \end{aligned} \tag{2.275}$$

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

$$\mathrm{E}\,\mathbf{x}\mathbf{x}^{\mathsf{T}} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.276}$$

(b)

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

$$\begin{aligned} p(\mathbf{x}_n) &= \mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \\ p(\mathbf{x}_m) &= \mathcal{N}\left(\mathbf{x_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \end{aligned} \tag{2.277}$$

If $n \neq m$, then

$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^{\mathsf{T}} = \mathrm{E}\,\mathbf{x}_n\,\mathrm{E}\,\mathbf{x}_m^{\mathsf{T}}. \tag{2.278}$$

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

$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^{\mathsf{T}} = I_{nm}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.279}$$

(c)

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

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.280}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}. \end{aligned} \tag{2.281}$$

Then,

$$\mathrm{E}\,\mathbf{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{E}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathsf{T}}. \tag{2.282}$$

The right hand side can be written as

$$\begin{aligned}
&\frac{1}{N} \sum_{n=1}^{N} \mathrm{E}\,\mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} - \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{E}\left(\sum_{n=1}^{N} \mathbf{x_n}\right) \mathbf{x}_n^{\mathsf{T}} - \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{E}\,\mathbf{x}_n \left(\sum_{n=1}^{N} \mathbf{x_n}\right)^{\mathsf{T}} \\
&+ \frac{1}{N^3} \sum_{n=1}^{N} \mathrm{E}\left(\sum_{n=1}^{N} \mathbf{x_n}\right)\left(\sum_{n=1}^{N} \mathbf{x_n}\right)^{\mathsf{T}}.
\end{aligned} \tag{2.283}$$

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

$$-\frac{1}{N}\left((\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}) + (N-1)\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right) = -\frac{1}{N}\mathbf{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.284}$$

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

$$\frac{1}{N^2}\left(N\left(\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right) + N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right) = \frac{1}{N}\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.285}$$

Then,

$$\mathrm{E}\,\mathbf{\Sigma}_{\mathrm{ML}} = (\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}) + 2\left(-\frac{1}{N}\mathbf{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}\right) + \frac{1}{N}\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}}. \tag{2.286}$$

Therefore,

$$\mathrm{E}\,\mathbf{\Sigma}_{\mathrm{ML}} = \frac{N-1}{N}\mathbf{\Sigma}. \tag{2.287}$$

2.36

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

$$p(x_n) = \mathcal{N}\left(x_n | \mu, \sigma^2\right). \tag{2.288}$$

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

$$\sigma_{\mathrm{ML}}^2{}^{(N)} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2. \tag{2.289}$$

The right hand side can be written as

$$\frac{1}{N}(x_N - \mu)^2 + \frac{1}{N}\sum_{n=1}^{N-1}(x_n - \mu)^2 = \frac{1}{N}(x_N - \mu)^2 + \frac{N-1}{N}{\sigma_{\mathrm{ML}}^2}^{(N-1)}. \quad (2.290)$$

Then,

$${\sigma_{\mathrm{ML}}^2}^{(N)} = {\sigma_{\mathrm{ML}}^2}^{(N-1)} + \frac{1}{N}\left((x_N - \mu)^2 - {\sigma_{\mathrm{ML}}^2}^{(N-1)}\right). \quad (2.291)$$

We have

$$\frac{\partial}{\partial \sigma^2}\left(-\ln p(x_n)\right) = \frac{1}{2\sigma^2} - \frac{1}{2(\sigma^2)^2}(x_n - \mu)^2. \quad (2.292)$$

Therefore,

$${\sigma_{\mathrm{ML}}^2}^{(N)} = {\sigma_{\mathrm{ML}}^2}^{(N-1)} - \frac{2\left({\sigma_{\mathrm{ML}}^2}^{(N-1)}\right)^2}{N}\left(\frac{\partial}{\partial \sigma^2}\left(-\ln p(x_N)\right)\right)\Bigg|_{\sigma^2 = {\sigma_{\mathrm{ML}}^2}^{(N-1)}}. \quad (2.293)$$

2.37

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

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \quad (2.294)$$

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

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}}. \quad (2.295)$$

The right hand side can be written as

$$\begin{aligned}
&\frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} + \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}} \\
&=\frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} + \frac{N-1}{N}\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}.
\end{aligned} \quad (2.296)$$

Then,

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N} \left((\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}} - \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} \right). \tag{2.297}$$

By 3.21**??**, we have

$$\frac{\partial}{\partial \boldsymbol{\Sigma}} \left(-\ln p(x_n) \right) = -\frac{1}{2} \left(\boldsymbol{\Sigma}^{-1} \right)^{\mathsf{T}} + \frac{1}{2} \left(\boldsymbol{\Sigma}^{-1} \right)^2 (\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\mathsf{T}}. \tag{2.298}$$

Therefore,

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} - \frac{\left(\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} \right)^2}{N} \left(\frac{\partial}{\partial \boldsymbol{\Sigma}} \left(-\ln p \left(\mathbf{x}_N \right) \right) \right) \Big|_{\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}}. \tag{2.299}$$

2.38

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

$$\begin{aligned} p(x_n | \mu) &= \mathcal{N} \left(x_n | \mu, \sigma^2 \right), \\ p(\mu) &= \mathcal{N} \left(\mu | \mu_0, \sigma_0^2 \right). \end{aligned} \tag{2.300}$$

By the Bayes' theorem,

$$p(\mu | \mathbf{x}) p(\mathbf{x}) = p(\mathbf{x} | \mu) p(\mu). \tag{2.301}$$

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

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2. \tag{2.302}$$

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

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n. \tag{2.303}$$

Then, the first term can be written as

$$\begin{aligned} &-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}} + \mu_{\mathrm{ML}} - \mu)^2 \\ =&-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 - \frac{\mu_{\mathrm{ML}} - \mu}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}}) - \frac{N}{2\sigma^2} (\mu_{\mathrm{ML}} - \mu)^2. \end{aligned} \tag{2.304}$$

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

$$
\begin{aligned}
&- \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2} (\mu_{\mathrm{ML}} - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \\
&= - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 - \frac{1}{2\sigma_N^2} (\mu - \mu_N)^2 + \frac{\mu_N^2}{2\sigma_N^2} - \frac{N\mu_{\mathrm{ML}}^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2},
\end{aligned}
\tag{2.305}
$$

where

$$
\begin{aligned}
\mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{\mathrm{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\
\sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}.
\end{aligned}
\tag{2.306}
$$

Therefore,

$$
p(\mu|\mathbf{x}) = \mathcal{N} \left(\mu | \mu_N, \sigma_N^2 \right).
\tag{2.307}
$$

2.39

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

$$
\begin{aligned}
p(x_n|\mu) &= \mathcal{N} \left(x_n | \mu, \sigma^2 \right), \\
p(\mu) &= \mathcal{N} \left(\mu | \mu_0, \sigma_0^2 \right).
\end{aligned}
\tag{2.308}
$$

(a)

By 2.38,

$$
p(\mu|\mathbf{x}) = \mathcal{N} \left(\mu | \mu_N, \sigma_N^2 \right),
\tag{2.309}
$$

where

$$
\begin{aligned}
\mu_N &= \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{n=1}^{N} x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0, \\
\sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}.
\end{aligned}
\tag{2.310}
$$

Then,

$$
\frac{1}{\sigma_N^2} = \frac{(N-1)\sigma_0^2 + \sigma^2}{\sigma^2 \sigma_0^2} + \frac{1}{\sigma^2}.
\tag{2.311}
$$

Therefore,

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \tag{2.312}$$

We have

$$\frac{\mu_N}{\sigma_N^2} = \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n + \frac{\mu_0}{\sigma_0^2}, \tag{2.313}$$

so that

$$\frac{\mu_{N-1}}{\sigma_{N-1}^2} = \frac{1}{\sigma^2} \sum_{n=1}^{N-1} x_n + \frac{\mu_0}{\sigma_0^2}. \tag{2.314}$$

Therefore,

$$\frac{\mu_N}{\sigma_N^2} = \frac{\mu_{N-1}}{\sigma_{N-1}^2} + \frac{x_N}{\sigma^2}. \tag{2.315}$$

(b)

By the Bayes' theorem,

$$p(\mu|\mathbf{x}_N)p(\mathbf{x}_N) = p(\mathbf{x}_N|\mu)p(\mu). \tag{2.316}$$

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

$$p(\mu|\mathbf{x}_N)p(x_N)p(\mathbf{x}_{N-1}) = p(x_N|\mu)p(\mathbf{x}_{N-1}|\mu)p(\mu). \tag{2.317}$$

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

$$p(x_N|\mu)p(\mu|\mathbf{x}_{N-1})p(\mathbf{x}_{N-1}). \tag{2.318}$$

Then,

$$p(\mu|\mathbf{x}_N)p(x_N) = p(\mu|\mathbf{x}_{N-1})p(x_N|\mu). \tag{2.319}$$

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

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

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

$$+ \frac{1}{2}\left(\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}\right)\left(\frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}\mu_{N-1} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}x_N\right)^2 - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}. \tag{2.320}$$

83

Then,

$$\mu_N = \frac{\frac{1}{\sigma_{N-1}^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} \mu_{N-1} + \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}} x_N,$$

$$\sigma_N^2 = \frac{1}{\frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}}. \tag{2.321}$$

Therefore,

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

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}. \tag{2.322}$$

2.40

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

$$p(\mathbf{x}_n|\boldsymbol{\mu}) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$
$$p(\boldsymbol{\mu}) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right). \tag{2.323}$$

By 2.34, the maximum likelihood solution for $\boldsymbol{\mu}$ is given by

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n, \tag{2.324}$$

By the Bayes' theorem,

$$p(\boldsymbol{\mu}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}). \tag{2.325}$$

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

$$-\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathsf{T}}. \tag{2.326}$$

The first term can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})$$

$$= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - (\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})$$

$$-\frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}). \tag{2.327}$$

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

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})$$

$$-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\mathsf{T}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_N)$$

$$+\frac{1}{2}\boldsymbol{\mu}_N^\mathsf{T}\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N - \frac{N}{2}\boldsymbol{\mu}_{\mathrm{ML}}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}}, \tag{2.328}$$

where

$$\boldsymbol{\mu}_N = \left(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right),$$
$$\boldsymbol{\Sigma}_N = \left(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}. \tag{2.329}$$

Therefore,

$$p(\boldsymbol{\mu}|\mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N). \tag{2.330}$$

2.41

By the definition,

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.331}$$

Then,

$$\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a-1}\exp(-b\lambda)d\lambda. \tag{2.332}$$

By the transformation
$$\lambda' = b\lambda, \qquad (2.333)$$

the right hand side can be written as
$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a-1} \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{\Gamma(a)} \int_0^\infty \lambda'^{a-1} \exp(-\lambda')d\lambda'. \qquad (2.334)$$

The right hand side can be written as
$$\frac{1}{\Gamma(a)}\Gamma(a) = 1. \qquad (2.335)$$

Therefore,
$$\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda = 1. \qquad (2.336)$$

2.42

Let λ be a variable such that
$$p(\lambda) = \mathrm{Gam}(\lambda|a,b). \qquad (2.337)$$

By the definition,
$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \qquad (2.338)$$

(a)

We have
$$\mathrm{E}\,\lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right) d\lambda. \qquad (2.339)$$

By the transformation
$$\lambda' = b\lambda, \qquad (2.340)$$

the right hand side can be written as
$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b\Gamma(a)} \int_0^\infty \lambda'^a \exp(-\lambda')d\lambda'. \qquad (2.341)$$

The right hand side can be written as
$$\frac{1}{b\Gamma(a)}\Gamma(a+1) = \frac{a}{b}. \qquad (2.342)$$

Therefore,
$$\mathrm{E}\,\lambda = \frac{a}{b}. \qquad (2.343)$$

(b)

We have

$$\mathrm{E}\,\lambda^2 = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1} \exp\left(-\frac{\lambda}{b}\right) d\lambda. \tag{2.344}$$

By the transformation

$$\lambda' = b\lambda, \tag{2.345}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b^2\Gamma(a)} \int_0^\infty \lambda'^{a+1} \exp(-\lambda')d\lambda'. \tag{2.346}$$

The right hand side can be written as

$$\frac{1}{b^2\Gamma(a)}\Gamma(a+2) = \frac{a(a+1)}{b^2}. \tag{2.347}$$

Then,

$$\mathrm{E}\,\lambda^2 = \frac{a(a+1)}{b^2}. \tag{2.348}$$

We have

$$\mathrm{var}\,\lambda = \mathrm{E}\,\lambda^2 - (\mathrm{E}\,\lambda)^2. \tag{2.349}$$

Therefore,

$$\mathrm{var}\,\lambda = \frac{a}{b^2}. \tag{2.350}$$

(c)

Setting the derivative of $\mathrm{Gam}(\lambda|a,b)$ with respect to λ to zero gives

$$0 = \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{\lambda} - b\right) \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right). \tag{2.351}$$

Therefore,

$$\mathrm{mode}\,\lambda = \frac{a-1}{b}. \tag{2.352}$$

2.43

Let

$$p\left(x|\sigma^2, q\right) = \frac{q}{2\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \exp\left(-\frac{|x|^q}{2\sigma^2}\right). \tag{2.353}$$

87

(a)

We have

$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = \frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^{\infty} \exp\left(-\frac{x^q}{2\sigma^2}\right) dx. \tag{2.354}$$

By the transformation

$$x' = \frac{x^q}{2\sigma^2}, \tag{2.355}$$

the right hand side can be written as

$$\frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^{\infty} \exp(-x') \left(2\sigma^2\right)^{\frac{1}{q}} \frac{1}{q} x'^{\frac{1}{q}-1} dx'$$

$$= \frac{1}{\Gamma(\frac{1}{q})} \int_0^{\infty} x'^{\frac{1}{q}-1} \exp(-x') dx'. \tag{2.356}$$

The right hand side can be written as

$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma\left(\frac{1}{q}\right) = 1. \tag{2.357}$$

Therefore,

$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = 1. \tag{2.358}$$

(b)

We have

$$p\left(x|\sigma^2, 2\right) = \frac{1}{\Gamma(\frac{1}{2})} \left(2\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{2.359}$$

Therefore,

$$p\left(x|\sigma^2, 2\right) = \mathcal{N}\left(x|0, \sigma^2\right). \tag{2.360}$$

(c)

Let $\mathbf{t} = (t_1, \cdots, t_N)^\intercal$ and $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ such that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \tag{2.361}$$

where

$$p(\epsilon_n) = p\left(\epsilon_n | \sigma^2, q\right). \tag{2.362}$$

Then, the logarithm of $p(\epsilon_n)$ except the terms independent of $\mathbf{w}$ and σ^2 can be written as

$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q}\ln\left(2\sigma^2\right). \tag{2.363}$$

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

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q}\ln\left(2\sigma^2\right). \tag{2.364}$$

2.44

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

$$\begin{aligned}
p(x_n | \mu, \tau) &= \mathcal{N}\left(x_n | \mu, \tau^{-1}\right), \\
p(\mu, \tau) &= \mathcal{N}\left(\mu | \mu_0, (\beta\tau)^{-1}\right)\mathrm{Gam}(\tau | a, b).
\end{aligned} \tag{2.365}$$

By the Bayes' theorem,

$$p(\mu, \tau | \mathbf{x})p(\mathbf{x}) = p(\mathbf{x} | \mu, \tau)p(\mu, \tau). \tag{2.366}$$

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

$$\begin{aligned}
&\frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 + \frac{1}{2}\ln\tau - \frac{\beta\tau}{2}(\mu - \mu_0)^2 + (a-1)\ln\tau - b\tau \\
&= \left(a + \frac{N-1}{2}\right)\ln\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{\beta\tau}{2}(\mu - \mu_0)^2 - b\tau.
\end{aligned} \tag{2.367}$$

Let

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N}x_n. \tag{2.368}$$

Then,

$$\sum_{n=1}^{N}(x_n - \mu)^2 = \sum_{n=1}^{N}(x_n - \bar{x} + \bar{x} - \mu)^2. \tag{2.369}$$

The right hand side can be written as

$$\sum_{n=1}^{N}(x_n - \bar{x})^2 + 2(\bar{x} - \mu)\sum_{n=1}^{N}(x_n - \bar{x}) + N(\bar{x} - \mu)^2$$
$$= \sum_{n=1}^{N}(x_n - \bar{x})^2 + N(\bar{x} - \mu)^2. \tag{2.370}$$

Then, the logarithm except the terms independent of $\mathbf{x}$, μ and τ can be written as

$$\left(a + \frac{N-1}{2}\right)\ln\tau - \frac{N\tau}{2}(\bar{x} - \mu)^2 - \frac{\beta\tau}{2}(\mu - \mu_0)^2 - b\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2. \tag{2.371}$$

The second and third terms can be written as

$$-\frac{N\tau}{2}(\bar{x} - \mu)^2 - \frac{\beta\tau}{2}(\mu - \mu_0)^2$$
$$= -\frac{(N+\beta)\tau}{2}\left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2 + \frac{(N\bar{x} + \beta\mu_0)^2\tau}{2(N+\beta)} - \frac{N\tau}{2}\bar{x}^2 - \frac{\beta\tau}{2}\mu_0^2. \tag{2.372}$$

The second, third and forth terms on the right hand side can be written as

$$\frac{(N\bar{x} + \beta\mu_0)^2\tau - (N+\beta)N\tau\bar{x}^2 - (N+\beta)\beta\tau\mu_0^2}{2(N+\beta)} = -\frac{N\beta\tau(\bar{x} - \mu_0)^2}{2(N+\beta)}. \tag{2.373}$$

Then, the logrithm except the terms independent of $\mathbf{x}$, μ and τ can be written as

$$-\frac{(N+\beta)\tau}{2}\left(\mu - \frac{N\bar{x} + \beta\mu_0}{N+\beta}\right)^2$$
$$+ \left(a + \frac{N-1}{2}\right)\ln\tau - \left(b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N+\beta)} + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2\right)\tau. \tag{2.374}$$

Therefore,

$$p(\mu, \tau|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \tau_N^{-1}\right)\mathrm{Gam}\left(\tau|a_N, b_N\right), \tag{2.375}$$

where

$$\mu_N = \frac{N\bar{x} + \beta\mu_0}{N + \beta},$$

$$\tau_N = (N + \beta)\tau,$$

$$a_N = a + \frac{N + 1}{2}, \tag{2.376}$$

$$b_N = b + \frac{N\beta(\bar{x} - \mu_0)^2}{2(N + \beta)} + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2.$$

2.45

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

$$p(\mathbf{x}_n|\mathbf{\Lambda}) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \mathbf{\Lambda}^{-1}\right),$$

$$p(\mathbf{\Lambda}) = \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu), \tag{2.377}$$

where

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu)|\det\mathbf{\Lambda}|^{\frac{\nu - D - 1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{-1}\mathbf{\Lambda}\right)\right). \tag{2.378}$$

By the Bayes' theorem,

$$p(\mathbf{\Lambda}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\mathbf{\Lambda})p(\mathbf{\Lambda}). \tag{2.379}$$

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

$$-\frac{N}{2}\ln\left|\det\left(\mathbf{\Lambda}^{-1}\right)\right| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}\mathbf{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})$$

$$+ \frac{\nu - D - 1}{2}\ln|\det\mathbf{\Lambda}| - \frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{-1}\mathbf{\Lambda}\right) \tag{2.380}$$

$$= \frac{\nu + N - D - 1}{2}\ln|\det\mathbf{\Lambda}| - \frac{1}{2}\operatorname{tr}\left(\left(\mathbf{W}^{-1} + \mathbf{S}\right)\mathbf{\Lambda}\right),$$

where

$$\mathbf{S} = \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\mathsf{T}. \tag{2.381}$$

Then,

$$p(\mathbf{\Lambda}|\mathbf{X}) = \mathcal{W}\left(\mathbf{\Lambda}|\left(\mathbf{W}^{-1} + \mathbf{S}\right)^{-1}, \nu + N\right). \tag{2.382}$$

Therefore, $\mathcal{W}$ is a conjugate prior distribution of $\mathbf{\Lambda}$.

2.46

Let x be a variable such that

$$p(x|\tau) = \mathcal{N}\left(x|\mu, \tau^{-1}\right),$$
$$p(\tau) = \mathrm{Gam}(\tau|a, b).$$
$$(2.383)$$

By marginalisation,

$$p(x) = \int_0^\infty p(x|\tau)p(\tau)d\tau. \tag{2.384}$$

The right hand side can be written as

$$\int_0^\infty \left(2\pi\tau^{-1}\right)^{-\frac{1}{2}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau)d\tau$$
$$= (2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp\left(-\left(b + \frac{(x-\mu)^2}{2}\right)\tau\right)d\tau. \tag{2.385}$$

By the transformation

$$\tau' = \left(b + \frac{(x-\mu)^2}{2}\right)\tau, \tag{2.386}$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{\tau'}{b + \frac{(x-\mu)^2}{2}}\right)^{a-\frac{1}{2}} \exp(-\tau') \frac{1}{b + \frac{(x-\mu)^2}{2}} d\tau'$$
$$= \Gamma\left(a + \frac{1}{2}\right)\left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.387}$$

Then,

$$p(x) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.388}$$

By the transformation

$$\nu = 2a,$$
$$\lambda = \frac{a}{b}, \tag{2.389}$$

the right hand side can be written as

$$(2\pi)^{-\frac{1}{2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\nu}{2\lambda}\right)^{-\frac{1}{2}} \left(1 + \frac{(x-\mu)^2}{\frac{\nu}{\lambda}}\right)^{-\frac{\nu+1}{2}}$$

$$= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda}{\nu}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}. \tag{2.390}$$

Therefore,

$$p(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda}{\nu}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}. \tag{2.391}$$

2.47

Let

$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda}{\nu}(x-\mu)^2\right)^{-\frac{\nu+1}{2}}. \tag{2.392}$$

By the transformation

$$\frac{1}{y} = \frac{\lambda}{\nu}(x-\mu)^2, \tag{2.393}$$

the right hand side except the terms independent of x can be written as

$$\left(1 + \frac{1}{y}\right)^{-\frac{\lambda(x-\mu)^2}{2}y - \frac{1}{2}}. \tag{2.394}$$

By the property

$$\lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^x = e, \tag{2.395}$$

we have

$$\lim_{y\to\infty} \left(1 + \frac{1}{y}\right)^{-\frac{\lambda(x-\mu)^2}{2}y - \frac{1}{2}} = \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right). \tag{2.396}$$

Therefore,

$$\lim_{\nu\to\infty} \mathrm{St}(x|\mu, \lambda, \nu) = \mathcal{N}(x|\mu, \lambda^{-1}). \tag{2.397}$$

2.48

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

$$p(\mathbf{x}|\eta) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right),$$
$$p(\eta) = \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right). \tag{2.398}$$

By marginalisation,

$$p(\mathbf{x}) = \int_0^\infty p(\mathbf{x}|\eta)p(\eta)d\eta. \tag{2.399}$$

The right hand side can be written as

$$\int_0^\infty (2\pi)^{-\frac{D}{2}}\left|\det(\eta\boldsymbol{\Lambda})^{-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{\eta}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right) \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\eta^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\eta\right)d\eta$$

$$=(2\pi)^{-\frac{D}{2}}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\int_0^\infty \eta^{\frac{D+\nu}{2}-1}\exp\left(-\frac{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2}\eta\right)d\eta. \tag{2.400}$$

By the transformation

$$\eta' = \frac{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2}\eta, \tag{2.401}$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{2\eta'}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}-1}\exp(-\eta')\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}d\eta'$$

$$=\left(\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}}\int_0^\infty \eta'^{\frac{D+\nu}{2}-1}\exp(-\eta')d\eta'. \tag{2.402}$$

Then,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\left(\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}}\Gamma\left(\frac{D+\nu}{2}\right). \tag{2.403}$$

The right hand side can be written as

$$(2\pi)^{-\frac{D}{2}}\frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\left(\frac{\nu}{2}\right)^{-\frac{D}{2}}\left(\frac{\nu}{2}\right)^{\frac{D+\nu}{2}}\left(\frac{2}{\nu+(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}}$$

$$=(2\pi)^{-\frac{D}{2}}\frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})}|\det\boldsymbol{\Lambda}|^{\frac{1}{2}}\left(\frac{\nu}{2}\right)^{-\frac{D}{2}}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \tag{2.404}$$

Therefore,

$$p(\mathbf{x}) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{|\det \mathbf{\Lambda}|^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \tag{2.405}$$

2.49

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

$$p(\mathbf{x}) = \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu), \tag{2.406}$$

where

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.407}$$

(a)

We have

$$\mathrm{E}\,\mathbf{x} = \int \mathbf{x}\,\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)d\mathbf{x}. \tag{2.408}$$

The right hand side can be written as

$$\int \mathbf{x} \left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta\right) d\mathbf{x}$$
$$= \int \left(\int \mathbf{x}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\mathbf{\Lambda})^{-1}\right) d\mathbf{x}\right) \mathrm{Gam}\left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.409}$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \mathrm{Gam}\left(\eta \,|\, \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \tag{2.410}$$

Therefore,

$$\mathrm{E}\,\mathbf{x} = \boldsymbol{\mu}. \tag{2.411}$$

(b)

By (a), we have

$$\mathrm{cov}\,\mathbf{x} = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)d\mathbf{x}. \tag{2.412}$$

95

The right hand side can be written as

$$\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x}$$
$$= \int \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) d\mathbf{x} \right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.413}$$

The right hand side can be written as

$$\int (\eta\boldsymbol{\Lambda})^{-1} \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\int \eta^{\frac{\nu}{2}-2} \exp\left(-\frac{\nu}{2}\eta\right) d\eta \right) \boldsymbol{\Lambda}^{-1}. \tag{2.414}$$

By the transformation

$$\eta' = \frac{\nu}{2}\eta, \tag{2.415}$$

the integral of the right hand side can be written as

$$\int \left(\frac{2}{\nu}\eta'\right)^{\frac{\nu}{2}-2} \exp(-\eta') \frac{2}{\nu} d\eta' = \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2}-1\right). \tag{2.416}$$

Then,

$$\mathrm{cov}\,\mathbf{x} = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2}-1\right) \boldsymbol{\Lambda}^{-1}. \tag{2.417}$$

Therefore,

$$\mathrm{cov}\,\mathbf{x} = \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1}. \tag{2.418}$$

(c)

Setting the derivative of $p(\mathbf{x})$ to zero gives

$$\mathbf{0} = -\frac{1}{2}\left(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\mathsf{T}\right)(\mathbf{x} - \boldsymbol{\mu}) \int \eta \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.419}$$

Therefore,

$$\mathrm{mode}\,\mathbf{x} = \boldsymbol{\mu}. \tag{2.420}$$

96

2.50

Let

$$\text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \quad (2.421)$$

By the transformation

$$y = \frac{\nu}{(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}, \quad (2.422)$$

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

$$\left(1 + \frac{1}{y}\right)^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{2}y - \frac{D}{2}}. \quad (2.423)$$

By the property

$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = e, \quad (2.424)$$

we have

$$\lim_{y \to \infty} \left(1 + \frac{1}{y}\right)^{-\frac{(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})}{2}y - \frac{D}{2}} = \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (2.425)$$

Therefore,

$$\lim_{\nu \to \infty} \text{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}). \quad (2.426)$$

2.51

(a)

We have

$$\exp(iA)\exp(-iA) = 1. \quad (2.427)$$

The left hand side can be written as

$$(\cos A + i \sin A)(\cos A - i \sin A) = \cos^2 A + \sin^2 A. \quad (2.428)$$

Therefore,

$$\cos^2 A + \sin^2 A = 1. \quad (2.429)$$

97

(b)

We have
$$\cos(A - B) = \operatorname{Re}\left(\exp\left(i(A - B)\right)\right). \tag{2.430}$$

The right hand side can be written as
$$\operatorname{Re}\left(\exp(iA)\exp(-iB)\right) = \operatorname{Re}\left((\cos A + i\sin A)(\cos B - i\sin B)\right). \tag{2.431}$$

Therefore,
$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \tag{2.432}$$

(c)

We have
$$\sin(A - B) = \operatorname{Im}\left(\exp\left(i(A - B)\right)\right). \tag{2.433}$$

The right hand side can be written as
$$\operatorname{Im}\left(\exp(iA)\exp(-iB)\right) = \left((\cos A + i\sin A)(\cos B - i\sin B)\right). \tag{2.434}$$

Therefore,
$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \tag{2.435}$$

2.52

Let
$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)}\exp\left(m\cos(\theta - \theta_0)\right), \tag{2.436}$$

where
$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi}\exp(m\cos\theta)d\theta. \tag{2.437}$$

By the Taylor series
$$\cos\alpha = 1 - \frac{1}{2}\alpha^2 + O\left(\alpha^4\right), \tag{2.438}$$

the right hand side can be written as
$$\begin{aligned}
&\frac{\exp\left(m\left(1 - \frac{1}{2}(\theta - \theta_0)^2 + O\left((\theta - \theta_0)^4\right)\right)\right)}{\int_0^{2\pi}\exp\left(m\left(1 - \frac{1}{2}\theta^2 + O(\theta^4)\right)\right)d\theta} \\
&= \exp\left(-\frac{m}{2}(\theta - \theta_0)^2\right)\frac{\exp\left(mO\left((\theta - \theta_0)^4\right)\right)}{\int_0^{2\pi}\exp\left(m\left(-\frac{1}{2}\theta^2 + O(\theta^4)\right)\right)d\theta}.
\end{aligned} \tag{2.439}$$

Therefore,
$$\lim_{m\to\infty} f(\theta|\theta_0, m) = \mathcal{N}\left(\theta|\theta_0, m^{-1}\right). \tag{2.440}$$

98

2.53

Let

$$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0. \tag{2.441}$$

The left hand side can be written as

$$\sum_{n=1}^{N} (\sin \theta_n \cos \theta_0 - \cos \theta_n \sin \theta_0) = \cos \theta_0 \sum_{n=1}^{N} \sin \theta_n - \sin \theta_0 \sum_{n=1}^{N} \cos \theta_n. \tag{2.442}$$

Therefore,

$$\theta_0 = \arctan \left(\frac{\sum_{n=1}^{N} \sin \theta_n}{\sum_{n=1}^{N} \cos \theta_n} \right). \tag{2.443}$$

2.54

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m \cos(\theta - \theta_0) \right), \tag{2.444}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \tag{2.445}$$

Setting the first and seconnd derivatives with respect to θ to zero gives

$$
\begin{aligned}
0 &= -m \sin(\theta - \theta_0) f(\theta|\theta_0, m), \\
0 &= \left(m^2 \sin^2(\theta - \theta_0) - m \cos(\theta - \theta_0) \right) f(\theta|\theta_0, m).
\end{aligned}
\tag{2.446}
$$

Therefore,

$$
\begin{aligned}
\operatorname*{argmax}_{\theta} f(\theta|\theta_0, m) &= \theta_0, \\
\operatorname*{argmin}_{\theta} f(\theta|\theta_0, m) &= \theta_0 - \pi \operatorname{sgn}(\theta_0 - \pi).
\end{aligned}
\tag{2.447}
$$

2.55

(a)

Let $\theta_1, \cdots, \theta_N$ be variables such that

$$p(\theta_n) = f(\theta_n|\theta_0, m), \tag{2.448}$$

where

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m \cos(\theta - \theta_0)\right),$$

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \qquad (2.449)$$

Then,

$$\ln\left(\prod_{n=1}^{N} p(\theta_n|\theta_0, m)\right) = -\frac{N}{2} \ln\left(2\pi I_0(m)\right) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0). \qquad (2.450)$$

Setting the derivative with respect to θ_0 to zero gives

$$0 = m \sum_{n=1}^{N} \sin(\theta_n - \theta_0). \qquad (2.451)$$

Therefore, by 2.53, the maximum likelihood solution for θ_0 is given by

$$\theta_0^{\mathrm{ML}} = \arctan\left(\frac{\sum_{n=1}^{N} \sin \theta_n}{\sum_{n=1}^{N} \cos \theta_n}\right). \qquad (2.452)$$

(b)

Let

$$\bar{r} \cos \bar{\theta} = \frac{1}{N} \sum_{n=1}^{N} \cos \theta_n,$$

$$\bar{r} \sin \bar{\theta} = \frac{1}{N} \sum_{n=1}^{N} \sin \theta_n. \qquad (2.453)$$

By (a),

$$\bar{\theta} = \theta_0^{\mathrm{ML}}. \qquad (2.454)$$

We have

$$\frac{1}{N} \sum_{n=1}^{N} \cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) = \left(\frac{1}{N} \sum_{n=1}^{N} \cos \theta_n\right) \cos \theta_0^{\mathrm{ML}} + \left(\frac{1}{N} \sum_{n=1}^{N} \sin \theta_n\right) \sin \theta_0^{\mathrm{ML}}. \qquad (2.455)$$

The right hand side can be written as

$$\bar{r} \cos^2 \bar{\theta} + \bar{r} \sin^2 \bar{\theta} = \bar{r}. \qquad (2.456)$$

100

Therefore,

$$\frac{1}{N}\sum_{n=1}^{N}\cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) = \bar{r}. \tag{2.457}$$

2.56

(a)

Let

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}. \tag{2.458}$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left((a-1)\ln\mu + (b-1)\ln(1-\mu)\right) \tag{2.459}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}. $$

(b)

Let

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.460}$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\exp\left((a-1)\ln\lambda - b\lambda\right). \tag{2.461}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ -b \end{bmatrix}. $$

101

(c)

Let

$$f(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m \cos(\theta - \theta_0)\right), \qquad (2.462)$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta, \qquad (2.463)$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m \cos \theta_0 \cos \theta + m \sin \theta_0 \sin \theta). \qquad (2.464)$$

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,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right). \qquad (2.465)$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}} \mathbf{u}(\mathbf{x})\right), \qquad (2.466)$$

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

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right). \qquad (2.467)$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = 1 \qquad (2.468)$$

with respect to $\boldsymbol{\eta}$ gives

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right) d\mathbf{x} + g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right) d\mathbf{x} = \mathbf{0}. \qquad (2.469)$$

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathrm{E}\,\mathbf{u}(\mathbf{x}). \qquad (2.470)$$

Therefore,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}. \qquad (2.471)$$

Thus,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}). \qquad (2.472)$$

Taking the second derivative with respect to $\boldsymbol{\eta}$ gives

$$\nabla\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right) d\mathbf{x} + 2\nabla g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})^{\mathsf{T}}h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right) d\mathbf{x}$$
$$+ g(\boldsymbol{\eta}) \int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathsf{T}}\mathbf{u}(\mathbf{x})\right) d\mathbf{x} = \mathbf{O}. \qquad (2.473)$$

The left hand side can be written as

$$\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \frac{2\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} \int \mathbf{u}(\mathbf{x})^{\mathsf{T}}p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x}$$
$$= \frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2\,\mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^{\mathsf{T}} + \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}\right). \qquad (2.474)$$

Therefore,

$$\mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}\right) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^{\mathsf{T}}}{g^2(\boldsymbol{\eta})}. \tag{2.475}$$

By the definition,

$$\mathrm{cov}\,\mathbf{u}(\mathbf{x}) = \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\mathsf{T}}\right) - \mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^{\mathsf{T}}. \tag{2.476}$$

Thus,

$$\mathrm{cov}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^{\mathsf{T}}}{g^2(\boldsymbol{\eta})}. \tag{2.477}$$

Hence,

$$\mathrm{cov}\,\mathbf{u}(\mathbf{x}) = -\nabla\nabla \ln g(\boldsymbol{\eta}). \tag{2.478}$$

2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right). \tag{2.479}$$

Then

$$\int p(x|\sigma)dx = \frac{1}{\sigma}\int f\left(\frac{x}{\sigma}\right)dx. \tag{2.480}$$

By the transformation

$$x' = \frac{x}{\sigma}, \tag{2.481}$$

the right hand side can be written as

$$\frac{1}{\sigma}\int f(x')\sigma dx' = \int f(x')dx'. \tag{2.482}$$

Therefore, $p(x|\sigma)$ will be normalised if $f(x)$ is normalised.

2.60

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

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \tag{2.483}$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \tag{2.484}$$

Since
$$\int p(\mathbf{x})d\mathbf{x} = 1, \tag{2.485}$$

we have
$$\sum_i h_i \Delta_i = 1. \tag{2.486}$$

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

$$\ln\left(\prod_i h_i{}^{n_i}\right) = \sum_i n_i \ln h_i, \tag{2.487}$$

where
$$\sum_i n_i = N. \tag{2.488}$$

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda\left(\sum_i h_i \Delta_i - 1\right) \tag{2.489}$$

with respect to h_i and λ to zero gives

$$\begin{aligned} \frac{n_i}{h_i} + \lambda \Delta_i &= 0, \\ \sum_i h_i \Delta_i - 1 &= 0. \end{aligned} \tag{2.490}$$

Then,
$$\begin{aligned} \lambda &= -N, \\ h_i &= \frac{n_i}{N\Delta_i}. \end{aligned} \tag{2.491}$$

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

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})}, \tag{2.492}$$

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}'-\mathbf{x}\|\leq\|\mathbf{x}_{(K)}-\mathbf{x}\|} d\mathbf{x}', \qquad (2.493)$$

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

(a)

We have

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \tag{3.1}$$

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

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{3.4}$$

(b)

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

$$w_0 f + \sum_{m=1}^{M} w_m \frac{1 + \tanh\left(\frac{x - \mu_m}{2s}\right)}{2}$$

$$= w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m + \frac{1}{2}\sum_{m=1}^{M} w_m \tanh\left(\frac{x - \mu_m}{2s}\right). \tag{3.6}$$

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

$$u_0 = w_0 + \frac{1}{2} \sum_{m=1}^{M} w_m,$$

$$u_m = \frac{1}{2} w_m. \tag{3.8}$$

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

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left(t_n - \mathbf{w}^{\mathsf{T}} \boldsymbol{\phi}_n \right)^2. \tag{3.11}$$

The right hand side can be written as

$$\frac{1}{2} \| \mathbf{t}' - \boldsymbol{\Phi}' \mathbf{w} \|^2, \tag{3.12}$$

where

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \vdots \\ \sqrt{r_N} t_N \end{bmatrix}, \boldsymbol{\Phi}' = \begin{bmatrix} \sqrt{r_1} \boldsymbol{\phi}(\mathbf{x}_1)^{\mathsf{T}} \\ \vdots \\ \sqrt{r_N} \boldsymbol{\phi}(\mathbf{x}_N)^{\mathsf{T}} \end{bmatrix}. \tag{3.13}$$

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

$$\mathbf{0} = -\boldsymbol{\Phi}'^{\mathsf{T}} (\mathbf{t}' - \boldsymbol{\Phi}' \mathbf{w}). \tag{3.14}$$

Therefore,

$$\underset{\mathbf{w}}{\operatorname{argmin}} E(\mathbf{w}) = \left(\boldsymbol{\Phi}'^{\mathsf{T}} \boldsymbol{\Phi}' \right)^{-1} \boldsymbol{\Phi}'^{\mathsf{T}} \mathbf{t}'. \tag{3.15}$$

3.4 (Incomplete)

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right)^2 , \tag{3.16}$$

where

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{m=1}^{M} w_m (x_m + \epsilon_m), \tag{3.17}$$

$$p(\epsilon_m) = \mathcal{N} \left(\epsilon_m | 0, \sigma^2 \right) .$$

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

$$\mathbf{0} = \sum_{n=1}^{N} \begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right) . \tag{3.18}$$

The right hand side can be written as

$$\sum_{n=1}^{N} \begin{bmatrix} 1 \\ \mathbf{x}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right) + \sum_{n=1}^{N} \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_n \end{bmatrix} \left(y(\mathbf{x}_n, \mathbf{w}) - t_n \right) . \tag{3.19}$$

3.5

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n) \right)^2 . \tag{3.20}$$

Then, the minimisation of $E(\mathbf{w})$ under the constraint

$$\sum_{m=1}^{M} |w_m|^q \leq \eta \tag{3.21}$$

reduces to the minimisation of

$$E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \tag{3.22}$$

with respect to $\mathbf{w}$ and λ. Then,

$$\eta = \sum_{m=1}^{M} |w_m^*(\lambda)|^q, \tag{3.23}$$

109

where

$$\mathbf{w}^*(\lambda) = \operatorname*{argmin}_{\mathbf{w}} \left(E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \right). \tag{3.24}$$

3.6

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in D dimensions such that

$$p(\mathbf{t}_n|\mathbf{y}_n) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}_n, \boldsymbol{\Sigma}\right), \tag{3.25}$$

where

$$\mathbf{y}_n = \mathbf{W}^\intercal \boldsymbol{\phi}_n. \tag{3.26}$$

Then,

$$\ln p(\mathbf{T}|\mathbf{Y}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(\det \boldsymbol{\Sigma})$$
$$-\frac{1}{2}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\intercal \boldsymbol{\phi}_n)^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^\intercal \boldsymbol{\phi}_n). \tag{3.27}$$

By 3.21, setting the derivatives with respect to $\mathbf{W}$ and $\boldsymbol{\Sigma}$ to zero gives

$$\mathbf{O} = -\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\intercal\right) \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\intercal \boldsymbol{\phi}_n)\,\boldsymbol{\phi}_n^\intercal,$$
$$\mathbf{O} = -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^\intercal + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2 \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^\intercal \boldsymbol{\phi}_n)(\mathbf{t}_n - \mathbf{W}^\intercal \boldsymbol{\phi}_n)^\intercal. \tag{3.28}$$

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

$$\mathbf{W}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\intercal \mathbf{t},$$
$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\intercal \boldsymbol{\phi}_n)(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\intercal \boldsymbol{\phi}_n)^\intercal. \tag{3.29}$$

3.7

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.30}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.31}$$

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

$$
\begin{aligned}
&-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n)^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\
&= -\frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v} + \mathbf{v}^\mathsf{T}\mathbf{u} - \frac{1}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0,
\end{aligned}
\tag{3.32}
$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} & -\beta\boldsymbol{\Phi}^\mathsf{T} \\ -\beta\boldsymbol{\Phi} & \beta\mathbf{I} \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}_0^{-1}\mathbf{m}_0 \\ \mathbf{0} \end{bmatrix}. \tag{3.33}$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \tag{3.34}$$

where

$$
\begin{aligned}
\mathbf{m} &= \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right), \\
\mathbf{S} &= \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}.
\end{aligned}
\tag{3.35}
$$

3.8

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).
\end{aligned}
\tag{3.36}
$$

By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \tag{3.37}$$

where

$$
\begin{aligned}
\mathbf{m} &= \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right), \\
\mathbf{S} &= \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}.
\end{aligned}
\tag{3.38}
$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}')p(\mathbf{t}') = p(\mathbf{t}'|\mathbf{w})p(\mathbf{w}), \tag{3.39}$$

111

where
$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}. \tag{3.40}$$

Since $\mathbf{t}$ and t_{N+1} are independent, it can be written as
$$p(\mathbf{w}|\mathbf{t}')p(t_{N+1})p(\mathbf{t}) = p(t_{N+1}|\mathbf{w})p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.41}$$

By the Bayes' theorem, the right hand side can be written as
$$p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})p(\mathbf{t}). \tag{3.42}$$

Then,
$$p(\mathbf{w}|\mathbf{t}')p(t_{N+1}) = p(\mathbf{w}|\mathbf{t})p(t_{N+1}|\mathbf{w}). \tag{3.43}$$

The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as
$$\begin{aligned}
&-\frac{1}{2}(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w}-\mathbf{m}) - \frac{\beta}{2}(t_{N+1}-\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_{N+1})^2 \\
&= -\frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v} + \mathbf{v}^\mathsf{T}\mathbf{u} - \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m},
\end{aligned} \tag{3.44}$$

where
$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\mathsf{T} & -\beta\boldsymbol{\phi}_{N+1} \\ -\beta\boldsymbol{\phi}_{N+1}^\mathsf{T} & \beta \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}^{-1}\mathbf{m} \\ 0 \end{bmatrix}. \tag{3.45}$$

Then,
$$p(\mathbf{w}|\mathbf{t}') = \mathcal{N}(\mathbf{w}|\mathbf{m}', \mathbf{S}'), \tag{3.46}$$

where
$$\begin{aligned}
\mathbf{m}' &= \mathbf{S}'\left(\mathbf{S}^{-1}\mathbf{m} + \beta t_{N+1}\boldsymbol{\phi}_{N+1}\right), \\
\mathbf{S}' &= \left(\mathbf{S}^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\mathsf{T}\right)^{-1}.
\end{aligned} \tag{3.47}$$

We have
$$\begin{aligned}
\mathbf{S}^{-1}\mathbf{m} + \beta t_{N+1}\boldsymbol{\phi}_{N+1} &= \mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}'^\mathsf{T}\mathbf{t}', \\
\mathbf{S}^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\mathsf{T} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}'^\mathsf{T}\boldsymbol{\Phi}',
\end{aligned} \tag{3.48}$$

where
$$\boldsymbol{\Phi}' = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\phi}_{N+1}^\mathsf{T} \end{bmatrix}. \tag{3.49}$$

Therefore,
$$\begin{aligned}
\mathbf{m}' &= \mathbf{S}'\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}'^\mathsf{T}\mathbf{t}'\right), \\
\mathbf{S}' &= \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}'^\mathsf{T}\boldsymbol{\Phi}'\right)^{-1}.
\end{aligned} \tag{3.50}$$

3.9

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.51}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S}), \tag{3.52}$$

where

$$\mathbf{m} = \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right),$$
$$\mathbf{S} = \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}. \tag{3.53}$$

Let

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\phi}_{N+1}^\mathsf{T} \end{bmatrix}. \tag{3.54}$$

Then,

$$p(\mathbf{w}|\mathbf{t}') = \mathcal{N}\left(\mathbf{w}|\mathbf{m}', \mathbf{t}'\right), \tag{3.55}$$

where

$$\mathbf{m}' = \mathbf{S}'\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}'^\mathsf{T}\mathbf{t}'\right),$$
$$\mathbf{S}' = \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}'^\mathsf{T}\boldsymbol{\Phi}'\right)^{-1}. \tag{3.56}$$

3.10

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right). \tag{3.57}$$

By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right), \tag{3.58}$$

where

$$\mathbf{m} = \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right),$$
$$\mathbf{S} = \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}. \tag{3.59}$$

By marginalisation,

$$p(t_{N+1}|\mathbf{t}) = \int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{3.60}$$

The logarithm of the integrand of the right hand side except the terms independent of t_{N+1} and $\mathbf{w}$ can be written as

$$
\begin{aligned}
&-\frac{\beta}{2}\left(t_{N+1}-\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_{N+1}\right)^2 - \frac{1}{2}(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w}-\mathbf{m}) \\
&= -\frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v} + \mathbf{v}^\mathsf{T}\mathbf{u} - \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m},
\end{aligned}
\tag{3.61}
$$

where

$$
\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\mathsf{T} & -\beta\boldsymbol{\phi}_{N+1} \\ -\beta\boldsymbol{\phi}_{N+1}^\mathsf{T} & \beta \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}^{-1}\mathbf{m} \\ 0 \end{bmatrix}. \tag{3.62}
$$

By 2.24,

$$
\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{S} & \mathbf{S}\boldsymbol{\phi}_{N+1} \\ \boldsymbol{\phi}_{N+1}^\mathsf{T}\mathbf{S} & \beta^{-1} + \boldsymbol{\phi}_{N+1}^\mathsf{T}\mathbf{S}\boldsymbol{\phi}_{N+1} \end{bmatrix}. \tag{3.63}
$$

Then,

$$
\mathbf{M}^{-1}\mathbf{u} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m}^\mathsf{T}\boldsymbol{\phi}_{N+1} \end{bmatrix}. \tag{3.64}
$$

Therefore,

$$
p(t_{N+1}|\mathbf{t}) = \mathcal{N}\left(t_{N+1}|\mathbf{m}^\mathsf{T}\boldsymbol{\phi}_{N+1}, \sigma^2(\boldsymbol{\phi}_{N+1})\right), \tag{3.65}
$$

where

$$
\sigma^2(\boldsymbol{\phi}) = \beta^{-1} + \boldsymbol{\phi}^\mathsf{T}\mathbf{S}\boldsymbol{\phi}. \tag{3.66}
$$

3.11

Let $t_1, \cdots, t_N$ be a variable such that

$$
\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right).
\end{aligned}
\tag{3.67}
$$

By 3.7,

$$
p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right), \tag{3.68}
$$

where

$$
\begin{aligned}
\mathbf{m} &= \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right), \\
\mathbf{S} &= \left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}.
\end{aligned}
\tag{3.69}
$$

By 3.10,

$$
p(t_{N+1}|\mathbf{t}) = \mathcal{N}\left(t_{N+1}|\mathbf{m}^\mathsf{T}\boldsymbol{\phi}_{N+1}, \sigma^2(\boldsymbol{\phi}_{N+1})\right), \tag{3.70}
$$

where
$$\sigma^2(\boldsymbol{\phi}) = \beta^{-1} + \boldsymbol{\phi}^\intercal \mathbf{S} \boldsymbol{\phi}. \tag{3.71}$$

Let
$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}, \mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi} \\ \boldsymbol{\phi}_{N+1}^\intercal \end{bmatrix}. \tag{3.72}$$

Then,
$$p(t_{N+2}|\mathbf{t}') = \mathcal{N}\left(t_{N+2}|\mathbf{m}'^\intercal \boldsymbol{\phi}_{N+2}, \sigma'^2(\boldsymbol{\phi}_{N+2})\right), \tag{3.73}$$

where
$$\begin{aligned}
\mathbf{m}' &= \mathbf{S}'\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \mathbf{\Phi}'^\intercal \mathbf{t}'\right), \\
\mathbf{S}' &= \left(\mathbf{S}_0^{-1} + \beta \mathbf{\Phi}'^\intercal \mathbf{\Phi}'\right)^{-1}, \\
\sigma'^2(\boldsymbol{\phi}) &= \beta^{-1} + \boldsymbol{\phi}^\intercal \mathbf{S}' \boldsymbol{\phi}.
\end{aligned} \tag{3.74}$$

We have
$$\sigma^2(\boldsymbol{\phi}) - \sigma'^2(\boldsymbol{\phi}) = \boldsymbol{\phi}^\intercal \left(\mathbf{S} - \mathbf{S}'\right) \boldsymbol{\phi}, \tag{3.75}$$

and
$$\mathbf{S}' = \left(\mathbf{S}^{-1} + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\intercal\right)^{-1}. \tag{3.76}$$

By 2.26,
$$\mathbf{S}' = \mathbf{S} - c\beta^2 \mathbf{S} \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}_{N+1}^\intercal \mathbf{S}, \tag{3.77}$$

where
$$c = \left(\beta + \beta^2 \boldsymbol{\phi}_{N+1}^\intercal \mathbf{S} \boldsymbol{\phi}_{N+1}\right)^{-1}. \tag{3.78}$$

Then,
$$\sigma^2(\boldsymbol{\phi}) - \sigma'^2(\boldsymbol{\phi}) = c\beta^2 \left(\boldsymbol{\phi}^\intercal \mathbf{S} \boldsymbol{\phi}_{N+1}\right)^2. \tag{3.79}$$

Therefore,
$$\sigma^2(\boldsymbol{\phi}) \geq \sigma'^2(\boldsymbol{\phi}). \tag{3.80}$$

3.12

Let $t_1, \cdots, t_N$ be variables such that
$$\begin{aligned}
p(t_n|\mathbf{w}, \beta) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right) \mathrm{Gam}(\beta|a_0, b_0),
\end{aligned} \tag{3.81}$$

where $\mathbf{w}$ is a vector in M dimensions.

(a)

By the Bayes' theorem,

$$p(\mathbf{w}, \beta|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}, \beta). \tag{3.82}$$

The logarithm of the right hand side except the terms independent of $\mathbf{t}$, $\mathbf{w}$ and β can be written as

$$-\frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n)^2$$
$$-\frac{M}{2}\ln\beta^{-1} - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathsf{T}}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) + (a_0 - 1)\ln\beta - b_0\beta. \tag{3.83}$$

We have

$$-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n)^2 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathsf{T}}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)$$
$$= -\frac{\beta}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v} + \beta\mathbf{v}^{\mathsf{T}}\mathbf{u} - \frac{\beta}{2}\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0^{-1}\mathbf{m}_0, \tag{3.84}$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} & -\boldsymbol{\Phi}^{\mathsf{T}} \\ -\boldsymbol{\Phi} & \mathbf{I} \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}_0^{-1}\mathbf{m}_0 \\ \mathbf{0} \end{bmatrix}. \tag{3.85}$$

The right hand side can be written as

$$-\frac{\beta}{2}(\mathbf{w} - \mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) + \frac{\beta}{2}\mathbf{m}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{m} - \frac{\beta}{2}\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0^{-1}\mathbf{m}_0, \tag{3.86}$$

where

$$\mathbf{m} = \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t}\right),$$
$$\mathbf{S} = \left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}. \tag{3.87}$$

Therefore,

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \beta^{-1}\mathbf{S}\right)\mathrm{Gam}\left(\beta|a, b\right), \tag{3.88}$$

where

$$a = a_0 + \frac{N}{2},$$
$$b = b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{1}{2}\mathbf{m}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{m}. \tag{3.89}$$

(b)

By the expression of the Byes' theorem in (a),

$$p(\mathbf{t}) = \frac{\mathcal{N}\left(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0)}{\mathcal{N}\left(\mathbf{w}|\mathbf{m}, \beta^{-1}\mathbf{S}\right)\mathrm{Gam}(\beta|a, b)}. \tag{3.90}$$

The logarithm of the right hand side can be written as

$$
\begin{aligned}
&-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 \\
&-\frac{M}{2}\ln(2\pi) - \frac{M}{2}\ln\beta^{-1} - \frac{1}{2}\det\mathbf{S}_0 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\
&+ a_0\ln b_0 - \ln\Gamma(a_0) + (a_0 - 1)\ln\beta - b_0\beta \\
&+ \frac{M}{2}\ln(2\pi) + \frac{M}{2}\ln\beta^{-1} + \frac{1}{2}\det\mathbf{S} + \frac{\beta}{2}(\mathbf{w} - \mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) \\
&- a\ln b + \ln\Gamma(a) - (a - 1)\ln\beta + b\beta \tag{3.91} \\
=&-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 \\
&- \frac{1}{2}\det\mathbf{S}_0 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) + a_0\ln b_0 - \ln\Gamma(a_0) \\
&+ \frac{1}{2}\det\mathbf{S} + \frac{\beta}{2}(\mathbf{w} - \mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) - a\ln b + \ln\Gamma(a) \\
&- (a - a_0)\ln\beta + (b - b_0)\beta.
\end{aligned}
$$

The right hand side can be written as

$$
\begin{aligned}
&-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\det\mathbf{S}_0 + a_0\ln b_0 - \ln\Gamma(a_0) \\
&\qquad\qquad + \frac{1}{2}\det\mathbf{S} - a\ln b + \ln\Gamma(a).
\end{aligned} \tag{3.92}
$$

Therefore,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}}\left(\frac{\det\mathbf{S}}{\det\mathbf{S}_0}\right)^{\frac{1}{2}}\frac{\Gamma(a)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b^a}. \tag{3.93}$$

3.13

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}, \beta) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0),
\end{aligned} \tag{3.94}
$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By marginalisation,

$$p(t_{N+1}|\mathbf{t}) = \iint p(t_{N+1}|\mathbf{w}, \beta)p(\mathbf{w}, \beta|\mathbf{t})d\mathbf{w}d\beta. \qquad (3.95)$$

By 3.12(a),

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \beta^{-1}\mathbf{S}\right)\mathrm{Gam}(\beta|a, b), \qquad (3.96)$$

where

$$\begin{aligned}
\mathbf{m} &= \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t}\right), \\
\mathbf{S} &= \left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}, \\
a &= a_0 + \frac{N}{2}, \\
b &= b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^{\mathsf{T}}\mathbf{S}_0\mathbf{m}_0 - \frac{1}{2}\mathbf{m}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{m}.
\end{aligned} \qquad (3.97)$$

Then, the integral can be written as

$$\int \left(\int \mathcal{N}\left(t_{N+1}|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}, \beta^{-1}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}, \beta^{-1}\mathbf{S}\right)d\mathbf{w}\right)\mathrm{Gam}(\beta|a, b)d\beta. \quad (3.98)$$

The logarithm of the integrand with respect to $\mathbf{w}$ except the terms indepndent of $\mathbf{w}$ can be written as

$$\begin{aligned}
&-\frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)^2 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) \\
&= -\frac{\beta}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v} + \beta\mathbf{v}^{\mathsf{T}}\mathbf{u} - \frac{\beta}{2}\mathbf{m}^{\mathsf{T}}\mathbf{S}^{-1}\mathbf{m},
\end{aligned} \qquad (3.99)$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}^{-1} + \boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathsf{T}} & -\boldsymbol{\phi}_{N+1} \\ -\boldsymbol{\phi}_{N+1}^{\mathsf{T}} & 1 \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}^{-1}\mathbf{m} \\ 0 \end{bmatrix}. \quad (3.100)$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{S} & \mathbf{S}\boldsymbol{\phi}_{N+1} \\ \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S} & 1 + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1} \end{bmatrix}. \qquad (3.101)$$

Then,

$$\mathbf{M}^{-1}\mathbf{u} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m}^{\mathsf{T}}\boldsymbol{\phi}_{N+1} \end{bmatrix}. \qquad (3.102)$$

Then, the integral with respect to $\mathbf{w}$ can be written as

$$\mathcal{N}\left(t_{N+1}|\mathbf{m}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}, \beta^{-1}\left(1 + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1}\right)\right). \qquad (3.103)$$

Then, the logarithm of the integrand with respect to β except the terms independent of β can be written as

$$
-\frac{1}{2}\ln\beta^{-1} - \frac{\beta}{2\left(1 + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1}\right)}\left(t_{N+1} - \mathbf{m}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)^2
$$
$$
+ (a-1)\ln\beta - b\beta \tag{3.104}
$$
$$
= \left(a + \frac{1}{2} - 1\right)\ln\beta - \left(b + \frac{\left(t_{N+1} - \mathbf{m}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)^2}{2\left(1 + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1}\right)}\right)\beta.
$$

Then, the integral with respect to β except the terms independent of t_{N+1} can be written as

$$
\left(b + \frac{\left(t_{N+1} - \mathbf{m}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)^2}{2\left(1 + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1}\right)}\right)^{-a-\frac{1}{2}}. \tag{3.105}
$$

Thererfore,

$$
p(t_{N+1}|\mathbf{t}) = \mathrm{St}(t_{N+1}|\mu, \lambda, \nu), \tag{3.106}
$$

where

$$
\mu = \mathbf{m}^{\mathsf{T}}\boldsymbol{\phi}_{N+1},
$$
$$
\lambda = \frac{a}{b}\left(1 + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1}\right)^{-1}, \tag{3.107}
$$
$$
\nu = 2a.
$$

3.14   (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$
p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n, \beta^{-1}\right),
$$
$$
p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right), \tag{3.108}
$$

where $\mathbf{w}$ is a vector in M dimensions.

(a)

By 3.7,

$$
p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right), \tag{3.109}
$$

where

$$
\mathbf{m} = \beta\mathbf{S}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t},
$$
$$
\mathbf{S} = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}. \tag{3.110}
$$

Let
$$y(\boldsymbol{\phi}, \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}. \tag{3.111}$$

Then,
$$y(\boldsymbol{\phi}, \mathbf{m}) = \beta\boldsymbol{\phi}^{\mathsf{T}}\mathbf{S}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t}. \tag{3.112}$$

The right hand side can be written as

$$\sum_{n=1}^{N} k(\boldsymbol{\phi}, \boldsymbol{\phi}_n)t_n, \tag{3.113}$$

where
$$k(\boldsymbol{\phi}, \boldsymbol{\phi}') = \beta\boldsymbol{\phi}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}'. \tag{3.114}$$

Let us suppose that $\phi_1, \cdots, \phi_M$ are linearly independent, $N > M$ and

$$\phi_1 = 1. \tag{3.115}$$

Then, we can construct a new basis set $\psi_1, \cdots, \psi_M$ such that

$$\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi} = \mathbf{I}? \tag{3.116}$$

where
$$\psi_1 = 1. \tag{3.117}$$

(b)

3.15

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n, \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right), \end{aligned} \tag{3.118}$$

where $\mathbf{w}$ is a vector in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right), \tag{3.119}$$

where
$$\begin{aligned} \mathbf{m} &= \beta\mathbf{S}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t}, \\ \mathbf{S} &= \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}. \end{aligned} \tag{3.120}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}) - E(\mathbf{m}), \qquad (3.121)$$

where

$$E(\mathbf{m}) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi m}\|^2 + \frac{\alpha}{2}\mathbf{m}^\intercal\mathbf{m}. \qquad (3.122)$$

By 3.22, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α and β to zero gives

$$\alpha = \frac{\gamma}{\mathbf{m}^\intercal\mathbf{m}},$$
$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \mathbf{\Phi m}\|^2}, \qquad (3.123)$$

where

$$\gamma = \sum_{m=1}^{M}\frac{\lambda_m}{\alpha + \lambda_m}, \qquad (3.124)$$

and $\lambda_1, \cdots, \lambda_M$ are the eigenvalues of $\beta\mathbf{\Phi}^\intercal\mathbf{\Phi}$. If α and β are set as above, then

$$E(\mathbf{m}) = \frac{N}{2}. \qquad (3.125)$$

3.16

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right), \qquad (3.126)$$

where $\mathbf{w}$ is a vector in M dimensions. By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \qquad (3.127)$$

The logarithm of the integrand of the right hand side except the terms independent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\intercal\boldsymbol{\phi}_n)^2 - \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w} = -\frac{1}{2}\mathbf{v}^\intercal\mathbf{M}\mathbf{v}, \qquad (3.128)$$

where
$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} & -\beta\mathbf{\Phi}^{\mathsf{T}} \\ -\beta\mathbf{\Phi} & \beta\mathbf{I} \end{bmatrix}. \tag{3.129}$$

By 2.24,
$$\mathbf{M}^{-1} = \begin{bmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\mathbf{\Phi}^{\mathsf{T}} \\ \alpha^{-1}\mathbf{\Phi} & \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}} + \beta^{-1}\mathbf{I} \end{bmatrix}. \tag{3.130}$$

Therefore,
$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}} + \beta^{-1}\mathbf{I}\right). \tag{3.131}$$

3.17

Let $t_1, \cdots, t_N$ be variables such that
$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n, \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.132}$$

where $\mathbf{w}$ is a vector in M dimensions. By marginalisation,
$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{3.133}$$

The logarithm of the integrand of the right hand side can be written as
$$\begin{aligned} &-\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n)^2 \\ &-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left(\alpha^{-1}\mathbf{I}\right)\right) - \frac{\alpha}{2}\|\mathbf{w}\|^2. \end{aligned} \tag{3.134}$$

Therefore,
$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int\exp\left(-E(\mathbf{w})\right)d\mathbf{w}, \tag{3.135}$$

where
$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2. \tag{3.136}$$

3.18

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right),
\end{aligned}
\tag{3.137}
$$

where $\mathbf{w}$ is a vector in M dimensions. By 3.17,

$$
p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left(-E(\mathbf{w})\right) d\mathbf{w},
\tag{3.138}
$$

where

$$
E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2.
\tag{3.139}
$$

We have

$$
\begin{aligned}
\nabla E(\mathbf{w}) &= \mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}), \\
\nabla\nabla E(\mathbf{w}) &= \mathbf{S}^{-1}.
\end{aligned}
\tag{3.140}
$$

where

$$
\begin{aligned}
\mathbf{m} &= \beta\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\
\mathbf{S}^{-1} &= \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \alpha\mathbf{I}.
\end{aligned}
\tag{3.141}
$$

Then,

$$
E(\mathbf{w}) = E(\mathbf{m}) + \frac{1}{2}(\mathbf{w} - \mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}).
\tag{3.142}
$$

3.19

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right),
\end{aligned}
\tag{3.143}
$$

where $\mathbf{w}$ is a vector in M dimensions. By 3.17,

$$
p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left(-E(\mathbf{w})\right) d\mathbf{w},
\tag{3.144}
$$

where

$$
E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2.
\tag{3.145}
$$

By 3.18,

$$E(\mathbf{w}) = E(\mathbf{m}) + \frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}), \qquad (3.146)$$

where

$$\mathbf{m} = \beta\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\mathbf{t},$$
$$\mathbf{S}^{-1} = \beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \alpha\mathbf{I}. \qquad (3.147)$$

Therefore, the integral in the expression above of $p(\mathbf{t})$ can be written as

$$\exp\left(-E(\mathbf{m})\right) \int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m})\right) d\mathbf{w}$$
$$= (2\pi)^{\frac{M}{2}} (\det \mathbf{S})^{\frac{1}{2}} \exp\left(-E(\mathbf{m})\right). \qquad (3.148)$$

Thus,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathbf{S}) - E(\mathbf{m}). \qquad (3.149)$$

3.20

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right), \qquad (3.150)$$

where $\mathbf{w}$ is a vector in M dimensions. By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathbf{S}) - E(\mathbf{m}), \qquad (3.151)$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2, \qquad (3.152)$$

and

$$\mathbf{m} = \beta\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\mathbf{t},$$
$$\mathbf{S}^{-1} = \beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \alpha\mathbf{I}. \qquad (3.153)$$

By 3.21, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α and β to zero gives

$$0 = \frac{M}{2\alpha} + \frac{1}{2}\operatorname{tr}\left(\mathbf{S}^{-1}\frac{\partial\mathbf{S}}{\partial\alpha}\right) - \frac{1}{2}\|\mathbf{m}\|^2,$$
$$0 = \frac{N}{2\beta} + \frac{1}{2}\operatorname{tr}\left(\mathbf{S}^{-1}\frac{\partial\mathbf{S}}{\partial\beta}\right) - \frac{1}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}\|^2. \qquad (3.154)$$

We have

$$\mathbf{O} = \frac{\partial \mathbf{S}^{-1}\mathbf{S}}{\partial \alpha},$$
$$\mathbf{O} = \frac{\partial \mathbf{S}^{-1}\mathbf{S}}{\partial \beta}. \tag{3.155}$$

The right hand side can be written as

$$\mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\partial \alpha} + \mathbf{S}\frac{\partial \mathbf{S}^{-1}}{\partial \alpha} = \mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\partial \alpha} + \mathbf{S},$$
$$\mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\partial \beta} + \mathbf{S}\frac{\partial \mathbf{S}^{-1}}{\partial \beta} = \mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\partial \beta} + \mathbf{S}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}. \tag{3.156}$$

Then,

$$\mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\partial \alpha} = -\mathbf{S},$$
$$\mathbf{S}^{-1}\frac{\partial \mathbf{S}}{\partial \beta} = -\mathbf{S}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}. \tag{3.157}$$

Therefore, the maximum likelihood solutions of $p(\mathbf{t})$ satisfy

$$\alpha = \frac{M}{\|\mathbf{m}\|^2 + \operatorname{tr}\mathbf{S}},$$
$$\beta = \frac{N}{\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}\|^2 + \operatorname{tr}(\mathbf{S}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})}. \tag{3.158}$$

3.21

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

$$\boldsymbol{\Sigma}\mathbf{u}_m = \lambda_m \mathbf{u}_m, \tag{3.159}$$

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

$$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_M),$$
$$\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_M]. \tag{3.160}$$

By 2.19,

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}},$$
$$\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}. \tag{3.161}$$

Then,

$$\det \mathbf{\Sigma} = \prod_{m=1}^{M} \lambda_m, \tag{3.162}$$

so that

$$\ln(\det \mathbf{\Sigma}) = \sum_{m=1}^{M} \ln \lambda_m. \tag{3.163}$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{\Sigma}) = \sum_{m=1}^{M} \frac{\partial \lambda_m}{\partial \alpha} \frac{1}{\lambda_m}. \tag{3.164}$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{\Sigma}) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} \frac{\partial \mathbf{\Lambda}}{\partial \alpha}\right). \tag{3.165}$$

The right hand side can be written as

$$\operatorname{tr}\left(\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\intercal}\frac{\partial \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\intercal}}{\partial \alpha}\right) = \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \alpha}\right). \tag{3.166}$$

Therefore,

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{\Sigma}) = \operatorname{tr}\left(\mathbf{\Sigma}^{-1}\frac{\partial \mathbf{\Sigma}}{\partial \alpha}\right). \tag{3.167}$$

3.22

Refer to 3.20.

3.23

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned}
p(t_n|\mathbf{w}, \beta) &= \mathcal{N}\left(t_n|\mathbf{w}^{\intercal}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\operatorname{Gam}(\beta|a_0, b_0),
\end{aligned} \tag{3.168}$$

where $\mathbf{w}$ is a vector in M dimensions. By marginalisation,

$$p(\mathbf{t}) = \iint p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}, \beta)d\mathbf{w}d\beta. \tag{3.169}$$

126

The right hand side can be written as

$$\int I(\beta)\mathrm{Gam}(\beta|a_0, b_0)d\beta, \tag{3.170}$$

where

$$I(\beta) = \int \left(\prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right)\right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right) d\mathbf{w}. \tag{3.171}$$

The logarithm of the integrand of the expression of $I(\beta)$ can be written as

$$
\begin{aligned}
&-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n)^2 \\
&-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(\beta^{-1}\mathbf{S}_0\right) - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^\mathsf{T}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\
=&-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) \\
&-\frac{\beta}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v} + \mathbf{v}^\mathsf{T}\mathbf{u} - \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0,
\end{aligned} \tag{3.172}
$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} & -\boldsymbol{\Phi}^\mathsf{T} \\ -\boldsymbol{\Phi} & \mathbf{I} \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}_0^{-1}\mathbf{m}_0 \\ \mathbf{0} \end{bmatrix}. \tag{3.173}$$

The right hand side can be written as

$$
\begin{aligned}
&-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) \\
&-\frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}(\mathbf{w}-\mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w}-\mathbf{m}) + \frac{\beta}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m} \quad -\frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0,
\end{aligned} \tag{3.174}
$$

where

$$
\begin{aligned}
\mathbf{m} &= \mathbf{S}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\mathsf{T}\mathbf{t}\right), \\
\mathbf{S} &= \left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}.
\end{aligned} \tag{3.175}
$$

Then, the logarithm of $I(\beta)$ can be written as

$$
\begin{aligned}
&-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}) \\
&-\frac{\beta}{2}\|\mathbf{t}\|^2 + \frac{\beta}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m} - \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0.
\end{aligned} \tag{3.176}
$$

127

Then, the logarithm of the integrand with respect to β can be written as

$$
\begin{aligned}
& -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}) \\
& + \frac{\beta}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m} - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0 \\
& - \ln\Gamma(a_0) + a_0\ln b_0 + (a_0 - 1)\ln\beta - b_0\beta \\
= & -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}) \\
& - \ln\Gamma(a_0) + a_0\ln b_0 + (a - 1)\ln\beta - b\beta,
\end{aligned} \tag{3.177}
$$

where

$$
\begin{aligned}
a &= a_0 + \frac{N}{2}, \\
b &= b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^\mathsf{T}\mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m}.
\end{aligned} \tag{3.178}
$$

Then, the logarithm of the integral with respect to β can be written as

$$
\begin{aligned}
& -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}) \\
& - \ln\Gamma(a_0) + a_0\ln b_0 + \ln\Gamma(a) - a\ln b.
\end{aligned} \tag{3.179}
$$

Therefore,

$$
p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left(\frac{\det \mathbf{S}}{\det \mathbf{S}_0}\right)^{\frac{1}{2}} \frac{\Gamma(a)}{\Gamma(a_0)} \frac{b_0{}^{a_0}}{b^a}. \tag{3.180}
$$

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

$$\sum_{m=1}^{M} \alpha_m = \sum_{n=1}^{N} \beta_n = 1, \tag{4.2}$$
$$\alpha_m \geq 0, \beta_n \geq 0.$$

Let us assume that $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ below are subject to the constraints above.

(a)

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

$$\hat{\mathbf{w}}^\mathsf{T} \mathbf{x}_m + w_0 > 0,$$
$$\hat{\mathbf{w}}^\mathsf{T} \mathbf{y}_n + w_0 < 0. \tag{4.6}$$

(b)

If there exist $\hat{\mathbf{w}}$ and w_0 such that

$$\hat{\mathbf{w}}^\mathsf{T}\mathbf{x}_m + w_0 > 0,$$
$$\hat{\mathbf{w}}^\mathsf{T}\mathbf{y}_n + w_0 < 0, \tag{4.7}$$

then

$$\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^\mathsf{T}\mathbf{x}_m + w_0\right) > 0,$$
$$\sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^\mathsf{T}\mathbf{y}_n + w_0\right) < 0. \tag{4.8}$$

The left hand sides can be written as

$$\hat{\mathbf{w}}^\mathsf{T} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0 \sum_{m=1}^{M} \alpha_m = \hat{\mathbf{w}}^\mathsf{T} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0,$$
$$\hat{\mathbf{w}}^\mathsf{T} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n = \hat{\mathbf{w}}^\mathsf{T} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0. \tag{4.9}$$

Then, there do not exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$\sum_{m=1}^{M} \alpha_m \mathbf{x}_m = \sum_{n=1}^{N} \beta_n \mathbf{y}_n. \tag{4.10}$$

Therefore, 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

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^\mathsf{T} \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\mathsf{T} \end{bmatrix}, \tilde{\mathbf{W}} = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}, \mathbf{T} = \begin{bmatrix} \mathbf{t}_1^\mathsf{T} \\ \vdots \\ \mathbf{t}_N^\mathsf{T} \end{bmatrix}. \tag{4.12}$$

130

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

$$\mathbf{O} = \tilde{\mathbf{X}}^\mathsf{T}(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}). \tag{4.13}$$

Therefore,

$$\operatorname*{argmin}_{\tilde{\mathbf{W}}} E(\tilde{\mathbf{W}}) = \left(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\mathsf{T}\mathbf{T}. \tag{4.14}$$

Let $\tilde{\mathbf{W}}^*$ denote the least-square solution above. Then,

$$(\tilde{\mathbf{W}}^*)^\mathsf{T}\tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{T}^\mathsf{T}\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}} - \mathbf{t}_n, \tag{4.15}$$

where $\tilde{\mathbf{x}}$ is a vector in $M + 1$ dimensions whose first element is 1. The right hand side can be written as

$$\mathbf{T}^\mathsf{T}\left(\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}} - \mathbf{v}_n\right) = \mathbf{0}? \tag{4.16}$$

where $\mathbf{v}_n$ is a vector in N dimensions whose n th element is 1 and other elements are zero. Therefore,

$$(\tilde{\mathbf{W}}^*)^\mathsf{T}\tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{0}. \tag{4.17}$$

Thus, if

$$\mathbf{a}^\mathsf{T}\mathbf{t}_n + b = 0, \tag{4.18}$$

then

$$\mathbf{a}^\mathsf{T}(\tilde{\mathbf{W}}^*)^\mathsf{T}\tilde{\mathbf{x}} + b = 0. \tag{4.19}$$

4.3   (Incomplete)

4.4

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n \in \mathcal{C}_k}\mathbf{x}_n, \tag{4.20}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Setting the derivatives of

$$\mathbf{w}^\mathsf{T}(\mathbf{m}_2 - \mathbf{m}_1) + \lambda\left(\|\mathbf{w}\|^2 - 1\right) \tag{4.21}$$

131

with respect to $\mathbf{w}$ and λ to zero gives

$$\mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} = \mathbf{0},$$
$$\|\mathbf{w}\|^2 - 1 = 0. \tag{4.22}$$

Therefore, $\mathbf{w}^\mathsf{T}(\mathbf{m}_2 - \mathbf{m}_1)$ under the constratint

$$\|\mathbf{w}\|^2 = 1 \tag{4.23}$$

is maximised if

$$\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1. \tag{4.24}$$

4.5

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \tag{4.25}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{4.26}$$

where

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2,$$
$$y_n = \mathbf{w}^\mathsf{T}\mathbf{x}_n, \tag{4.27}$$
$$m_k = \mathbf{w}^\mathsf{T}\mathbf{m}_k.$$

Then, $J(\mathbf{w})$ can be written as

$$\frac{(\mathbf{w}^\mathsf{T}(\mathbf{m}_2 - \mathbf{m}_1))^2}{\sum_{n \in \mathcal{C}_1} (\mathbf{w}^\mathsf{T}(\mathbf{x}_n - \mathbf{m}_1))^2 + \sum_{n \in \mathcal{C}_2} (\mathbf{w}^\mathsf{T}(\mathbf{x}_n - \mathbf{m}_2))^2} = \frac{\mathbf{w}^\mathsf{T}\mathbf{S}_\mathrm{B}\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{S}_\mathrm{W}\mathbf{w}}, \tag{4.28}$$

where

$$\mathbf{S}_\mathrm{B} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\mathsf{T},$$
$$\mathbf{S}_\mathrm{W} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathsf{T} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathsf{T}. \tag{4.29}$$

4.6

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \tag{4.30}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{4.31}$$

where

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2,$$
$$y_n = \mathbf{w}^\mathsf{T} \mathbf{x}_n, \tag{4.32}$$
$$m_k = \mathbf{w}^\mathsf{T} \mathbf{m}_k.$$

By 4.5,

$$J(\mathbf{w}) = \frac{\mathbf{w}^\mathsf{T} \mathbf{S}_\mathrm{B} \mathbf{w}}{\mathbf{w}^\mathsf{T} \mathbf{S}_\mathrm{W} \mathbf{w}}, \tag{4.33}$$

where

$$\mathbf{S}_\mathrm{B} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\mathsf{T},$$
$$\mathbf{S}_\mathrm{W} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathsf{T} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathsf{T}. \tag{4.34}$$

Let

$$E = \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n + w_0 - t_n\right)^2, \tag{4.35}$$

where

$$t_n = \begin{cases} \frac{N}{N_1}, & n \in \mathcal{C}_1, \\ -\frac{N}{N_2}, & n \in \mathcal{C}_2. \end{cases} \tag{4.36}$$

Setting the derivative with respect to $\mathbf{w}$ and w_0 gives

$$0 = \sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n + w_0 - t_n\right),$$
$$\mathbf{0} = \sum_{n=1}^{N} \left(\mathbf{w}^\mathsf{T} \mathbf{x}_n + w_0 - t_n\right) \mathbf{x}_n. \tag{4.37}$$

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

$$\mathbf{w}^{\mathsf{T}} \sum_{n=1}^{N} \mathbf{x}_n + N w_0 - \sum_{n=1}^{N} t_n = N \left(\mathbf{w}^{\mathsf{T}} \mathbf{m} + w_0 \right), \tag{4.38}$$

where

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n. \tag{4.39}$$

Then,

$$w_0 = -\mathbf{w}^{\mathsf{T}} \mathbf{m}. \tag{4.40}$$

Then, the right hand side of the second equation above can be written as

$$\sum_{n=1}^{N} \left(\mathbf{w}^{\mathsf{T}} \left(\mathbf{x}_n - \mathbf{m} \right) - t_n \right) \mathbf{x}_n$$

$$= \sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^{\mathsf{T}} (\mathbf{x}_n - \mathbf{m}) - \frac{N}{N_1} \right) \mathbf{x}_n + \sum_{n \in \mathcal{C}_2} \left(\mathbf{w}^{\mathsf{T}} (\mathbf{x}_n - \mathbf{m}) + \frac{N}{N_2} \right) \mathbf{x}_n. \tag{4.41}$$

We have

$$\mathbf{m} = \frac{N_1}{N} \mathbf{m}_1 + \frac{N_2}{N} \mathbf{m}_2,$$

$$\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1) = \mathbf{0}. \tag{4.42}$$

Then, the first term of the right hand side can be written as

$$\sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^{\mathsf{T}} \left(\mathbf{x}_n - \mathbf{m}_1 + \frac{N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \right) - \frac{N}{N_1} \right) (\mathbf{x}_n - \mathbf{m}_1 + \mathbf{m}_1)$$

$$= \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathsf{T}} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^{\mathsf{T}} \mathbf{w} - N \mathbf{m}_1. \tag{4.43}$$

Similarly, the second term can be written as

$$\left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathsf{T}} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^{\mathsf{T}} \mathbf{w} - N \mathbf{m}_2. \tag{4.44}$$

134

Then,

$$\mathbf{0} = \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\mathsf{T} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2)\mathbf{m}_1^\mathsf{T}\mathbf{w}$$

$$- N\mathbf{m}_1 + \left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\mathsf{T} \right) \mathbf{w} \tag{4.45}$$

$$+ \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1)\mathbf{m}_2^\mathsf{T}\mathbf{w} - N\mathbf{m}_2.$$

Therefore,

$$\left(\mathbf{S}_\mathrm{W} + \frac{N_1 N_2}{N} \mathbf{S}_\mathrm{B} \right) \mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2). \tag{4.46}$$

4.7

Let

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{4.47}$$

(a)

We have

$$\sigma(-a) = \frac{1}{1 + \exp(a)}. \tag{4.48}$$

The right hand side can be written as

$$1 - \frac{\exp(a)}{1 + \exp(a)} = 1 - \frac{1}{1 + \exp(-a)}. \tag{4.49}$$

Therefore,

$$\sigma(-a) = 1 - \sigma(a). \tag{4.50}$$

(b)

We have

$$\exp(-a) = \frac{1}{\sigma(a)} - 1. \tag{4.51}$$

Then,

$$a = -\ln\left(\frac{1}{\sigma(a)} - 1 \right). \tag{4.52}$$

135

Therefore,

$$\sigma^{-1}(y) = \ln\left(\frac{y}{1-y}\right). \tag{4.53}$$

4.8

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

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right), \tag{4.54}$$

where

$$p(\mathcal{C}_1) + p(\mathcal{C}_2) = 1. \tag{4.55}$$

By the Bayes' theorem,

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}. \tag{4.56}$$

The right hand side can be written as

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{4.57}$$

where

$$a = \ln\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right). \tag{4.58}$$

Substituting the expressions above of $p(\mathbf{x}|\mathcal{C}_k)$, we have

$$\begin{aligned} a = &-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \ln p(\mathcal{C}_1) \\ &+ \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) - \ln p(\mathcal{C}_2). \end{aligned} \tag{4.59}$$

Therefore,

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma\left(\mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0\right), \tag{4.60}$$

where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2). \end{aligned} \tag{4.61}$$

4.9

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{4.62}$$

where

$$\begin{aligned} y_{nk} &= p(\boldsymbol{\phi}_n, \mathcal{C}_k), \\ p(\mathcal{C}_k) &= \pi_k, \\ \sum_{k=1}^{K} \pi_k &= 1. \end{aligned} \tag{4.63}$$

Then,

$$p(\mathbf{T}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{4.64}$$

By the Bayes' theorem,

$$y_{nk} = \pi_k p(\boldsymbol{\phi}_n | \mathcal{C}_k). \tag{4.65}$$

Then,

$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \pi_k + \ln p(\boldsymbol{\phi}_n | \mathcal{C}_k) \right). \tag{4.66}$$

Setting the derivatives of

$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{4.67}$$

with respect to π_k and λ to zero gives

$$\begin{aligned} 0 &= \frac{1}{\pi_k} \sum_{n=1}^{N} t_{nk} + \lambda, \\ 0 &= \sum_{k=1}^{K} \pi_k - 1. \end{aligned} \tag{4.68}$$

Then,

$$\lambda = - \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}. \tag{4.69}$$

Since $\mathbf{t}_1, \cdots, \mathbf{t}_N$ are from the standard basis in K dimensions, the right hand side can be written as $-N$. Therefore, the maximum likelihood solution for π_k is given by

$$\pi_{k\,\mathrm{ML}} = \frac{N_k}{N}, \tag{4.70}$$

where

$$N_k = \sum_{n=1}^{N} t_{nk}. \tag{4.71}$$

4.10

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{4.72}$$

where

$$\begin{aligned}
y_{nk} &= p(\boldsymbol{\phi}_n, \mathcal{C}_k), \\
p(\boldsymbol{\phi}_n | \mathcal{C}_k) &= \mathcal{N}\left(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right), \\
p(\mathcal{C}_k) &= \pi_k, \\
\sum_{k=1}^{K} \pi_k &= 1.
\end{aligned} \tag{4.73}$$

Then,

$$p(\mathbf{T}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.74}$$

By the Bayes' theorem,

$$y_{nk} = \pi_k \mathcal{N}\left(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right). \tag{4.75}$$

Then,

$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \mathcal{N}\left(\boldsymbol{\phi}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right) + \ln \pi_k\right). \tag{4.76}$$

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

$$\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(-\frac{1}{2} \ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)\right). \tag{4.77}$$

138

By 3.21??, setting the derivatives with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ to zero gives

$$\mathbf{0} = \frac{1}{2}\sum_{n=1}^{N} t_{nk}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^{\intercal}\right)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k),$$

$$\mathbf{O} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left(\left(\boldsymbol{\Sigma}^{-1}\right)^{\intercal} - \left(\boldsymbol{\Sigma}^{-1}\right)^2(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\intercal}\right). \tag{4.78}$$

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

$$\boldsymbol{\mu}_{k\mathrm{ML}} = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}\boldsymbol{\phi}_n,$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{K} N_k \mathbf{S}_k, \tag{4.79}$$

where

$$N_k = \sum_{n=1}^{N} t_{nk},$$

$$\mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\intercal}. \tag{4.80}$$

4.11

Let $\boldsymbol{\phi}$ be a variable in M dimensions whose each component is a binary code with length L such that

$$p(\boldsymbol{\phi}|\mathcal{C}_k) = \prod_{m=1}^{M}\prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}}, \tag{4.81}$$

where

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1. \tag{4.82}$$

By the Bayes' theorem,

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = \frac{p(\boldsymbol{\phi}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{k'=1}^{K} p(\boldsymbol{\phi}|\mathcal{C}_{k'})p(\mathcal{C}_{k'})}. \tag{4.83}$$

Therefore,

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = \frac{\exp\left(a_k(\boldsymbol{\phi})\right)}{\sum_{k'=1}^{K}\exp\left(a_{k'}(\boldsymbol{\phi})\right)}, \tag{4.84}$$

where

$$a_k(\boldsymbol{\phi}) = \left(\sum_{m=1}^{M}\sum_{l=1}^{L}\phi_{ml}\ln\mu_{kml}\right) + \ln p(\mathcal{C}_k). \tag{4.85}$$

4.12

Let

$$\sigma(a) = \frac{1}{1+\exp(-a)}. \tag{4.86}$$

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,

$$\frac{d\sigma(a)}{da} = \sigma(a)\left(1-\sigma(a)\right). \tag{4.89}$$

4.13

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n}(1-y_n)^{1-t_n}, \tag{4.90}$$

where

$$y_n = \sigma\left(\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\right),$$
$$\sigma(a) = \frac{1}{1+\exp(-a)}. \tag{4.91}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{4.92}$$

The right hand side can be written as

$$-\ln\left(\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\right) = -\sum_{n=1}^{N}\left(t_n \ln y_n + (1-t_n)\ln(1-y_n)\right). \quad (4.93)$$

By 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{t_n}{y_n}y_n(1-y_n)\boldsymbol{\phi}_n - \frac{1-t_n}{1-y_n}y_n(1-y_n)\boldsymbol{\phi}_n\right). \quad (4.94)$$

The right hand side can be written as

$$-\sum_{n=1}^{N}\left(t_n(1-y_n)\boldsymbol{\phi}_n - (1-t_n)y_n\boldsymbol{\phi}_n\right) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \quad (4.95)$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \quad (4.96)$$

4.14

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n}(1-y_n)^{1-t_n}, \quad (4.97)$$

where

$$y_n = \sigma\left(\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\right),$$
$$\sigma(a) = \frac{1}{1+\exp(-a)}. \quad (4.98)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \quad (4.99)$$

By 4.13, setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \quad (4.100)$$

If $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N$ are linearly independent, then

$$y_n = t_n. \quad (4.101)$$

141

Then,

$$\sigma\left(\mathbf{w}^\intercal\boldsymbol{\phi}_n\right) = \begin{cases} 1, & t_n = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.102}$$

Threrefore,

$$\mathbf{w}^\intercal\boldsymbol{\phi}_n = \begin{cases} \infty, & t_n = 1, \\ -\infty, & \text{otherwise.} \end{cases} \tag{4.103}$$

4.15

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n}(1 - y_n)^{1-t_n}, \tag{4.104}$$

where

$$y_n = \sigma\left(\mathbf{w}^\intercal\boldsymbol{\phi}_n\right),$$
$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{4.105}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{4.106}$$

By 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n. \tag{4.107}$$

By 4.12,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\intercal. \tag{4.108}$$

The right hand side can be written as

$$\mathbf{H} = \boldsymbol{\Phi}^\intercal\mathbf{R}\boldsymbol{\Phi}, \tag{4.109}$$

where

$$R_{nn'} = y_n(1 - y_n)I_{nn'}, \tag{4.110}$$

Then,

$$\mathbf{u}^\intercal\mathbf{H}\mathbf{u} = \mathbf{v}^\intercal\mathbf{R}\mathbf{v}, \tag{4.111}$$

where

$$\mathbf{v} = \boldsymbol{\Phi}\mathbf{u}. \tag{4.112}$$

142

The right hand side can be written as

$$\sum_{n=1}^{N}\sum_{n'=1}^{N} v_n y_n (1 - y_n) I_{nn'} v_{n'} = \sum_{n=1}^{N} y_n (1 - y_n) v_n^2. \tag{4.113}$$

We have

$$y_n (1 - y_n) > 0. \tag{4.114}$$

Then,

$$\mathbf{u}^\mathsf{T} \mathbf{H} \mathbf{u} > 0, \tag{4.115}$$

so that $\mathbf{H}$ is positive definite. Therefore, E is a convex function of $\mathbf{w}$ and it has a unique minimum.

4.16

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n = 1 | \boldsymbol{\phi}_n) = \pi_n. \tag{4.116}$$

Then,

$$p(t_n | \boldsymbol{\phi}_n) = \pi_n^{t_n} (1 - \pi_n)^{1 - t_n}, \tag{4.117}$$

so that

$$p(\mathbf{t} | \boldsymbol{\Phi}) = \prod_{n=1}^{N} \pi_n^{t_n} (1 - \pi_n)^{1 - t_n}. \tag{4.118}$$

Therefore,

$$- \ln p(\mathbf{t} | \boldsymbol{\Phi}) = - \sum_{n=1}^{N} \left(t_n \ln \pi_n + (1 - t_n) \ln(1 - \pi_n) \right). \tag{4.119}$$

4.17

Let

$$y_k = \frac{\exp(a_k)}{\sum_{k'=1}^{K} \exp(a_{k'})}. \tag{4.120}$$

Then,

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k)}{\sum_{k'=1}^{K} \exp(a_{k'})} - \frac{\exp(2a_k)}{\left(\sum_{k'=1}^{K} \exp(a_{k'}) \right)^2}. \tag{4.121}$$

The right hand side can be written as $y_k(1 - y_k)$. If $k \neq k'$, then

$$\frac{\partial y_k}{\partial a_{k'}} = -\frac{\exp(a_k + a_{k'})}{\left(\sum_{k''=1}^{K} \exp(a_{k''})\right)^2}. \tag{4.122}$$

The right hand side can be written as $-y_k y_{k'}$. Therefore,

$$\frac{\partial y_k}{\partial a_{k'}} = y_k(I_{kk'} - y_{k'}). \tag{4.123}$$

4.18

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n | \mathbf{W}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{4.124}$$

where

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_{k'=1}^{K} \exp(a_{nk'})}, \tag{4.125}$$

$$a_{nk} = \mathbf{w}_k^\mathsf{T} \boldsymbol{\phi}_n.$$

Then,

$$p(\mathbf{T} | \mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.126}$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T} | \mathbf{W}). \tag{4.127}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{4.128}$$

By 4.17,

$$\nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk}(I_{kk'} - y_{nk'}) \frac{t_{nk}}{y_{nk}} \boldsymbol{\phi}_n. \tag{4.129}$$

144

Since $\mathbf{t}_1, \cdots, \mathbf{t}_N$ are variables from the standard basis in K dimensions, the right hand side can be written as

$$-\sum_{n=1}^{N} \left(\sum_{k=1}^{K} (I_{kk'} - y_{nk'}) t_{nk} \right) \boldsymbol{\phi}_n = -\sum_{n=1}^{N} (t_{nk'} - y_{nk'}) \boldsymbol{\phi}_n. \qquad (4.130)$$

Therefore,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^{N} (y_{nk} - t_{nk}) \boldsymbol{\phi}_n. \qquad (4.131)$$

4.19

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n = 1 | a_n) = \Phi(a_n), \qquad (4.132)$$

where

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta | 0, 1) d\theta, \qquad (4.133)$$
$$a_n = \mathbf{w}^{\mathsf{T}} \boldsymbol{\phi}_n.$$

(a)

We have

$$p(t_n | \boldsymbol{\phi}_n) = (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \qquad (4.134)$$

Then,

$$p(\mathbf{t} | \boldsymbol{\Phi}) = \prod_{n=1}^{N} (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \qquad (4.135)$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t} | \boldsymbol{\Phi}). \qquad (4.136)$$

The right hand side can be written as

$$-\sum_{n=1}^{N} (t_n \ln \Phi(a_n) + (1 - t_n) \ln (1 - \Phi(a_n))). \qquad (4.137)$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \frac{\mathcal{N}(a_n | 0, 1)}{\Phi(a_n)} - (1 - t_n) \frac{\mathcal{N}(a_n | 0, 1)}{1 - \Phi(a_n)} \right) \boldsymbol{\phi}_n. \qquad (4.138)$$

145

The right hand side can be written as

$$-\sum_{n=1}^{N}\left(\frac{t_n}{\Phi(a_n)}-\frac{1-t_n}{1-\Phi(a_n)}\right)\mathcal{N}(a_n|0,1)\phi_n$$

$$=\sum_{n=1}^{N}\frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\left(\Phi(a_n)-t_n\right)\phi_n. \tag{4.139}$$

Therefore,

$$\nabla E(\mathbf{w})=\sum_{n=1}^{N}\frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\left(\Phi(a_n)-t_n\right)\phi_n. \tag{4.140}$$

(b)

We have

$$\nabla\nabla E(\mathbf{w})=\sum_{n=1}^{N}\frac{-a_n\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\left(\Phi(a_n)-t_n\right)\phi_n\phi_n^{\mathsf{T}}$$

$$-\sum_{n=1}^{N}\frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\left(\Phi(a_n)\right)^2\left(1-\Phi(a_n)\right)}\left(\Phi(a_n)-t_n\right)\phi_n\phi_n^{\mathsf{T}}$$

$$+\sum_{n=1}^{N}\frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\Phi(a_n)\left(1-\Phi(a_n)\right)^2}\left(\Phi(a_n)-t_n\right)\phi_n\phi_n^{\mathsf{T}}$$

$$+\sum_{n=1}^{N}\frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\phi_n\phi_n^{\mathsf{T}}. \tag{4.141}$$

Therefore,

$$\nabla\nabla E(\mathbf{w})=\sum_{n=1}^{N}b_n\phi_n\phi_n^{\mathsf{T}}, \tag{4.142}$$

where

$$b_n=\left(\frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\right)^2\left(\left(\Phi(a_n)\right)^2-2t_n\Phi(a_n)+t_n\right)$$

$$-\frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}a_n\left(\Phi(a_n)-t_n\right). \tag{4.143}$$

146

4.20

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n|\mathbf{W}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{4.144}$$

where

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_{k'=1}^{K} \exp(a_{nk'})}, \tag{4.145}$$
$$a_{nk} = \mathbf{w}_k^\mathsf{T} \boldsymbol{\phi}_n.$$

Then,

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.146}$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{W}). \tag{4.147}$$

By 4.18,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^{N} \left(y_{nk} - t_{nk} \right) \boldsymbol{\phi}_n. \tag{4.148}$$

By 4.17,

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = \sum_{n=1}^{N} y_{nk}(I_{kk'} - y_{nk'})\boldsymbol{\phi}_n \boldsymbol{\phi}_n^\mathsf{T}. \tag{4.149}$$

The right hand side can be written as

$$\mathbf{H}_{kk'} = \boldsymbol{\Phi}^\mathsf{T} \mathbf{R}_{kk'} \boldsymbol{\Phi}, \tag{4.150}$$

where

$$R_{kk'nn'} = y_{nk}(I_{kk'} - y_{nk'})I_{nn'}. \tag{4.151}$$

Let

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{K1} & \cdots & \mathbf{H}_{KK} \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_K \end{bmatrix}, \tag{4.152}$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_K$ are vectors in the same dimension as $\mathbf{w}$. Then,

$$\mathbf{u}^\mathsf{T} \mathbf{H} \mathbf{u} = \sum_{k=1}^{K} \sum_{k'=1}^{K} \mathbf{v}_k^\mathsf{T} \mathbf{R}_{kk'} \mathbf{v}_{k'}, \tag{4.153}$$

where

$$\mathbf{v}_k = \mathbf{\Phi}\mathbf{u}_k. \tag{4.154}$$

The right hand side can be written as

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{n=1}^{N}\sum_{n'=1}^{N} v_{kn}y_{nk}(I_{kk'} - y_{nk'})I_{nn'}v_{k'n'} \\
&= \sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{n=1}^{N} v_{kn}y_{nk}(I_{kk'} - y_{nk'})v_{k'n}.
\end{aligned}
\tag{4.155}
$$

The right hand side can be written as

$$\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}v_{kn}\left(v_{kn} - \sum_{k'=1}^{K} y_{nk'}v_{k'n}\right). \tag{4.156}$$

By the Jensen's inequality,

$$
\begin{aligned}
&\sum_{k=1}^{K} y_{nk}v_{kn}\left(v_{kn} - \sum_{k'=1}^{K} y_{nk'}v_{k'n}\right) \\
&\geq \left(\sum_{k=1}^{K} y_{nk}v_{kn}\right)\left(\sum_{k=1}^{K} y_{nk}v_{kn} - \sum_{k'=1}^{K} y_{nk'}v_{k'n}\right),
\end{aligned}
\tag{4.157}
$$

where the right hand side is zero. Then,

$$\mathbf{u}^{\mathsf{T}}\mathbf{H}\mathbf{u} \geq 0. \tag{4.158}$$

Therefore, $\mathbf{H}$ is positive semidifinite.

4.21

Let

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta. \tag{4.159}$$

The right hand side can be written as

$$\int_{-\infty}^{0} \mathcal{N}(\theta|0,1)d\theta + \int_{0}^{a} \mathcal{N}(\theta|0,1)d\theta = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{a} \exp\left(-\frac{\theta^2}{2}\right)d\theta. \tag{4.160}$$

148

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

$$\frac{1}{\sqrt{2\pi}} \int_0^{\frac{a}{\sqrt{2}}} \exp\left(-t^2\right) \sqrt{2} dt = \frac{1}{2} \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right), \tag{4.161}$$

where

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp\left(-t^2\right) dt. \tag{4.162}$$

Therefore,

$$\Phi(a) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right). \tag{4.163}$$

4.22

Let $\mathcal{D}$ be a set of variables dependent on $\boldsymbol{\theta}$ in M dimensions. By marginalisation,

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.164}$$

Let $\boldsymbol{\theta}_{\mathrm{MAP}}$ be a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then, we have the Taylor series

$$\begin{aligned} &\ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) \\ &\simeq \ln\left(p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}})\right) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})^\mathsf{T}\mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}}), \end{aligned} \tag{4.165}$$

where

$$\mathbf{A} = -\nabla\nabla \ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MAP}}}. \tag{4.166}$$

Then,

$$\begin{aligned} &\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &\simeq p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})^\mathsf{T}\mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}})\right) d\boldsymbol{\theta}. \end{aligned} \tag{4.167}$$

The integral of the right hand side can be written as

$$(2\pi)^{\frac{M}{2}}\left(\det \mathbf{A}^{-1}\right)^{\frac{1}{2}} = (2\pi)^{\frac{M}{2}}(\det \mathbf{A})^{-\frac{1}{2}}. \tag{4.168}$$

Therefore,

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}})(2\pi)^{\frac{M}{2}}(\det \mathbf{A})^{-\frac{1}{2}}, \tag{4.169}$$

so that

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{A}). \tag{4.170}$$

4.23

Let $\mathcal{D}$ be a set of N independent and identically distributed variables dependent on $\boldsymbol{\theta}$ in M dimensions such that

$$p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0\right). \tag{4.171}$$

By 4.22,

$$
\begin{aligned}
\ln p(\mathcal{D}) \simeq{} & \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) \\
& -\frac{1}{2}\ln\left(\det\left(\mathbf{H} + \mathbf{V}_0^{-1}\right)\right),
\end{aligned} \tag{4.172}
$$

where $\boldsymbol{\theta}_{\mathrm{MAP}}$ is a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and

$$\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MAP}}}. \tag{4.173}$$

The right hand side of the approximation can be written as

$$
\begin{aligned}
& \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{V}_0) \\
& -\frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})^\mathsf{T}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) + \frac{M}{2}\ln(2\pi) \\
& -\frac{1}{2}\ln\left(\det\left(\mathbf{H} + \mathbf{V}_0^{-1}\right)\right) \\
={} & \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{1}{2}\ln\left(\det \mathbf{V}_0^{-1}\right) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})^\mathsf{T}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) \\
& -\frac{1}{2}\ln\left(\det\left(\mathbf{H} + \mathbf{V}_0^{-1}\right)\right).
\end{aligned} \tag{4.174}
$$

If $\mathbf{V}_0^{-1}$ can be neglected, the right hand side can be approximated as

$$\ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}\ln(\det \mathbf{H}). \tag{4.175}$$

Since the elements of $\mathcal{D}$ are independent and identically distributed,

$$\mathbf{H} = N\bar{\mathbf{H}}, \tag{4.176}$$

where

$$\bar{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n, \tag{4.177}$$

150

and $\mathbf{H}_1, \cdots, \mathbf{H}_N$ are respectively the one for each element. Then,

$$\det \mathbf{H} = N^M \det \bar{\mathbf{H}}. \tag{4.178}$$

Threrefore,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{M}{2}\ln N. \tag{4.179}$$

4.24

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n}(1-y_n)^{1-t_n}, \\ p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0), \tag{4.180}$$

where

$$y_n = \sigma\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n\right), \\ \sigma(a) = \frac{1}{1+\exp(-a)}. \tag{4.181}$$

By marginalisation,

$$p(t_{N+1}|\mathbf{t}) = \int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{4.182}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{4.183}$$

The lograrithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as $-E(\mathbf{w})$ where

$$E(\mathbf{w}) = \sum_{n=1}^{N}\left(t_n\ln y_n + (1-t_n)\ln(1-y_n)\right) \\ - \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathsf{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0). \tag{4.184}$$

Let $\mathbf{w}_{\mathrm{MAP}}$ be a stationary point of E. Then, we have the Taylor series

$$E(\mathbf{w}) = E(\mathbf{w}_{\mathrm{MAP}}) + \frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}}) \\ + O(\|\mathbf{w}-\mathbf{w}_{\mathrm{MAP}}\|^3), \tag{4.185}$$

151

where

$$\mathbf{A} = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\mathrm{MAP}}}. \tag{4.186}$$

Then,

$$p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{A}^{-1}\right). \tag{4.187}$$

Then,

$$\int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}$$
$$\simeq \int \sigma\left(a_{N+1}\right)^{t_{N+1}} \left(1 - \sigma(a_{N+1})\right)^{1-t_{N+1}} p(a_{N+1})da_{N+1}, \tag{4.188}$$

where

$$p(a_{N+1}) = \int \delta\left(a_{N+1} - \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{A}^{-1}\right)d\mathbf{w}, \tag{4.189}$$

and δ is the Dirac delta function. Since t_{N+1} is a binary variable, the right hand side of the approximation can be written as

$$\upsilon_{N+1}^{t_{N+1}}(1 - \upsilon_{N+1})^{1-t_{N+1}}, \tag{4.190}$$

where

$$\upsilon_{N+1} = \int \sigma(a_{N+1})p(a_{N+1})da_{N+1}. \tag{4.191}$$

We have

$$\mathrm{E}\,a_{N+1} = \int a_{N+1}p(a_{N+1})da_{N+1}. \tag{4.192}$$

The right hand side can be written as

$$\int \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{A}^{-1}\right)d\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}. \tag{4.193}$$

Then,

$$\mathrm{var}\,a_{N+1} = \int \left(a_{N+1} - \mathbf{w}_{\mathrm{MAP}}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)^2 p(a_{N+1})da_{N+1}. \tag{4.194}$$

The right hand side can be written as

$$\int \left((\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\boldsymbol{\phi}_{N+1}\right)^2 p(a_{N+1})da_{N+1} = \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\phi}_{N+1}. \tag{4.195}$$

152

Then,

$$p(a_{N+1}) = \mathcal{N}\left(a_{N+1}|\mathbf{w}_{\text{MAP}}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}, \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\phi}_{N+1}\right). \tag{4.196}$$

Therefore,

$$p(t_{N+1}|\mathbf{t}) = v_{N+1}^{t_{N+1}}(1 - v_{N+1})^{1-t_{N+1}}, \tag{4.197}$$

where

$$v_{N+1} = \int \sigma(a)\mathcal{N}\left(a|\mathbf{w}_{\text{MAP}}^{\mathsf{T}}\boldsymbol{\phi}_{N+1}, \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\phi}_{N+1}\right)da. \tag{4.198}$$

4.25

Let λ be a constant such that

$$\left.\frac{d\sigma(a)}{da}\right|_{a=0} = \left.\frac{d\Phi(\lambda a)}{da}\right|_{a=0}, \tag{4.199}$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)},$$
$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1)d\theta. \tag{4.200}$$

By 4.12,

$$\frac{d\sigma(a)}{da} = \sigma(a)\left(1 - \sigma(a)\right). \tag{4.201}$$

We have

$$\frac{d\Phi(\lambda a)}{da} = \lambda\mathcal{N}(\lambda a|0, 1). \tag{4.202}$$

Then,

$$\frac{1}{4} = \lambda(2\pi)^{-\frac{1}{2}}. \tag{4.203}$$

Therefore,

$$\lambda^2 = \frac{\pi}{8}. \tag{4.204}$$

4.26

Let

$$I(\mu) = \int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da, \tag{4.205}$$

where

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1) d\theta. \tag{4.206}$$

By the transformation

$$z = \frac{\theta - \mu}{\sigma}, \tag{4.207}$$

the right hand side can be written as

$$\int \Phi\left(\lambda(\mu + \sigma z)\right) \mathcal{N}(\mu + \sigma z|\mu, \sigma^2) \sigma dz$$
$$= \int \Phi\left(\lambda(\mu + \sigma z)\right) \mathcal{N}(z|0, 1) dz. \tag{4.208}$$

Then,

$$\frac{\partial}{\partial \mu} I(\mu) = \lambda \int \mathcal{N}\left(\lambda(\mu + \sigma z)|0, 1\right) \mathcal{N}(z|0, 1) dz. \tag{4.209}$$

The logarithm of the integrand of the right hand side can be written as

$$-\frac{1}{2}\ln(2\pi) - \frac{\lambda^2(\mu + \sigma z)^2}{2} - \frac{1}{2}\ln(2\pi) - \frac{z^2}{2}$$
$$= -\ln(2\pi) - \frac{1 + \sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1 + \sigma^2\lambda^2}\right)^2 + \frac{\mu^2\sigma^2\lambda^4}{2\left(1 + \sigma^2\lambda^2\right)} - \frac{\mu^2\lambda^2}{2}. \tag{4.210}$$

The right hand side can be written as

$$-\ln(2\pi) - \frac{1 + \sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1 + \sigma^2\lambda^2}\right)^2 - \frac{\mu^2\lambda^2}{2\left(1 + \sigma^2\lambda^2\right)}$$
$$= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(1 + \sigma^2\lambda^2\right)^{-1} - \frac{1 + \sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1 + \sigma^2\lambda^2}\right)^2 \tag{4.211}$$
$$-\ln\lambda - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\lambda^{-2} + \sigma^2\right) - \frac{\mu^2}{2\left(\lambda^{-2} + \sigma^2\right)}.$$

Then, the integral can be written as

$$\frac{1}{\lambda} \mathcal{N}\left(\mu \mid 0, \lambda^{-2} + \sigma^2\right) \int \mathcal{N}\left(z \mid -\frac{\mu \sigma \lambda^2}{1 + \sigma^2 \lambda^2}, \left(1 + \sigma^2 \lambda^2\right)^{-1}\right) dz$$
$$= \frac{1}{\lambda} \mathcal{N}\left(\mu \mid 0, \lambda^{-2} + \sigma^2\right). \tag{4.212}$$

Then,

$$\frac{\partial}{\partial \mu} I(\mu) = \mathcal{N}\left(\mu \mid 0, \lambda^{-2} + \sigma^2\right), \tag{4.213}$$

so that

$$I(\mu) = \int_{-\infty}^{\mu} \mathcal{N}\left(m \mid 0, \lambda^{-2} + \sigma^2\right) dm. \tag{4.214}$$

By the transformation

$$m' = \left(\lambda^{-2} + \sigma^2\right)^{-\frac{1}{2}} m, \tag{4.215}$$

the right hand side can be written as

$$\int_{-\infty}^{\left(\lambda^{-2} + \sigma^2\right)^{-\frac{1}{2}} \mu} \left(\lambda^{-2} + \sigma^2\right)^{-\frac{1}{2}} \mathcal{N}\left(m' \mid 0, 1\right) \left(\lambda^{-2} + \sigma^2\right)^{\frac{1}{2}} dm'$$
$$= \int_{-\infty}^{\left(\lambda^{-2} + \sigma^2\right)^{-\frac{1}{2}} \mu} \mathcal{N}\left(m' \mid 0, 1\right) dm'. \tag{4.216}$$

Therefore,

$$I(\mu) = \Phi\left(\left(\lambda^{-2} + \sigma^2\right)^{-\frac{1}{2}} \mu\right). \tag{4.217}$$

5 Neural Networks

5.1

Let

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\sum_{m=1}^{M} w_{km}^{(2)} \sigma \left(\sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) + w_{k0}^{(2)} \right), \qquad (5.1)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \qquad (5.2)$$

We have

$$\sigma(a) = \frac{\exp\left(\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)}. \qquad (5.3)$$

The right hand side can be written as

$$\frac{\exp\left(\frac{a}{2}\right) - \exp\left(-\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)} + \frac{\exp\left(-\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)} = \tanh\left(\frac{a}{2}\right) + 1 - \sigma(a). \qquad (5.4)$$

Then,

$$\sigma(a) = \frac{1}{2}\left(1 + \tanh\left(\frac{a}{2}\right)\right). \qquad (5.5)$$

Then, the argument of the expression of $y_k(\mathbf{x}, \mathbf{w})$ can be written as

$$\begin{aligned}
&\sum_{m=1}^{M} w_{km}^{(2)} \left(\frac{1}{2} \left(1 + \tanh \left(\frac{1}{2} \left(\sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)} \right) \right) \right) \right) + w_{k0}^{(2)} \\
&= \frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} \tanh \left(\frac{1}{2} \sum_{d=1}^{D} w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)} \right) + \frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} + w_{k0}^{(2)}.
\end{aligned} \qquad (5.6)$$

Therefore,

$$\begin{aligned}
&y_k(\mathbf{x}, \mathbf{w}) \\
&= \sigma \left(\frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} \tanh \left(\frac{1}{2} \sum_{d=1}^{D} w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)} \right) + \frac{1}{2} \sum_{m=1}^{M} w_{km}^{(2)} + w_{k0}^{(2)} \right).
\end{aligned} \qquad (5.7)$$

156

5.2

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

$$p(\mathbf{t}_n|\mathbf{w}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}_n, \beta^{-1}\mathbf{I}\right), \tag{5.8}$$

where

$$\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n, \mathbf{w}). \tag{5.9}$$

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

$$-\frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\|^2. \tag{5.10}$$

Therefore, maximising $p(\mathbf{T}|\mathbf{w})$ is equivalent to minimising $E(\mathbf{w})$, where

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{5.11}$$

5.3

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

$$p(\mathbf{t}_n|\mathbf{w}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}_n, \boldsymbol{\Sigma}\right), \tag{5.12}$$

where

$$\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n, \mathbf{w}). \tag{5.13}$$

(a)

The logarithm of $p(\mathbf{T}|\mathbf{w})$ except the terms independent of $\mathbf{w}$ and $\boldsymbol{\Sigma}$ can be written as

$$-\frac{1}{2}\ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}_n). \tag{5.14}$$

Setting the derivatives with respect to $\mathbf{w}$ and $\boldsymbol{\Sigma}$ to zero gives

$$\mathbf{0} = -\sum_{n=1}^{N}\frac{\partial \mathbf{y}_n}{\partial \mathbf{w}}\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}_n),$$

$$\mathbf{O} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^\mathsf{T}. \tag{5.15}$$

Therefore, the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)(\mathbf{t}_n - \mathbf{y}_n)^{\mathsf{T}}. \tag{5.16}$$

(b)

If $\boldsymbol{\Sigma}$ is fixed and known, then the maximum likelihood solution for $\mathbf{w}$ is given by minimising $E(\mathbf{w})$ where

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}_n)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{y}_n). \tag{5.17}$$

5.4

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n = 1 | \mathbf{w}) = (1 - \epsilon)y_n + \epsilon(1 - y_n), \tag{5.18}$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \tag{5.19}$$

Then,

$$p(t_n | \mathbf{w}) = ((1 - \epsilon)y_n + \epsilon(1 - y_n))^{t_n} (\epsilon y_n + (1 - \epsilon)(1 - y_n))^{1 - t_n}. \tag{5.20}$$

Therefore,

$$\begin{aligned}
-\ln p(\mathbf{t} | \mathbf{w}) = &-\sum_{n=1}^{N} t_n \ln((1 - \epsilon)y_n + \epsilon(1 - y_n)) \\
&- \sum_{n=1}^{N} (1 - t_n) \ln(\epsilon y_n + (1 - \epsilon)(1 - y_n)).
\end{aligned} \tag{5.21}$$

5.5

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n | \mathbf{w}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{5.22}$$

158

where
$$y_{nk} = y_k(\mathbf{x}_n, \mathbf{w}). \tag{5.23}$$

Therefore,
$$\ln p(\mathbf{T}|\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.24}$$

5.6

Let $t_1, \cdots, t_N$ be binary variables such that
$$p(t_n|\mathbf{w}) = y_n^{t_n}(1 - y_n)^{1-t_n}, \tag{5.25}$$

where
$$y_n = \sigma(a_n),$$
$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{5.26}$$
$$a_n = a(\mathbf{x}_n, \mathbf{w}).$$

Let
$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{5.27}$$

Then,
$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right). \tag{5.28}$$

By 4.12,
$$\frac{\partial E(\mathbf{w})}{\partial a_n} = -y_n(1 - y_n)\left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n} \right). \tag{5.29}$$

Therefore,
$$\frac{\partial E(\mathbf{w})}{\partial a_n} = y_n - t_n. \tag{5.30}$$

5.7

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimenstions such that
$$p(\mathbf{t}_n|\mathbf{w}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{5.31}$$

159

where

$$y_{nk} = y_k(\mathbf{a}_n),$$

$$y_k(\mathbf{a}) = \frac{\exp(a_k)}{\sum_{k'=1}^{K} \exp(a_{k'})}, \tag{5.32}$$

$$\mathbf{a}_n = \mathbf{a}(\mathbf{x}_n, \mathbf{w}).$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{T}|\mathbf{w}). \tag{5.33}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.34}$$

Then,

$$\frac{\partial E(\mathbf{w})}{\partial a_{nk'}} = -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} \frac{\partial y_{nk}}{\partial a_{nk'}}. \tag{5.35}$$

By 4.17,

$$\frac{\partial y_{nk}}{\partial a_{nk'}} = y_{nk}(I_{kk'} - y_{nk'}). \tag{5.36}$$

Then,

$$\frac{\partial E(\mathbf{w})}{\partial a_{nk'}} = \sum_{k=1}^{K} t_{nk}(y_{nk'} - I_{kk'}). \tag{5.37}$$

The right hand side can be written as

$$y_{nk'} \sum_{k=1}^{K} t_{nk} - t_{nk'}. \tag{5.38}$$

Since $\mathbf{t}_1, \cdots, \mathbf{t}_N$ are from the standard basis in K dimensions,

$$\sum_{k=1}^{K} t_{nk} = 1. \tag{5.39}$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_{nk}} = y_{nk} - t_{nk}. \tag{5.40}$$

160

5.8

We have

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \tag{5.41}$$

Then,

$$\frac{d}{da} \tanh a = 1 - \left(\frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \right)^2. \tag{5.42}$$

Therefore,

$$\frac{d}{da} \tanh a = 1 - (\tanh a)^2. \tag{5.43}$$

5.9

Let $t_1, \cdots, t_N$ be variables from $\{-1, 1\}$ such that

$$p(t_n|\mathbf{w}) = \left(\frac{1 + y_n}{2} \right)^{\frac{1+t_n}{2}} \left(\frac{1 - y_n}{2} \right)^{\frac{1-t_n}{2}}, \tag{5.44}$$

where

$$\begin{aligned} y_n &= y(\mathbf{x}_n, \mathbf{w}), \\ -1 &\leq y(\mathbf{x}_n, \mathbf{w}) \leq 1. \end{aligned} \tag{5.45}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{5.46}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left(\frac{1 + t_n}{2} \ln \frac{1 + y_n}{2} + \frac{1 - t_n}{2} \ln \frac{1 - y_n}{2} \right). \tag{5.47}$$

The appropriate choice of y is tanh.

5.10

Let

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}), \tag{5.48}$$

where E is a real function of real vectors. Then, $\mathbf{H}$ is a real symmetric matrix. Therefore, by 2.20, $\mathbf{H}$ is positive definite if and only if all of its eigenvalues are positive.

5.11

Let $\mathbf{w}$ be a real vector in M dimensions. We have the Taylor series

$$E(\mathbf{w}) \simeq E\left(\mathbf{w}^*\right) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^*\right)^{\mathsf{T}} \mathbf{H}\left(\mathbf{w} - \mathbf{w}^*\right) + O\left(\|\mathbf{w} - \mathbf{w}^*\|^3\right), \qquad (5.49)$$

where

$$\begin{aligned}
\nabla E\left(\mathbf{w}^*\right) &= \mathbf{0}, \\
\mathbf{H} &= \left.\nabla\nabla E(\mathbf{w})\right|_{\mathbf{w}=\mathbf{w}^*}.
\end{aligned} \qquad (5.50)$$

Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ be eigenvectors of $\mathbf{H}$ such that

$$\mathbf{H}\mathbf{u}_m = \lambda_m \mathbf{u}_m. \qquad (5.51)$$

Then, there exists $\alpha_1, \cdots, \alpha_M$ such that

$$\mathbf{w} - \mathbf{w}^* = \sum_{m=1}^{M} \alpha_m \mathbf{u}_m. \qquad (5.52)$$

By 2.19, since $\mathbf{H}$ is a real symmetric matrix,

$$\mathbf{u}_m^{\mathsf{T}} \mathbf{u}_{m'} = I_{mm'}. \qquad (5.53)$$

Then,

$$E(\mathbf{w}) \simeq E\left(\mathbf{w}^*\right) + \frac{1}{2}\sum_{m=1}^{M} \lambda_m \alpha_m^2. \qquad (5.54)$$

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. We have the Taylor series

$$E(\mathbf{w}) = E\left(\mathbf{w}^*\right) + \frac{1}{2}\left(\mathbf{w} - \mathbf{w}^*\right)^{\mathsf{T}} \mathbf{H}\left(\mathbf{w} - \mathbf{w}^*\right) + O\left(\|\mathbf{w} - \mathbf{w}^*\|^3\right), \qquad (5.55)$$

where

$$\begin{aligned}
\nabla E\left(\mathbf{w}^*\right) &= \mathbf{0}, \\
\mathbf{H} &= \left.\nabla\nabla E(\mathbf{w})\right|_{\mathbf{w}=\mathbf{w}^*}.
\end{aligned} \qquad (5.56)$$

162

If $\mathbf{H}$ is positive definite, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \tag{5.57}$$

Then, $\mathbf{w}^*$ is a local minimum of E. On the other hand, if $\mathbf{w}^*$ is a local minimum of E, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \tag{5.58}$$

Then, $\mathbf{H}$ is positive definite. Therefore, the necessary and sufficient condition for $\mathbf{w}^*$ to be a local minimum is that $\mathbf{H}$ is positive definite.

5.13

Let $\mathbf{w}$ be a vector in M dimensions. We have the Taylor series

$$\begin{aligned} E(\mathbf{w}) =& E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^\mathsf{T} \mathbf{b} + \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\mathsf{T} \mathbf{H}(\mathbf{w} - \hat{\mathbf{w}}) \\ &+ O\left(\|\mathbf{w} - \hat{\mathbf{w}}\|^3\right), \end{aligned} \tag{5.59}$$

where

$$\begin{aligned} \mathbf{b} &= \nabla E(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}. \\ \mathbf{H} &= \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}. \end{aligned} \tag{5.60}$$

Since $\mathbf{b}$ is a vector in M dimensions and $\mathbf{H}$ is a $M \times M$ symmetric matrix, the number of independent elements of the right hand side is

$$M + \frac{M(M+1)}{2} = \frac{M(M+3)}{2}. \tag{5.61}$$

5.14

Let w be a variable. We have the Taylor series

$$\begin{aligned} E_n(w_{mm'} + \epsilon) &= E_n(w_{mm'}) + \left.\frac{\partial E_n}{\partial w}\right|_{w=w_{mm'}} \epsilon + O\left(\epsilon^2\right), \\ E_n(w_{mm'} - \epsilon) &= E_n(w_{mm'}) - \left.\frac{\partial E_n}{\partial w}\right|_{w=w_{mm'}} \epsilon + O\left(\epsilon^2\right). \end{aligned} \tag{5.62}$$

Therefore,

$$\left.\frac{\partial E_n}{\partial w}\right|_{w=w_{mm'}} = \frac{E_n(w_{mm'} + \epsilon) - E_n(w_{mm'} - \epsilon)}{2\epsilon} + O\left(\epsilon^2\right). \tag{5.63}$$

5.15   (Incomplete)

Let

$$J_{km} = \frac{\partial y_k}{\partial x_m}. \tag{5.64}$$

The right hand side can be written as

$$\sum_{m'=1}^{M} W_{mm'} \frac{\partial y_k}{\partial a_{m'}}, \tag{5.65}$$

where

$$W_{mm'} = \frac{\partial a_{m'}}{\partial x_m}. \tag{5.66}$$

We have

$$\frac{\partial y_k}{\partial a_{m'}} = \sum_{m''=1}^{M} \frac{\partial y_k}{\partial a_{m''}} \frac{\partial a_{m''}}{\partial a_{m'}}. \tag{5.67}$$

5.16

Let $\mathbf{y}_1, \cdots, \mathbf{y}_N$ be vectors dependent on $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{t}_n\|^2. \tag{5.68}$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (\nabla \mathbf{y}_n)^{\mathsf{T}} (\mathbf{y}_n - \mathbf{t}_n). \tag{5.69}$$

Therefore,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\nabla \operatorname{vec} (\nabla \mathbf{y}_n)^{\mathsf{T}} \right)^{\mathsf{T}} (\mathbf{y}_n - \mathbf{t}_n) + \sum_{n=1}^{N} (\nabla \mathbf{y}_n)^{\mathsf{T}} (\nabla \mathbf{y}_n). \tag{5.70}$$

5.17

Let y be a scalar dependent on $\mathbf{x}$ and $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \int \int (y - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt. \tag{5.71}$$

Then,

$$\nabla E(\mathbf{w}) = \int \int (y - t) p(\mathbf{x}, t) \nabla y d\mathbf{x} dt. \tag{5.72}$$

The right hand side can be written as

$$\int y \nabla y \left(\int p(\mathbf{x}, t) dt \right) d\mathbf{x} - \int \nabla y \left(\int t p(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x}$$

$$= \int y \nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla y \, \mathrm{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5.73}$$

Then,

$$\nabla \nabla E(\mathbf{w}) = \int \nabla y \left(\nabla y \right)^\mathsf{T} p(\mathbf{x}) d\mathbf{x} + \int y \nabla \nabla y p(\mathbf{x}) d\mathbf{x}$$

$$- \int \nabla \nabla y \, \mathrm{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5.74}$$

The second and the third terms of the right hand side can be written as

$$\mathrm{E} \left(y \nabla \nabla y \right) - \mathrm{E} \left(\nabla \nabla y \, \mathrm{E}(t|\mathbf{x}) \right) = \mathrm{E} \left((y - \mathrm{E}(t|\mathbf{x})) \nabla \nabla y \right). \tag{5.75}$$

Therefore, if

$$y = \mathrm{E}(t|\mathbf{x}), \tag{5.76}$$

then

$$\nabla \nabla E(\mathbf{w}) = \int \nabla y \left(\nabla y \right)^\mathsf{T} p(\mathbf{x}) d\mathbf{x}. \tag{5.77}$$

5.18

Let $y_1, \cdots, y_N$ be scalars such that

$$y_n = \mathbf{w}_n^{(2)\mathsf{T}} \mathbf{z} + \mathbf{w}_n^{(0)\mathsf{T}} \mathbf{x},$$

$$z_m = \tanh \left(\mathbf{w}_m^{(1)\mathsf{T}} \mathbf{x} \right). \tag{5.78}$$

Let

$$E = \frac{1}{2} \sum_{n=1}^{N} (y_n - t_n)^2. \tag{5.79}$$

165

Then,

$$\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} = \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}},$$

$$\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} = \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{w}_m^{(1)}}, \qquad (5.80)$$

$$\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} = \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.$$

Therefore,

$$\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} = (y_n - t_n)\mathbf{x},$$

$$\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} = (y_n - t_n)\mathbf{A}\mathbf{w}_n^{(2)}, \qquad (5.81)$$

$$\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} = (y_n - t_n)\mathbf{z},$$

where

$$A_{mm'} = \left(1 - z_m^2\right) x_{m'}. \qquad (5.82)$$

5.19

Let $t_1, \cdots, t_N$ be binary variables such that

$$p(t_n|\mathbf{w}) = y_n^{t_n}(1 - y_n)^{1-t_n}, \qquad (5.83)$$

where

$$y_n = \sigma(a_n),$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \qquad (5.84)$$

$$a_n = a(\mathbf{x}_n, \mathbf{w}).$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \qquad (5.85)$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\right). \qquad (5.86)$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n}\right)\nabla y_n. \qquad (5.87)$$

166

We have

$$\nabla y_n = \frac{\partial y_n}{\partial a_n} \nabla a_n. \tag{5.88}$$

By 4.12,

$$\frac{\partial y_n}{\partial a_n} = y_n(1 - y_n). \tag{5.89}$$

Then,

$$\nabla y_n = y_n(1 - y_n)\nabla a_n. \tag{5.90}$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n(1 - y_n) - (1 - t_n)y_n \right) \nabla a_n, \tag{5.91}$$

so that

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\nabla a_n. \tag{5.92}$$

Then,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \nabla a_n \left(\nabla y_n\right)^{\mathsf{T}} + \sum_{n=1}^{N} (y_n - t_n)\nabla\nabla a_n. \tag{5.93}$$

Therefore,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\nabla a_n(\nabla a_n)^{\mathsf{T}} + \sum_{n=1}^{N} (y_n - t_n)\nabla\nabla a_n. \tag{5.94}$$

5.20

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n|\mathbf{w}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{5.95}$$

where

$$y_{nk} = y_k(\mathbf{a}_n),$$
$$y_k(\mathbf{a}) = \frac{\exp(a_k)}{\sum_{k'=1}^{K} \exp(a_{k'})}, \tag{5.96}$$
$$\mathbf{a}_n = \mathbf{a}(\mathbf{x}_n, \mathbf{w})$$

Let
$$E(\mathbf{w}) = -\ln p(\mathbf{T}|\mathbf{w}). \tag{5.97}$$

Then,
$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.98}$$

Then,
$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}}\nabla y_{nk}. \tag{5.99}$$

We have
$$\nabla y_{nk} = \sum_{k'=1}^{K} \frac{\partial y_{nk}}{\partial a_{nk'}}\nabla a_{nk'}. \tag{5.100}$$

By 4.17,
$$\frac{\partial y_{nk}}{\partial a_{nk'}} = y_{nk}(I_{kk'} - y_{nk'}). \tag{5.101}$$

Then,
$$\nabla y_{nk} = y_{nk}\sum_{k'=1}^{K}(I_{kk'} - y_{nk'})\nabla a_{nk'}. \tag{5.102}$$

Then,
$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\sum_{k'=1}^{K}\left(I_{kk'} - y_{nk'}\right)\nabla a_{nk}, \tag{5.103}$$

so that
$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla a_{nk}. \tag{5.104}$$

Then,
$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\nabla a_{nk}\left(\nabla y_{nk}\right)^{\mathsf{T}} + \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla\nabla a_{nk}. \tag{5.105}$$

Therefore,
$$\begin{aligned} \nabla\nabla E(\mathbf{w}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}\nabla a_{nk}\sum_{k'=1}^{K}(I_{kk'} - y_{nk'})\left(\nabla a_{nk'}\right)^{\mathsf{T}} \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla\nabla a_{nk}. \end{aligned} \tag{5.106}$$

168

5.21

Let $\mathbf{y}_1, \cdots, \mathbf{y}_N$ be vectors dependent on $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{t}_n\|^2. \qquad (5.107)$$

By 5.16,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\nabla \operatorname{vec}\left(\nabla \mathbf{y}_n\right)^\mathsf{T}\right)^\mathsf{T} \left(\mathbf{y}_n - \mathbf{t}_n\right) + \mathbf{H}, \qquad (5.108)$$

where

$$\mathbf{H} = \sum_{n=1}^{N} \left(\nabla \mathbf{y}_n\right)^\mathsf{T} \left(\nabla \mathbf{y}_n\right). \qquad (5.109)$$

Let

$$\mathbf{H}' = \sum_{n=1}^{N+1} \left(\nabla \mathbf{y}_n\right)^\mathsf{T} \left(\nabla \mathbf{y}_n\right). \qquad (5.110)$$

Then,

$$\mathbf{H}' = \mathbf{H} + \mathbf{B}^\mathsf{T} \mathbf{B}, \qquad (5.111)$$

where

$$\mathbf{B} = \nabla \mathbf{y}_{N+1}. \qquad (5.112)$$

Therefore, by 2.26,

$$\mathbf{H'}^{-1} = \mathbf{H}^{-1} - \mathbf{H}^{-1} \mathbf{B}^\mathsf{T} \mathbf{M} \mathbf{B} \mathbf{H}^{-1}, \qquad (5.113)$$

where

$$\mathbf{M} = \left(\mathbf{I} + \mathbf{B} \mathbf{H}^{-1} \mathbf{B}^\mathsf{T}\right)^{-1}. \qquad (5.114)$$

5.22

Let $a_1, \cdots, a_N$ be variables such that

$$\begin{aligned} a_n &= \mathbf{w}_n^{(2)\mathsf{T}} \mathbf{z}, \\ z_m &= h(b_m), \\ b_m &= \mathbf{w}_m^{(1)\mathsf{T}} \mathbf{x}. \end{aligned} \qquad (5.115)$$

Let E be a function of $a_1, \cdots, a_N$. Then,

$$\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} = \sum_{n=1}^{N} \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial z_m} \frac{\partial z_m}{\partial b_m} \frac{\partial b_m}{\partial \mathbf{w}_m^{(1)}},$$

$$\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} = \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial \mathbf{w}_n^{(2)}}. \tag{5.116}$$

Then,

$$\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} = h'(b_m) \mathbf{x} \sum_{n=1}^{N} w_{nm}^{(2)} \frac{\partial E}{\partial a_n},$$

$$\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} = \frac{\partial E}{\partial a_n} \mathbf{z}. \tag{5.117}$$

Therefore,

$$\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} = C_{mm'}^{(11)} \mathbf{x} \mathbf{x}^\mathsf{T},$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} = C_{mn}^{(12)} \mathbf{x} \mathbf{z}^\mathsf{T} + h'(b_m) \frac{\partial E}{\partial a_n} \mathbf{x} \mathbf{v}^\mathsf{T}, \tag{5.118}$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} = C_{nn'}^{(22)} \mathbf{z} \mathbf{z}^\mathsf{T},$$

where $\mathbf{v}$ is a vector from the standard basis in M dimensions such that

$$v_m = 1, \tag{5.119}$$

and

$$C_{mm'}^{(11)} = h''(b_m) I_{mm'} \sum_{n=1}^{N} w_{nm}^{(2)} \frac{\partial E}{\partial a_n}$$

$$+ h'(b_m) h'(b_{m'}) \sum_{n=1}^{N} w_{nm}^{(2)} \sum_{n'=1}^{N} w_{n'm'}^{(2)} \frac{\partial^2 E}{\partial y_n \partial a_{n'}}, \tag{5.120}$$

$$C_{mn}^{(12)} = h'(b_m) \sum_{n'=1}^{N} w_{n'm}^{(2)} \frac{\partial^2 E}{\partial a_n \partial a_{n'}},$$

$$C_{nn'}^{(22)} = \frac{\partial^2 E}{\partial a_n \partial a_{n'}}.$$

170

5.23

Let $a_1, \cdots, a_N$ be variables such that

$$
\begin{aligned}
a_n &= \mathbf{w}_n^{(2)\mathsf{T}}\mathbf{z} + \mathbf{w}_n^{(0)\mathsf{T}}\mathbf{x}, \\
z_m &= h(b_m), \\
b_m &= \mathbf{w}_m^{(1)\mathsf{T}}\mathbf{x}.
\end{aligned}
\tag{5.121}
$$

Let E be a function of $a_1, \cdots, a_N$. Then,

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial a_n}\frac{\partial a_n}{\partial \mathbf{w}_n^{(0)}}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \sum_{n=1}^{N} \frac{\partial E}{\partial a_n}\frac{\partial a_n}{\partial z_m}\frac{\partial z_m}{\partial b_m}\frac{\partial b_m}{\partial \mathbf{w}_m^{(1)}}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial a_n}\frac{\partial a_n}{\partial \mathbf{w}_n^{(2)}}.
\end{aligned}
\tag{5.122}
$$

Then,

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial a_n}\mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= h'(b_m)\mathbf{x}\sum_{n=1}^{N} w_{nm}^{(2)}\frac{\partial E}{\partial a_n}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial a_n}\mathbf{z}.
\end{aligned}
\tag{5.123}
$$

Therefore,

$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(0)}} = C_{nn'}^{(22)} \mathbf{x}\mathbf{x}^\mathsf{T},$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_m^{(1)}} = C_{nm}^{(12)} \mathbf{x}\mathbf{x}^\mathsf{T},$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(2)}} = C_{nn'}^{(22)} \mathbf{x}\mathbf{z}^\mathsf{T},$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} = C_{mm'}^{(11)} \mathbf{x}\mathbf{x}^\mathsf{T}, \tag{5.124}$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} = C_{mn}^{(12)} \mathbf{x}\mathbf{z}^\mathsf{T} + h'(b_m)\frac{\partial E}{\partial a_n}\mathbf{x}\mathbf{v}^\mathsf{T},$$

$$\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} = C_{nn'}^{(22)} \mathbf{z}\mathbf{z}^\mathsf{T},$$

where $\mathbf{v}$ is a vector from the standard basis in M dimensions such that

$$v_m = 1, \tag{5.125}$$

and

$$C_{mm'}^{(11)} = h''(b_m) I_{mm'} \sum_{n=1}^{N} w_{nm}^{(2)} \frac{\partial E}{\partial a_n}$$

$$+ h'(b_m)h'(b_{m'}) \sum_{n=1}^{N} w_{nm}^{(2)} \sum_{n'=1}^{N} w_{n'm'}^{(2)} \frac{\partial^2 E}{\partial a_n a_{n'}},$$

$$C_{nm}^{(12)} = h'(b_m) \sum_{n'=1}^{N} w_{n'm}^{(2)} \frac{\partial^2 E}{\partial a_n \partial a_n'}, \tag{5.126}$$

$$C_{nn'}^{(22)} = \frac{\partial^2 E}{\partial a_n \partial a_{n'}}.$$

5.24

Let $y_1, \cdots, y_N$ be variables such that

$$y_n = \mathbf{w}_n^\mathsf{T}\mathbf{z} + w_{n0},$$

$$z_m = h\left(\mathbf{w}_m^\mathsf{T}\mathbf{x} + w_{m0}\right). \tag{5.127}$$

(a)

Let
$$\tilde{\mathbf{x}} = a\mathbf{x} + b\mathbf{v}, \tag{5.128}$$

where
$$\mathbf{v} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Then,
$$z_m = h\left(\frac{1}{a}\mathbf{w}_m^\mathsf{T}\left(\tilde{\mathbf{x}} - b\mathbf{v}\right) + w_{m0}\right). \tag{5.129}$$

Therefore,
$$\tilde{z}_m = h\left(\tilde{\mathbf{w}}_m^\mathsf{T}\mathbf{x} + \tilde{w}_{m0}\right), \tag{5.130}$$

where
$$\tilde{\mathbf{w}}_m = \frac{1}{a}\mathbf{w}_m,$$
$$\tilde{w}_{m0} = w_{m0} - \frac{b}{a}\mathbf{w}_m^\mathsf{T}\mathbf{v}. \tag{5.131}$$

(b)

Let
$$\tilde{y}_n = cy_n + d. \tag{5.132}$$

Then,
$$\frac{\tilde{y}_n - d}{c} = \mathbf{w}_n^\mathsf{T}\mathbf{z} + w_{n0}. \tag{5.133}$$

Therefore,
$$\tilde{y}_n = \tilde{\mathbf{w}}_n^\mathsf{T}\mathbf{z} + \tilde{w}_{n0}, \tag{5.134}$$

where
$$\tilde{\mathbf{w}}_n = c\mathbf{w}_n,$$
$$\tilde{w}_{n0} = cw_{n0} + d. \tag{5.135}$$

5.25

Let E be a quadratic error function such that

$$E(\mathbf{w}) = E_0 + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\mathsf{T} \mathbf{H} (\mathbf{w} - \mathbf{w}^*), \tag{5.136}$$

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

$$\begin{aligned} \mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \rho \nabla E, \\ \mathbf{w}^{(0)} &= \mathbf{0}. \end{aligned} \tag{5.137}$$

(a)

We have

$$\nabla E = \mathbf{H} (\mathbf{w} - \mathbf{w}^*). \tag{5.138}$$

Then,

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho \mathbf{H} \left(\mathbf{w}^{(\tau-1)} - \mathbf{w}^* \right). \tag{5.139}$$

so that

$$\mathbf{u}_m^\mathsf{T} \mathbf{w}^{(\tau)} = \mathbf{u}_m^\mathsf{T} \mathbf{w}^{(\tau-1)} - \rho \mathbf{u}_m^\mathsf{T} \mathbf{H} \left(\mathbf{w}^{(\tau-1)} - \mathbf{w}^* \right). \tag{5.140}$$

Since $\mathbf{H}$ is symmetric,

$$\mathbf{u}_m^\mathsf{T} \mathbf{H} \mathbf{w} = \mathbf{w}^\mathsf{T} \mathbf{H} \mathbf{u}_m. \tag{5.141}$$

The right hand side can be written as $\eta_m w_m$, where

$$w_m = \mathbf{w}^\mathsf{T} \mathbf{u}_m. \tag{5.142}$$

Then,

$$w_m^{(\tau)} = w_m^{(\tau-1)} - \rho \eta_m \left(w_m^{(\tau-1)} - w_m^* \right), \tag{5.143}$$

so that

$$w_m^{(\tau)} - w_m^* = (1 - \rho \eta_m) \left(w_m^{(\tau-1)} - w_m^* \right). \tag{5.144}$$

Therefore,

$$w_m^{(\tau)} = w_m^* + (1 - \rho \eta_m)^\tau \left(w_m^{(0)} - w_m^* \right), \tag{5.145}$$

so that

$$w_m^{(\tau)} = (1 - (1 - \rho \eta_m)^\tau) w_m^*. \tag{5.146}$$

174

(b)

By (a), if

$$|1 - \rho\eta_m| < 1, \tag{5.147}$$

then

$$\lim_{\tau \to \infty} w_m^{(\tau)} = w_m^*. \tag{5.148}$$

(c)

By (a) and the Taylor series

$$(1 - \rho\eta_m)^\tau = 1 - \rho\tau\eta_m + \rho^2\tau^2 O\left(\eta_m^2\right), \tag{5.149}$$

we have

$$w_m^{(\tau)} = \left(\rho\tau\eta_m - \rho^2\tau^2 O\left(\eta_m^2\right)\right) w_m^*, \tag{5.150}$$

so that

$$w_m^{(\tau)} = \rho\tau\eta_m \left(1 - \rho\tau\eta_m O\left(\eta_m\right)\right) w_m^*. \tag{5.151}$$

Therefore,

$$\left|w_m^{(\tau)}\right| \ll |w_m|^*, \quad \text{if} \quad \eta_m \ll (\rho\tau)^{-1}, \tag{5.152}$$

while

$$w_m^{(\tau)} \simeq w_m^*, \quad \text{if} \quad \eta_m \gg (\rho\tau)^{-1}. \tag{5.153}$$

5.26 (Incomplete)

Let

$$\tilde{E} = E + \lambda\Omega, \tag{5.154}$$

where

$$\Omega = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\frac{\partial y_{nk}}{\partial \xi}\right)^2. \tag{5.155}$$

(a)

We have

$$\frac{\partial y_{nk}}{\partial \xi} = \mathcal{G} y_{nk}, \tag{5.156}$$

where

$$\mathcal{G} = \sum_{d=1}^{D} \tau_d \frac{\partial}{\partial x_d},$$

$$\tau_d = \frac{\partial x_d}{\partial \xi}. \tag{5.157}$$

Therefore,

$$\Omega = \sum_{n=1}^{N} \Omega_n, \tag{5.158}$$

where

$$\Omega_n = \frac{1}{2} \sum_{k=1}^{K} (\mathcal{G} y_{nk})^2. \tag{5.159}$$

(b)

$$\alpha_j = \mathcal{G} z_j = h'(a_j)\beta_j?$$

$$\beta_j = \mathcal{G} a_j = \sum_{d=1}^{D} w_{jd}\alpha_d? \tag{5.160}$$

5.27

Let

$$\tilde{E} = \frac{1}{2} \iiint \left(y\left(\mathbf{x} + \boldsymbol{\xi}\right) - t \right)^2 p(t|\mathbf{x}) p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} dt d\boldsymbol{\xi}, \tag{5.161}$$

where

$$\mathrm{E}\,\boldsymbol{\xi} = 0,$$

$$\mathrm{var}\,\boldsymbol{\xi} = \lambda \mathbf{I}. \tag{5.162}$$

By the Taylor series

$$y(\mathbf{x} + \boldsymbol{\xi}) = y(\mathbf{x}) + \rho(\mathbf{x}, \boldsymbol{\xi}) + O\left(\|\boldsymbol{\xi}\|^3\right), \tag{5.163}$$

where

$$\rho(\mathbf{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\mathsf{T} \nabla y(\mathbf{x}) + \frac{1}{2}\boldsymbol{\xi}^\mathsf{T} \nabla\nabla y(\mathbf{x})\boldsymbol{\xi}, \tag{5.164}$$

the integrand can be written as

$$(y(\mathbf{x}) - t)^2 + 2\left(y(\mathbf{x}) - t\right)\left(\rho(\mathbf{x}, \boldsymbol{\xi}) + O\left(\|\boldsymbol{\xi}\|^3\right)\right)$$

$$+ \left(\rho(\mathbf{x}, \boldsymbol{\xi}) + O\left(\|\boldsymbol{\xi}\|^3\right)\right)^2. \tag{5.165}$$

Omitting the terms of $O(\|\boldsymbol{\xi}\|^3)$, the integral can be approximated as

$$2E + 2 \iiint (y(\mathbf{x}) - t)\, \rho(\mathbf{x}, \boldsymbol{\xi}) p(t|\mathbf{x}) p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} dt d\boldsymbol{\xi}$$
$$+ \iint \rho(\mathbf{x}, \boldsymbol{\xi})^2 p(\mathbf{x}) p(\boldsymbol{\xi}) d\mathbf{x} d\boldsymbol{\xi}, \tag{5.166}$$

where

$$E = \frac{1}{2} \iint (y(\mathbf{x}) - t)^2\, p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt. \tag{5.167}$$

Omitting the terms of $O(\|\boldsymbol{\xi}\|^3)$, the second and third terms can be approximated as

$$2 \left(\int \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi} \right)^{\mathsf{T}} \iint (y(\mathbf{x}) - t)\, \nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt$$
$$+ \int \boldsymbol{\xi}^{\mathsf{T}} \left(\iint (y(\mathbf{x}) - t)\, \nabla\nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \right) \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi} \tag{5.168}$$
$$+ \int \boldsymbol{\xi}^{\mathsf{T}} \left(\int (\nabla y(\mathbf{x})) (\nabla y(\mathbf{x}))^{\mathsf{T}}\, p(\mathbf{x}) d\mathbf{x} \right) \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi}.$$

Since

$$\mathrm{E}\, \boldsymbol{\xi} = \mathbf{0}, \tag{5.169}$$

the first term can be written as $\mathbf{0}$. Since

$$\mathrm{E}\, \boldsymbol{\xi} \boldsymbol{\xi}^{\mathsf{T}} = \lambda \mathbf{I}, \tag{5.170}$$

The second term can be written as

$$\lambda \operatorname{tr} \left(\iint (y(\mathbf{x}) - t)\, \nabla\nabla y(\mathbf{x}) p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \right), \tag{5.171}$$

where the integral can be written as

$$\int \left(y(\mathbf{x}) \int p(t|\mathbf{x}) dt - \int t p(t|\mathbf{x}) dt \right) \nabla\nabla y(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$
$$= \int (y(\mathbf{x}) - \mathrm{E}(t|\mathbf{x}))\, \nabla\nabla y(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5.172}$$

Similarly, the third term can be written as $2\lambda\Omega$, where

$$\Omega = \frac{1}{2} \operatorname{tr} \left(\int \nabla y(\mathbf{x}) (\nabla y(\mathbf{x}))^{\mathsf{T}}\, p(\mathbf{x}) d\mathbf{x}. \right). \tag{5.173}$$

177

Therefore, if
$$y(\mathbf{x}) = \mathrm{E}(t|\mathbf{x}), \tag{5.174}$$
then
$$\tilde{E} \simeq E + \lambda \Omega. \tag{5.175}$$

5.28 (Incomplete)

5.29

Let $\mathbf{w}$ be a variable in M dimensions such that
$$p(w_m) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(w_m|\mu_k, \sigma_k^2\right). \tag{5.176}$$

Let
$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \tag{5.177}$$
where
$$\Omega(\mathbf{w}) = -\sum_{m=1}^{M} \ln p(w_m). \tag{5.178}$$

We have
$$\frac{\partial \Omega}{\partial w_m} = \frac{\partial \Omega}{\partial p_m} \frac{\partial p_m}{\partial w_m}, \tag{5.179}$$
where
$$p_m = p(w_m). \tag{5.180}$$

We have
$$\frac{\partial \Omega}{\partial p_m} = -\frac{1}{p_m},$$
$$\frac{\partial p_m}{\partial w_m} = -\sum_{k=1}^{K} \frac{w_m - \mu_k}{\sigma_k^2} p_{mk}, \tag{5.181}$$
where
$$p_{mk} = \pi_k \mathcal{N}\left(w_m|\mu_k, \sigma_k^2\right). \tag{5.182}$$

Therefore,
$$\frac{\partial \tilde{E}}{\partial w_m} = \frac{\partial E}{\partial w_m} + \lambda \sum_{k=1}^{K} \gamma_{mk} \frac{w_m - \mu_k}{\sigma_k^2}, \tag{5.183}$$
where
$$\gamma_{mk} = \frac{p_{mk}}{p_m}. \tag{5.184}$$

5.30

Let $\mathbf{w}$ be a variable in M dimensions such that

$$p(w_m) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right). \tag{5.185}$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \tag{5.186}$$

where

$$\Omega(\mathbf{w}) = -\sum_{m=1}^{M} \ln p(w_m). \tag{5.187}$$

We have

$$\frac{\partial \Omega}{\partial \mu_k} = \sum_{m=1}^{M} \frac{\partial \Omega}{\partial p_m} \frac{\partial p_m}{\partial \mu_k}, \tag{5.188}$$

where

$$p_m = p(w_m). \tag{5.189}$$

We have

$$\begin{aligned} \frac{\partial \Omega}{\partial p_m} &= -\frac{1}{p_m}, \\ \frac{\partial p_m}{\partial \mu_k} &= -\frac{\mu_k - w_m}{\sigma_k^2} p_{mk}, \end{aligned} \tag{5.190}$$

where

$$p_{mk} = \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right). \tag{5.191}$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial \mu_k} = \lambda \sum_{m=1}^{M} \gamma_{mk} \frac{\mu_k - w_m}{\sigma_k^2}, \tag{5.192}$$

where

$$\gamma_{mk} = \frac{p_{mk}}{p_m}. \tag{5.193}$$

5.31

Let $\mathbf{w}$ be a variable in M dimensions such that

$$p(w_m) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right). \tag{5.194}$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \tag{5.195}$$

where

$$\Omega(\mathbf{w}) = -\sum_{m=1}^{M} \ln p(w_m). \tag{5.196}$$

We have

$$\frac{\partial \Omega}{\partial \sigma_k} = \sum_{m=1}^{M} \frac{\partial \Omega}{\partial p_m} \frac{\partial p_m}{\partial \sigma_k}, \tag{5.197}$$

where

$$p_m = p(w_m). \tag{5.198}$$

We have

$$\begin{aligned}
\frac{\partial \Omega}{\partial p_m} &= -\frac{1}{p_m}, \\
\frac{\partial p_m}{\partial \sigma_k} &= \left(-\frac{1}{\sigma_k} + \frac{(w_m - \mu_k)^2}{\sigma_k^3}\right) p_{mk},
\end{aligned} \tag{5.199}$$

where

$$p_{mk} = \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right). \tag{5.200}$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial \sigma_k} = \lambda \sum_{m=1}^{M} \gamma_{mk} \left(\frac{1}{\sigma_k} - \frac{(w_m - \mu_k)^2}{\sigma_k^3}\right), \tag{5.201}$$

where

$$\gamma_{mk} = \frac{p_{mk}}{p_m}. \tag{5.202}$$

5.32

Let $\mathbf{w}$ be variable in M dimensions such that

$$p(w_m) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right), \qquad (5.203)$$

where

$$\pi_k = \frac{\exp(\eta_k)}{\sum_{k'=1}^{K} \exp(\eta_{k'})}. \qquad (5.204)$$

Let

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \lambda \Omega(\mathbf{w}), \qquad (5.205)$$

where

$$\Omega(\mathbf{w}) = -\sum_{m=1}^{M} \ln p(w_m). \qquad (5.206)$$

(a)

If $k \neq k'$, then

$$\frac{\partial \pi_k}{\partial \eta_{k'}} = -\pi_k \pi_{k'}. \qquad (5.207)$$

We have

$$\frac{\partial \pi_k}{\partial \eta_k} = \pi_k - \pi_k^2. \qquad (5.208)$$

Therefore,

$$\frac{\partial \pi_k}{\partial \eta_{k'}} = I_{kk'} \pi_k - \pi_k \pi_{k'}. \qquad (5.209)$$

(b)

We have

$$\frac{\partial \Omega}{\partial \eta_k} = \sum_{m=1}^{M} \sum_{k'=1}^{K} \frac{\partial \Omega}{\partial p_m} \frac{\partial p_m}{\partial \pi_{k'}} \frac{\partial \pi_{k'}}{\partial \eta_k}, \qquad (5.210)$$

where

$$p_m = p(w_m). \qquad (5.211)$$

We have

$$\frac{\partial \Omega}{\partial p_m} = -\frac{1}{p_m},$$

$$\frac{\partial p_m}{\partial \pi_{k'}} = \frac{p_{mk'}}{\pi_{k'}}, \tag{5.212}$$

where

$$p_{mk} = \pi_k \mathcal{N}\left(w_m | \mu_k, \sigma_k^2\right). \tag{5.213}$$

By (a),

$$\frac{\partial \pi_{k'}}{\partial \eta_k} = I_{k'k}\pi_{k'} - \pi_{k'}\pi_{k'}. \tag{5.214}$$

Then,

$$\frac{\partial \Omega}{\partial \eta_k} = -\sum_{m=1}^{M}\sum_{k'=1}^{K} \frac{\gamma_{mk'}}{\pi_{k'}}\left(I_{k'k}\pi_{k'} - \pi_k\pi_{k'}\right), \tag{5.215}$$

where

$$\gamma_{mk} = \frac{p_{mk}}{p_m}. \tag{5.216}$$

The right hand side can be written as

$$-\sum_{m=1}^{M}\sum_{k'=1}^{K} \gamma_{mk'}\left(I_{k'k} - \pi_k\right) = \sum_{m=1}^{M}\left(\pi_k - \gamma_{mk}\right). \tag{5.217}$$

Therefore,

$$\frac{\partial \tilde{E}}{\partial \eta_k} = \lambda \sum_{m=1}^{M}\left(\pi_k - \gamma_{mk}\right). \tag{5.218}$$

5.33

Let (x_1, x_2) be the Cartesian coordinates of the end-effector of a two-link robot arm whose joint angles are θ_1 and θ_2 and whose arm lengths are L_1 and L_2. Let us asuume that the origin of the coordinate system is given by the attachment point of the lower arm. We have

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} L_1 \cos\theta_1 \\ L_1 \sin\theta_1 \end{bmatrix} + \begin{bmatrix} L_2 \cos\left(\theta_2 - (\pi - \theta_1)\right) \\ L_2 \sin\left(\theta_2 - (\pi - \theta_1)\right) \end{bmatrix}. \tag{5.219}$$

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

$$\begin{bmatrix} L_2 \cos\left(\theta_1 + \theta_2 - \pi\right) \\ L_2 \sin\left(\theta_1 + \theta_2 - \pi\right) \end{bmatrix} = \begin{bmatrix} -L_2 \cos(\theta_1 + \theta_2) \\ -L_2 \sin(\theta_1 + \theta_2) \end{bmatrix}. \tag{5.220}$$

Therefore,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} L_1 \cos \theta_1 - L_2 \cos(\theta_1 + \theta_2) \\ L_1 \sin \theta_1 - L_2 \sin(\theta_1 + \theta_2) \end{bmatrix}. \tag{5.221}$$

5.34

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

$$p(\mathbf{t}_n | \mathbf{w}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{nk}, \tag{5.222}$$

where

$$\pi_k = \frac{\exp(a_k)}{\sum_{k'=1}^{K} \exp(a_{k'})}, \tag{5.223}$$
$$\mathcal{N}_{nk} = \mathcal{N}\left(\mathbf{t}_n | \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w})\mathbf{I}\right).$$

Let

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}), \tag{5.224}$$

where

$$E_n(\mathbf{w}) = -\ln p(\mathbf{t}_n | \mathbf{w}). \tag{5.225}$$

We have

$$\frac{\partial E_n}{\partial a_k} = \sum_{k'=1}^{K} \frac{\partial E_n}{\partial p_n} \frac{\partial p_n}{\partial \pi_{k'}} \frac{\partial \pi_{k'}}{\partial a_k}, \tag{5.226}$$

where

$$p_n = p(\mathbf{t}_n | \mathbf{w}). \tag{5.227}$$

We have

$$\frac{\partial E_n}{\partial p_n} = -\frac{1}{p_n}, \\ \frac{\partial p_n}{\partial \pi_{k'}} = \frac{p_{nk'}}{\pi_{k'}}, \tag{5.228}$$

where

$$p_{nk} = \pi_k \mathcal{N}_{nk}. \tag{5.229}$$

By 5.32(a),

$$\frac{\partial \pi_{k'}}{\partial a_k} = I_{k'k} \pi_{k'} - \pi_k \pi_{k'}. \tag{5.230}$$

Then,

$$\frac{\partial E_n}{\partial a_k} = -\frac{1}{p_n} \sum_{k'=1}^{K} \frac{p_{nk'}}{\pi_{k'}} (I_{k'k}\pi_{k'} - \pi_k\pi_{k'}). \tag{5.231}$$

The right hand side can be written as

$$-\sum_{k'=1}^{K} \gamma_{nk'} (I_{kk'} - \pi_k) = \pi_k - \gamma_{nk}, \tag{5.232}$$

where

$$\gamma_{nk} = \frac{p_{nk}}{p_n}, \tag{5.233}$$

Therefore,

$$\frac{\partial E_n}{\partial a_k} = \pi_k - \gamma_{nk}. \tag{5.234}$$

5.35

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

$$p(\mathbf{t}_n|\mathbf{w}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{nk}, \tag{5.235}$$

where

$$\mathcal{N}_{nk} = \mathcal{N}\left(\mathbf{t}_n | \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w})\mathbf{I}\right). \tag{5.236}$$

Let

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}), \tag{5.237}$$

where

$$E_n(\mathbf{w}) = -\ln p(\mathbf{t}_n|\mathbf{w}). \tag{5.238}$$

Then,

$$\frac{\partial E_n}{\partial \boldsymbol{\mu}_k} = \frac{\partial E_n}{\partial p_n} \frac{\partial p_n}{\partial \boldsymbol{\mu}_k}, \tag{5.239}$$

where

$$p_n = p(t_n|\mathbf{w}). \tag{5.240}$$

184

We have

$$\frac{\partial E_n}{\partial p_n} = -\frac{1}{p_n},$$
$$\frac{\partial p_n}{\partial \boldsymbol{\mu}_k} = -\sigma_k^{-2}(\boldsymbol{\mu}_k - \mathbf{t}_n)p_{nk}, \tag{5.241}$$

where

$$p_{nk} = \pi_k \mathcal{N}_{nk}. \tag{5.242}$$

Therefore,

$$\frac{\partial E_n}{\partial \boldsymbol{\mu}_k} = \gamma_{nk}\sigma_k^{-2}(\boldsymbol{\mu}_k - \mathbf{t}_n), \tag{5.243}$$

where

$$\gamma_{nk} = \frac{p_{nk}}{p_n}. \tag{5.244}$$

5.36

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in M dimensions such that

$$p(\mathbf{t}_n|\mathbf{w}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{nk}, \tag{5.245}$$

where

$$\mathcal{N}_{nk} = \mathcal{N}\left(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w})\mathbf{I}\right),$$
$$\sigma_k = \exp\left(a_k^\sigma\right). \tag{5.246}$$

Let

$$E(\mathbf{w}) = \sum_{n=1}^{N} E_n(\mathbf{w}), \tag{5.247}$$

where

$$E_n(\mathbf{w}) = -\ln p(\mathbf{t}_n|\mathbf{w}). \tag{5.248}$$

Then,

$$\frac{\partial E_n}{\partial a_k^\sigma} = \frac{\partial E_n}{\partial p_n}\frac{\partial p_n}{\partial \sigma_k}\frac{\partial \sigma_k}{\partial a_k^\sigma}, \tag{5.249}$$

where

$$p_n = p(\mathbf{t}_n|\mathbf{w}). \tag{5.250}$$

We have

$$\frac{\partial E_n}{\partial p_n} = -\frac{1}{p_n},$$

$$\frac{\partial p_n}{\partial \sigma_k} = \left(-\frac{M}{\sigma_k} + \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3}\right) p_{nk}, \tag{5.251}$$

$$\frac{\partial \sigma_k}{\partial a_k^\sigma} = \sigma_k,$$

where

$$p_{nk} = \pi_k \mathcal{N}_{nk}. \tag{5.252}$$

Therefore,

$$\frac{\partial E_n}{\partial a_k^\sigma} = \gamma_{nk} \left(M - \frac{\|\mathbf{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2}\right), \tag{5.253}$$

where

$$\gamma_{nk} = \frac{p_{nk}}{p_n}. \tag{5.254}$$

5.37

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

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right). \tag{5.255}$$

(a)

We have

$$\mathrm{E}(\mathbf{t}|\mathbf{x}) = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \tag{5.256}$$

The right hand side can be written as

$$\sum_{k=1}^{K} \pi_k(\mathbf{x}) \int \mathbf{t} \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t} = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}). \tag{5.257}$$

Therefore,

$$\mathrm{E}(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \boldsymbol{\mu}_k(\mathbf{x}). \tag{5.258}$$

186

(b)

We have

$$\text{var}(\mathbf{t}|\mathbf{x}) = \int \left(\mathbf{t} - \text{E}(\mathbf{t}|\mathbf{x})\right) \left(\mathbf{t} - \text{E}(\mathbf{t}|\mathbf{x})\right)^\mathsf{T} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \tag{5.259}$$

The right hand side can be written as

$$\begin{aligned}
&\int \mathbf{t}\mathbf{t}^\mathsf{T} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \text{E}(\mathbf{t}|\mathbf{x}) \left(\int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}\right)^\mathsf{T} \\
&- \left(\int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}\right) \text{E}(\mathbf{t}|\mathbf{x})^\mathsf{T} + \text{E}(\mathbf{t}|\mathbf{x}) \text{E}(\mathbf{t}|\mathbf{x})^\mathsf{T} \int p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \\
&= \int \mathbf{t}\mathbf{t}^\mathsf{T} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \text{E}(\mathbf{t}|\mathbf{x}) \text{E}(\mathbf{t}|\mathbf{x})^\mathsf{T}.
\end{aligned} \tag{5.260}$$

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

$$\sum_{k=1}^{K} \pi_k(\mathbf{x}) \int \mathbf{t}\mathbf{t}^\mathsf{T} \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}. \tag{5.261}$$

The integral of the right hand side can be written as

$$\begin{aligned}
&\int \left(\mathbf{t}' + \boldsymbol{\mu}_k(\mathbf{x})\right) \left(\mathbf{t}' + \boldsymbol{\mu}_k(\mathbf{x})\right)^\mathsf{T} \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t} \\
&= \int \mathbf{t}'\mathbf{t}'^\mathsf{T} \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t} \\
&\quad + \boldsymbol{\mu}_k(\mathbf{x}) \left(\int \mathbf{t}' \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}\right)^\mathsf{T} \\
&\quad + \left(\int \mathbf{t}' \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}\right) \boldsymbol{\mu}_k(\mathbf{x})^\mathsf{T} \\
&\quad + \boldsymbol{\mu}_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})^\mathsf{T} \int \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t},
\end{aligned} \tag{5.262}$$

where

$$\mathbf{t}' = \mathbf{t} - \boldsymbol{\mu}_k(\mathbf{x}). \tag{5.263}$$

The right hand side can be written as

$$
\int \mathbf{t}'\mathbf{t}'^{\mathsf{T}}\mathcal{N}\left(\mathbf{t}'|\mathbf{0}, \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}'
$$
$$
+ \boldsymbol{\mu}_k(\mathbf{x})\left(\int \mathbf{t}'\mathcal{N}\left(\mathbf{t}'|\mathbf{0}, \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}'\right)^{\mathsf{T}}
$$
$$
+ \left(\int \mathbf{t}'\mathcal{N}\left(\mathbf{t}'|\mathbf{0}, \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}'\right)\boldsymbol{\mu}_k(\mathbf{x})^{\mathsf{T}} \tag{5.264}
$$
$$
+ \boldsymbol{\mu}_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})^{\mathsf{T}}\int \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I}\right) d\mathbf{t}
$$
$$
=\sigma_k^2(\mathbf{x})\mathbf{I} + \boldsymbol{\mu}_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})^{\mathsf{T}}.
$$

Therefore, by (a),

$$
\text{cov}(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K}\pi_k(\mathbf{x})\left(\sigma_k^2(\mathbf{x})\mathbf{I} + \boldsymbol{\mu}_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})^{\mathsf{T}}\right)
$$
$$
- \left(\sum_{k=1}^{K}\pi_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})\right)\left(\sum_{k=1}^{K}\pi_k(\mathbf{x})\boldsymbol{\mu}_k(\mathbf{x})\right)^{\mathsf{T}}. \tag{5.265}
$$

5.38

Let $t_1, \cdots, t_N$ be variables such that

$$
p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\right),
$$
$$
p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right). \tag{5.266}
$$

By marginalisation,

$$
p(t_{N+1}|\mathbf{t}) = \int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{5.267}
$$

By the Bayes' theorem,

$$
p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{5.268}
$$

The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as $-E(\mathbf{w})$ where

$$
E(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(\mathbf{x}_n, \mathbf{w}))^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2. \tag{5.269}
$$

188

Let $\mathbf{w}_{\mathrm{MAP}}$ be a stationary point of $p(\mathbf{w}|\mathbf{t})$. Then, we have the Taylor series

$$
\begin{aligned}
E(\mathbf{w}) =& E(\mathbf{w}_{\mathrm{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}) \\
& + O\left(\|\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}\|^3\right),
\end{aligned}
\tag{5.270}
$$

where

$$
\mathbf{A} = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\mathrm{MAP}}}.
\tag{5.271}
$$

Then,

$$
p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{A}^{-1}\right).
\tag{5.272}
$$

By the Taylor series

$$
y(\mathbf{x}_{N+1}, \mathbf{w}) = y(\mathbf{x}_{N+1}, \mathbf{w}_{\mathrm{MAP}}) + \mathbf{g}^{\mathsf{T}}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}) + O\left(\|\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}\|^2\right),
\tag{5.273}
$$

where

$$
\mathbf{g} = \nabla_{\mathbf{w}}y(\mathbf{x}_{N+1}, \mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\mathrm{MAP}}},
\tag{5.274}
$$

we have

$$
p(t_{N+1}|\mathbf{w}) \simeq \mathcal{N}\left(t_{N+1}|y(\mathbf{x}_{N+1}, \mathbf{w}_{\mathrm{MAP}}) + \mathbf{g}^{\mathsf{T}}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}), \beta^{-1}\right).
\tag{5.275}
$$

Then, the logarithm of the integrand except the terms independent of $\mathbf{w}$ can be approximated as

$$
\begin{aligned}
& -\frac{\beta}{2}\left(t_{N+1} - y(\mathbf{x}_{N+1}, \mathbf{w}_{\mathrm{MAP}}) - \mathbf{g}^{\mathsf{T}}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})\right)^2 \\
& -\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}}) \\
& = -\frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v},
\end{aligned}
\tag{5.276}
$$

where

$$
\mathbf{v} = \begin{bmatrix} \mathbf{w} - \mathbf{w}_{\mathrm{MAP}} \\ t_{N+1} - y(\mathbf{x}_{N+1}, \mathbf{w}_{\mathrm{MAP}}) \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{A} + \beta\mathbf{g}\mathbf{g}^{\mathsf{T}} & -\beta\mathbf{g} \\ -\beta\mathbf{g}^{\mathsf{T}} & \beta \end{bmatrix}.
\tag{5.277}
$$

By 2.24,

$$
\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{g} \\ \mathbf{g}^{\mathsf{T}}\mathbf{A}^{-1} & \beta^{-1} + \mathbf{g}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{g} \end{bmatrix}.
\tag{5.278}
$$

Therefore,

$$
p(t_{N+1}|\mathbf{t}) \simeq \mathcal{N}\left(t_{N+1}|y(\mathbf{x}_{N+1}, \mathbf{w}_{\mathrm{MAP}}), \beta^{-1} + \mathbf{g}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{g}\right).
\tag{5.279}
$$

5.39

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right), \tag{5.280}$$

where $\mathbf{w}$ is in M dimensions. By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{5.281}$$

The logarithm of the integrand of the right hand side can be written as

$$
\begin{aligned}
& -\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(\mathbf{x}_n, \mathbf{w}))^2 \\
& -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left(\alpha^{-1}\mathbf{I}\right)\right) - \frac{\alpha}{2}\|\mathbf{w}\|^2 \\
& = -E(\mathbf{w}) - \frac{N+M}{2}\ln(2\pi) + \frac{N}{2}\ln\beta + \frac{M}{2}\ln\alpha,
\end{aligned}
\tag{5.282}
$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(\mathbf{x}_n, \mathbf{w}))^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2. \tag{5.283}$$

Let $\mathbf{w}_{\text{MAP}}$ be a stationary point of E. Then, we have the Taylor series

$$
\begin{aligned}
E(\mathbf{w}) = & E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\
& + O\left(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3\right),
\end{aligned}
\tag{5.284}
$$

where

$$\mathbf{A} = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}. \tag{5.285}$$

Then, the logarithm of the integrand can be approximated as

$$
\begin{aligned}
& -E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\
& -\frac{N+M}{2}\ln(2\pi) + \frac{N}{2}\ln\beta + \frac{M}{2}\ln\alpha.
\end{aligned}
\tag{5.286}
$$

We have

$$\int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})\right) d\mathbf{w}$$
$$= (2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}^{-1}\right)^{\frac{1}{2}}. \tag{5.287}$$

Therefore,

$$\ln p(\mathbf{t}) \simeq -E(\mathbf{w}_{\mathrm{MAP}}) - \frac{1}{2}\ln|\det \mathbf{A}| - \frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln \beta + \frac{M}{2}\ln \alpha. \tag{5.288}$$

5.40

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n|\mathbf{w}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \tag{5.289}$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right),$$

where

$$y_{nk} = y_k(\mathbf{a}_n),$$
$$y_k(\mathbf{a}) = \frac{\exp(a_k)}{\sum_{k'=1}^{K}\exp(a_{k'})}, \tag{5.290}$$
$$\mathbf{a}_n = \mathbf{a}(\mathbf{x}_n, \mathbf{w}).$$

By marginalisation,

$$p(\mathbf{t}_{N+1}|\mathbf{T}) = \int p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{T})d\mathbf{w}. \tag{5.291}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{T})p(\mathbf{T}) = p(\mathbf{T}|\mathbf{w})p(\mathbf{w}). \tag{5.292}$$

The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as $-E(\mathbf{w})$ where

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk} + \frac{\alpha}{2}\|\mathbf{w}\|^2. \tag{5.293}$$

Let $\mathbf{w}_{\text{MAP}}$ be a stationary point of E. Then, we have the Taylor series

$$
\begin{aligned}
E(\mathbf{w}) =& E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\
& + O\left(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3\right),
\end{aligned}
\tag{5.294}
$$

where

$$
\mathbf{A} = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}.
\tag{5.295}
$$

Then,

$$
p(\mathbf{w}|\mathbf{T}) \simeq \mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}\right).
\tag{5.296}
$$

We have

$$
p(\mathbf{t}_{N+1}|\mathbf{w}) = \prod_{k=1}^{K} y_{N+1,k}^{t_{N+1,k}},
\tag{5.297}
$$

where

$$
y_{N+1,k} = y_k\left(\mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w})\right).
\tag{5.298}
$$

We have the Taylor series

$$
\begin{aligned}
\mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}) =& \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) + \mathbf{B}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) \\
& + O\left(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^2\right),
\end{aligned}
\tag{5.299}
$$

where

$$
\mathbf{B} = \nabla_{\mathbf{w}}\mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}.
\tag{5.300}
$$

Then,

$$
p(\mathbf{t}_{N+1}|\mathbf{w}) \simeq \int \delta\left(\mathbf{a}_{N+1} - \mathbf{c}_{N+1}\right)\left(\prod_{k=1}^{K} y_k(\mathbf{a}_{N+1})^{t_{N+1,k}}\right) d\mathbf{a}_{N+1},
\tag{5.301}
$$

where δ is the Dirac delta function and

$$
\mathbf{c}_{N+1} = \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}) + \mathbf{B}(\mathbf{w} - \mathbf{w}_{\text{MAP}}).
\tag{5.302}
$$

Then,

$$
\int p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{T})d\mathbf{w} \simeq \int \left(\prod_{k=1}^{K} y_k(\mathbf{a}_{N+1})^{t_{N+1,k}}\right) p(\mathbf{a}_{N+1})d\mathbf{a}_{N+1},
\tag{5.303}
$$

where

$$p(\mathbf{a}_{N+1}) = \int \delta\left(\mathbf{a}_{N+1} - \mathbf{c}_{N+1}\right) \mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}\right) d\mathbf{w}. \tag{5.304}$$

Since $\mathbf{t}_{N+1}$ is a variable from the standard basis in K dimensions, the right hand side of the approximation can be written as

$$\prod_{k=1}^{K} \left(\int y_k(\mathbf{a}_{N+1}) p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1}\right)^{t_{N+1,k}}. \tag{5.305}$$

We have

$$\mathrm{E}\,\mathbf{a}_{N+1} = \int \mathbf{a}_{N+1} p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1}. \tag{5.306}$$

The right hand side can be written as

$$\int \mathbf{c}_{N+1} \mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}\right) d\mathbf{w} = \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}). \tag{5.307}$$

Then,

$$\mathrm{var}\,\mathbf{a}_{N+1} = \int \left(\mathbf{a}_{N+1} - \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}})\right)^2 p(\mathbf{a}_{N+1}) d\mathbf{a}_{N+1}. \tag{5.308}$$

The right hand side can be written as

$$\int \left(\mathbf{c}_{N+1} - \mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}})\right)^2 \mathcal{N}\left(\mathbf{w}|\mathbf{w}_{\text{MAP}}, \mathbf{A}^{-1}\right) d\mathbf{w} = \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{\mathsf{T}}. \tag{5.309}$$

Then,

$$p(\mathbf{a}_{N+1}) = \mathcal{N}\left(\mathbf{a}_{N+1}|\mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}), \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{\mathsf{T}}\right). \tag{5.310}$$

Therefore,

$$p(\mathbf{t}_{N+1}|\mathbf{T}) \simeq \prod_{k=1}^{K} v_{N+1,k}^{t_{N+1,k}}, \tag{5.311}$$

where

$$v_{N+1,k} = \int y_k(\mathbf{a}) \mathcal{N}\left(\mathbf{a}|\mathbf{a}(\mathbf{x}_{N+1}, \mathbf{w}_{\text{MAP}}), \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^{\mathsf{T}}\right) d\mathbf{a}. \tag{5.312}$$

5.41

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables from the standard basis in K dimensions such that

$$p(\mathbf{t}_n|\mathbf{w}) = \prod_{k=1}^{K} y_{nk}^{t_{nk}},$$

$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right),$$

(5.313)

where $\mathbf{w}$ is in M dimensions and

$$y_{nk} = y_k(\mathbf{x}_n, \mathbf{w}).$$

(5.314)

By marginalisation,

$$p(\mathbf{T}) = \int p(\mathbf{T}|\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

(5.315)

The logarithm of the integrand of the right hand side can be written as

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk} - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left(\alpha^{-1}\mathbf{I}\right)\right) - \frac{\alpha}{2}\|\mathbf{w}\|^2$$

$$= -E(\mathbf{w}) - \frac{M}{2}\ln(2\pi) + \frac{M}{2}\ln\alpha,$$

(5.316)

where

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk} + \frac{\alpha}{2}\|\mathbf{w}\|^2.$$

(5.317)

Let $\mathbf{w}_{\text{MAP}}$ be a stationary point of E. Then, we have the Taylor series

$$E(\mathbf{w}) = E(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}})$$

$$+ O\left(\|(\mathbf{w} - \mathbf{w}_{\text{MAP}})\|^3\right),$$

(5.318)

where

$$\mathbf{A} = \nabla\nabla E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}}.$$

(5.319)

Then, the logarithm of the integrand can be approximated as

$$-E(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) - \frac{M}{2}\ln(2\pi) + \frac{M}{2}\ln\alpha. \quad (5.320)$$

We have

$$\int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})^{\mathsf{T}} \mathbf{A}(\mathbf{w} - \mathbf{w}_{\mathrm{MAP}})\right) d\mathbf{w}$$
$$=(2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}^{-1}\right)^{\frac{1}{2}}. \tag{5.321}$$

Therefore,

$$\ln p(\mathbf{T}) \simeq -E(\mathbf{w}_{\mathrm{MAP}}) - \frac{1}{2}\ln|\det \mathbf{A}| + \frac{M}{2}\ln \alpha. \tag{5.322}$$

6 Kernel Methods

6.1

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, 1\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}\right). \tag{6.1}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{6.2}$$

The logarithm of the right hand side except the terms independent of $\mathbf{w}$ can be written as $-E_{\mathbf{w}}(\mathbf{w})$ where

$$E_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{6.3}$$

so that

$$E_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{6.4}$$

(a)

Let $\mathbf{w}_{\text{MAP}}$ be a stationary point of $E_{\mathbf{w}}$. Setting the derivative of $E_{\mathbf{w}}$ with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = -\boldsymbol{\Phi}^\mathsf{T}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) + \lambda\mathbf{w}. \tag{6.5}$$

Therefore,

$$\mathbf{w}_{\text{MAP}} = \boldsymbol{\Phi}^\mathsf{T}\mathbf{a}_{\text{MAP}}, \tag{6.6}$$

where

$$\mathbf{a}_{\text{MAP}} = \frac{1}{\lambda}\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{\text{MAP}}\right). \tag{6.7}$$

(b)

Let

$$\mathbf{w} = \boldsymbol{\Phi}^\mathsf{T}\mathbf{a}. \tag{6.8}$$

Then,
$$E_{\mathbf{w}}(\mathbf{w}) = E_{\mathbf{a}}(\mathbf{a}), \tag{6.9}$$
where
$$E_{\mathbf{a}}(\mathbf{a}) = \frac{1}{2}\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{a}\|^2 + \frac{\lambda}{2}\mathbf{a}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{a}. \tag{6.10}$$
The right hand side can be written as
$$\frac{1}{2}\|\mathbf{t}\|^2 - \mathbf{t}^{\mathsf{T}}\mathbf{K}\mathbf{a} + \frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{K}\mathbf{K}\mathbf{a} + \frac{\lambda}{2}\mathbf{a}^{\mathsf{T}}\mathbf{K}\mathbf{a}, \tag{6.11}$$
where
$$\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}}. \tag{6.12}$$
Let $\mathbf{a}_{\mathrm{MAP}}$ be a stationary point of $E_{\mathbf{a}}$. Setting the derivative of $E_{\mathbf{a}}$ with respect to $\mathbf{a}$ to zero gives
$$\mathbf{0} = -\mathbf{K}\mathbf{t} + \mathbf{K}\mathbf{K}\mathbf{a} + \lambda\mathbf{K}\mathbf{a}. \tag{6.13}$$
Therefore,
$$\mathbf{a}_{\mathrm{MAP}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{t}. \tag{6.14}$$

(c)

By (a) and (b),
$$\frac{1}{\lambda}\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{\mathrm{MAP}}\right) = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{t}, \tag{6.15}$$
so that
$$\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{\mathrm{MAP}} = \mathbf{N}\mathbf{t}, \tag{6.16}$$
where
$$\mathbf{N} = \left(\mathbf{I} + \lambda^{-1}\mathbf{K}\right)^{-1}. \tag{6.17}$$
By 2.26, the right hand side can be written as
$$\mathbf{I} - \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^{\mathsf{T}}, \tag{6.18}$$
where
$$\mathbf{M} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1}. \tag{6.19}$$
Then,
$$\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{\mathrm{MAP}} = \left(\mathbf{I} - \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^{\mathsf{T}}\right)\mathbf{t}, \tag{6.20}$$
so that
$$\boldsymbol{\Phi}\left(\mathbf{w}_{\mathrm{MAP}} - \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t}\right) = \mathbf{0}. \tag{6.21}$$
Therefore,
$$\mathbf{w}_{\mathrm{MAP}} = \mathbf{M}\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{t}. \tag{6.22}$$

6.2   (Incomplete)

Let $t_1, \cdots, t_N$ be vaariables such that

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

$$p(t_n|\mathbf{w}) = \left(\frac{1 + y_n}{2}\right)^{\frac{1+t_n}{2}} \left(\frac{1 - y_n}{2}\right)^{\frac{1-t_n}{2}}, \tag{6.23}$$

where

$$y_n = f\left(\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\right),$$

$$f(a) = \begin{cases} 1, & a \geq 0, \\ -1, & \text{otherwise.} \end{cases} \tag{6.24}$$

Let

$$E(\mathbf{w}) = -\sum_{n \in \mathcal{M}} \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n t_n, \tag{6.25}$$

where $\mathcal{M}$ is the set of all misclassification. We have the Taylor series

$$E(\mathbf{w}^{(\tau)}) = E(\mathbf{w}^{(\tau-1)}) + \mathbf{b}^\mathsf{T}\left(\mathbf{w}^{(\tau)} - \mathbf{w}^{(\tau-1)}\right) + O\left(\left\|\mathbf{w}^{(\tau)} - \mathbf{w}^{(\tau-1)}\right\|^2\right), \tag{6.26}$$

where

$$\mathbf{b} = -\sum_{n \in \mathcal{M}} t_n \boldsymbol{\phi}_n. \tag{6.27}$$

6.3

We have

$$\|\mathbf{x} - \mathbf{x}_n\|^2 = \|\mathbf{x}\|^2 - 2\mathbf{x}^\mathsf{T}\mathbf{x}_n + \|\mathbf{x}_n\|^2. \tag{6.28}$$

The right hand side can be written as

$$k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{x}_n) + k(\mathbf{x}_n, \mathbf{x}_n), \tag{6.29}$$

where

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T}\mathbf{x}'. \tag{6.30}$$

6.4

Let $\mathbf{A}$ be a 2×2 matrix with positive eigenvalues and with at least one negative element. We have

$$\det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - A_{11})(\lambda - A_{22}) - A_{12}A_{21}. \qquad (6.31)$$

Seeing the left hand side to zero gives

$$0 = \lambda^2 - (A_{11} + A_{22})\lambda + A_{11}A_{22} - A_{12}A_{21}. \qquad (6.32)$$

For the equation to have two positive solutions, we have the conditions

$$\begin{aligned}
(A_{11} + A_{22})^2 - 4(A_{11}A_{22} - A_{12}A_{21}) &> 0, \\
A_{11} + A_{22} &> 0, \qquad (6.33) \\
A_{11}A_{22} - A_{12}A_{21} &> 0.
\end{aligned}$$

Therefore, an example of A is given by

$$\begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix}. \qquad (6.34)$$

6.5

Let

$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}_1(\mathbf{x}'). \qquad (6.35)$$

(a)

Let

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}'), \qquad (6.36)$$

where c is a positive constant. The right hand side can be written as

$$c\boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}_1(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}'), \qquad (6.37)$$

where

$$\boldsymbol{\phi} = \sqrt{c}\boldsymbol{\phi}_1. \qquad (6.38)$$

Therefore, k is a valid kernel.

(b)

Let
$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}'). \tag{6.39}$$
The right hand side can be written as
$$f(\mathbf{x})\boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}_1(\mathbf{x}')f(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}'), \tag{6.40}$$
where
$$\boldsymbol{\phi} = f\boldsymbol{\phi}_1. \tag{6.41}$$
Therefore, k is a valid kernel.

6.6

Let
$$k_1(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}_1(\mathbf{x}'), \tag{6.42}$$
where $\boldsymbol{\phi}_1$ is a vector in M dimensions.

(a)

Let
$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=0}^{J} c_j q_j(\mathbf{x}, \mathbf{x}'), \tag{6.43}$$
where $c_0, \cdots, c_J$ are nonnegative constants and
$$q_j(\mathbf{x}, \mathbf{x}') = (k_1(\mathbf{x}, \mathbf{x}'))^j. \tag{6.44}$$
We have
$$q_0(\mathbf{x}, \mathbf{x}') = 1. \tag{6.45}$$
Let us asuume that q_j is a valid kernel. Then,
$$q_{j+1}(\mathbf{x}, \mathbf{x}') = q_j(\mathbf{x}, \mathbf{x}')\boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}_1(\mathbf{x}'). \tag{6.46}$$
The right hand side can be written as
$$q_j(\mathbf{x}, \mathbf{x}')\sum_{m=1}^{M} \phi_{1m}(\mathbf{x})\phi_{1m}(\mathbf{x}') = \sum_{m=1}^{M} \phi_{1m}(\mathbf{x})q_j(\mathbf{x}, \mathbf{x}')\phi_{1m}(\mathbf{x}'). \tag{6.47}$$

By 6.5(b) and 6.7(a), the right hand side is a valid kernel. Then, q_{j+1} is a valid kernel. Then, the assumption is proved by induction on j. Therefore, by 6.5(a) and 6.7(a), k is a valid kernel.

(b)

Let
$$k(\mathbf{x}, \mathbf{x}') = \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right). \tag{6.48}$$

The right hand side can be written as

$$\sum_{j=0}^{\infty} \frac{1}{j!} \left(k_1(\mathbf{x}, \mathbf{x}')\right)^j. \tag{6.49}$$

Therefore, by (a), k is a valid kernel.

6.7

Let
$$\begin{aligned} k_1(\mathbf{x}, \mathbf{x}') &= \boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}_1(\mathbf{x}'), \\ k_2(\mathbf{x}, \mathbf{x}') &= \boldsymbol{\phi}_2(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}_2(\mathbf{x}'). \end{aligned} \tag{6.50}$$

(a)

Let
$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}'). \tag{6.51}$$

The right hand side can be written as

$$\boldsymbol{\phi}_1(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}_1(\mathbf{x}') + \boldsymbol{\phi}_2(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}_2(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}'), \tag{6.52}$$

where
$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{bmatrix}. \tag{6.53}$$

Therefore, k is a valid kernel.

(b)

Let
$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}'). \tag{6.54}$$

Then,
$$\ln k(\mathbf{x}, \mathbf{x}') = \ln k_1(\mathbf{x}, \mathbf{x}') + \ln k_2(\mathbf{x}, \mathbf{x}'). \tag{6.55}$$

By (a) and 6.6(b), the right hand side is a valid kernel. Then, $\ln k$ is a valid kernel. Therefore, by 6.6(b), k is a valid kernel.

6.8

(a)

Let
$$k(\mathbf{x}, \mathbf{x}') = k_1 \left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \right), \tag{6.56}$$

where k_1 is a valid kernel. Therefore, k is a valid kernel.

(b)

Let
$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x}', \tag{6.57}$$

where $\mathbf{A}$ is a symmetric positive semidefinite matrix. Then, there exists $\mathbf{M}$ such that
$$\mathbf{A} = \mathbf{M}^\mathsf{T} \mathbf{M}. \tag{6.58}$$

Then,
$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}'), \tag{6.59}$$

where
$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{M} \mathbf{x}. \tag{6.60}$$

Therefore, k is a valid kernel.

6.9

Let
$$\begin{aligned} k_a(\mathbf{x}_a, \mathbf{x}'_a) &= \boldsymbol{\phi}_a(\mathbf{x}_a)^\mathsf{T} \boldsymbol{\phi}_a(\mathbf{x}'_a), \\ k_b(\mathbf{x}_b, \mathbf{x}'_b) &= \boldsymbol{\phi}_b(\mathbf{x}_b)^\mathsf{T} \boldsymbol{\phi}_b(\mathbf{x}'_b). \end{aligned} \tag{6.61}$$

(a)

Let
$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b), \tag{6.62}$$

where
$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}. \tag{6.63}$$

The right hand side can be written as
$$\boldsymbol{\phi}_a(\mathbf{x}_a)^\mathsf{T} \boldsymbol{\phi}_a(\mathbf{x}'_a) + \boldsymbol{\phi}_b(\mathbf{x}_b)^\mathsf{T} \boldsymbol{\phi}_b(\mathbf{x}'_b) = \boldsymbol{\phi}(\mathbf{x})^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}'), \tag{6.64}$$

where

$$\phi(\mathbf{x}) = \begin{bmatrix} \phi_a(\mathbf{x}_a) \\ \phi_b(\mathbf{x}_b) \end{bmatrix}. \tag{6.65}$$

Therefore, k is a valid kernel.

(b)

Let

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) k_b(\mathbf{x}_b, \mathbf{x}'_b), \tag{6.66}$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}. \tag{6.67}$$

Then,

$$\ln k(\mathbf{x}, \mathbf{x}') = \ln k_a(\mathbf{x}_a, \mathbf{x}'_a) + \ln k_b(\mathbf{x}_b, \mathbf{x}'_b). \tag{6.68}$$

By (a) and 6.6(b), the right hand side is a valid kernel. Then, $\ln k$ is a valid kernel. Therefore, by 6.6(b), k is a valid kernel.

6.10 (Incomplete)

Let

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) f(\mathbf{x}'). \tag{6.69}$$

6.11

Let

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right). \tag{6.70}$$

The right hand side can be written as

$$\exp\left(-\frac{\mathbf{x}^\mathsf{T}\mathbf{x}}{2\sigma^2}\right) \exp\left(\frac{\mathbf{x}^\mathsf{T}\mathbf{x}'}{\sigma^2}\right) \exp\left(-\frac{\mathbf{x}'^\mathsf{T}\mathbf{x}'}{2\sigma^2}\right). \tag{6.71}$$

We have

$$\exp\left(\frac{\mathbf{x}^\mathsf{T}\mathbf{x}'}{\sigma^2}\right) = \sum_{j=0}^{\infty} q_j(\mathbf{x}, \mathbf{x}'), \tag{6.72}$$

where

$$q_j(\mathbf{x}, \mathbf{x}') = \frac{1}{j!} \left(\frac{\mathbf{x}^\mathsf{T}\mathbf{x}'}{\sigma^2} \right)^j. \tag{6.73}$$

By 6.6(a), q_j is a valid kernel. Then, there exists $\boldsymbol{\psi}_j$ such that

$$q_j(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}_j(\mathbf{x})^\mathsf{T}\boldsymbol{\psi}_j(\mathbf{x}'). \tag{6.74}$$

Then,

$$\exp\left(\frac{\mathbf{x}^\mathsf{T}\mathbf{x}'}{\sigma^2} \right) = \sum_{j=0}^{\infty} \boldsymbol{\psi}_j(\mathbf{x})^\mathsf{T}\boldsymbol{\psi}_j(\mathbf{x}'). \tag{6.75}$$

The right hand side can be written as $\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}')$, where

$$\boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\psi}_0 \\ \vdots \\ \boldsymbol{\psi}_j \\ \vdots \end{bmatrix}. \tag{6.76}$$

Therefore, k can be written as the inner product of an infinite-dimensional feature vector.

6.12

Let D be a set. Let $A_1, \cdots, A_{2^{|D|}}$ be all the subsets of D where $|.|$ is the number of elements in $(.)$. Let

$$k(A_m, A_{m'}) = 2^{|A_m \cap A_{m'}|}. \tag{6.77}$$

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

$$\phi_m(A) = \begin{cases} 1, & \text{if } A_m \subseteq A, \\ 0, & \text{otherwise.} \end{cases} \tag{6.78}$$

If

$$|A_m \cap A_{m'}| = 0, \tag{6.79}$$

then

$$k(A_m, A_{m'}) = 1. \tag{6.80}$$

204

The right hand side can be written as

$$\phi(A_m)^\mathsf{T}\phi(A_{m'}).\tag{6.81}$$

Let us assume that if

$$|A_m \cap A_{m'}| = n,\tag{6.82}$$

then

$$k(A_m, A_{m'}) = \phi(A_m)^\mathsf{T}\phi(A_{m'}).\tag{6.83}$$

Let

$$A_{m''} = A_{m'} \cup \{x\},\tag{6.84}$$

where x is an element such that

$$x \in A_m - A_{m'}.\tag{6.85}$$

Then,

$$|A_m \cap A_{m''}| = n + 1,\tag{6.86}$$

so that

$$k(A_m, A_{m''}) = 2^{n+1}.\tag{6.87}$$

We have

$$\phi(A_m)^\mathsf{T}\phi(A_{m''}) = \phi(A_m)^\mathsf{T}\phi(A_{m'}) + \phi(A_m)^\mathsf{T}\left(\phi(A_{m''}) - \phi(A_{m'})\right).\tag{6.88}$$

By the assumption, the first term of the right hand side is 2^n. By the assumption, $\phi(A_m)$ and $\phi(A_{m'})$ have a set of 2^n 1s in common. Then, $\phi(A_m)$ and $\phi(A_{m''}) - \phi(A_{m'})$ have another set of 2^n 1s in common. Then, the second term of the right hand side is 2^n. Then,

$$\phi(A_m)^\mathsf{T}\phi(A_{m''}) = 2^{n+1},\tag{6.89}$$

so that

$$k(A_m, A_{m''}) = \phi(A_m)^\mathsf{T}\phi(A_{m''}).\tag{6.90}$$

Therefore, the assumption is proved by induction on n.

6.13

Let

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}'), \tag{6.91}$$

where

$$\begin{aligned} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) &= \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}), \\ \mathbf{F} &= \mathrm{E}_{\mathbf{x}} \left(\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} \right). \end{aligned} \tag{6.92}$$

Let $\boldsymbol{\psi}$ be an invertible and differentiable function of $\boldsymbol{\theta}$. Let

$$k_{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{x}') = \mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} \mathbf{F}_{\boldsymbol{\psi}}^{-1} \mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x}'), \tag{6.93}$$

where

$$\begin{aligned} \mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x}) &= \nabla_{\boldsymbol{\psi}} \ln p(\mathbf{x}|\boldsymbol{\theta}), \\ \mathbf{F}_{\boldsymbol{\psi}} &= \mathrm{E}_{\mathbf{x}} \left(\mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} \right). \end{aligned} \tag{6.94}$$

We have

$$\mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x}) = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\psi}} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}). \tag{6.95}$$

Then,

$$\mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x}) = \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}). \tag{6.96}$$

We have

$$\mathbf{F}_{\boldsymbol{\psi}} = \int \mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x}. \tag{6.97}$$

The right hand side can be written as

$$\begin{aligned} &\int \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} \left(\left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x} \\ &= \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \left(\int \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} p(\mathbf{x}) d\mathbf{x} \right) \left(\left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^{\mathsf{T}}. \end{aligned} \tag{6.98}$$

Then,

$$\mathbf{F}_{\boldsymbol{\psi}} = \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{F} \left(\left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^{\mathsf{T}}. \tag{6.99}$$

Then,

$$\begin{aligned} &k_{\boldsymbol{\psi}}(\mathbf{x}, \mathbf{x}') \\ &= \mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^{\mathsf{T}} \left(\left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \right)^{\mathsf{T}} \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{\mathsf{T}} \mathbf{F}^{-1} \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}'). \end{aligned} \tag{6.100}$$

206

The right hand side can be written as

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x})^\mathsf{T} \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'). \tag{6.101}$$

Therefore,

$$k_\psi(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}'). \tag{6.102}$$

6.14

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

$$p(\mathbf{x}|\boldsymbol{\mu}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}). \tag{6.103}$$

Let

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^\mathsf{T} \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}'), \tag{6.104}$$

where

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}|\boldsymbol{\mu}), \\ \mathbf{F} = \mathrm{E}_\mathbf{x}\left(\mathbf{g}(\boldsymbol{\mu}, \mathbf{x})\mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^\mathsf{T}\right). \tag{6.105}$$

We have

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x}) = -\mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{6.106}$$

Then,

$$\mathbf{F} = \int \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{S}^{-1} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) d\mathbf{x}. \tag{6.107}$$

The right hand side can be written as

$$\mathbf{S}^{-1} \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S}) d\mathbf{x} \right) \mathbf{S}^{-1} = \mathbf{S}^{-1}. \tag{6.108}$$

Then,

$$\mathbf{F} = \mathbf{S}^{-1}, \tag{6.109}$$

so that

$$\mathbf{g}(\boldsymbol{\mu}, \mathbf{x})^\mathsf{T} \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\mu}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{6.110}$$

Therefore,

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{6.111}$$

6.15

Let k be a positive definite kernel function. Then,

$$\sum_{n=1}^{N}\sum_{n'=1}^{N} c_n c_{n'} k(x_n, x_{n'}) > 0. \tag{6.112}$$

Then,

$$\mathbf{c}^{\mathsf{T}}\mathbf{K}\mathbf{c} \geq 0, \tag{6.113}$$

where

$$\mathbf{c} = \begin{bmatrix} c \\ c' \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(x, x) & k(x, x') \\ k(x, x') & k(x', x') \end{bmatrix}. \tag{6.114}$$

Then,

$$\det \mathbf{K} \geq 0. \tag{6.115}$$

Therefore,

$$k(x, x')^2 \leq k(x, x)k(x', x'). \tag{6.116}$$

6.16

Let

$$J(\mathbf{w}) = f\left(\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}_1), \cdots, \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}_N)\right) + g\left(\|\mathbf{w}\|^2\right), \tag{6.117}$$

where g is a monotonically increasing function. Setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0} = \sum_{n=1}^{N}(\nabla_n f)\boldsymbol{\phi}(\mathbf{x}_n) + 2g'\left(\|\mathbf{w}\|^2\right)\mathbf{w}, \tag{6.118}$$

where ∇_n is the partial derivative with respect to the nth argument. Then,

$$\mathbf{w} = -\frac{1}{2g'(\|\mathbf{w}\|^2)}\sum_{n=1}^{N}(\nabla_n f)\boldsymbol{\phi}(\mathbf{x}_n). \tag{6.119}$$

Therefore, J is minimised by $\mathbf{w}$ in the form of a linear combination of $\boldsymbol{\phi}(\mathbf{x}_1), \cdots, \boldsymbol{\phi}(\mathbf{x}_N)$.

6.17

Let
$$E_{\boldsymbol{\xi}} = \frac{1}{2} \sum_{n=1}^{N} \int \left(y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n \right)^2 p(\boldsymbol{\xi}) d\boldsymbol{\xi}. \tag{6.120}$$

By the transformation
$$\mathbf{z}_n = \mathbf{x}_n + \boldsymbol{\xi}, \tag{6.121}$$

we have
$$E = \frac{1}{2} \sum_{n=1}^{N} \int \left(y(\mathbf{z}_n) - t_n \right)^2 p(\mathbf{z}_n - \mathbf{x}_n) d\mathbf{z}_n. \tag{6.122}$$

Setting the variation with respect to y to zero gives
$$0 = \sum_{n=1}^{N} \int \left(y(\mathbf{z}_n) - t_n \right) p(\mathbf{z}_n - \mathbf{x}_n) \delta y(\mathbf{z}_n) d\mathbf{z}_n. \tag{6.123}$$

Then,
$$\sum_{n=1}^{N} \left(y(\mathbf{z}_n) - t_n \right) p(\mathbf{z}_n - \mathbf{x}_n) = 0, \tag{6.124}$$

so that
$$\sum_{n=1}^{N} y(\mathbf{z}_n) p(\mathbf{z}_n - \mathbf{x}_n) = \sum_{n=1}^{N} t_n p(\mathbf{z}_n - \mathbf{x}_n). \tag{6.125}$$

Therefore,
$$y(\mathbf{x}) = \sum_{n=1}^{N} t_n k(\mathbf{x}, \mathbf{x}_n), \tag{6.126}$$

where
$$k(\mathbf{x}, \mathbf{x}_n) = \frac{p(\mathbf{x} - \mathbf{x}_n)}{\sum_{n=1}^{N} p(\mathbf{x} - \mathbf{x}_n)}. \tag{6.127}$$

6.18

Let x be a variable such that
$$p(x, t) = \frac{1}{N} \sum_{n=1}^{N} f(x - x_n, t - t_n), \tag{6.128}$$

where
$$f(x, t) = \mathcal{N}\left(x | 0, \sigma^2 \right) \mathcal{N}\left(t | 0, \sigma^2 \right). \tag{6.129}$$

(a)

By the Bayes' theorem,

$$p(t|x) = \frac{p(x,t)}{p(x)}. \tag{6.130}$$

By marginalisation,

$$p(x) = \int p(x,t)dt. \tag{6.131}$$

The right hand side can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(x - x_n|0,\sigma^2\right)\int\mathcal{N}\left(t - t_n|0,\sigma^2\right)dt$$
$$= \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(x - x_n|0,\sigma^2\right). \tag{6.132}$$

Then,

$$p(t|x) = \frac{\sum_{n=1}^{N}\mathcal{N}\left(x - x_n|0,\sigma^2\right)\mathcal{N}\left(t - t_n|0,\sigma^2\right)}{\sum_{n=1}^{N}\mathcal{N}\left(x - x_n|0,\sigma^2\right)}. \tag{6.133}$$

Therefore,

$$p(t|x) = \sum_{n=1}^{N}k(x,x_n)\mathcal{N}\left(t - t_n|0,\sigma^2\right), \tag{6.134}$$

where

$$k(x,x_n) = \frac{\mathcal{N}\left(x - x_n|0,\sigma^2\right)}{\sum_{n=1}^{N}\mathcal{N}\left(x - x_n|0,\sigma^2\right)}. \tag{6.135}$$

(b)

We have

$$\mathrm{E}(t|x) = \int tp(t|x)dt. \tag{6.136}$$

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

$$\sum_{n=1}^{N}k(x,x_n)\int t\mathcal{N}\left(t - t_n|0,\sigma^2\right)dt. \tag{6.137}$$

By the transformation

$$t' = t - t_n, \tag{6.138}$$

the integral can be written as

$$\int (t'+t_n)\mathcal{N}\left(t'|0,\sigma^2\right)dt' = \int t'\mathcal{N}\left(t'|0,\sigma^2\right)dt' + t_n\int \mathcal{N}\left(t'|0,\sigma^2\right)dt'. \quad (6.139)$$

The right hand side can be written as t_n. Therefore,

$$E(t|x) = \sum_{n=1}^{N} k(x,x_n)t_n. \quad (6.140)$$

(c)

We have

$$\text{var}(t|x) = \int \left(t - E(t|x)\right)^2 p(t|x)dt. \quad (6.141)$$

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

$$\sum_{n=1}^{N} k(x,x_n)\int \left(t - \sum_{n=1}^{N} k(x,x_n)t_n\right)^2 \mathcal{N}\left(t - t_n|0,\sigma^2\right)dt. \quad (6.142)$$

By the transformation

$$t' = t - t_n, \quad (6.143)$$

the integral can be written as

$$\int \left(t' + t_n - \sum_{n=1}^{N} k(x,x_n)t_n\right)^2 \mathcal{N}\left(t'|0,\sigma^2\right)dt'$$

$$= \int t'^2\mathcal{N}\left(t'|0,\sigma^2\right)dt' + 2\left(t_n - \sum_{n=1}^{N} k(x,x_n)t_n\right)\int t'\mathcal{N}\left(t'|0,\sigma^2\right)dt' \quad (6.144)$$

$$+ \left(t_n - \sum_{n=1}^{N} k(x,x_n)t_n\right)^2 \int \mathcal{N}\left(t'|0,\sigma^2\right)dt'.$$

The right hand side can be written as

$$\sigma^2 + \left(t_n - \sum_{n=1}^{N} k(x,x_n)t_n\right)^2. \quad (6.145)$$

Therefore,

$$\text{var}(t|x) = \sum_{n=1}^{N} k(x, x_n) \left(\sigma^2 + \left(t_n - \sum_{n'=1}^{N} k(x, x_{n'}) t_{n'} \right)^2 \right), \tag{6.146}$$

so that

$$\text{var}(t|x) = \sigma^2 + \sum_{n=1}^{N} k(x, x_n) \left(t_n - \sum_{n'=1}^{N} k(x, x_{n'}) t_{n'} \right)^2. \tag{6.147}$$

6.19

Let

$$E = \frac{1}{2} \sum_{n=1}^{N} \int \left(y(\mathbf{x}_n - \boldsymbol{\xi}_n) - t_n \right)^2 p(\boldsymbol{\xi}_n) d\boldsymbol{\xi}_n. \tag{6.148}$$

By the transformation

$$\mathbf{z}_n = \mathbf{x}_n - \boldsymbol{\xi}_n, \tag{6.149}$$

we have

$$E = \frac{1}{2} \sum_{n=1}^{N} \int \left(y(\mathbf{z}_n) - t_n \right)^2 p(\mathbf{x}_n - \mathbf{z}_n) d\mathbf{z}_n. \tag{6.150}$$

Setting the variation with respect to y to zero gives

$$0 = \sum_{n=1}^{N} \int \left(y(\mathbf{z}_n) - t_n \right) p(\mathbf{x}_n - \mathbf{z}_n) \delta y(\mathbf{z}_n) d\mathbf{z}_n. \tag{6.151}$$

Then,

$$0 = \sum_{n=1}^{N} \left(y(\mathbf{z}_n) - t_n \right) p(\mathbf{x}_n - \mathbf{z}_n), \tag{6.152}$$

so that

$$\sum_{n=1}^{N} y(\mathbf{z}_n) p(\mathbf{x}_n - \mathbf{z}_n) = \sum_{n=1}^{N} t_n p(\mathbf{x}_n - \mathbf{z}_n). \tag{6.153}$$

Therefore,

$$y(\mathbf{x}) = \sum_{n=1}^{N} t_n k(\mathbf{x}, \mathbf{x}_n), \tag{6.154}$$

where

$$k(\mathbf{x}, \mathbf{x}_n) = \frac{p(\mathbf{x}_n - \mathbf{x})}{\sum_{n=1}^{N} p(\mathbf{x}_n - \mathbf{x})}. \tag{6.155}$$

6.20

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}\right), \\
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}),
\end{aligned}
\tag{6.156}
$$

where

$$
K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}). \tag{6.157}
$$

By the Bayes' theorem,

$$
p(t_{N+1}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}'), \tag{6.158}
$$

where

$$
\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \end{bmatrix}. \tag{6.159}
$$

By marginalisation,

$$
p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y}. \tag{6.160}
$$

The logarighm of the integrand of the right hand side except the terms independent of $\mathbf{y}_N$ can be written as

$$
-\frac{\beta}{2}(\mathbf{t} - \mathbf{y})^{\mathsf{T}}(\mathbf{t} - \mathbf{y}) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{K}^{-1}\mathbf{y} = -\frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v}, \tag{6.161}
$$

where

$$
\mathbf{v} = \begin{bmatrix} \mathbf{y} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I} \\ -\beta\mathbf{I} & \beta\mathbf{I} \end{bmatrix}. \tag{6.162}
$$

By 2.24,

$$
\begin{bmatrix} \beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I} \\ -\beta\mathbf{I} & \beta\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{K} & \mathbf{K} \\ \mathbf{K} & \mathbf{K} + \beta^{-1}\mathbf{I} \end{bmatrix}. \tag{6.163}
$$

Then,

$$
p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \mathbf{C}\right), \tag{6.164}
$$

where

$$
\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}. \tag{6.165}
$$

Then,

$$
p(\mathbf{t}') = \mathcal{N}\left(\mathbf{t}'|\mathbf{0}, \mathbf{C}'\right), \tag{6.166}
$$

where

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\mathsf{T} & c \end{bmatrix}, \tag{6.167}$$

where

$$\begin{aligned} k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}. \end{aligned} \tag{6.168}$$

By 2.24,

$$\begin{bmatrix} c & \mathbf{k}^\mathsf{T} \\ \mathbf{k} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} s^{-1} & -s^{-1}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} \\ -s^{-1}\mathbf{C}^{-1}\mathbf{k} & \mathbf{C}^{-1} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} \end{bmatrix}, \tag{6.169}$$

where

$$s = c - \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{k}. \tag{6.170}$$

Therefore,

$$p(t_{N+1}|\mathbf{t}) = \mathcal{N}\left(t_{N+1}|m, s\right), \tag{6.171}$$

where

$$m = \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{t}. \tag{6.172}$$

6.21

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{6.173}$$

Then,

$$\begin{aligned} p(\mathbf{t}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}\right), \\ p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}\right), \end{aligned} \tag{6.174}$$

where

$$\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\mathsf{T}. \tag{6.175}$$

By 6.20,

$$p(t_{N+1}|\mathbf{t}) = \mathcal{N}\left(t_{N+1}|m, s\right), \tag{6.176}$$

where

$$\begin{aligned} m &= \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{t}, \\ s &= c - \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{k}, \end{aligned} \tag{6.177}$$

where
$$\mathbf{k} = \mathbf{\Phi}\boldsymbol{\phi}_{N+1},$$
$$\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}, \tag{6.178}$$
$$c = \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\boldsymbol{\phi}_{N+1} + \beta^{-1}.$$

By 2.26,
$$\mathbf{C}^{-1} = \beta\mathbf{I} - \beta^2\mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}, \tag{6.179}$$

where
$$\mathbf{S} = \left(\mathbf{I} + \beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}\right)^{-1}. \tag{6.180}$$

Then,
$$m = \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\left(\beta\mathbf{I} - \beta^2\mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\right)\mathbf{t},$$
$$s = \beta^{-1} + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\boldsymbol{\phi}_{N+1} - \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\left(\beta\mathbf{I} - \beta^2\mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\right)\mathbf{\Phi}\boldsymbol{\phi}_{N+1}. \tag{6.181}$$

We have
$$\beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} = \mathbf{S}^{-1} - \mathbf{I}. \tag{6.182}$$

Then,
$$m = \beta\boldsymbol{\phi}_{N+1}^{\mathsf{T}}\left(\mathbf{I} - \left(\mathbf{S}^{-1} - \mathbf{I}\right)\mathbf{S}\right)\mathbf{\Phi}^{\mathsf{T}}\mathbf{t},$$
$$s = \beta^{-1} + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\left(\mathbf{I} - \left(\mathbf{S}^{-1} - \mathbf{I}\right) + \left(\mathbf{S}^{-1} - \mathbf{I}\right)\mathbf{S}\left(\mathbf{S}^{-1} - \mathbf{I}\right)\right)\boldsymbol{\phi}_{N+1}. \tag{6.183}$$

Therefore,
$$m = \beta\boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\mathbf{t},$$
$$s = \beta^{-1} + \boldsymbol{\phi}_{N+1}^{\mathsf{T}}\mathbf{S}\boldsymbol{\phi}_{N+1}. \tag{6.184}$$

6.22

Let $t_1, \cdots, t_N$ be variables such that
$$p(\mathbf{t}|\mathbf{y}) = \mathcal{N}\left(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}\right),$$
$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}\right), \tag{6.185}$$

where
$$K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}). \tag{6.186}$$

(a)

By the Bayes' theorem,

$$p(t_{N+1}, \cdots, t_{N+L}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}'), \qquad (6.187)$$

where

$$\mathbf{t}' = \begin{bmatrix} \mathbf{t} \\ t_{N+1} \\ \vdots \\ t_{N+L} \end{bmatrix}. \qquad (6.188)$$

By 6.20,

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}), \qquad (6.189)$$

where

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{K}. \qquad (6.190)$$

Let $\mathbf{K}'$ and $\mathbf{\Gamma}$ be matrices where

$$\begin{aligned} K'_{nl} &= k(\mathbf{x}_n, \mathbf{x}_{N+l}), \\ \Gamma_{ll'} &= \beta^{-1}I_{ll'} + k(\mathbf{x}_{N+l}, \mathbf{x}_{N+l'}). \end{aligned} \qquad (6.191)$$

Then,

$$p(\mathbf{t}') = \mathcal{N}(\mathbf{t}'|\mathbf{0}, \mathbf{C}'), \qquad (6.192)$$

where

$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{K}' \\ \mathbf{K}'^{\mathsf{T}} & \mathbf{\Gamma} \end{bmatrix}. \qquad (6.193)$$

By 2.24,

$$\begin{bmatrix} \mathbf{\Gamma} & \mathbf{K}'^{\mathsf{T}} \\ \mathbf{K}' & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{K}'^{\mathsf{T}}\mathbf{C}^{-1} \\ -\mathbf{C}^{-1}\mathbf{K}'\mathbf{S}^{-1} & \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{K}'\mathbf{S}^{-1}\mathbf{K}'^{\mathsf{T}}\mathbf{C}^{-1} \end{bmatrix}, \qquad (6.194)$$

where

$$\mathbf{S} = \mathbf{\Gamma} - \mathbf{K}'^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{K}'. \qquad (6.195)$$

Therefore,

$$p(t_{N+1}, \cdots, t_{N+L}|\mathbf{t}_N) = \mathcal{N}(t_{N+1}, \cdots, t_{N+L}|\mathbf{m}, \mathbf{S}), \qquad (6.196)$$

where

$$\mathbf{m} = \mathbf{K}'^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{t}. \qquad (6.197)$$

216

(b)

By (a),

$$p(t_{N+1}|\mathbf{t}) = \mathcal{N}\left(t_{N+1}|m, s\right), \tag{6.198}$$

where

$$\begin{aligned} m &= \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{t}, \\ s &= c - \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{k}, \end{aligned} \tag{6.199}$$

and

$$\begin{aligned} k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= \beta^{-1} + k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}). \end{aligned} \tag{6.200}$$

6.23

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

$$\begin{aligned} p(\mathbf{t}_n|\mathbf{y}_n) &= \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}_n, \beta^{-1}\mathbf{I}\right), \\ p(\mathbf{Y}) &= \mathcal{N}\left(\mathbf{Y}|\mathbf{O}, \mathbf{K}\right), \end{aligned} \tag{6.201}$$

where

$$K_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}). \tag{6.202}$$

By the Bayes' theorem,

$$p(\mathbf{t}_{N+1}|\mathbf{T})p(\mathbf{T}) = p(\mathbf{T}'). \tag{6.203}$$

By marginalisation,

$$p(\mathbf{T}) = \int p(\mathbf{T}|\mathbf{Y})p(\mathbf{Y})d\mathbf{Y}. \tag{6.204}$$

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

$$-\frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{t}_n - \mathbf{y}_n\|^2 - \frac{1}{2}\operatorname{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{K}^{-1}\mathbf{Y}\right) = -\frac{1}{2}\operatorname{tr}\mathbf{M}, \tag{6.205}$$

where

$$\mathbf{M} = \begin{bmatrix}\mathbf{Y}\\\mathbf{T}\end{bmatrix}^\mathsf{T}\begin{bmatrix}\beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I}\\-\beta\mathbf{I} & \beta\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{Y}\\\mathbf{T}\end{bmatrix}. \tag{6.206}$$

By 2.24,

$$\begin{bmatrix}\beta\mathbf{I} + \mathbf{K}^{-1} & -\beta\mathbf{I}\\-\beta\mathbf{I} & \beta\mathbf{I}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{K} & \mathbf{K}\\\mathbf{K} & \beta^{-1}\mathbf{I} + \mathbf{K}\end{bmatrix}. \tag{6.207}$$

Then,
$$p(\mathbf{T}) = \mathcal{N}\left(\mathbf{T}|\mathbf{O},\mathbf{C}\right), \tag{6.208}$$

where
$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{K}. \tag{6.209}$$

Then,
$$p(\mathbf{T}') = \mathcal{N}\left(\mathbf{T}'|\mathbf{O},\mathbf{C}'\right), \tag{6.210}$$

where
$$\mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\mathsf{T} & c \end{bmatrix}, \tag{6.211}$$

and
$$\begin{aligned} k_n &= k(\mathbf{x}_n, \mathbf{x}_{N+1}), \\ c &= \beta^{-1} + k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}). \end{aligned} \tag{6.212}$$

By 2.24,
$$\begin{bmatrix} c & \mathbf{k}^\mathsf{T} \\ \mathbf{k} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} s^{-1} & -s^{-1}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} \\ -s^{-1}\mathbf{C}^{-1}\mathbf{k} & \mathbf{C}^{-1} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} \end{bmatrix}, \tag{6.213}$$

where
$$s = c - \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{k}. \tag{6.214}$$

Therefore,
$$p(\mathbf{t}_{N+1}|\mathbf{T}) = \mathcal{N}\left(\mathbf{t}_{N+1}|\mathbf{m}, s\mathbf{I}\right), \tag{6.215}$$

where
$$\mathbf{m} = \mathbf{T}^\mathsf{T}\mathbf{C}^{-1}\mathbf{k}. \tag{6.216}$$

6.24

(a)

Let $\mathbf{M}$ be a $D \times D$ diagonal matrix such that $0 < M_{dd} < 1$. For any vector $\mathbf{v}$ in D dimensions,
$$\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v} = \sum_{d=1}^{D}\sum_{d'=1}^{D} v_d M_{dd'} v_{d'}. \tag{6.217}$$

The right hand side can be written as
$$\sum_{d=1}^{D} M_{dd} v_d^2 > 0. \tag{6.218}$$

Therefore, $\mathbf{M}$ is positive definite.

(b)

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be $D \times D$ positive definite matrices. Then, for any vector $\mathbf{v}$ in D dimensions,

$$\begin{aligned} \mathbf{v}^\mathsf{T} \mathbf{M}_1 \mathbf{v} &> 0, \\ \mathbf{v}^\mathsf{T} \mathbf{M}_2 \mathbf{v} &> 0. \end{aligned} \tag{6.219}$$

Then,

$$\mathbf{v}^\mathsf{T} (\mathbf{M}_1 + \mathbf{M}_2) \mathbf{v} > 0. \tag{6.220}$$

Therefore, $\mathbf{M}_1 + \mathbf{M}_2$ is positive definite.

6.25

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n | a_n) &= \sigma(a_n)^{t_n} \left(1 - \sigma(a_n)\right)^{1 - t_n}, \\ p(\mathbf{a}) &= \mathcal{N}\left(\mathbf{a} | \mathbf{0}, \mathbf{C}\right), \end{aligned} \tag{6.221}$$

where

$$\begin{aligned} \sigma(a) &= \frac{1}{1 + \exp(-a)}, \\ C_{nn'} &= k(\mathbf{x}_n, \mathbf{x}_{n'}) + \nu I_{nn'}. \end{aligned} \tag{6.222}$$

(a)

By the Bayes' theorem,

$$p(\mathbf{a} | \mathbf{t}) p(\mathbf{t}) = p(\mathbf{t} | \mathbf{a}) p(\mathbf{a}). \tag{6.223}$$

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

$$\Psi(\mathbf{a}) = \sum_{n=1}^{N} \left(t_n \ln \sigma(a_n) + (1 - t_n) \ln\left(1 - \sigma(a_n)\right)\right) - \frac{1}{2} \mathbf{a}^\mathsf{T} \mathbf{C}^{-1} \mathbf{a}. \tag{6.224}$$

In order to maximise $p(\mathbf{a} | \mathbf{t})$ with respect to $\mathbf{a}$, Ψ needs to be maximised. We have

$$\frac{d}{da} \sigma(a) = \sigma(a) \left(1 - \sigma(a)\right). \tag{6.225}$$

Then,

$$(\nabla\Psi(\mathbf{a}))_n = (t_n\left(1 - \sigma(a_n)\right) - (1 - t_n)\sigma(a_n)) - \left(\mathbf{C}^{-1}\mathbf{a}\right)_n, \tag{6.226}$$

so that

$$\nabla\Psi(\mathbf{a}) = \mathbf{t} - \boldsymbol{\sigma} - \mathbf{C}^{-1}\mathbf{a}, \tag{6.227}$$

where

$$\sigma_n = \sigma(a_n). \tag{6.228}$$

Since $\boldsymbol{\sigma}$ non-linearly depends on $\mathbf{a}$, the Taylor series

$$\nabla\Psi(\mathbf{a}^{\text{new}}) = \nabla\Psi(\mathbf{a}) + \nabla\nabla\Psi(\mathbf{a})\left(\mathbf{a}^{\text{new}} - \mathbf{a}\right) + O\left(\|\mathbf{a} - \mathbf{a}^{\text{new}}\|^2\right). \tag{6.229}$$

is used to iteratively find $\mathbf{a}$ which maximises Ψ. Setting the left hand side to zero and omitting the second order term of the right hand side gives

$$\mathbf{0} = \nabla\Psi(\mathbf{a}) + \nabla\nabla\Psi(\mathbf{a})\left(\mathbf{a}^{\text{new}} - \mathbf{a}\right). \tag{6.230}$$

Then,

$$\mathbf{a}^{\text{new}} = \mathbf{a} - \left(\nabla\nabla\Psi(\mathbf{a})\right)^{-1}\nabla\Psi(\mathbf{a}). \tag{6.231}$$

We have

$$\nabla\nabla\Psi(\mathbf{a}) = -\mathbf{W} - \mathbf{C}^{-1}, \tag{6.232}$$

where

$$W_{nn'} = \sigma(a_n)\left(1 - \sigma(a_n)\right)I_{nn'}. \tag{6.233}$$

Then,

$$\mathbf{a}^{\text{new}} = \mathbf{a} + \left(\mathbf{W} + \mathbf{C}^{-1}\right)^{-1}\left(\mathbf{t} - \boldsymbol{\sigma} - \mathbf{C}^{-1}\mathbf{a}\right). \tag{6.234}$$

The right hand side can be written as

$$\begin{aligned}
&\left(\mathbf{W} + \mathbf{C}^{-1}\right)^{-1}\left(\left(\mathbf{W} + \mathbf{C}^{-1}\right)\mathbf{a} + \mathbf{t} - \boldsymbol{\sigma} - \mathbf{C}^{-1}\mathbf{a}\right)\\
=&\mathbf{C}\left(\mathbf{W}\mathbf{C} + \mathbf{I}\right)^{-1}\left(\mathbf{W}\mathbf{a} + \mathbf{t} - \boldsymbol{\sigma}\right).
\end{aligned} \tag{6.235}$$

Therefore,

$$\mathbf{a}^{\text{new}} = \mathbf{C}\left(\mathbf{W}\mathbf{C} + \mathbf{I}\right)^{-1}\left(\mathbf{W}\mathbf{a} + \mathbf{t} - \boldsymbol{\sigma}\right). \tag{6.236}$$

(b)

Let $\mathbf{a}_{\mathrm{MAP}}$ be a stationary point of $p(\mathbf{a}|\mathbf{t})$ which is found by the iterative process given in (a). We have the Taylor series

$$
\begin{aligned}
\Psi(\mathbf{a}) =\,& \Psi(\mathbf{a}_{\mathrm{MAP}}) + \nabla\Psi(\mathbf{a}_{\mathrm{MAP}})(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}) \\
& + \frac{1}{2}(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}})^{\mathsf{T}}\nabla\nabla\Psi(\mathbf{a}_{\mathrm{MAP}})(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}) \\
& + O\left(\|\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}\|^3\right).
\end{aligned}
\tag{6.237}
$$

Then,

$$
\Psi(\mathbf{a}) \simeq \Psi(\mathbf{a}_{\mathrm{MAP}}) - \frac{1}{2}(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}})^{\mathsf{T}}\mathbf{H}(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}),
\tag{6.238}
$$

where

$$
\mathbf{H} = \mathbf{W}\left.\right|_{\mathbf{a}=\mathbf{a}_{\mathrm{MAP}}} + \mathbf{C}^{-1}.
\tag{6.239}
$$

Therefore,

$$
p(\mathbf{a}|\mathbf{t}) \simeq \mathcal{N}\left(\mathbf{a}|\mathbf{a}_{\mathrm{MAP}}, \mathbf{H}^{-1}\right).
\tag{6.240}
$$

(c)

By (a),

$$
\mathbf{a}_{\mathrm{MAP}} = \mathbf{C}\left(\mathbf{WC} + \mathbf{I}\right)^{-1}\left(\mathbf{Wa}_{\mathrm{MAP}} + \mathbf{t} - \boldsymbol{\sigma}\right).
\tag{6.241}
$$

Then,

$$
\left(\mathbf{WC} + \mathbf{I}\right)\mathbf{C}^{-1}\mathbf{a}_{\mathrm{MAP}} = \mathbf{Wa}_{\mathrm{MAP}} + \mathbf{t} - \boldsymbol{\sigma}.
\tag{6.242}
$$

Therefore,

$$
\mathbf{a}_{\mathrm{MAP}} = \mathbf{C}(\mathbf{t} - \boldsymbol{\sigma}).
\tag{6.243}
$$

6.26

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|a_n) &= \sigma(a_n)^{t_n}\left(1 - \sigma(a_n)\right)^{1-t_n}, \\
p(\mathbf{a}) &= \mathcal{N}\left(\mathbf{a}|\mathbf{0}, \mathbf{C}\right),
\end{aligned}
\tag{6.244}
$$

where

$$
\begin{aligned}
\sigma(a) &= \frac{1}{1 + \exp(-a)}, \\
C_{nn'} &= k(\mathbf{x}_n, \mathbf{x}_{n'}) + \nu I_{nn'}.
\end{aligned}
\tag{6.245}
$$

By marginalisation,

$$p(a_{N+1}|\mathbf{t}) = \int p(a_{N+1}|\mathbf{a})p(\mathbf{a}|\mathbf{t})d\mathbf{a}. \tag{6.246}$$

We have

$$p(\mathbf{a}') = \mathcal{N}\left(\mathbf{a}'|\mathbf{0}, \mathbf{C}'\right), \tag{6.247}$$

where

$$\mathbf{a}' = \begin{bmatrix} \mathbf{a} \\ a_{N+1} \end{bmatrix}, \mathbf{C}' = \begin{bmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^\mathsf{T} & c \end{bmatrix}, \tag{6.248}$$

and

$$k_n = k(\mathbf{x}_n, \mathbf{x}_{N+1}), \tag{6.249}$$
$$c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \nu.$$

By 2.24,

$$\begin{bmatrix} c & \mathbf{k}^\mathsf{T} \\ \mathbf{k} & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} s^{-1} & -s^{-1}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} \\ -s^{-1}\mathbf{C}^{-1}\mathbf{k} & \mathbf{C}^{-1} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} \end{bmatrix}, \tag{6.250}$$

where

$$s = c - \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{k}. \tag{6.251}$$

Then,

$$p(a_{N+1}|\mathbf{a}) = \mathcal{N}\left(a_{N+1}|m, s\right), \tag{6.252}$$

where

$$m = \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{a}. \tag{6.253}$$

By 6.25(b),

$$p(\mathbf{a}|\mathbf{t}) \simeq \mathcal{N}\left(\mathbf{a}|\mathbf{a}_{\mathrm{MAP}}, \mathbf{H}^{-1}\right), \tag{6.254}$$

where $\mathbf{a}_{\mathrm{MAP}}$ is a stationary point of $p(\mathbf{a}|\mathbf{t})$ and

$$\mathbf{H} = \mathbf{W}\left.\right|_{\mathbf{a}=\mathbf{a}_{\mathrm{MAP}}} + \mathbf{C}^{-1},$$
$$W_{nn'} = \sigma(a_n)\left(1 - \sigma(a_n)\right)I_{nn'}. \tag{6.255}$$

Then, the logarithm of the integrand except the terms independent of $\mathbf{t}$ and $\mathbf{a}$ can be approximated as

$$-\frac{1}{2s}\left(a_{N+1} - m\right)^2 - \frac{1}{2}\left(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}\right)^\mathsf{T}\mathbf{H}\left(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}\right)$$
$$= -\frac{1}{2}\mathbf{a}'^\mathsf{T}\mathbf{M}\mathbf{a}' + \mathbf{a}'^\mathsf{T}\mathbf{v} - \frac{1}{2}\mathbf{a}_{\mathrm{MAP}}^\mathsf{T}\mathbf{H}\mathbf{a}_{\mathrm{MAP}}. \tag{6.256}$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{H} + s^{-1}\mathbf{C}^{-1}\mathbf{k}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} & -s^{-1}\mathbf{C}^{-1}\mathbf{k} \\ -s^{-1}\mathbf{k}^\mathsf{T}\mathbf{C}^{-1} & s^{-1} \end{bmatrix}, \mathbf{v} = \begin{bmatrix} \mathbf{H}\mathbf{a}_{\mathrm{MAP}} \\ 0 \end{bmatrix}. \qquad (6.257)$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{H}^{-1} & \mathbf{H}^{-1}\mathbf{C}^{-1}\mathbf{k} \\ \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{H}^{-1} & s + \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{H}^{-1}\mathbf{C}^{-1}\mathbf{k} \end{bmatrix}, \qquad (6.258)$$

so that

$$\mathbf{M}^{-1}\mathbf{v} = \begin{bmatrix} \mathbf{a}_{\mathrm{MAP}} \\ \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{a}_{\mathrm{MAP}} \end{bmatrix}. \qquad (6.259)$$

Then,

$$p(a_{N+1}|\mathbf{t}) \simeq \mathcal{N}\left(a_{N+1}|\mu, \sigma^2\right), \qquad (6.260)$$

where

$$\begin{aligned} \mu &= \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{a}_{\mathrm{MAP}}, \\ \sigma^2 &= s + \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\mathbf{H}^{-1}\mathbf{C}^{-1}\mathbf{k}. \end{aligned} \qquad (6.261)$$

By 6.25(c),

$$\mathbf{a}_{\mathrm{MAP}} = \mathbf{C}(\mathbf{t} - \boldsymbol{\sigma}), \qquad (6.262)$$

where

$$\sigma_n = \sigma(a_n). \qquad (6.263)$$

Therefore,

$$\begin{aligned} \mu &= \mathbf{k}^\mathsf{T}(\mathbf{t} - \boldsymbol{\sigma}), \\ \sigma^2 &= c - \mathbf{k}^\mathsf{T}\mathbf{C}^{-1}\left(\mathbf{C} - \mathbf{H}^{-1}\right)\mathbf{C}^{-1}\mathbf{k}. \end{aligned} \qquad (6.264)$$

6.27   (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|a_n) &= \sigma(a_n)^{t_n}\left(1 - \sigma(a_n)\right)^{1-t_n}, \\ p(\mathbf{a}|\boldsymbol{\theta}) &= \mathcal{N}\left(\mathbf{a}|\mathbf{0}, \mathbf{C}\right), \end{aligned} \qquad (6.265)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \qquad (6.266)$$

and $\mathbf{C}$ is dependent on $\boldsymbol{\theta}$ with M dimensions.

(a)

By marginalisation,

$$p(\mathbf{t}|\boldsymbol{\theta}) = \int p(\mathbf{t}|\mathbf{a})p(\mathbf{a}|\boldsymbol{\theta})d\mathbf{a}. \tag{6.267}$$

Let $\mathbf{a}_{\mathrm{MAP}}$ be a stationary point of $p(\mathbf{a}|\mathbf{t})$ and

$$\Psi(\mathbf{a}) = \ln p(\mathbf{t}|\mathbf{a}) + \ln p(\mathbf{a}|\boldsymbol{\theta}). \tag{6.268}$$

By 6.25(b),

$$\Psi(\mathbf{a}) \simeq \Psi(\mathbf{a}_{\mathrm{MAP}}) - \frac{1}{2}\left(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}\right)^{\mathsf{T}}\mathbf{H}\left(\mathbf{a} - \mathbf{a}_{\mathrm{MAP}}\right), \tag{6.269}$$

where

$$\mathbf{H} = \mathbf{W}\,|_{\mathbf{a}=\mathbf{a}_{\mathrm{MAP}}} + \mathbf{C}^{-1},$$
$$W_{nn'} = \sigma(a_n)\left(1 - \sigma(a_n)\right)I_{nn'}. \tag{6.270}$$

Therefore,

$$\ln p(\mathbf{t}|\boldsymbol{\theta}) \simeq \Psi(\mathbf{a}_{\mathrm{MAP}}) + \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{H}). \tag{6.271}$$

(b)

We have

$$\frac{\partial\Psi(\mathbf{a}_{\mathrm{MAP}})}{\partial\theta_m} = \sum_{n=1}^{N}\sum_{n'=1}^{N}\frac{\partial\Psi(\mathbf{a}_{\mathrm{MAP}})}{\partial C_{nn'}}\frac{\partial C_{nn'}}{\partial\theta_m}. \tag{6.272}$$

The right hand side can be written as

$$\frac{1}{2}\operatorname{tr}\left(\mathbf{a}_{\mathrm{MAP}}\mathbf{a}_{\mathrm{MAP}}^{\mathsf{T}}\left(\mathbf{C}^{-1}\right)^2\frac{\partial\mathbf{C}}{\partial\theta_m}\right). \tag{6.273}$$

We have

$$\frac{\partial\ln(\det\mathbf{H})}{\partial\theta_m} = \sum_{n=1}^{N}\sum_{n'=1}^{N}\frac{\partial\ln(\det\mathbf{H})}{\partial H_{nn'}}\frac{\partial H_{nn'}}{\partial\theta_m}. \tag{6.274}$$

By 3.21**??**, the right hand side can be written as

$$\operatorname{tr}\left(\mathbf{H}^{-1}\frac{\partial\mathbf{C}^{-1}}{\partial\theta_m}\right). \tag{6.275}$$

Therefore,

$$\frac{\partial\ln p(\mathbf{t}|\boldsymbol{\theta})}{\partial\theta_m} \simeq \frac{1}{2}\operatorname{tr}\left(\mathbf{a}_{\mathrm{MAP}}\mathbf{a}_{\mathrm{MAP}}^{\mathsf{T}}\left(\mathbf{C}^{-1}\right)^2\frac{\partial\mathbf{C}}{\partial\theta_m} - \mathbf{H}^{-1}\frac{\partial\mathbf{C}^{-1}}{\partial\theta_m}\right). \tag{6.276}$$

7 Sparse Kernel Machines

7.1 (Incomplete)

7.2

Let $t_1, \cdots, t_N$ be variables. In order to minimise $\|\mathbf{w}\|^2$ under the constraint

$$t_n \left(\mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n) + b\right) \geq \gamma, \tag{7.1}$$

for $\gamma > 0$, let

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left(t_n \left(\mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}_n) + b\right) - \gamma\right). \tag{7.2}$$

Setting the derivatives of L with respect to $\mathbf{w}$ and b to zero gives

$$\begin{aligned}
\mathbf{0} &= \mathbf{w} - \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n), \\
0 &= -\sum_{n=1}^{N} a_n t_n.
\end{aligned} \tag{7.3}$$

Therefore,

$$\begin{aligned}
\mathbf{w} &= \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n), \\
0 &= \sum_{n=1}^{N} a_n t_n,
\end{aligned} \tag{7.4}$$

which is independent of γ.

7.3 (Incomplete)

7.4 (Incomplete)

7.5 (Incomplete)

7.6

Let $t_1, \cdots, t_N$ be a variable such that

$$
t_n \in \{-1, 1\},
$$
$$
p(t_n|y_n) = \sigma(y_n)^{\frac{1+t_n}{2}} \left(1 - \sigma(y_n)\right)^{\frac{1-t_n}{2}}, \tag{7.5}
$$

where

$$
\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{7.6}
$$
$$
y_n = \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n) + b.
$$

We have

$$
1 - \sigma(a) = \sigma(-a). \tag{7.7}
$$

Then,

$$
p(t_n|y_n) = \sigma(y_n t_n). \tag{7.8}
$$

Then,

$$
p(\mathbf{t}|\mathbf{y}) = \prod_{n=1}^{N} \sigma(y_n t_n), \tag{7.9}
$$

so that

$$
-\ln p(\mathbf{t}|\mathbf{y}) = -\sum_{n=1}^{N} \ln \sigma(y_n t_n). \tag{7.10}
$$

Therefore,

$$
-\ln p(\mathbf{t}|\mathbf{y}) = \sum_{n=1}^{N} \ln\left(1 + \exp(-y_n t_n)\right). \tag{7.11}
$$

7.7

Let $t_1, \cdots, t_N$ be vaariables. Let

$$
E = C \sum_{n=1}^{N} E_\epsilon(y_n - t_n) + \frac{1}{2}\|\mathbf{w}\|^2, \tag{7.12}
$$

226

where $\epsilon > 0$ and

$$E_\epsilon(a) = \begin{cases} 0, & |a| < \epsilon, \\ |a| - \epsilon, & \text{otherwise,} \end{cases} \tag{7.13}$$

$$y_n = \mathbf{w}^\mathsf{T}\phi(\mathbf{x}_n) + b.$$

In order to minimise E under the constraints

$$t_n \leq y_n + \epsilon + \xi_n,$$
$$t_n \geq y_n - \epsilon - \hat{\xi}_n, \tag{7.14}$$

where $\xi_n \geq 0$ and $\hat{\xi}_n \geq 0$, let us minimise

$$\begin{aligned} L =& C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}\left(\mu_n\xi_n + \hat{\mu}_n\hat{\xi}_n\right) \\ & - \sum_{n=1}^{N}a_n(\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N}\hat{a}_n(\epsilon + \hat{\xi}_n - y_n + t_n). \end{aligned} \tag{7.15}$$

Setting the derivatives of L with respect to $\mathbf{w}$, b, ξ_n and $\hat{\xi}_n$ to zero gives

$$\begin{aligned} \mathbf{0} &= \mathbf{w} - \sum_{n=1}^{N}a_n\phi(\mathbf{x}_n) + \sum_{n=1}^{N}\hat{a}_n\phi(\mathbf{x}_n), \\ 0 &= -\sum_{n=1}^{N}a_n + \sum_{n=1}^{N}\hat{a}_n, \\ 0 &= C - \mu_n - a_n, \\ 0 &= C - \hat{\mu}_n - \hat{a}_n. \end{aligned} \tag{7.16}$$

Then, L can be written as

$$\begin{aligned} & \sum_{n=1}^{N}\left((\mu_n + a_n)\xi_n + (\hat{\mu}_n + \hat{a}_n)\hat{\xi}_n\right) + \frac{1}{2}\left\|\sum_{n=1}^{N}(\hat{a}_n - a_n)\phi(\mathbf{x}_n)\right\|^2 \\ & - \sum_{n=1}^{N}\left(\mu_n\xi_n + \hat{\mu}_n\hat{\xi}_n\right) - \sum_{n=1}^{N}a_n(\epsilon + \xi_n + y_n - t_n) \\ & - \sum_{n=1}^{N}\hat{a}_n(\epsilon + \hat{\xi}_n - y_n + t_n). \end{aligned} \tag{7.17}$$

Therefore, minimising L is equilvalent to maximising

$$
\tilde{L} = -\frac{1}{2} \sum_{n=1}^{N} \sum_{n'=1}^{N} (\hat{a}_n - a_n)(\hat{a}_{n'} - a_{n'}) k(\mathbf{x}_n, \mathbf{x}_{n'})
$$
$$
+ \epsilon \sum_{n=1}^{N} (a_n + \hat{a}_n) + \sum_{n=1}^{N} (a_n - \hat{a}_n)(y_n - t_n),
$$

(7.18)

where

$$
k(\mathbf{x}_n, \mathbf{x}_{n'}) = \phi(\mathbf{x}_n)^{\mathsf{T}} \phi(\mathbf{x}_{n'}).
$$

(7.19)

7.8   (Incomplete)

Let $t_1, \cdots, t_N$ be vaariables. Let

$$
E = C \sum_{n=1}^{N} E_\epsilon(y_n - t_n) + \frac{1}{2} \|\mathbf{w}\|^2,
$$

(7.20)

where $\epsilon > 0$ and

$$
E_\epsilon(a) = \begin{cases} 0, & |a| < \epsilon, \\ |a| - \epsilon, & \text{otherwise,} \end{cases}
$$

(7.21)

$$
y_n = \mathbf{w}^{\mathsf{T}} \phi(\mathbf{x}_n) + b.
$$

(a)

In order to minimise E under the constraints

$$
t_n \leq y_n + \epsilon + \xi_n,
$$

(7.22)

where $\xi_n \geq 0$, let us minimise

$$
L = C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^{N} \mu_n \xi_n - \sum_{n=1}^{N} a_n(\epsilon + \xi_n + y_n - t_n).
$$

(7.23)

Setting the derivatives of L with respect to $\mathbf{w}$, b and ξ_n to zero gives

$$
\mathbf{0} = \mathbf{w} - \sum_{n=1}^{N} a_n \phi(\mathbf{x}_n),
$$
$$
0 = -\sum_{n=1}^{N} a_n,
$$
$$
0 = C - \mu_n - a_n.
$$

(7.24)

Then, L can be written as

$$
\sum_{n=1}^{N}(\mu_n + a_n)\xi_n + \frac{1}{2}\left\|\sum_{n=1}^{N} a_n\phi(\mathbf{x}_n)\right\|^2 - \sum_{n=1}^{N}\mu_n\xi_n
$$
$$
- \sum_{n=1}^{N} a_n(\epsilon + \xi_n + y_n - t_n). \tag{7.25}
$$

Therefore, minimising L is equilvalent to maximising

$$
\tilde{L} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N} a_n a_{n'} k(\mathbf{x}_n, \mathbf{x}_{n'}) + \sum_{n=1}^{N} a_n(y_n - t_n), \tag{7.26}
$$

where

$$
k(\mathbf{x}_n, \mathbf{x}_{n'}) = \phi(\mathbf{x}_n)^{\mathsf{T}}\phi(\mathbf{x}_{n'}). \tag{7.27}
$$

7.9

Let $t_1, \cdots, t_N$ be variables such that

$$
p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\phi_n, \beta^{-1}\right),
$$
$$
p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}\right). \tag{7.28}
$$

By the Bayes' theorem,

$$
p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{7.29}
$$

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

$$
-\frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{w}^{\mathsf{T}}\phi_n\|^2 - \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w} = -\frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{M}\mathbf{v}, \tag{7.30}
$$

where

$$
\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \mathbf{A} & -\beta\mathbf{\Phi}^{\mathsf{T}} \\ -\beta\mathbf{\Phi} & \beta\mathbf{I} \end{bmatrix}. \tag{7.31}
$$

Therefore,

$$
p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right), \tag{7.32}
$$

where

$$
\mathbf{m} = \beta\mathbf{S}\mathbf{\Phi}^{\mathsf{T}}\mathbf{t},
$$
$$
\mathbf{S} = \left(\beta\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi} + \mathbf{A}\right)^{-1}. \tag{7.33}
$$

7.10

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}\right). \tag{7.34}$$

By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{7.35}$$

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

$$-\frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\|^2 - \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w} = -\frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v}, \tag{7.36}$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}^{-1} & -\beta\boldsymbol{\Phi}^\mathsf{T} \\ -\beta\boldsymbol{\Phi} & \beta\mathbf{I} \end{bmatrix}, \tag{7.37}$$

and

$$\mathbf{S}^{-1} = \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \mathbf{A}. \tag{7.38}$$

The right hand side can be written as

$$-\frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{S}^{-1}\mathbf{u} - \frac{1}{2}\mathbf{t}^\mathsf{T}\mathbf{N}\mathbf{t}, \tag{7.39}$$

where

$$\mathbf{u} = \mathbf{w} - \beta\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\mathbf{t},$$
$$\mathbf{N} = \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}. \tag{7.40}$$

By 2.26,

$$\mathbf{N}^{-1} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T}. \tag{7.41}$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T}\right). \tag{7.42}$$

7.11

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}\right). \tag{7.43}$$

By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{7.44}$$

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

$$-\frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\|^2 - \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w} = -\frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v}, \tag{7.45}$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ \mathbf{t} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \mathbf{A} & -\beta\boldsymbol{\Phi}^\mathsf{T} \\ -\beta\boldsymbol{\Phi} & \beta\mathbf{I} \end{bmatrix}. \tag{7.46}$$

By 2.24,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T} \\ \boldsymbol{\Phi}\mathbf{A}^{-1} & \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T} \end{bmatrix}. \tag{7.47}$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\mathsf{T}\right). \tag{7.48}$$

7.12

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}) = \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}\right), \tag{7.49}$$

where

$$\mathbf{A} = \mathrm{diag}(\alpha_m). \tag{7.50}$$

By marginalisation,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{7.51}$$

231

The logarithm of the integrand of the right hand side can be written as

$$-\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\|^2$$

$$-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\mathbf{A}^{-1}\right) - \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w} \tag{7.52}$$

$$= -\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{M}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{A}) - E(\mathbf{w}),$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}. \tag{7.53}$$

Then,

$$p(\mathbf{t}) = \left(2\pi\beta^{-1}\right)^{-\frac{N}{2}}(2\pi)^{-\frac{M}{2}}(\det\mathbf{A})^{\frac{1}{2}}\int\exp\left(-E(\mathbf{w})\right)d\mathbf{w}. \tag{7.54}$$

We have

$$\nabla E(\mathbf{w}) = \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}),$$
$$\nabla\nabla E(\mathbf{w}) = \boldsymbol{\Sigma}^{-1}. \tag{7.55}$$

where

$$\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^\mathsf{T}\mathbf{t},$$
$$\boldsymbol{\Sigma}^{-1} = \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \alpha\mathbf{I}. \tag{7.56}$$

Then,

$$E(\mathbf{w}) = E(\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}), \tag{7.57}$$

so that

$$\int\exp\left(-E(\mathbf{w})\right)d\mathbf{w} = (2\pi)^{\frac{M}{2}}(\det\boldsymbol{\Sigma})^{\frac{1}{2}}\exp\left(-E(\boldsymbol{\mu})\right). \tag{7.58}$$

Then,

$$\ln p(\mathbf{t}) = -\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) + \frac{1}{2}\ln(\det\mathbf{A}) + \frac{1}{2}\ln(\det\boldsymbol{\Sigma}) - E(\boldsymbol{\mu}). \tag{7.59}$$

By 3.21, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α_m and β to zero gives

$$0 = \frac{1}{2\alpha_m} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\alpha_m}\right) - \frac{\mu_m^2}{2},$$

$$0 = \frac{N}{2\beta} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\beta}\right) - \frac{\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{2}. \tag{7.60}$$

We have

$$\mathbf{O} = \frac{\partial \mathbf{\Sigma}^{-1} \mathbf{\Sigma}}{\partial \alpha_m},$$
$$\mathbf{O} = \frac{\partial \mathbf{\Sigma}^{-1} \mathbf{\Sigma}}{\partial \beta}. \tag{7.61}$$

The right hand side can be written as

$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m} + \mathbf{\Sigma} \frac{\partial \mathbf{\Sigma}^{-1}}{\partial \alpha_m} = \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m} + \mathbf{\Sigma},$$
$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta} + \mathbf{\Sigma} \frac{\partial \mathbf{\Sigma}^{-1}}{\partial \beta} = \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta} + \mathbf{\Sigma}\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}. \tag{7.62}$$

Then,

$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m} = -\mathbf{\Sigma},$$
$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta} = -\mathbf{\Sigma}\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi}. \tag{7.63}$$

Therefore, the maximum likelihood solutions of $p(\mathbf{t})$ satisfy

$$\alpha_m = \frac{1}{\mu_m^2 + \operatorname{tr}\mathbf{\Sigma}},$$
$$\beta = \frac{N}{\|\mathbf{t} - \mathbf{\Phi}\boldsymbol{\mu}\|^2 + \operatorname{tr}(\mathbf{\Sigma}\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi})}. \tag{7.64}$$

7.13   (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$p(t_n|\mathbf{w}, \beta) = \mathcal{N}\left(t_n|\mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}_n, \beta^{-1}\right),$$
$$p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}\right),$$
$$p(\beta) = \operatorname{Gam}(\beta|a_0, b_0), \tag{7.65}$$

where

$$\mathbf{A} = \operatorname{diag}(\alpha_m). \tag{7.66}$$

By marginalisation,

$$p(\mathbf{t}) = \iint p(\mathbf{t}|\mathbf{w}, \beta) p(\mathbf{w}) p(\beta) d\mathbf{w} d\beta. \tag{7.67}$$

The logarithm of the integrand of the right hand side can be written as

$$-\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n\|^2$$
$$-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\mathbf{A}^{-1}\right) - \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}$$
$$+ a_0\ln b_0 - \ln\Gamma(a_0) + (a_0 - 1)\ln\beta - b_0\beta \qquad (7.68)$$
$$= -\frac{N+M}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{A}) + a_0\ln b_0 - \ln\Gamma(a_0)$$
$$- E(\mathbf{w}, \beta) + G(\beta),$$

where

$$E(\mathbf{w}, \beta) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w},$$
$$G(\beta) = (a - 1)\ln\beta - b_0\beta, \qquad (7.69)$$

and

$$a = a_0 + \frac{N}{2}. \qquad (7.70)$$

Then,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N+M}{2}} (\det\mathbf{A})^{\frac{1}{2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \int I(\beta)\exp\left(-G(\beta)\right)d\beta, \qquad (7.71)$$

where

$$I(\beta) = \int \exp\left(-E(\mathbf{w}, \beta)\right)d\mathbf{w}. \qquad (7.72)$$

We have

$$\nabla_\mathbf{w}E(\mathbf{w}, \beta) = \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}),$$
$$\nabla_\mathbf{w}\nabla_\mathbf{w}E(\mathbf{w}, \beta) = \boldsymbol{\Sigma}^{-1}. \qquad (7.73)$$

where

$$\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^\mathsf{T}\mathbf{t},$$
$$\boldsymbol{\Sigma}^{-1} = \beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \alpha\mathbf{I}. \qquad (7.74)$$

Then,

$$E(\mathbf{w}, \beta) = E(\boldsymbol{\mu}, \beta) + \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu}), \qquad (7.75)$$

so that

$$I(\beta) = (2\pi)^{\frac{M}{2}} (\det\boldsymbol{\Sigma})^{\frac{1}{2}} \exp\left(-E(\boldsymbol{\mu}, \beta)\right). \qquad (7.76)$$

Then,

$$p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} (\det \mathbf{A})^{\frac{1}{2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \int (\det \mathbf{\Sigma})^{\frac{1}{2}} \exp\left(-E(\boldsymbol{\mu}, \beta) + G(\beta)\right) d\beta. \quad (7.77)$$

The integral can be written as

$$\int (\det \mathbf{\Sigma})^{\frac{1}{2}} \exp\left(-E(\boldsymbol{\mu}, \beta) + G(\beta)\right) d\beta = . \quad (7.78)$$

Then,

$$\ln p(\mathbf{t}) = . \quad (7.79)$$

By 3.21, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α_m and β to zero gives

$$0 = \frac{1}{2\alpha_m} + \frac{1}{2} \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m}\right) - \frac{\mu_m^2}{2},$$
$$0 = \frac{N}{2\beta} + \frac{1}{2} \operatorname{tr}\left(\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta}\right) - \frac{\|\mathbf{t} - \mathbf{\Phi}\boldsymbol{\mu}\|^2}{2}. \quad (7.80)$$

We have

$$\mathbf{O} = \frac{\partial \mathbf{\Sigma}^{-1} \mathbf{\Sigma}}{\partial \alpha_m},$$
$$\mathbf{O} = \frac{\partial \mathbf{\Sigma}^{-1} \mathbf{\Sigma}}{\partial \beta}. \quad (7.81)$$

The right hand side can be written as

$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m} + \mathbf{\Sigma} \frac{\partial \mathbf{\Sigma}^{-1}}{\partial \alpha_m} = \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m} + \mathbf{\Sigma},$$
$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta} + \mathbf{\Sigma} \frac{\partial \mathbf{\Sigma}^{-1}}{\partial \beta} = \mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta} + \mathbf{\Sigma}\mathbf{\Phi}^{\intercal}\mathbf{\Phi}. \quad (7.82)$$

Then,

$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \alpha_m} = -\mathbf{\Sigma},$$
$$\mathbf{\Sigma}^{-1} \frac{\partial \mathbf{\Sigma}}{\partial \beta} = -\mathbf{\Sigma}\mathbf{\Phi}^{\intercal}\mathbf{\Phi}. \quad (7.83)$$

Therefore, the maximum likelihood solutions of $p(\mathbf{t})$ satisfy

$$\alpha_m = \frac{1}{\mu_m^2 + \operatorname{tr}\mathbf{\Sigma}},$$
$$\beta = \frac{N}{\|\mathbf{t} - \mathbf{\Phi}\boldsymbol{\mu}\|^2 + \operatorname{tr}(\mathbf{\Sigma}\mathbf{\Phi}^{\intercal}\mathbf{\Phi})}. \quad (7.84)$$

7.14

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\mathsf{T}\boldsymbol{\phi}_n, \beta^{-1}\right), \\
p(\mathbf{w}) &= \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}\right).
\end{aligned} \tag{7.85}
$$

By marginalisation,

$$
p(t_{N+1}|\mathbf{t}) = \int p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{7.86}
$$

By 7.9,

$$
p(\mathbf{w}|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}, \mathbf{S}\right), \tag{7.87}
$$

where

$$
\begin{aligned}
\mathbf{m} &= \beta\mathbf{S}\boldsymbol{\Phi}^\mathsf{T}\mathbf{t}, \\
\mathbf{S} &= \left(\beta\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} + \mathbf{A}\right)^{-1}.
\end{aligned} \tag{7.88}
$$

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

$$
\begin{aligned}
&-\frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}_{N+1}\right)^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m})^\mathsf{T}\mathbf{S}^{-1}(\mathbf{w} - \mathbf{m}) \\
&= -\frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{v} + \mathbf{v}^\mathsf{T}\mathbf{u} - \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m},
\end{aligned} \tag{7.89}
$$

where

$$
\mathbf{v} = \begin{bmatrix} \mathbf{w} \\ t_{N+1} \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{S}^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^\mathsf{T} & -\beta\boldsymbol{\phi}_{N+1} \\ -\beta\boldsymbol{\phi}_{N+1}^\mathsf{T} & \beta \end{bmatrix}, \mathbf{u} = \begin{bmatrix} \mathbf{S}^{-1}\mathbf{m} \\ 0 \end{bmatrix}. \tag{7.90}
$$

By 2.24,

$$
\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{S} & \mathbf{S}\boldsymbol{\phi}_{N+1} \\ \boldsymbol{\phi}_{N+1}^\mathsf{T}\mathbf{S} & \beta^{-1} + \boldsymbol{\phi}_{N+1}^\mathsf{T}\mathbf{S}\boldsymbol{\phi}_{N+1} \end{bmatrix}, \tag{7.91}
$$

so that

$$
\mathbf{M}^{-1}\mathbf{u} = \begin{bmatrix} \mathbf{m} \\ \mathbf{m}^\mathsf{T}\boldsymbol{\phi}_{N+1} \end{bmatrix}. \tag{7.92}
$$

Therefore,

$$
p(t_{N+1}|\mathbf{t}) = \mathcal{N}\left(t_{N+1}|\mathbf{m}^\mathsf{T}\boldsymbol{\phi}_{N+1}, \beta^{-1} + \boldsymbol{\phi}_{N+1}^\mathsf{T}\mathbf{S}\boldsymbol{\phi}_{N+1}\right). \tag{7.93}
$$