

# Solutions Manual to Pattern Recognition and Machine Learning

Hiromichi Inawashiro

June 28, 2024

# 1 Introduction

## 1.1

Let

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

To minimise it, setting the derivative to zero gives

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

Substituting

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

gives

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

Therefore,

$$\sum_{j=0}^M A_{ij} w_j = T_i \quad (1.5)$$

where

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

## 1.2

Let

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

To minimise it, setting the derivative to zero gives

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

Substituting

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

gives

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

Therefore,

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

where

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

### 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

$$p(a) = p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g). \quad (1.13)$$

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

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

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

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

Here,

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

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,

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

## 1.4

Let

$$x = g(y) \quad (1.18)$$

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

$$p_y(y) = p_x(g(y)) |g'(y)| \quad (1.19)$$

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

$$0 = g'(\hat{y})p'_x(g(\hat{y})) + p_x(g(\hat{y}))g''(\hat{y}). \quad (1.20)$$

Therefore, in general,

$$\hat{x} \neq g(\hat{y}). \quad (1.21)$$

Here, let us assume that

$$g(y) = ay + b. \quad (1.22)$$

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

$$0 = p'_x(g(\hat{y})). \quad (1.23)$$

Therefore,

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

## 1.5

By the definition,

$$\text{var} f(x) = E(f(x) - Ef(x))^2. \quad (1.25)$$

The right hand side can be written as

$$E((f(x))^2 - 2f(x)Ef(x) + (Ef(x))^2) = E(f(x))^2 - (Ef(x))^2. \quad (1.26)$$

Therefore,

$$\text{var} f(x) = E(f(x))^2 - (Ef(x))^2. \quad (1.27)$$

## 1.6

By the definition,

$$\text{cov}(x, y) = E((x - Ex)(y - Ey)). \quad (1.28)$$

The right hand side can be written as

$$Exy - E(xEy) - E(yEx) + E(ExEy) = Exy - ExEy. \quad (1.29)$$

The right hand side can be written as

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

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

$$f(x, y) = f(x)f(y). \quad (1.31)$$

Then,

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

Therefore,

$$\text{cov}(x, y) = 0. \quad (1.33)$$

## 1.7

Let

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

Then

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

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

$$\int_0^{\infty} \int_0^{2\pi} \exp\left(-\frac{1}{2\sigma^2}r^2\right) \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} dr d\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr. \quad (1.36)$$

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

$$2\pi\sigma^2 \int_0^{\infty} \exp\left(-\frac{1}{2}s^2\right) s ds = 2\pi\sigma^2 \left[-\exp\left(-\frac{1}{2}s^2\right)\right]_0^{\infty}. \quad (1.37)$$

Therefore,

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

By the definition,

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

Then

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

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

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

Therefore,

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

## 1.8

Let  $x$  be a variable under the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . Then

$$\mathbb{E}x = \int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.43)$$

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

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx. \quad (1.44)$$

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

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} (y + \mu) \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy. \quad (1.45)$$

Since

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} y \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = 0, \quad (1.46)$$

and

$$(2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mu \exp\left(-\frac{1}{2\sigma^2}y^2\right) dy = \mu \int_{-\infty}^{\infty} \mathcal{N}(y|\mu, \sigma^2) dy, \quad (1.47)$$

we have

$$\mathbb{E}x = \mu. \quad (1.48)$$

By the definition,

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1 \quad (1.49)$$

can be written as

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

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

$$\begin{aligned} & (2\pi)^{-\frac{1}{2}} \left(-\frac{1}{2}\right) (\sigma^2)^{-\frac{3}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ & + (2\pi\sigma^2)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{2} (\sigma^2)^{-2} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = 0. \end{aligned} \quad (1.51)$$

The left hand side can be written as

$$\begin{aligned} -\frac{1}{2}(\sigma^2)^{-1} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2}(\sigma^2)^{-2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx \\ = -\frac{1}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2} \text{var}x. \end{aligned} \quad (1.52)$$

Therefore,

$$\text{var}x = \sigma^2. \quad (1.53)$$

## 1.9

Let

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.54)$$

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

$$0 = (2\pi\sigma^2)^{-\frac{1}{2}} \left(-\frac{1}{\sigma^2}(x - \mu)\right) \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.55)$$

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

Similarly, let

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1.56)$$

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

$$\mathbf{0} = -(2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} (\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^\top) (\mathbf{x} - \boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (1.57)$$

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

## 1.10

By the definition,

$$\mathbb{E}(x + y) = \int \int (x + y)p(x, y)dx dy. \quad (1.58)$$



The right hand side can be written as

$$\int x \left( \int p(x, y) dy \right) dx + \int y \left( \int p(x, y) dx \right) dy = \int xp(x)dx + \int yp(y)dy. \quad (1.59)$$

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

$$Ex + Ey. \quad (1.60)$$

Therefore,

$$E(x + y) = Ex + Ey. \quad (1.61)$$

Similarly, by the definition,

$$\text{var}(x + y) = E(x + y - E(x + y))^2 \quad (1.62)$$

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

$$\begin{aligned} E(x - Ex)^2 + 2E((x - Ex)(y - Ey)) + E(y - Ey)^2 \\ = \text{var}x + 2\text{cov}(x, y) + \text{var}y. \end{aligned} \quad (1.63)$$

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

$$\text{cov}(x, y) = 0, \quad (1.64)$$

by 1.6. Therefore,

$$\text{var}(x + y) = \text{var}x + \text{var}y. \quad (1.65)$$

## 1.11

Let

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2. \quad (1.66)$$

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

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu), \\ 0 &= -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{n=1}^N (x_n - \mu)^2. \end{aligned} \quad (1.67)$$

Therefore,

$$\begin{aligned}\mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^N x_n, \\ \sigma_{\text{ML}}^2 &= \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2.\end{aligned}\tag{1.68}$$

## 1.12

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

$$\mathbb{E}x_mx_n = \mathbb{E}x_m\mathbb{E}x_n.\tag{1.69}$$

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,

$$\mathbb{E}x_n^2 = \text{var}x_n + (\mathbb{E}x_n)^2.\tag{1.70}$$

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,

$$\mathbb{E}x_mx_n = \mu^2 + \delta_{mn}\sigma^2.\tag{1.71}$$

Here, since

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n,\tag{1.72}$$

we have

$$\mathbb{E}\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbb{E}x_n.\tag{1.73}$$

Therefore,

$$\mathbb{E}\mu_{\text{ML}} = \mu.\tag{1.74}$$

Similarly, since

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2,\tag{1.75}$$

we have

$$\mathbb{E}\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \mathbb{E}(x_n - \mu_{\text{ML}})^2.\tag{1.76}$$

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n^2 - 2\mu_{\text{ML}}x_n + \mu_{\text{ML}}^2) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n^2 - \frac{2}{N} \mathbb{E} \left( \mu_{\text{ML}} \left( \sum_{n=1}^N x_n \right) \right) + \mathbb{E} \mu_{\text{ML}}^2. \quad (1.77)$$

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

$$\frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) = \mu^2 + \sigma^2, \quad (1.78)$$

while the second and third terms can be written as

$$-2\mathbb{E} \mu_{\text{ML}}^2 + \mathbb{E} \mu_{\text{ML}}^2 = -\mathbb{E} \mu_{\text{ML}}^2. \quad (1.79)$$

Here,

$$\mathbb{E} \mu_{\text{ML}}^2 = \mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N x_n \right)^2. \quad (1.80)$$

The right hand side can be written as

$$\frac{1}{N^2} \sum_{n=1}^N \mathbb{E} x_n^2 + \frac{2}{N^2} \sum_{1 \leq m < n \leq N} \mathbb{E} x_m x_n = \frac{1}{N} (\mu^2 + \sigma^2) + \frac{N-1}{N} \mu^2. \quad (1.81)$$

Therefore,

$$\mathbb{E} \mu_{\text{ML}}^2 = \mu^2 + \frac{1}{N} \sigma^2. \quad (1.82)$$

Thus,

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{N-1}{N} \sigma^2. \quad (1.83)$$

### 1.13

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

$$\mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n - \mu)^2. \quad (1.84)$$

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n^2 - 2\mu x_n + \mu^2) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n^2 - \frac{2\mu}{N} \sum_{n=1}^N \mathbb{E} x_n + \mu^2. \quad (1.85)$$

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

$$\frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2) = \mu^2 + \sigma^2, \quad (1.86)$$

while the second term can be written as

$$-\frac{2\mu}{N} \sum_{n=1}^N \mu = -2\mu^2. \quad (1.87)$$

Therefore,

$$\mathbb{E} \left( \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right) = \sigma^2. \quad (1.88)$$

## 1.14

Let

$$\begin{aligned} w_{ij}^S &= \frac{1}{2}(w_{ij} + w_{ji}), \\ w_{ij}^A &= \frac{1}{2}(w_{ij} - w_{ji}). \end{aligned} \quad (1.89)$$

Then

$$\begin{aligned} w_{ij} &= w_{ij}^S + w_{ij}^A, \\ w_{ij}^S &= w_{ji}^S, \\ w_{ij}^A &= -w_{ji}^A. \end{aligned} \quad (1.90)$$

Here,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = \frac{1}{2} \sum_{i=1}^D \sum_{j=1}^D (w_{ij} - w_{ji}) x_i x_j. \quad (1.91)$$

The right hand side can be written as

$$\frac{1}{2} \left( \sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j - \sum_{i=1}^D \sum_{j=1}^D w_{ji} x_i x_j \right) = 0. \quad (1.92)$$

Therefore,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = 0. \quad (1.93)$$

Additionally,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = \sum_{i=1}^D \sum_{j=1}^D (w_{ij}^S + w_{ij}^A) x_i x_j. \quad (1.94)$$

The right hand side can be written as

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j + \sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = \sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j, \quad (1.95)$$

where the result above is used. Therefore,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = \sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j. \quad (1.96)$$

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

### 1.15 (Incomplete)

### 1.16 (Incomplete)

### 1.17

Let

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \quad (1.97)$$

Then

$$\Gamma(x+1) = \int_0^\infty u^x \exp(-u) du. \quad (1.98)$$

The right hand side can be written as

$$[-u^x \exp(-u)]_{u=0}^{u=\infty} + \int_0^\infty x u^{x-1} \exp(-u) du = x \Gamma(x). \quad (1.99)$$

Therefore,

$$\Gamma(x+1) = x \Gamma(x). \quad (1.100)$$

Since

$$\Gamma(1) = \int_0^1 \exp(-u) du, \quad (1.101)$$

and the right hand side can be written as 1,

$$\Gamma(1) = 0!. \quad (1.102)$$

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

$$\Gamma(x) = (x-1)!. \quad (1.103)$$

Then,

$$\Gamma(x+1) = x\Gamma(x), \quad (1.104)$$

where the right hand side can be written as

$$x(x-1)! = x!. \quad (1.105)$$

Therefore,

$$\Gamma(x+1) = x!. \quad (1.106)$$

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

## 1.18

Let us consider the transformation from Cartesian to polar coordinates

$$\prod_{i=1}^D \int_{-\infty}^{\infty} \exp(-x_i^2) dx_i = S_D \int_0^{\infty} \exp(-r^2) r^{D-1} dr, \quad (1.107)$$

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

$$\frac{S_D}{2} \int_0^{\infty} \exp(-s) s^{\frac{D-1}{2}} s^{-\frac{1}{2}} ds = \frac{S_D}{2} \Gamma\left(\frac{D}{2}\right). \quad (1.108)$$

Therefore,

$$S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}\right)}. \quad (1.109)$$

Additionally, the volume of the sphere can be written as

$$V_D = S_D \int_0^1 r^{D-1} dr. \quad (1.110)$$

The right hand side can be written as

$$S_D \left[ \frac{r^D}{D} \right]_{r=0}^{r=1} = \frac{S_D}{D}. \quad (1.111)$$

Therefore,

$$V_D = \frac{S_D}{D}. \quad (1.112)$$

Finally, the results above reduce to

$$\begin{aligned} S_2 &= \frac{2\pi}{\Gamma(1)}, \\ V_2 &= \frac{S_2}{2}. \end{aligned} \quad (1.113)$$

Therefore,

$$\begin{aligned} S_2 &= 2\pi, \\ V_2 &= \pi. \end{aligned} \quad (1.114)$$

Similarly,

$$\begin{aligned} S_3 &= \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})}, \\ V_3 &= \frac{S_3}{3}. \end{aligned} \quad (1.115)$$

Therefore,

$$\begin{aligned} S_3 &= 4\pi, \\ V_3 &= \frac{4}{3}\pi. \end{aligned} \quad (1.116)$$

## 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

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} \Gamma(\frac{D}{2})}, \quad (1.117)$$

by 1.18.

Additionally, by Sterling's formula

$$\Gamma(x+1) \simeq (2\pi)^{\frac{1}{2}} \exp(-x) x^{\frac{x+1}{2}}, \quad (1.118)$$

the ratio can be approximated as

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D 2^{D-1} (2\pi)^{\frac{1}{2}} \exp\left(1 - \frac{D}{2}\right) \left(\frac{D}{2} - 1\right)^{\frac{D}{4}}}. \quad (1.119)$$

The right hand side can be written as

$$\frac{1}{2e(2\pi)^{\frac{1}{2}}} \frac{1}{D} \left( \frac{e^2 \pi^2}{8D - 16} \right)^{\frac{D}{4}}. \quad (1.120)$$

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

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

$$\frac{\sqrt{\sum_{i=1}^D 1^2}}{1} = \sqrt{D}. \quad (1.121)$$

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

## 1.20

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

$$p(\mathbf{x}) = (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \quad (1.122)$$

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

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} = \int_r^{r+\epsilon} \int (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r'^2}{2\sigma^2}\right) J dr' d\phi, \quad (1.123)$$

where  $\phi$  is the vector of the angular components of the polar coordinate and  $J$  is the Jacobian of the transformation from the Cartesian to polar coordinate.



For a sufficiently small  $\epsilon$ , the right hand side can be approximated as

$$\begin{aligned} & (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_r^{r+\epsilon} \int J dr' d\phi \\ &= (2\pi\sigma^2)^{-\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} d\mathbf{x}. \end{aligned} \quad (1.124)$$

Therefore,

$$\int_{r \leq \|\mathbf{x}\| \leq r+\epsilon} p(\mathbf{x}) d\mathbf{x} \simeq p(r)\epsilon, \quad (1.125)$$

where

$$p(r) = (2\pi\sigma^2)^{-\frac{D}{2}} S_D r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad (1.126)$$

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

$$0 = (2\pi\sigma^2)^{-\frac{D}{2}} S_D \left( (D-1)r^{D-2} - \frac{r^D}{\sigma^2} \right) \exp\left(-\frac{r^2}{2\sigma^2}\right). \quad (1.127)$$

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

$$\hat{r} = \sqrt{D-1}\sigma. \quad (1.128)$$

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

$$\frac{p(\hat{r} + \epsilon)}{p(\hat{r})} = \left( \frac{\hat{r} + \epsilon}{\hat{r}} \right)^{D-1} \exp\left(-\frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \quad (1.129)$$

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

$$\begin{aligned} & \exp\left((D-1)\ln\left(1 + \frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\hat{r}^2}{\sigma^2}\ln\left(1 + \frac{\epsilon}{\hat{r}}\right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2}\right). \end{aligned} \quad (1.130)$$

By the Taylor series

$$\ln(1+x) = x - \frac{1}{2}x^2 + o(x^3), \quad (1.131)$$

the right hand side can be approximated as

$$\exp \left( \frac{\hat{r}^2}{\sigma^2} \left( \frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2} \right) - \frac{2\hat{r}\epsilon + \epsilon^2}{2\sigma^2} \right) = \exp \left( -\frac{\epsilon^2}{\sigma^2} \right). \quad (1.132)$$

Therefore,

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

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

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp \left( \frac{\hat{r}^2}{2\sigma^2} \right). \quad (1.134)$$

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

$$\frac{p(\mathbf{0})}{p(\hat{\mathbf{r}})} = \exp \left( \frac{D-1}{2} \right). \quad (1.135)$$

## 1.21

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

$$0 \leq a(b-a). \quad (1.136)$$

Therefore,

$$a \leq (ab)^{\frac{1}{2}}. \quad (1.137)$$

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,

$$p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2) \Rightarrow \mathbf{x} \in \mathcal{C}_1, \quad (1.138)$$

and

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

Then, using the inequality above,

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} \leq \int_{\mathcal{R}_1} (p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2))^{\frac{1}{2}} d\mathbf{x}, \quad (1.140)$$

and

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

Therefore,

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

## 1.22

Let

$$EL = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.143)$$

If

$$L_{kj} = 1 - \delta_{kj}, \quad (1.144)$$

then the right hand side can be written as

$$\sum_k \sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k p(\mathbf{x}, \mathcal{C}_k) - p(\mathbf{x}, \mathcal{C}_j) \right) d\mathbf{x}. \quad (1.145)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} (p(\mathbf{x}) - p(\mathbf{x}, \mathcal{C}_j)) d\mathbf{x} = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathbf{x}, \mathcal{C}_j) d\mathbf{x}. \quad (1.146)$$

Therefore,

$$EL = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.147)$$

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

## 1.23

Let

$$EL = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}. \quad (1.148)$$

The right hand side can be written as

$$\sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_j \int_{\mathcal{R}_j} \left( \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.149)$$

Therefore,

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

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

## 1.24 (Incomplete)

### 1.25

Let

$$EL(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.151)$$

Then

$$\frac{\delta EL(\mathbf{t}, \mathbf{y}(\mathbf{x}))}{\delta \mathbf{y}(\mathbf{x})} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \quad (1.152)$$

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

$$\mathbf{0} = \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t}. \quad (1.153)$$

The right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \quad (1.154)$$

Thus,

$$\mathbf{y}(\mathbf{x}) = E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \quad (1.155)$$

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

$$y(\mathbf{x}) = E_t(t|\mathbf{x}). \quad (1.156)$$

### 1.26

Let

$$EL(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.157)$$

The right hand side can be written as

$$\begin{aligned} & \int \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &= \int \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &+ 2 \int \int (\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top (E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &+ \int \int \|E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \end{aligned} \quad (1.158)$$

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

$$\int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 \left( \int p(\mathbf{x}, \mathbf{t}) d\mathbf{t} \right) d\mathbf{x} = \int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x}. \quad (1.159)$$

The second term can be written as

$$2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top \left( \int (\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.160)$$

Since

$$\begin{aligned} \int \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) \frac{\int p(\mathbf{x}, \mathbf{t}) d\mathbf{t}}{p(\mathbf{x})}, \\ \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} &= \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}), \end{aligned} \quad (1.161)$$

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

$$\int \left( \int \|\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x} = \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.162)$$

Therefore,

$$EL(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.163)$$

Thus,  $EL(\mathbf{t}, \mathbf{y}(\mathbf{x}))$  is minimised if

$$\mathbf{y}(\mathbf{x}) = \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \quad (1.164)$$

## 1.27

Let

$$EL_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \quad (1.165)$$

Then

$$\frac{\delta EL_q}{\delta y(\mathbf{x})} = \int q |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t) p(\mathbf{x}, t) dt. \quad (1.166)$$

To minimise  $EL_q$ , setting the left hand side to zero gives

$$0 = \int |y(\mathbf{x}) - t|^{q-1} \text{sign}(y(\mathbf{x}) - t) p(t|\mathbf{x}) dt. \quad (1.167)$$

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

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

$$0 = \int_{y(\mathbf{x})}^{\infty} p(t|\mathbf{x})dt - \int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})dt. \quad (1.168)$$

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

## 1.28

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \quad (1.169)$$

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

$$h(p^2) = h(p) + h(p). \quad (1.170)$$

Therefore,

$$h(p^2) = 2h(p). \quad (1.171)$$

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

$$h(p^n) = nh(p). \quad (1.172)$$

Then, by the first assumption,

$$h(p^{n+1}) = h(p^n) + h(p). \quad (1.173)$$

Therefore,

$$h(p^{n+1}) = (n+1)h(p). \quad (1.174)$$

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

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

$$h(p^n) = h(p^{\frac{n}{m}m}). \quad (1.175)$$

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

$$h(p^{\frac{n}{m}}) = \frac{n}{m}h(p). \quad (1.176)$$

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

$$h(p^a) = ah(p). \quad (1.177)$$

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

$$(p \ln p)h'(p) = h(p). \quad (1.178)$$

Therefore,

$$\int \frac{h'(p)}{h(p)} dp = \int \frac{1}{p \ln p} dp + C, \quad (1.179)$$

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

$$h(p) \propto \ln p. \quad (1.180)$$

## 1.29

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

$$H(x) = - \sum_{i=1}^M p(x_i) \ln p(x_i), \quad (1.181)$$

where

$$\sum_{i=1}^M p(x_i) = 1. \quad (1.182)$$

By Jensen's inequality,

$$\sum_{i=1}^M p(x_i) \ln \frac{1}{p(x_i)} \leq \ln \left( \sum_{i=1}^M 1 \right). \quad (1.183)$$

Therefore,

$$H(x) \leq \ln M. \quad (1.184)$$

### 1.30

Let

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu, \sigma^2), \\ q(x) &= \mathcal{N}(x|m, s^2). \end{aligned} \quad (1.185)$$

Then, by the definition,

$$\text{KL}(p||q) = - \int p(x) \ln \frac{q(x)}{p(x)} dx. \quad (1.186)$$

The right hand side can be written as

$$\begin{aligned} & - \int_{-\infty}^{\infty} p(x) \ln \frac{(2\pi s^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)}{(2\pi \sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} dx \\ &= - \int_{-\infty}^{\infty} p(x) \left( -\frac{1}{2} \ln \frac{s^2}{\sigma^2} - \frac{(x-m)^2}{2s^2} + \frac{(x-\mu)^2}{2\sigma^2} \right) dx. \end{aligned} \quad (1.187)$$

The right hand side can be written as

$$\ln \frac{s}{\sigma} \int_{-\infty}^{\infty} p(x) dx + \frac{1}{2s^2} \int_{-\infty}^{\infty} (x-m)^2 p(x) dx - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx. \quad (1.188)$$

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

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

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

$$\text{KL}(p||q) = \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{2s^2} - \frac{1}{2}. \quad (1.190)$$

### 1.31

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

$$\begin{aligned} H(\mathbf{x}) &= - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}, \\ H(\mathbf{y}) &= - \int p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y}, \\ H(\mathbf{x}, \mathbf{y}) &= - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \end{aligned} \quad (1.191)$$



Note that

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

Therefore,

$$H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) = - \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}. \quad (1.193)$$

Since

$$\int \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = 1, \quad (1.194)$$

Jensen's inequality can be used to write that

$$- \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y} \geq - \ln \left( \int \int p(\mathbf{x})p(\mathbf{y}) d\mathbf{x} d\mathbf{y} \right). \quad (1.195)$$

The right hand side can be written as

$$- \ln \left( \int p(\mathbf{x}) d\mathbf{x} \int p(\mathbf{y}) d\mathbf{y} \right) = 0. \quad (1.196)$$

Thus,

$$H(\mathbf{x}, \mathbf{y}) \leq H(\mathbf{x}) + H(\mathbf{y}). \quad (1.197)$$

## 1.32

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (1.198)$$

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

$$H(\mathbf{y}) = - \int p_y(\mathbf{y}) \ln p_y(\mathbf{y}) d\mathbf{y}. \quad (1.199)$$

By the transformation

$$p_y(\mathbf{y}) = p_x(\mathbf{A}\mathbf{x}) |\det \mathbf{A}^{-1}|, \quad (1.200)$$

the right hand side can be written as

$$- \int p_x(\mathbf{Ax}) \ln p_x(\mathbf{Ax}) |\det \mathbf{A}| d\mathbf{x} - \ln |\det \mathbf{A}^{-1}| \int p_y(\mathbf{y}) d\mathbf{y}. \quad (1.201)$$

By the transformation

$$\mathbf{x}' = \mathbf{Ax}, \quad (1.202)$$

the first term can be written as

$$- \int p_x(\mathbf{x}') \ln p_x(\mathbf{x}') d\mathbf{x}' = H(\mathbf{x}), \quad (1.203)$$

and the second term can be written as

$$- \ln |\det \mathbf{A}^{-1}| = \ln |\det \mathbf{A}|. \quad (1.204)$$

Therefore,

$$H(\mathbf{y}) = H(\mathbf{x}) + \ln |\det \mathbf{A}|. \quad (1.205)$$

### 1.33

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

$$H(y|x) = - \sum_i \sum_j p(x_i, y_j) \ln p(y_j|x_i). \quad (1.206)$$

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

$$0 = - \sum_i p(x_i) \sum_j p(y_j|x_i) \ln p(y_j|x_i). \quad (1.207)$$

Since

$$\begin{aligned} p(x_i) &\geq 0, \\ p(y_j|x_i) \ln p(y_j|x_i) &\leq 0. \end{aligned} \quad (1.208)$$

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

$$p(y_j|x_i) \ln p(y_j|x_i) = 0. \quad (1.209)$$

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

$$\sum_j p(y_j|x_i) = 1, \quad (1.210)$$

it can be written that

$$p(y_j|x_i) = \delta_{jj'(i)}, \quad (1.211)$$

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

### 1.34

Let

$$\begin{aligned} L(p(x)) = & - \int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1 \left( \int_{-\infty}^{\infty} p(x) dx - 1 \right) \\ & + \lambda_2 \left( \int_{-\infty}^{\infty} xp(x) dx - \mu \right) + \lambda_3 \left( \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right). \end{aligned} \quad (1.212)$$

Then

$$\frac{\delta L(p(x))}{\delta p(x)} = \int_{-\infty}^{\infty} (-\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2) dx. \quad (1.213)$$

Setting the left hand side to zero gives

$$p(x) = \exp \left( -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 \right). \quad (1.214)$$

Therefore,

$$p(x) = \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} + \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right). \quad (1.215)$$

Substituting it to

$$\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= 1, \\ \int_{-\infty}^{\infty} xp(x) dx &= \mu, \\ \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx &= \sigma^2, \end{aligned} \quad (1.216)$$

gives

$$\begin{aligned} \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= 1, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} x \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= \mu, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} (x - \mu)^2 \exp \left( \lambda_3 \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right)^2 \right) dx &= \sigma^2. \end{aligned} \quad (1.217)$$

By the transformation

$$y = \sqrt{-\lambda_3} \left( x - \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) \right), \quad (1.218)$$

they can be written as

$$\begin{aligned} \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= 1, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \left( (-\lambda_3)^{-\frac{1}{2}} y + \mu - \frac{\lambda_2}{2\lambda_3} \right) \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= \mu, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \int_{-\infty}^{\infty} \left( (-\lambda_3)^{-\frac{1}{2}} y - \frac{\lambda_2}{2\lambda_3} \right)^2 \exp(-y^2) (-\lambda_3)^{-\frac{1}{2}} dy &= \sigma^2. \end{aligned} \quad (1.219)$$

Since

$$\begin{aligned} \int_{-\infty}^{\infty} \exp(-y^2) dy &= \Gamma \left( \frac{1}{2} \right), \\ \int_{-\infty}^{\infty} y \exp(-y^2) dy &= 0, \\ \int_{-\infty}^{\infty} y^2 \exp(-y^2) dy &= \Gamma \left( \frac{3}{2} \right), \end{aligned} \quad (1.220)$$

they can be written as

$$\begin{aligned} \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) (-\lambda_3)^{-\frac{1}{2}} \Gamma \left( \frac{1}{2} \right) &= 1, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \left( \mu - \frac{\lambda_2}{2\lambda_3} \right) (-\lambda_3)^{-\frac{1}{2}} \Gamma \left( \frac{1}{2} \right) &= \mu, \\ \exp \left( -1 + \lambda_1 - \frac{\lambda_2^2}{4\lambda_3} \right) \left( (-\lambda_3)^{-\frac{3}{2}} \Gamma \left( \frac{3}{2} \right) + (-\lambda_3)^{-\frac{1}{2}} \frac{\lambda_2^2}{4\lambda_3^2} \Gamma \left( \frac{1}{2} \right) \right) &= \sigma^2. \end{aligned} \quad (1.221)$$

Thus,

$$\begin{aligned} \lambda_1 &= 1 - \frac{1}{2} \ln(2\pi\sigma^2), \\ \lambda_2 &= 0, \\ \lambda_3 &= -\frac{1}{2\sigma^2}, \end{aligned} \quad (1.222)$$

so that

$$p(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right). \quad (1.223)$$

### 1.35

Let  $x$  be a variable under the Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . Then, by the definition,

$$H(x) = - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \ln \mathcal{N}(x|\mu, \sigma^2) dx, \quad (1.224)$$

where

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (1.225)$$

Therefore,

$$H(x) = - \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2 \right) dx. \quad (1.226)$$

The right hand side can be written as

$$\frac{1}{2} \ln(2\pi\sigma^2) \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.227)$$

Thus,

$$H(x) = \frac{1}{2} (1 + \ln(2\pi\sigma^2)). \quad (1.228)$$

### 1.36 (Incomplete)

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

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b), \quad (1.229)$$

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

$$x = \lambda a + (1 - \lambda)b. \quad (1.230)$$

Then, the inequality can be written as

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

Let

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

Then,

$$g(x) \geq 0. \quad (1.233)$$

Additionally, for  $x > a$ ,

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

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

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

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

$$f'(y) \leq f'(c). \quad (1.236)$$

### 1.37

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

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

The right hand side can be written as

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

The right hand side can be written as

$$H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \quad (1.239)$$

Therefore,

$$H(\mathbf{x}, \mathbf{y}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}). \quad (1.240)$$

### 1.38

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

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \quad (1.241)$$

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

$$f\left(\sum_{i=1}^M \lambda_i x_i\right) \leq \sum_{i=1}^M \lambda_i f(x_i), \quad (1.242)$$

where  $\lambda_i \geq 0$  and

$$\sum_{i=1}^M \lambda_i = 1. \quad (1.243)$$

Here, let  $\lambda_i \geq 0$  and

$$\sum_{i=1}^{M+1} \lambda_i = 1. \quad (1.244)$$

Then, by the definition,

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

By the assumption,

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

Therefore,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \lambda_{M+1} f(x_{M+1}) + (1 - \lambda_{M+1}) \sum_{i=1}^M \frac{\lambda_i}{1 - \lambda_{M+1}} f(x_i). \quad (1.247)$$

Thus,

$$f\left(\sum_{i=1}^{M+1} \lambda_i x_i\right) \leq \sum_{i=1}^{M+1} \lambda_i f(x_i). \quad (1.248)$$

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

### 1.39

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

$$\begin{aligned}p(x = 0, y = 0) &= \frac{1}{3}, \\p(x = 0, y = 1) &= \frac{1}{3}, \\p(x = 1, y = 0) &= 0, \\p(x = 1, y = 1) &= \frac{1}{3}.\end{aligned}\tag{1.249}$$

(a)

By the definition,

$$H(x) = - \sum p(x) \ln p(x).\tag{1.250}$$

By the distribution,

$$\begin{aligned}p(x = 0) &= \frac{2}{3}, \\p(x = 1) &= \frac{1}{3}.\end{aligned}\tag{1.251}$$

Therefore,

$$H(x) = \ln 3 - \frac{2}{3} \ln 2.\tag{1.252}$$

(b)

By the definition,

$$H(y) = - \sum p(y) \ln p(y).\tag{1.253}$$

By the distribution,

$$\begin{aligned}p(y = 0) &= \frac{1}{3}, \\p(y = 1) &= \frac{2}{3}.\end{aligned}\tag{1.254}$$

Therefore,

$$H(y) = \ln 3 - \frac{2}{3} \ln 2.\tag{1.255}$$



(c)

By the definition,

$$H(y|x) = - \sum p(x, y) \ln p(y|x). \quad (1.256)$$

By the definition,

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

Then, by the distribution,

$$\begin{aligned} p(y = 0|x = 0) &= \frac{1}{2}, \\ p(y = 0|x = 1) &= 0, \\ p(y = 1|x = 0) &= \frac{1}{2}, \\ p(y = 1|x = 1) &= 1. \end{aligned} \quad (1.258)$$

Therefore,

$$H(y|x) = \frac{2}{3} \ln 2. \quad (1.259)$$

(d)

By the definition,

$$H(x|y) = - \sum p(x, y) \ln p(x|y). \quad (1.260)$$

By the definition,

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

Then, by the distribution,

$$\begin{aligned}
p(x=0|y=0) &= 1, \\
p(x=0|y=1) &= \frac{1}{2}, \\
p(x=1|y=0) &= 0, \\
p(x=1|y=1) &= \frac{1}{2}.
\end{aligned} \tag{1.262}$$

Therefore,

$$H(x|y) = \frac{2}{3} \ln 2. \tag{1.263}$$

(e)

By the definition,

$$H(x, y) = - \sum p(x, y) \ln p(x, y). \tag{1.264}$$

Therefore,

$$H(x, y) = \ln 3. \tag{1.265}$$

(f)

By the definition,

$$I(x, y) = - \sum p(x, y) \ln \frac{p(x)p(y)}{p(x, y)}. \tag{1.266}$$

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

$$H(x) + H(y) - H(x, y). \quad (1.267)$$

Therefore,

$$I(x, y) = \ln 3 - \frac{4}{3} \ln 2. \quad (1.268)$$

## 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

$$\sum_{i=1}^M \lambda_i = 1. \quad (1.269)$$

By Jensen's inequality,

$$\sum_{i=1}^M \lambda_i \ln x_i \leq \ln \left( \sum_{i=1}^M \lambda_i x_i \right). \quad (1.270)$$

Therefore,

$$\prod_{i=1}^M x_i^{\lambda_i} \leq \sum_{i=1}^M \lambda_i x_i. \quad (1.271)$$

Substituting

$$\lambda_i = \frac{1}{M} \quad (1.272)$$

gives

$$\left( \prod_{i=1}^M x_i \right)^{\frac{1}{M}} \leq \frac{1}{M} \sum_{i=1}^M x_i. \quad (1.273)$$