Solutions Manual to
Pattern Recognition and Machine Learning

Hiromichi Inawashiro

June 26, 2024

1 Introduction

1.1

Let

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

To minimise it, setting the derivative to zero gives

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

Substituting

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

gives

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

Therefore,

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

where

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

1.2

Let

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

To minimise it, setting the derivative to zero gives

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

Substituting

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

gives

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

Therefore,

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

where

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

1.3

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

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

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

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

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

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

Here,

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

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

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

1.4

Let

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

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

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

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

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

Therefore, in general,

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

Here, let us assume that

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

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

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

Therefore,

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

1.5

By the definition,

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

The right hand side can be written as

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

Therefore,

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

1.6

By the definition,

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

The right hand side can be written as

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

The right hand side can be written as

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

If x and y are independent, by the definition,

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

Then,

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

Therefore,

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

1.7

Let

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

Then

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

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

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

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

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

Therefore,

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

By the definition,

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

Then

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

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

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

Therefore,

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

1.8

Let x be a variable under the Gaussian distribution with mean μ and variance σ^2. Then

$$\mathrm{E}x = \int_{-\infty}^{\infty} x\mathcal{N}\left(x|\mu,\sigma^2\right) dx. \tag{1.43}$$

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

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

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

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

Since

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

and

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

we have

$$\mathrm{E}x = \mu. \tag{1.48}$$

By the definition,

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

can be written as

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

Differentiating both sides with respect to σ^2 gives

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

The left hand side can be written as

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

Therefore,

$$\mathrm{var}x=\sigma^2. \tag{1.53}$$

1.9

Let

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

Setting its derivative with respect to x to zero gives

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

Therefore, the mode is given by μ.

Similarly, let

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

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

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

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

1.10

By the definition,

$$\mathrm{E}(x+y)=\int\int(x+y)p(x,y)dxdy. \tag{1.58}$$

The right hand side can be written as

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

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

$$\mathrm{E}x+\mathrm{E}y. \tag{1.60}$$

Therefore,

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

Similarly, by the definition,

$$\mathrm{var}(x+y)=\mathrm{E}\left(x+y-\mathrm{E}(x+y)\right)^2 \tag{1.62}$$

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

$$\begin{aligned}\mathrm{E}\left(x-\mathrm{E}x\right)^2+2\mathrm{E}\left(\left(x-\mathrm{E}x\right)\left(y-\mathrm{E}y\right)\right)+\mathrm{E}\left(y-\mathrm{E}y\right)^2 \\ =\mathrm{var}x+2\mathrm{cov}(x,y)+\mathrm{var}y.\end{aligned} \tag{1.63}$$

If x and y are independent, then

$$\mathrm{cov}(x,y)=0, \tag{1.64}$$

by 1.6. Therefore,

$$\mathrm{var}(x+y)=\mathrm{var}x+\mathrm{var}y. \tag{1.65}$$

1.11

Let

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

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

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

Therefore,

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

1.12

Let x_m and x_n be independent variables. Then

$$\mathrm{E}x_m x_n = \mathrm{E}x_m \mathrm{E}x_n. \tag{1.69}$$

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,

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

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,

$$\mathrm{E}x_m x_n = \mu^2 + \delta_{mn}\sigma^2. \tag{1.71}$$

Here, since

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

we have

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

Therefore,

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

Similarly, since

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

we have

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

The right hand side can be writen as

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

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

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

while the second and third terms can be writen as

$$-2\mathrm{E}\mu_{\mathrm{ML}}^2+\mathrm{E}\mu_{\mathrm{ML}}^2=-\mathrm{E}\mu_{\mathrm{ML}}^2. \tag{1.79}$$

Here,

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

The right hand side can be written as

$$\frac{1}{N^2}\sum_{n=1}^{N}\mathrm{E}x_n^2+\frac{2}{N^2}\sum_{1\le m<n\le N}\mathrm{E}x_mx_n=\frac{1}{N}(\mu^2+\sigma^2)+\frac{N-1}{N}\mu^2. \tag{1.81}$$

Therefore,

$$\mathrm{E}\mu_{\mathrm{ML}}^2=\mu^2+\frac{1}{N}\sigma^2. \tag{1.82}$$

Thus,

$$\mathrm{E}\sigma_{\mathrm{ML}}^2=\frac{N-1}{N}\sigma^2. \tag{1.83}$$

1.13

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

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

The right hand side can be writen as

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

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

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

while the second term can be writen as

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

Therefore,

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

1.14

Let

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

Then

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

Here,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_ix_j=\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}-w_{ji})x_ix_j. \tag{1.91}$$

The right hand side can be written as

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

Therefore,

$$\sum_{i=1}^{D}\sum_{j=1}^{D}w_{ij}^{\mathrm{A}}x_ix_j=0. \tag{1.93}$$

Additionally,

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

The right hand side can be written as

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

where the result above is used. Therefore,

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

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

1.17

Let

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

Then

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

The right hand side can be written as

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

Therefore,

$$\Gamma(x+1) = x\Gamma(x). \tag{1.100}$$

Since

$$\Gamma(1) = \int_0^1 \exp(-u)du, \tag{1.101}$$

and the right hand side can be written as 1,

$$\Gamma(1) = 0!. \tag{1.102}$$

For a positive integer x, let us assume that

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

Then,

$$\Gamma(x+1) = x\Gamma(x), \tag{1.104}$$

where the right hand side can be written as

$$x(x-1)! = x!. \tag{1.105}$$

Therefore,

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

Thus, the assumption is proved by induction on x.

1.18

Let us consider the transformation from Cartesian to polar coordinates

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

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

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

Therefore,

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

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

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

The right hand side can be written as

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

Therefore,

$$V_D = \frac{S_D}{D}. \tag{1.112}$$

Finally, the results above reduce to

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

Therefore,

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

Similarly,

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

Therefore,

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

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

$$\frac{V_D}{2^D} = \frac{\pi^{\frac{D}{2}}}{D2^{D-1}\Gamma\left(\frac{D}{2}\right)}, \tag{1.117}$$

by 1.18.

Additionally, by Stering's formula

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

the ratio can be approximated as

$$\frac{V_D}{2^D} \simeq \frac{\pi^{\frac{D}{2}}}{D2^{D-1}(2\pi)^{\frac{1}{2}} \exp\left(1-\frac{D}{2}\right)\left(\frac{D}{2}-1\right)^{\frac{D}{4}}}. \tag{1.119}$$

The right hand side can be written as

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

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

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

$$\frac{\sqrt{\sum_{i=1}^{D} 1^2}}{1} = \sqrt{D}. \tag{1.121}$$

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

1.20

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

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

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

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

where $\boldsymbol{\phi}$ is the vector of the angular components of the polar corrdinate and J is the Jacobian of the transformation from the Cartesian to polar coordinate. For a sufficiently small ϵ, the right hand side can be approximated as

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

Therefore,

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

where

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

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

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

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

$$\hat{r} = \sqrt{D-1}\sigma. \tag{1.128}$$

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

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

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

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

By the Taylor series

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

the right hand side can be approximated as

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

Therefore,

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

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

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

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

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

1.21

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

$$0 \leq a(b-a). \tag{1.136}$$

Therefore,

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

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,

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

and

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

Then, using the inequality above,

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

and

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

Therefore,

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

1.22

Let

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

If

$$L_{kj} = 1 - \delta_{kj}, \tag{1.144}$$

then the right hand side can be written as

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

The right hand side can be written as

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

Therefore,

$$\mathrm{E}L = 1 - \sum_j \int_{\mathcal{R}_j} p(\mathcal{C}_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{1.147}$$

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

1.23

Let

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

The right hand side can be written as

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

Therefore,

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

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

1.25

Let

$$\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.151}$$

Then

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

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

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

The right hand side can be written as

$$\mathbf{y}(\mathbf{x}) \int p(\mathbf{t}|\mathbf{x}) d\mathbf{t} - \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}). \tag{1.154}$$

Thus,

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

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

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

1.26

Let

$$\mathrm{E}L\left(\mathbf{t}, \mathbf{y}(\mathbf{x})\right) = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}. \tag{1.157}$$

The right hand side can be written as

$$\begin{aligned}
&\int \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) + \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
&= \int \int \|\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
&+ 2 \int \int \left(\mathbf{y}(\mathbf{x}) - \mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x})\right)^{\intercal} \left(\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\right) p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\
&+ \int \int \|\mathrm{E}_{\mathbf{t}}(\mathbf{t}|\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}.
\end{aligned} \tag{1.158}$$

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

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

The second term can be written as

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

Since

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

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

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

Therefore,

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

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

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

1.27

Let

$$\mathrm{E}L_q = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt. \tag{1.165}$$

Then

$$\frac{\delta \mathrm{E}L_q}{\delta y(\mathbf{x})} = \int q|y(\mathbf{x}) - t|^{q-1} \mathrm{sign}(y(\mathbf{x}) - t) p(\mathbf{x}, t) dt. \tag{1.166}$$

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

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

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

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

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

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

1.28

Let us assume that

$$p(x, y) = p(x)p(y) \Rightarrow h(x, y) = h(x) + h(y). \tag{1.169}$$

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

$$h\left(p^2\right) = h(p) + h(p). \tag{1.170}$$

Therefore,

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

Let us assume that, for a positive integer n,

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

Then, by the first assumption,

$$h\left(p^{n+1}\right) = h\left(p^n\right) + h(p). \tag{1.173}$$

Therefore,

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

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

Additionally, for positive integers m and n,

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

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

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

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

$$h\left(p^a\right) = ah(p). \tag{1.177}$$

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

$$(p \ln p)h'(p) = h(p). \tag{1.178}$$

Therefore,

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

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

$$h(p) \propto \ln p. \tag{1.180}$$

1.29

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

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

where

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

By Jensen's inequality,

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

Therefore,

$$\mathrm{H}(x) \leq \ln M. \tag{1.184}$$

1.30

Let

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

Then, by the definition,

$$\mathrm{KL}\left(p||q\right) = -\int p(x) \ln \frac{q(x)}{p(x)} dx. \tag{1.186}$$

The right hand side can be written as

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

The right hand side can be written as

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

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

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

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

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

1.31

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

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

Note that

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

Therefore,

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

Since

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

Jensen's inequality can be used to write that

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

The right hand side can be written as

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

Thus,

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

1.32

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1.198}$$

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

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

By the transformation

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

the right hand side can be written as

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

By the transformation

$$\mathbf{x}' = \mathbf{A}\mathbf{x}, \tag{1.202}$$

the first term can be written as

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

and the second term can be written as

$$-\ln \left|\det \mathbf{A}^{-1}\right| = \ln |\det \mathbf{A}|. \tag{1.204}$$

Therefore,

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

1.33

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

$$\mathrm{H}(y|x) = -\sum_i \sum_j p(x_i, y_j) \ln p(y_j|x_i). \tag{1.206}$$

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

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

Since

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

for all i and j, the equation reduces to

$$p(y_j|x_i) \ln p(y_j|x_i) = 0. \tag{1.209}$$

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

$$\sum_j p(y_j|x_i) = 1, \tag{1.210}$$

it can be written that

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

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

1.34

Let

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

Then

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

Setting the left hand side to zero gives

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

Therefore,

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

Substituting it to

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

gives

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

By the transformation

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

they can be written as

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

Since

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

they can be written as

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

Thus,

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

so that

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

1.35

Let x be a variable under the Gaussian distribution with mean μ and variance σ^2. Then, by the definition,

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

where

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

Therefore,

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

The right hand side can be written as

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

Thus,

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

1.36

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

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

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

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

Then, the inequality can be written as

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

Let

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

Then,

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

Additionally, for $x > a$,

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

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

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

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

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

1.37

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

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

The right hand side can be written as

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

The right hand side can be written as

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

Therefore,

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

1.38

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

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda f(x_1) + (1-\lambda)f(x_2), \tag{1.241}$$

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

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

where $\lambda_i \geq 0$ and

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

Here, let $\lambda_i \geq 0$ and

$$\sum_{i=1}^{M+1}\lambda_i = 1. \tag{1.244}$$

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.