

Solutions Manual to Pattern Recognition and Machine Learning

Hiromichi Inawashiro

July 22, 2024

Contents

1	Introduction	1
2	Probability Distributions	36

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, θ) , the right hand side can be written as

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

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 such that

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

Then

$$\mathbb{E}x = \int_{-\infty}^{\infty} x \mathcal{N}(x|\mu, \sigma^2) dx. \quad (1.44)$$

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

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

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

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

Since

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

and

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

we have

$$\mathbb{E}x = \mu. \quad (1.49)$$

By the definition,

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

can be written as

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

Differentiating both sides with respect to σ^2 gives

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

The left hand side can be written as

$$\begin{aligned} -\frac{1}{2}(\sigma^2)^{-1} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx + \frac{1}{2}(\sigma^2)^{-2} \int_{-\infty}^{\infty} (x - \mu)^2 \mathcal{N}(x|\mu, \sigma^2) dx \\ = -\frac{1}{2}(\sigma^2)^{-1} + \frac{1}{2}(\sigma^2)^{-2} \text{var} x. \end{aligned} \quad (1.53)$$

Therefore,

$$\text{var} x = \sigma^2. \quad (1.54)$$

1.9

Let

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

Setting its derivative with respect to x to zero gives

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

Therefore, the mode is given by μ .

Similarly, let

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

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

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

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

1.10

By the definition,

$$\mathbb{E}(x + y) = \int \int (x + y) p(x, y) dx dy. \quad (1.59)$$

The right hand side can be written as

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

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

$$Ex + Ey. \quad (1.61)$$

Therefore,

$$E(x + y) = Ex + Ey. \quad (1.62)$$

Similarly, by the definition,

$$\text{var}(x + y) = E(x + y - E(x + y))^2 \quad (1.63)$$

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

$$\begin{aligned} E(x - Ex)^2 + 2E((x - Ex)(y - Ey)) + E(y - Ey)^2 \\ = \text{var}x + 2\text{cov}(x, y) + \text{var}y. \end{aligned} \quad (1.64)$$

If x and y are independent, then

$$\text{cov}(x, y) = 0, \quad (1.65)$$

by 1.6. Therefore,

$$\text{var}(x + y) = \text{var}x + \text{var}y. \quad (1.66)$$

1.11

Let

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

To maximise it with respect to μ and σ^2 , setting the partial derivatives to zero gives

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

Therefore,

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

1.12

Let x_m and x_n be independent variables. Then

$$\mathbb{E}x_mx_n = \mathbb{E}x_m\mathbb{E}x_n.\tag{1.70}$$

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

$$\mathbb{E}x_n^2 = \text{var}x_n + (\mathbb{E}x_n)^2.\tag{1.71}$$

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

$$\mathbb{E}x_mx_n = \mu^2 + \delta_{mn}\sigma^2.\tag{1.72}$$

Here, since

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

we have

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

Therefore,

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

Similarly, since

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

we have

$$\mathbb{E}\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \mathbb{E}(x_n - \mu_{\text{ML}})^2.\tag{1.77}$$

The right hand side can be written as

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

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

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

while the second and third terms can be written as

$$-2\mathbb{E} \mu_{\text{ML}}^2 + \mathbb{E} \mu_{\text{ML}}^2 = -\mathbb{E} \mu_{\text{ML}}^2. \quad (1.80)$$

Here,

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

The right hand side can be written as

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

Therefore,

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

Thus,

$$\mathbb{E} \sigma_{\text{ML}}^2 = \frac{N-1}{N} \sigma^2. \quad (1.84)$$

1.13

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

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

The right hand side can be written as

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E} (x_n^2 - 2\mu x_n + \mu^2) = \frac{1}{N} \sum_{n=1}^N \mathbb{E} x_n^2 - \frac{2\mu}{N} \sum_{n=1}^N \mathbb{E} x_n + \mu^2. \quad (1.86)$$

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

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

while the second term can be written as

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

Therefore,

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

1.14

Let

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

Then

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

Here,

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

The right hand side can be written as

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

Therefore,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij}^A x_i x_j = 0. \quad (1.94)$$

Additionally,

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

The right hand side can be written as

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

where the result above is used. Therefore,

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = \sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j. \quad (1.97)$$

Finally, since the matrix w_{ij}^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

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

Then

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

The right hand side can be written as

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

Therefore,

$$\Gamma(x+1) = x \Gamma(x). \quad (1.101)$$

Since

$$\Gamma(1) = \int_0^1 \exp(-u) du, \quad (1.102)$$

and the right hand side can be written as 1,

$$\Gamma(1) = 0!. \quad (1.103)$$

For a positive integer x , let us assume that

$$\Gamma(x) = (x-1)!. \quad (1.104)$$

Then,

$$\Gamma(x+1) = x\Gamma(x), \quad (1.105)$$

where the right hand side can be written as

$$x(x-1)! = x!. \quad (1.106)$$

Therefore,

$$\Gamma(x+1) = x!. \quad (1.107)$$

Thus, the assumption is proved by induction on x .

1.18

Let us consider the transformation from Cartesian to polar coordinates

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

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

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

Therefore,

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

Additionally, the volume of the sphere can be written as

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

The right hand side can be written as

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

Therefore,

$$V_D = \frac{S_D}{D}. \quad (1.113)$$

Finally, the results above reduce to

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

Therefore,

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

Similarly,

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

Therefore,

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

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

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

by 1.18.

Additionally, by Sterling's formula

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

the ratio can be approximated as

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

The right hand side can be written as

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

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

$$\frac{\sqrt{\sum_{i=1}^D 1^2}}{1} = \sqrt{D}. \quad (1.122)$$

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

1.20

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

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

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

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

where ϕ 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 ϵ , the right hand side can be approximated as

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

Therefore,

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

where

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

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

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

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

$$\hat{r} = \sqrt{D-1}\sigma. \quad (1.129)$$

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

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

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

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

By the Taylor series

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

the right hand side can be approximated as

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

Therefore,

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

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

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

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

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

1.21

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

$$0 \leq a(b-a). \quad (1.137)$$

Therefore,

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

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,

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

and

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

Then, using the inequality above,

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

and

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

Therefore,

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

1.22

Let

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

If

$$L_{kj} = 1 - \delta_{kj}, \quad (1.145)$$

then the right hand side can be written as

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

The right hand side can be written as

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

Therefore,

$$EL = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j | \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.148)$$

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

1.23

Let

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

The right hand side can be written as

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

Therefore,

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

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

1.24 (Incomplete)

1.25

Let

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

Then

$$\frac{\delta EL(\mathbf{t}, \mathbf{y}(\mathbf{x}))}{\delta \mathbf{y}(\mathbf{x})} = 2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{t}. \quad (1.153)$$

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

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

The right hand side can be written as

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

Thus,

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

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

$$y(\mathbf{x}) = E_t(t|\mathbf{x}). \quad (1.157)$$

1.26

Let

$$EL(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \quad (1.158)$$

The right hand side can be written as

$$\begin{aligned} & \int \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &= \int \int \|\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &+ 2 \int \int (\mathbf{y}(\mathbf{x}) - E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top (E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &+ \int \int \|E_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \end{aligned} \quad (1.159)$$

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

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

The second term can be written as

$$2 \int (\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}))^\top \left(\int (\mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}) p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \right) p(\mathbf{x}) d\mathbf{x}. \quad (1.161)$$

Since

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

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

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

Therefore,

$$EL(\mathbf{t}, \mathbf{y}(\mathbf{x})) = \int \|\mathbf{y}(\mathbf{x}) - \mathbf{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}) d\mathbf{x} + \int \text{var}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1.164)$$

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

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

1.27

Let

$$EL_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \quad (1.166)$$

Then

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

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

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

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

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

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

1.28

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \quad (1.170)$$

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

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

Therefore,

$$h(p^2) = 2h(p). \quad (1.172)$$

Let us assume that, for a positive integer n ,

$$h(p^n) = nh(p). \quad (1.173)$$

Then, by the first assumption,

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

Therefore,

$$h(p^{n+1}) = (n+1)h(p). \quad (1.175)$$

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

Additionally, for positive integers m and n ,

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

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

$$h(p^{\frac{n}{m}}) = \frac{n}{m}h(p). \quad (1.177)$$

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

$$h(p^a) = ah(p). \quad (1.178)$$

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

$$(p \ln p)h'(p) = h(p). \quad (1.179)$$

Therefore,

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

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,

$$h(p) \propto \ln p. \quad (1.181)$$

1.29

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

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

where

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

By Jensen's inequality,

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

Therefore,

$$H(x) \leq \ln M. \quad (1.185)$$

1.30

Let

$$\begin{aligned} p(x) &= \mathcal{N}(x|\mu, \sigma^2), \\ q(x) &= \mathcal{N}(x|m, s^2). \end{aligned} \quad (1.186)$$

Then, by the definition,

$$\text{KL}(p||q) = - \int p(x) \ln \frac{q(x)}{p(x)} dx. \quad (1.187)$$

The right hand side can be written as

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

The right hand side can be written as

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

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

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

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

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

1.31

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

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

Note that

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

Therefore,

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

Since

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

Jensen's inequality can be used to write that

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

The right hand side can be written as

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

Thus,

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

1.32

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad (1.199)$$

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

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

By the transformation

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

the right hand side can be written as

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

By the transformation

$$\mathbf{x}' = \mathbf{Ax}, \quad (1.203)$$

the first term can be written as

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

and the second term can be written as

$$- \ln |\det \mathbf{A}^{-1}| = \ln |\det \mathbf{A}|. \quad (1.205)$$

Therefore,

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

1.33

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

$$H(y|x) = - \sum_i \sum_j p(x_i, y_j) \ln p(y_j|x_i). \quad (1.207)$$

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

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

Since

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

for all i and j , the equation reduces to

$$p(y_j|x_i) \ln p(y_j|x_i) = 0. \quad (1.210)$$

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

$$\sum_j p(y_j|x_i) = 1, \quad (1.211)$$

it can be written that

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

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

1.34

Let

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

Then

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

Setting the left hand side to zero gives

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

Therefore,

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

Substituting it to

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

gives

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

By the transformation

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

they can be written as

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

Since

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

they can be written as

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

Therefore,

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

Thus,

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

1.35

Let x be a variable such that

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

Then, by the definition,

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

The right hand side can be written as

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

Therefore,

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

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

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

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

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

1.38

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

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2), \quad (1.240)$$

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

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

where $\lambda_i \geq 0$ and

$$\sum_{i=1}^M \lambda_i = 1. \quad (1.242)$$

Here, let $\lambda_i \geq 0$ and

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

Then, by the definition,

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

By the assumption,

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

Therefore,

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

Thus,

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

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,

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

By the distribution,

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

Therefore,

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

(b)

By the definition,

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

By the distribution,

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

Therefore,

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

(c)

By the definition,

$$H(y|x) = - \sum p(x, y) \ln p(y|x). \quad (1.255)$$

By the definition,

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

Then, by the distribution,

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

Therefore,

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

(d)

By the definition,

$$H(x|y) = - \sum p(x, y) \ln p(x|y). \quad (1.259)$$

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,

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

(e)

By the definition,

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

Therefore,

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

(f)

By the definition,

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

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

$$H(x) + H(y) - H(x, y). \quad (1.266)$$

Therefore,

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

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

$$\sum_{i=1}^M \lambda_i = 1. \quad (1.268)$$

By Jensen's inequality,

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

Therefore,

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

Substituting

$$\lambda_i = \frac{1}{M} \quad (1.271)$$

gives

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

1.41

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

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

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 $H(\mathbf{x})$ and the second term can be written as $-H(\mathbf{x}|\mathbf{y})$. Therefore,

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

By the definition,

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

Thus,

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

2 Probability Distributions

2.1

Let x be a variable such that

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

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

$$\sum_x p(x|\mu) = 1. \quad (2.2)$$

By the definition,

$$\begin{aligned} \mathbb{E}x &= \mu, \\ \mathbb{E}x^2 &= \mu, \end{aligned} \quad (2.3)$$

Since

$$\text{var}x = \mathbb{E}x^2 - (\mathbb{E}x)^2, \quad (2.4)$$

we have

$$\text{var}x = \mu(1 - \mu). \quad (2.5)$$

By the definition,

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

Therefore,

$$\mathbb{H}(x) = -\mu \ln \mu - (1 - \mu) \ln(1 - \mu). \quad (2.7)$$

2.2

Let x be a variable such that

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

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

$$\sum_x p(x|\mu) = 1. \quad (2.9)$$

By the definition,

$$\begin{aligned} \mathbb{E}x &= \mu, \\ \mathbb{E}x^2 &= 1, \end{aligned} \quad (2.10)$$

Since

$$\text{var}x = \text{E}x^2 - (\text{E}x)^2, \quad (2.11)$$

we have

$$\text{var}x = 1 - \mu^2. \quad (2.12)$$

By the definition,

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

Therefore,

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

2.3

By the definition,

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

Therefore,

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

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

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

Thus,

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}. \quad (2.18)$$

Note that

$$1+x = \sum_{m=0}^1 \binom{1}{m} x^m. \quad (2.19)$$

Let us assume that

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

Then,

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

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

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

Therefore,

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

Thus, the assumption is proved by induction on N .

Finally, let m be a variable such that

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

Then

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

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

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

Therefore,

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

2.4

Let m be a variable such that

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

Then

$$Em = \sum_{m=0}^N m \binom{N}{m} \mu^m (1-\mu)^{N-m}. \quad (2.29)$$

Differentiating both sides of

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

with respect to μ gives

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

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

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

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

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

Therefore,

$$Em = N\mu. \quad (2.34)$$

Differentiating both sides of

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

twice with respect to μ gives

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

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

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

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

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

Therefore,

$$Em(m-1) = N(N-1)\mu^2. \quad (2.39)$$

Thus, since

$$\text{var } m = Em(m-1) + Em - (Em)^2, \quad (2.40)$$

we have

$$\text{var } m = N\mu(1-\mu). \quad (2.41)$$

2.5

By the definition,

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

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

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

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

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

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

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

2.6

Let μ be a variable such that

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

Then

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

Since

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

we have

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

Since

$$\text{var}\mu = E\mu^2 - (E\mu)^2, \quad (2.50)$$

we have

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

Since

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

we have

$$\text{mode}\mu = \frac{a-1}{a+b-2}. \quad (2.53)$$

2.7

Let m and l be a variable such that

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

where

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

By 2.6,

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

Note that

$$\mu_{\text{ML}} = \frac{m}{m+l}. \quad (2.57)$$

Since

$$p(\mu|m, l, a, b) \propto p(m, l|\mu)p(\mu|a, b), \quad (2.58)$$

we have

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

Therefore, by 2.6,

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

Thus,

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

where

$$\lambda = \frac{m+l}{m+l+a+b}. \quad (2.62)$$

2.8 (Incomplete)

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

$$Ex = \int xp(x)dx. \quad (2.63)$$

The right hand side can be written as

$$\int x \left(\int p(x, y)dy \right) dx = \int \left(\int xp(x|y)dx \right) p(y)dy. \quad (2.64)$$

Therefore,

$$Ex = E_y (E_x(x|y)) . \quad (2.65)$$

By the definition,

$$\text{var} x = E (x - Ex)^2 . \quad (2.66)$$

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

$$\begin{aligned} & E_y (E_x ((x - E_x(x|y) + E_x(x|y) - Ex)^2 | y)) \\ &= E_y (E_x ((x - E_x(x|y))^2 | y)) \\ &+ 2E_y (E_x ((x - E_x(x|y)) (E_x(x|y) - Ex) | y)) \\ &+ E_y (E_x ((E_x(x|y) - Ex)^2 | y)) \end{aligned} \quad (2.67)$$

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

$$2E_y ((E_x(x|y) - Ex) E_x ((x - E_x(x|y)) | y)) \quad (2.68)$$

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

$$E_y (E_x(x|y) - E_y (E_x(x|y)))^2 = \text{var}_y (E_x(x|y)) . \quad (2.69)$$

Therefore,

$$\text{var} x = E_y (\text{var}_x(x|y)) + \text{var}_y (E_x(x|y)) . \quad (2.70)$$

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

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

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

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

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

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

By the assumption, it can be written as

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

Therefore,

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

Thus, the assumption is proved by induction on M .

2.10

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

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

Then

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

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

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

Therefore,

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

Since

$$\begin{aligned} \text{var}\mu_j &= E\mu_j^2 - (E\mu_j)^2, \\ \text{cov}(\mu_j, \mu_l) &= E\mu_j \mu_l - E\mu_j E\mu_l, \end{aligned} \quad (2.78)$$

we have

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

2.11

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

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

Then

$$\mathbb{E} \ln \mu_j = \int (\ln \mu_j) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}. \quad (2.81)$$

Since

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

we have

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

where

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

Therefore,

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

2.12

Let x be a variable such that

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

where $a < b$. Then

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

Note that

$$\mathbb{E} x = \int_a^b x p(x|a, b) dx, \quad (2.88)$$

$$\mathbb{E} x^2 = \int_a^b x^2 p(x|a, b) dx.$$

The right hand sides can be written as

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

Therefore,

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

Since

$$\text{var}x = \mathbb{E}x^2 - (\mathbb{E}x)^2, \quad (2.91)$$

we have

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

2.13

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

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

Then, by the definition,

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

Since

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

The right hand side can be written as

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

Let us look at each term. Since

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

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

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

the second term can be written as

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

Since

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

it can be written as

$$\frac{1}{2} \text{tr} (\mathbf{L}^{-1} \mathbf{\Sigma}) + \frac{1}{2} (\boldsymbol{\mu} - \mathbf{m})^\top \mathbf{L}^{-1} (\boldsymbol{\mu} - \mathbf{m}). \quad (2.101)$$

Since

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

the third term can be written as

$$-\frac{1}{2} \text{tr} (\mathbf{\Sigma}^{-1} \mathbf{\Sigma}) = -\frac{D}{2} \quad (2.103)$$

Therefore,

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

2.14

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

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

Then

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

Setting the left hand side to zero gives

$$p(\mathbf{x}) = \exp(-1 + \lambda + \mathbf{l}^\top \mathbf{x} + \mathbf{m}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{m}), \quad (2.107)$$

so that

$$p(\mathbf{x}) = \exp(-1 + \lambda - \mathbf{l}^\top \mathbf{M} \mathbf{l} + (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M} \mathbf{l})^\top \mathbf{M}^{-1} (\mathbf{x} - \boldsymbol{\mu} - \mathbf{M} \mathbf{l})), \quad (2.108)$$

where

$$\mathbf{M} = (\mathbf{m} \mathbf{m}^\top)^{-1}. \quad (2.109)$$

Substituting it to

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

and the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} - \mathbf{M} \mathbf{l} \quad (2.111)$$

gives

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

Since

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

they can be written as

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

Therefore,

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

Thus,

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

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,

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

The right hand side can be written as

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

Since

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

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

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

Therefore,

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

2.16

Let

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

where x_1 and x_2 are variables such that

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

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

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

The right hand side can be written as

$$\begin{aligned}
& \int_{-\infty}^{\infty} \frac{(\tau_1 \tau_2)^{\frac{1}{2}}}{2\pi} \exp\left(-\frac{\tau_1 + \tau_2}{2} \left(x_2 - \frac{\tau_1(x - \mu_1) + \tau_2 \mu_2}{\tau_1 + \tau_2}\right)^2\right) dx_2 \\
& \exp\left(-\frac{\tau_1}{2}(x - \mu_1)^2 - \frac{\tau_2}{2}\mu_2^2 + \frac{\tau_1 + \tau_2}{2} \left(\frac{\tau_1(x - \mu_1) + \tau_2 \mu_2}{\tau_1 + \tau_2}\right)^2\right) \\
& \propto \exp\left(-\frac{\tau_1 \tau_2}{2(\tau_1 + \tau_2)}(x - \mu_1 - \mu_2)^2\right).
\end{aligned} \tag{2.128}$$

Therefore,

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

Thus, by 1.35,

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

2.17

Let Σ be a matrix and

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

Then

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

Therefore,

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

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

$$\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma}^{-1})^\top (\mathbf{x} - \boldsymbol{\mu}). \quad (2.134)$$

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

$$-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))^\top (\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (2.135)$$

Thus,

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

Hence

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

2.18

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

$$\boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.138)$$

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

$$\bar{\mathbf{u}}_i^\top \boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i. \quad (2.139)$$

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

$$\bar{\mathbf{u}}_i^\top \boldsymbol{\Sigma}^\top \mathbf{u}_i = (\bar{\boldsymbol{\Sigma} \mathbf{u}}_i)^\top \mathbf{u}_i. \quad (2.140)$$

The right hand side can be written as

$$\bar{\lambda}_i \bar{\mathbf{u}}_i^\top \mathbf{u}_i = \bar{\lambda}_i. \quad (2.141)$$

Therefore,

$$\lambda_i = \bar{\lambda}_i. \quad (2.142)$$

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

$$\mathbf{u}_j^\top \boldsymbol{\Sigma} \mathbf{u}_i = \lambda_i \mathbf{u}_j^\top \mathbf{u}_i. \quad (2.143)$$

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

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

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

$$\lambda_i \mathbf{u}_j^\top \mathbf{u}_i = \lambda_j \mathbf{u}_j^\top \mathbf{u}_i. \quad (2.145)$$

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

$$\mathbf{u}_j^\top \mathbf{u}_i = 0. \quad (2.146)$$

2.19

Let Σ be a $D \times D$ real symmetric matrix such that

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.147)$$

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

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

Then

$$\Sigma \mathbf{U} = \mathbf{U} \Lambda. \quad (2.149)$$

By 2.18,

$$\mathbf{U}^\top \mathbf{U} = \mathbf{I}. \quad (2.150)$$

Therefore,

$$\begin{aligned} \Sigma &= \mathbf{U} \Lambda \mathbf{U}^\top, \\ \Sigma^{-1} &= \mathbf{U} \Lambda^{-1} \mathbf{U}^\top, \end{aligned} \quad (2.151)$$

Thus,

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

2.20

Let Σ be a $D \times D$ real symmetric matrix such that

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad (2.153)$$

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

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

By 2.19,

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

where

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

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

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

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 $\frac{D(D+1)}{2}$.