Solutions Manual to
Pattern Recognition and Machine Learning

Hiromichi Inawashiro

April 20, 2025

Contents

1 Introduction

1.1

Let

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

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

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

If

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

then

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

Therefore,

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

Thus,

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

where

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

If

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

then

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

1.2

Let

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

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

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

If

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

then

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

Therefore,

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

Thus,

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

where

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

If

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

then

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

1.3

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

(a)

The probability that an apple is selected is given by

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

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

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

(b)

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

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

Here,

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

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

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

Therefore,

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

1.4

Let

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

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

(a)

Let us assume that there exists $\epsilon > 0$ such that $g'(y) \neq 0$ for $|y - \hat{y}| < \epsilon$. Then, Taking the derivative of the transoformation

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

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

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

Therefore, in general,

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

(b)

Let us assume that

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

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

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

Therefore,

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

1.5

By the definition,

$$\operatorname{var} f(x) = \mathrm{E}\left(f(x) - \mathrm{E}\, f(x)\right)^2 . \tag{1.30}$$

The right hand side can be written as

$$\mathrm{E}\left((f(x))^2 - 2f(x)\,\mathrm{E}\, f(x) + (\mathrm{E}\, f(x))^2\right) = \mathrm{E}\left(f(x)\right)^2 - \left(\mathrm{E}\, f(x)\right)^2 . \tag{1.31}$$

Therefore,

$$\operatorname{var} f(x) = \mathrm{E}\left(f(x)\right)^2 - \left(\mathrm{E}\, f(x)\right)^2 . \tag{1.32}$$

1.6

By the definition,

$$\operatorname{cov}(x, y) = \mathrm{E}\left((x - \mathrm{E}\, x)\left(y - \mathrm{E}\, y\right)\right). \tag{1.33}$$

The right hand side can be written as

$$\mathrm{E}\, xy - \mathrm{E}\left(x\,\mathrm{E}\, y\right) - \mathrm{E}\left(y\,\mathrm{E}\, x\right) + \mathrm{E}\left(\mathrm{E}\, x\,\mathrm{E}\, y\right) = \mathrm{E}\, xy - \mathrm{E}\, x\,\mathrm{E}\, y. \tag{1.34}$$

The right hand side can be written as

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

If x and y are independent, by the definition,

$$f(x,y) = f(x)f(y). \tag{1.36}$$

Then,

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

Therefore,

$$\operatorname{cov}(x, y) = 0. \tag{1.38}$$

1.7

(a)

Let

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

Then,

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

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

$$\int_{0}^{\infty}\int_{0}^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix}\cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta\end{vmatrix} drd\theta = 2\pi \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdr. \tag{1.41}$$

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

$$2\pi\sigma^2 \int_{0}^{\infty} \exp\left(-\frac{1}{2}s^2\right) sds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_{0}^{\infty}. \tag{1.42}$$

Therefore,

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

(b)

By the definition,

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

Then,

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

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

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

Therefore,

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

1.8

(a)

Let x be a variable such that

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

Then

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

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

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

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

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

Since

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

and

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

we have

$$\mathrm{E}\, x = \mu. \tag{1.54}$$

(b)

The property

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

can be written as

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

Taking the derivative with respect to σ^2 gives

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

The left hand side can be written as

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

Therefore,

$$\operatorname{var}x=\sigma^2. \tag{1.59}$$

1.9

(a)

Let x be a variable such that

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

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

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

Therefore,

$$\operatorname{mode}x=\mu. \tag{1.62}$$

(b)

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

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

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

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

Therefore,

$$\operatorname{mode}\mathbf{x} = \boldsymbol{\mu}. \tag{1.65}$$

1.10

(a)

By the definition,

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

The right hand side can be written as

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

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

$$\mathrm{E}\,x + \mathrm{E}\,y. \tag{1.68}$$

Therefore,

$$\mathrm{E}(x+y) = \mathrm{E}\,x + \mathrm{E}\,y. \tag{1.69}$$

(b)

By the definition,

$$\operatorname{var}(x+y) = \mathrm{E}\left(x+y-\mathrm{E}(x+y)\right)^2 \tag{1.70}$$

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

$$\begin{aligned} &\mathrm{E}\left(x-\mathrm{E}\,x\right)^2 + 2\,\mathrm{E}\left(\left(x-\mathrm{E}\,x\right)\left(y-\mathrm{E}\,y\right)\right) + \mathrm{E}\left(y-\mathrm{E}\,y\right)^2 \\ &= \operatorname{var}x + 2\operatorname{cov}(x,y) + \operatorname{var}y. \end{aligned} \tag{1.71}$$

By 1.6, if x and y are independent, then

$$\operatorname{cov}(x,y) = 0. \tag{1.72}$$

Therefore,

$$\operatorname{var}(x+y) = \operatorname{var}x + \operatorname{var}y. \tag{1.73}$$

1.11

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

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

Then,

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

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

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

Therefore,

$$\begin{aligned} \mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}x_n, \\ \sigma^2_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2. \end{aligned} \tag{1.77}$$

1.12

(a)

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

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

Then,

$$\mathrm{E}\, x_n x_{n'} = \mu^2. \tag{1.79}$$

By the property

$$\mathrm{E}\, x_n^2 = \operatorname{var} x_n + \left(\mathrm{E}\, x_n\right)^2, \tag{1.80}$$

we have

$$\mathrm{E}\, x_n^2 = \sigma^2 + \mu^2. \tag{1.81}$$

Therefore,

$$\mathrm{E}\, x_n x_{n'} = \mu^2 + I_{nn'}\sigma^2. \tag{1.82}$$

(b)

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

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

Then, by 1.11,

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{1.84}$$

Then,

$$\mathrm{E}\,\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\,x_n. \tag{1.85}$$

Therefore,

$$\mathrm{E}\,\mu_{\mathrm{ML}} = \mu. \tag{1.86}$$

Similarly, by 1.11,

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \mu_{\mathrm{ML}}\right)^2. \tag{1.87}$$

Then,

$$\mathrm{E}\,\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left(x_n - \mu_{\mathrm{ML}}\right)^2. \tag{1.88}$$

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left(x_n^2 - 2\mu_{\mathrm{ML}}x_n + \mu^2_{\mathrm{ML}}\right) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\,x_n^2 - \frac{2}{N}\,\mathrm{E}\left(\mu_{\mathrm{ML}}\left(\sum_{n=1}^{N} x_n\right)\right) + \mathrm{E}\,\mu^2_{\mathrm{ML}}. \tag{1.89}$$

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

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

while the second and third terms can be writen as

$$-2\,\mathrm{E}\,\mu^2_{\mathrm{ML}} + \mathrm{E}\,\mu^2_{\mathrm{ML}} = -\,\mathrm{E}\,\mu^2_{\mathrm{ML}}. \tag{1.91}$$

Here,

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

The right hand side can be written as

$$\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\,x_n^2 + \frac{2}{N^2}\sum_{1\le n<n'\le N}\mathrm{E}\,x_n x_{n'} = \frac{1}{N}\left(\mu^2+\sigma^2\right) + \frac{N-1}{N}\mu^2. \tag{1.93}$$

Therefore,

$$\mathrm{E}\,\mu_{\mathrm{ML}}^2 = \mu^2 + \frac{1}{N}\sigma^2. \tag{1.94}$$

Thus,

$$\mathrm{E}\,\sigma_{\mathrm{ML}}^2 = \frac{N-1}{N}\sigma^2. \tag{1.95}$$

1.13

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

$$\begin{aligned}\mathrm{E}\,x_n &= \mu,\\ \operatorname{var} x_n &= \sigma^2.\end{aligned} \tag{1.96}$$

Here,

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

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N}\operatorname{var} x_n = \sigma^2. \tag{1.98}$$

Therefore,

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

1.14

(a)

Let

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

Then,

$$\begin{aligned} w_{dd'} &= w^{\mathrm{S}}_{dd'} + w^{\mathrm{A}}_{dd'}, \\ w^{\mathrm{S}}_{dd'} &= w^{\mathrm{S}}_{d'd}, \\ w^{\mathrm{A}}_{dd'} &= -w^{\mathrm{A}}_{d'd}. \end{aligned} \tag{1.101}$$

(b)

Here,

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

The right hand side can be written as

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

Therefore,

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w^{\mathrm{A}}_{dd'} x_d x_{d'} = 0. \tag{1.104}$$

(c)

We have

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'} x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} \left(w^{\mathrm{S}}_{dd'} + w^{\mathrm{A}}_{dd'}\right) x_d x_{d'}. \tag{1.105}$$

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

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w^{\mathrm{S}}_{dd'} x_d x_{d'} + \sum_{d=1}^{D}\sum_{d'=1}^{D} w^{\mathrm{A}}_{dd'} x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} w^{\mathrm{S}}_{dd'} x_d x_{d'}, \tag{1.106}$$

Therefore,

$$\sum_{d=1}^{D}\sum_{d'=1}^{D} w_{dd'}x_d x_{d'} = \sum_{d=1}^{D}\sum_{d'=1}^{D} w^{\mathrm{S}}_{dd'}x_d x_{d'}. \tag{1.107}$$

(d)

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

$$n(1, M) = n(1, M-1) = 1. \tag{1.108}$$

Let us assume that

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

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

$$n(D+1, M) = n(D, M) + n(D+1, M-1). \tag{1.110}$$

Thus,

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

Hence, the assumption is proved by induction on D.

(b)

Note that

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

Let us assume that

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

Then

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

The right hand side can be written as

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

Therefore, the assumption is proved by induction on D.

(c)

By 1.14(d),

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

Let us assume that

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

Then, by (a),

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

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

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

Therefore, the assumption is proved by induction on M.

1.16

(a)

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

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

where

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

(b)

By (a),

$$N(D, 0) = 1. \tag{1.122}$$

Let us assume that

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

Then,

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

The right hand side can be written as

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

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

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

(c)

By the approximation

$$n! \simeq n^n \exp(-n), \tag{1.127}$$

we have

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

The right hand side can be written as

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

Therefore,

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

(d)

By (b),

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

1.17

(a)

Let

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

Then

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

The right hand side can be written as

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

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.135}$$

(b)

Since

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

we have

$$\Gamma(1) = 0!. \tag{1.137}$$

For a positive integer x, let us assume that

$$\Gamma(x) = (x-1)!. \tag{1.138}$$

Then,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.139}$$

Therefore,

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

Thus, the assumption is proved by induction on x.

1.18

(a)

Let us consider the transformation from Cartesian to polar coordinates

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

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

$$s = r^2, \tag{1.142}$$

the right hand side can be written as

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

Therefore,

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

(b)

The volume of the sphere can can be written as

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

Therefore,

$$V_D = \frac{S_D}{D}. \tag{1.146}$$

(c)

By (a) and (b),

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

Similarly,

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

1.19

(a)

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

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

(b)

By the Stering's formula

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

we have

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

The right hand side can be written as

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

Therefore,

$$\lim_{D\to\infty}\frac{V_D}{2^D}=0. \tag{1.153}$$

(c)

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

$$\frac{\sqrt{\sum_{i=1}^{D}1^2}}{1}=\sqrt{D}. \tag{1.154}$$

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

1.20

(a)

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

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

Then

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

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

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

Therefore,

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

where

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

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

(b)

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

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

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

$$\hat{r} = \sqrt{D-1}\sigma. \tag{1.161}$$

(c)

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

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

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

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

By the Taylor series

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

the right hand side can be approximated as

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

Therefore,

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

(d)

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

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

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

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

1.21

(a)

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

$$0 \leq a(b-a). \tag{1.169}$$

Therefore,

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

(b)

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

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

Then, by (a),

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

Therefore,

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

1.22

Let

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

If

$$L_{kj} = 1 - I_{kj}, \tag{1.175}$$

then the right hand side can be written as

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

The right hand side can be written as

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

Therefore,

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

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

1.23

Let

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

The right hand side can be written as

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

Therefore,

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

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

1.24 (Incomplete)

Let

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

1.25

Let

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

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

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

The integral of the right hand side can be written as

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

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

$$\mathbf{0} = p(\mathbf{x}) \left(\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\right). \tag{1.186}$$

Thus,

$$\operatorname*{argmin}_{\mathbf{y}(\mathbf{x})} \mathrm{E}\,L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.187}$$

For a single target variable t, it reduces to

$$\operatorname*{argmin}_{\mathbf{y}(\mathbf{x})} \mathrm{E}\,L(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \mathrm{E}_t(t|\mathbf{x}). \tag{1.188}$$

1.26

Let

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

The right hand side can be written as

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

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

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

The integral of the second term can be written as

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

Since

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

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

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

Therefore,

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

Thus,

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

1.27 (Incomplete)

(a)

Let

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

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

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

Therefore,

$$\operatorname*{argmin}_{y(\mathbf{x})} \mathrm{E}\, L_q = \left\{ y(\mathbf{x}) \mid \int |y(\mathbf{x}) - t|^{q-1} \operatorname{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt = 0 \right\}. \tag{1.199}$$

(b)

We have

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

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

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

Therefore,

$$\operatorname*{argmin}_{y(\mathbf{x})} \mathrm{E}\, L_1 = \operatorname{median}(t|\mathbf{x}). \tag{1.202}$$

(c)

We have

$$\lim_{q \to 0} \left(\operatorname*{argmin}_{y(\mathbf{x})} \mathrm{E}\, L_q \right) = \operatorname{mode}(t|\mathbf{x})? \tag{1.203}$$

1.28

(a)

Let us assume that

$$p(x,y) = p(x)p(y) \Rightarrow h(x,y) = h(x) + h(y). \tag{1.204}$$

Then,

$$h\left(p^2\right) = 2h(p). \tag{1.205}$$

Let us assume that, for a positive integer n,

$$h\left(p^n\right) = nh(p). \tag{1.206}$$

Then, by the first assumption,

$$h\left(p^{n+1}\right) = h\left(p^n\right) + h(p), \tag{1.207}$$

so that

$$h\left(p^{n+1}\right) = (n+1)h(p). \tag{1.208}$$

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

(b)

For positive integers m and n,

$$h\left(p^n\right) = h\left(p^{\frac{n}{m}m}\right). \tag{1.209}$$

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

$$h\left(p^{\frac{n}{m}}\right) = \frac{n}{m}h(p). \tag{1.210}$$

(c)

By the continuity, for a positive real number a,

$$h\left(p^a\right) = ah(p). \tag{1.211}$$

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

$$(p \ln p)h'(p) = h(p). \tag{1.212}$$

Then,

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

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

$$h(p) \propto \ln p. \tag{1.214}$$

1.29

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

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

where

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

By the Jensen's inequality,

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

Therefore,

$$\mathrm{H}(x) \leq \ln M. \tag{1.218}$$

1.30

Let

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

By the definition,

$$\mathrm{KL}\left(p||q\right) = -\int p(x) \ln \frac{q(x)}{p(x)} dx. \tag{1.220}$$

The right hand side can be written as

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

The right hand side can be written as

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

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

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

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

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

1.31

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

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

Note that

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

Therefore,

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

Since

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

The Jensen's inequality can be used to write that

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

The right hand side can be written as

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

Therefore,

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

1.32

Let $\mathbf{x}$ be a vector and let

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1.232}$$

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

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

we have

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

so that

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

By the definition,

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

Then, the right hand side can be written as

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

By the transformation

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}, \tag{1.238}$$

the first term can be written as

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

Since

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

we have

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

Then, the first term can be written as

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

Therefore,

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

1.33

Let x and y be two discrete variables with K and L states. By the definition,

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

If $\mathrm{H}(y|x)$ is zero, then

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

Since

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

the equation reduces to

$$p(y_l|x_k) \ln p(y_l|x_k) = 0. \tag{1.247}$$

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

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

we have

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

1.34

Let x be a variable. By the definition,

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

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

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

let

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

Setting the variation with respect to p to zero gives

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

Therefore,

$$p(x) = \exp\left(-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right), \tag{1.254}$$

so that

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

where

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

Substituting it to the constraints gives

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

By the transformation

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

they can be written as

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

Since

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

they can be written as

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

Then,

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

Therefore,

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

1.35

Let x be a variable such that

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

Then, by the definition,

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

The right hand side can be written as

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

Therefore,

$$
\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln\left(2\pi\sigma^2\right)\right). \tag{1.267}
$$

1.36 (Incomplete)

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

$$f\left(\lambda a + (1-\lambda)b\right) \leq \lambda f(a) + (1-\lambda)f(b), \tag{1.268}$$

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

$$x = \lambda a + (1-\lambda)b. \tag{1.269}$$

Then, the inequality can be written as

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

Let

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

Then,

$$g(x) \geq 0. \tag{1.272}$$

Additionally, for $x > a$,

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

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

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

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

$$f'(y) \leq f'(c). \tag{1.275}$$

1.37

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

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

The right hand side can be written as

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

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

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

1.38

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

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

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

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

where $\lambda_m\geq 0$ and

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

Here, let $\lambda_m\geq 0$ and

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

Then, by the definition,

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

By the assumption,

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

Then,

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

Then,

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

Therefore, the assumption is proved by induction on M.

1.39

Let x and y be two binary variables where

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

(a)

By the definition,

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

By the distribution,

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

Therefore,

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

(b)

By the definition,

$$\mathrm{H}(y) = -\sum p(y) \ln p(y). \tag{1.291}$$

By the distribution,

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

Therefore,

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

(c)

By the definition,

$$\mathrm{H}(y|x) = -\sum p(x, y) \ln p(y|x). \tag{1.294}$$

By the definition,

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

Then, by the distribution,

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

Therefore,

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

(d)

By the definition,

$$\mathrm{H}(x|y) = -\sum p(x,y) \ln p(x|y). \tag{1.298}$$

By the definition,

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

Then, by the distribution,

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

Therefore,

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

(e)

By the definition,

$$\mathrm{H}(x,y) = -\sum p(x,y) \ln p(x,y). \tag{1.302}$$

Therefore,

$$\mathrm{H}(x,y) = \ln 3. \tag{1.303}$$

(f)

By the definition,

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

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

$$\mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x,y). \tag{1.305}$$

Therefore,

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

1.40

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

$$\sum_{m=1}^{M} \lambda_m = 1. \tag{1.307}$$

By the Jensen's inequality,

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

so that

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

Substituting

$$\lambda_m = \frac{1}{M} \tag{1.310}$$

to the inequality gives

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

1.41

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

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

The right hand side can be written as

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

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

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

By the definition,

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

Therefore,

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

2 Probability Distributions

2.1

Let x be a variable such that

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

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

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

By the definition,

$$\begin{aligned} \mathrm{E}\, x &= \mu, \\ \mathrm{E}\, x^2 &= \mu, \end{aligned} \tag{2.3}$$

Since

$$\operatorname{var} x = \mathrm{E}\, x^2 - (\mathrm{E}\, x)^2, \tag{2.4}$$

we have

$$\operatorname{var} x = \mu(1-\mu). \tag{2.5}$$

By the definition,

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

Therefore,

$$\mathrm{H}(x) = -\mu \ln \mu - (1-\mu) \ln(1-\mu). \tag{2.7}$$

2.2

Let x be a variable such that

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

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

$$\sum_x p(x|\mu) = 1. \tag{2.9}$$

By the definition,

$$\begin{aligned} \mathrm{E}\, x &= \mu, \\ \mathrm{E}\, x^2 &= 1, \end{aligned} \tag{2.10}$$

Since

$$\operatorname{var} x = \operatorname{E} x^2 - (\operatorname{E} x)^2 \,, \tag{2.11}$$

we have

$$\operatorname{var} x = 1 - \mu^2. \tag{2.12}$$

By the definition,

$$\operatorname{H}(x) = -\sum_x p(x|\mu) \ln p(x|\mu). \tag{2.13}$$

Therefore,

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

2.3

By the definition,

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

Therefore,

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

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

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

Thus,

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}. \tag{2.18}$$

Note that

$$1 + x = \sum_{m=0}^{1} \binom{1}{m} x^m. \tag{2.19}$$

Let us assume that

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

Then,

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

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

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

Therefore,

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

Thus, the assumption is proved by induction on N.

Finally, let m be a variable such that

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

Then

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

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

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

Therefore,

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

2.4

Let m be a variable such that

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

Then

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

Taking the derivative of

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

with respect to μ gives

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

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

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

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

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

Therefore,

$$\mathrm{E}\, m = N\mu. \tag{2.34}$$

Taking the second derivative of

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

with respect to μ gives

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

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

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

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

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

Therefore,

$$\mathrm{E}\,m(m-1) = N(N-1)\mu^2. \tag{2.39}$$

Thus, since

$$\operatorname{var} m = \mathrm{E}\,m(m-1) + \mathrm{E}\,m - (\mathrm{E}\,m)^2, \tag{2.40}$$

we have

$$\operatorname{var} m = N\mu(1-\mu). \tag{2.41}$$

2.5

By the definition,

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

By the transformation

$$t = x + y, \tag{2.43}$$

the right hand side can be written as

$$\begin{aligned} &\int_0^\infty x^{a-1}\left(\int_x^\infty (t-x)^{b-1}\exp(-t)dt\right)dx \\ &= \int_0^\infty \left(\int_0^t x^{a-1}(t-x)^{b-1}dx\right)\exp(-t)dt. \end{aligned} \tag{2.44}$$

By the transformation

$$x = t\mu, \tag{2.45}$$

the right hand side can be written as

$$\begin{aligned} &\int_0^\infty \left(\int_0^1 (t\mu)^{a-1} t^{b-1} (1-\mu)^{b-1} t d\mu \right) \exp(-t) dt \\ &= \int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu \int_0^\infty t^{a+b-1} \exp(-t) dt. \end{aligned} \tag{2.46}$$

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

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2.47}$$

2.6

Let μ be a variable such that

$$p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}. \tag{2.48}$$

Then

$$\begin{aligned} \mathrm{E}\,\mu &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu, \\ \mathrm{E}\,\mu^2 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu. \end{aligned} \tag{2.49}$$

Since

$$\begin{aligned} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}, \\ \int_0^1 \mu^{a+1} (1-\mu)^{b-1} d\mu &= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}, \end{aligned} \tag{2.50}$$

we have

$$\begin{aligned} \mathrm{E}\,\mu &= \frac{a}{a+b}, \\ \mathrm{E}\,\mu^2 &= \frac{a(a+1)}{(a+b)(a+b+1)}. \end{aligned} \tag{2.51}$$

Since

$$\operatorname{var}\mu = \mathrm{E}\,\mu^2 - (\mathrm{E}\mu)^2, \tag{2.52}$$

we have

$$\operatorname{var}\mu = \frac{ab}{(a+b)^2(a+b+1)}. \tag{2.53}$$

Since

$$\frac{\partial}{\partial\mu}p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}\left(\frac{a-1}{\mu} - \frac{b-1}{1-\mu}\right), \tag{2.54}$$

we have

$$\operatorname{mode}\mu = \frac{a-1}{a+b-2}. \tag{2.55}$$

2.7

Let m and l be a variable such that

$$p(m,l|\mu) = \binom{m+l}{m}\mu^m(1-\mu)^l, \tag{2.56}$$

where

$$p(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}. \tag{2.57}$$

By 2.6,

$$\mathrm{E}(\mu|a,b) = \frac{a}{a+b}. \tag{2.58}$$

Note that

$$\mu_{\mathrm{ML}} = \frac{m}{m+l}. \tag{2.59}$$

Since

$$p(\mu|m,l,a,b) \propto p(m,l|\mu)p(\mu|a,b), \tag{2.60}$$

we have

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}. \tag{2.61}$$

Therefore, by 2.6,

$$\mathrm{E}(\mu|m,l,a,b) = \frac{m+a}{m+l+a+b}. \tag{2.62}$$

Thus,

$$\mathrm{E}(\mu|m,l,a,b) = \lambda\mu_{\mathrm{ML}} + (1-\lambda)\,\mathrm{E}(\mu|a,b), \tag{2.63}$$

where

$$\lambda = \frac{m+l}{m+l+a+b}. \tag{2.64}$$

2.8

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

$$\mathrm{E}\, x = \int x p(x) dx. \tag{2.65}$$

The right hand side can be written as

$$\int x \left(\int p(x, y) dy \right) dx = \int \left(\int x p(x|y) dx \right) p(y) dy. \tag{2.66}$$

Therefore,

$$\mathrm{E}\, x = \mathrm{E}_y \left(\mathrm{E}_x(x|y) \right). \tag{2.67}$$

Additionally, by the definition,

$$\operatorname{var} x = \mathrm{E} \left(x - \mathrm{E}\, x \right)^2. \tag{2.68}$$

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

$$\begin{aligned} &\mathrm{E}_y \left(\mathrm{E}_x \left(\left(x - \mathrm{E}_x(x|y) + \mathrm{E}_x(x|y) - \mathrm{E}\, x \right)^2 |y \right) \right) \\ =& \mathrm{E}_y \left(\mathrm{E}_x \left(\left(x - \mathrm{E}_x(x|y) \right)^2 |y \right) \right) \\ &+ 2\, \mathrm{E}_y \left(\left(\left(\mathrm{E}_x(x|y) - \mathrm{E}\, x \right) \mathrm{E}_x \left(x - \mathrm{E}_x(x|y) \right) |y \right) \right) \\ &+ \mathrm{E}_y \left(\left(\mathrm{E}_x(x|y) - \mathrm{E}\, x \right)^2 |y \right). \end{aligned} \tag{2.69}$$

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

$$2\, \mathrm{E}_y \left(\left(\mathrm{E}_x(x|y) - \mathrm{E}\, x \right) \left(\mathrm{E}_x(x|y) - \mathrm{E}_x(x|y) \right) \right) = 0. \tag{2.70}$$

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

$$\mathrm{E}_y \left(\mathrm{E}_x(x|y) - \mathrm{E}_y \left(\mathrm{E}_x(x|y) \right) \right)^2 = \operatorname{var}_y \left(\mathrm{E}_x(x|y) \right). \tag{2.71}$$

Therefore,

$$\operatorname{var} x = \mathrm{E}_y \left(\operatorname{var}_x(x|y) \right) + \operatorname{var}_y \left(\mathrm{E}_x(x|y) \right). \tag{2.72}$$

2.9

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

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

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

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

Under the constraint

$$\sum_{k=1}^{M+1}\mu_k = 1, \tag{2.73}$$

we have

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

By the transformation

$$\mu'_{M+1} = \frac{\mu_{M+1}}{1-\sum_{k=1}^{M-1}\mu_k}, \tag{2.75}$$

the integral of the right hand side can be written as

$$\begin{aligned}
&\int_0^1 \left(\left(1-\sum_{k=1}^{M-1}\mu_k\right)\mu'_{M+1}\right)^{\alpha_{M+1}-1} \left(\left(1-\sum_{k=1}^{M-1}\mu_k\right)(1-\mu'_{M+1})\right)^{\alpha_M-1} \left(1-\sum_{k=1}^{M-1}\mu_k\right)d\mu'_{M+1} \\
&= \left(1-\sum_{k=1}^{M-1}\mu_k\right)^{\alpha_M+\alpha_{M+1}-1} \int_0^1 \mu'_{M+1}{}^{\alpha_{M+1}-1}(1-\mu'_{M+1})^{\alpha_M-1}d\mu'_{M+1}.
\end{aligned} \tag{2.76}$$

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

$$\frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M+\alpha_{M+1})}. \tag{2.77}$$

Therefore,

$$\int_0^{1-\sum_{k=1}^{M-1}\mu_k} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1} d\mu_{M+1} = \left(\prod_{k=1}^{M-1} {\mu_k}^{\alpha_k-1}\right)\left(1-\sum_{k=1}^{M-1}\mu_k\right)^{\alpha_M+\alpha_{M+1}-1} \frac{\Gamma(\alpha_M)\Gamma(\alpha_{M+1})}{\Gamma(\alpha_M+\alpha_{M+1})}. \tag{2.78}$$

By the assumption,

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

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

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

The right hand side can be written as

$$\frac{\prod_{k=1}^{M+1}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{M+1}\alpha_k)}. \tag{2.79}$$

Hence, the assumption is proved by induction on M.

2.10

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

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k-1}. \tag{2.80}$$

Then

$$\begin{aligned} \mathrm{E}\,\mu_j &= \int \mu_j p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \\ \mathrm{E}\,\mu_j^2 &= \int \mu_j^2 p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \\ \mathrm{E}\,\mu_j\mu_l &= \int \mu_j\mu_l p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}. \end{aligned} \tag{2.81}$$

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

$$\begin{aligned}
\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k + 1)} &= \frac{\alpha_j}{\sum_{k=1}^K \alpha_k}, \\
\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+2)}{\Gamma(\alpha_j)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + 2\right)} &= \frac{\alpha_j(\alpha_j+1)}{\sum_{k=1}^K \alpha_k(\sum_{k=1}^K \alpha_k + 1)}, \\
\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+1)\Gamma(\alpha_l+1)}{\Gamma(\alpha_j)\Gamma(\alpha_l)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + 2\right)} &= \frac{\alpha_j \alpha_l}{\sum_{k=1}^K \alpha_k(\sum_{k=1}^K \alpha_k + 1)}.
\end{aligned} \tag{2.82}$$

Therefore,

$$\begin{aligned}
\mathrm{E}\,\mu_j &= \frac{\alpha_j}{\sum_{k=1}^K \alpha_k}. \\
\mathrm{E}\,\mu_j^2 &= \frac{\alpha_j(\alpha_j+1)}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1\right)}, \\
\mathrm{E}\,\mu_j \mu_l &= \frac{\alpha_j \alpha_l}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1\right)}.
\end{aligned} \tag{2.83}$$

Since

$$\begin{aligned}
\operatorname{var} \mu_j &= \mathrm{E}\,\mu_j^2 - (\mathrm{E}\,\mu_j)^2, \\
\operatorname{cov}(\mu_j, \mu_l) &= \mathrm{E}\,\mu_j\mu_l - \mathrm{E}\,\mu_j\,\mathrm{E}\,\mu_l,
\end{aligned} \tag{2.84}$$

we have

$$\begin{aligned}
\operatorname{var} \mu_j &= \frac{\alpha_j(\sum_{k=1}^K \alpha_k - \alpha_j)}{\sum_{k=1}^K \alpha_k)^2(\sum_{k=1}^K \alpha_k + 1)}, \\
\operatorname{cov}(\mu_j, \mu_l) &= -\frac{\alpha_j \alpha_l}{(\sum_{k=1}^K \alpha_k)^2(\sum_{k=1}^K \alpha_k + 1)}.
\end{aligned} \tag{2.85}$$

2.11

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

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}. \tag{2.86}$$

Then

$$\mathrm{E} \ln \mu_j = \int (\ln \mu_j)\, p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}. \tag{2.87}$$

Since

$$\frac{\partial}{\partial \alpha_j} p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \left(\frac{\Gamma'\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} - \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} + \ln \mu_j \right) p(\boldsymbol{\mu}|\boldsymbol{\alpha}), \tag{2.88}$$

we have

$$\mathrm{E} \ln \mu_j = \frac{\partial}{\partial \alpha_j} \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} + \left(\psi(\alpha_j) - \psi\left(\sum_{k=1}^K \alpha_k \right) \right) \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \tag{2.89}$$

where

$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \tag{2.90}$$

Therefore,

$$\mathrm{E} \ln \mu_j = \psi(\alpha_j) - \psi\left(\sum_{k=1}^K \alpha_k \right). \tag{2.91}$$

2.12

Let x be a variable such that

$$p(x|a, b) = \frac{1}{b-a}, \tag{2.92}$$

where $a < b$. Then

$$\int_a^b p(x|a, b) dx = 1. \tag{2.93}$$

Note that

$$\begin{aligned} \mathrm{E}\, x &= \int_a^b x p(x|a, b) dx, \\ \mathrm{E}\, x^2 &= \int_a^b x^2 p(x|a, b) dx. \end{aligned} \tag{2.94}$$

The right hand sides can be written as

$$\begin{aligned} \frac{1}{b-a} \int_a^b x dx &= \frac{1}{2}(a+b), \\ \frac{1}{b-a} \int_a^b x^2 dx &= \frac{1}{3}\left(a^2 + ab + b^2\right). \end{aligned} \tag{2.95}$$

Therefore,

$$\begin{aligned} \mathrm{E}\, x &= \frac{1}{2}(a+b), \\ \mathrm{E}\, x^2 &= \frac{1}{3}\left(a^2+ab+b^2\right). \end{aligned} \tag{2.96}$$

Since

$$\operatorname{var} x = \mathrm{E}\, x^2 - (\mathrm{E}\, x)^2, \tag{2.97}$$

we have

$$\operatorname{var} x = \frac{1}{12}(b-a)^2. \tag{2.98}$$

2.13

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}). \end{aligned} \tag{2.99}$$

Then, by the definition,

$$\mathrm{KL}(p||q) = -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}. \tag{2.100}$$

Since

$$\ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \frac{(2\pi)^{-\frac{D}{2}} \left(|\det \mathbf{L}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right)}{(2\pi)^{-\frac{D}{2}} \left(|\det \boldsymbol{\Sigma}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}, \tag{2.101}$$

The right hand side can be written as

$$\begin{aligned} &\frac{1}{2} \ln \left| \frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}} \right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &+ \frac{1}{2} \int (\mathbf{x}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &- \frac{1}{2} \int (\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \end{aligned} \tag{2.102}$$

Let us look at each term. Since

$$\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = 1, \tag{2.103}$$

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

$$(\mathbf{x}-\mathbf{m})^\intercal\mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) = (\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\mathbf{m})^\intercal\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\mathbf{m}), \quad (2.104)$$

the second term can be written as

$$\begin{aligned}
&\frac{1}{2}\int(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} \\
&+(\boldsymbol{\mu}-\mathbf{m})^\intercal\mathbf{L}^{-1}\int(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} \\
&+\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^\intercal\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m})\int\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}.
\end{aligned} \quad (2.105)$$

Since

$$\begin{aligned}
\int\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} &= 1, \\
\int\mathbf{x}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} &= \boldsymbol{\mu}, \\
\int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} &= \boldsymbol{\Sigma},
\end{aligned} \quad (2.106)$$

it can be written as

$$\frac{1}{2}\operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right)+\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^\intercal\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}). \quad (2.107)$$

Since

$$\int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} = \boldsymbol{\Sigma}, \quad (2.108)$$

the third term can be written as

$$-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) = -\frac{D}{2}. \quad (2.109)$$

Therefore,

$$\mathrm{KL}(p||q) = \frac{1}{2}\left(\ln\left|\frac{\det\mathbf{L}}{\det\boldsymbol{\Sigma}}\right|+\operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right)+(\boldsymbol{\mu}-\mathbf{m})^\intercal\mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m})-D\right). \quad (2.110)$$

2.14

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

$$L(p(\mathbf{x})) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} + \lambda \left(\int p(\mathbf{x}) d\mathbf{x} - 1 \right) + \mathbf{l}^{\intercal} \left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\mu} \right) + \mathbf{m}^{\intercal} \left(\int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\intercal} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\Sigma} \right) \mathbf{m}. \tag{2.111}$$

Then

$$\frac{\delta L(p(\mathbf{x}))}{\delta p(\mathbf{x})} = -\ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^{\intercal}\mathbf{x} + \mathbf{m}^{\intercal}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\intercal}\mathbf{m}. \tag{2.112}$$

Setting the left hand side to zero gives

$$p(\mathbf{x}) = \exp\left(-1 + \lambda + \mathbf{l}^{\intercal}\mathbf{x} + \mathbf{m}^{\intercal}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\intercal}\mathbf{m}\right), \tag{2.113}$$

so that

$$p(\mathbf{x}) = \exp\left(-1 + \lambda - \mathbf{l}^{\intercal}\mathbf{M}\mathbf{l} + (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})^{\intercal}\, \mathbf{M}^{-1}\, (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})\right), \tag{2.114}$$

where

$$\mathbf{M} = \left(\mathbf{m}\mathbf{m}^{\intercal}\right)^{-1}. \tag{2.115}$$

Substituting it to

$$\begin{aligned} \int p(\mathbf{x}) d\mathbf{x} &= 1, \\ \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} &= \boldsymbol{\mu}, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\intercal} p(\mathbf{x}) d\mathbf{x} &= \boldsymbol{\Sigma}, \end{aligned} \tag{2.116}$$

and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l} \tag{2.117}$$

gives

$$\begin{aligned} \exp\left(-1 + \lambda - \mathbf{l}^{\intercal}\mathbf{M}\mathbf{l}\right) \int \exp\left(-\mathbf{y}^{\intercal}\mathbf{M}^{-1}\mathbf{y}\right) d\mathbf{y} &= 1, \\ \exp\left(-1 + \lambda - \mathbf{l}^{\intercal}\mathbf{M}\mathbf{l}\right) \int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \exp\left(-\mathbf{y}^{\intercal}\mathbf{M}^{-1}\mathbf{y}\right) d\mathbf{y} &= \boldsymbol{\mu}, \\ \exp\left(-1 + \lambda - \mathbf{l}^{\intercal}\mathbf{M}\mathbf{l}\right) \int (\mathbf{y} + \mathbf{M}\mathbf{l})(\mathbf{y} + \mathbf{M}\mathbf{l})^{\intercal} \exp\left(-\mathbf{y}^{\intercal}\mathbf{M}^{-1}\mathbf{y}\right) d\mathbf{y} &= \boldsymbol{\Sigma}. \end{aligned} \tag{2.118}$$

Since

$$\begin{aligned}
\int \exp\left(-\mathbf{y}^\intercal \mathbf{y}\right) d\mathbf{y} &= \left(\Gamma\left(\frac{1}{2}\right)\right)^D, \\
\int \mathbf{y} \exp\left(-\mathbf{y}^\intercal \mathbf{y}\right) d\mathbf{y} &= \mathbf{0}, \\
\int \mathbf{y}\mathbf{y}^\intercal \exp\left(-\mathbf{y}^\intercal \mathbf{y}\right) d\mathbf{y} &= \Gamma\left(\frac{3}{2}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1} \mathbf{I},
\end{aligned} \tag{2.119}$$

they can be written as

$$\begin{aligned}
\exp\left(-1+\lambda-\mathbf{l}^\intercal \mathbf{M}\mathbf{l}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^D (\det \mathbf{M})^{\frac{1}{2}} &= 1, \\
\exp\left(-1+\lambda-\mathbf{l}^\intercal \mathbf{M}\mathbf{l}\right)(\boldsymbol{\mu}+\mathbf{M}\mathbf{l})\left(\Gamma\left(\frac{1}{2}\right)\right)^D (\det \mathbf{M})^{\frac{1}{2}} &= \boldsymbol{\mu}, \\
\exp\left(-1+\lambda-\mathbf{l}^\intercal \mathbf{M}\mathbf{l}\right)\left(\Gamma\left(\frac{3}{2}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1}\mathbf{M} + \mathbf{M}\mathbf{l}(\mathbf{M}\mathbf{l})^\intercal\left(\Gamma\left(\frac{1}{2}\right)\right)^D\right)(\det \mathbf{M})^{\frac{1}{2}} &= \boldsymbol{\Sigma}.
\end{aligned} \tag{2.120}$$

Therefore,

$$\begin{aligned}
\lambda &= 1 - \frac{D}{2}\ln\pi - \frac{1}{2}\ln(\det \mathbf{M}), \\
\mathbf{l} &= \mathbf{0}, \\
\mathbf{M} &= 2\boldsymbol{\Sigma}.
\end{aligned} \tag{2.121}$$

Thus,

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}}(\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.122}$$

2.15

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.123}$$

Then, by the definition,

$$\mathrm{H}(\mathbf{x}) = -\int \mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) \ln \mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) d\mathbf{x}. \tag{2.124}$$

The right hand side can be written as

$$
\begin{aligned}
&-\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\ln|\det\boldsymbol{\Sigma}|-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)d\mathbf{x}\\
&=\left(\frac{D}{2}\ln(2\pi)+\frac{1}{2}\ln|\det\boldsymbol{\Sigma}|\right)\int\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}\\
&\quad+\frac{1}{2}\int(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}.
\end{aligned}
\tag{2.125}
$$

Since

$$
\begin{aligned}
\int\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}&=1,\\
\int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}&=\boldsymbol{\Sigma},
\end{aligned}
\tag{2.126}
$$

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

$$\frac{D}{2}\ln(2\pi)+\frac{1}{2}\ln|\det\boldsymbol{\Sigma}| \tag{2.127}$$

and

$$\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right)=\frac{D}{2}. \tag{2.128}$$

Therefore,

$$\mathrm{H}(\mathbf{x})=\frac{D}{2}\left(1+\ln(2\pi)\right)+\frac{1}{2}\ln|\det\boldsymbol{\Sigma}|. \tag{2.129}$$

2.16

Let x be a variable such that

$$x=x_1+x_2, \tag{2.130}$$

where

$$
\begin{aligned}
p(x_1)&=\mathcal{N}\left(x_1|\mu_1,\tau_1^{-1}\right),\\
p(x_2)&=\mathcal{N}\left(x_2|\mu_2,\tau_2^{-1}\right).
\end{aligned}
\tag{2.131}
$$

By marginalisation,

$$p(x)=\int_{-\infty}^{\infty}p(x|x_2)p(x_2)dx_2. \tag{2.132}$$

The right hand side can be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu_1 + x_2, \tau_1^{-1}\right) \mathcal{N}\left(x_2|\mu_2, \tau_2^{-1}\right) dx_2 \\
&= \int_{-\infty}^{\infty} \left(\frac{\tau_1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_1}{2}(x - \mu_1 - x_2)^2\right) \left(\frac{\tau_2}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_2}{2}(x_2 - \mu_2)^2\right) dx_2.
\end{aligned} \tag{2.133}$$

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

$$\begin{aligned}
&-\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1(x-\mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1}{2}(x-\mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 \\
&+\frac{\tau_1 + \tau_2}{2}\left(\frac{\tau_1(x-\mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 \\
=&-\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1(x-\mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2.
\end{aligned} \tag{2.134}$$

Therefore,

$$p(x) = \mathcal{N}\left(x|\mu_1 + \mu_2, \tau_1^{-1} + \tau_2^{-1}\right). \tag{2.135}$$

Thus, by 1.35,

$$\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln(2\pi) + \ln\left(\tau_1^{-1} + \tau_2^{-1}\right)\right). \tag{2.136}$$

2.17

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

$$\begin{aligned}
\mathbf{S} &= \frac{1}{2}\left(\mathbf{\Sigma}^{-1} + \left(\mathbf{\Sigma}^{-1}\right)^\intercal\right), \\
\mathbf{A} &= \frac{1}{2}\left(\mathbf{\Sigma}^{-1} - \left(\mathbf{\Sigma}^{-1}\right)^\intercal\right).
\end{aligned} \tag{2.137}$$

Then

$$\mathbf{\Sigma}^{-1} = \mathbf{S} + \mathbf{A}. \tag{2.138}$$

Therefore,

$$(\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{S}(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.139}$$

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

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\left(\boldsymbol{\Sigma}^{-1}\right)^\intercal(\mathbf{x}-\boldsymbol{\mu}). \tag{2.140}$$

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

$$-\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^\intercal(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.141}$$

Thus,

$$(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) = 0. \tag{2.142}$$

Hence

$$(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{S}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.143}$$

2.18

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.144}$$

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

$$\overline{\mathbf{u}_d}^\intercal\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d, \tag{2.145}$$

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

$$\overline{\mathbf{u}_d}^\intercal\overline{\boldsymbol{\Sigma}}^\intercal\mathbf{u}_d = \left(\overline{\boldsymbol{\Sigma}}\overline{\mathbf{u}_d}\right)^\intercal\mathbf{u}_d. \tag{2.146}$$

The right hand side can be writtet as

$$\overline{\lambda_d}\overline{\mathbf{u}_d}^\intercal\mathbf{u}_d = \overline{\lambda_d}. \tag{2.147}$$

Therefore,

$$\lambda_d = \overline{\lambda_d}. \tag{2.148}$$

Additionally, for $d \neq d'$, taking the inner product with $\mathbf{u}'_d$ on n both sides of the original equation gives

$$\mathbf{u}_{d'}^\intercal\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_{d'}^\intercal\mathbf{u}_d. \tag{2.149}$$

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

$$\mathbf{u}_{d'}^\intercal\boldsymbol{\Sigma}^\intercal\mathbf{u}_d = \left(\boldsymbol{\Sigma}\mathbf{u}_{d'}\right)^\intercal\mathbf{u}_d. \tag{2.150}$$

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

$$\lambda_d\mathbf{u}_{d'}^\intercal\mathbf{u}_d = \lambda_{d'}\mathbf{u}_{d'}^\intercal\mathbf{u}_d. \tag{2.151}$$

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

$$\mathbf{u}_{d'}^\intercal\mathbf{u}_d = 0. \tag{2.152}$$

2.19

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.153}$$

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

$$\begin{aligned}\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \cdots, \lambda_D),\\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D].\end{aligned} \tag{2.154}$$

Then

$$\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}. \tag{2.155}$$

By 2.18,

$$\mathbf{U}^\intercal\mathbf{U} = \mathbf{I}. \tag{2.156}$$

Therefore,

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\intercal,\\ \boldsymbol{\Sigma}^{-1} &= \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\intercal,\end{aligned} \tag{2.157}$$

Thus,

$$\begin{aligned}\boldsymbol{\Sigma} &= \sum_{d=1}^{D}\lambda_d\mathbf{u}_d\mathbf{u}_d^\intercal,\\ \boldsymbol{\Sigma}^{-1} &= \sum_{d=1}^{D}\frac{1}{\lambda_d}\mathbf{u}_d\mathbf{u}_d^\intercal.\end{aligned} \tag{2.158}$$

2.20

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

$$\boldsymbol{\Sigma}\mathbf{u}_d = \lambda_d\mathbf{u}_d, \tag{2.159}$$

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

$$\begin{aligned} \mathbf{\Lambda} &= \mathrm{diag}(\lambda_1, \cdots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \tag{2.160}$$

By 2.19,

$$\mathbf{a}^\intercal \mathbf{\Sigma} \mathbf{a} = \mathbf{b}^\intercal \mathbf{\Lambda} \mathbf{b}, \tag{2.161}$$

where

$$\mathbf{b} = \mathbf{U}^\intercal \mathbf{a}. \tag{2.162}$$

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

$$\mathbf{a}^\intercal \mathbf{\Sigma} \mathbf{a} > 0 \tag{2.163}$$

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

$$\lambda_d > 0. \tag{2.164}$$

2.21

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

2.22

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

$$\mathbf{\Sigma} \mathbf{\Lambda} = \mathbf{I}. \tag{2.165}$$

Taking the transpose of the both sides gives

$$\mathbf{\Lambda}^\intercal \mathbf{\Sigma} = \mathbf{I}. \tag{2.166}$$

Therefore,

$$\mathbf{\Lambda}^\intercal = \mathbf{\Lambda}. \tag{2.167}$$

2.23

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

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

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

$$\begin{aligned} \mathbf{\Lambda}' &= \operatorname{diag}\left(\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_D^{-\frac{1}{2}}\right), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \tag{2.169}$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\intercal \mathbf{U}\mathbf{\Lambda}'\mathbf{\Lambda}'^\intercal \mathbf{U}^\intercal (\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \tag{2.170}$$

By the transformation

$$\mathbf{y} = \mathbf{\Lambda}'^\intercal \mathbf{U}^\intercal (\mathbf{x} - \boldsymbol{\mu}) \tag{2.171}$$

and the property

$$\mathbf{U}^\intercal \mathbf{U} = \mathbf{I}, \tag{2.172}$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left|\det\left(\mathbf{U}\mathbf{\Lambda}'^{-1}\right)\right| d\mathbf{y} = |\det \mathbf{\Sigma}|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \tag{2.173}$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\intercal \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det \mathbf{\Sigma}|^{\frac{1}{2}} \Delta^D V_D, \tag{2.174}$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \tag{2.175}$$

2.24

Let

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

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

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

Therefore,

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

Thus,

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

2.25

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.176}$$

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

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \tag{2.177}$$

where

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

Then

$$\begin{aligned} &-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\ =&-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\intercal \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\intercal \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\intercal \boldsymbol{\Lambda}_{ac}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\ &-\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\intercal \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\intercal \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^\intercal \boldsymbol{\Lambda}_{bc}(\mathbf{x}_c - \boldsymbol{\mu}_c) \\ &-\frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Lambda}_{ca}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Lambda}_{cb}(\mathbf{x}_b - \boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_c - \boldsymbol{\mu}_c)^\intercal \boldsymbol{\Lambda}_{cc}(\mathbf{x}_c - \boldsymbol{\mu}_c). \end{aligned} \tag{2.178}$$

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

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c})^\intercal \boldsymbol{\Sigma}_{a|b,c}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_{a|b,c}), \tag{2.179}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\left(\mathbf{x}_c - \boldsymbol{\mu}_c\right), \\ \boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \tag{2.180}$$

Therefore,

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

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

$$p(\mathbf{x}_a|\mathbf{x}_b) = \int \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c})p(\mathbf{x}_c)d\mathbf{x}_c. \tag{2.182}$$

Thus,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \tag{2.183}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right) + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\mu}_c, \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \tag{2.184}$$

2.26

We have

$$\begin{aligned} &\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\right)(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}) \\ =&\mathbf{I} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D} + \mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D} \\ &- \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\mathbf{B}\mathbf{C}\mathbf{D}. \end{aligned} \tag{2.185}$$

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

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

The right hand sidde can be written as

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

Therefore,

$$\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}\right)(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}) = \mathbf{I}. \tag{2.188}$$

Thus,

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}. \tag{2.189}$$

2.27

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

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \int\int (\mathbf{x} + \mathbf{z}) p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \tag{2.190}$$

The right hand side can be written as

$$\int \mathbf{x}\left(\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}\right) d\mathbf{x} + \int \mathbf{z}\left(\int p(\mathbf{x}, \mathbf{z}) d\mathbf{x}\right) d\mathbf{z} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} + \int \mathbf{z} p(\mathbf{z}) d\mathbf{z}. \tag{2.191}$$

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

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \mathrm{E}\,\mathbf{x} + \mathrm{E}\,\mathbf{z}. \tag{2.192}$$

Additionally,

$$\operatorname{cov}(\mathbf{x} + \mathbf{z}) = \int\int (\mathbf{x} + \mathbf{z} - \mathrm{E}(\mathbf{x} + \mathbf{z}))(\mathbf{x} + \mathbf{z} - \mathrm{E}(\mathbf{x} + \mathbf{z}))^\intercal p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \tag{2.193}$$

The right hand side can be written as

$$\begin{aligned}&\int\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathbf{x} - \mathrm{E}\,\mathbf{x})^\intercal p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} + \int\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})(\mathbf{z} - \mathrm{E}\,\mathbf{z})^\intercal p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}\\ &+ \int\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathbf{x} - \mathrm{E}\,\mathbf{x})^\intercal p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} + \int\int (\mathbf{z} - \mathrm{E}\,\mathbf{z})(\mathbf{z} - \mathrm{E}\,\mathbf{z})^\intercal p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}.\end{aligned} \tag{2.194}$$

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

$$\begin{aligned}\int (\mathbf{x} - \mathrm{E}\,\mathbf{x})\, p(\mathbf{x}) d\mathbf{x} \int (\mathbf{z} - \mathrm{E}\,\mathbf{z})^\intercal p(\mathbf{z}) d\mathbf{z} = \mathbf{O},\\ \int (\mathbf{z} - \mathrm{E}\,\mathbf{z})\, p(\mathbf{z}) d\mathbf{z} \int (\mathbf{x} - \mathrm{E}\,\mathbf{x})^\intercal p(\mathbf{x}) d\mathbf{x} = \mathbf{O}.\end{aligned} \tag{2.195}$$

Therefore,

$$\operatorname{cov}(\mathbf{x} + \mathbf{z}) = \operatorname{cov}\mathbf{x} + \operatorname{cov}\mathbf{z}. \tag{2.196}$$

2.28

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

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

where

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

and

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

Then, by 2.29,

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

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

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\left(\boldsymbol{\Lambda} + \mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathbf{A}^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}) \\
&+\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\intercal}\mathbf{L}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}) \\
=&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\left(\boldsymbol{\Lambda} + \mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}) + \\
&-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right)^{\intercal}\mathbf{L}\left(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right) \\
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}).
\end{aligned} \tag{2.197}$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right)^{\intercal}\mathbf{L}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right). \tag{2.198}$$

Therefore,

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right), \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right). \end{aligned} \tag{2.199}$$

2.29

Let

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

Then, by 2.24,

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

2.30

Let

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

Then

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

2.31

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

$$\mathbf{y} = \mathbf{x} + \mathbf{z}, \tag{2.200}$$

where

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}\right), \\ p(\mathbf{z}) &= \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z}, \mathbf{\Sigma}_\mathbf{z}\right). \end{aligned} \tag{2.201}$$

By marginalisation,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{2.202}$$

The right hand side can be written as

$$\int \mathcal{N}\left(\mathbf{y}|\mathbf{x} + \boldsymbol{\mu}_\mathbf{z}, \mathbf{\Sigma}_\mathbf{z}\right) \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}\right) d\mathbf{x}. \tag{2.203}$$

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

$$-\frac{1}{2}\left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_\mathbf{z}\right)^\intercal \mathbf{\Sigma}_\mathbf{z}^{-1} \left(\mathbf{y} - \mathbf{x} - \boldsymbol{\mu}_\mathbf{z}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^\intercal \mathbf{\Sigma}_\mathbf{x}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}). \tag{2.204}$$

The first and second order terms can be written as

$$-\mathbf{x}^{\intercal}\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}}-\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\right)+\mathbf{y}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}}=\mathbf{u}^{\intercal}\mathbf{v} \tag{2.205}$$

and

$$-\frac{1}{2}\mathbf{x}^{\intercal}\left(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}+\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\right)\mathbf{x}+\frac{1}{2}\mathbf{x}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{y}+\frac{1}{2}\mathbf{y}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{x}-\frac{1}{2}\mathbf{y}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{y}=-\frac{1}{2}\mathbf{u}^{\intercal}\mathbf{R}\mathbf{u}, \tag{2.206}$$

respectively, where

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

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

$$-\frac{1}{2}\left(\mathbf{u}-\mathbf{R}^{-1}\mathbf{v}\right)^{\intercal}\mathbf{R}\left(\mathbf{u}-\mathbf{R}^{-1}\mathbf{v}\right), \tag{2.207}$$

where

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

by 2.29 and 2.30. Thus,

$$p(\mathbf{y})=\mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{x}}+\boldsymbol{\mu}_{\mathbf{z}},\boldsymbol{\Sigma}_{\mathbf{x}}+\boldsymbol{\Sigma}_{\mathbf{z}}\right). \tag{2.208}$$

2.32

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

$$\begin{aligned}p(\mathbf{x})&=\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right),\\ p(\mathbf{y}|\mathbf{x})&=\mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1}\right).\end{aligned} \tag{2.209}$$

By the Bayes' theorem,

$$p(\mathbf{y}|\mathbf{x})p(\mathbf{x})=p(\mathbf{x}|\mathbf{y})p(\mathbf{y}). \tag{2.210}$$

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

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

Since the first term can be written as

$$\begin{aligned}
&-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}\right)^{\intercal}\mathbf{L}\left(\mathbf{y}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}\right)\\
=&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})+(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathbf{A}^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})\\
&-\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}),
\end{aligned} \tag{2.212}$$

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

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})^{\intercal}\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})+\frac{1}{2}\mathbf{z}^{\intercal}\left(\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}+\boldsymbol{\Lambda}\right)\mathbf{z}\\
&-\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})\\
=&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})^{\intercal}\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})\\
&-\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^{\intercal}\mathbf{M}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}),
\end{aligned} \tag{2.213}$$

where

$$\begin{aligned}
\mathbf{z}&=\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}),\\
\mathbf{M}&=\mathbf{L}-\mathbf{L}\mathbf{A}\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^{\intercal}\mathbf{L}.
\end{aligned} \tag{2.214}$$

we have

$$\boldsymbol{\mu}+\mathbf{z}=\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right). \tag{2.215}$$

By 2.26,

$$\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)^{-1}=\boldsymbol{\Lambda}^{-1}-\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\intercal}\left(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\intercal}\right)^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}. \tag{2.216}$$

Therefore,

$$\mathbf{M}=\left(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\intercal}\right)^{-1}. \tag{2.217}$$

Thus,

$$\begin{aligned}
p(\mathbf{x}|\mathbf{y})&=\mathcal{N}\left(\mathbf{x}\middle|\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^{\intercal}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right),\left(\boldsymbol{\Lambda}+\mathbf{A}^{\intercal}\mathbf{L}\mathbf{A}\right)^{-1}\right),\\
p(\mathbf{y})&=\mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\intercal}\right).
\end{aligned} \tag{2.218}$$

2.33

Refer to 2.32.

2.34

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

$$\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.219}$$

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

$$\begin{aligned}\mathbf{0} &= \sum_{n=1}^{N}\left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^\intercal\right)(\mathbf{x}_n - \boldsymbol{\mu}),\\ \mathbf{O} &= -\frac{N}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^\intercal + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\intercal.\end{aligned} \tag{2.220}$$

Therefore,

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n,\\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\intercal.\end{aligned} \tag{2.221}$$

2.35

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.222}$$

Then

$$\mathrm{E}\,\mathbf{x}\mathbf{x}^\intercal = \int \mathbf{x}\mathbf{x}^\intercal\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)d\mathbf{x}. \tag{2.223}$$

The right hand side can be written as

$$\begin{aligned}&\int(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu} + \boldsymbol{\mu})^\intercal\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)d\mathbf{x}\\ =&\int(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)d\mathbf{x} + \boldsymbol{\mu}\int(\mathbf{x} - \boldsymbol{\mu})^\intercal\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)d\mathbf{x}\\ &+\left(\int(\mathbf{x} - \boldsymbol{\mu})\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)d\mathbf{x}\right)\boldsymbol{\mu}^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\int\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)d\mathbf{x}.\end{aligned} \tag{2.224}$$

Since

$$\begin{aligned}
\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} &= 1, \\
\int \mathbf{x}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} &= \boldsymbol{\mu}, \\
\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} &= \boldsymbol{\Sigma},
\end{aligned} \tag{2.225}$$

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

$$\mathrm{E}\,\mathbf{x}\mathbf{x}^{\intercal} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}. \tag{2.226}$$

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

$$\begin{aligned}
p(\mathbf{x}_n) &= \mathcal{N}\left(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \\
p(\mathbf{x}_m) &= \mathcal{N}\left(\mathbf{x_m}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).
\end{aligned} \tag{2.227}$$

If $n \neq m$, then

$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^{\intercal} = \mathrm{E}\,\mathbf{x}_n\,\mathrm{E}\,\mathbf{x}_m^{\intercal}. \tag{2.228}$$

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

$$\mathrm{E}\,\mathbf{x}_n\mathbf{x}_m^{\intercal} = \delta_{nm}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}. \tag{2.229}$$

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

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.230}$$

By 2.34,

$$\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \\
\boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\intercal}.
\end{aligned} \tag{2.231}$$

Then

$$\mathrm{E}\,\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\intercal}. \tag{2.232}$$

The right hand side can be written as

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\,\mathbf{x}_n\mathbf{x}_n^\intercal-\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\left(\sum_{n=1}^{N}\mathbf{x_n}\right)\mathbf{x}_n^\intercal-\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\,\mathbf{x}_n\left(\sum_{n=1}^{N}\mathbf{x_n}\right)^\intercal\\&+\frac{1}{N^3}\sum_{n=1}^{N}\mathrm{E}\left(\sum_{n=1}^{N}\mathbf{x_n}\right)\left(\sum_{n=1}^{N}\mathbf{x_n}\right)^\intercal.\end{aligned}\tag{2.233}$$

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

$$-\frac{1}{N}\left((\mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\intercal)+(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right)=-\frac{1}{N}\mathbf{\Sigma}-\boldsymbol{\mu}\boldsymbol{\mu}^\intercal.\tag{2.234}$$

The fourth term can be written as

$$\frac{1}{N^2}\left(N\left(\mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right)+N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right)=\frac{1}{N}\mathbf{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\intercal.\tag{2.235}$$

Therefore,

$$\mathrm{E}\,\mathbf{\Sigma}_{\mathrm{ML}}=\frac{N-1}{N}\mathbf{\Sigma}.\tag{2.236}$$

2.36

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

$$p(x_n)=\mathcal{N}\left(x_n|\mu,\sigma^2\right).\tag{2.237}$$

Let us assume that μ is known. Then, by 2.34,

$${\sigma^2_{\mathrm{ML}}}^{(N)}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2.\tag{2.238}$$

The right hand side can be written as

$$\frac{1}{N}(x_N-\mu)^2+\frac{1}{N}\sum_{n=1}^{N-1}(x_n-\mu)^2=\frac{1}{N}(x_N-\mu)^2+\frac{N-1}{N}{\sigma^2_{\mathrm{ML}}}^{(N-1)}.\tag{2.239}$$

Therefore,

$${\sigma^2_{\mathrm{ML}}}^{(N)}={\sigma^2_{\mathrm{ML}}}^{(N-1)}+\frac{1}{N}\left((x_N-\mu)^2-{\sigma^2_{\mathrm{ML}}}^{(N-1)}\right).\tag{2.240}$$

Since

$$\frac{\partial}{\partial \sigma^2}\left(-\ln p\left(x_n|\sigma^2\right)\right) = \frac{1}{2\sigma^2} - \frac{1}{2\left(\sigma^2\right)^2}(x_n-\mu)^2, \tag{2.241}$$

we have

$$\sigma^2_{\mathrm{ML}}{}^{(N)} = \sigma^2_{\mathrm{ML}}{}^{(N-1)} - \frac{\sigma^2_{\mathrm{ML}}{}^{(N-1)}}{N} \frac{\partial}{\partial \sigma^2_{\mathrm{ML}}{}^{(N-1)}}\left(-\ln p\left(x_N|\sigma^2_{\mathrm{ML}}{}^{(N-1)}\right)\right). \tag{2.242}$$

2.37

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

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.243}$$

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

$$\boldsymbol{\Sigma}^{(N)}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^\intercal. \tag{2.244}$$

The right hand side can be written as

$$\begin{aligned}
&\frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^\intercal + \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^\intercal \\
&= \frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^\intercal + \frac{N-1}{N}\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}.
\end{aligned} \tag{2.245}$$

Therefore,

$$\boldsymbol{\Sigma}^{(N)}_{\mathrm{ML}} = \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}} + \frac{1}{N}\left((\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^\intercal - \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right). \tag{2.246}$$

Since

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\left(-\ln p(x_n|\boldsymbol{\Sigma})\right) = -\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^\intercal + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^\intercal, \tag{2.247}$$

we have

$$\boldsymbol{\Sigma}^{(N)}_{\mathrm{ML}} = \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}} - \frac{\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}}{N}\frac{\partial}{\partial \boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}}\left(-\ln p\left(\mathbf{x}_N|\boldsymbol{\Sigma}^{(N-1)}_{\mathrm{ML}}\right)\right). \tag{2.248}$$

2.38

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

$$\begin{aligned} p(x_n|\mu) &= \mathcal{N}\left(x_n|\mu, \sigma^2\right), \\ p(\mu) &= \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right). \end{aligned} \tag{2.249}$$

By the Bayes' theorem,

$$p(\mu|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu)p(\mu). \tag{2.250}$$

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

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2. \tag{2.251}$$

The first term can be written as

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}} + \mu_{\mathrm{ML}} - \mu)^2 = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2}(\mu_{\mathrm{ML}} - \mu)^2. \tag{2.252}$$

where

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{2.253}$$

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

$$\begin{aligned} &-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2}(\mu_{\mathrm{ML}} - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 \\ &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 - \frac{1}{2\sigma_N^2}\left(\mu - \mu_N\right)^2 + \frac{\mu_N^2}{2\sigma_N^2}, \end{aligned} \tag{2.254}$$

where

$$\begin{aligned} \mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\ \sigma_N^2 &= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}. \end{aligned} \tag{2.255}$$

Therefore,

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu \mid \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}\right). \tag{2.256}$$

2.39 (Incomplete)

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

$$\begin{aligned} p(x_n|\mu) &= \mathcal{N}\left(x_n|\mu, \sigma^2\right), \\ p(\mu) &= \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right). \end{aligned} \tag{2.257}$$

Then, by 2.38,

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right), \tag{2.258}$$

where

$$\begin{aligned} \mu_N &= \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{n=1}^{N} x_n + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\ \sigma_N^2 &= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}. \end{aligned} \tag{2.259}$$

Then

$$\begin{aligned} \mu_N &= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2}x_N + \frac{\sigma^2 - \sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\ \sigma_N^2 &= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2}\sigma_{N-1}^2. \end{aligned} \tag{2.260}$$

Additionally, we have

$$\begin{aligned} p(\mu|x_1, \cdots, x_{N-1}) &= \mathcal{N}\left(\mu|\mu_{N-1}, \sigma_{N-1}^2\right), \\ p(x_N|\mu) &= \mathcal{N}\left(x_N|\mu, \sigma^2\right). \end{aligned} \tag{2.261}$$

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

$$\begin{aligned} & -\frac{1}{2\sigma_{N-1}^2}(\mu - \mu_{N-1})^2 - \frac{1}{2\sigma^2}(x_N - \mu)^2 \\ =& -\frac{1}{2\frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2+\sigma^2}}\left(\mu - \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} - \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N\right)^2 \\ & + \frac{1}{2\frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2+\sigma^2}}\left(\frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N\right) - \frac{\mu_{N-1}^2}{2\sigma_{N-1}^2} - \frac{x_N^2}{2\sigma^2}. \end{aligned} \tag{2.262}$$

Therefoere,

$$\begin{aligned} \mu_N &= \frac{\sigma^2}{\sigma_{N-1}^2 + \sigma^2}\mu_{N-1} + \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2 + \sigma^2}x_N, \\ \sigma_N^2 &= \frac{\sigma_{N-1}^2\sigma^2}{\sigma_{N-1}^2 + \sigma^2}. \end{aligned} \tag{2.263}$$

2.40

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

$$\begin{aligned} p(\mathbf{x}_n|\boldsymbol{\mu}) &= \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \\ p(\boldsymbol{\mu}) &= \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}\right). \end{aligned} \tag{2.264}$$

By the Bayes' theorem,

$$p(\boldsymbol{\mu}|\mathbf{X})p(\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}). \tag{2.265}$$

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

$$-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\intercal. \tag{2.266}$$

The first term can be written as

$$\begin{aligned} &-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}} + \boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}) \\ &= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}}) - \frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}} - \boldsymbol{\mu}). \end{aligned} \tag{2.267}$$

where

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \tag{2.268}$$

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

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})-\frac{N}{2}(\boldsymbol{\mu}_{\mathrm{ML}}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\mathrm{ML}}-\boldsymbol{\mu})\\
&-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\intercal\\
=&-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)^\intercal\boldsymbol{\Sigma}_N^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_N)\\
&+\frac{1}{2}\boldsymbol{\mu}_N^\intercal\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N,
\end{aligned}
\tag{2.269}
$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_N &= \left(N\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_{\mathrm{ML}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right),\\
\boldsymbol{\Sigma}_N &= \left(N\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}.
\end{aligned}
\tag{2.270}
$$

Therefore,

$$
p(\boldsymbol{\mu}|\mathbf{X})=\mathcal{N}\left(\boldsymbol{\mu}\mid\left(N\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(N\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_{\mathrm{ML}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0\right),\left(N\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\right).
\tag{2.271}
$$

2.41

By the definition,

$$
\mathrm{Gam}(\lambda|a,b)=\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda).
\tag{2.272}
$$

Then

$$
\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda=\frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{a-1}\exp(-b\lambda)d\lambda.
\tag{2.273}
$$

By the transformation

$$
\lambda'=b\lambda,
\tag{2.274}
$$

the right hand side can be written as

$$
\frac{b^a}{\Gamma(a)}\int_0^\infty\left(\frac{\lambda'}{b}\right)^{a-1}\exp(-\lambda')\frac{1}{b}d\lambda'=\frac{1}{\Gamma(a)}\int_0^\infty \lambda'^{a-1}\exp(-\lambda')d\lambda'.
\tag{2.275}
$$

The right hand side can be written as

$$\frac{1}{\Gamma(a)}\Gamma(a) = 1. \tag{2.276}$$

Therefore,

$$\int_0^\infty \mathrm{Gam}(\lambda|a,b)d\lambda = 1. \tag{2.277}$$

2.42

Let λ be a variable such that

$$p(\lambda) = \mathrm{Gam}(\lambda|a,b). \tag{2.278}$$

By the definition,

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.279}$$

Then

$$\mathrm{E}\,\lambda = \frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^a \exp\left(-\frac{\lambda}{b}\right)d\lambda. \tag{2.280}$$

By the transformation

$$\lambda' = b\lambda, \tag{2.281}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\int_0^\infty \left(\frac{\lambda'}{b}\right)^a \exp(-\lambda')\frac{1}{b}d\lambda' = \frac{1}{b\Gamma(a)}\int_0^\infty \lambda'^a \exp(-\lambda')d\lambda'. \tag{2.282}$$

The right hand side can be written as

$$\frac{1}{b\Gamma(a)}\Gamma(a+1) = \frac{a}{b}. \tag{2.283}$$

Therefore,

$$\mathrm{E}\,\lambda = \frac{a}{b}. \tag{2.284}$$

Additionally,

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

By the transformation

$$\lambda' = b\lambda, \tag{2.286}$$

the right hand side can be written as

$$\frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{\lambda'}{b}\right)^{a+1} \exp(-\lambda') \frac{1}{b} d\lambda' = \frac{1}{b^2\Gamma(a)} \int_0^\infty \lambda'^{a+1} \exp(-\lambda') d\lambda'. \tag{2.287}$$

The right hand side can be written as

$$\frac{1}{b^2\Gamma(a)} \Gamma(a+2) = \frac{a(a+1)}{b^2}. \tag{2.288}$$

Therefore,

$$\mathrm{E}\,\lambda^2 = \frac{a(a+1)}{b^2}. \tag{2.289}$$

By the definition,

$$\operatorname{var}\lambda = \mathrm{E}\,\lambda^2 - (\mathrm{E}\,\lambda)^2. \tag{2.290}$$

Therefore,

$$\operatorname{var}\lambda = \frac{a}{b^2}. \tag{2.291}$$

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

$$0 = \frac{b^a}{\Gamma(a)} \left(\frac{a-1}{\lambda} - b\right) \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right). \tag{2.292}$$

Therefore,

$$\operatorname{mode}\lambda = \frac{a-1}{b}. \tag{2.293}$$

2.43

Let

$$p\left(x|\sigma^2, q\right) = \frac{q}{2\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \exp\left(-\frac{|x|^q}{2\sigma^2}\right). \tag{2.294}$$

Then

$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = \frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^\infty \exp\left(-\frac{x^q}{2\sigma^2}\right) dx. \tag{2.295}$$

By the transformation

$$x' = \frac{x^q}{2\sigma^2}, \tag{2.296}$$

the right hand side can be written as

$$\begin{aligned}
&\frac{q}{\Gamma(\frac{1}{q})} \left(2\sigma^2\right)^{-\frac{1}{q}} \int_0^\infty \exp(-x') \left(2\sigma^2\right)^{\frac{1}{q}} \frac{1}{q} x^{\frac{1}{q}-1} dx' \\
&= \frac{1}{\Gamma(\frac{1}{q})} \int_0^\infty x^{\frac{1}{q}-1} \exp(-x') dx'.
\end{aligned} \tag{2.297}$$

The right hand side can be written as

$$\frac{1}{\Gamma(\frac{1}{q})} \Gamma\left(\frac{1}{q}\right) = 1. \tag{2.298}$$

Therefore,

$$\int_{-\infty}^{\infty} p\left(x|\sigma^2, q\right) dx = 1. \tag{2.299}$$

Additionally,

$$p\left(x|\sigma^2, 2\right) = \frac{1}{\Gamma(\frac{1}{2})} \left(2\sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{2.300}$$

Therefore,

$$p\left(x|\sigma^2, 2\right) = \mathcal{N}\left(x|0, \sigma^2\right). \tag{2.301}$$

Finally, let $\mathbf{t} = (t_1, \cdots, t_N)^\intercal$ and $\mathbf{X} = \{x_1, \cdots, x_N\}$ such that

$$t_n = y(\mathbf{x}_n, \mathbf{w}) + \epsilon_n, \tag{2.302}$$

where

$$p(\epsilon_n) = p\left(\epsilon_n|\sigma^2, q\right). \tag{2.303}$$

Therefore, the logarithm of $p(\epsilon_n)$ except the terms independent of $\mathbf{w}$ and σ^2 can be written as

$$-\frac{|\epsilon_n|^q}{2\sigma^2} - \frac{1}{q} \ln\left(2\sigma^2\right). \tag{2.304}$$

Thus, 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

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln\left(2\sigma^2\right). \tag{2.305}$$

2.44

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

$$\begin{aligned} p(x_n|\mu,\tau) &= \mathcal{N}\left(x_n|\mu,\tau^{-1}\right), \\ p(\mu,\tau) &= \mathcal{N}\left(\mu|\mu_0,(\beta\tau)^{-1}\right)\mathrm{Gam}(\tau|a,b). \end{aligned} \tag{2.306}$$

By the Bayes' theorem,

$$p(\mu,\tau|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mu,\tau)p(\mu,\tau). \tag{2.307}$$

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

$$\begin{aligned} &\frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{1}{2}\ln\tau - \frac{\beta\tau}{2}(\mu-\mu_0)^2 + (a-1)\ln\tau - b\tau \\ =&\left(a+\frac{N-1}{2}\right)\ln\tau - \frac{N\tau}{2}(\bar{x}-\mu)^2 - \frac{\beta\tau}{2}(\mu-\mu_0)^2 - b\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\bar{x})^2, \end{aligned} \tag{2.308}$$

where

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N}x_n. \tag{2.309}$$

Since

$$-\frac{N\tau}{2}(\bar{x}-\mu)^2 - \frac{\beta\tau}{2}(\mu-\mu_0)^2 = -\frac{(N+\beta)\tau}{2}\left(\mu-\frac{N\bar{x}+\beta\mu_0}{N+\beta}\right)^2 - \frac{N\beta\tau(\bar{x}-\mu_0)^2}{2(N+\beta)}, \tag{2.310}$$

the right hand side can be written as

$$\begin{aligned} &-\frac{(N+\beta)\tau}{2}\left(\mu-\frac{N\bar{x}+\beta\mu_0}{N+\beta}\right)^2 \\ &+\left(a+\frac{N-1}{2}\right)\ln\tau - \left(b+\frac{N\beta(\bar{x}-\mu_0)^2}{2(N+\beta)}+\frac{1}{2}\sum_{n=1}^{N}(x_n-\bar{x})^2\right)\tau. \end{aligned} \tag{2.311}$$

Therefore,

$$\begin{aligned} p(\mu,\tau|\mathbf{x}) =& \mathcal{N}\left(\mu \mid \frac{N\bar{x}+\beta\mu_0}{N+\beta}, ((N+\beta)\tau)^{-1}\right) \\ &\mathrm{Gam}\left(\tau \mid a+\frac{N+1}{2}, b+\frac{N\beta(\bar{x}-\mu_0)^2}{2(N+\beta)}+\frac{1}{2}\sum_{n=1}^{N}(x_n-\bar{x})^2\right). \end{aligned} \tag{2.312}$$

2.45

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right). \tag{2.313}$$

Then

$$p(\mathbf{X}|\boldsymbol{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right). \tag{2.314}$$

The right hand side exept the terms independent of $\boldsymbol{\Lambda}$ can be written as

$$(\det \boldsymbol{\Lambda})^{\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\intercal}\boldsymbol{\Lambda}(\mathbf{x}_n - \boldsymbol{\mu})\right) = (\det \boldsymbol{\Lambda})^{\frac{N}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{S}\boldsymbol{\Lambda})\right), \tag{2.315}$$

where

$$\mathbf{S} = \sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^{\intercal}. \tag{2.316}$$

Therefore,

$$p(\mathbf{X}|\boldsymbol{\Lambda}) \propto (\det \boldsymbol{\Lambda})^{\frac{N}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\mathbf{S}\boldsymbol{\Lambda})\right). \tag{2.317}$$

Let us assume that a prior distribution of $\boldsymbol{\Lambda}$ is given by

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = B(\mathbf{W}, \nu)(\det \boldsymbol{\Lambda})^{\frac{\nu-D-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\mathbf{W}^{-1}\boldsymbol{\Lambda}\right)\right). \tag{2.318}$$

Then, by the definition,

$$p(\boldsymbol{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) \propto p(\mathbf{X}|\boldsymbol{\Lambda})\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu), \tag{2.319}$$

where the right hand side except the terms independent of $\boldsymbol{\Lambda}$ can be written as

$$(\det \boldsymbol{\Lambda})^{\frac{\nu+N-D-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\left(\mathbf{W}^{-1} + \mathbf{S}\right)\boldsymbol{\Lambda}\right)\right). \tag{2.320}$$

Therefore,

$$p(\boldsymbol{\Lambda}|\mathbf{X}, \mathbf{W}, \nu) = \mathcal{W}\left(\boldsymbol{\Lambda} \mid \left(\mathbf{W}^{-1} + \mathbf{S}\right)^{-1}, \nu + N\right). \tag{2.321}$$

Thus, $\mathcal{W}$ is a conjugate prior distribution of $\boldsymbol{\Lambda}$.

2.46

Let x be a variable such that

$$p(x|\mu, \tau, a, b) = \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathrm{Gam}(\tau|a, b). \tag{2.322}$$

Then

$$p(x|\mu, a, b) = \int_0^\infty \mathcal{N}\left(x|\mu, \tau^{-1}\right) \mathrm{Gam}(\tau|a, b) d\tau. \tag{2.323}$$

The right hand side can be written as

$$\begin{aligned}
&\int_0^\infty \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) d\tau \\
=&(2\pi)^{-\frac{1}{2}} \frac{b^a}{\Gamma(a)} \int_0^\infty \tau^{a-\frac{1}{2}} \exp\left(-\left(b + \frac{(x-\mu)^2}{2}\right)\tau\right) d\tau.
\end{aligned} \tag{2.324}$$

By the transformation

$$\tau' = \left(b + \frac{(x-\mu)^2}{2}\right)\tau, \tag{2.325}$$

the integral of the right hand side can be written as

$$\int_0^\infty \left(\frac{\tau'}{b + \frac{(x-\mu)^2}{2}}\right)^{a-\frac{1}{2}} \exp(-\tau') \frac{d\tau'}{b + \frac{(x-\mu)^2}{2}} = \Gamma\left(a + \frac{1}{2}\right)\left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.326}$$

Therefore,

$$p(x|\mu, \tau, a, b) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(a + \frac{1}{2})}{\Gamma(a)} b^a \left(b + \frac{(x-\mu)^2}{2}\right)^{-a-\frac{1}{2}}. \tag{2.327}$$

Let

$$\begin{aligned}
\nu &= 2a, \\
\lambda &= \frac{a}{b}.
\end{aligned} \tag{2.328}$$

Then

$$p(x|\mu, \lambda, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{2.329}$$

2.47

By the definition,

$$\mathrm{St}(x|\mu,\lambda,\nu)=\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \tag{2.330}$$

By the transformation

$$y=\frac{\lambda(x-\mu)^2}{\nu}, \tag{2.331}$$

the right hand side except the terms independent of x can be written as

$$(1+y)^{-\frac{\lambda(x-\mu)^2}{2y}-\frac{1}{2}}. \tag{2.332}$$

In the limit $y\to\infty$, it becomes

$$\exp\left(-\frac{\lambda}{2}(x-\mu)^2\right). \tag{2.333}$$

Therefore, in the limit $\nu\to\infty$, $\mathrm{St}(x|\mu,\lambda,\nu)$ becomes $\mathcal{N}(x|\mu,\lambda^{-1})$.

2.48

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

$$p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Lambda},\eta,\nu)=\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\mathbf{\Lambda})^{-1}\right)\mathrm{Gam}\left(\eta\mid\frac{\nu}{2},\frac{\nu}{2}\right). \tag{2.334}$$

Then

$$p(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Lambda},\nu)=\int_0^\infty\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},(\eta\mathbf{\Lambda})^{-1}\right)\mathrm{Gam}\left(\eta\mid\frac{\nu}{2},\frac{\nu}{2}\right)d\eta. \tag{2.335}$$

The right hand side can be written as

$$\begin{aligned}&\int_0^\infty(2\pi)^{-\frac{D}{2}}(\det(\eta\mathbf{\Lambda}))^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\eta\mathbf{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\eta^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\eta\right)d\eta\\ =&(2\pi)^{-\frac{D}{2}}\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\det\mathbf{\Lambda})^{\frac{1}{2}}\int_0^\infty\eta^{\frac{D+\nu}{2}-1}\exp\left(-\frac{1}{2}\left(\nu+(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\eta\right)d\eta.\end{aligned} \tag{2.336}$$

By the transformation

$$\eta' = \frac{1}{2}\left(\nu + (\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right)\eta, \tag{2.337}$$

the integral of the right hand side can be written as

$$\begin{aligned}
&\int_0^\infty \left(\frac{2\eta'}{\nu + (\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}-1} \exp(-\eta') \frac{2}{\nu + (\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})} d\eta' \\
&= \left(\frac{2}{\nu + (\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}\right)^{\frac{D+\nu}{2}} \Gamma\left(\frac{D+\nu}{2}\right).
\end{aligned} \tag{2.338}$$

Therefore,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\det \boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}} \left(1 + \frac{(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \tag{2.339}$$

2.49

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

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu). \tag{2.340}$$

By the definition,

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.341}$$

First,

$$\mathrm{E}\,\mathbf{x} = \int \mathbf{x}\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \tag{2.342}$$

The right hand side can be written as

$$\begin{aligned}
&\int \mathbf{x} \left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta\right) d\mathbf{x} \\
&= \int \left(\int \mathbf{x}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) d\mathbf{x}\right) \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta.
\end{aligned} \tag{2.343}$$

The right hand side can be written as

$$\boldsymbol{\mu} \int \mathrm{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\mu}. \tag{2.344}$$

Therefore,

$$\operatorname{E}\mathbf{x} = \boldsymbol{\mu}. \tag{2.345}$$

Additionally,

$$\operatorname{cov}\mathbf{x} = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal} \operatorname{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) d\mathbf{x}. \tag{2.346}$$

The right hand side can be written as

$$\begin{aligned} &\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal} \left(\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \operatorname{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta \right) d\mathbf{x} \\ &= \int \left(\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) d\mathbf{x} \right) \operatorname{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \end{aligned} \tag{2.347}$$

The right hand side can be written as

$$\int (\eta\boldsymbol{\Lambda})^{-1} \operatorname{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta = \boldsymbol{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \int \eta^{\frac{\nu}{2}-2} \exp\left(-\frac{\nu}{2}\eta\right) d\eta. \tag{2.348}$$

By the transformation

$$\eta' = \frac{\nu}{2}\eta, \tag{2.349}$$

the integral of the right hand side can be written as

$$\int \left(\frac{2}{\nu}\eta'\right)^{-\frac{\nu}{2}-2} \exp(-\eta') \frac{2}{\nu} d\eta' = \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2}-1\right). \tag{2.350}$$

Therefore, the right hand side can be written as

$$\boldsymbol{\Lambda}^{-1} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left(\frac{2}{\nu}\right)^{\frac{\nu}{2}-1} \Gamma\left(\frac{\nu}{2}-1\right) = \frac{\frac{\nu}{2}}{\frac{\nu}{2}-1} \boldsymbol{\Lambda}^{-1}. \tag{2.351}$$

Thus,

$$\operatorname{cov}\mathbf{x} = \frac{\nu}{\nu-2} \boldsymbol{\Lambda}^{-1}. \tag{2.352}$$

Finally, setting the derivative of $\operatorname{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$ with respect to $\mathbf{x}$ to zero gives

$$\mathbf{0} = -\frac{1}{2}\left(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^{\intercal}\right)(\mathbf{x}-\boldsymbol{\mu}) \int \eta \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1}\right) \operatorname{Gam}\left(\eta \mid \frac{\nu}{2}, \frac{\nu}{2}\right) d\eta. \tag{2.353}$$

Therefore,

$$\operatorname{mode}\mathbf{x} = \boldsymbol{\mu}. \tag{2.354}$$

2.50

By the definition,

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)=\frac{\Gamma(\frac{D+\nu}{2})}{\Gamma(\frac{\nu}{2})}\frac{(\det\boldsymbol{\Lambda})^{\frac{1}{2}}}{(\pi\nu)^{\frac{D}{2}}}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{D+\nu}{2}}. \tag{2.355}$$

By the transformation

$$y=\frac{(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{\nu}, \tag{2.356}$$

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

$$(1+y)^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}{2y}-\frac{D}{2}}. \tag{2.357}$$

In the limit $y\to\infty$, it becomes

$$\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.358}$$

Therefore, in the limit $\nu\to\infty$, $\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)$ becomes $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$.

2.51

We have

$$\exp(iA)\exp(-iA)=1. \tag{2.359}$$

The left hand side can be written as

$$(\cos A+i\sin A)(\cos A-i\sin A)=\cos^2 A+\sin^2 A. \tag{2.360}$$

Therefore,

$$\cos^2 A+\sin^2 A=1. \tag{2.361}$$

Additionally,

$$\cos(A-B)=\mathrm{Re}\left(\exp\left(i(A-B)\right)\right). \tag{2.362}$$

The right hand side can be written as

$$\mathrm{Re}\left(\exp(iA)\exp(-iB)\right)=\mathrm{Re}\left((\cos A+i\sin A)(\cos B-i\sin B)\right). \tag{2.363}$$

The right hand side can be written as $\cos A \cos B + \sin A \sin B$. Therefore,

$$\cos(A - B) = \cos A \cos B + \sin A \sin B. \tag{2.364}$$

Finally,

$$\sin(A - B) = \mathrm{Im}\left(\exp\left(i(A - B)\right)\right). \tag{2.365}$$

The right hand side can be written as

$$\mathrm{Im}\left(\exp(iA)\exp(-iB)\right) = \left((\cos A + i \sin A)(\cos B - i \sin B)\right). \tag{2.366}$$

The right hand side can be written as $\sin A \cos B - \cos A \sin B$. Therefore,

$$\sin(A - B) = \sin A \cos B - \cos A \sin B. \tag{2.367}$$

2.52 (Incomplete)

Let θ be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m \cos(\theta - \theta_0)\right), \tag{2.368}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m \cos \theta) d\theta. \tag{2.369}$$

By the Taylor series

$$\cos \alpha = 1 - \frac{1}{2}\alpha^2 + O\left(\alpha^4\right) \tag{2.370}$$

and the transformation

$$\xi = m^{\frac{1}{2}}(\theta - \theta_0), \tag{2.371}$$

we have

$$\exp\left(m \cos(\theta - \theta_0)\right) = \exp\left(m\left(1 - \frac{1}{2}(\theta - \theta_0)^2 + O\left((\theta - \theta_0)^4\right)\right)\right). \tag{2.372}$$

2.53

Let θ_0 be a parameter such that

$$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0. \tag{2.373}$$

The left hand side can be written as

$$\sum_{n=1}^{N}(\sin\theta_n\cos\theta_0 - \cos\theta_n\sin\theta_0) = \cos\theta_0\sum_{n=1}^{N}\sin\theta_n - \sin\theta_0\sum_{n=1}^{N}\cos\theta_n. \quad (2.374)$$

Therefore,

$$\theta_0 = \arctan\left(\frac{\sum_{n=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}\right). \quad (2.375)$$

2.54

Let θ be a variable such that

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)}\exp\left(m\cos(\theta-\theta_0)\right), \quad (2.376)$$

where

$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi}\exp(m\cos\theta)d\theta. \quad (2.377)$$

Setting the first and seconnd derivatives with respect to θ to zero gives

$$\begin{aligned} 0 &= -m\sin(\theta-\theta_0)p(\theta|\theta_0, m), \\ 0 &= \left(m^2\sin^2(\theta-\theta_0) - m\cos(\theta-\theta_0)\right)p(\theta|\theta_0, m). \end{aligned} \quad (2.378)$$

Therefore,

$$\begin{aligned} \operatorname*{argmax}_\theta p(\theta|\theta_0, m) &= \theta_0, \\ \operatorname*{argmin}_\theta p(\theta|\theta_0, m) &= \theta_0 - \pi\,\mathrm{sgn}(\theta_0-\pi). \end{aligned} \quad (2.379)$$

2.55

Let

$$\theta_0^{\mathrm{ML}} = \arctan\left(\frac{\sum_{n=1}^{N}\sin\theta_n}{\sum_{n=1}^{N}\cos\theta_n}\right). \quad (2.380)$$

Let

$$\begin{aligned} \bar{r}\cos\bar{\theta} &= \frac{1}{N}\sum_{n=1}^{N}\cos\theta_n, \\ \bar{r}\sin\bar{\theta} &= \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n. \end{aligned} \quad (2.381)$$

Then

$$\theta_0^{\mathrm{ML}} = \bar{\theta}. \tag{2.382}$$

Here,

$$\frac{1}{N}\sum_{n=1}^{N}\cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) = \left(\frac{1}{N}\sum_{n=1}^{N}\cos\theta_n\right)\cos\theta_0^{\mathrm{ML}} + \left(\frac{1}{N}\sum_{n=1}^{N}\sin\theta_n\right)\sin\theta_0^{\mathrm{ML}}. \tag{2.383}$$

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

$$\bar{r}\cos^2\bar{\theta} + \bar{r}\sin^2\bar{\theta} = \bar{r}. \tag{2.384}$$

Therefore,

$$\frac{1}{N}\sum_{n=1}^{N}\cos\left(\theta_n - \theta_0^{\mathrm{ML}}\right) = \bar{r}. \tag{2.385}$$

2.56

By the definition,

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}. \tag{2.386}$$

The right hand side can be written as

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\exp\left((a-1)\ln\mu + (b-1)\ln(1-\mu)\right) \tag{2.387}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}.$$

Additionally, by the definition,

$$\mathrm{Gam}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda). \tag{2.388}$$

The right hand side can be written as

$$\frac{b^a}{\Gamma(a)}\exp\left((a-1)\ln\lambda - b\lambda\right). \tag{2.389}$$

Therefore, the natural parameters are given by

$$\boldsymbol{\eta} = \begin{bmatrix} a-1 \\ -b \end{bmatrix}.$$

Finally, for

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\left(m\cos(\theta - \theta_0)\right), \tag{2.390}$$

where

$$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp(m\cos\theta) d\theta, \tag{2.391}$$

the right hand side can be written as

$$\frac{1}{2\pi I_0(m)} \exp(m\cos\theta_0 \cos\theta + m\sin\theta_0 \sin\theta). \tag{2.392}$$

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,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} (\det \boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.393}$$

Therefore,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right), \tag{2.394}$$

where

$$\begin{aligned} h(\mathbf{x}) &= (2\pi)^{-\frac{D}{2}}, \\ g(\boldsymbol{\eta}) &= (\det(-2\boldsymbol{\eta}_2))^{-\frac{1}{2}} \exp\left(\frac{1}{4}\boldsymbol{\eta}_1^\intercal \boldsymbol{\eta}_2^{-1} \boldsymbol{\eta}_1\right), \\ \boldsymbol{\eta} &= \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix}, \\ \mathbf{u}(\mathbf{x}) &= \begin{bmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}^\intercal \end{bmatrix}. \end{aligned}$$

2.58

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

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right). \tag{2.395}$$

Then, taking the first derivative of

$$\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = 1 \tag{2.396}$$

with respect to $\boldsymbol{\eta}$ gives

$$\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right)d\mathbf{x} + g(\boldsymbol{\eta})\int \mathbf{u}(\mathbf{x})h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right)d\mathbf{x} = \mathbf{0}. \tag{2.397}$$

The left hand side can be written as

$$\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} = \frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \mathrm{E}\,\mathbf{u}(\mathbf{x}). \tag{2.398}$$

Therefore,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}. \tag{2.399}$$

Thus,

$$\mathrm{E}\,\mathbf{u}(\mathbf{x}) = -\nabla \ln g(\boldsymbol{\eta}). \tag{2.400}$$

Taking the second derivative with respect to $\boldsymbol{\eta}$ gives

$$\begin{aligned}\nabla\nabla g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right)d\mathbf{x} + 2\nabla g(\boldsymbol{\eta})\int \mathbf{u}(\mathbf{x})^\intercal h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right)d\mathbf{x}\\ + g(\boldsymbol{\eta})\int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\intercal h(\mathbf{x})\exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right)d\mathbf{x} = \mathbf{O}.\end{aligned} \tag{2.401}$$

The left hand side can be written as

$$\begin{aligned}&\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\int p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \frac{2\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})}\int \mathbf{u}(\mathbf{x})^\intercal p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x} + \int \mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\intercal p(\mathbf{x}|\boldsymbol{\eta})d\mathbf{x}\\ =&\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} - 2\,\mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^\intercal + \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^\intercal\right).\end{aligned} \tag{2.402}$$

Therefore,

$$\mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\intercal}\right) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{2\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^{\intercal}}{g^2(\boldsymbol{\eta})}. \tag{2.403}$$

By the definition,

$$\operatorname{cov}\mathbf{u}(\mathbf{x}) = \mathrm{E}\left(\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})^{\intercal}\right) - \mathrm{E}\,\mathbf{u}(\mathbf{x})\,\mathrm{E}\,\mathbf{u}(\mathbf{x})^{\intercal}. \tag{2.404}$$

Thus,

$$\operatorname{cov}\mathbf{u}(\mathbf{x}) = -\frac{\nabla\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \frac{\nabla g(\boldsymbol{\eta})(\nabla g(\boldsymbol{\eta}))^{\intercal}}{g^2(\boldsymbol{\eta})}. \tag{2.405}$$

Hence,

$$\operatorname{cov}\mathbf{u}(\mathbf{x}) = -\nabla\nabla\ln g(\boldsymbol{\eta}). \tag{2.406}$$

2.59

Let

$$p(x|\sigma) = \frac{1}{\sigma}f\left(\frac{x}{\sigma}\right). \tag{2.407}$$

Then

$$\int p(x|\sigma)dx = \frac{1}{\sigma}\int f\left(\frac{x}{\sigma}\right)dx. \tag{2.408}$$

By the transformation

$$x' = \frac{x}{\sigma}, \tag{2.409}$$

the right hand side can be written as

$$\frac{1}{\sigma}\int f(x')\sigma dx' = \int f(x')dx'. \tag{2.410}$$

Therefore, $p(x|\sigma)$ will be normalised if $f(x)$ is normalised.

2.60

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

$$\mathbf{x} \in \mathcal{R}_i \Rightarrow p(\mathbf{x}) = h_i, \tag{2.411}$$

where

$$\int_{\mathcal{R}_i} d\mathbf{x} = \Delta_i. \tag{2.412}$$

Since

$$\int p(\mathbf{x})d\mathbf{x} = 1, \tag{2.413}$$

we have

$$\sum_i h_i \Delta_i = 1. \tag{2.414}$$

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

$$\ln\left(\prod_i {h_i}^{n_i}\right) = \sum_i n_i \ln h_i, \tag{2.415}$$

where

$$\sum_i n_i = N. \tag{2.416}$$

Setting the derivatives of

$$\sum_i n_i \ln h_i + \lambda\left(\sum_i h_i\Delta_i - 1\right) \tag{2.417}$$

with respect to h_i and λ to zero gives

$$\begin{aligned} \frac{n_i}{h_i} + \lambda\Delta_i &= 0, \\ \sum_i h_i\Delta_i - 1 &= 0. \end{aligned} \tag{2.418}$$

Then,

$$\begin{aligned} \lambda &= -N, \\ h_i &= \frac{n_i}{N\Delta_i}. \end{aligned} \tag{2.419}$$

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

$$p(\mathbf{x}) = \frac{K}{NV(\mathbf{x})}, \tag{2.420}$$

where

$$V(\mathbf{x}) = \int_{\|\mathbf{x}'-\mathbf{x}\| \leq \|\mathbf{x}_{(K)}-\mathbf{x}\|} d\mathbf{x}', \tag{2.421}$$

K is a constant and $\mathbf{x}_{(K)}$ is the Kth nearest observation from the point $\mathbf{x}$.

3 Linear Models for Regression

3.1

By the definition,

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}. \tag{3.1}$$

The right hand side can be written as

$$\frac{1 - \exp(-2a)}{1 + \exp(-2a)} = \frac{2}{1 + \exp(-2a)} - 1. \tag{3.2}$$

Therefore,

$$\tanh a = 2\sigma(2a) - 1, \tag{3.3}$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{3.4}$$

Let

$$y(x_n, \mathbf{w}) = w_0 + \sum_{m=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right). \tag{3.5}$$

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

$$w_0 + \sum_{m=1}^{M} w_m \frac{1 + \tanh\left(\frac{x-\mu_m}{2s}\right)}{2} = w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m + \frac{1}{2}\sum_{m=1}^{M} w_m \tanh\left(\frac{x - \mu_m}{2s}\right). \tag{3.6}$$

Therefore, $y(x_n, \mathbf{w})$ is equivalent to

$$y(x_n, \mathbf{u}) = u_0 + \sum_{m=1}^{M} u_m \tanh\left(\frac{x - \mu_m}{2s}\right), \tag{3.7}$$

where

$$\begin{aligned} u_0 &= w_0 + \frac{1}{2}\sum_{m=1}^{M} w_m, \\ u_m &= \frac{1}{2} w_m. \end{aligned} \tag{3.8}$$

3.2 (Incomplete)

Let $\mathbf{\Phi}$ be an $N \times M$ matarix. Then, for any vector $\mathbf{v}$ in N dimensions,

$$\mathbf{\Phi}\left(\mathbf{\Phi}^\intercal \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\intercal \mathbf{v} \tag{3.9}$$

is a projection of $\mathbf{v}$ onto the space spanned by the columns of $\mathbf{\Phi}$?

Additionally, for a vector $\mathbf{t}$ in N dimensions,

$$\left(\mathbf{\Phi}^\intercal \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\intercal \mathbf{t} \tag{3.10}$$

is an orthogonal projection of $\mathbf{t}$ onto the space spanned by the columns of $\mathbf{\Phi}$?

3.3

Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left(t_n - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n)\right)^2 . \tag{3.11}$$

The right hand side can be written as

$$\frac{1}{2} \|\mathbf{t}' - \mathbf{\Phi}' \mathbf{w}\|^2, \tag{3.12}$$

where

$$\mathbf{t}' = \begin{bmatrix} \sqrt{r_1} t_1 \\ \vdots \\ \sqrt{r_N} t_N \end{bmatrix}, \mathbf{\Phi}' = \begin{bmatrix} \sqrt{r_1} \boldsymbol{\phi}(\mathbf{x}_1)^\intercal \\ \vdots \\ \sqrt{r_N} \boldsymbol{\phi}(\mathbf{x}_N)^\intercal \end{bmatrix}.$$

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

$$\mathbf{0} = -\mathbf{\Phi}'^\intercal (\mathbf{t}' - \mathbf{\Phi}' \mathbf{w}). \tag{3.13}$$

Therefore,

$$\operatorname*{argmin}_{\mathbf{w}} E(\mathbf{w}) = \left(\mathbf{\Phi}'^\intercal \mathbf{\Phi}'\right)^{-1} \mathbf{\Phi}'^\intercal \mathbf{t}'. \tag{3.14}$$

3.4 (Incomplete)

Let

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(\mathbf{x}_n, \mathbf{w}) - t_n\right)^2, \tag{3.15}$$

where

$$\begin{aligned} y(\mathbf{x}, \mathbf{w}) &= w_0 + \sum_{m=1}^{M} w_m(x_m + \epsilon_m), \\ p(\epsilon_m) &= \mathcal{N}\left(\epsilon_m | 0, \sigma^2\right). \end{aligned} \tag{3.16}$$

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

$$\mathbf{0} = \sum_{n=1}^{N}\begin{bmatrix} 1 \\ \mathbf{x}_n + \boldsymbol{\epsilon}_n \end{bmatrix}\left(y(\mathbf{x}_n, \mathbf{w}) - t_n\right).$$

The right hand side can be written as

$$\sum_{n=1}^{N}\begin{bmatrix} 1 \\ \mathbf{x}_n \end{bmatrix}\left(y(\mathbf{x}_n, \mathbf{w}) - t_n\right) + \sum_{n=1}^{N}\begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_n \end{bmatrix}\left(y(\mathbf{x}_n, \mathbf{w}) - t_n\right).$$

3.5

Let

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^{\intercal}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2. \tag{3.17}$$

Then, the minimisation of $E(\mathbf{w})$ under the constraint

$$\sum_{m=1}^{M} |w_m|^q \le \eta \tag{3.18}$$

reduces to the minimisation of

$$E(\mathbf{w}) + \lambda\left(\sum_{m=1}^{M} |w_m|^q - \eta\right) \tag{3.19}$$

with respect to $\mathbf{w}$ and λ. Then,

$$\eta = \sum_{m=1}^{M} |w_m^*(\lambda)|^q, \tag{3.20}$$

where

$$\mathbf{w}^*(\lambda) = \underset{\mathbf{w}}{\operatorname{argmin}} \left(E(\mathbf{w}) + \lambda \left(\sum_{m=1}^{M} |w_m|^q - \eta \right) \right). \tag{3.21}$$

3.6

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in D dimensions such that

$$p(\mathbf{t}_n|\mathbf{W}, \mathbf{\Sigma}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{W}), \mathbf{\Sigma}\right), \tag{3.22}$$

where

$$\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^\intercal \boldsymbol{\phi}(\mathbf{x}_n). \tag{3.23}$$

Then,

$$\begin{aligned}
&\ln\left(\prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{W}, \mathbf{\Sigma})\right) \\
&= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(\det\mathbf{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)^\intercal \mathbf{\Sigma}^{-1}\left(\mathbf{t}_n - \mathbf{W}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right).
\end{aligned} \tag{3.24}$$

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

$$\begin{aligned}
\mathbf{O} &= -\frac{1}{2}\left(\mathbf{\Sigma}^{-1} + \left(\mathbf{\Sigma}^{-1}\right)^\intercal\right)\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\boldsymbol{\phi}(\mathbf{x}_n)\right)^\intercal, \\
\mathbf{O} &= -\frac{N}{2}\left(\mathbf{\Sigma}^{-1}\right)^\intercal + \frac{1}{2}\left(\mathbf{\Sigma}^{-1}\right)^2\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)^\intercal.
\end{aligned} \tag{3.25}$$

Therefore,

$$\begin{aligned}
\mathbf{W}_{\mathrm{ML}} &= \left(\mathbf{\Phi}^\intercal\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^\intercal\mathbf{t}, \\
\mathbf{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)^\intercal,
\end{aligned} \tag{3.26}$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\intercal \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\intercal \end{bmatrix}.$$

3.7

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \tag{3.27}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.28}$$

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

$$\begin{aligned} &-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^\intercal \mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ =&-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^\intercal(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}) - \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^\intercal \mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0), \end{aligned} \tag{3.29}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^\intercal \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^\intercal \end{bmatrix}.$$

The right hand side can be written as

$$-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\intercal \mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) + \frac{1}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\mathbf{t}^\intercal\mathbf{t} - \frac{1}{2}\mathbf{m}_0^\intercal \mathbf{S}_0^{-1}\mathbf{m}_0, \tag{3.30}$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\intercal\mathbf{t}\right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}. \end{aligned} \tag{3.31}$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N). \tag{3.32}$$

3.8

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \tag{3.33}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.34}$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta \mathbf{\Phi}_N^\intercal \mathbf{t}_N \right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta \mathbf{\Phi}_N^\intercal \mathbf{\Phi}_N. \end{aligned} \tag{3.35}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \tag{3.36}$$

The right hand side can be written as

$$p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \tag{3.37}$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N). \tag{3.38}$$

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

$$\begin{aligned} &-\frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2 - \frac{1}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\intercal \mathbf{S}_N^{-1} \left(\mathbf{w} - \mathbf{m}_N\right) \\ =&-\frac{1}{2}\left(\mathbf{w} - \boldsymbol{\mu}_{N+1}\right)^\intercal \mathbf{\Lambda}_{N+1} \left(\mathbf{w} - \boldsymbol{\mu}_{N+1}\right) + \frac{1}{2}\boldsymbol{\mu}_{N+1}^\intercal \mathbf{\Lambda}_{N+1} \boldsymbol{\mu}_{N+1} \\ &-\frac{1}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1} \mathbf{m}_N - \frac{\beta}{2} t_{N+1}^2, \end{aligned} \tag{3.39}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{N+1} &= \mathbf{\Lambda}_{N+1}^{-1} \left(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta t_{N+1} \boldsymbol{\phi}(\mathbf{x}_{N+1}) \right), \\ \mathbf{\Lambda}_{N+1} &= \mathbf{S}_N^{-1} + \beta \boldsymbol{\phi}(\mathbf{x}_{N+1}) \boldsymbol{\phi}(\mathbf{x}_{N+1})^\intercal. \end{aligned} \tag{3.40}$$

Therefore,

$$\begin{aligned} \boldsymbol{\mu}_{N+1} &= \mathbf{m}_{N+1}, \\ \mathbf{\Lambda}_{N+1} &= \mathbf{S}_{N+1}^{-1}. \end{aligned} \tag{3.41}$$

Thus,

$$p(\mathbf{w}|\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}). \tag{3.42}$$

3.9 (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \tag{3.43}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.44}$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^\intercal \mathbf{t}_N\right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^\intercal\boldsymbol{\Phi}_N. \end{aligned} \tag{3.45}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(\mathbf{t}_{N+1}) = p(\mathbf{t}_{N+1}|\mathbf{w})p(\mathbf{w}). \tag{3.46}$$

The right hand side can be written as

$$p(t_{N+1}|\mathbf{w})p(\mathbf{t}_N|\mathbf{w})p(\mathbf{w}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N)p(\mathbf{t}_N). \tag{3.47}$$

Therefore,

$$p(\mathbf{w}|\mathbf{t}_{N+1})p(t_{N+1}) = p(t_{N+1}|\mathbf{w})p(\mathbf{w}|\mathbf{t}_N). \tag{3.48}$$

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

$$-\frac{\beta}{2}\left(t_{N+1} - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2 - \frac{1}{2}\left(\mathbf{w} - \mathbf{m}_N\right)^\intercal \mathbf{S}_N^{-1}\left(\mathbf{w} - \mathbf{m}_N\right). \tag{3.49}$$

3.10

Let t be a variable such that

$$\begin{aligned} p(t|\mathbf{w}) &= \mathcal{N}\left(t|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \tag{3.50}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.51}$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\intercal \mathbf{t}\right), \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}. \end{aligned} \tag{3.52}$$

By marginalisation,

$$p(t|\mathbf{t}) = \int p(t|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{3.53}$$

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

$$-\frac{\beta}{2}\left(t - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2 - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.54}$$

It can be written as

$$-\frac{1}{2}\begin{bmatrix}\mathbf{w}\\ t\end{bmatrix}^\intercal \begin{bmatrix}\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\beta\boldsymbol{\phi}(\mathbf{x})\\ -\beta\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ t\end{bmatrix} + \begin{bmatrix}\mathbf{w}\\ t\end{bmatrix}^\intercal \begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\ 0\end{bmatrix} - \frac{1}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\beta\boldsymbol{\phi}(\mathbf{x})\\ -\beta\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{S}_N & \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\\ \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N & \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Therefore,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\boldsymbol{\phi}(\mathbf{x})\\ \beta\boldsymbol{\phi}(\mathbf{x})^\intercal & \beta\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\ 0\end{bmatrix} = \begin{bmatrix}\mathbf{m}_N\\ \mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Thus,

$$p(t|\mathbf{t}) = \mathcal{N}\left(t|\mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right), \tag{3.55}$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}). \tag{3.56}$$

3.11

Let t be a variable such that

$$\begin{aligned} p(t|\mathbf{w}) &= \mathcal{N}\left(t|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right),\\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \end{aligned} \tag{3.57}$$

Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.58}$$

where

$$
\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}_N^\intercal \mathbf{t}_N\right),\\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}_N^\intercal\boldsymbol{\Phi}_N.
\end{aligned} \tag{3.59}
$$

Then, by 3.10,

$$
p(t|\mathbf{t}) = \mathcal{N}\left(t \mid \mathbf{m}_N^\intercal\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right), \tag{3.60}
$$

where

$$
\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}). \tag{3.61}
$$

Then,

$$
\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\intercal\left(\mathbf{S}_N - \mathbf{S}_{N+1}\right)\boldsymbol{\phi}(\mathbf{x}). \tag{3.62}
$$

By the expression of $\mathbf{S}_N$ above,

$$
\mathbf{S}_{N+1} = \left(\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^\intercal\right)^{-1}. \tag{3.63}
$$

By the identity

$$
\left(\mathbf{M} + \mathbf{v}\mathbf{v}^\intercal\right)^{-1} = \mathbf{M}^{-1} - \frac{\left(\mathbf{M}^{-1}\mathbf{v}\right)\left(\mathbf{v}^\intercal\mathbf{M}^{-1}\right)}{1 + \mathbf{v}^\intercal\mathbf{M}^{-1}\mathbf{v}}, \tag{3.64}
$$

the right hand side can be written as

$$
\mathbf{S}_N - \frac{\beta\left(\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)\left(\boldsymbol{\phi}(\mathbf{x}_{N+1})^\intercal\mathbf{S}_N\right)}{1 + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})^\intercal}. \tag{3.65}
$$

Therefore,

$$
\boldsymbol{\phi}(\mathbf{x})^\intercal(\mathbf{S}_N - \mathbf{S}_{N+1})\boldsymbol{\phi}(\mathbf{x}) = \frac{\beta\left(\boldsymbol{\phi}(\mathbf{x})^\intercal\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})\right)^2}{1 + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_{N+1})^\intercal}. \tag{3.66}
$$

Thus,

$$
\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}). \tag{3.67}
$$

3.12

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w}, \beta) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right),\\
p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0),
\end{aligned} \tag{3.68}
$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{w}, \beta|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}, \beta). \tag{3.69}$$

The logarithm of the right hand side except the terms independent of $\mathbf{t}$, $\mathbf{w}$ and β can be written as

$$\begin{aligned}
&-\frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 - \frac{M}{2}\ln\beta^{-1} - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^\intercal\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\\
&+(a_0-1)\ln\beta - b_0\beta\\
=&-\frac{M}{2}\ln\beta - \frac{\beta}{2}\mathbf{w}^\intercal\left(\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^\intercal\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\intercal\mathbf{t}\right) - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0\\
&+\left(a_0 + \frac{N}{2} - 1\right)\ln\beta - b_0\beta.
\end{aligned} \tag{3.70}$$

The right hand side can be written as

$$-\frac{M}{2}\ln\beta - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^\intercal\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) + (a_N-1)\ln\beta - b_N\beta, \tag{3.71}$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\intercal\mathbf{t}\right),\\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi},\\
a_N &= a_0 + \frac{N}{2},\\
b_N &= b_0 + \frac{1}{2}\|\mathbf{t}\|^2 + \frac{1}{2}\mathbf{m}_0^\intercal\mathbf{S}_0\mathbf{m}_0 - \frac{1}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N.
\end{aligned} \tag{3.72}$$

Therefore,

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N, b_N). \tag{3.73}$$

Substituting it to the result of the Bayes' theorem above, we have

$$p(\mathbf{t}) = \frac{\mathcal{N}\left(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}\right)\mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0, b_0)}{\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N, b_N)}. \tag{3.74}$$

The logarithm of the right hand side can be written as

$$
\begin{aligned}
&-\frac{N}{2}\ln(2\pi)-\frac{N}{2}\ln\beta^{-1}-\frac{\beta}{2}\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)^{\intercal}\left(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}\right)\\
&-\frac{M}{2}\ln(2\pi)-\frac{M}{2}\ln\beta^{-1}-\frac{1}{2}\det\mathbf{S}_0-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\intercal}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\\
&+a_0\ln b_0-\ln\Gamma(a_0)+(a_0-1)\ln\beta-b_0\beta\\
&+\frac{M}{2}\ln(2\pi)+\frac{M}{2}\ln\beta^{-1}+\frac{1}{2}\det\mathbf{S}_N+\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^{\intercal}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_0)\\
&-a_N\ln b_N+\ln\Gamma(a_N)-(a_N-1)\ln\beta+b_N\beta\\
=&-\frac{N}{2}\ln(2\pi)-\frac{1}{2}\det\mathbf{S}_0+a_0\ln b_0-\ln\Gamma(a_0)+\frac{1}{2}\det\mathbf{S}_N-a_N\ln b_N+\ln\Gamma(a_N).
\end{aligned}
\tag{3.75}
$$

Therefore,

$$
p(\mathbf{t})=(2\pi)^{-\frac{N}{2}}\left(\frac{\det\mathbf{S}_N}{\det\mathbf{S}_0}\right)^{\frac{1}{2}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{b_0^{a_0}}{b_N^{a_N}}. \tag{3.76}
$$

3.13

Let $t_1,\cdots,t_N$ be variables such that

$$
\begin{aligned}
p(t_n|\mathbf{w},\beta)&=\mathcal{N}\left(t_n|\mathbf{w}^{\intercal}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}\right),\\
p(\mathbf{w},\beta)&=\mathcal{N}\left(\mathbf{w}|\mathbf{m}_0,\beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}(\beta|a_0,b_0),
\end{aligned}
\tag{3.77}
$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. Then, by 3.12,

$$
p(\mathbf{w},\beta|\mathbf{t})=\mathcal{N}\left(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}(\beta|a_N,b_N), \tag{3.78}
$$

where

$$
\begin{aligned}
\mathbf{m}_N&=\mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\boldsymbol{\Phi}^{\intercal}\mathbf{t}\right),\\
\mathbf{S}_N^{-1}&=\mathbf{S}_0^{-1}+\boldsymbol{\Phi}^{\intercal}\boldsymbol{\Phi},\\
a_N&=a_0+\frac{N}{2},\\
b_N&=b_0+\frac{1}{2}\|\mathbf{t}\|^2+\frac{1}{2}\mathbf{m}_0^{\intercal}\mathbf{S}_0\mathbf{m}_0-\frac{1}{2}\mathbf{m}_N^{\intercal}\mathbf{S}_N^{-1}\mathbf{m}_N.
\end{aligned}
\tag{3.79}
$$

By marginalisation,

$$
p(t|\mathbf{t})=\int\int p(t|\mathbf{w},\beta)p(\mathbf{w},\beta|\mathbf{t})d\mathbf{w}d\beta. \tag{3.80}
$$

The right hand side can be written as

$$\int \left(\int \mathcal{N}\left(t|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\right) \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N\right) d\mathbf{w} \right) \mathrm{Gam}(\beta|a_N, b_N) d\beta. \tag{3.81}$$

The logarithm of the integrand with respect to $\mathbf{w}$ except the terms indepn-dent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\left(t - \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2 - \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.82}$$

It can be written as

$$-\frac{\beta}{2}\begin{bmatrix}\mathbf{w}\\ t\end{bmatrix}^\intercal \begin{bmatrix}\mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x})\\ -\boldsymbol{\phi}(\mathbf{x})^\intercal & 1\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ t\end{bmatrix} + \beta\begin{bmatrix}\mathbf{w}\\ t\end{bmatrix}^\intercal \begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\ 0\end{bmatrix} - \frac{\beta}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N.$$

By 2.24,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x})\\ -\boldsymbol{\phi}(\mathbf{x})^\intercal & 1\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{S}_N & \mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\\ \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N & 1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Then,

$$\begin{bmatrix}\mathbf{S}_N^{-1} + \boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^\intercal & -\boldsymbol{\phi}(\mathbf{x})\\ -\boldsymbol{\phi}(\mathbf{x})^\intercal & 1\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{S}_N^{-1}\mathbf{m}_N\\ 0\end{bmatrix} = \begin{bmatrix}\mathbf{m}_N\\ \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\end{bmatrix}.$$

Therefore, the integral with respect to $\mathbf{w}$ can be written as

$$\mathcal{N}\left(t|\mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)\right). \tag{3.83}$$

Then, the logarithm of the integrand with respect to β except the terms independent of β can be written as

$$\begin{aligned} &-\frac{1}{2}\ln\beta^{-1} - \frac{\beta}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)}\left(t - \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2 + (a_N - 1)\ln\beta - b_N\beta \\ &= \left(a_N + \frac{1}{2} - 1\right)\ln\beta - \left(b_N + \frac{\left(t - \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)}\right)\beta. \end{aligned} \tag{3.84}$$

Therefore, the integral with respect to β except the terms independent of t can be written as

$$\left(b_N + \frac{\left(t - \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x})\right)^2}{2\left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)}\right)^{-a_N - \frac{1}{2}}. \tag{3.85}$$

Thus,

$$p(t|\mathbf{x}, \mathbf{t}) = \mathrm{St}(t|\mu, \lambda, \nu), \tag{3.86}$$

where

$$\begin{aligned} \mu &= \mathbf{m}_N^\intercal \boldsymbol{\phi}(\mathbf{x}), \\ \lambda &= \frac{a_N}{b_N} \left(1 + \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right)^{-1}, \\ \nu &= 2a_N. \end{aligned} \tag{3.87}$$

3.14 (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.88}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. Then, by 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.89}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\intercal \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}. \end{aligned} \tag{3.90}$$

Let

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}). \tag{3.91}$$

Then,

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n, \tag{3.92}$$

where

$$k(\mathbf{x}, \mathbf{x}') = \beta \boldsymbol{\phi}(\mathbf{x})^\intercal \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}'). \tag{3.93}$$

Let us suppose that $\phi_j(\mathbf{x})$ are linearly independent, $N > M$ and

$$\phi_0(\mathbf{x}) = 1. \tag{3.94}$$

Then, we can construct a new basis set $\psi_j(\mathbf{x})$ such that

$$\boldsymbol{\Psi}^\intercal \boldsymbol{\Psi} = \mathbf{I}? \tag{3.95}$$

$$\sum_{n=1}^{N} \psi_j(\mathbf{x}_n)\psi_k(\mathbf{x}_n) = I_{jk}? \tag{3.96}$$

where

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\psi}(\mathbf{x}_1)^\intercal \\ \vdots \\ \boldsymbol{\psi}(\mathbf{x}_N)^\intercal \end{bmatrix}$$

and

$$\psi_0(\mathbf{x}) = 1. \tag{3.97}$$

Under the basis set, if $\alpha = 0$, then

$$\mathbf{S}_N^{-1} = \beta \mathbf{I}, \tag{3.98}$$

so that

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\intercal \boldsymbol{\psi}(\mathbf{x}'). \tag{3.99}$$

Then,

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^{N} \sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) = 1? \tag{3.100}$$

3.15

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.101}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.102}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\intercal \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}. \end{aligned} \tag{3.103}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) + \frac{1}{2} \ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.104}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\intercal \mathbf{m}_N. \tag{3.105}$$

By 3.22, setting the derivatives of $\ln p(\mathbf{t})$ with respect to α and β to zero gives

$$\begin{aligned} \alpha &= \frac{\gamma}{\mathbf{m}_N^\intercal \mathbf{m}_N}, \\ \beta &= \frac{N - \gamma}{\|\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N\|^2}, \end{aligned} \tag{3.106}$$

where

$$\gamma = \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m} \tag{3.107}$$

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

$$E(\mathbf{m}_N) = \frac{N}{2}. \tag{3.108}$$

3.16

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.109}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.110}$$

Integrating both sides with respect to $\mathbf{w}$ gives

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{3.111}$$

The logarithm of the integrand of the right hand side except the terms independent of $\mathbf{w}$ can be written as

$$-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n))^2 - \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w} = -\frac{1}{2}\begin{bmatrix}\mathbf{w}\\ \mathbf{t}\end{bmatrix}^\intercal \begin{bmatrix}\alpha\mathbf{I} + \beta\mathbf{\Phi}^\intercal\mathbf{\Phi} & -\beta\mathbf{\Phi}^\intercal \\ -\beta\mathbf{\Phi} & \beta\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ \mathbf{t}\end{bmatrix}.$$

By 2.24,

$$\begin{bmatrix} \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi} & -\beta\boldsymbol{\Phi}^\intercal \\ -\beta\boldsymbol{\Phi} & \beta\mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \alpha^{-1}\mathbf{I} & \alpha^{-1}\boldsymbol{\Phi}^\intercal \\ \alpha^{-1}\boldsymbol{\Phi} & \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \beta^{-1}\mathbf{I} \end{bmatrix}.$$

Therefore,

$$p(\mathbf{t}) = \mathcal{N}\left(\mathbf{t}|\mathbf{0}, \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \beta^{-1}\mathbf{I}\right). \tag{3.112}$$

3.17

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.113}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By the Bayes' theorem,

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{3.114}$$

Then,

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w})d\mathbf{w}. \tag{3.115}$$

The logarithm of the integrand of the right hand side can be written as

$$-\frac{N}{2}\ln\left(2\pi\beta^{-1}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 - \frac{M}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln\left(\det\left(\alpha^{-1}\mathbf{I}\right)\right) - \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w}. \tag{3.116}$$

Therefore,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int\exp\left(-E(\mathbf{w})\right)d\mathbf{w}, \tag{3.117}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w}. \tag{3.118}$$

3.18

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.119}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.120}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\intercal \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}. \end{aligned} \tag{3.121}$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}} \int \exp\left(-E(\mathbf{w})\right) d\mathbf{w}, \tag{3.122}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\intercal \mathbf{w}. \tag{3.123}$$

The first term of the definition of $E(\mathbf{w})$ can be written as

$$\begin{aligned} &\frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N - \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N)\|^2 \\ =&\frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 - \beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^\intercal \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \frac{\beta}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N). \end{aligned} \tag{3.124}$$

Similarly, the second term can be written as

$$\begin{aligned} &\frac{\alpha}{2}(\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N)^\intercal (\mathbf{w} - \mathbf{m}_N + \mathbf{m}_N) \\ =&\frac{\alpha}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal (\mathbf{w} - \mathbf{m}_N) + \alpha \mathbf{m}_N^\intercal (\mathbf{w} - \mathbf{m}_N) + \frac{\alpha}{2}\mathbf{m}_N^\intercal \mathbf{m}_N. \end{aligned} \tag{3.125}$$

Here,

$$\begin{aligned} &-\beta(\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N)^\intercal \boldsymbol{\Phi}(\mathbf{w} - \mathbf{m}_N) + \alpha \mathbf{m}_N^\intercal (\mathbf{w} - \mathbf{m}_N) \\ =&\left(-\beta \boldsymbol{\Phi}^\intercal \mathbf{t} + \beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}\mathbf{m}_N + \alpha \mathbf{m}_N\right)^\intercal (\mathbf{w} - \mathbf{m}_N). \end{aligned} \tag{3.126}$$

By the definitions of $\mathbf{m}_N$ and $\mathbf{S}_N$ above, the right hand can be written as

$$\left(-\beta \boldsymbol{\Phi}^\intercal \mathbf{t} + \mathbf{S}_N^{-1}\mathbf{m}_N\right)^\intercal (\mathbf{w} - \mathbf{m}_N) = 0. \tag{3.127}$$

Therefore,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal \mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.128}$$

3.19

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.129}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.130}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\intercal\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}. \end{aligned} \tag{3.131}$$

By 3.17,

$$p(\mathbf{t}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int \exp\left(-E(\mathbf{w})\right) d\mathbf{w}, \tag{3.132}$$

where

$$E(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\intercal\mathbf{w}. \tag{3.133}$$

By 3.18,

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N). \tag{3.134}$$

Therefore, the integral in the expression above of $p(\mathbf{t})$ can be written as

$$\begin{aligned} &\exp\left(-E(\mathbf{m}_N)\right)\int \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\intercal\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N)\right) d\mathbf{w} \\ =&(2\pi)^{\frac{M}{2}}(\det\mathbf{S}_N)^{\frac{1}{2}}\exp\left(-E(\mathbf{m}_N)\right). \end{aligned} \tag{3.135}$$

Thus,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N). \tag{3.136}$$

3.20

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.137}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.138}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\intercal \mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}. \end{aligned} \tag{3.139}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.140}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\intercal \mathbf{m}_N. \tag{3.141}$$

Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ be eigenvectors of $\beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}$ such that

$$\beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi} \mathbf{u}_m = \lambda_m \mathbf{u}_m. \tag{3.142}$$

Then,

$$\mathbf{S}_N^{-1}\mathbf{u}_m = (\alpha + \lambda_m)\mathbf{u}_m, \tag{3.143}$$

so that

$$\det \mathbf{S}_N = \prod_{m=1}^{M} \frac{1}{\alpha + \lambda_m}. \tag{3.144}$$

Therefore, setting the derivative of $\ln p(\mathbf{t}|\alpha, \beta)$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\sum_{m=1}^{M} \frac{1}{\alpha + \lambda_m} - \frac{1}{2}\mathbf{m}_N^\intercal \mathbf{m}_N. \tag{3.145}$$

Multiplying both sides by 2α gives

$$\alpha \mathbf{m}_N^\intercal \mathbf{m}_N = M - \sum_{m=1}^{M} \frac{\alpha}{\alpha + \lambda_m}. \tag{3.146}$$

The right hand side can be written as

$$\sum_{m=1}^{M}\left(1 - \frac{\alpha}{\alpha + \lambda_m}\right) = \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m}. \tag{3.147}$$

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\intercal}\mathbf{m}_N}, \tag{3.148}$$

where

$$\gamma = \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m}. \tag{3.149}$$

3.21

(a)

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

$$\mathbf{\Sigma}\mathbf{u}_m = \lambda_m \mathbf{u}_m, \tag{3.150}$$

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

$$\begin{aligned} \mathbf{\Lambda} &= \operatorname{diag}(\lambda_1, \cdots, \lambda_M), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_M]. \end{aligned} \tag{3.151}$$

By 2.19,

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\intercal}, \\ \mathbf{U}^{\intercal}\mathbf{U} &= \mathbf{I}. \end{aligned} \tag{3.152}$$

Therefore,

$$\det \mathbf{\Sigma} = \prod_{m=1}^{M} \lambda_m, \tag{3.153}$$

so that

$$\ln(\det \mathbf{\Sigma}) = \sum_{m=1}^{M} \ln \lambda_i. \tag{3.154}$$

Then,

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{\Sigma}) = \sum_{m=1}^{M} \frac{\partial \lambda_m}{\partial \alpha} \frac{1}{\lambda_m}. \tag{3.155}$$

Therefore,

$$\frac{\partial}{\partial \alpha} \ln(\det \mathbf{\Sigma}) = \operatorname{tr}\left(\mathbf{\Lambda}^{-1} \frac{\partial \mathbf{\Lambda}}{\partial \alpha}\right). \tag{3.156}$$

The right hand side can be written as

$$\operatorname{tr}\left(\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\intercal}\frac{\partial \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\intercal}}{\partial\alpha}\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\alpha}\right). \tag{3.157}$$

Therefore,

$$\frac{\partial}{\partial\alpha}\ln(\det\boldsymbol{\Sigma}) = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\alpha}\right). \tag{3.158}$$

(b)

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^{\intercal}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.159}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.160}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^{\intercal}\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\intercal}\boldsymbol{\Phi}. \end{aligned} \tag{3.161}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.162}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\intercal}\mathbf{m}_N. \tag{3.163}$$

By 3.21(a),

$$\frac{\partial}{\partial\alpha}\ln\left(\det\mathbf{S}_N^{-1}\right) = \operatorname{tr}\left(\mathbf{S}_N\right). \tag{3.164}$$

The right hand side can be written as

$$\sum_{m=1}^{M}\frac{1}{\alpha+\lambda_m}, \tag{3.165}$$

where $\lambda_1, \cdots, \lambda_M$ are eigenvalues of $\beta\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to α to zero gives

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\sum_{m=1}^{M}\frac{1}{\alpha+\lambda_m} - \frac{1}{2}\mathbf{m}_N^\intercal\mathbf{m}_N, \tag{3.166}$$

Thus,

$$\alpha = \frac{\gamma}{\mathbf{m}_N^\intercal\mathbf{m}_N}, \tag{3.167}$$

where

$$\gamma = \sum_{m=1}^{M}\frac{\lambda_m}{\alpha+\lambda_m}. \tag{3.168}$$

3.22

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n|\mathbf{w}) &= \mathcal{N}\left(t_n|\mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \end{aligned} \tag{3.169}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By 3.7,

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{3.170}$$

where

$$\begin{aligned} \mathbf{m}_N &= \beta\mathbf{S}_N\boldsymbol{\Phi}^\intercal\mathbf{t}, \\ \mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}. \end{aligned} \tag{3.171}$$

By 3.19,

$$\ln p(\mathbf{t}) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\mathbf{S}_N) - E(\mathbf{m}_N), \tag{3.172}$$

where

$$E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\intercal\mathbf{m}_N. \tag{3.173}$$

By 3.21(a),

$$\frac{\partial}{\partial\beta}\ln\!\left(\det\mathbf{S}_N^{-1}\right) = \operatorname{tr}\left(\mathbf{S}_N\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\right). \tag{3.174}$$

Since

$$\mathbf{S}_N \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi} = \frac{1}{\beta} \left(\mathbf{I} - \alpha \mathbf{S}_N \right), \tag{3.175}$$

the right hand side can be written as

$$\frac{1}{\beta} \left(M - \alpha \sum_{m=1}^{M} \frac{1}{\alpha + \lambda_m} \right) = \frac{1}{\beta} \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m}, \tag{3.176}$$

where $\lambda_1, \cdots, \lambda_M$ are eigenvalues of $\beta \boldsymbol{\Phi}^\intercal \boldsymbol{\Phi}$. Therefore, setting the derivative of $\ln p(\mathbf{t})$ with respect to β to zero gives

$$0 = \frac{N}{2\beta} - \frac{1}{2\beta} \sum_{m=1}^{M} \frac{\lambda_i}{\alpha + \lambda_m} - \frac{1}{2} \|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2. \tag{3.177}$$

Thus,

$$\beta = \frac{N - \gamma}{\|\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N\|^2}, \tag{3.178}$$

where

$$\gamma = \sum_{m=1}^{M} \frac{\lambda_m}{\alpha + \lambda_m}. \tag{3.179}$$

3.23

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} p(t_n | \mathbf{w}, \beta) &= \mathcal{N}\left(t_n | \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \\ p(\mathbf{w}, \beta) &= \mathcal{N}\left(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0\right) \mathrm{Gam}(\beta | a_0, b_0), \end{aligned} \tag{3.180}$$

where $\mathbf{w}$ and $\boldsymbol{\phi}$ are vectors in M dimensions. By marginalisation,

$$p(\mathbf{t}) = \int \int p(\mathbf{t} | \mathbf{w}, \beta) p(\mathbf{w}, \beta) d\mathbf{w} d\beta. \tag{3.181}$$

The right hand side can be written as

$$\int \left(\int \left(\prod_{n=1}^{N} \mathcal{N}\left(t_n | \mathbf{w}^\intercal \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right) \right) \mathcal{N}\left(\mathbf{w} | \mathbf{m}_0, \beta^{-1} \mathbf{S}_0\right) d\mathbf{w} \right) \mathrm{Gam}(\beta | a_0, b_0) d\beta. \tag{3.182}$$

The logarithm of the integrand with respect to $\mathbf{w}$ can be written as

$$\begin{aligned}
&-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\beta^{-1} - \frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\intercal\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 \\
&-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\det\bigl(\beta^{-1}\mathbf{S}_0\bigr) - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^\intercal\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\
=&-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) \\
&-\frac{\beta}{2}\mathbf{w}^\intercal\left(\mathbf{S}_0^{-1}+\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\right)\mathbf{w} + \beta\mathbf{w}^\intercal\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\intercal\mathbf{t}\right) - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0.
\end{aligned} \tag{3.183}$$

The right hand side can be written as

$$\begin{aligned}
&-\frac{N+M}{2}\ln(2\pi) + \frac{N+M}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) \\
&-\frac{\beta}{2}\left(\mathbf{w}-\mathbf{m}_N\right)^\intercal\mathbf{S}_N^{-1}\left(\mathbf{w}-\mathbf{m}_N\right) + \frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0,
\end{aligned} \tag{3.184}$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \boldsymbol{\Phi}^\intercal\mathbf{t}\right), \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}.
\end{aligned} \tag{3.185}$$

Therefore, the logarithm of the integral with respect to $\mathbf{w}$ can be written as

$$\begin{aligned}
&-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}_N) \\
&+\frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0.
\end{aligned} \tag{3.186}$$

Then, the logarithm of the integrand with respect to β can be written as

$$\begin{aligned}
&-\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}_N) \\
&+\frac{\beta}{2}\mathbf{m}_N^\intercal\mathbf{S}_N^{-1}\mathbf{m}_N - \frac{\beta}{2}\|\mathbf{t}\|^2 - \frac{\beta}{2}\mathbf{m}_0^\intercal\mathbf{S}_0^{-1}\mathbf{m}_0 \\
&-\ln\Gamma(a_0) + a_0\ln b_0 + (a_0-1)\ln\beta - b_0\beta \\
=&-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det\mathbf{S}_0) + \frac{1}{2}\ln(\det\mathbf{S}_N) \\
&-\ln\Gamma(a_0) + a_0\ln b_0 + (a_N-1)\ln\beta - b_N\beta,
\end{aligned} \tag{3.187}$$

where

$$
\begin{aligned}
a_N &= a_0 + \frac{N}{2}, \\
b_N &= b_0 + \frac{\beta}{2}\|\mathbf{t}\|^2 + \frac{\beta}{2}\mathbf{m}_0^\intercal \mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{\beta}{2}\mathbf{m}_N^\intercal \mathbf{S}_N^{-1}\mathbf{m}_N.
\end{aligned}
\tag{3.188}
$$

Therefore, the logarithm of the integral with respect to β can be written as

$$
-\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{S}_0) + \frac{1}{2}\ln(\det \mathbf{S}_N) - \ln\Gamma(a_0) + a_0 \ln b_0 + \ln\Gamma(a_N) - a_N \ln b_N. \tag{3.189}
$$

Thus,

$$
p(\mathbf{t}) = (2\pi)^{-\frac{N}{2}} \left(\frac{\det \mathbf{S}_N}{\det \mathbf{S}_0}\right)^{\frac{1}{2}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{{b_0}^{a_0}}{{b_N}^{a_N}}. \tag{3.190}
$$

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

$$\begin{aligned}\mathbf{x} &= \sum_{m=1}^{M} \alpha_m \mathbf{x}_m, \\ \mathbf{y} &= \sum_{n=1}^{N} \beta_n \mathbf{y}_n,\end{aligned} \tag{4.1}$$

where

$$\begin{gathered}\sum_{m=1}^{M} \alpha_m = \sum_{n=1}^{N} \beta_n = 1, \\ \alpha_m \geq 0, \beta_n \geq 0.\end{gathered} \tag{4.2}$$

Let us assume that $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ below are subject to the constraints above.

If the convex hulls intersect, then there exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$\sum_{m=1}^{M} \alpha_m \mathbf{x}_m = \sum_{n=1}^{N} \beta_n \mathbf{y}_n. \tag{4.3}$$

Then,

$$\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_m + w_0\right) = \hat{\mathbf{w}}^{\mathsf{T}} \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0 \sum_{m} \alpha_m, \tag{4.4}$$

for any $\hat{\mathbf{w}}$ and w_0. The right hand side can be written as

$$\hat{\mathbf{w}}^{\mathsf{T}} \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n = \sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^{\mathsf{T}} \mathbf{y}_n + w_0\right). \tag{4.5}$$

Therefore, there do not exist $\hat{\mathbf{w}}$ and w_0 such that

$$\begin{aligned}\hat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_m + w_0 &> 0, \\ \hat{\mathbf{w}}^{\mathsf{T}} \mathbf{y}_n + w_0 &< 0.\end{aligned} \tag{4.6}$$

Conversely, if there exist $\hat{\mathbf{w}}$ and w_0 such that

$$\begin{aligned}\hat{\mathbf{w}}^\intercal \mathbf{x}_m + w_0 > 0,\\ \hat{\mathbf{w}}^\intercal \mathbf{y}_n + w_0 < 0,\end{aligned} \tag{4.7}$$

then

$$\begin{aligned}\sum_{m=1}^{M} \alpha_m \left(\hat{\mathbf{w}}^\intercal \mathbf{x}_m + w_0\right) > 0,\\ \sum_{n=1}^{N} \beta_n \left(\hat{\mathbf{w}}^\intercal \mathbf{y}_n + w_0\right) < 0.\end{aligned} \tag{4.8}$$

The left hand sides can be written as

$$\begin{aligned}\hat{\mathbf{w}}^\intercal \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0 \sum_{m=1}^{M} \alpha_m = \hat{\mathbf{w}}^\intercal \sum_{m=1}^{M} \alpha_m \mathbf{x}_m + w_0,\\ \hat{\mathbf{w}}^\intercal \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0 \sum_{n=1}^{N} \beta_n = \hat{\mathbf{w}}^\intercal \sum_{n=1}^{N} \beta_n \mathbf{y}_n + w_0.\end{aligned} \tag{4.9}$$

Therefore, there do not exist $\alpha_1, \cdots, \alpha_M$ and $\beta_1, \cdots, \beta_N$ such that

$$\sum_{m=1}^{M} \alpha_m \mathbf{x}_m = \sum_{n=1}^{N} \beta_n \mathbf{y}_n. \tag{4.10}$$

Thus, the convex hulls do not intersect.

4.2 (Incomplete)

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ and $\mathbf{w}_1, \cdots, \mathbf{w}_K$ are variables in M dimensions and $\mathbf{t}_1, \cdots, \mathbf{t}_N$ are ones in K dimensions. Let

$$E(\tilde{\mathbf{W}}) = \frac{1}{2} \operatorname{tr} \left((\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^\intercal (\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}) \right), \tag{4.11}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_1^\intercal \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\intercal \end{bmatrix},$$

$$\tilde{\mathbf{W}} = \begin{bmatrix} w_{10} & \cdots & w_{K0} \\ \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^\intercal \\ \vdots \\ \mathbf{t}_N^\intercal \end{bmatrix}.$$

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

$$\mathbf{O} = \tilde{\mathbf{X}}^\intercal(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T}). \tag{4.12}$$

Therefore,

$$\operatorname*{argmin}_{\tilde{\mathbf{W}}} E(\tilde{\mathbf{W}}) = \left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^\intercal\mathbf{T}. \tag{4.13}$$

Let $\tilde{\mathbf{W}}^*$ denote the least-square solution above. Then,

$$(\tilde{\mathbf{W}}^*)^\intercal\tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{T}^\intercal\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}} - \mathbf{t}_n, \tag{4.14}$$

where $\tilde{\mathbf{x}}$ is a vector in $M+1$ dimensions whose first element is 1. The right hand side can be written as

$$\mathbf{T}^\intercal\left(\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{x}} - \mathbf{v}_n\right) = \mathbf{0}? \tag{4.15}$$

where $\mathbf{v}_n$ is a vector in N dimensions whose n th element is 1 and other elements are zero. Therefore,

$$(\tilde{\mathbf{W}}^*)^\intercal\tilde{\mathbf{x}} - \mathbf{t}_n = \mathbf{0}. \tag{4.16}$$

Thus, if

$$\mathbf{a}^\intercal\mathbf{t}_n + b = 0, \tag{4.17}$$

then

$$\mathbf{a}^\intercal(\tilde{\mathbf{W}}^*)^\intercal\tilde{\mathbf{x}} + b = 0. \tag{4.18}$$

4.3 (Incomplete)

4.4

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n, \tag{4.19}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Setting the derivatives of

$$\mathbf{w}^\intercal(\mathbf{m}_2 - \mathbf{m}_1) + \lambda\left(\|\mathbf{w}\|^2 - 1\right) \tag{4.20}$$

with respect to $\mathbf{w}$ and λ to zero gives

$$\begin{aligned} \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} &= \mathbf{0}, \\ \|\mathbf{w}\|^2 - 1 &= 0. \end{aligned} \tag{4.21}$$

Therefore, $\mathbf{w}^\intercal(\mathbf{m}_2 - \mathbf{m}_1)$ under the constratint

$$\|\mathbf{w}\|^2 = 1 \tag{4.22}$$

is maximised if

$$\mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1. \tag{4.23}$$

4.5

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n, \tag{4.24}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{4.25}$$

where

$$\begin{aligned} s_k^2 &= \sum_{n\in\mathcal{C}_k}(y_n - m_k)^2, \\ y_n &= \mathbf{w}^\intercal\mathbf{x}_n, \\ m_k &= \mathbf{w}^\intercal\mathbf{m}_k. \end{aligned} \tag{4.26}$$

Then, $J(\mathbf{w})$ can be written as

$$\frac{\left(\mathbf{w}^\intercal(\mathbf{m}_2 - \mathbf{m}_1)\right)^2}{\sum_{n\in\mathcal{C}_1}\left(\mathbf{w}^\intercal(\mathbf{x}_n - \mathbf{m}_1)\right)^2 + \sum_{n\in\mathcal{C}_2}\left(\mathbf{w}^\intercal(\mathbf{x}_n - \mathbf{m}_2)\right)^2} = \frac{\mathbf{w}^\intercal\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^\intercal\mathbf{S}_{\mathrm{W}}\mathbf{w}}, \tag{4.27}$$

where

$$\begin{aligned} \mathbf{S}_{\mathrm{B}} &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\intercal, \\ \mathbf{S}_{\mathrm{W}} &= \sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\intercal + \sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\intercal. \end{aligned} \tag{4.28}$$

4.6

Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be variables and let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n, \tag{4.29}$$

where N_k is the number of $\mathbf{x}_n$ such that n is in $\mathcal{C}_k$. Let

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{4.30}$$

where

$$\begin{aligned} s_k^2 &= \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2, \\ y_n &= \mathbf{w}^\intercal \mathbf{x}_n, \\ m_k &= \mathbf{w}^\intercal \mathbf{m}_k. \end{aligned} \tag{4.31}$$

Then, by 4.5,

$$J(\mathbf{w}) = \frac{\mathbf{w}^\intercal \mathbf{S}_\mathrm{B} \mathbf{w}}{\mathbf{w}^\intercal \mathbf{S}_\mathrm{W} \mathbf{w}}, \tag{4.32}$$

where

$$\begin{aligned} \mathbf{S}_\mathrm{B} &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\intercal, \\ \mathbf{S}_\mathrm{W} &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\intercal + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\intercal. \end{aligned} \tag{4.33}$$

Let

$$E = \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{w}^\intercal \mathbf{x}_n + w_0 - t_n \right)^2, \tag{4.34}$$

where

$$t_n = \begin{cases} \frac{N}{N_1}, & n \in \mathcal{C}_1, \\ -\frac{N}{N_2}, & n \in \mathcal{C}_2. \end{cases} \tag{4.35}$$

Setting the derivative with respect to $\mathbf{w}$ and w_0 gives

$$\begin{aligned} 0 &= \sum_{n=1}^{N} \left(\mathbf{w}^\intercal \mathbf{x}_n + w_0 - t_n \right), \\ \mathbf{0} &= \sum_{n=1}^{N} \left(\mathbf{w}^\intercal \mathbf{x}_n + w_0 - t_n \right) \mathbf{x}_n. \end{aligned} \tag{4.36}$$

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

$$\mathbf{w}^{\mathsf{T}} \sum_{n=1}^{N} \mathbf{x}_n + N w_0 - \sum_{n=1}^{N} t_n = N \left(\mathbf{w}^{\mathsf{T}} \mathbf{m} + w_0 \right), \tag{4.37}$$

where

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n. \tag{4.38}$$

Therefore,

$$w_0 = -\mathbf{w}^{\mathsf{T}} \mathbf{m}. \tag{4.39}$$

Then, the right hand side of the second equation above can be written as

$$\begin{aligned} &\sum_{n=1}^{N} \left(\mathbf{w}^{\mathsf{T}} \left(\mathbf{x}_n - \mathbf{m} \right) - t_n \right) \mathbf{x}_n \\ &= \sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^{\mathsf{T}} (\mathbf{x}_n - \mathbf{m}) - \frac{N}{N_1} \right) \mathbf{x}_n + \sum_{n \in \mathcal{C}_2} \left(\mathbf{w}^{\mathsf{T}} (\mathbf{x}_n - \mathbf{m}) + \frac{N}{N_2} \right) \mathbf{x}_n. \end{aligned} \tag{4.40}$$

Since

$$\begin{aligned} \mathbf{m} &= \frac{N_1}{N} \mathbf{m}_1 + \frac{N_2}{N} \mathbf{m}_2, \\ \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1) &= \mathbf{0}, \end{aligned} \tag{4.41}$$

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

$$\begin{aligned} &\sum_{n \in \mathcal{C}_1} \left(\mathbf{w}^{\mathsf{T}} \left(\mathbf{x}_n - \mathbf{m}_1 + \frac{N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \right) - \frac{N}{N_1} \right) (\mathbf{x}_n - \mathbf{m}_1 + \mathbf{m}_1) \\ &= \left(\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathsf{T}} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_1 - \mathbf{m}_2) \mathbf{m}_1^{\mathsf{T}} \mathbf{w} - N \mathbf{m}_1. \end{aligned} \tag{4.42}$$

Similarly, the second term can be written as

$$\left(\sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathsf{T}} \right) \mathbf{w} + \frac{N_1 N_2}{N} (\mathbf{m}_2 - \mathbf{m}_1) \mathbf{m}_2^{\mathsf{T}} \mathbf{w} - N \mathbf{m}_2. \tag{4.43}$$

Therefore,

$$\begin{aligned}\mathbf{0} = &\left(\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\intercal\right)\mathbf{w} + \frac{N_1N_2}{N}(\mathbf{m}_1 - \mathbf{m}_2)\mathbf{m}_1^\intercal\mathbf{w} - N\mathbf{m}_1 \\ &+ \left(\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\intercal\right)\mathbf{w} + \frac{N_1N_2}{N}(\mathbf{m}_2 - \mathbf{m}_1)\mathbf{m}_2^\intercal\mathbf{w} - N\mathbf{m}_2.\end{aligned} \tag{4.44}$$

Thus,

$$\left(\mathbf{S}_\mathrm{W} + \frac{N_1N_2}{N}\mathbf{S}_\mathrm{B}\right)\mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2). \tag{4.45}$$

4.7

Let

$$\sigma(a) = \frac{1}{1+\exp(-a)}. \tag{4.46}$$

Then,

$$\sigma(-a) = \frac{1}{1+\exp(a)}. \tag{4.47}$$

The right hand side can be written as

$$1 - \frac{\exp(a)}{1+\exp(a)} = 1 - \frac{1}{1+\exp(-a)}. \tag{4.48}$$

Therefore,

$$\sigma(-a) = 1 - \sigma(a). \tag{4.49}$$

Additionally,

$$\exp(-a) = \frac{1}{\sigma(a)} - 1. \tag{4.50}$$

Then,

$$a = -\ln\left(\frac{1}{\sigma(a)} - 1\right). \tag{4.51}$$

Therefore,

$$\sigma^{-1}(y) = \ln\left(\frac{y}{1-y}\right). \tag{4.52}$$

4.8

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

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right), \tag{4.53}$$

where

$$p(\mathcal{C}_1) + p(\mathcal{C}_2) = 1. \tag{4.54}$$

By the Bayes' theorem,

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}. \tag{4.55}$$

The right hand side can be written as

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{4.56}$$

where

$$a = \ln\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}\right). \tag{4.57}$$

Substituting the expressions above of $p(\mathbf{x}|\mathcal{C}_k)$, we have

$$\begin{aligned} a = &-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(\det\Sigma) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \ln p(\mathcal{C}_1) \\ &+ \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\Sigma) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) - \ln p(\mathcal{C}_2). \end{aligned} \tag{4.58}$$

Therefore,

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma\left(\mathbf{w}^\intercal\mathbf{x} + w_0\right), \tag{4.59}$$

where

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2). \end{aligned} \tag{4.60}$$

4.9

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}, \tag{4.61}$$

where

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1. \tag{4.62}$$

Then,

$$p(\mathbf{T}, \mathbf{\Phi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}. \tag{4.63}$$

If

$$p(\mathcal{C}_k) = \pi_k, \tag{4.64}$$

then, by the Bayes' theorem,

$$\ln p(\mathbf{T}, \mathbf{\Phi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \pi_k + \ln p(\boldsymbol{\phi}_n | \mathcal{C}_k)\right). \tag{4.65}$$

Setting the derivatives of

$$\ln p(\mathbf{T}, \mathbf{\Phi}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{4.66}$$

with respect to π_k and λ to zero gives

$$\begin{aligned} 0 &= \frac{1}{\pi_k} \sum_{n=1}^{N} t_{nk} + \lambda, \\ 0 &= \sum_{k=1}^{K} \pi_k - 1. \end{aligned} \tag{4.67}$$

Then,

$$\lambda = -\sum_{k=1}^{K} \sum_{n=1}^{N} t_{nk}. \tag{4.68}$$

The right hand side can be written as $-N$. Therefore, the maximum likelihood solution for π_k is given by

$$\pi_k = \frac{N_k}{N}, \tag{4.69}$$

where

$$N_k = \sum_{n=1}^{N} t_{nk}. \tag{4.70}$$

4.10

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in K dimensions such that

$$p(\mathbf{t}_n, \boldsymbol{\phi}_n) = \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}, \tag{4.71}$$

where

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1. \tag{4.72}$$

Then,

$$p(\mathbf{T}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(p(\boldsymbol{\phi}_n, \mathcal{C}_k)\right)^{t_{nk}}. \tag{4.73}$$

If

$$p(\boldsymbol{\phi}_n|\mathcal{C}_k) = \mathcal{N}(\boldsymbol{\phi}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \tag{4.74}$$

then, by the Bayes' theorem,

$$\ln p(\mathbf{T}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \mathcal{N}(\boldsymbol{\phi}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) + \ln p(\mathcal{C}_k)\right). \tag{4.75}$$

The right hand side can be written as

$$\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) + \ln p(\mathcal{C}_k)\right). \tag{4.76}$$

By 3.21(a), setting the derivatives of $\ln p(\mathbf{T}, \boldsymbol{\Phi})$ with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ to zero gives

$$\begin{aligned}
\mathbf{0} &= \frac{1}{2}\sum_{n=1}^{N} t_{nk} \left(\boldsymbol{\Sigma}^{-1} + \left(\boldsymbol{\Sigma}^{-1}\right)^\intercal\right)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k), \\
\mathbf{O} &= -\frac{1}{2}\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\left(\boldsymbol{\Sigma}^{-1}\right)^\intercal - \left(\boldsymbol{\Sigma}^{-1}\right)^2 (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^\intercal\right).
\end{aligned} \tag{4.77}$$

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

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}\boldsymbol{\phi}_n,\\ \boldsymbol{\Sigma} &= \frac{1}{N}\sum_{k=1}^{K} N_k\mathbf{S}_k,\end{aligned} \tag{4.78}$$

where

$$\begin{aligned}N_k &= \sum_{n=1}^{N} t_{nk},\\ \mathbf{S}_k &= \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\intercal}.\end{aligned} \tag{4.79}$$

4.11

Let $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_M$ be variables such that

$$p(\boldsymbol{\phi}_m|\mathcal{C}_k) = \prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}}, \tag{4.80}$$

where

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1. \tag{4.81}$$

Then,

$$p(\boldsymbol{\Phi}|\mathcal{C}_k) = \prod_{m=1}^{M}\prod_{l=1}^{L} \mu_{kml}^{\phi_{ml}}. \tag{4.82}$$

By the Bayes' theorem,

$$p(\mathcal{C}_k|\boldsymbol{\Phi}) = \frac{p(\boldsymbol{\Phi}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_{k=1}^{K} p(\boldsymbol{\Phi}|\mathcal{C}_k)p(\mathcal{C}_k)}. \tag{4.83}$$

Therefore,

$$p(\mathcal{C}_k|\boldsymbol{\Phi}) = \frac{\exp\left(a_k(\boldsymbol{\Phi})\right)}{\sum_{k=1}^{K} \exp\left(a_k(\boldsymbol{\Phi})\right)}, \tag{4.84}$$

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 variables such that

$$\begin{gathered} t_n \in \{0,1\}, \\ p(t_n|\mathbf{w}) = y_n^{t_n}(1-y_n)^{1-t_n}, \end{gathered} \tag{4.90}$$

where

$$\begin{gathered} y_n = \sigma\left(\mathbf{w}^\intercal \boldsymbol{\phi}_n\right), \\ \sigma(a) = \frac{1}{1+\exp(-a)}. \end{gathered} \tag{4.91}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{4.92}$$

The right hand side can be written as

$$-\ln\left(\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\right) = -\sum_{n=1}^{N}\left(t_n \ln y_n + (1-t_n)\ln(1-y_n)\right). \tag{4.93}$$

Then, by 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{t_n}{y_n} y_n(1-y_n)\boldsymbol{\phi}_n - \frac{1-t_n}{1-y_n} y_n(1-y_n)\boldsymbol{\phi}_n\right). \tag{4.94}$$

The right hand side can be written as

$$-\sum_{n=1}^{N}(t_n(1-y_n)\boldsymbol{\phi}_n-(1-t_n)y_n\boldsymbol{\phi}_n)=\sum_{n=1}^{N}(y_n-t_n)\boldsymbol{\phi}_n. \tag{4.95}$$

Therefore,

$$\nabla E(\mathbf{w})=\sum_{n=1}^{N}(y_n-t_n)\boldsymbol{\phi}_n. \tag{4.96}$$

4.14

Let $t_1,\cdots,t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0,1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n}(1-y_n)^{1-t_n}, \end{aligned} \tag{4.97}$$

where

$$\begin{aligned} y_n &= \sigma\left(\mathbf{w}^\intercal\boldsymbol{\phi}_n\right), \\ \sigma(a) &= \frac{1}{1+\exp(-a)}. \end{aligned} \tag{4.98}$$

Let

$$E(\mathbf{w})=-\ln p(\mathbf{t}|\mathbf{w}). \tag{4.99}$$

By 4.13, setting the derivative with respect to $\mathbf{w}$ to zero gives

$$\mathbf{0}=\sum_{n=1}^{N}\left(y_n-t_n\right)\boldsymbol{\phi}_n. \tag{4.100}$$

If $\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_N$ are linearly independent, then

$$y_n=t_n. \tag{4.101}$$

Then,

$$\sigma\left(\mathbf{w}^\intercal\boldsymbol{\phi}_n\right)=\begin{cases}1, & t_n=1, \\ 0, & \text{otherwise.}\end{cases} \tag{4.102}$$

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 variables such that

$$\begin{aligned} t_n &\in \{0,1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n}(1-y_n)^{1-t_n}, \end{aligned} \tag{4.104}$$

where

$$\begin{aligned} y_n &= \sigma\left(\mathbf{w}^\intercal \boldsymbol{\phi}_n\right), \\ \sigma(a) &= \frac{1}{1+\exp(-a)}. \end{aligned} \tag{4.105}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{4.106}$$

By 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n. \tag{4.107}$$

Then, by 4.12,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^\intercal. \tag{4.108}$$

The right hand side can be written as

$$\mathbf{H} = \boldsymbol{\Phi}^\intercal \mathbf{R} \boldsymbol{\Phi}, \tag{4.109}$$

where

$$R_{nn'} = \begin{cases} y_n(1-y_n), & n = n', \\ 0, & \text{otherwise}. \end{cases} \tag{4.110}$$

Then,

$$\mathbf{u}^\intercal \mathbf{H} \mathbf{u} = \left(\boldsymbol{\Phi}\mathbf{u}\right)^\intercal \mathbf{R} \left(\boldsymbol{\Phi}\mathbf{u}\right). \tag{4.111}$$

Since

$$y_n(1-y_n) > 0, \tag{4.112}$$

we have

$$\mathbf{u}^\intercal \mathbf{H} \mathbf{u} > 0. \tag{4.113}$$

Therefore, $\mathbf{H}$ is positive definite. Thus, E is a convex function of $\mathbf{w}$ and it has a unique minimum.

4.16

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1|\boldsymbol{\phi}_n) &= \pi_n. \end{aligned} \tag{4.114}$$

Then,

$$p(t_n|\boldsymbol{\phi}_n) = \pi_n^{t_n}(1 - \pi_n)^{1-t_n}. \tag{4.115}$$

Therefore,

$$p(\mathbf{t}|\boldsymbol{\Phi}) = \prod_{n=1}^{N} \pi_n^{t_n}(1 - \pi_n)^{1-t_n}. \tag{4.116}$$

Thus,

$$-\ln p(\mathbf{t}|\boldsymbol{\Phi}) = -\sum_{n=1}^{N} \left(t_n \ln \pi_n + (1 - t_n) \ln(1 - \pi_n)\right). \tag{4.117}$$

4.17

Let

$$y_k = \frac{\exp(a_k)}{\sum_{k=1}^{K} \exp(a_k)}. \tag{4.118}$$

Then,

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k)}{\sum_{k=1}^{K} \exp(a_k)} - \frac{\exp(2a_k)}{\left(\sum_{k=1}^{K} \exp(a_k)\right)^2}. \tag{4.119}$$

The right hand side can be written as $y_k(1 - y_k)$. If $k \neq k'$, then

$$\frac{\partial y_k}{\partial a_{k'}} = -\frac{\exp(a_k + a_{k'})}{\left(\sum_{k=1}^{K} \exp(a_k)\right)^2}. \tag{4.120}$$

The right hand side can be written as $-y_k y_{k'}$. Therefore,

$$\frac{\partial y_k}{\partial a_{k'}} = y_k(I_{kk'} - y_{k'}). \tag{4.121}$$

4.18

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

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(\mathbf{t}_n|\mathbf{W}) &= \prod_{k=1}^{K} y_{nk}^{t_{nk}}, \end{aligned} \tag{4.122}$$

where

$$\begin{aligned} y_{nk} &= \frac{\exp(a_{nk})}{\sum_{k=1}^{K} \exp(a_{nk})}, \\ a_{nk} &= \mathbf{w}_k^\intercal \boldsymbol{\phi}_n. \end{aligned} \tag{4.123}$$

Then,

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.124}$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{W}). \tag{4.125}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{4.126}$$

Then, by 4.17,

$$\nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} y_{nk}(I_{kk'} - y_{nk'}) \frac{t_{nk}}{y_{nk}} \boldsymbol{\phi}_n. \tag{4.127}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} \left(\sum_{k=1}^{K} (I_{kk'} - y_{nk'}) t_{nk} \right) \boldsymbol{\phi}_n = -\sum_{n=1}^{N} (t_{nk'} - y_{nk'}) \boldsymbol{\phi}_n. \tag{4.128}$$

Therefore,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^{N} (y_{nk} - t_{nk}) \boldsymbol{\phi}_n. \tag{4.129}$$

4.19

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1|a_n) &= \Phi(a_n), \end{aligned} \tag{4.130}$$

where

$$\begin{aligned} \Phi(a) &= \int_{-\infty}^{a} \mathcal{N}(\theta|0,1) d\theta, \\ a_n &= \mathbf{w}^\intercal \boldsymbol{\phi}_n. \end{aligned} \tag{4.131}$$

Then,

$$p(t_n|\boldsymbol{\phi}_n) = (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \tag{4.132}$$

Therefore,

$$p(\mathbf{t}|\boldsymbol{\Phi}) = \prod_{n=1}^{N} (\Phi(a_n))^{t_n} (1 - \Phi(a_n))^{1-t_n}. \tag{4.133}$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\boldsymbol{\phi}). \tag{4.134}$$

The right hand side can be written as

$$-\sum_{n=1}^{N} (t_n \ln \Phi(a_n) + (1 - t_n) \ln (1 - \Phi(a_n))). \tag{4.135}$$

Then,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)} - (1 - t_n) \frac{\mathcal{N}(a_n|0,1)}{1 - \Phi(a_n)} \right) \boldsymbol{\phi}_n. \tag{4.136}$$

The right hand side can be written as

$$\begin{aligned} &-\sum_{n=1}^{N} \left(\frac{t_n}{\Phi(a_n)} - \frac{1 - t_n}{1 - \Phi(a_n)} \right) \mathcal{N}(a_n|0,1) \boldsymbol{\phi}_n \\ =&\sum_{n=1}^{N} \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)(1 - \Phi(a_n))} (\Phi(a_n) - t_n) \boldsymbol{\phi}_n. \end{aligned} \tag{4.137}$$

Therefore,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n. \tag{4.138}$$

Then,

$$\begin{aligned}
\nabla\nabla E(\mathbf{w}) = &\sum_{n=1}^{N} \frac{-a_n\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathsf{T}} \\
&- \sum_{n=1}^{N} \frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\left(\Phi(a_n)\right)^2\left(1-\Phi(a_n)\right)} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathsf{T}} \\
&+ \sum_{n=1}^{N} \frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\Phi(a_n)\left(1-\Phi(a_n)\right)^2} \left(\Phi(a_n) - t_n\right) \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathsf{T}} \\
&+ \sum_{n=1}^{N} \frac{\left(\mathcal{N}(a_n|0,1)\right)^2}{\Phi(a_n)\left(1-\Phi(a_n)\right)} \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathsf{T}}.
\end{aligned} \tag{4.139}$$

Therefore,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} b_n \boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathsf{T}}, \tag{4.140}$$

where

$$\begin{aligned}
b_n = &\left(\frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)}\right)^2 \left(\left(\Phi(a_n)\right)^2 - 2t_n\Phi(a_n) + t_n\right) \\
&- \frac{\mathcal{N}(a_n|0,1)}{\Phi(a_n)\left(1-\Phi(a_n)\right)} a_n \left(\Phi(a_n) - t_n\right).
\end{aligned} \tag{4.141}$$

4.20 (Incomplete)

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

$$\begin{aligned}
t_{nk} &\in \{0,1\}, \\
p(\mathbf{t}_n|\mathbf{W}) &= \prod_{k=1}^{K} y_{nk}^{t_{nk}},
\end{aligned} \tag{4.142}$$

where

$$y_{nk} = \frac{\exp(a_{nk})}{\sum_{k=1}^{K} \exp(a_{nk})}, \qquad a_{nk} = \mathbf{w}_k^\intercal \boldsymbol{\phi}_n. \tag{4.143}$$

Then,

$$p(\mathbf{T}|\mathbf{W}) = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{4.144}$$

Let

$$E(\mathbf{W}) = -\ln p(\mathbf{T}|\mathbf{W}). \tag{4.145}$$

By 4.18,

$$\nabla_{\mathbf{w}_k} E(\mathbf{W}) = \sum_{n=1}^{N} (y_{nk} - t_{nk})\, \boldsymbol{\phi}_n. \tag{4.146}$$

Additionally, by 4.17,

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_{k'}} E(\mathbf{W}) = \sum_{n=1}^{N} y_{nk}(I_{kk'} - y_{nk'}) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^\intercal. \tag{4.147}$$

The right hand side can be written as

$$\mathbf{H}_{kk'} = \boldsymbol{\Phi}^\intercal \mathbf{R}_{kk'} \boldsymbol{\Phi}, \tag{4.148}$$

where

$$R_{kk'nn'} = \begin{cases} y_{nk}(I_{kk'} - y_{nk'}), & n = n', \\ 0, & \text{otherwise.} \end{cases} \tag{4.149}$$

Let

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{K1} & \cdots & \mathbf{H}_{KK} \end{bmatrix},$$

and

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_K \end{bmatrix},$$

where $\mathbf{u}_1, \cdots, \mathbf{u}_K$ are vectors in the same dimension as $\mathbf{w}$. Then,

$$\mathbf{u}^\intercal \mathbf{H} \mathbf{u} = \sum_{k=1}^{K} \sum_{k'=1}^{K} \mathbf{u}_k^\intercal \mathbf{H}_{kk'} \mathbf{u}_{k'}, \tag{4.150}$$

Then, the right hand side can be written as

$$\sum_{k=1}^{K} \sum_{k'=1}^{K} (\mathbf{\Phi} \mathbf{u}_k)^\intercal \mathbf{R}_{kk'} (\mathbf{\Phi} \mathbf{u}_{k'}). \tag{4.151}$$

4.21

Let

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1) d\theta. \tag{4.152}$$

The right hand side can be written as

$$\int_{-\infty}^{0} \mathcal{N}(\theta|0,1) d\theta + \int_{0}^{a} \mathcal{N}(\theta|0,1) d\theta = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{a} \exp\left(-\frac{\theta^2}{2}\right) d\theta. \tag{4.153}$$

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

$$\frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{a}{\sqrt{2}}} \exp\left(-t^2\right) \sqrt{2} dt = \frac{1}{2} \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right), \tag{4.154}$$

where

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp\left(-t^2\right) dt. \tag{4.155}$$

Therefore,

$$\Phi(a) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right). \tag{4.156}$$

4.22

Let $\boldsymbol{\theta}$ be a variable in M dimensions. By a Taylor expansion,

$$\ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) \simeq \ln\left(p(\mathcal{D}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)\right) + \mathbf{v}(\boldsymbol{\theta}_0)^\intercal (\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \mathbf{A}(\boldsymbol{\theta}_0) (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{4.157}$$

where

$$\begin{aligned}\mathbf{v}(\boldsymbol{\theta}) &= \nabla \ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right), \\ \mathbf{A}(\boldsymbol{\theta}) &= -\nabla\nabla \ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right).\end{aligned} \tag{4.158}$$

Let $\boldsymbol{\theta}_{\mathrm{MAP}}$ be a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then,

$$\ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) \simeq \ln\left(p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}})\right) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})^{\intercal}\mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}}), \tag{4.159}$$

so that

$$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})^{\intercal}\mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})\right). \tag{4.160}$$

By marginalisation, integrating both sides with respect to $\boldsymbol{\theta}$ gives

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) \int \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})^{\intercal}\mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MAP}})\right) d\boldsymbol{\theta}. \tag{4.161}$$

The integral of the right hand side can be written as

$$(2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})^{-1}\right)^{\frac{1}{2}} = (2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})\right)^{-\frac{1}{2}}. \tag{4.162}$$

Therefore,

$$p(\mathcal{D}) \simeq p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}})(2\pi)^{\frac{M}{2}} \left(\det \mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})\right)^{-\frac{1}{2}}, \tag{4.163}$$

so that

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det \mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})\right). \tag{4.164}$$

4.23

Let $\boldsymbol{\theta}$ be a variable in M dimensions. By 4.22,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det \mathbf{A}(\boldsymbol{\theta}_{\mathrm{MAP}})\right), \tag{4.165}$$

where $\boldsymbol{\theta}_{\mathrm{MAP}}$ is a stationary point of $p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and

$$\mathbf{A}(\boldsymbol{\theta}) = -\nabla\nabla \ln\left(p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right). \tag{4.166}$$

If

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0), \tag{4.167}$$

then

$$\nabla\nabla \ln p(\theta) = -\mathbf{V}_0^{-1}, \tag{4.168}$$

so that

$$\mathbf{A}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta}) + \mathbf{V}_0^{-1}, \tag{4.169}$$

where

$$\mathbf{H}(\boldsymbol{\theta}) = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}). \tag{4.170}$$

Then, the right hand side of the approximation above can be written as

$$\begin{aligned}
&\ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln(\det \mathbf{V}_0) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}_0)^{\intercal}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}_0) \\
&+ \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\det\left(\mathbf{H}(\boldsymbol{\theta}_{\mathrm{MAP}}) + \mathbf{V}_0^{-1}\right)\right) \\
=& \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{1}{2}\ln\left(\det \mathbf{V}_0^{-1}\right) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}_0)^{\intercal}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}_0) \\
&- \frac{1}{2}\ln\left(\det\left(\mathbf{H}(\boldsymbol{\theta}_{\mathrm{MAP}}) + \mathbf{V}_0^{-1}\right)\right).
\end{aligned} \tag{4.171}$$

If $\mathbf{V}_0^{-1}$ can be neglected, the right hand side can be written as

$$\ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}\ln\left(\det \mathbf{H}(\boldsymbol{\theta}_{\mathrm{MAP}})\right). \tag{4.172}$$

If each data point is independent and identically distributed, then

$$\mathbf{H}(\boldsymbol{\theta}) = N\bar{\mathbf{H}}(\boldsymbol{\theta}), \tag{4.173}$$

where

$$\bar{\mathbf{H}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}_n(\boldsymbol{\theta}), \tag{4.174}$$

and $\mathbf{H}_n(\boldsymbol{\theta})$ is the one for each data point. Then,

$$\det \mathbf{H}(\boldsymbol{\theta}_{\mathrm{MAP}}) = N^M \det \bar{\mathbf{H}}(\boldsymbol{\theta}_{\mathrm{MAP}}). \tag{4.175}$$

Threrefore,

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{M}{2}\ln N. \tag{4.176}$$

4.24 (Incomplete)

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0,1\}, \\ p(t_n|\mathbf{w}) &= y_n^{t_n}(1-y_n)^{1-t_n}, \end{aligned} \tag{4.177}$$

where

$$\begin{aligned} y_n &= \sigma\left(\mathbf{w}^\intercal \boldsymbol{\phi}_n\right), \\ \sigma(a) &= \frac{1}{1+\exp(-a)}. \end{aligned} \tag{4.178}$$

By the Bayes' theorem,

$$p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}). \tag{4.179}$$

If

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0), \tag{4.180}$$

then the lograrithm of the right hand side except the terms independent of $\mathbf{w}$ and $\mathbf{t}$ can be written as

$$\sum_{n=1}^{N} \left(t_n \ln y_n + (1-t_n)\ln(1-y_n)\right) - \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^\intercal \mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0). \tag{4.181}$$

Then, by 4.22,

$$p(\mathbf{w}|\mathbf{t}) \simeq \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)? \tag{4.182}$$

where $\mathbf{w}_{\mathrm{MAP}}$ is the maximum likelihood solution for $p(\mathbf{w})$ and

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}). \tag{4.183}$$

By marginalisation,

$$p(\mathcal{C}_1|\mathbf{t}) = \int p(\mathcal{C}_1|\mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}. \tag{4.184}$$

The logarithm of the integrand of the right hand side except the terms independent of $\mathbf{w}$ can be approximated as

$$-\ln\left(1+\exp\left(-\mathbf{w}^\intercal \boldsymbol{\phi}\right)\right) - \frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^\intercal \mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}}) = \tag{4.185}$$

4.25

Let

$$\begin{aligned}\sigma(a) &= \frac{1}{1+\exp(-a)}, \\ \Phi(a) &= \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta.\end{aligned} \tag{4.186}$$

By 4.12,

$$\frac{d\sigma(a)}{da} = \sigma(a)\left(1-\sigma(a)\right). \tag{4.187}$$

On the other hand, the right hand side can be written as

$$\frac{d\Phi(\lambda a)}{da} = \lambda\mathcal{N}(\lambda a|0,1). \tag{4.188}$$

Let us assume that

$$\left.\frac{d\sigma(a)}{da}\right|_{a=0} = \left.\frac{d\Phi(\lambda a)}{da}\right|_{a=0}. \tag{4.189}$$

Then,

$$\frac{1}{4} = \lambda(2\pi)^{-\frac{1}{2}}. \tag{4.190}$$

Therefore,

$$\lambda^2 = \frac{\pi}{8}. \tag{4.191}$$

4.26

Let

$$I(\mu) = \int \Phi(\lambda a)\mathcal{N}(a|\mu,\sigma^2)da, \tag{4.192}$$

where

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta. \tag{4.193}$$

By the transformation

$$z = \frac{a-\mu}{\sigma}, \tag{4.194}$$

the right hand side can be written as

$$\int \Phi\left(\lambda(\mu+\sigma z)\right)\mathcal{N}(\mu+\sigma z|\mu,\sigma^2)\sigma dz = \int \Phi\left(\lambda(\mu+\sigma z)\right)\mathcal{N}(z|0,1)dz. \tag{4.195}$$

Then,

$$\frac{\partial}{\partial \mu} I(\mu) = \lambda \int \mathcal{N}\left(\lambda(\mu + \sigma z)|0,1\right) \mathcal{N}(z|0,1) dz. \tag{4.196}$$

The logarithm of the integrand of the right hand side can be written as

$$\begin{aligned} & -\frac{1}{2}\ln(2\pi) - \frac{\lambda^2(\mu+\sigma z)^2}{2} - \frac{1}{2}\ln(2\pi) - \frac{z^2}{2} \\ =& -\ln(2\pi) - \frac{1+\sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}\right)^2 + \frac{\mu^2\sigma^2\lambda^4}{2\left(1+\sigma^2\lambda^2\right)} - \frac{\mu^2\lambda^2}{2}. \end{aligned} \tag{4.197}$$

The right hand side can be written as

$$\begin{aligned} & -\ln(2\pi) - \frac{1+\sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}\right)^2 - \frac{\mu^2\lambda^2}{2\left(1+\sigma^2\lambda^2\right)} \\ =& -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(1+\sigma^2\lambda^2\right)^{-1} - \frac{1+\sigma^2\lambda^2}{2}\left(z + \frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}\right)^2 \\ & -\ln\lambda - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\lambda^{-2}+\sigma^2\right) - \frac{\mu^2}{2\left(\lambda^{-2}+\sigma^2\right)}. \end{aligned} \tag{4.198}$$

Then, the integral can be written as

$$\begin{aligned} & \int \mathcal{N}\left(z\middle| -\frac{\mu\sigma\lambda^2}{1+\sigma^2\lambda^2}, \left(1+\sigma^2\lambda^2\right)^{-1}\right) \frac{1}{\lambda}\mathcal{N}\left(\mu\middle|\, 0, \lambda^{-2}+\sigma^2\right) dz \\ =& \frac{1}{\lambda}\mathcal{N}\left(\mu\middle|\, 0, \lambda^{-2}+\sigma^2\right). \end{aligned} \tag{4.199}$$

Therefore,

$$\frac{\partial}{\partial \mu} I(\mu) = \mathcal{N}\left(\mu|0, \lambda^{-2}+\sigma^2\right). \tag{4.200}$$

Integrating both sides with respect to μ gives

$$I(\mu) = \int_{-\infty}^{\mu} \mathcal{N}\left(m|0, \lambda^{-2}+\sigma^2\right) dm. \tag{4.201}$$

By the transformation

$$m' = \frac{m}{\left(\lambda^{-2}+\sigma^2\right)^{\frac{1}{2}}}, \tag{4.202}$$

the right hand side can be written as

$$\int_{-\infty}^{\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}} \left(\lambda^{-2}+\sigma^2\right)^{-\frac{1}{2}} \mathcal{N}\left(m'|0,1\right)\left(\lambda^{-2}+\sigma^2\right)^{\frac{1}{2}} dm' = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}\right). \tag{4.203}$$

Therefore,

$$I(\mu) = \Phi\left(\frac{\mu}{(\lambda^{-2}+\sigma^2)^{\frac{1}{2}}}\right). \tag{4.204}$$

5 Neural Networks

5.1

Let

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{m=1}^{M} w_{km}^{(2)} \sigma\left(\sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)}\right) + w_{k0}^{(2)}\right), \tag{5.1}$$

where

$$\sigma(a) = \frac{1}{1+\exp(-a)}. \tag{5.2}$$

Here,

$$\sigma(a) = \frac{\exp\left(\frac{a}{2}\right)}{\exp\left(\frac{a}{2}\right) + \exp\left(-\frac{a}{2}\right)}. \tag{5.3}$$

The right hand side can be written as

$$\tanh\left(\frac{a}{2}\right) + \sigma(-a) = \tanh\left(\frac{a}{2}\right) + 1 - \sigma(a). \tag{5.4}$$

Therefore,

$$\sigma(a) = \frac{1}{2}\left(1 + \tanh\left(\frac{a}{2}\right)\right). \tag{5.5}$$

Then, the argument of the right hand side can be written as

$$\begin{aligned}
&\sum_{m=1}^{M} w_{km}^{(2)} \left(\frac{1}{2}\left(1 + \tanh\left(\frac{1}{2}\left(\sum_{d=1}^{D} w_{md}^{(1)} x_d + w_{m0}^{(1)}\right)\right)\right)\right) + w_{k0}^{(2)} \\
=&\frac{1}{2}\sum_{m=1}^{M} w_{km}^{(2)} \tanh\left(\frac{1}{2}\sum_{d=1}^{D} w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)}\right) + \frac{1}{2}\sum_{m=1}^{M} w_{km}^{(2)} + w_{k0}^{(2)}.
\end{aligned} \tag{5.6}$$

Therefore,

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\frac{1}{2}\sum_{m=1}^{M} w_{km}^{(2)} \tanh\left(\frac{1}{2}\sum_{d=1}^{D} w_{md}^{(1)} x_d + \frac{1}{2} w_{m0}^{(1)}\right) + \frac{1}{2}\sum_{m=1}^{M} w_{km}^{(2)} + w_{k0}^{(2)}\right). \tag{5.7}$$

5.2

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

$$p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}\right). \tag{5.8}$$

Then, the logarithm of the likelihood except the terms independent of $\mathbf{w}$ can be written as

$$-\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\intercal}\left(\beta^{-1}\mathbf{I}\right)^{-1}\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right) = -\frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{5.9}$$

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

$$\mathbf{0} = -\beta\sum_{n=1}^{N}\frac{\partial\mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial\mathbf{w}}\left(\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\right). \tag{5.10}$$

Let

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2. \tag{5.11}$$

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

$$\mathbf{0} = \sum_{n=1}^{N}\frac{\partial\mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial\mathbf{w}}\left(\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\right). \tag{5.12}$$

Therefore, maximising the likelihood is equivalent to minimising $E(\mathbf{w})$.

5.3

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

$$p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \mathbf{\Sigma}\right). \tag{5.13}$$

Then, the logarithm of the likelihood except the terms independent of $\mathbf{w}$ and $\mathbf{\Sigma}$ can be written as

$$-\frac{1}{2}\ln(\det\mathbf{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\intercal}\mathbf{\Sigma}^{-1}\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right). \tag{5.14}$$

Setting the derivatives with respect to $\mathbf{w}$ and $\boldsymbol{\Sigma}$ to zero gives

$$\begin{aligned}\mathbf{0} &= -\sum_{n=1}^{N} \frac{\partial \mathbf{y}(\mathbf{x}_n, \mathbf{w})}{\partial \mathbf{w}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right), \\ \mathbf{O} &= -\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^2 \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\intercal}.\end{aligned} \tag{5.15}$$

Therefore, the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)\left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\intercal}. \tag{5.16}$$

On the other hand, if $\boldsymbol{\Sigma}$ is fixed and known, then the maximum likelihood solution for $\mathbf{w}$ is given by minimising

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\intercal} \boldsymbol{\Sigma}^{-1} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right). \tag{5.17}$$

5.4

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{gathered} t_n \in \{0, 1\}, \\ p(t_n = 1|\mathbf{x}_n) = (1-\epsilon)y_n + \epsilon\left(1 - y_n\right), \end{gathered} \tag{5.18}$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \tag{5.19}$$

Then,

$$p(t_n|\mathbf{x}_n) = \left((1-\epsilon)y_n + \epsilon\left(1-y_n\right)\right)^{t_n} \left(\epsilon y_n + (1-\epsilon)\left(1-y_n\right)\right)^{1-t_n}. \tag{5.20}$$

Therefore,

$$\begin{aligned}&-\ln\left(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n)\right) \\ &= -\sum_{n=1}^{N} \left(t_n \ln\left((1-\epsilon)y_n + \epsilon(1-y_n)\right) + (1-t_n)\ln\left(\epsilon y_n + (1-\epsilon)(1-y_n)\right)\right).\end{aligned} \tag{5.21}$$

5.5

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be variables in K dimensions such that

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1 | \mathbf{x}_n) &= y_{nk}, \end{aligned} \tag{5.22}$$

where

$$y_{nk} = y_k(\mathbf{x}_n, \mathbf{w}). \tag{5.23}$$

Then,

$$p(t_{nk} | \mathbf{x}_n) = y_{nk}^{t_{nk}}, \tag{5.24}$$

so that

$$p(\mathbf{t}_n | \mathbf{x}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{5.25}$$

Therefore,

$$\ln \left(\prod_{n=1}^{N} p(\mathbf{t}_n | \mathbf{x}_n) \right) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.26}$$

5.6

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{0, 1\}, \\ p(t_n = 1 | \mathbf{x}_n) &= y_n, \end{aligned} \tag{5.27}$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \tag{5.28}$$

Then,

$$p(t_n | \mathbf{x}_n) = y_n^{t_n} (1 - y_n)^{1 - t_n}. \tag{5.29}$$

Let

$$E(\mathbf{w}) = -\ln \left(\prod_{n=1}^{N} p(t_n | \mathbf{x}_n) \right). \tag{5.30}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \left(t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right). \tag{5.31}$$

If

$$y_n = \sigma(a_n), \tag{5.32}$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}, \tag{5.33}$$

then, by 4.12,

$$\frac{\partial E(\mathbf{w})}{\partial a_n} = -y_n(1 - y_n)\left(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n}\right). \tag{5.34}$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_n} = y_n - t_n. \tag{5.35}$$

5.7

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

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1 | \mathbf{x}_n) &= y_{nk}, \end{aligned} \tag{5.36}$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^{K} y_{nk} &= 1. \end{aligned} \tag{5.37}$$

Then,

$$p(\mathbf{t}_n | \mathbf{x}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{5.38}$$

Let

$$E(\mathbf{w}) = -\ln\left(\prod_{n=1}^{N} p(\mathbf{t}_n | \mathbf{x}_n)\right). \tag{5.39}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}. \tag{5.40}$$

If

$$y_{nk} = \frac{\exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}{\sum_{k=1}^{K} \exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}, \tag{5.41}$$

then, by 4.17,

$$\frac{\partial E(\mathbf{w})}{\partial a_{k'}} = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} y_{nk} (I_{kk'} - y_{nk}) \frac{1}{y_{nk}}. \tag{5.42}$$

The right hand side can be written as

$$-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(I_{kk'} - y_{nk}) = -\sum_{n=1}^{N}\left(\sum_{k=1}^{K} t_{nk} y_{nk} - t_{nk'}\right). \tag{5.43}$$

Therefore,

$$\frac{\partial E(\mathbf{w})}{\partial a_k} = \sum_{n=1}^{N}(y_{nk} - t_{nk}). \tag{5.44}$$

5.8

Setting the derivative of

$$\tanh a = \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \tag{5.45}$$

gives

$$\frac{d}{da}\tanh a = 1 - \left(\frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)}\right)^2. \tag{5.46}$$

Therefore,

$$\frac{d}{da}\tanh a = 1 - (\tanh a)^2. \tag{5.47}$$

5.9

Let $t_1, \cdots, t_N$ be variables such that

$$\begin{aligned} t_n &\in \{-1, 1\}, \\ p(t_n = 1|\mathbf{x}_n) &= \frac{1 + y_n}{2}, \end{aligned} \tag{5.48}$$

where

$$y_n = y(\mathbf{x}_n, \mathbf{w}). \tag{5.49}$$

Then,

$$p(t_n|\mathbf{x}_n) = \left(\frac{1+y_n}{2}\right)^{\frac{1+t_n}{2}} \left(\frac{1-y_n}{2}\right)^{\frac{1-t_n}{2}}. \tag{5.50}$$

Let

$$E(\mathbf{w}) = -\ln\left(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n)\right). \tag{5.51}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{1+t_n}{2}\ln\frac{1+y_n}{2} + \frac{1-t_n}{2}\ln\frac{1-y_n}{2}\right). \tag{5.52}$$

The appropriate choice of y is tanh.

5.10

Let

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}), \tag{5.53}$$

where E is a real function of real vectors. Then, $\mathbf{H}$ is a real symmetric matrix. Therefore, by 2.20, $\mathbf{H}$ is positive if and only if its eigenvalues are positive.

5.11

Let $\mathbf{w}$ be a real vector in M dimensions. Let E be a real function of $\mathbf{w}$. Let $\mathbf{w}^*$ be a vector such that

$$\nabla E\left(\mathbf{w}^*\right) = \mathbf{0}. \tag{5.54}$$

Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E\left(\mathbf{w}^*\right) + \frac{1}{2}\left(\mathbf{w}-\mathbf{w}^*\right)^{\intercal}\mathbf{H}\left(\mathbf{w}-\mathbf{w}^*\right), \tag{5.55}$$

where

$$\mathbf{H} = \left.\nabla\nabla E(\mathbf{w})\right|_{\mathbf{w}=\mathbf{w}^*}. \tag{5.56}$$

Let $\mathbf{u}_1, \cdots, \mathbf{u}_M$ be eigenvectors such that

$$\mathbf{H}\mathbf{u}_m = \lambda_m\mathbf{u}_m. \tag{5.57}$$

Note that $\mathbf{H}$ is a real symmetric matrix. Then, by **??**, we have

$$\mathbf{u}_m^\intercal \mathbf{u}_{m'} = I_{mm'}. \tag{5.58}$$

Therefore, there exists $\alpha_1, \cdots, \alpha_M$ such that

$$\mathbf{w} - \mathbf{w}^* = \sum_{m=1}^{M} \alpha_m \mathbf{u}_m. \tag{5.59}$$

Then, the approximation can be written as

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} \sum_{m=1}^{M} \lambda_m \alpha_m^2. \tag{5.60}$$

Therefore, the contours of constant E are ellipses whose axes are aligned with $\mathbf{u}_1, \cdots, \mathbf{u}_M$ with lengths which are proportional to $\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_M^{-\frac{1}{2}}$.

5.12

Let $\mathbf{w}$ be a real vector. Let E be a real function of $\mathbf{w}$. Let $\mathbf{w}^*$ be a vector such that

$$\nabla E(\mathbf{w}^*) = \mathbf{0}. \tag{5.61}$$

Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^\intercal \mathbf{H} (\mathbf{w} - \mathbf{w}^*), \tag{5.62}$$

where

$$\mathbf{H} = \nabla\nabla E(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}^*}. \tag{5.63}$$

If $\mathbf{H}$ is positive definite, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \tag{5.64}$$

Therefore, $\mathbf{w}^*$ is a local minimum of the right hand side. On the other hand, if $\mathbf{w}^*$ is a local minimum of the right hand side, then the second term of the right hand side is positive unless

$$\mathbf{w} = \mathbf{w}^*. \tag{5.65}$$

Therefore, $\mathbf{H}$ is positive definite. Thus, the necessary and sufficient condition for $\mathbf{w}^*$ to be a local minimum is that $\mathbf{H}$ be positive definite.

5.13

Let $\mathbf{w}$ be a vector in M dimensions. Let E be a function of $\mathbf{w}$. Then, by a Taylor expansion,

$$E(\mathbf{w}) \simeq E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^\intercal \mathbf{b} + \frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^\intercal \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}}), \tag{5.66}$$

where

$$\begin{aligned} \mathbf{b} &= \nabla E(\mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} . \\ \mathbf{H} &= \nabla\nabla E(\mathbf{w})\big|_{\mathbf{w}=\hat{\mathbf{w}}} . \end{aligned} \tag{5.67}$$

Since $\mathbf{b}$ is a vector in M dimensions and $\mathbf{H}$ is a $M \times M$ symmetric matrix, the number of independent elements of the right hand side is

$$M + \frac{M(M+1)}{2} = \frac{M(M+3)}{2}. \tag{5.68}$$

5.14

Let w be a variable. Let E_n be a function of w. Then, by a Taylor expansion,

$$\begin{aligned} E_n(w_{mm'} + \epsilon) &= E_n(w_{mm'}) + \left.\frac{\partial E_n}{\partial w}\right|_{w=w_{mm'}} \epsilon + O\left(\epsilon^2\right), \\ E_n(w_{mm'} - \epsilon) &= E_n(w_{mm'}) - \left.\frac{\partial E_n}{\partial w}\right|_{w=w_{mm'}} \epsilon + O\left(\epsilon^2\right). \end{aligned} \tag{5.69}$$

Therefore,

$$\left.\frac{\partial E_n}{\partial w}\right|_{w=w_{mm'}} = \frac{E_n(w_{mm'} + \epsilon) - E_n(w_{mm'} - \epsilon)}{2\epsilon} + O\left(\epsilon^2\right). \tag{5.70}$$

5.15 (Incomplete)

5.16

Let $\mathbf{t}_1, \cdots, \mathbf{t}_N$ be vectors. Let $\mathbf{y}_1, \cdots, \mathbf{y}_N$ be vectors which are dependent on a vector $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{t}_n\|^2. \tag{5.71}$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (\nabla \mathbf{y}_n)^\intercal (\mathbf{y}_n - \mathbf{t}_n), \tag{5.72}$$

so that

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \left(\nabla \operatorname{vec} (\nabla \mathbf{y}_n)^\intercal\right)^\intercal (\mathbf{y}_n - \mathbf{t}_n) + \sum_{n=1}^{N} (\nabla \mathbf{y}_n)^\intercal (\nabla \mathbf{y}_n) . \tag{5.73}$$

5.17

Let t be a variable. Let y be a function of a vector $\mathbf{x}$ and a vector $\mathbf{w}$. Let

$$E(\mathbf{w}) = \frac{1}{2} \int\int (y - t)^2 p(\mathbf{x}, t) d\mathbf{x} dt. \tag{5.74}$$

Then,

$$\nabla E(\mathbf{w}) = \int\int (y - t) p(\mathbf{x}, t) \nabla y d\mathbf{x} dt. \tag{5.75}$$

The right hand side can be written as

$$\begin{aligned} &\int y \nabla y \left(\int p(\mathbf{x}, t) dt \right) d\mathbf{x} - \int \nabla y \left(\int tp(t|\mathbf{x}) dt \right) p(\mathbf{x}) d\mathbf{x} \\ =& \int y \nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla y \, \mathrm{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{5.76}$$

Then,

$$\nabla\nabla E(\mathbf{w}) = \int \nabla y \, (\nabla y)^\intercal \, p(\mathbf{x}) d\mathbf{x} + \int y \nabla\nabla y p(\mathbf{x}) d\mathbf{x} - \int \nabla\nabla y \, \mathrm{E}(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{5.77}$$

The second and the third terms of the right hand side can be written as

$$\mathrm{E}\left(y \nabla\nabla y\right) - \mathrm{E}\left(\nabla\nabla y \, \mathrm{E}(t|\mathbf{x})\right) = \mathrm{E}\left(\left(y - \mathrm{E}(t|\mathbf{x})\right) \nabla\nabla y\right). \tag{5.78}$$

Therefore, if

$$y = \mathrm{E}(t|\mathbf{x}), \tag{5.79}$$

then

$$\nabla\nabla E(\mathbf{w}) = \int \nabla y \, (\nabla y)^\intercal \, p(\mathbf{x}) d\mathbf{x}. \tag{5.80}$$

5.18

Let $t_1, \cdots, t_N$ be variables. Let $y_1, \cdots, y_N$ be variablees such that

$$
\begin{aligned}
y_n &= \mathbf{w}_n^{(2)\intercal}\mathbf{z} + \mathbf{w}_n^{(0)\intercal}\mathbf{x}, \\
z_m &= \tanh\left(\mathbf{w}_m^{(1)\intercal}\mathbf{x}\right).
\end{aligned} \tag{5.81}
$$

Let

$$
E = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2. \tag{5.82}
$$

Then,

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial \mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{w}_m^{(1)}}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.
\end{aligned} \tag{5.83}
$$

Therefore,

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= (y_n - t_n)\mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= (y_n - t_n)\mathbf{A}\mathbf{w}_n^{(2)}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= (y_n - t_n)\mathbf{z},
\end{aligned} \tag{5.84}
$$

where

$$
A_{mm'} = \left(1 - z_m^2\right) x_{m'}. \tag{5.85}
$$

5.19

Let $t_1, \cdots, t_N$ be variables such that

$$
\begin{aligned}
t_n &\in \{0, 1\}, \\
p(t_n|\mathbf{w}) &= y_n^{t_n}(1 - y_n)^{1-t_n},
\end{aligned} \tag{5.86}
$$

where

$$
\begin{aligned}
y_n &= \sigma\left(a_n(\mathbf{w})\right), \\
\sigma(a) &= \frac{1}{1 + \exp(-a)},
\end{aligned} \tag{5.87}
$$

Let

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}). \tag{5.88}$$

The right hand side can be written as

$$-\ln\left(\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\right) = -\sum_{n=1}^{N}\left(t_n \ln y_n + (1-t_n)\ln(1-y_n)\right). \tag{5.89}$$

Then, by 4.12,

$$\nabla E(\mathbf{w}) = -\sum_{n=1}^{N}\left(\frac{t_n}{y_n} y_n(1-y_n)\nabla a_n - \frac{1-t_n}{1-y_n} y_n(1-y_n)\nabla a_n\right). \tag{5.90}$$

The right hand side can be written as

$$-\sum_{n=1}^{N}\left(t_n(1-y_n)\nabla a_n - (1-t_n)y_n\nabla a_n\right) = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n. \tag{5.91}$$

Then, by 4.13,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\nabla a_n. \tag{5.92}$$

Therefore, by 4.12,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\nabla a_n(\nabla a_n)^{\intercal} + \sum_{n=1}^{N}(y_n - t_n)\nabla\nabla a_n. \tag{5.93}$$

5.20

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

$$\begin{aligned} t_{nk} &\in \{0, 1\}, \\ p(t_{nk} = 1|\mathbf{x}_n) &= y_{nk}, \end{aligned} \tag{5.94}$$

where

$$\begin{aligned} y_{nk} &= y_k(\mathbf{x}_n, \mathbf{w}), \\ \sum_{k=1}^{K} y_{nk} &= 1. \end{aligned} \tag{5.95}$$

Then,

$$p(\mathbf{t}_n|\mathbf{x}_n) = \prod_{k=1}^{K} y_{nk}^{t_{nk}}. \tag{5.96}$$

Let

$$E(\mathbf{w}) = -\ln\left(\prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n)\right). \tag{5.97}$$

Then,

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}. \tag{5.98}$$

If

$$y_{nk} = \frac{\exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}{\sum_{k=1}^{K}\exp\left(a_k(\mathbf{x}_n, \mathbf{w})\right)}, \tag{5.99}$$

then, by 5.7,

$$\frac{\partial E(\mathbf{w})}{\partial a_k} = \sum_{n=1}^{N}(y_{nk} - t_{nk}). \tag{5.100}$$

Then,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla a_k. \tag{5.101}$$

Then,

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}(1 - y_{nk})\nabla a_k\left(\nabla a_k\right)^{\mathsf{T}} + \sum_{n=1}^{N}\sum_{k=1}^{K}(y_{nk} - t_{nk})\nabla\nabla a_k. \tag{5.102}$$

5.21 (Incomplete)

5.22

Let $y_1, \cdots, y_N$ be variables such that

$$\begin{aligned} y_n &= \mathbf{w}_n^{(2)\mathsf{T}}\mathbf{z}, \\ z_m &= h(a_m), \\ a_m &= \mathbf{w}_m^{(1)\mathsf{T}}\mathbf{x}. \end{aligned} \tag{5.103}$$

Let E be a function of $y_1, \cdots, y_N$. Then,

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial z_m}\frac{\partial z_m}{\partial a_m}\frac{\partial a_m}{\partial \mathbf{w}_m^{(1)}}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.
\end{aligned} \tag{5.104}$$

Therefore,

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} w_{nm}^{(2)} h'(a_m)\mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n}\mathbf{z}.
\end{aligned} \tag{5.105}$$

Thus,

$$\begin{aligned}
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= w_{nm}^{(2)}\mathbf{x}\left(\frac{\partial^2 E}{\partial y_n^2} w_{nm'}^{(2)} h'(a_{m'})h'(a_m)\mathbf{x} + \frac{\partial E}{\partial y_n} h''(a_m) I_{mm'}\mathbf{x}\right)^{\intercal}, \\
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= h'(a_m)\mathbf{x}\left(\frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)}\mathbf{z} + \frac{\partial E}{\partial y_n}\mathbf{v}\right)^{\intercal}, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}}\mathbf{z}\mathbf{z}^{\intercal},
\end{aligned} \tag{5.106}$$

where

$$v_m = \begin{cases} 1, & m = n, \\ 0 & \text{otherwise.} \end{cases} \tag{5.107}$$

5.23

Let $y_1, \cdots, y_N$ be variables such that

$$\begin{aligned}
y_n &= \mathbf{w}_n^{(2)\intercal}\mathbf{z} + \mathbf{w}_n^{(0)\intercal}\mathbf{x}, \\
z_m &= h(a_m), \\
a_m &= \mathbf{w}_m^{(1)\intercal}\mathbf{x}.
\end{aligned} \tag{5.108}$$

Let E be a function of $y_1, \cdots, y_N$. Then,

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(0)}}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial z_m} \frac{\partial z_m}{\partial a_m} \frac{\partial a_m}{\partial \mathbf{w}_m^{(1)}}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial \mathbf{w}_n^{(2)}}.
\end{aligned} \tag{5.109}$$

Therefore,

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{w}_n^{(0)}} &= \frac{\partial E}{\partial y_n} \mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_m^{(1)}} &= \frac{\partial E}{\partial y_n} w_{nm}^{(2)} h'(a_m) \mathbf{x}, \\
\frac{\partial E}{\partial \mathbf{w}_n^{(2)}} &= \frac{\partial E}{\partial y_n} \mathbf{z}.
\end{aligned} \tag{5.110}$$

Thus,

$$\begin{aligned}
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(0)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{x}\mathbf{x}^\intercal, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_m^{(1)}} &= \frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} h'(a_m) \mathbf{x}\mathbf{x}^\intercal, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(0)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{x}\mathbf{z}^\intercal, \\
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_{m'}^{(1)}} &= w_{nm}^{(2)} \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm'}^{(2)} h'(a_{m'}) h'(a_m) \mathbf{x} + \frac{\partial E}{\partial y_n} h''(a_m) I_{mm'} \mathbf{x} \right)^\intercal, \\
\frac{\partial^2 E}{\partial \mathbf{w}_m^{(1)} \partial \mathbf{w}_n^{(2)}} &= h'(a_m) \mathbf{x} \left(\frac{\partial^2 E}{\partial y_n^2} w_{nm}^{(2)} \mathbf{z} + \frac{\partial E}{\partial y_n} \mathbf{v} \right)^\intercal, \\
\frac{\partial^2 E}{\partial \mathbf{w}_n^{(2)} \partial \mathbf{w}_{n'}^{(2)}} &= \frac{\partial^2 E}{\partial y_n \partial y_{n'}} \mathbf{z}\mathbf{z}^\intercal,
\end{aligned} \tag{5.111}$$

where

$$v_m = \begin{cases} 1, & m = n, \\ 0 & \text{otherwise.} \end{cases} \tag{5.112}$$

5.24

Let $y_1, \cdots, y_n$ be variables such that

$$\begin{aligned} y_n &= \mathbf{w}_n^\intercal \mathbf{z} + w_{n0}, \\ z_m &= h\left(\mathbf{w}_m^\intercal \mathbf{x} + w_{m0}\right). \end{aligned} \tag{5.113}$$

(a)

Let

$$\tilde{\mathbf{x}} = a\mathbf{x} + b\mathbf{v}, \tag{5.114}$$

where

$$\mathbf{v} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Then,

$$z_m = h\left(\frac{1}{a}\mathbf{w}_m^\intercal \left(\tilde{\mathbf{x}} - b\mathbf{v}\right) + w_{m0}\right). \tag{5.115}$$

Therefore,

$$\tilde{z}_m = h\left(\tilde{\mathbf{w}}_m^\intercal \mathbf{x} + \tilde{w}_{m0}\right). \tag{5.116}$$

where

$$\begin{aligned} \tilde{\mathbf{w}}_m &= \frac{1}{a}\mathbf{w}_m, \\ \tilde{w}_{m0} &= w_{m0} - \frac{b}{a}\mathbf{w}_m^\intercal \mathbf{v}. \end{aligned} \tag{5.117}$$

(b)

Let

$$\tilde{y}_n = cy_n + d. \tag{5.118}$$

Then,

$$\frac{\tilde{y}_n - d}{c} = \mathbf{w}_n^\intercal \mathbf{z} + w_{n0}. \tag{5.119}$$

Therefore,

$$\tilde{y}_n = \tilde{\mathbf{w}}_n^\intercal \mathbf{z} + \tilde{w}_{n0}, \tag{5.120}$$

where

$$\begin{aligned} \tilde{\mathbf{w}}_n &= c\mathbf{w}_n, \\ \tilde{w}_{n0} &= cw_{n0} + d. \end{aligned} \tag{5.121}$$