# Solutions Manual to
# Pattern Recognition and Machine Learning

Hiromichi Inawashiro

August 17, 2024

# Contents

# 1 Introduction

## 1.1

Let

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

To minimise it, setting the derivative to zero gives

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

Substituting

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

gives

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

Therefore,

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.5}$$

where

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

## 1.2

Let

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

To minimise it, setting the derivative to zero gives

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

Substituting

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

gives

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

Therefore,

$$\sum_{j=0}^{M} \tilde{A}_{ij} w_j = T_i \tag{1.11}$$

where

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

## 1.3

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

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \tag{1.13}$$

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

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

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

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

Here,

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

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

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

## 1.4

Let

$$x = g(y) \tag{1.18}$$

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

$$p_y(y) = p_x\left(g(y)\right)|g'(y)| \tag{1.19}$$

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

$$0 = g'(\hat{y})p_x'\left(g\left(\hat{y}\right)\right) + p_x\left(g\left(\hat{y}\right)\right)g''\left(\hat{y}\right). \tag{1.20}$$

Therefore, in general,

$$\hat{x} \neq g\left(\hat{y}\right). \tag{1.21}$$

Here, let us assume that

$$g(y) = ay + b. \tag{1.22}$$

Then, differentiating both sides of the transformation and substituting $y = \hat{y}$ gives

$$0 = p_x'\left(g\left(\hat{y}\right)\right). \tag{1.23}$$

Therefore,

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

## 1.5

By the definition,

$$\mathrm{var} f(x) = \mathrm{E}\left(f(x) - \mathrm{E}f(x)\right)^2. \tag{1.25}$$

The right hand side can be written as

$$\mathrm{E}\left((f(x))^2 - 2f(x)\mathrm{E}f(x) + (\mathrm{E}f(x))^2\right) = \mathrm{E}\left(f(x)\right)^2 - (\mathrm{E}f(x))^2. \tag{1.26}$$

Therefore,

$$\mathrm{var} f(x) = \mathrm{E}\left(f(x)\right)^2 - (\mathrm{E}f(x))^2. \tag{1.27}$$

## 1.6

By the definition,

$$\mathrm{cov}(x, y) = \mathrm{E}\left((x - \mathrm{E}x)\left(y - \mathrm{E}y\right)\right). \tag{1.28}$$

The right hand side can be written as

$$\mathrm{E}xy - \mathrm{E}\left(x\mathrm{E}y\right) - \mathrm{E}\left(y\mathrm{E}x\right) + \mathrm{E}\left(\mathrm{E}x\mathrm{E}y\right) = \mathrm{E}xy - \mathrm{E}x\mathrm{E}y. \tag{1.29}$$

The right hand side can be written as

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

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

$$f(x, y) = f(x)f(y). \tag{1.31}$$

Then,

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

Therefore,

$$\mathrm{cov}(x, y) = 0. \tag{1.33}$$

## 1.7

Let

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

Then

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

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

$$\int_{0}^{\infty}\int_{0}^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} drd\theta = 2\pi \int_{0}^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) rdr. \tag{1.36}$$

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

$$2\pi\sigma^2 \int_{0}^{\infty} \exp\left(-\frac{1}{2}s^2\right) sds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_{0}^{\infty}. \tag{1.37}$$

Therefore,

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

By the definition,

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

Then

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

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

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

Therefore,

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

## 1.8

Let $x$ be a variable such that

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

Then

$$\mathrm{E}x = \int_{-\infty}^{\infty} x\mathcal{N}\left(x|\mu, \sigma^2\right) dx. \tag{1.44}$$

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

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

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

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

Since

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

and

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

we have

$$\mathrm{E}x = \mu. \tag{1.49}$$

By the definition,

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

can be written as

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

Differentiating both sides with respect to $\sigma^2$ gives

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

The left hand side can be written as

$$-\frac{1}{2}\left(\sigma^2\right)^{-1}\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)dx+\frac{1}{2}\left(\sigma^2\right)^{-2}\int_{-\infty}^{\infty}(x-\mu)^2\mathcal{N}\left(x|\mu,\sigma^2\right)dx$$
$$=-\frac{1}{2}\left(\sigma^2\right)^{-1}+\frac{1}{2}\left(\sigma^2\right)^{-2}\mathrm{var}x. \tag{1.53}$$

Therefore,

$$\mathrm{var}x=\sigma^2. \tag{1.54}$$

## 1.9

Let

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

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

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

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

Similarly, let

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

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

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

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

## 1.10

By the definition,

$$\mathrm{E}(x+y)=\int\int(x+y)p(x,y)dxdy. \tag{1.59}$$

The right hand side can be written as

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

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

$$\mathrm{E}x + \mathrm{E}y. \tag{1.61}$$

Therefore,

$$\mathrm{E}(x+y) = \mathrm{E}x + \mathrm{E}y. \tag{1.62}$$

Similarly, by the definition,

$$\mathrm{var}(x+y) = \mathrm{E}\left(x + y - \mathrm{E}(x+y)\right)^2 \tag{1.63}$$

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

$$\begin{aligned} \mathrm{E}\left(x - \mathrm{E}x\right)^2 + 2\mathrm{E}\left(\left(x - \mathrm{E}x\right)\left(y - \mathrm{E}y\right)\right) + \mathrm{E}\left(y - \mathrm{E}y\right)^2 \\ = \mathrm{var}x + 2\mathrm{cov}(x,y) + \mathrm{var}y. \end{aligned} \tag{1.64}$$

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

$$\mathrm{cov}(x,y) = 0, \tag{1.65}$$

by 1.6. Therefore,

$$\mathrm{var}(x+y) = \mathrm{var}x + \mathrm{var}y. \tag{1.66}$$

## 1.11

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

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{1.67}$$

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

$$\begin{aligned} 0 &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu), \\ 0 &= -\frac{N}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{n=1}^{N}(x_n - \mu)^2. \end{aligned} \tag{1.68}$$

Therefore,

$$
\begin{aligned}
\mu_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} x_n, \\
\sigma^2_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2.
\end{aligned} \tag{1.69}
$$

## 1.12

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

$$
\mathrm{E}x_m x_n = \mathrm{E}x_m \mathrm{E}x_n. \tag{1.70}
$$

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

$$
\mathrm{E}x_n^2 = \mathrm{var}x_n + \left(\mathrm{E}x_n\right)^2. \tag{1.71}
$$

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

$$
\mathrm{E}x_m x_n = \mu^2 + \delta_{mn}\sigma^2. \tag{1.72}
$$

Here, since

$$
\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n, \tag{1.73}
$$

we have

$$
\mathrm{E}\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}x_n. \tag{1.74}
$$

Therefore,

$$
\mathrm{E}\mu_{\mathrm{ML}} = \mu. \tag{1.75}
$$

Similarly, since

$$
\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \mu_{\mathrm{ML}}\right)^2, \tag{1.76}
$$

we have

$$
\mathrm{E}\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left(x_n - \mu_{\mathrm{ML}}\right)^2. \tag{1.77}
$$

The right hand side can be writen as

$$\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(x_n^2-2\mu_{\mathrm{ML}}x_n+\mu_{\mathrm{ML}}^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}x_n^2-\frac{2}{N}\mathrm{E}\left(\mu_{\mathrm{ML}}\left(\sum_{n=1}^{N}x_n\right)\right)+\mathrm{E}\mu_{\mathrm{ML}}^2. \tag{1.78}$$

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

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2\right)=\mu^2+\sigma^2, \tag{1.79}$$

while the second and third terms can be writen as

$$-2\mathrm{E}\mu_{\mathrm{ML}}^2+\mathrm{E}\mu_{\mathrm{ML}}^2=-\mathrm{E}\mu_{\mathrm{ML}}^2. \tag{1.80}$$

Here,

$$\mathrm{E}\mu_{\mathrm{ML}}^2=\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2. \tag{1.81}$$

The right hand side can be written as

$$\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}x_n^2+\frac{2}{N^2}\sum_{1\leq m<n\leq N}\mathrm{E}x_mx_n=\frac{1}{N}(\mu^2+\sigma^2)+\frac{N-1}{N}\mu^2. \tag{1.82}$$

Therefore,

$$\mathrm{E}\mu_{\mathrm{ML}}^2=\mu^2+\frac{1}{N}\sigma^2. \tag{1.83}$$

Thus,

$$\mathrm{E}\sigma_{\mathrm{ML}}^2=\frac{N-1}{N}\sigma^2. \tag{1.84}$$

## 1.13

Let $\{x_n\}$ be a set of variables whose mean is $\mu$ and variance is $\sigma^2$. Then

$$\mathrm{E}\left(\frac{1}{N}\sum_{n=1}^{N}\left(x_n-\mu\right)^2\right)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\left(x_n-\mu\right)^2. \tag{1.85}$$

The right hand side can be writen as

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

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

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2\right)=\mu^2+\sigma^2, \tag{1.87}$$

while the second term can be writen as

$$-\frac{2\mu}{N}\sum_{n=1}^{N}\mu=-2\mu^2. \tag{1.88}$$

Therefore,

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

## 1.14

Let

$$\begin{aligned} w_{ij}^{\mathrm{S}} &= \frac{1}{2}(w_{ij}+w_{ji}),\\ w_{ij}^{\mathrm{A}} &= \frac{1}{2}(w_{ij}-w_{ji}). \end{aligned} \tag{1.90}$$

Then

$$\begin{aligned} w_{ij} &= w_{ij}^{\mathrm{S}}+w_{ij}^{\mathrm{A}},\\ w_{ij}^{\mathrm{S}} &= w_{ji}^{\mathrm{S}},\\ w_{ij}^{\mathrm{A}} &= -w_{ji}^{\mathrm{A}}. \end{aligned} \tag{1.91}$$

Here,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_ix_j=\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}-w_{ji})x_ix_j. \tag{1.92}$$

The right hand side can be written as

$$\frac{1}{2}\left(\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}x_ix_j-\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ji}x_ix_j\right)=0. \tag{1.93}$$

Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_ix_j=0. \tag{1.94}$$

Additionally,

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} \left(w_{ij}^{\mathrm{S}} + w_{ij}^{\mathrm{A}}\right) x_i x_j. \tag{1.95}$$

The right hand side can be written as

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{A}} x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}} x_i x_j, \tag{1.96}$$

where the result above is used. Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^{\mathrm{S}} x_i x_j. \tag{1.97}$$

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

## 1.15 (Incomplete)

## 1.16 (Incomplete)

## 1.17

Let

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \tag{1.98}$$

Then

$$\Gamma(x+1) = \int_0^\infty u^{x} \exp(-u) du. \tag{1.99}$$

The right hand side can be written as

$$\left[-u^x \exp(-u)\right]_{u=0}^{u=\infty} + \int_0^\infty x u^{x-1} \exp(-u) du = x\Gamma(x). \tag{1.100}$$

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.101}$$

Since

$$\Gamma(1) = \int_0^1 \exp(-u) du, \tag{1.102}$$

and the right hand side can be written as 1,

$$\Gamma(1) = 0!. \tag{1.103}$$

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

$$\Gamma(x) = (x-1)!. \tag{1.104}$$

Then,

$$\Gamma(x+1) = x\Gamma(x), \tag{1.105}$$

where the right hand side can be written as

$$x(x-1)! = x!. \tag{1.106}$$

Therefore,

$$\Gamma(x+1) = x!. \tag{1.107}$$

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

## 1.18

Let us consider the transformation from Cartesian to polar coordinates

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} \exp\left(-x_i^2\right) dx_i = S_D \int_0^{\infty} \exp\left(-r^2\right) r^{D-1} dr, \tag{1.108}$$

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

$$\frac{S_D}{2} \int_0^{\infty} \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \tag{1.109}$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \tag{1.110}$$

Additionally, the volume of the sphere can can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \tag{1.111}$$

The right hand side can be written as

$$S_D \left[ \frac{r^D}{D} \right]_{r=0}^{r=1} = \frac{S_D}{D}. \tag{1.112}$$

Therefore,

$$V_D = \frac{S_D}{D}. \tag{1.113}$$

Finally, the results above reduce to

$$\begin{aligned} S_2 &= \frac{2\pi}{\Gamma(1)}, \\ V_2 &= \frac{S_2}{2}. \end{aligned} \tag{1.114}$$

Therefore,

$$\begin{aligned} S_2 &= 2\pi, \\ V_2 &= \pi. \end{aligned} \tag{1.115}$$

Similarly,

$$\begin{aligned} S_3 &= \frac{2\pi^{\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)}, \\ V_3 &= \frac{S_3}{3}. \end{aligned} \tag{1.116}$$

Therefore,

$$\begin{aligned} S_3 &= 4\pi, \\ V_3 &= \frac{4}{3}\pi. \end{aligned} \tag{1.117}$$

## 1.19

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

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} \Gamma\left(\frac{D}{2}\right)}, \tag{1.118}$$

by 1.18.

Additionally, by Stering's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}} \exp(-x) x^{\frac{x+1}{2}}, \tag{1.119}$$

the ratio can be approximated as

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} (2\pi)^{\frac{1}{2}} \exp\left(1 - \frac{D}{2}\right) \left(\frac{D}{2} - 1\right)^{\frac{D}{4}}}. \tag{1.120}$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}} \frac{1}{D} \left( \frac{e^2 \pi^2}{8D - 16} \right)^{\frac{D}{4}}. \tag{1.121}$$

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

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

$$\frac{\sqrt{\sum_{i=1}^{D} 1^2}}{1} = \sqrt{D}. \tag{1.122}$$

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

## 1.20

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

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left( -\frac{\|\mathbf{x}\|^2}{2\sigma^2} \right). \tag{1.123}$$

Integrating both sides from $\|\mathbf{x}\| = r$ to $\|\mathbf{x}\| = r + \epsilon$ gives

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_{r}^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left( -\frac{r'^2}{2\sigma^2} \right) J dr' d\boldsymbol{\phi}, \tag{1.124}$$

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 $\epsilon$, the right hand side can be approximated as

$$\begin{aligned}
&(2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon} \int J dr' d\boldsymbol{\phi} \\
&= (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} d\mathbf{x}.
\end{aligned} \tag{1.125}$$

Therefore,

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} \simeq p(r)\epsilon, \tag{1.126}$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{1.127}$$

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

Secondly, to maximise $p(r)$, setting the derivative to zero gives

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left((D-1)r^{D-2} - \frac{r^D}{\sigma^2}\right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{1.128}$$

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

$$\hat{r} = \sqrt{D-1}\sigma. \tag{1.129}$$

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

$$\frac{p(\hat{r}+\epsilon)}{p(\hat{r})} = \left(\frac{\hat{r}+\epsilon}{\hat{r}}\right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right). \tag{1.130}$$

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

$$\begin{aligned}
&\exp\left((D-1)\ln\left(1+\frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{\hat{r}^2}{\sigma^2}\ln\left(1+\frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right).
\end{aligned} \tag{1.131}$$

By the Taylor series

$$\ln(1+x) = x - \frac{1}{2}x^2 + o\left(x^3\right), \tag{1.132}$$

the right hand side can be approximated as

$$\exp\left(\frac{\hat{r}^2}{\sigma^2}\left(\frac{\epsilon}{\hat{r}}-\frac{\epsilon^2}{2\hat{r}^2}\right)-\frac{2\hat{r}\epsilon+\epsilon^2}{2\sigma^2}\right)=\exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.133}$$

Therefore,

$$p\left(\hat{r}+\epsilon\right)\simeq p\left(\hat{r}\right)\exp\left(-\frac{\epsilon^2}{\sigma^2}\right). \tag{1.134}$$

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

$$\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)}=\exp\left(\frac{\hat{r}^2}{2\sigma^2}\right). \tag{1.135}$$

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

$$\frac{p(\mathbf{0})}{p\left(\hat{\mathbf{r}}\right)}=\exp\left(\frac{D-1}{2}\right). \tag{1.136}$$

## 1.21

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

$$0\leq a(b-a). \tag{1.137}$$

Therefore,

$$a\leq (ab)^{\frac{1}{2}}. \tag{1.138}$$

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

$$p(\mathbf{x},\mathcal{C}_1)>p(\mathbf{x},\mathcal{C}_2)\Rightarrow \mathbf{x}\in\mathcal{C}_1, \tag{1.139}$$

and

$$p(\mathbf{x},\mathcal{C}_2)>p(\mathbf{x},\mathcal{C}_1)\Rightarrow \mathbf{x}\in\mathcal{C}_2. \tag{1.140}$$

Then, using the inequality above,

$$\int_{\mathcal{R}_1}p(\mathbf{x},\mathcal{C}_2)d\mathbf{x}\leq\int_{\mathcal{R}_1}\left(p(\mathbf{x},\mathcal{C}_1)p(\mathbf{x},\mathcal{C}_2)\right)^{\frac{1}{2}}d\mathbf{x}, \tag{1.141}$$

and

$$\int_{\mathcal{R}_2}p(\mathbf{x},\mathcal{C}_1)d\mathbf{x}\leq\int_{\mathcal{R}_2}\left(p(\mathbf{x},\mathcal{C}_1)p(\mathbf{x},\mathcal{C}_2)\right)^{\frac{1}{2}}d\mathbf{x}. \tag{1.142}$$

Therefore,

$$\int_{\mathcal{R}_1}p(\mathbf{x},\mathcal{C}_2)d\mathbf{x}+\int_{\mathcal{R}_2}p(\mathbf{x},\mathcal{C}_1)d\mathbf{x}\leq\int\left(p(\mathbf{x},\mathcal{C}_1)p(\mathbf{x},\mathcal{C}_2)\right)^{\frac{1}{2}}d\mathbf{x}. \tag{1.143}$$

## 1.22

Let

$$\mathrm{E}L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.144}$$

If

$$L_{kj} = 1 - \delta_{kj}, \tag{1.145}$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} \left(p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)\right) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \tag{1.146}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \left(p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)\right) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \tag{1.147}$$

Therefore,

$$\mathrm{E}L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.148}$$

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

## 1.23

Let

$$\mathrm{E}L = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \tag{1.149}$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.150}$$

Therefore,

$$\mathrm{E}L = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.151}$$

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

## 1.24 (Incomplete)

## 1.25

Let

$$\mathrm{E}L\left(\mathbf{t},\mathbf{y}(\mathbf{x})\right)=\int\int\|\mathbf{y}(\mathbf{x})-\mathbf{t}\|^2 p(\mathbf{x},\mathbf{t})d\mathbf{x}d\mathbf{t}. \tag{1.152}$$

Then

$$\frac{\delta \mathrm{E}L\left(\mathbf{t},\mathbf{y}(\mathbf{x})\right)}{\delta \mathbf{y}(\mathbf{x})}=2\int\left(\mathbf{y}(\mathbf{x})-\mathbf{t}\right)p(\mathbf{x},\mathbf{t})d\mathbf{t}. \tag{1.153}$$

To minimise $\mathrm{E}L\left(\mathbf{t},\mathbf{y}(\mathbf{x})\right)$, setting the left hand side to zero gives

$$\mathbf{0}=\int\left(\mathbf{y}(\mathbf{x})-\mathbf{t}\right)p(\mathbf{t}|\mathbf{x})d\mathbf{t}. \tag{1.154}$$

The right hand side can be written as

$$\mathbf{y}(\mathbf{x})\int p(\mathbf{t}|\mathbf{x})d\mathbf{t}-\int \mathbf{t}p(\mathbf{t}|\mathbf{x})d\mathbf{t}=\mathbf{y}(\mathbf{x})-\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.155}$$

Thus,

$$\mathbf{y}(\mathbf{x})=\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.156}$$

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

$$\mathbf{y}(\mathbf{x})=\mathrm{E}_{t}(t|\mathbf{x}). \tag{1.157}$$

## 1.26

Let

$$\mathrm{E}L\left(\mathbf{t},\mathbf{y}(\mathbf{x})\right)=\int\int\|\mathbf{y}(\mathbf{x})-\mathbf{t}\|^2 p(\mathbf{x},\mathbf{t})d\mathbf{x}d\mathbf{t}. \tag{1.158}$$

The right hand side can be written as

$$\begin{aligned}
&\int\int\|\mathbf{y}(\mathbf{x})-\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})+\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})-\mathbf{t}\|^2 p(\mathbf{x},\mathbf{t})d\mathbf{x}d\mathbf{t}\\
&=\int\int\|\mathbf{y}(\mathbf{x})-\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x},\mathbf{t})d\mathbf{x}d\mathbf{t}\\
&+2\int\int\left(\mathbf{y}(\mathbf{x})-\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\right)^{\intercal}\left(\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})-\mathbf{t}\right)p(\mathbf{x},\mathbf{t})d\mathbf{x}d\mathbf{t}\\
&+\int\int\|\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})-\mathbf{t}\|^2 p(\mathbf{x},\mathbf{t})d\mathbf{x}d\mathbf{t}.
\end{aligned} \tag{1.159}$$

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

$$\int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 \left( \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \tag{1.160}$$

The second term can be written as

$$2 \int (\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\intercal \left( \int (\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t})\, p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \tag{1.161}$$

Since

$$\begin{aligned} \int \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t}}{p(\mathbf{x})}, \\ \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}), \end{aligned} \tag{1.162}$$

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

$$\int \left( \int \|\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x} = \int \mathrm{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.163}$$

Therefore,

$$\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \int \mathrm{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.164}$$

Thus, $\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right)$ is mimimised if

$$\mathbf{y}(\mathbf{x}) = \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.165}$$

## 1.27

Let

$$\mathrm{E}L_q = \int\int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \tag{1.166}$$

Then

$$\frac{\delta \mathrm{E}L_q}{\delta y(\mathbf{x})} = \int q |y(\mathbf{x}) - t|^{q-1} \mathrm{sign}(y(\mathbf{x}) - t) p(\mathbf{x}, t) dt. \tag{1.167}$$

To minimise $\mathrm{E}L_q$, setting the left hand side to zero gives

$$0 = \int |y(\mathbf{x}) - t|^{q-1} \mathrm{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt. \tag{1.168}$$

This is the condition that $y(\mathbf{x})$ must satisfy in order to minimise $\mathrm{E}L_q$.

If $q = 1$, the condition can be written as

$$0 = \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})dt. \tag{1.169}$$

Therefore, $y(\mathbf{x})$ is given by the conditional median.

## 1.28

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \tag{1.170}$$

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

$$h\left(p^2\right) = h(p) + h(p). \tag{1.171}$$

Therefore,

$$h\left(p^2\right) = 2h(p). \tag{1.172}$$

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

$$h\left(p^n\right) = nh(p). \tag{1.173}$$

Then, by the first assumption,

$$h\left(p^{n+1}\right) = h\left(p^n\right) + h(p). \tag{1.174}$$

Therefore,

$$h\left(p^{n+1}\right) = (n+1)h(p). \tag{1.175}$$

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

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

$$h\left(p^n\right) = h\left(p^{\frac{n}{m}m}\right). \tag{1.176}$$

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

$$h\left(p^{\frac{n}{m}}\right) = \frac{n}{m}h(p). \tag{1.177}$$

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

$$h\left(p^a\right) = ah(p). \tag{1.178}$$

Differentiating both sides with respect to $a$ and substituting $a = 1$ gives

$$(p \ln p)h'(p) = h(p). \tag{1.179}$$

Therefore,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + C, \tag{1.180}$$

where $C$ is a constant. Ignorting the constants, the left hand side can be written as $\ln h(p)$ and the right hand side can be written as $\ln(\ln p)$. Thus,

$$h(p) \propto \ln p. \tag{1.181}$$

## 1.29

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

$$\mathrm{H}(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i), \tag{1.182}$$

where

$$\sum_{i=1}^{M} p(x_i) = 1. \tag{1.183}$$

By Jensen's inequality,

$$\sum_{i=1}^{M} p(x_i) \ln \frac{1}{p(x_i)} \leq \ln \left( \sum_{i=1}^{M} 1 \right). \tag{1.184}$$

Therefore,

$$\mathrm{H}(x) \leq \ln M. \tag{1.185}$$

## 1.30

Let

$$p(x) = \mathcal{N}\left(x|\mu, \sigma^2\right),$$
$$q(x) = \mathcal{N}\left(x|m, s^2\right). \tag{1.186}$$

Then, by the definition,

$$\mathrm{KL}\left(p||q\right) = -\int p(x) \ln \frac{q(x)}{p(x)} dx. \tag{1.187}$$

The right hand side can be written as

$$\begin{aligned} &-\int_{-\infty}^{\infty} p(x) \ln \frac{\left(2\pi s^2\right)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{\left(2\pi \sigma^2\right)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\ &= -\int_{-\infty}^{\infty} p(x) \left(-\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2}\right) dx. \end{aligned} \tag{1.188}$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x) dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x) dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx. \tag{1.189}$$

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

$$\frac{1}{2s^2} \int_{-\infty}^{\infty} (x-\mu+\mu-m)^2 p(x) dx = \frac{\sigma^2 + (\mu-m)^2}{2s^2}. \tag{1.190}$$

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

$$\mathrm{KL}\left(p||q\right) = \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu-m)^2}{2s^2} - \frac{1}{2}. \tag{1.191}$$

## 1.31

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

$$\begin{aligned} \mathrm{H}(\mathbf{x}) &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}, \\ \mathrm{H}(\mathbf{y}) &= -\int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \\ \mathrm{H}(\mathbf{x}, \mathbf{y}) &= -\int\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \end{aligned} \tag{1.192}$$

Note that

$$\begin{aligned}\mathrm{H}(\mathbf{x}) &= -\int\left(\int p(\mathbf{x},\mathbf{y})d\mathbf{y}\right)\ln p(\mathbf{x})d\mathbf{x},\\ \mathrm{H}(\mathbf{y}) &= -\int\left(\int p(\mathbf{x},\mathbf{y})d\mathbf{x}\right)\ln p(\mathbf{y})d\mathbf{y}.\end{aligned}\tag{1.193}$$

Therefore,

$$\mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{x},\mathbf{y}) = -\int\int p(\mathbf{x},\mathbf{y})\ln\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}d\mathbf{x}d\mathbf{y}.\tag{1.194}$$

Since

$$\int\int p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} = 1,\tag{1.195}$$

Jensen's inequality can be used to write that

$$-\int\int p(\mathbf{x},\mathbf{y})\ln\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}d\mathbf{x}d\mathbf{y} \geq -\ln\left(\int\int p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y}\right).\tag{1.196}$$

The right hand side can be written as

$$-\ln\left(\int p(\mathbf{x})d\mathbf{x}\int p(\mathbf{y})d\mathbf{y}\right) = 0.\tag{1.197}$$

Thus,

$$\mathrm{H}(\mathbf{x},\mathbf{y}) \leq \mathrm{H}(\mathbf{x}) + \mathrm{H}(\mathbf{y}).\tag{1.198}$$

## 1.32

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

$$\mathbf{y} = \mathbf{A}\mathbf{x},\tag{1.199}$$

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

$$\mathrm{H}(\mathbf{y}) = -\int p_y(\mathbf{y})\ln p_y(\mathbf{y})d\mathbf{y}.\tag{1.200}$$

By the transformation

$$p_y(\mathbf{y}) = p_x(\mathbf{A}\mathbf{x})\left|\det\mathbf{A}^{-1}\right|,\tag{1.201}$$

the right hand side can be written as

$$-\int p_x(\mathbf{Ax}) \ln p_x(\mathbf{Ax}) |\det \mathbf{A}| d\mathbf{x} - \ln \left|\det \mathbf{A}^{-1}\right| \int p_y(\mathbf{y}) d\mathbf{y}. \tag{1.202}$$

By the transformation

$$\mathbf{x}' = \mathbf{Ax}, \tag{1.203}$$

the first term can be written as

$$-\int p_x(\mathbf{x}') \ln p_x(\mathbf{x}') d\mathbf{x}' = \mathrm{H}(\mathbf{x}), \tag{1.204}$$

and the second term can be written as

$$-\ln \left|\det \mathbf{A}^{-1}\right| = \ln |\det \mathbf{A}|. \tag{1.205}$$

Therefore,

$$\mathrm{H}(\mathbf{y}) = \mathrm{H}(\mathbf{x}) + \ln |\det \mathbf{A}|. \tag{1.206}$$

## 1.33

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

$$\mathrm{H}(y|x) = -\sum_i \sum_j p(x_i, y_j) \ln p(y_j|x_i). \tag{1.207}$$

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

$$0 = -\sum_i p(x_i) \sum_j p(y_j|x_i) \ln p(y_j|x_i). \tag{1.208}$$

Since

$$\begin{aligned} p(x_i) &\geq 0, \\ p(y_j|x_i) \ln p(y_j|x_i) &\leq 0. \end{aligned} \tag{1.209}$$

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

$$p(y_j|x_i) \ln p(y_j|x_i) = 0. \tag{1.210}$$

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

$$\sum_j p(y_j|x_i) = 1, \tag{1.211}$$

it can be written that

$$p(y_j|x_i) = \delta_{jj'(i)}, \tag{1.212}$$

where $j'(i)$ is unique for each $i$.

## 1.34

Let

$$\begin{aligned} L(p(x)) = &- \int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1 \left( \int_{-\infty}^{\infty} p(x) dx - 1 \right) \\ &+ \lambda_2 \left( \int_{-\infty}^{\infty} x p(x) dx - \mu \right) + \lambda_3 \left( \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx - \sigma^2 \right). \end{aligned} \tag{1.213}$$

Then

$$\frac{\delta L(p(x))}{\delta p(x)} = - \ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2. \tag{1.214}$$

Setting the left hand side to zero gives

$$p(x) = \exp\left(-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right). \tag{1.215}$$

Therefore,

$$p(x) = \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} + \lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right). \tag{1.216}$$

Substituting it to

$$\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= 1, \\ \int_{-\infty}^{\infty} x p(x) dx &= \mu, \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx &= \sigma^2, \end{aligned} \tag{1.217}$$

gives

$$\begin{aligned} \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx &= 1, \\ \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} x \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx &= \mu, \\ \exp\left(-1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3}\right) \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(\lambda_3 \left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right)^2\right) dx &= \sigma^2. \end{aligned} \tag{1.218}$$

By the transformation

$$y = \sqrt{-\lambda_3}\left(x - \left(\mu - \frac{\lambda_2}{2\lambda_3}\right)\right), \tag{1.219}$$

they can be written as

$$\begin{aligned}
\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\int_{-\infty}^{\infty}\exp\left(-y^2\right)(-\lambda_3)^{-\frac{1}{2}}dy &= 1,\\
\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\int_{-\infty}^{\infty}\left((-\lambda_3)^{-\frac{1}{2}}y+\mu-\frac{\lambda_2}{2\lambda_3}\right)\exp\left(-y^2\right)(-\lambda_3)^{-\frac{1}{2}}dy &= \mu,\\
\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\int_{-\infty}^{\infty}\left((-\lambda_3)^{-\frac{1}{2}}y-\frac{\lambda_2}{2\lambda_3}\right)^2\exp\left(-y^2\right)(-\lambda_3)^{-\frac{1}{2}}dy &= \sigma^2.
\end{aligned} \tag{1.220}$$

Since

$$\begin{aligned}
\int_{-\infty}^{\infty}\exp\left(-y^2\right)dy &= \Gamma\left(\frac{1}{2}\right),\\
\int_{-\infty}^{\infty}y\exp\left(-y^2\right)dy &= 0,\\
\int_{-\infty}^{\infty}y^2\exp\left(-y^2\right)dy &= \Gamma\left(\frac{3}{2}\right),
\end{aligned} \tag{1.221}$$

they can be written as

$$\begin{aligned}
\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right) &= 1,\\
\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\left(\mu-\frac{\lambda_2}{2\lambda_3}\right)(-\lambda_3)^{-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right) &= \mu,\\
\exp\left(-1+\lambda_1-\frac{\lambda_2^2}{4\lambda_3}\right)\left((-\lambda_3)^{-\frac{3}{2}}\Gamma\left(\frac{3}{2}\right)+(-\lambda_3)^{-\frac{1}{2}}\frac{\lambda_2^2}{4\lambda_3^2}\Gamma\left(\frac{1}{2}\right)\right) &= \sigma^2.
\end{aligned} \tag{1.222}$$

Therefore,

$$\begin{aligned}
\lambda_1 &= 1-\frac{1}{2}\ln\left(2\pi\sigma^2\right),\\
\lambda_2 &= 0,\\
\lambda_3 &= -\frac{1}{2\sigma^2}.
\end{aligned} \tag{1.223}$$

Thus,

$$p(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{1.224}$$

## 1.35

Let $x$ be a variable such that

$$p(x) = \mathcal{N}(x|\mu, \sigma^2). \tag{1.225}$$

Then, by the definition,

$$\mathrm{H}(x) = -\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) \ln \mathcal{N}\left(x|\mu, \sigma^2\right) dx. \tag{1.226}$$

The right hand side can be written as

$$\begin{aligned}
&-\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) \left(-\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\
&= \frac{1}{2}\ln\left(2\pi\sigma^2\right) \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) dx + \frac{1}{2\sigma^2}\int_{-\infty}^{\infty} (x-\mu)^2 \mathcal{N}\left(x|\mu, \sigma^2\right) dx.
\end{aligned} \tag{1.227}$$

Therefore,

$$\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln\left(2\pi\sigma^2\right)\right). \tag{1.228}$$

## 1.36 (Incomplete)

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

$$f\left(\lambda a + (1-\lambda)b\right) \leq \lambda f(a) + (1-\lambda)f(b), \tag{1.229}$$

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

$$x = \lambda a + (1-\lambda)b. \tag{1.230}$$

Then, the inequality can be written as

$$f(x) \leq \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b). \tag{1.231}$$

Let

$$g(x) = \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b) - f(x). \tag{1.232}$$

Then,

$$g(x) \geq 0. \tag{1.233}$$

Additionally, for $x > a$,

$$g(x) = (x - a)\left(\frac{f(b) - f(a)}{b - a} - \frac{f(x) - f(a)}{x - a}\right). \tag{1.234}$$

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

$$\begin{aligned} f'(c) &= \frac{f(b) - f(a)}{b - a}, \\ f'(y) &= \frac{f(x) - f(a)}{x - a}. \end{aligned} \tag{1.235}$$

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

$$f'(y) \leq f'(c). \tag{1.236}$$

## 1.37

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

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) = -\int\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \tag{1.237}$$

The right hand side can be written as

$$\begin{aligned} &-\int\int p(\mathbf{x}, \mathbf{y}) \left(\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\right) d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int \left(\int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}\right) \ln p(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{1.238}$$

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

$$\mathrm{H}(\mathbf{x}, \mathbf{y}) = \mathrm{H}(\mathbf{y}|\mathbf{x}) + \mathrm{H}(\mathbf{x}). \tag{1.239}$$

## 1.38

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

$$f\left(\lambda x_1 + (1 - \lambda)x_2\right) \leq \lambda f(x_1) + (1 - \lambda) f(x_2), \tag{1.240}$$

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

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i), \tag{1.241}$$

where $\lambda_i \geq 0$ and

$$\sum_{i=1}^{M} \lambda_i = 1. \tag{1.242}$$

Here, let $\lambda_i \geq 0$ and

$$\sum_{i=1}^{M+1} \lambda_i = 1. \tag{1.243}$$

Then, by the definition,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) f\left(\sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right). \tag{1.244}$$

By the assumption,

$$f\left(\sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} x_i\right) \leq \sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \tag{1.245}$$

Therefore,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{i=1}^{M} \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \tag{1.246}$$

Thus,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \sum_{i=1}^{M+1} \lambda_i f(x_i). \tag{1.247}$$

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

## 1.39

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

$$\begin{aligned} p(x=0,y=0) &= \frac{1}{3}, \\ p(x=0,y=1) &= \frac{1}{3}, \\ p(x=1,y=0) &= 0, \\ p(x=1,y=1) &= \frac{1}{3}. \end{aligned} \tag{1.248}$$

### (a)

By the definition,

$$\mathrm{H}(x) = -\sum p(x)\ln p(x). \tag{1.249}$$

By the distribution,

$$\begin{aligned} p(x=0) &= \frac{2}{3}, \\ p(x=1) &= \frac{1}{3}. \end{aligned} \tag{1.250}$$

Therefore,

$$\mathrm{H}(x) = \ln 3 - \frac{2}{3}\ln 2. \tag{1.251}$$

### (b)

By the definition,

$$\mathrm{H}(y) = -\sum p(y)\ln p(y). \tag{1.252}$$

By the distribution,

$$\begin{aligned} p(y=0) &= \frac{1}{3}, \\ p(y=1) &= \frac{2}{3}. \end{aligned} \tag{1.253}$$

Therefore,

$$\mathrm{H}(y) = \ln 3 - \frac{2}{3}\ln 2. \tag{1.254}$$

**(c)**

By the definition,

$$\mathrm{H}(y|x) = -\sum p(x, y) \ln p(y|x). \tag{1.255}$$

By the definition,

$$\begin{aligned}
p(y=0|x=0) &= \frac{p(x=0, y=0)}{p(x=0)}, \\
p(y=0|x=1) &= \frac{p(x=1, y=0)}{p(x=1)}, \\
p(y=1|x=0) &= \frac{p(x=0, y=1)}{p(x=0)}, \\
p(y=1|x=1) &= \frac{p(x=1, y=1)}{p(x=1)}.
\end{aligned} \tag{1.256}$$

Then, by the distribution,

$$\begin{aligned}
p(y=0|x=0) &= \frac{1}{2}, \\
p(y=0|x=1) &= 0, \\
p(y=1|x=0) &= \frac{1}{2}, \\
p(y=1|x=1) &= 1.
\end{aligned} \tag{1.257}$$

Therefore,

$$\mathrm{H}(y|x) = \frac{2}{3} \ln 2. \tag{1.258}$$

**(d)**

By the definition,

$$\mathrm{H}(x|y) = -\sum p(x, y) \ln p(x|y). \tag{1.259}$$

By the definition,

$$\begin{aligned}
p(x=0|y=0) &= \frac{p(x=0,y=0)}{p(y=0)}, \\
p(x=0|y=1) &= \frac{p(x=0,y=1)}{p(y=1)}, \\
p(x=1|y=0) &= \frac{p(x=1,y=0)}{p(y=0)}, \\
p(x=1|y=1) &= \frac{p(x=1,y=1)}{p(y=1)}.
\end{aligned} \tag{1.260}$$

Then, by the distribution,

$$\begin{aligned}
p(x=0|y=0) &= 1, \\
p(x=0|y=1) &= \frac{1}{2}, \\
p(x=1|y=0) &= 0, \\
p(x=1|y=1) &= \frac{1}{2}.
\end{aligned} \tag{1.261}$$

Therefore,

$$\mathrm{H}(x|y) = \frac{2}{3}\ln 2. \tag{1.262}$$

**(e)**

By the definition,

$$\mathrm{H}(x,y) = -\sum p(x,y)\ln p(x,y). \tag{1.263}$$

Therefore,

$$\mathrm{H}(x,y) = \ln 3. \tag{1.264}$$

**(f)**

By the definition,

$$\mathrm{I}(x,y) = -\sum p(x,y)\ln\frac{p(x)p(y)}{p(x,y)}. \tag{1.265}$$

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

$$\mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x, y). \tag{1.266}$$

Therefore,

$$\mathrm{I}(x, y) = \ln 3 - \frac{4}{3} \ln 2. \tag{1.267}$$

## 1.40

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

$$\sum_{i=1}^{M} \lambda_i = 1. \tag{1.268}$$

By Jensen's inequality,

$$\sum_{i=1}^{M} \lambda_i \ln x_i \leq \ln \left( \sum_{i=1}^{M} \lambda_i x_i \right). \tag{1.269}$$

Therefore,

$$\prod_{i=1}^{M} x_i^{\lambda_i} \leq \sum_{i=1}^{M} \lambda_i x_i. \tag{1.270}$$

Substituting

$$\lambda_i = \frac{1}{M} \tag{1.271}$$

gives

$$\left( \prod_{i=1}^{M} x_i \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{i=1}^{M} x_i. \tag{1.272}$$

## 1.41

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

$$\mathrm{I}(\mathbf{x}, \mathbf{y}) = -\int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}. \tag{1.273}$$

The right hand side can be written as

$$
\begin{aligned}
&-\int\int p(\mathbf{x},\mathbf{y})\left(\ln p(\mathbf{x}) + \ln\frac{p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y} \\
&= -\int\left(\int p(\mathbf{x},\mathbf{y})d\mathbf{y}\right)\ln p(\mathbf{x})d\mathbf{x} + \int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{x}|\mathbf{y})d\mathbf{x}d\mathbf{y}.
\end{aligned}
\tag{1.274}
$$

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

$$
\mathrm{I}(\mathbf{x},\mathbf{y}) = \mathrm{H}(\mathbf{x}) - \mathrm{H}(\mathbf{x}|\mathbf{y}). \tag{1.275}
$$

By the definition,

$$
\mathrm{I}(\mathbf{x},\mathbf{y}) = \mathrm{I}(\mathbf{y},\mathbf{x}). \tag{1.276}
$$

Thus,

$$
\mathrm{I}(\mathbf{x},\mathbf{y}) = \mathrm{H}(\mathbf{y}) - \mathrm{H}(\mathbf{y}|\mathbf{x}). \tag{1.277}
$$

# 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

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

we have

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

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

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

we have

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

By the definition,

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

Therefore,

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

## 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

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

Differentiating both sides of

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

with respect to $\mu$ gives

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

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,

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

Differentiating both sides of

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

twice with respect to $\mu$ gives

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

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

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

we have

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

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

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

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

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

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

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

## 2.6

Let $\mu$ be a variable such that

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

Then

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

Since

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

we have

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

Since

$$\mathrm{var}\mu = \mathrm{E}\mu^2 - (\mathrm{E}\mu)^2\,, \tag{2.50}$$

we have

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

Since

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

we have

$$\mathrm{mode}\mu = \frac{a-1}{a+b-2}. \tag{2.53}$$

## 2.7

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

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

where

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

By 2.6,

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

Note that

$$\mu_{\mathrm{ML}} = \frac{m}{m+l}. \tag{2.57}$$

Since

$$p(\mu|m, l, a, b) \propto p(m, l|\mu) p(\mu|a, b), \tag{2.58}$$

we have

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

Therefore, by 2.6,

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

Thus,

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

where

$$\lambda = \frac{m+l}{m+l+a+b}. \tag{2.62}$$

## 2.8 (Incomplete)

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

$$\mathrm{E}x = \int x p(x) dx. \tag{2.63}$$

The right hand side can be written as

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

Therefore,

$$\mathrm{E}x = \mathrm{E}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.65}$$

By the definition,

$$\mathrm{var}x = \mathrm{E}\left(x - \mathrm{E}x\right)^2. \tag{2.66}$$

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

$$\begin{aligned}
&\mathrm{E}_y\left(\mathrm{E}_x\left(\left(x - \mathrm{E}_x(x|y) + \mathrm{E}_x(x|y) - \mathrm{E}x\right)^2|y\right)\right)\\
&= \mathrm{E}_y\left(\mathrm{E}_x\left(\left(x - \mathrm{E}_x(x|y)\right)^2|y\right)\right)\\
&+ 2\mathrm{E}_y\left(\mathrm{E}_x\left(\left(x - \mathrm{E}_x(x|y)\right)\left(\mathrm{E}_x(x|y) - \mathrm{E}x\right)|y\right)\right)\\
&+ \mathrm{E}_y\left(\mathrm{E}_x\left(\left(\mathrm{E}_x(x|y) - \mathrm{E}x\right)^2|y\right)\right)
\end{aligned} \tag{2.67}$$

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

$$2\mathrm{E}_y\left(\left(\mathrm{E}_x(x|y) - \mathrm{E}x\right)\mathrm{E}_x\left(\left(x - \mathrm{E}_x(x|y)\right)|y\right)\right) \tag{2.68}$$

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

$$\mathrm{E}_y\left(\mathrm{E}_x(x|y) - \mathrm{E}_y\left(\mathrm{E}_x(x|y)\right)\right)^2 = \mathrm{var}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.69}$$

Therefore,

$$\mathrm{var}x = \mathrm{E}_y\left(\mathrm{var}_x(x|y)\right) + \mathrm{var}_y\left(\mathrm{E}_x(x|y)\right). \tag{2.70}$$

## 2.9 (Incomplete)

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

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

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

$$\int_{\substack{\sum_{k=1}^{M+1} \mu_k=1\\ \mu_k\geq 0}} \prod_{k=1}^{M+1} \mu_k^{\alpha_k-1}d\boldsymbol{\mu} = \int_0^1 \mu_{M+1}^{\alpha_{M+1}-1}\left(\int_{\substack{\sum_{k=1}^M \mu'_k=1-\mu_{M+1}\\ \mu'_k\geq 0}} \prod_{k=1}^M {\mu'_k}^{\alpha_k-1}d\boldsymbol{\mu}'\right)d\mu_{M+1}.$$

where $\boldsymbol{\mu}'$ is the vector of the first $M$ elements of $\boldsymbol{\mu}$. By the transformation

$$\boldsymbol{\mu}'' = \frac{1}{1-\mu_{M+1}}\boldsymbol{\mu}', \tag{2.71}$$

the right hand side can be written as

$$\int_0^1 \mu_{M+1}^{\alpha_{M+1}-1}\left(\int_{\substack{\sum_{k=1}^M \mu_k''=1\\ \mu_k''\geq 0}}\left(\prod_{k=1}^{M}((1-\mu_{M+1})\mu_k'')^{\alpha_k-1}\right)(1-\mu_{M+1})^M d\boldsymbol{\mu}''\right) d\mu_{M+1},$$

so that

$$\int_0^1 \mu_{M+1}^{\alpha_{M+1}-1}(1-\mu_{M+1})^{\Sigma_{k=1}^M \alpha_k}\left(\int_{\substack{\sum_{k=1}^M \mu_k''=1\\ \mu_k''\geq 0}}\prod_{k=1}^{M}\mu_k''^{\alpha_k-1}d\boldsymbol{\mu}''\right) d\mu_{M+1}.$$

By the assumption, it can be written as

$$\frac{\prod_{k=1}^M \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^M \alpha_k\right)}\frac{\Gamma(\alpha_{M+1})\Gamma\left(\sum_{k=1}^M \alpha_k+1\right)}{\Gamma\left(\sum_{k=1}^{M+1}\alpha_k+1\right)} = \frac{\sum_{k=1}^M \alpha_k}{\sum_{k=1}^{M+1}\alpha_k}\frac{\prod_{k=1}^{M+1}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{M+1}\alpha_k\right)}. \tag{2.72}$$

Therefore,

$$\int\prod_{k=1}^{M+1}\mu_k^{\alpha_k-1}d\boldsymbol{\mu} = \frac{\prod_{k=1}^{M+1}\Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{M+1}\alpha_k\right)}? \tag{2.73}$$

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

## 2.10

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

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}. \tag{2.74}$$

Then

$$\begin{aligned}\mathrm{E}\mu_j &= \int \mu_j p(\boldsymbol{\mu}|\boldsymbol{\alpha})d\boldsymbol{\mu},\\ \mathrm{E}\mu_j^2 &= \int \mu_j^2 p(\boldsymbol{\mu}|\boldsymbol{\alpha})d\boldsymbol{\mu},\\ \mathrm{E}\mu_j\mu_l &= \int \mu_j\mu_l p(\boldsymbol{\mu}|\boldsymbol{\alpha})d\boldsymbol{\mu}.\end{aligned} \tag{2.75}$$

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

$$
\begin{aligned}
\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + 1\right)} &= \frac{\alpha_j}{\sum_{k=1}^K \alpha_k}, \\
\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+2)}{\Gamma(\alpha_j)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + 2\right)} &= \frac{\alpha_j(\alpha_j+1)}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1\right)}, \\
\frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\frac{\Gamma(\alpha_j+1)\Gamma(\alpha_l+1)}{\Gamma(\alpha_j)\Gamma(\alpha_l)} \prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k + 2\right)} &= \frac{\alpha_j \alpha_l}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1\right)}.
\end{aligned} \tag{2.76}
$$

Therefore,

$$
\begin{aligned}
\mathrm{E}\mu_j &= \frac{\alpha_j}{\sum_{k=1}^K \alpha_k}. \\
\mathrm{E}\mu_j^2 &= \frac{\alpha_j(\alpha_j+1)}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1\right)}, \\
\mathrm{E}\mu_j \mu_l &= \frac{\alpha_j \alpha_l}{\sum_{k=1}^K \alpha_k \left(\sum_{k=1}^K \alpha_k + 1\right)}.
\end{aligned} \tag{2.77}
$$

Since

$$
\begin{aligned}
\mathrm{var}\mu_j &= \mathrm{E}\mu_j^2 - \left(\mathrm{E}\mu_j\right)^2, \\
\mathrm{cov}\left(\mu_j, \mu_l\right) &= \mathrm{E}\mu_j\mu_l - \mathrm{E}\mu_j \mathrm{E}\mu_l,
\end{aligned} \tag{2.78}
$$

we have

$$
\begin{aligned}
\mathrm{var}\mu_j &= \frac{\alpha_j \left(\sum_{k=1}^K \alpha_k - \alpha_j\right)}{\left(\sum_{k=1}^K \alpha_k\right)^2 \left(\sum_{k=1}^K \alpha_k + 1\right)}, \\
\mathrm{cov}\left(\mu_j, \mu_l\right) &= -\frac{\alpha_j \alpha_l}{\left(\sum_{k=1}^K \alpha_k\right)^2 \left(\sum_{k=1}^K \alpha_k + 1\right)}.
\end{aligned} \tag{2.79}
$$

## 2.11

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

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

Then

$$\mathrm{E} \ln \mu_j = \int (\ln \mu_j)\, p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}. \tag{2.81}$$

Since

$$\frac{\partial}{\partial \alpha_j} p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \left( \frac{\Gamma'\left(\sum_{k=1}^K \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} - \frac{\Gamma'(\alpha_j)}{\Gamma(\alpha_j)} + \ln \mu_j \right) p(\boldsymbol{\mu}|\boldsymbol{\alpha}), \tag{2.82}$$

we have

$$\mathrm{E} \ln \mu_j = \frac{\partial}{\partial \alpha_j} \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} + \left( \psi(\alpha_j) - \psi\left( \sum_{k=1}^K \alpha_k \right) \right) \int p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}, \tag{2.83}$$

where

$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \tag{2.84}$$

Therefore,

$$\mathrm{E} \ln \mu_j = \psi(\alpha_j) - \psi\left( \sum_{k=1}^K \alpha_k \right). \tag{2.85}$$

## 2.12

Let $x$ be a variable such that

$$p(x|a, b) = \frac{1}{b-a}, \tag{2.86}$$

where $a < b$. Then

$$\int_a^b p(x|a, b) dx = 1. \tag{2.87}$$

Note that

$$\begin{aligned} \mathrm{E}x &= \int_a^b x p(x|a, b) dx, \\ \mathrm{E}x^2 &= \int_a^b x^2 p(x|a, b) dx. \end{aligned} \tag{2.88}$$

The right hand sides can be written as

$$\begin{aligned} \frac{1}{b-a} \int_a^b x dx &= \frac{1}{2}(a+b), \\ \frac{1}{b-a} \int_a^b x^2 dx &= \frac{1}{3}\left(a^2 + ab + b^2\right). \end{aligned} \tag{2.89}$$

Therefore,

$$\begin{aligned} \mathrm{E}x &= \frac{1}{2}(a+b), \\ \mathrm{E}x^2 &= \frac{1}{3}\left(a^2+ab+b^2\right). \end{aligned} \tag{2.90}$$

Since

$$\mathrm{var}x = \mathrm{E}x^2 - (\mathrm{E}x)^2, \tag{2.91}$$

we have

$$\mathrm{var}x = \frac{1}{12}(b-a)^2. \tag{2.92}$$

## 2.13

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ q(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L}). \end{aligned} \tag{2.93}$$

Then, by the definition,

$$\mathrm{KL}(p||q) = -\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} d\mathbf{x}. \tag{2.94}$$

Since

$$\ln \frac{\mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{L})}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \ln \frac{(2\pi)^{-\frac{D}{2}} \left(|\det \mathbf{L}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m})\right)}{(2\pi)^{-\frac{D}{2}} \left(|\det \boldsymbol{\Sigma}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}, \tag{2.95}$$

The right hand side can be written as

$$\begin{aligned} &\frac{1}{2}\ln\left|\frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}}\right| \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &+ \frac{1}{2}\int (\mathbf{x}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \\ &- \frac{1}{2}\int (\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}. \end{aligned} \tag{2.96}$$

Let us look at each term. Since

$$\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} = 1, \tag{2.97}$$

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

$$(\mathbf{x}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\mathbf{m}) = (\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\mathbf{m}), \tag{2.98}$$

the second term can be written as

$$\begin{aligned}
&\frac{1}{2}\int (\mathbf{x}-\boldsymbol{\mu})^\intercal \mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} \\
&+ (\boldsymbol{\mu}-\mathbf{m})^\intercal \mathbf{L}^{-1}\int (\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} \\
&+ \frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m})\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x}.
\end{aligned} \tag{2.99}$$

Since

$$\begin{aligned}
\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} &= 1, \\
\int \mathbf{x}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} &= \boldsymbol{\mu}, \\
\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} &= \boldsymbol{\Sigma},
\end{aligned} \tag{2.100}$$

it can be written as

$$\frac{1}{2}\operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right) + \frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}). \tag{2.101}$$

Since

$$\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} = \boldsymbol{\Sigma}, \tag{2.102}$$

the third term can be written as

$$-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) = -\frac{D}{2} \tag{2.103}$$

Therefore,

$$\mathrm{KL}(p||q) = \frac{1}{2}\left(\ln\left|\frac{\det \mathbf{L}}{\det \boldsymbol{\Sigma}}\right| + \operatorname{tr}\left(\mathbf{L}^{-1}\boldsymbol{\Sigma}\right) + (\boldsymbol{\mu}-\mathbf{m})^\intercal \mathbf{L}^{-1}(\boldsymbol{\mu}-\mathbf{m}) - D\right). \tag{2.104}$$

## 2.14

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

$$\begin{aligned}L(p(\mathbf{x})) = &- \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} + \lambda \left( \int p(\mathbf{x}) d\mathbf{x} - 1 \right) \\ &+ \mathbf{l}^\intercal \left( \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\mu} \right) + \mathbf{m}^\intercal \left( \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal p(\mathbf{x}) d\mathbf{x} - \boldsymbol{\Sigma} \right) \mathbf{m}.\end{aligned} \tag{2.105}$$

Then

$$\frac{\delta L(p(\mathbf{x}))}{\delta p(\mathbf{x})} = - \ln p(\mathbf{x}) - 1 + \lambda + \mathbf{l}^\intercal \mathbf{x} + \mathbf{m}^\intercal (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{m}. \tag{2.106}$$

Setting the left hand side to zero gives

$$p(\mathbf{x}) = \exp\left(-1 + \lambda + \mathbf{l}^\intercal \mathbf{x} + \mathbf{m}^\intercal (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal \mathbf{m}\right), \tag{2.107}$$

so that

$$p(\mathbf{x}) = \exp\left(-1 + \lambda - \mathbf{l}^\intercal \mathbf{M} \mathbf{l} + (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})^\intercal \mathbf{M}^{-1} (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l})\right), \tag{2.108}$$

where

$$\mathbf{M} = \left(\mathbf{m}\mathbf{m}^\intercal\right)^{-1}. \tag{2.109}$$

Substituting it to

$$\begin{aligned}\int p(\mathbf{x}) d\mathbf{x} &= 1, \\ \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} &= \boldsymbol{\mu}, \\ \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\intercal p(\mathbf{x}) d\mathbf{x} &= \boldsymbol{\Sigma},\end{aligned} \tag{2.110}$$

and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{M}\mathbf{l} \tag{2.111}$$

gives

$$\begin{aligned}\exp\left(-1 + \lambda - \mathbf{l}^\intercal \mathbf{M} \mathbf{l}\right) \int \exp\left(-\mathbf{y}^\intercal \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} &= 1, \\ \exp\left(-1 + \lambda - \mathbf{l}^\intercal \mathbf{M} \mathbf{l}\right) \int (\mathbf{y} + \boldsymbol{\mu} + \mathbf{M}\mathbf{l}) \exp\left(-\mathbf{y}^\intercal \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} &= \boldsymbol{\mu}, \\ \exp\left(-1 + \lambda - \mathbf{l}^\intercal \mathbf{M} \mathbf{l}\right) \int (\mathbf{y} + \mathbf{M}\mathbf{l}) (\mathbf{y} + \mathbf{M}\mathbf{l})^\intercal \exp\left(-\mathbf{y}^\intercal \mathbf{M}^{-1} \mathbf{y}\right) d\mathbf{y} &= \boldsymbol{\Sigma}.\end{aligned} \tag{2.112}$$

Since

$$
\begin{aligned}
\int \exp\left(-\mathbf{y}^{\intercal}\mathbf{y}\right) d\mathbf{y} &= \left(\Gamma\left(\frac{1}{2}\right)\right)^{D}, \\
\int \mathbf{y} \exp\left(-\mathbf{y}^{\intercal}\mathbf{y}\right) d\mathbf{y} &= \mathbf{0}, \\
\int \mathbf{y}\mathbf{y}^{\intercal} \exp\left(-\mathbf{y}^{\intercal}\mathbf{y}\right) d\mathbf{y} &= \Gamma\left(\frac{3}{2}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1}\mathbf{I},
\end{aligned}
\tag{2.113}
$$

they can be written as

$$
\begin{aligned}
\exp\left(-1+\lambda-\mathbf{l}^{\intercal}\mathbf{M}\mathbf{l}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D}(\det\mathbf{M})^{\frac{1}{2}} &= 1, \\
\exp\left(-1+\lambda-\mathbf{l}^{\intercal}\mathbf{M}\mathbf{l}\right)\left(\boldsymbol{\mu}+\mathbf{M}\mathbf{l}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D}(\det\mathbf{M})^{\frac{1}{2}} &= \boldsymbol{\mu}, \\
\exp\left(-1+\lambda-\mathbf{l}^{\intercal}\mathbf{M}\mathbf{l}\right)\left(\Gamma\left(\frac{3}{2}\right)\left(\Gamma\left(\frac{1}{2}\right)\right)^{D-1}\mathbf{M}+\mathbf{M}\mathbf{l}(\mathbf{M}\mathbf{l})^{\intercal}\left(\Gamma\left(\frac{1}{2}\right)\right)^{D}\right)(\det\mathbf{M})^{\frac{1}{2}} &= \boldsymbol{\Sigma}.
\end{aligned}
\tag{2.114}
$$

Therefore,

$$
\begin{aligned}
\lambda &= 1 - \frac{D}{2}\ln\pi - \frac{1}{2}\ln(\det\mathbf{M}), \\
\mathbf{l} &= \mathbf{0}, \\
\mathbf{M} &= 2\boldsymbol{\Sigma}.
\end{aligned}
\tag{2.115}
$$

Thus,

$$
p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}}(\det\boldsymbol{\Sigma})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.116}
$$

## 2.15

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

$$
p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{2.117}
$$

Then, by the definition,

$$
\mathrm{H}(\mathbf{x}) = -\int \mathcal{N}\left(\mathbf{x}|\mu,\boldsymbol{\Sigma}\right)\ln\mathcal{N}\left(\mathbf{x}|\mu,\boldsymbol{\Sigma}\right) d\mathbf{x}. \tag{2.118}
$$

The right hand side can be written as

$$
\begin{aligned}
&-\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\det\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x} \\
&= \left(\frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\det\boldsymbol{\Sigma}|\right)\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{x} \\
&\quad + \frac{1}{2}\int (\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{x}.
\end{aligned}
\tag{2.119}
$$

Since

$$
\begin{aligned}
\int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{x} = 1, \\
\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{x} = \boldsymbol{\Sigma},
\end{aligned}
\tag{2.120}
$$

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

$$
\frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\det\boldsymbol{\Sigma}| \tag{2.121}
$$

and

$$
\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) = \frac{D}{2}. \tag{2.122}
$$

Therefore,

$$
\mathrm{H}(\mathbf{x}) = \frac{D}{2}\left(1 + \ln(2\pi)\right) + \frac{1}{2}\ln|\det\boldsymbol{\Sigma}|. \tag{2.123}
$$

## 2.16

Let $x$ be a variable such that

$$
x = x_1 + x_2, \tag{2.124}
$$

where

$$
\begin{aligned}
p(x_1) = \mathcal{N}\left(x_1|\mu_1, \tau_1^{-1}\right), \\
p(x_2) = \mathcal{N}\left(x_2|\mu_2, \tau_2^{-1}\right).
\end{aligned}
\tag{2.125}
$$

Then

$$
p(x) = \int_{-\infty}^{\infty} p(x|x_2)p(x_2)dx_2. \tag{2.126}
$$

The right hand side can be written as

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu_1 + x_2, \tau_1^{-1})\mathcal{N}\left(x_2|\mu_2, \tau_2^{-1}\right) dx_2$$
$$= \int_{-\infty}^{\infty} \left(\frac{\tau_1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_1}{2}(x - \mu_1 - x_2)^2\right) \left(\frac{\tau_2}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau_2}{2}(x_2 - \mu_2)^2\right) dx_2. \tag{2.127}$$

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

$$- \frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 + \frac{\tau_1 + \tau_2}{2}\left(\frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2$$
$$= -\frac{\tau_1 + \tau_2}{2}\left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2\mu_2}{\tau_1 + \tau_2}\right)^2 - \frac{\tau_1\tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2. \tag{2.128}$$

Therefore,

$$p(x) = \mathcal{N}\left(x \,\middle|\, \mu_1 + \mu_2, \tau_1^{-1} + \tau_2^{-1}\right). \tag{2.129}$$

Thus, by 1.35,

$$\mathrm{H}(x) = \frac{1}{2}\left(1 + \ln(2\pi) + \ln\left(\tau_1^{-1} + \tau_2^{-1}\right)\right). \tag{2.130}$$

## 2.17

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

$$\begin{aligned} \mathbf{S} &= \frac{1}{2}\left(\mathbf{\Sigma}^{-1} + \left(\mathbf{\Sigma}^{-1}\right)^\intercal\right), \\ \mathbf{A} &= \frac{1}{2}\left(\mathbf{\Sigma}^{-1} - \left(\mathbf{\Sigma}^{-1}\right)^\intercal\right). \end{aligned} \tag{2.131}$$

Then

$$\mathbf{\Sigma}^{-1} = \mathbf{S} + \mathbf{A}. \tag{2.132}$$

Therefore,

$$(\mathbf{x} - \boldsymbol{\mu})^\intercal\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^\intercal\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}) + (\mathbf{x} - \boldsymbol{\mu})^\intercal\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.133}$$

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

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\intercal\left(\mathbf{\Sigma}^{-1}\right)^\intercal(\mathbf{x} - \boldsymbol{\mu}). \tag{2.134}$$

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

$$-\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{\intercal}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.135}$$

Thus,

$$(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) = 0. \tag{2.136}$$

Hence

$$(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathbf{S}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.137}$$

## 2.18

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

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \tag{2.138}$$

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

$$\overline{\mathbf{u}_i}^{\intercal}\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i. \tag{2.139}$$

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

$$\overline{\mathbf{u}_i}^{\intercal}\overline{\boldsymbol{\Sigma}}^{\intercal}\mathbf{u}_i = \left(\overline{\boldsymbol{\Sigma}}\overline{\mathbf{u}}_i\right)^{\intercal}\mathbf{u}_i. \tag{2.140}$$

The right hand side can be writtet as

$$\overline{\lambda}_i\overline{\mathbf{u}}_i^{\intercal}\mathbf{u}_i = \overline{\lambda}_i. \tag{2.141}$$

Therefore,

$$\lambda_i = \overline{\lambda}_i. \tag{2.142}$$

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

$$\mathbf{u}_j^{\intercal}\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_j^{\intercal}\mathbf{u}_i. \tag{2.143}$$

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

$$\mathbf{u}_j^{\intercal}\boldsymbol{\Sigma}^{\intercal}\mathbf{u}_i = (\boldsymbol{\Sigma}\mathbf{u}_j)^{\intercal}\mathbf{u}_i. \tag{2.144}$$

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

$$\lambda_i\mathbf{u}_j^{\intercal}\mathbf{u}_i = \lambda_j\mathbf{u}_j^{\intercal}\mathbf{u}_i. \tag{2.145}$$

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

$$\mathbf{u}_j^{\intercal}\mathbf{u}_i = 0. \tag{2.146}$$

## 2.19

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

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \tag{2.147}$$

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

$$\begin{aligned} \mathbf{\Lambda} &= \operatorname{diag}(\lambda_1, \cdots, \lambda_D), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \tag{2.148}$$

Then

$$\mathbf{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}. \tag{2.149}$$

By 2.18,

$$\mathbf{U}^\intercal \mathbf{U} = \mathbf{I}. \tag{2.150}$$

Therefore,

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal, \\ \mathbf{\Sigma}^{-1} &= \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal, \end{aligned} \tag{2.151}$$

Thus,

$$\begin{aligned} \mathbf{\Sigma} &= \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^\intercal, \\ \mathbf{\Sigma}^{-1} &= \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\intercal. \end{aligned} \tag{2.152}$$

## 2.20

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

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \tag{2.153}$$

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

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

By 2.19,

$$\mathbf{a}^\intercal \mathbf{\Sigma} \mathbf{a} = \mathbf{b}^\intercal \mathbf{\Lambda} \mathbf{b}, \tag{2.155}$$

where

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

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

$$\mathbf{a}^{\intercal}\mathbf{\Sigma}\mathbf{a} > 0 \tag{2.157}$$

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

$$\lambda_i > 0. \tag{2.158}$$

## 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

$$\mathbf{\Sigma}\mathbf{\Lambda} = \mathbf{I}. \tag{2.159}$$

Taking the transpose of the both sides gives

$$\mathbf{\Lambda}^{\intercal}\mathbf{\Sigma} = \mathbf{I}. \tag{2.160}$$

Therefore,

$$\mathbf{\Lambda}^{\intercal} = \mathbf{\Lambda}. \tag{2.161}$$

## 2.23

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

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \tag{2.162}$$

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

$$\begin{aligned} \mathbf{\Lambda}' &= \operatorname{diag}\left(\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_D^{-\frac{1}{2}}\right), \\ \mathbf{U} &= [\mathbf{u}_1 \cdots \mathbf{u}_D]. \end{aligned} \tag{2.163}$$

By 2.19,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = \int_{(\mathbf{x}-\boldsymbol{\mu})^\intercal \mathbf{U}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}'^\intercal \mathbf{U}^\intercal (\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x}. \tag{2.164}$$

By the transformation

$$\mathbf{y} = \boldsymbol{\Lambda}'^\intercal \mathbf{U}^\intercal (\mathbf{x} - \boldsymbol{\mu}) \tag{2.165}$$

and the property

$$\mathbf{U}^\intercal \mathbf{U} = \mathbf{I}, \tag{2.166}$$

the right hand side can be written as

$$\int_{\|\mathbf{y}\|^2=\Delta} \left|\det\left(\mathbf{U}\boldsymbol{\Lambda}'^{-1}\right)\right| d\mathbf{y} = |\det \boldsymbol{\Sigma}|^{\frac{1}{2}} \int_{\|\mathbf{y}\|^2=\Delta} d\mathbf{y}. \tag{2.167}$$

Therefore,

$$\int_{(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\Delta} d\mathbf{x} = |\det \boldsymbol{\Sigma}|^{\frac{1}{2}} \Delta^D V_D, \tag{2.168}$$

where

$$V_D = \int_{\|\mathbf{x}\|=1} d\mathbf{x}. \tag{2.169}$$

## 2.24

Let

$$\begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \mathbf{W} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1}.$$

Then

$$\begin{aligned} \mathbf{XA} + \mathbf{YC} &= \mathbf{I}, \\ \mathbf{XB} + \mathbf{YD} &= \mathbf{O}, \\ \mathbf{ZA} + \mathbf{WC} &= \mathbf{O}, \\ \mathbf{ZB} + \mathbf{WD} &= \mathbf{I}. \end{aligned} \tag{2.170}$$

By the second and third equations,

$$\begin{aligned} \mathbf{Y} &= -\mathbf{XBD}^{-1}, \\ \mathbf{W} &= -\mathbf{ZAC}^{-1}. \end{aligned} \tag{2.171}$$

Substituting them to the first and fourth equation gives

$$\begin{aligned} \mathbf{X}\left(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}\right) &= \mathbf{I}, \\ \mathbf{Z}\left(\mathbf{B} - \mathbf{AC}^{-1}\mathbf{D}\right) &= \mathbf{I}. \end{aligned} \tag{2.172}$$

Therefore,

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

## 2.25

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.173}$$

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

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \tag{2.174}$$

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

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
=&-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^\intercal\boldsymbol{\Lambda}_{ac}(\mathbf{x}_c-\boldsymbol{\mu}_c) \\
&-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^\intercal\boldsymbol{\Lambda}_{bc}(\mathbf{x}_c-\boldsymbol{\mu}_c) \\
&-\frac{1}{2}(\mathbf{x}_c-\boldsymbol{\mu}_c)^\intercal\boldsymbol{\Lambda}_{ca}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_c-\boldsymbol{\mu}_c)^\intercal\boldsymbol{\Lambda}_{cb}(\mathbf{x}_b-\boldsymbol{\mu}_b)-\frac{1}{2}(\mathbf{x}_c-\boldsymbol{\mu}_c)^\intercal\boldsymbol{\Lambda}_{cc}(\mathbf{x}_c-\boldsymbol{\mu}_c).
\end{aligned} \tag{2.175}$$

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

$$-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_{a|b,c})^\intercal\boldsymbol{\Sigma}_{a|b,c}^{-1}(\mathbf{x}_a-\boldsymbol{\mu}_{a|b,c}), \tag{2.176}$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{a|b,c} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b-\boldsymbol{\mu}_b\right) - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\left(\mathbf{x}_c-\boldsymbol{\mu}_c\right), \\
\boldsymbol{\Sigma}_{a|b,c} &= \boldsymbol{\Lambda}_{aa}^{-1}.
\end{aligned} \tag{2.177}$$

Therefore,

$$p(\mathbf{x}_a|\mathbf{x}_b, \mathbf{x}_c) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b,c}, \boldsymbol{\Sigma}_{a|b,c}). \tag{2.178}$$

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

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

Thus,

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}), \tag{2.180}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) + \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ac}\boldsymbol{\mu}_c, \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1}. \end{aligned} \tag{2.181}$$

## 2.26 (Incomplete)

Since

$$\mathbf{I} = (\mathbf{A} + \mathbf{BCD})\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{BCDA}^{-1}\right) + \left(\mathbf{BCDA}^{-1}\right)^2, \tag{2.182}$$

we have

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}\right)\mathbf{DA}^{-1}. \tag{2.183}$$

## 2.27

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

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \int\int (\mathbf{x} + \mathbf{z}) p(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}. \tag{2.184}$$

The right hand side can be written as

$$\int \mathbf{x}\left(\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}\right) d\mathbf{x} + \int \mathbf{z}\left(\int p(\mathbf{x}, \mathbf{z}) d\mathbf{x}\right) d\mathbf{z} = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} + \int \mathbf{z} p(\mathbf{z}) d\mathbf{z}. \tag{2.185}$$

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

$$\mathrm{E}(\mathbf{x} + \mathbf{z}) = \mathrm{E}\mathbf{x} + \mathrm{E}\mathbf{z}. \tag{2.186}$$

Additionally,

$$\mathrm{cov}(\mathbf{x}+\mathbf{z}) = \int\int (\mathbf{x}+\mathbf{z}-\mathrm{E}(\mathbf{x}+\mathbf{z}))(\mathbf{x}+\mathbf{z}-\mathrm{E}(\mathbf{x}+\mathbf{z}))^\intercal p(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z}. \tag{2.187}$$

The right hand side can be written as

$$\begin{aligned}&\int\int (\mathbf{x}-\mathrm{E}\mathbf{x})(\mathbf{x}-\mathrm{E}\mathbf{x})^\intercal p(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z} + \int\int (\mathbf{x}-\mathrm{E}\mathbf{x})(\mathbf{z}-\mathrm{E}\mathbf{z})^\intercal p(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z}\\ &+ \int\int (\mathbf{z}-\mathrm{E}\mathbf{z})(\mathbf{x}-\mathrm{E}\mathbf{x})^\intercal p(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z} + \int\int (\mathbf{z}-\mathrm{E}\mathbf{z})(\mathbf{z}-\mathrm{E}\mathbf{z})^\intercal p(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z}.\end{aligned} \tag{2.188}$$

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

$$\begin{aligned}\int (\mathbf{x}-\mathrm{E}\mathbf{x})\, p(\mathbf{x})d\mathbf{x} \int (\mathbf{z}-\mathrm{E}\mathbf{z})^\intercal p(\mathbf{z})d\mathbf{z} = \mathbf{O},\\ \int (\mathbf{z}-\mathrm{E}\mathbf{z})\, p(\mathbf{z})d\mathbf{z} \int (\mathbf{x}-\mathrm{E}\mathbf{x})^\intercal p(\mathbf{x})d\mathbf{x} = \mathbf{O}.\end{aligned} \tag{2.189}$$

Therefore,

$$\mathrm{cov}(\mathbf{x}+\mathbf{z}) = \mathrm{cov}\mathbf{x} + \mathrm{cov}\mathbf{z}. \tag{2.190}$$

## 2.28 (Incomplete)

Let

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

where

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

and

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

Then

$$
\begin{aligned}
\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\intercal p(\mathbf{x})d\mathbf{x} &= \boldsymbol{\Lambda}^{-1}, \\
\int\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^\intercal p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} &= \boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal, \\
\int\int (\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})(\mathbf{x}-\boldsymbol{\mu})^\intercal p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} &= \mathbf{A}\boldsymbol{\Lambda}^{-1}, \\
\int (\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^\intercal p(\mathbf{y})d\mathbf{y} &= \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal.
\end{aligned}
\tag{2.191}
$$

## 2.29

Let

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

Then, by 2.24,

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

## 2.30

Let

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

Then

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

## 2.31

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

$$
\mathbf{y} = \mathbf{x}+\mathbf{z}, \tag{2.192}
$$

where

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_\mathbf{x},\boldsymbol{\Sigma}_\mathbf{x}\right), \\
p(\mathbf{z}) &= \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_\mathbf{z},\boldsymbol{\Sigma}_\mathbf{z}\right).
\end{aligned}
\tag{2.193}
$$

By the definition,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{2.194}$$

The right hand side can be written as

$$\int \mathcal{N}\left(\mathbf{y}|\mathbf{x}+\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}\right) \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}\right) d\mathbf{x}. \tag{2.195}$$

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

$$-\frac{1}{2}\left(\mathbf{y}-\mathbf{x}-\boldsymbol{\mu}_{\mathbf{z}}\right)^{\intercal} \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\left(\mathbf{y}-\mathbf{x}-\boldsymbol{\mu}_{\mathbf{z}}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^{\intercal}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}). \tag{2.196}$$

The first and second order terms can be written as

$$-\mathbf{x}^{\intercal}\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}} - \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\right) + \mathbf{y}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\mu}_{\mathbf{z}} = \mathbf{u}^{\intercal}\mathbf{v} \tag{2.197}$$

and

$$-\frac{1}{2}\mathbf{x}^{\intercal}\left(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\right)\mathbf{x} + \frac{1}{2}\mathbf{x}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{y}^{\intercal}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\mathbf{y} = -\frac{1}{2}\mathbf{u}^{\intercal}\mathbf{R}\mathbf{u}, \tag{2.198}$$

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

$$-\frac{1}{2}\left(\mathbf{u}-\mathbf{R}^{-1}\mathbf{v}\right)^{\intercal}\mathbf{R}\left(\mathbf{u}-\mathbf{R}^{-1}\mathbf{v}\right), \tag{2.199}$$

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,

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{x}}+\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{x}}+\boldsymbol{\Sigma}_{\mathbf{z}}). \tag{2.200}$$

## 2.32

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

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right), \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right). \end{aligned} \tag{2.201}$$

By the definition,

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

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

$$-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right)^\intercal \mathbf{L}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.203}$$

Since the first term can be written as

$$\begin{aligned} &-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}\right)^\intercal \mathbf{L}\left(\mathbf{y}-\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}\right) \\ =&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{A}^\intercal\mathbf{L}\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^\intercal\mathbf{A}^\intercal\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}) \\ &-\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^\intercal\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}), \end{aligned} \tag{2.204}$$

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

$$\begin{aligned} &-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})^\intercal\left(\boldsymbol{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z}) + \frac{1}{2}\mathbf{z}^\intercal\left(\mathbf{A}^\intercal\mathbf{L}\mathbf{A}+\boldsymbol{\Lambda}\right)\mathbf{z} \\ &-\frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^\intercal\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}) \\ =&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z})^\intercal\left(\boldsymbol{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)(\mathbf{x}-\boldsymbol{\mu}-\mathbf{z}) - \frac{1}{2}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b})^\intercal\mathbf{M}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}), \end{aligned} \tag{2.205}$$

where

$$\begin{aligned} \mathbf{z} &= \left(\boldsymbol{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\intercal\mathbf{L}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}-\mathbf{b}), \\ \mathbf{M} &= \mathbf{L}-\mathbf{L}\mathbf{A}\left(\boldsymbol{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)^{-1}\mathbf{A}^\intercal\mathbf{L}. \end{aligned} \tag{2.206}$$

we have

$$\boldsymbol{\mu}+\mathbf{z} = \left(\boldsymbol{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^\intercal\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right). \tag{2.207}$$

By 2.26,

$$\left(\boldsymbol{\Lambda}+\mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)^{-1} = \boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\left(\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\right)^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}. \tag{2.208}$$

Therefore,

$$\mathbf{M} = \left(\mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\intercal\right)^{-1}. \tag{2.209}$$

Thus,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}\left(\mathbf{x}|\left(\mathbf{\Lambda} + \mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)^{-1}\left(\mathbf{A}^\intercal\mathbf{L}(\mathbf{y}-\mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\right), \left(\mathbf{\Lambda} + \mathbf{A}^\intercal\mathbf{L}\mathbf{A}\right)^{-1}\right), \\ p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^\intercal\right). \end{aligned} \tag{2.210}$$

## 2.33

Refer to 2.32.

## 2.34

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

$$\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}\right) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(\det\mathbf{\Sigma}) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\intercal\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.211}$$

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

$$\begin{aligned} \mathbf{0} &= \sum_{n=1}^{N}\left(\mathbf{\Sigma}^{-1} + \left(\mathbf{\Sigma}^{-1}\right)^\intercal\right)(\mathbf{x}_n - \boldsymbol{\mu}), \\ \mathbf{O} &= -\frac{N}{2}\left(\mathbf{\Sigma}^{-1}\right)^\intercal + \frac{1}{2}\left(\mathbf{\Sigma}^{-1}\right)^2\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^\intercal. \end{aligned} \tag{2.212}$$

Therefore,

$$\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \\ \mathbf{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\intercal. \end{aligned} \tag{2.213}$$

## 2.35

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

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.214}$$

Then

$$\mathrm{E}\mathbf{x}\mathbf{x}^{\intercal} = \int \mathbf{x}\mathbf{x}^{\intercal}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x}. \tag{2.215}$$

The right hand side can be written as

$$\begin{aligned}
&\int (\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu}+\boldsymbol{\mu})^{\intercal}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} \\
&= \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} + \boldsymbol{\mu}\int (\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} \\
&\quad + \left(\int (\mathbf{x}-\boldsymbol{\mu})\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x}\right)\boldsymbol{\mu}^{\intercal} + \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x}.
\end{aligned} \tag{2.216}$$

Since

$$\begin{aligned}
\int \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} &= 1, \\
\int \mathbf{x}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} &= \boldsymbol{\mu}, \\
\int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) d\mathbf{x} &= \boldsymbol{\Sigma},
\end{aligned} \tag{2.217}$$

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

$$\mathrm{E}\mathbf{x}\mathbf{x}^{\intercal} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}. \tag{2.218}$$

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

$$\begin{aligned}
p(\mathbf{x}_n) &= \mathcal{N}\left(\mathbf{x}_{\mathbf{n}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \\
p(\mathbf{x}_m) &= \mathcal{N}\left(\mathbf{x}_{\mathbf{m}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).
\end{aligned} \tag{2.219}$$

If $n \neq m$, then

$$\mathrm{E}\mathbf{x}_n\mathbf{x}_m^{\intercal} = \mathrm{E}\mathbf{x}_n\mathrm{E}\mathbf{x}_m^{\intercal}. \tag{2.220}$$

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

$$\mathrm{E}\mathbf{x}_n\mathbf{x}_m^{\intercal} = \delta_{nm}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}. \tag{2.221}$$

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

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

By 2.34,

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\intercal.\end{aligned} \tag{2.223}$$

Then

$$\mathrm{E}\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^\intercal. \tag{2.224}$$

The right hand side can be written as

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N}\mathrm{E}\mathbf{x}_n\mathbf{x}_n^\intercal - \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\left(\sum_{n=1}^{N}\mathbf{x_n}\right)\mathbf{x}_n^\intercal - \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}\mathbf{x}_n\left(\sum_{n=1}^{N}\mathbf{x_n}\right)^\intercal \\ &+ \frac{1}{N^3}\sum_{n=1}^{N}\mathrm{E}\left(\sum_{n=1}^{N}\mathbf{x_n}\right)\left(\sum_{n=1}^{N}\mathbf{x_n}\right)^\intercal.\end{aligned} \tag{2.225}$$

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

$$-\frac{1}{N}\left((\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal) + (N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right) = -\frac{1}{N}\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal. \tag{2.226}$$

The fourth term can be written as

$$\frac{1}{N^2}\left(N\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right) + N(N-1)\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\right) = \frac{1}{N}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal. \tag{2.227}$$

Therefore,

$$\mathrm{E}\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{N-1}{N}\boldsymbol{\Sigma}. \tag{2.228}$$

## 2.36

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

$$p(x_n) = \mathcal{N}\left(x_n|\mu, \sigma^2\right). \tag{2.229}$$

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

$$\sigma_{\mathrm{ML}}^{2}{}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2. \tag{2.230}$$

The right hand side can be written as

$$\frac{1}{N}(x_N-\mu)^2 + \frac{1}{N}\sum_{n=1}^{N-1}(x_n-\mu)^2 = \frac{1}{N}(x_N-\mu)^2 + \frac{N-1}{N}\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}. \tag{2.231}$$

Therefore,

$$\sigma_{\mathrm{ML}}^{2}{}^{(N)} = \sigma_{\mathrm{ML}}^{2}{}^{(N-1)} + \frac{1}{N}\left((x_N-\mu)^2 - \sigma_{\mathrm{ML}}^{2}{}^{(N-1)}\right). \tag{2.232}$$

Since

$$\frac{\partial}{\partial\sigma^2}\left(-\ln p\left(x_n|\sigma^2\right)\right) = \frac{1}{2\sigma^2} - \frac{1}{2\left(\sigma^2\right)^2}(x_n-\mu)^2, \tag{2.233}$$

we have

$$\sigma_{\mathrm{ML}}^{2}{}^{(N)} = \sigma_{\mathrm{ML}}^{2}{}^{(N-1)} - \frac{\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}}{N}\frac{\partial}{\partial\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}}\left(-\ln p\left(x_N|\sigma_{\mathrm{ML}}^{2}{}^{(N-1)}\right)\right). \tag{2.234}$$

## 2.37

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

$$p(\mathbf{x}_n) = \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right). \tag{2.235}$$

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

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\intercal}. \tag{2.236}$$

The right hand side can be written as

$$\begin{aligned}&\frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\intercal} + \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\intercal}\\ &= \frac{1}{N}(\mathbf{x}_N-\boldsymbol{\mu})(\mathbf{x}_N-\boldsymbol{\mu})^{\intercal} + \frac{N-1}{N}\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}.\end{aligned} \tag{2.237}$$

Therefore,

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}\left((\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\intercal} - \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right). \tag{2.238}$$

Since

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\left(-\ln p(x_n|\boldsymbol{\Sigma})\right) = -\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{\intercal} + \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\right)^{2}(\mathbf{x}_N - \boldsymbol{\mu})(\mathbf{x}_N - \boldsymbol{\mu})^{\intercal}, \tag{2.239}$$

we have

$$\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N)} = \boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)} - \frac{\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}}{N}\frac{\partial}{\partial \boldsymbol{\sigma}_{\mathrm{ML}}^{(N-1)}}\left(-\ln p\left(\mathbf{x}_N|\boldsymbol{\Sigma}_{\mathrm{ML}}^{(N-1)}\right)\right). \tag{2.240}$$