Solution Manual For Pattern Recognition and Machine Learning

Edited By

ZHENGQI GAO

Information Science and Technology School
Fudan University

Nov.2017

0.1 Introduction

Problem 1.1 Solution

We let the derivative of *error function* E with respect to vector $\mathbf{w}$ equals to $\mathbf{0}$, (i.e. $\frac{\partial E}{\partial \mathbf{w}} = 0$), and this will be the solution of $\mathbf{w} = \{w_i\}$ which minimizes *error function* E. To solve this problem, we will calculate the derivative of E with respect to every w_i , and let them equal to 0 instead. Based on (1.1) and (1.2) we can obtain :

=>
$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\} x_n^i = 0$$

=>
$$\sum_{n=1}^{N} y(x_n, \mathbf{w}) x_n^i = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{n=1}^{N} (\sum_{j=0}^{M} w_j x_n^j) x_n^i = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{n=1}^{N} \sum_{j=0}^{M} w_j x_n^{(j+i)} = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{j=0}^{M} \sum_{n=1}^{N} x_n^{(j+i)} w_j = \sum_{n=1}^{N} x_n^i t_n$$

If we denote $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$ and $T_i = \sum_{n=1}^{N} {x_n}^i t_n$, the equation above can be written exactly as (1.222), Therefore the problem is solved.

Problem 1.2 Solution

This problem is similar to Prob.1.1, and the only difference is the last term on the right side of (1.4), the penalty term. So we will do the same thing as in Prob.1.1 :

=>
$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\} x_n^i + \lambda w_i = 0$$

=>
$$\sum_{j=0}^{M} \sum_{n=1}^{N} x_n^{(j+i)} w_j + \lambda w_i = \sum_{n=1}^{N} x_n^i t_n$$

=>
$$\sum_{j=0}^{M} \{\sum_{n=1}^{N} x_n^{(j+i)} + \delta_{ji}\lambda\} w_j = \sum_{n=1}^{N} x_n^i t_n$$

where

$$\delta_{ji} \begin{cases} 0 & \text{j} \neq \text{i} \\ 1 & \text{j} = \text{i} \end{cases}$$

Problem 1.3 Solution

This problem can be solved by *Bayes' theorem*. The probability of selecting an apple $P(a)$:

$$P(a) = P(a|r)P(r)+P(a|b)P(b)+P(a|g)P(g) = \frac{3}{10}\times 0.2+\frac{1}{2}\times 0.2+\frac{3}{10}\times 0.6 = 0.34$$

Based on *Bayes' theorem*, the probability of an selected orange coming from the green box $P(g|o)$:

$$P(g|o) = \frac{P(o|g)P(g)}{P(o)}$$

We calculate the probability of selecting an orange $P(o)$ first :

$$P(o) = P(o|r)P(r)+P(o|b)P(b)+P(o|g)P(g) = \frac{4}{10}\times 0.2+\frac{1}{2}\times 0.2+\frac{3}{10}\times 0.6 = 0.36$$

Therefore we can get :

$$P(g|o) = \frac{P(o|g)P(g)}{P(o)} = \frac{\frac{3}{10}\times 0.6}{0.36} = 0.5$$

Problem 1.4 Solution

This problem needs knowledge about *calculus*, especially about *Chain rule*. We calculate the derivative of $P_y(y)$ with respect to y, according to (1.27) :

$$\frac{dp_y(y)}{dy} = \frac{d(p_x(g(y))|g^{`}(y)|)}{dy} = \frac{dp_x(g(y))}{dy}|g^{`}(y)| + p_x(g(y))\frac{d|g^{`}(y)|}{dy} \qquad (*)$$

The first term in the above equation can be further simplified:

$$\frac{dp_x(g(y))}{dy}|g^{`}(y)| = \frac{dp_x(g(y))}{dg(y)}\frac{dg(y)}{dy}|g^{`}(y)| \qquad (**)$$

If $\hat{x}$ is the maximum of density over x, we can obtain :

$$\frac{dp_x(x)}{dx}\Big|_{\hat{x}} = 0$$

Therefore, when $y = \hat{y}, s.t. \hat{x} = g(\hat{y})$, the first term on the right side of $(**)$ will be 0, leading the first term in $(*)$ equals to 0, however because of the existence of the second term in $(*)$, the derivative may not equal to 0. But

when linear transformation is applied, the second term in $(*)$ will vanish, (e.g. $x = ay + b$). A simple example can be shown by :

$$p_x(x) = 2x, \quad x \in [0,1] \quad => \quad \hat{x} = 1$$

And given that:

$$x = sin(y)$$

Therefore, $p_y(y) = 2\sin(y)|\cos(y)|$, $y \in [0, \frac{\pi}{2}]$, which can be simplified :

$$p_y(y) = \sin(2y), \quad y \in [0, \frac{\pi}{2}] \quad => \quad \hat{y} = \frac{\pi}{4}$$

However, it is quite obvious :

$$\hat{x} \neq \sin(\hat{y})$$

Problem 1.5 Solution

This problem takes advantage of the property of expectation:

$$\begin{aligned} var[f] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 \\ => var[f] &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \end{aligned}$$

Problem 1.6 Solution

Based on (1.41), we only need to prove when x and y is independent, $\mathbb{E}_{x,y}[xy] = \mathbb{E}[x]\mathbb{E}[y]$. Because x and y is independent, we have :

$$p(x,y) = p_x(x)\,p_y(y)$$

Therefore:

$$\begin{aligned} \int\int xyp(x,y)dxdy &= \int\int xyp_x(x)p_y(y)dxdy \\ &= (\int xp_x(x)dx)(\int yp_y(y)dy) \\ => \mathbb{E}_{x,y}[xy] &= \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Problem 1.7 Solution

This problem should take advantage of *Integration by substitution*.

$$\begin{aligned} I^2 &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} exp(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2)dxdy \\ &= \int_0^{2\pi}\int_0^{+\infty} exp(-\frac{1}{2\sigma^2}r^2)r\,dr\,d\theta \end{aligned}$$

Here we utilize :

$$x = r\cos\theta, \quad y = r\sin\theta$$

Based on the fact :

$$\int_0^{+\infty} exp(-\frac{1}{2\sigma^2}) r\,dr = -\sigma^2 exp(-\frac{r^2}{2\sigma^2})\Big|_0^{+\infty} = -\sigma^2(0-(-1)) = \sigma^2$$

Therefore, I can be solved :

$$I^2 = \int_0^{2\pi} \sigma^2\,d\theta = 2\pi\sigma^2, \quad => I = \sqrt{2\pi}\sigma$$

And next,we will show that Gaussian distribution $\mathcal{N}(x|\mu,\sigma^2)$ is normalized, (i.e. $\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)dx = 1$) :

$$\begin{aligned}
\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)dx &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}dx \\
&= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}dy \quad (y = x-\mu) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty}exp\{-\frac{1}{2\sigma^2}y^2\}dy \\
&= 1
\end{aligned}$$

Problem 1.8 Solution

The first question will need the result of Prob.1.7 :

$$\begin{aligned}
\int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)x\,dx &= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}x\,dx \\
&= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}(y+\mu)dy \quad (y = x-\mu) \\
&= \mu\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}dy + \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}exp\{-\frac{1}{2\sigma^2}y^2\}y\,dy \\
&= \mu + 0 = \mu
\end{aligned}$$

The second problem has already be given hint in the description. Given that :

$$\frac{d(fg)}{dx} = f\frac{dg}{dx} + g\frac{df}{dx}$$

We differentiate both side of (1.127) with respect to σ^2, we will obtain :

$$\int_{-\infty}^{+\infty}(-\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4})\mathcal{N}(x|\mu,\sigma^2)dx = 0$$

Provided the fact that $\sigma \neq 0$, we can get:

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

So the equation above has actually proven (1.51), according to the definition:

$$var[x] = \int_{-\infty}^{+\infty} (x-\mathbb{E}[x])^2 \mathcal{N}(x|\mu,\sigma^2)\,dx$$

Where $\mathbb{E}[x] = \mu$ has already been proved. Therefore :

$$var[x] = \sigma^2$$

Finally,

$$\mathbb{E}[x^2] = var[x] + \mathbb{E}[x]^2 = \sigma^2 + \mu^2$$

Problem 1.9 Solution

Here we only focus on (1.52), because (1.52) is the general form of (1.42). Based on the definition : The maximum of distribution is known as its mode and (1.52), we can obtain :

$$\begin{aligned} \frac{\partial \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \mathbf{x}} &= -\frac{1}{2}[\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1})^T](\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= -\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \end{aligned}$$

Where we take advantage of :

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}+\mathbf{A}^T)\mathbf{x} \quad \text{and} \quad (\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$$

Therefore,

$$\text{only when} \quad \mathbf{x} = \boldsymbol{\mu}, \ \frac{\partial \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \mathbf{x}} = 0$$

Note: You may also need to calculate *Hessian Matrix* to prove that it is maximum. However, here we find that the first derivative only has one root. Based on the description in the problem, this point should be maximum point.

Problem 1.10 Solution

We will solve this problem based on the definition of *expectation*, *variation*

and *independence*.

$$\begin{aligned}
\mathbb{E}[x+z] &= \int\int (x+z)p(x,z)\,dx\,dz \\
&= \int\int (x+z)p(x)p(z)\,dx\,dz \\
&= \int\int xp(x)p(z)\,dx\,dz + \int\int zp(x)p(z)\,dx\,dz \\
&= \int (\int p(z)\,dz)xp(x)\,dx + \int (\int p(x)\,dx)zp(z)\,dz \\
&= \int xp(x)\,dx + \int zp(z)\,dz \\
&= \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned}$$

$$\begin{aligned}
var[x+z] &= \int\int (x+z-\mathbb{E}[x+z])^2 p(x,z)\,dx\,dz \\
&= \int\int \{(x+z)^2 - 2(x+z)\mathbb{E}[x+z]) + \mathbb{E}^2[x+z]\}p(x,z)\,dx\,dz \\
&= \int\int (x+z)^2 p(x,z)\,dx\,dz - 2\mathbb{E}[x+z]\int\int (x+z)p(x,z)\,dx\,dz + \mathbb{E}^2[x+z] \\
&= \int\int (x+z)^2 p(x,z)\,dx\,dz - \mathbb{E}^2[x+z] \\
&= \int\int (x^2+2xz+z^2)p(x)p(z)\,dx\,dz - \mathbb{E}^2[x+z] \\
&= \int (\int p(z)dz)x^2 p(x)\,dx + \int\int 2xz\,p(x)p(z)\,dx\,dz + \int (\int p(x)dx)z^2 p(z)\,dz - \mathbb{E}^2[x+z] \\
&= \mathbb{E}[x^2]+\mathbb{E}[z^2]-\mathbb{E}^2[x+z]+\int\int 2xz\,p(x)p(z)\,dx\,dz \\
&= \mathbb{E}[x^2]+\mathbb{E}[z^2]-(\mathbb{E}[x]+\mathbb{E}[z])^2+\int\int 2xz\,p(x)p(z)\,dx\,dz \\
&= \mathbb{E}[x^2]-\mathbb{E}^2[x]+\mathbb{E}[z^2]-\mathbb{E}^2[z]-2\mathbb{E}[x]\mathbb{E}[z]+2\int\int xz\,p(x)p(z)\,dx\,dz \\
&= var[x]+var[z]-2\mathbb{E}[x]\mathbb{E}[z]+2(\int xp(x)\,dx)(\int zp(z)\,dz) \\
&= var[x]+var[z]
\end{aligned}$$

Problem 1.11 Solution

Based on prior knowledge that μ_{ML} and σ^2_{ML} will decouple. We will first calculate μ_{ML} :

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

We let :

$$\frac{d(ln\,p(\mathbf{x}|\mu,\sigma^2))}{d\mu} = 0$$

Therefore :

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

And because:

$$\frac{d(ln\, p(\mathbf{x}|\mu,\sigma^2))}{d\sigma^2} = \frac{1}{2\sigma^4}(\sum_{n=1}^{N}(x_n-\mu)^2 - N\sigma^2)$$

We let :

$$\frac{d(ln\, p(\mathbf{x}|\mu,\sigma^2))}{d\sigma^2} = 0$$

Therefore :

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

Problem 1.12 Solution

It is quite straightforward for $\mathbb{E}[\mu_{ML}]$, with the prior knowledge that x_n is i.i.d. and it also obeys Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$.

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N} x_n] = \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N} x_n] = \mathbb{E}[x_n] = \mu$$

For $\mathbb{E}[\sigma^2_{ML}]$, we need to take advantage of (1.56) and what has been given in the problem :

$$\begin{aligned}
\mathbb{E}[\sigma^2_{ML}] &= \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n-\mu_{ML})^2] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n^2 - 2x_n\mu_{ML}+\mu^2_{ML})] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}x_n^2] - \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}2x_n\mu_{ML}] + \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}\mu^2_{ML}] \\
&= \mu^2+\sigma^2 - \frac{2}{N}\mathbb{E}[\sum_{n=1}^{N}x_n(\frac{1}{N}\sum_{n=1}^{N}x_n)] + \mathbb{E}[\mu^2_{ML}] \\
&= \mu^2+\sigma^2 - \frac{2}{N^2}\mathbb{E}[\sum_{n=1}^{N}x_n(\sum_{n=1}^{N}x_n)] + \mathbb{E}[(\frac{1}{N}\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2+\sigma^2 - \frac{2}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] + \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2+\sigma^2 - \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] \\
&= \mu^2+\sigma^2 - \frac{1}{N^2}[N(N\mu^2+\sigma^2)]
\end{aligned}$$

Therefore we have:

$$\mathbb{E}[\sigma^2_{ML}] = (\frac{N-1}{N})\sigma^2$$

Problem 1.13 Solution

This problem can be solved in the same method used in Prob.1.12 :

$$\begin{aligned}
\mathbb{E}[\sigma^2_{ML}] &= \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2] \quad \text{(Because here we use } \mu \text{ to replace } \mu_{ML}) \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n-\mu)^2] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}(x_n^2-2x_n\mu+\mu^2)] \\
&= \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}x_n^2] - \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}2x_n\mu] + \frac{1}{N}\mathbb{E}[\sum_{n=1}^{N}\mu^2] \\
&= \mu^2+\sigma^2-\frac{2\mu}{N}\mathbb{E}[\sum_{n=1}^{N}x_n]+\mu^2 \\
&= \mu^2+\sigma^2-2\mu^2+\mu^2 \\
&= \sigma^2
\end{aligned}$$

Note: The biggest difference between Prob.1.12 and Prob.1.13 is that the mean of Gaussian Distribution is known previously (in Prob.1.13) or not (in Prob.1.12). In other words, the difference can be shown by the following equations:

$$\mathbb{E}[\mu^2] = \mu^2 \quad (\mu \text{ is determined, i.e. its } \textit{expectation} \text{ is itself, also true for } \mu^2)$$

$$\mathbb{E}[\mu^2_{ML}] = \mathbb{E}[(\frac{1}{N}\sum_{n=1}^{N}x_n)^2] = \frac{1}{N^2}\mathbb{E}[(\sum_{n=1}^{N}x_n)^2] = \frac{1}{N^2}N(N\mu^2+\sigma^2) = \mu^2+\frac{\sigma^2}{N}$$

Problem 1.14 Solution

This problem is quite similar to the fact that *any function $f(x)$ can be written into the sum of an odd function and an even function*. If we let:

$$w^S_{ij} = \frac{w_{ij}+w_{ji}}{2} \quad \text{and} \quad w^A_{ij} = \frac{w_{ij}-w_{ji}}{2}$$

It is obvious that they satisfy the constraints described in the problem, which are :

$$w_{ij} = w^S_{ij} + w^A_{ij}, \quad w^S_{ij} = w^S_{ji}, \quad w^A_{ij} = -w^A_{ji}$$

To prove (1.132), we only need to simplify it :

$$\begin{aligned}\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^S + w_{ij}^A)x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j\end{aligned}$$

Therefore, we only need to prove that the second term equals to 0, and here we use a simple trick: we will prove twice of the second term equals to 0 instead.

$$\begin{aligned}2\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^A + w_{ij}^A)x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D}(w_{ij}^A - w_{ji}^A)x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j - \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ji}^A x_i x_j \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^A x_i x_j - \sum_{j=1}^{D}\sum_{i=1}^{D} w_{ji}^A x_j x_i \\ &= 0\end{aligned}$$

Therefore, we choose the coefficient matrix to be symmetric as described in the problem. Considering about the symmetry, we can see that if and only if for $i = 1, 2, ..., D$ and $i \leq j$, w_{ij} is given, the whole matrix will be determined. Hence, the number of independent parameters are given by :

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

Note: You can view this intuitively by considering if the upper triangular part of a symmetric matrix is given, the whole matrix will be determined.

Problem 1.15 Solution

This problem is a more general form of Prob.1.14, so the method can also be used here: we will find a way to use $w_{i_1 i_2 ... i_M}$ to represent $\widetilde{w}_{i_1 i_2 ... i_M}$.

We begin by introducing a mapping function:

$$F(x_{i1}x_{i2}...x_{iM}) = x_{j1}x_{j2}..., x_{jM}$$

$$s.t. \bigcup_{k=1}^{M} x_{ik} = \bigcup_{k=1}^{M} x_{jk}, \quad \text{and} \quad x_{j1} \geq x_{j2} \geq x_{j3}... \geq x_{jM}$$

It is complexed to write F in mathematical form. Actually this function does a simple work: it rearranges the element in a decreasing order based on its subindex. Several examples are given below, when $D = 5$, $M = 4$:

$$F(x_5x_2x_3x_2) = x_5x_3x_2x_2$$

$$F(x_1x_3x_3x_2) = x_3x_3x_2x_1$$

$$F(x_1x_4x_2x_3) = x_4x_3x_2x_1$$

$$F(x_1x_1x_5x_2) = x_5x_2x_1x_1$$

After introducing F, the solution will be very simple, based on the fact that F will not change the value of the term, but only rearrange it.

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1i_2\ldots i_M}x_{i1}x_{i2}\ldots x_{iM} = \sum_{j_1=1}^{D}\sum_{j_2=1}^{j_1}\cdots\sum_{j_M=1}^{j_{M-1}} \widetilde{w}_{j_1j_2\ldots j_M}x_{j1}x_{j2}\ldots x_{jM}$$

where $$\widetilde{w}_{j_1j_2\ldots j_M} = \sum_{w\in\Omega} w$$

$$\Omega = \{w_{i_1i_2\ldots i_M} \,\big|\, F(x_{i1}x_{i2}\ldots x_{iM}) = x_{j1}x_{j2}\ldots x_{jM},\ \forall x_{i1}x_{i2}\ldots x_{iM}\}$$

By far, we have already proven (1.134). *Mathematical induction* will be used to prove (1.135) and we will begin by proving $D = 1$, i.e. $n(1,M) = n(1,M-1)$. When $D = 1$, (1.134) will degenerate into $\widetilde{w}x_1^M$, i.e., it only has one term, whose coefficient is govern by $\widetilde{w}$ regardless the value of M.

Therefore, we have proven when $D = 1$, $n(D,M) = 1$. Suppose (1.135) holds for D, let's prove it will also hold for $D+1$, and then (1.135) will be proved based on *Mathematical induction*.

Let's begin based on (1.134):

$$\sum_{i_1=1}^{D+1}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}} \widetilde{w}_{i_1i_2\ldots i_M}x_{i1}x_{i2}\ldots x_{iM} \qquad (*)$$

We divide $(*)$ into two parts based on the first summation: the first part is made up of $i_i = 1,2,\ldots,D$ and the second part $i_1 = D+1$. After division, the first part corresponds to $n(D,M)$, and the second part corresponds to $n(D+1,M-1)$. Therefore we obtain:

$$n(D+1,M) = n(D,M) + n(D+1,M-1) \qquad (**)$$

And given the fact that (1.135) holds for D :

$$n(D,M) = \sum_{i=1}^{D} n(i,M-1)$$

Therefore,we substitute it into $(**)$

$$n(D+1,M) = \sum_{i=1}^{D} n(i,M-1) + n(D+1,M-1) = \sum_{i=1}^{D+1} n(i,M-1)$$

We will prove (1.136) in a different but simple way. We rewrite (1.136) in *Permutation and Combination* view:

$$\sum_{i=1}^{D} C_{i+M-2}^{M-1} = C_{D+M-1}^{M}$$

Firstly, We expand the summation.

$$C_{M-1}^{M-1} + C_{M}^{M-1} + ... C_{D+M-2}^{M-1} = C_{D+M-1}^{M}$$

Secondly, we rewrite the first term on the left side to C_M^M, because $C_{M-1}^{M-1} = C_M^M = 1$. In other words, we only need to prove:

$$C_{M}^{M} + C_{M}^{M-1} + ... C_{D+M-2}^{M-1} = C_{D+M-1}^{M}$$

Thirdly, we take advantage of the property : $C_N^r = C_{N-1}^r + C_{N-1}^{r-1}$. So we can recursively combine the first term and the second term on the left side, and it will ultimately equal to the right side.

(1.137) gives the mathematical form of $n(D,M)$, and we need all the conclusions above to prove it.

Let's give some intuitive concepts by illustrating $M = 0,1,2$. When $M = 0$, (1.134) will consist of only a constant term, which means $n(D,0) = 1$. When $M = 1$,it is obvious $n(D,1) = D$, because in this case (1.134) will only have D terms if we expand it. When $M = 2$, it degenerates to Prob.1.14, so $n(D,2) = \frac{D(D+1)}{2}$ is also obvious. Suppose (1.137) holds for $M-1$, let's prove it will also hold for M.

$$\begin{aligned}
n(D,M) &= \sum_{i=1}^{D} n(i,M-1) \quad \text{(based on (1.135))} \\
&= \sum_{i=1}^{D} C_{i+M-2}^{M-1} \quad \text{(based on (1.137) holds for } M-1 \text{)} \\
&= C_{M-1}^{M-1} + C_{M}^{M-1} + C_{M+1}^{M-1} ... + C_{D+M-2}^{M-1} \\
&= (C_{M}^{M} + C_{M}^{M-1}) + C_{M+1}^{M-1} ... + C_{D+M-2}^{M-1} \\
&= (C_{M+1}^{M} + C_{M+1}^{M-1}) ... + C_{D+M-2}^{M-1} \\
&= C_{M+2}^{M} ... + C_{D+M-2}^{M-1} \\
&\quad ... \\
&= C_{D+M-1}^{M}
\end{aligned}$$

By far, all have been proven.

Problem 1.16 Solution

This problem can be solved in the same way as the one in Prob.1.15. Firstly, we should write the expression consisted of all the independent terms up to Mth order corresponding to $N(D,M)$. By adding a summation regarding to M on the left side of (1.134), we obtain:

$$\sum_{m=0}^{M}\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_m=1}^{i_{m-1}}\widetilde{w}_{i_1 i_2 \ldots i_m}x_{i1}x_{i2}\ldots x_{im} \qquad (*)$$

(1.138) is quite obvious if we view m as an looping variable, iterating through all the possible orders less equal than M, and for every possible oder m, the independent parameters are given by $n(D,m)$.

Let's prove (1.138) in a formal way by using *Mathematical Induction*. When $M = 1$,$(*)$ will degenerate to two terms: $m = 0$, corresponding to $n(D,0)$ and $m = 1$, corresponding to $n(D,1)$. Therefore $N(D,1) = n(D,0) + n(D,1)$. Suppose (1.138) holds for M, we will see that it will also hold for $M+1$. Let's begin by writing all the independent terms based on $(*)$:

$$\sum_{m=0}^{M+1}\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_m=1}^{i_{m-1}}\widetilde{w}_{i_1 i_2 \ldots i_m}x_{i1}x_{i2}\ldots x_{im} \qquad (**)$$

Using the same technique as in Prob.1.15, we divide $(**)$ to two parts based on the summation regarding to m: the first part consisted of $m = 0,1,\ldots,M$ and the second part $m = M+1$. Hence, the first part will correspond to $N(D,M)$ and the second part will correspond to $n(D,M+1)$. So we obtain:

$$N(D,M+1) = N(D,M) + n(D,M+1)$$

Then we substitute (1.138) into the equation above :

$$\begin{aligned} N(D,M+1) &= \sum_{m=0}^{M} n(D,m) + n(D,M+1) \\ &= \sum_{m=0}^{M+1} n(D,m) \end{aligned}$$

To prove (1.139), we will also use the same technique in Prob.1.15 instead of *Mathematical Induction*. We begin based on already proved (1.138):

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

We then take advantage of (1.137):

$$\begin{aligned} N(D,M) &= \sum_{m=0}^{M} C_{D+m-1}^{m} \\ &= C_{D-1}^{0} + C_{D}^{1} + C_{D+1}^{2} + ... + C_{D+M-1}^{M} \\ &= (C_{D}^{0} + C_{D}^{1}) + C_{D+1}^{2} + ... + C_{D+M-1}^{M} \\ &= (C_{D+1}^{1} + C_{D+1}^{2}) + ... + C_{D+M-1}^{M} \\ &= ... \\ &= C_{D+M}^{M} \end{aligned}$$

Here as asked by the problem, we will view the growing speed of $N(D,M)$. We should see that in $n(D,M)$, D and M are symmetric, meaning that we only need to prove when $D \gg M$, it will grow like D^M, and then the situation of $M \gg D$ will be solved by symmetry.

$$\begin{aligned} N(D,M) &= \frac{(D+M)!}{D!M!} \approx \frac{(D+M)^{D+M}}{D^D M^M} \\ &= \frac{1}{M^M}(\frac{D+M}{D})^D (D+M)^M \\ &= \frac{1}{M^M}[(1+\frac{M}{D})^{\frac{D}{M}}]^M (D+M)^M \\ &\approx (\frac{e}{M})^M (D+M)^M \\ &= \frac{e^M}{M^M}(1+\frac{M}{D})^M D^M \\ &= \frac{e^M}{M^M}[(1+\frac{M}{D})^{\frac{D}{M}}]^{\frac{M^2}{D}} D^M \\ &\approx \frac{e^{M+\frac{M^2}{D}}}{M^M} D^M \approx \frac{e^M}{M^M} D^M \end{aligned}$$

Where we use *Stirling's approximation*, $\lim_{n\to+\infty}(1+\frac{1}{n})^n = e$ and $e^{\frac{M^2}{D}} \approx e^0 = 1$. According to the description in the problem, When $D \gg M$, we can actually view $\frac{e^M}{M^M}$ as a constant, so $N(D,M)$ will grow like D^M in this case. And by symmetry, $N(D,M)$ will grow like M^D, when $M \gg D$.

Finally, we are asked to calculate $N(10,3)$ and $N(100,3)$:

$$N(10,3) = C_{13}^{3} = 286$$

$$N(100,3) = C_{103}^{3} = 176851$$

Problem 1.17 Solution

$$\begin{aligned}\Gamma(x+1) &= \int_0^{+\infty} u^x e^{-u} du \\ &= \int_0^{+\infty} -u^x de^{-u} \\ &= -u^x e^{-u}\Big|_0^{+\infty} - \int_0^{+\infty} e^{-u} d(-u^x) \\ &= -u^x e^{-u}\Big|_0^{+\infty} + x\int_0^{+\infty} e^{-u} u^{x-1} du \\ &= -u^x e^{-u}\Big|_0^{+\infty} + x\Gamma(x)\end{aligned}$$

Where we have taken advantage of *Integration by parts* and according to the equation above, we only need to prove the first term equals to 0. Given *L'Hospital's Rule*:

$$\lim_{u\to+\infty} -\frac{u^x}{e^u} = \lim_{u\to+\infty} -\frac{x!}{e^u} = 0$$

And also when $u = 0, -u^x e^u = 0$, so we have proved $\Gamma(x+1) = x\Gamma(x)$. Based on the definition of $\Gamma(x)$, we can write:

$$\Gamma(1) = \int_0^{+\infty} e^{-u} du = -e^{-u}\Big|_0^{+\infty} = -(0-1) = 1$$

Therefore when x is an integer:

$$\Gamma(x) = (x-1)\Gamma(x-1) = (x-1)(x-2)\Gamma(x-2) = \ldots = x!\,\Gamma(1) = x!$$

Problem 1.18 Solution

Based on (1.124) and (1.126) and by substituting x to $\sqrt{2}\sigma y$, it is quite obvious to obtain :

$$\int_{-\infty}^{+\infty} e^{-x_i^2} dx_i = \sqrt{\pi}$$

Therefore, the left side of (1.42) will equal to $\pi^{\frac{D}{2}}$. For the right side of (1.42):

$$\begin{aligned}S_D \int_0^{+\infty} e^{-r^2} r^{D-1} dr &= S_D \int_0^{+\infty} e^{-u} u^{\frac{D-1}{2}} d\sqrt{u} \quad (u = r^2) \\ &= \frac{S_D}{2}\int_0^{+\infty} e^{-u} u^{\frac{D}{2}-1} du \\ &= \frac{S_D}{2}\Gamma(\frac{D}{2})\end{aligned}$$

Hence, we obtain:

$$\pi^{\frac{D}{2}} = \frac{S_D}{2}\Gamma(\frac{D}{2}) \quad => \quad S_D = \frac{2\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2})}$$

S_D has given the expression of the surface area with radius 1 in dimension D, we can further expand the conclusion: the surface area with radius r in dimension D will equal to $S_D \cdot r^{D-1}$, and when $r = 1$, it will reduce to S_D. This conclusion is naive, if you find that the surface area of different sphere in dimension D is proportion to the $D-1$th power of radius, i.e. r^{D-1}. Considering the relationship between V and S of a sphere with arbitrary radius in dimension D: $\frac{dV}{dr} = S$, we can obtain :

$$V = \int S\,dr = \int S_D r^{D-1}\,dr = \frac{S_D}{D} r^D$$

The equation above gives the expression of the volume of a sphere with radius r in dimension D, so we let $r = 1$:

$$V_D = \frac{S_D}{D}$$

For $D = 2$ and $D = 3$:

$$V_2 = \frac{S_2}{2} = \frac{1}{2} \cdot \frac{2\pi}{\Gamma(1)} = \pi$$

$$V_3 = \frac{S_3}{3} = \frac{1}{3} \cdot \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} = \frac{1}{3} \cdot \frac{2\pi^{\frac{3}{2}}}{\frac{\sqrt{\pi}}{2}} = \frac{4}{3}\pi$$

Problem 1.19 Solution

We have already given a hint in the solution of Prob.1.18, and here we will make it more clearly: the volume of a sphere with radius r is $V_D \cdot r^D$. This is quite similar with the conclusion we obtained in Prob.1.18 about the surface area except that it is proportion to Dth power of its radius, i.e. r^D not r^{D-1}.

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{V_D a^D}{(2a)^D} = \frac{S_D}{2^D D} = \frac{\pi^{\frac{D}{2}}}{2^{D-1} D\,\Gamma(\frac{D}{2})} \qquad (*)$$

Where we have used the result of (1.143). And when $D \to +\infty$, we will use a simple method to show that $(*)$ will converge to 0. We rewrite it :

$$(*) = \frac{2}{D} \cdot (\frac{\pi}{4})^{\frac{D}{2}} \cdot \frac{1}{\Gamma(\frac{D}{2})}$$

Hence, it is now quite obvious, all the three terms will converge to 0 when $D \to +\infty$. Therefore their product will also converge to 0. The last problem is quite simple :

$$\frac{\text{center to one corner}}{\text{center to one side}} = \frac{\sqrt{a^2 \cdot D}}{a} = \sqrt{D} \quad \text{and} \quad \lim_{D \to +\infty} \sqrt{D} = +\infty$$

Problem 1.20 Solution

The density of probability in a thin shell with radius r and thickness ϵ can be viewed as a constant. And considering that a sphere in dimension D with radius r has surface area $S_D r^{D-1}$, which has already been proved in Prob.1.19 :

$$\int_{shell} p(\mathbf{x})d\mathbf{x} = p(\mathbf{x})\int_{shell} d\mathbf{x} = \frac{exp(-\frac{r^2}{2\sigma^2})}{(2\pi\sigma^2)^{\frac{D}{2}}} \cdot V(\text{shell}) = \frac{exp(-\frac{r^2}{2\sigma^2})}{(2\pi\sigma^2)^{\frac{D}{2}}} S_D r^{D-1}\epsilon$$

Thus we denote :

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{\frac{D}{2}}} exp(-\frac{r^2}{2\sigma^2})$$

We calculate the derivative of (1.148) with respect to r :

$$\frac{dp(r)}{dr} = \frac{S_D}{(2\pi\sigma^2)^{\frac{D}{2}}} r^{D-2} exp(-\frac{r^2}{2\sigma^2})(D-1-\frac{r^2}{\sigma^2}) \qquad (*)$$

We let the derivative equal to 0, we will obtain its unique root(stationary point) $\hat{r} = \sqrt{D-1}\sigma$, because $r \in [0, +\infty]$. When $r < \hat{r}$, the derivative is large than 0, $p(r)$ will increase as $r \uparrow$, and when $r > \hat{r}$, the derivative is less than 0, $p(r)$ will decrease as $r \uparrow$. Therefore $\hat{r}$ will be the only maximum point. And it is obvious when $D \gg 1$, $\hat{r} \approx \sqrt{D}\sigma$.

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

We process for the exponential term by using *Taylor Theorems*.

$$\begin{aligned}
-\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)ln(1+\frac{\epsilon}{\hat{r}}) &\approx -\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + (D-1)(\frac{\epsilon}{\hat{r}} - \frac{\epsilon^2}{2\hat{r}^2}) \\
&= -\frac{2\epsilon\hat{r}+\epsilon^2}{2\sigma^2} + \frac{2\hat{r}\epsilon-\epsilon^2}{2\sigma^2} \\
&= -\frac{\epsilon^2}{\sigma^2}
\end{aligned}$$

Therefore, $p(\hat{r}+\epsilon) = p(\hat{r})exp(-\frac{\epsilon^2}{\sigma^2})$. **Note: Here I draw a different conclusion compared with (1.149),** but I do not think there is any mistake in my deduction.

Finally, we see from (1.147) :

$$p(\mathbf{x})\Big|_{\mathbf{x}=0} = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}}$$

$$p(\mathbf{x})\Big|_{||\mathbf{x}||^2=\hat{r}^2} = \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} exp(-\frac{\hat{r}^2}{2\sigma^2}) \approx \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}} exp(-\frac{D}{2})$$

Problem 1.21 Solution

The first question is rather simple :

$$(ab)^{\frac{1}{2}} - a = a^{\frac{1}{2}}(b^{\frac{1}{2}} - a^{\frac{1}{2}}) \geq 0$$

Where we have taken advantage of $b \geq a \geq 0$. And based on (1.78):

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in R_1, C_2) + p(\mathbf{x} \in R_2, C_1) \\ &= \int_{R_1} p(\mathbf{x}, C_2)dx + \int_{R_2} p(\mathbf{x}, C_1)dx \end{aligned}$$

Recall that the decision rule which can minimize misclassification is that if $p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2)$, for a given value of $\mathbf{x}$, we will assign that $\mathbf{x}$ to class C_1. We can see that in decision area R_1, it should satisfy $p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2)$. Therefore, using what we have proved, we can obtain :

$$\int_{R_1} p(\mathbf{x}, C_2)dx \leq \int_{R_1} \{p(\mathbf{x}, C_1)p(\mathbf{x}, C_2)\}^{\frac{1}{2}} dx$$

It is the same for decision area R_2. Therefore we can obtain:

$$p(\text{mistake}) \leq \int \{p(\mathbf{x}, C_1)p(\mathbf{x}, C_2)\}^{\frac{1}{2}} dx$$

Problem 1.22 Solution

We need to deeply understand (1.81). When $L_{kj} = 1 - I_{kj}$:

$$\sum_k L_{kj} p(C_k|\mathbf{x}) = \sum_k p(C_k|\mathbf{x}) - p(C_j|\mathbf{x})$$

Given a specific $\mathbf{x}$, the first term on the right side is a constant, which equals to 1, no matter which class C_j we assign $\mathbf{x}$ to. Therefore if we want to minimize the loss, we will maximize $p(C_j|\mathbf{x})$. Hence, we will assign $\mathbf{x}$ to class C_j, which can give the biggest posterior probability $p(C_j|\mathbf{x})$.

The explanation of the loss matrix is quite simple. If we label correctly, there is no loss. Otherwise, we will incur a loss, in the same degree whichever class we label it to. The loss matrix is given below to give you an intuitive view:

$$\begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 0 \end{bmatrix}$$

Problem 1.23 Solution

$$\mathbb{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj} p(\mathbf{x}, C_k)\, d\mathbf{x} = \sum_k \sum_j \int_{R_j} L_{kj} p(C_k) p(\mathbf{x}|C_k)\, d\mathbf{x}$$

If we denote a new loss matrix by $L^{\star}_{jk} = L_{jk} p(C_k)$, we can obtain a new equation :

$$\mathbb{E}[L] = \sum_k \sum_j \int_{R_j} L^{\star}_{kj} p(\mathbf{x}|C_k)\, d\mathbf{x}$$

Problem 1.24 Solution

This description of the problem is a little confusing, and what it really mean is that λ is the parameter governing the loss, just like θ governing the posterior probability $p(C_k|\mathbf{x})$ when we introduce the reject option. Therefore the reject option can be written in a new way when we view it from the view of λ and the loss:

$$\text{choice} \begin{cases} \text{class } C_j & \min\limits_l \sum_k L_{kl} p(C_k|x) < \lambda \\ \text{reject} & \text{else} \end{cases}$$

Where C_j is the class that can obtain the minimum. If $L_{kj} = 1 - I_{kj}$, according to what we have proved in Prob.1.22 :

$$\sum_k L_{kj} p(C_k|\mathbf{x}) = \sum_k p(C_k|\mathbf{x}) - p(C_j|\mathbf{x}) = 1 - p(C_j|\mathbf{x})$$

Therefore, the reject criterion from the view of λ above is actually equivalent to the largest posterior probability is larger than $1 - \lambda$:

$$\min_l \sum_k L_{kl} p(C_k|x) < \lambda \quad <=> \quad \max_l p(C_l|x) > 1 - \lambda$$

And from the view of θ and posterior probability, we label a class for $\mathbf{x}$ (i.e. we do not reject) is given by the constrain :

$$\max_l p(C_l|x) > \theta$$

Hence from the two different views, we can see that λ and θ are correlated with:

$$\lambda + \theta = 1$$

Problem 1.25 Solution

We can prove this informally by dealing with one dimension once a time just as the same process in (1.87) - (1.89) until all has been done, due to the fact that the total loss E can be divided to the summation of loss on every

dimension, and what's more they are independent. Here, we will use a more informal way to prove this. In this case, the expected loss can be written :

$$\mathbb{E}[L] = \int\int \{\mathbf{y}(\mathbf{x}) - \mathbf{t}\}^2 p(\mathbf{x},\mathbf{t})\, d\mathbf{t}\, d\mathbf{x}$$

Therefore, just as the same process in (1.87) - (1.89):

$$\frac{\partial \mathbb{E}[L]}{\partial y(\mathbf{x})} = 2\int \{\mathbf{y}(\mathbf{x}) - \mathbf{t}\} p(\mathbf{x},\mathbf{t})\, d\mathbf{t} = \mathbf{0}$$

$$=> \quad \mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t}\, p(\mathbf{x},\mathbf{t})\, d\mathbf{t}}{p(\mathbf{x})} = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$$

Problem 1.26 Solution

The process is identical as the deduction we conduct for (1.90). We will not repeat here. And what we should emphasize is that $\mathbb{E}[\mathbf{t}|\mathbf{x}]$ is a function of $\mathbf{x}$, not $\mathbf{t}$. Thus the integral over $\mathbf{t}$ and $\mathbf{x}$ can be simplified based on *Integration by parts* and that is how we obtain (1.90).

]**Note**: There is a mistake in (1.90), i.e. the second term on the right side is wrong. You can view (3.37) on P148 for reference. It should be :

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, d\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x} - t]\}^2 p(\mathbf{x},t)\, d\mathbf{x}\, dt$$

Problem 1.27 Solution

We deal with this problem based on *Calculus of Variations*.

$$\frac{\partial \mathbb{E}[L_q]}{\partial y(\mathbf{x})} = q\int [y(\mathbf{x}) - t]^{q-1} sign(y(\mathbf{x}) - t)\, p(\mathbf{x},t)\, dt = 0$$

$$=> \quad \int_{-\infty}^{y(\mathbf{x})} [y(\mathbf{x}) - t]^{q-1} p(\mathbf{x},t)\, dt = \int_{y(\mathbf{x})}^{+\infty} [y(\mathbf{x}) - t]^{q-1} p(\mathbf{x},t)\, dt$$

$$=> \quad \int_{-\infty}^{y(\mathbf{x})} [y(\mathbf{x}) - t]^{q-1} p(t|\mathbf{x})\, dt = \int_{y(\mathbf{x})}^{+\infty} [y(\mathbf{x}) - t]^{q-1} p(t|\mathbf{x})\, dt$$

Where we take advantage of $p(\mathbf{x},t) = p(t|\mathbf{x})p(\mathbf{x})$ and the property of *sign function*. Hence, when $q = 1$, the equation above will reduce to :

$$\int_{-\infty}^{y(\mathbf{x})} p(t|\mathbf{x})\, dt = \int_{y(\mathbf{x})}^{+\infty} p(t|\mathbf{x})\, dt$$

In other words, when $q = 1$, the optimal $y(\mathbf{x})$ will be given by conditional median. When $q = 0$, it is non-trivial. We need to rewrite (1.91) :

$$\begin{aligned} \mathbb{E}[L_q] &= \int \left\{ \int |y(\mathbf{x}) - t|^q\, p(t|\mathbf{x}) p(\mathbf{x})\, dt \right\} d\mathbf{x} \\ &= \int \left\{ p(\mathbf{x}) \int |y(\mathbf{x}) - t|^q\, p(t|\mathbf{x})\, dt \right\} d\mathbf{x} \quad (*) \end{aligned}$$

If we want to minimize $\mathbb{E}[L_q]$, we only need to minimize the integrand of $(*)$:

$$\int |y(\mathbf{x})-t|^q \, p(t|\mathbf{x})\,dt \qquad (**)$$

When $q=0$, $|y(\mathbf{x})-t|^q$ is close to 1 everywhere except in the neighborhood around $t=y(\mathbf{x})$ (This can be seen from Fig1.29). Therefore:

$$(**) \approx \int_{\mathscr{U}} p(t|\mathbf{x})\,dt - \int_{\epsilon}(1-|y(\mathbf{x})-t|^q)\,p(t|\mathbf{x})\,dt \approx \int_{\mathscr{U}} p(t|\mathbf{x})\,dt - \int_{\epsilon} p(t|\mathbf{x})\,dt$$

Where ϵ means the small neighborhood,$\mathscr{U}$ means the whole space $\mathbf{x}$ lies in. Note that $y(\mathbf{x})$ has no correlation with the first term, but the second term (because how to choose $y(\mathbf{x})$ will affect the location of ϵ). Hence we will put ϵ at the location where $p(t|\mathbf{x})$ achieve its largest value, i.e. the mode, because in this way we can obtain the largest reduction. Therefore, it is natural we choose $y(\mathbf{x})$ equals to t that maximize $p(t|\mathbf{x})$ for every $\mathbf{x}$.

Problem 1.28 Solution

Basically this problem is focused on the definition of *Information Content*, i.e.$h(x)$. We will rewrite the problem more precisely. In *Information Theory*, $h(\cdot)$ is also called *Information Content* and denoted as $I(\cdot)$. Here we will still use $h(\cdot)$ for consistency. The whole problem is about the property of $h(x)$. Based on our knowledge that $h(\cdot)$ is a monotonic function of the probability $p(x)$, we can obtain:

$$h(x) = f(p(x))$$

The equation above means that the *Information* we obtain for a specific value of a random variable x is correlated with its occurring probability $p(x)$, and its relationship is given by a mapping function $f(\cdot)$. Suppose C is the intersection of two independent event A and B, then the information of event C occurring is the compound message of both independent events A and B occurring:

$$h(C) = h(A \cap B) = h(A) + h(B) \qquad (*)$$

Because A and B is independent:

$$P(C) = P(A) \cdot P(B)$$

We apply function $f(\cdot)$ to both side:

$$f(P(C)) = f(P(A) \cdot P(B)) \qquad (**)$$

Moreover, the left side of $(*)$ and $(**)$ are equivalent by definition, so we can obtain:

$$h(A) + h(B) = f(P(A) \cdot P(B))$$

$$=> \quad f(p(A)) + f(p(B)) = f(P(A) \cdot P(B))$$

We obtain an important property of function $f(\cdot)$: $f(x \cdot y) = f(x) + f(y)$. Note: In problem (1.28), what it really wants us to prove is about the form and property of function f in our formulation, because there is one sentence in the description of the problem : "In this exercise, we derive the relation between h and p in the form of a function $h(p)$", (i.e. $f(\cdot)$ in our formulation is equivalent to $h(p)$ in the description).

At present, what we know is the property of function $f(\cdot)$:

$$f(xy) = f(x) + f(y) \qquad (*)$$

Firstly, we choose $x = y$, and then it is obvious : $f(x^2) = 2f(x)$. Secondly, it is obvious $f(x^n) = nf(x)$, $n \in \mathbb{N}$ is true for $n = 1$, $n = 2$. Suppose it is also true for n, we will prove it is true for $n + 1$:

$$f(x^{n+1}) = f(x^n) + f(x) = nf(x) + f(x) = (n+1)f(x)$$

Therefore, $f(x^n) = nf(x)$, $n \in \mathbb{N}$ has been proved. For an integer m, we rewrite x^n as $(x^{\frac{n}{m}})^m$, and take advantage of what we have proved, we will obtain:

$$f(x^n) = f((x^{\frac{n}{m}})^m) = mf(x^{\frac{n}{m}})$$

Because $f(x^n)$ also equals to $nf(x)$, therefore $nf(x) = mf(x^{\frac{n}{m}})$. We simplify the equation and obtain:

$$f(x^{\frac{n}{m}}) = \frac{n}{m} f(x)$$

For an arbitrary positive x , $x \in \mathbb{R}^+$, we can find two positive rational array $\{y_n\}$ and $\{z_n\}$, which satisfy:

$$y_1 < y_2 < ... < y_N < x \quad \text{and} \quad \lim_{N \to +\infty} y_N = x$$

$$z_1 > z_2 > ... > z_N > x, \quad \text{and} \quad \lim_{N \to +\infty} z_N = x$$

We take advantage of function $f(\cdot)$ is monotonic:

$$y_N f(p) = f(p^{y_N}) \le f(p^x) \le f(p^{z_N}) = z_N f(p)$$

And when $N \to +\infty$, we will obtain: $f(p^x) = xf(p)$, $x \in \mathbb{R}^+$. We let $p = e$, it can be rewritten as : $f(e^x) = xf(e)$. Finally, We denote $y = e^x$:

$$f(y) = ln(y)f(e)$$

Where $f(e)$ is a constant once function $f(\cdot)$ is decided. Therefore $f(x) \propto ln(x)$.

Problem 1.29 Solution

This problem is a little bit tricky. The entropy for a M-state discrete random variable x can be written as :

$$H[x] = -\sum_{i}^{M} \lambda_i ln(\lambda_i)$$

Where λ_i is the probability that x choose state i. Here we choose a concave function $f(\cdot) = ln(\cdot)$, we rewrite *Jensen's inequality*, i.e.(1.115):

$$ln(\sum_{i=1}^{M} \lambda_i x_i) \geq \sum_{i=1}^{M} \lambda_i ln(x_i)$$

We choose $x_i = \frac{1}{\lambda_i}$ and simplify the equation above, we will obtain :

$$lnM \geq -\sum_{i=1}^{M} \lambda_i ln(\lambda_i) = H[x]$$

Problem 1.30 Solution

Based on definition :

$$\begin{aligned} ln\{\frac{p(x)}{q(x)}\} &= ln(\frac{s}{\sigma}) - [\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2s^2}(x-m)^2] \\ &= ln(\frac{s}{\sigma}) - [(\frac{1}{2\sigma^2} - \frac{1}{2s^2})x^2 - (\frac{\mu}{\sigma^2} - \frac{m}{s^2})x + (\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2})] \end{aligned}$$

We will take advantage of the following equations to solve this problem.

$$\mathbb{E}[x^2] = \int x^2 \mathcal{N}(x|\mu,\sigma^2)dx = \mu^2 + \sigma^2$$

$$\mathbb{E}[x] = \int x \mathcal{N}(x|\mu,\sigma^2)dx = \mu$$

$$\int \mathcal{N}(x|\mu,\sigma^2)dx = 1$$

Given the equations above, it is easy to see :

$$\begin{aligned} KL(p||q) &= -\int p(x) ln\{\frac{q(x)}{p(x)}\}dx \\ &= \int \mathcal{N}(x|\mu,\sigma) ln\{\frac{p(x)}{q(x)}\}dx \\ &= ln(\frac{s}{\sigma}) - (\frac{1}{2\sigma^2} - \frac{1}{2s^2})(\mu^2+\sigma^2) + (\frac{\mu}{\sigma^2} - \frac{m}{s^2})\mu - (\frac{\mu^2}{2\sigma^2} - \frac{m^2}{2s^2}) \\ &= ln(\frac{s}{\sigma}) + \frac{\sigma^2 + (\mu-m)^2}{2s^2} - \frac{1}{2} \end{aligned}$$

We will discuss this result in more detail. Firstly, if KL distance is defined in *Information Theory*, the first term of the result will be $log_2(\frac{s}{\sigma})$ instead of $ln(\frac{s}{\sigma})$. Secondly, if we denote $x = \frac{s}{\sigma}$, KL distance can be rewritten as :

$$KL(p||q) = ln(x) + \frac{1}{2x^2} - \frac{1}{2} + a, \quad \text{where } a = \frac{(\mu - m)^2}{2s^2}$$

We calculate the derivative of KL with respect to x, and let it equal to 0:

$$\frac{d(KL)}{dx} = \frac{1}{x} - x^{-3} = 0 \quad => \quad x = 1 \, (\because s, \sigma > 0)$$

When $x < 1$ the derivative is less than 0, and when $x > 1$, it is greater than 0, which makes $x = 1$ the global minimum. When $x = 1$, $KL(p||q) = a$. What's more, when $\mu = m$, a will achieve its minimum 0. In this way, we have shown that the KL distance between two Gaussian Distributions is not less than 0, and only when the two Gaussian Distributions are identical, i.e. having same mean and variance, KL distance will equal to 0.

Problem 1.31 Solution

We evaluate $H[\mathbf{x}] + H[\mathbf{y}] - H[\mathbf{x},\mathbf{y}]$ by definition. Firstly, let's calculate $H[\mathbf{x},\mathbf{y}]$:

$$\begin{aligned} H[\mathbf{x},\mathbf{y}] &= -\int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x},\mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= -\int p(\mathbf{x}) ln p(\mathbf{x}) d\mathbf{x} - \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= H[\mathbf{x}] + H[\mathbf{y}|\mathbf{x}] \end{aligned}$$

Where we take advantage of $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$, $\int p(\mathbf{x},\mathbf{y}) d\mathbf{y} = p(\mathbf{x})$ and (1.111). Therefore, we have actually solved Prob.1.37 here. We will continue our proof for this problem, based on what we have proved:

$$\begin{aligned} H[\mathbf{x}] + H[\mathbf{y}] - H[\mathbf{x},\mathbf{y}] &= H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}] \\ &= -\int p(\mathbf{y}) ln p(\mathbf{y}) d\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{y}) d\mathbf{x} d\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x},\mathbf{y}) ln \big(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\big) d\mathbf{x} d\mathbf{y} \\ &= KL(p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})) = I(\mathbf{x},\mathbf{y}) \geq 0 \end{aligned}$$

Where we take advantage of the following properties:

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

$$\frac{p(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})} = \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}$$

Moreover, it is straightforward that if and only if $\mathbf{x}$ and $\mathbf{y}$ is statistically independent, the equality holds, due to the property of *KL distance*. You can also view this result by :

$$\begin{aligned} H[\mathbf{x},\mathbf{y}] &= -\int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x},\mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= -\int p(\mathbf{x}) ln p(\mathbf{x}) d\mathbf{x} - \int\int p(\mathbf{y}) ln p(\mathbf{y}) d\mathbf{y} \\ &= H[\mathbf{x}] + H[\mathbf{y}] \end{aligned}$$

Problem 1.32 Solution

It is straightforward based on definition and note that if we want to change variable in integral, we have to introduce a redundant term called *Jacobian Determinant*.

$$\begin{aligned} H[\mathbf{y}] &= -\int p(\mathbf{y}) ln p(\mathbf{y}) d\mathbf{y} \\ &= -\int \frac{p(\mathbf{x})}{|\mathbf{A}|} ln \frac{p(\mathbf{x})}{|\mathbf{A}|} |\frac{\partial \mathbf{y}}{\partial \mathbf{x}}| d\mathbf{x} \\ &= -\int p(\mathbf{x}) ln \frac{p(\mathbf{x})}{|\mathbf{A}|} d\mathbf{x} \\ &= -\int p(\mathbf{x}) ln p(\mathbf{x}) d\mathbf{x} - \int p(\mathbf{x}) ln \frac{1}{|\mathbf{A}|} d\mathbf{x} \\ &= H[\mathbf{x}] + ln|\mathbf{A}| \end{aligned}$$

Where we have taken advantage of the following equations:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A} \quad \text{and} \quad p(\mathbf{x}) = p(\mathbf{y})|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}| = p(\mathbf{y})|\mathbf{A}|$$

$$\int p(\mathbf{x}) d\mathbf{x} = 1$$

Problem 1.33 Solution

Based on the definition of *Entropy*, we write:

$$H[y|x] = -\sum_{x_i}\sum_{y_j} p(x_i, y_j) ln p(y_j|x_i)$$

Considering the property of *probability*, we can obtain that $0 \le p(y_j|x_i) \le 1$, $0 \le p(x_i, y_j) \le 1$. Therefore, we can see that $-p(x_i, y_j) ln p(y_j|x_i) \ge 0$ when $0 < p(y_j|x_i) \le 1$. And when $p(y_j|x_i) = 0$, provided with the fact that $\lim_{p\to 0} p ln p =$

0, we can see that $-p(x_i,y_j)lnp(y_j|x_i) = -p(x_i)p(y_j|x_i)lnp(y_j|x_i) \approx 0$, (here we view $p(x)$ as a constant). Hence for an arbitrary term in the equation above, we have proved that it can not be less than 0. In other words, if and only if every term of $H[y|x]$ equals to 0, $H[y|x]$ will equal to 0.

Therefore, for each possible value of random variable x, denoted as x_i :

$$-\sum_{y_j} p(x_i,y_j)lnp(y_j|x_i) = 0 \qquad (*)$$

If there are more than one possible value of random variable y given $x = x_i$, denoted as y_j, such that $p(y_j|x_i) \neq 0$ (Because x_i, y_j are both "possible", $p(x_i,y_j)$ will also not equal to 0), constrained by $0 \leq p(y_j|x_i) \leq 1$ and $\sum_j p(y_j|x_i) = 1$, there should be at least two value of y satisfied $0 < p(y_j|x_i) < 1$, which ultimately leads to $(*) > 0$.

Therefore, for each possible value of x, there will only be one y such that $p(y|x) \neq 0$. In other words, y is determined by x. Note: This result is quite straightforward. If y is a function of x, we can obtain the value of y as soon as observing a x. Therefore we will obtain no additional information when observing a y_j given an already observed x.

Problem 1.34 Solution

This problem is complicated. We will explain it in detail. According to Appenddix D, we can obtain the relation,i.e. (D.3) :

$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \int \frac{\partial F}{\partial y}\epsilon\eta(x)dx \qquad (**)$$

Where $y(x)$ can be viewed as an operator that for any input x it will give an output value y, and equivalently, $F[y(x)]$ can be viewed as an functional operator that for any input value $y(x)$, it will give an ouput value $F[y(x)]$. Then we consider a functional operator:

$$I[p(x)] = \int p(x)f(x)dx$$

Under a small variation $p(x) \rightarrow p(x)+\epsilon\eta(x)$, we will obtain :

$$I[p(x)+\epsilon\eta(x)] = \int p(x)f(x)dx + \int \epsilon\eta(x)f(x)dx$$

Comparing the equation above and $(*)$, we can draw a conclusion :

$$\frac{\partial I}{\partial p(x)} = f(x)$$

Similarly, let's consider another functional operator:

$$J[p(x)] = \int p(x)lnp(x)dx$$

Then under a small variation $p(x) \rightarrow p(x) + \epsilon\eta(x)$:

$$\begin{aligned} J[p(x)+\epsilon\eta(x)] &= \int (p(x)+\epsilon\eta(x))\, ln(p(x)+\epsilon\eta(x))\, dx \\ &= \int p(x)\, ln(p(x)+\epsilon\eta(x))\, dx + \int \epsilon\eta(x)\, ln(p(x)+\epsilon\eta(x))\, dx \end{aligned}$$

Note that $\epsilon\eta(x)$ is much smaller than $p(x)$, we will write its *Taylor Theorems* at point $p(x)$:

$$ln(p(x)+\epsilon\eta(x)) = lnp(x) + \frac{\epsilon\eta(x)}{p(x)} + O(\epsilon\eta(x)^2)$$

Therefore, we substitute the equation above into $J[p(x)+\epsilon\eta(x)]$:

$$J[p(x)+\epsilon\eta(x)] = \int p(x)\, lnp(x) dx + \epsilon\eta(x) \int (lnp(x)+1) dx + O(\epsilon^2)$$

Therefore, we also obtain :

$$\frac{\partial J}{\partial p(x)} = lnp(x) + 1$$

Now we can go back to (1.108). Based on $\frac{\partial J}{\partial p(x)}$ and $\frac{\partial I}{\partial p(x)}$, we can calculate the derivative of the expression just before (1.108) and let it equal to 0:

$$-lnp(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

Hence we rearrange it and obtain (1.108). From (1.108) we can see that $p(x)$ should take the form of a Gaussian distribution. So we rewrite it into Gaussian form and then compare it to a Gaussian distribution with mean μ and variance σ^2, it is straightforward:

$$exp(-1+\lambda_1) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \quad , \quad exp(\lambda_2 x + \lambda_3 (x-\mu)^2) = exp\{\frac{(x-\mu)^2}{2\sigma^2}\}$$

Finally, we obtain :

$$\lambda_1 = 1 - \ln(2\pi\sigma^2)$$

$$\lambda_2 = 0$$

$$\lambda_3 = \frac{1}{2\sigma^2}$$

Problem 1.35 Solution

If $p(x) = \mathcal{N}(\mu, \sigma^2)$, we write its entropy:

$$\begin{aligned} H[x] &= -\int p(x) ln p(x) dx \\ &= -\int p(x) ln\{\frac{1}{2\pi\sigma^2}\} dx - \int p(x)\{-\frac{(x-\mu)^2}{2\sigma^2}\} dx \\ &= -ln\{\frac{1}{2\pi\sigma^2}\} + \frac{\sigma^2}{2\sigma^2} \\ &= \frac{1}{2}\{1 + ln(2\pi\sigma^2)\} \end{aligned}$$

Where we have taken advantage of the following properties of a Gaussian distribution:

$$\int p(x) dx = 1 \text{ and } \int (x-\mu)^2 p(x) dx = \sigma^2$$

Problem 1.36 Solution

Here we should make it clear that if the second derivative is strictly positive, the function must be strictly convex. However, the converse may not be true. For example $f(x) = x^4$, $g(x) = x^2$, $x \in \mathscr{R}$ are both strictly convex by definition, but their second derivatives at $x = 0$ are both indeed 0 (See keyword convex function on Wikipedia or Page 71 of the book Convex Optimization written by Boyd, Vandenberghe for more details). Hence, here more precisely we will prove that a convex function is equivalent to its second derivative is non-negative by first considering *Taylor Theorems*:

$$f(x+\epsilon) = f(x) + \frac{f'(x)}{1!}\epsilon + \frac{f''(x)}{2!}\epsilon^2 + \frac{f'''(x)}{3!}\epsilon^3 + ...$$

$$f(x-\epsilon) = f(x) - \frac{f'(x)}{1!}\epsilon + \frac{f''(x)}{2!}\epsilon^2 - \frac{f'''(x)}{3!}\epsilon^3 + ...$$

Then we can obtain the expression of $f''(x)$:

$$f''(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) + f(x-\epsilon) - 2f(x)}{\epsilon^2}$$

Where $O(\epsilon^4)$ is neglected and if $f(x)$ is convex, we can obtain:

$$f(x) = f(\frac{1}{2}(x+\epsilon) + \frac{1}{2}(x-\epsilon)) \leq \frac{1}{2}f(x+\epsilon) + \frac{1}{2}f(x-\epsilon)$$

Hence $f''(x) \geq 0$. The converse situation is a little bit complex, we will use *Lagrange form of Taylor Theorems* to rewrite the Taylor Series Expansion above :

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x^\star)}{2}(x-x_0)$$

Where $x^\star$ lies between x and x_0. By hypothesis, $f''(x) \geq 0$, the last term is non-negative for all x. We let $x_0 = \lambda x_1 + (1-\lambda)x_2$, and $x = x_1$:

$$f(x_1) \geq f(x_0) + (1-\lambda)(x_1 - x_2)f'(x_0) \qquad (*)$$

And then, we let $x = x_2$:

$$f(x_2) \geq f(x_0) + \lambda(x_2 - x_1)f'(x_0) \qquad (**)$$

We multiply $(*)$ by λ, $(**)$ by $1-\lambda$ and then add them together, we will see :

$$\lambda f(x_1) + (1-\lambda)f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$$

Problem 1.37 Solution

See Prob.1.31.

Problem 1.38 Solution

When $M = 2$, (1.115) will reduce to (1.114). We suppose (1.115) holds for M, we will prove that it will also hold for $M+1$.

$$\begin{aligned} f(\sum_{m=1}^{M} \lambda_m x_m) &= f(\lambda_{M+1}x_{M+1} + (1-\lambda_{M+1}) \sum_{m=1}^{M} \frac{\lambda_m}{1-\lambda_{M+1}} x_m) \\ &\leq \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1}) f(\sum_{m=1}^{M} \frac{\lambda_m}{1-\lambda_{M+1}} x_m) \\ &\leq \lambda_{M+1} f(x_{M+1}) + (1-\lambda_{M+1}) \sum_{m=1}^{M} \frac{\lambda_m}{1-\lambda_{M+1}} f(x_m) \\ &\leq \sum_{m=1}^{M+1} \lambda_m f(x_m) \end{aligned}$$

Hence, *Jensen's Inequality*, i.e. (1.115), has been proved.

Problem 1.39 Solution

It is quite straightforward based on definition.

$$H[x] = -\sum_i p(x_i) ln p(x_i) = -\frac{2}{3} ln \frac{2}{3} - \frac{1}{3} ln \frac{1}{3} = 0.6365$$

$$H[y] = -\sum_i p(y_i) ln p(y_i) = -\frac{2}{3} ln \frac{2}{3} - \frac{1}{3} ln \frac{1}{3} = 0.6365$$

$$H[x,y] = -\sum_{i,j} p(x_i, y_j) ln p(x_i, y_j) = -3 \cdot \frac{1}{3} ln \frac{1}{3} - 0 = 1.0986$$

$$H[x|y] = -\sum_{i,j} p(x_i, y_j) ln p(x_i|y_j) = -\frac{1}{3} ln 1 - \frac{1}{3} ln \frac{1}{2} - \frac{1}{3} ln \frac{1}{2} = 0.4621$$

$$H[y|x] = -\sum_{i,j} p(x_i, y_j) ln p(y_j|x_i) = -\frac{1}{3} ln \frac{1}{2} - -\frac{1}{3} ln \frac{1}{2} - \frac{1}{3} ln 1 = 0.4621$$

$$\begin{aligned} I[x,y] &= -\sum_{i,j} p(x_i, y_j) ln \frac{p(x_i)p(y_j)}{p(x_i, y_j)} \\ &= -\frac{1}{3} ln \frac{\frac{2}{3} \cdot \frac{1}{3}}{1/3} - \frac{1}{3} ln \frac{\frac{2}{3} \cdot \frac{2}{3}}{1/3} - \frac{1}{3} ln \frac{\frac{1}{3} \cdot \frac{2}{3}}{1/3} = 0.1744 \end{aligned}$$

Their relations are given below, diagrams omitted.

$$I[x,y] = H[x] - H[x|y] = H[y] - H[y|x]$$

$$H[x,y] = H[y|x] + H[x] = H[x|y] + H[y]$$

Problem 1.40 Solution

$f(x) = lnx$ is actually a strict concave function, therefore we take advantage of *Jensen's Inequality* to obtain:

$$f(\sum_{i=1}^{M} \lambda_m x_m) \geq \sum_{i=1}^{M} \lambda_m f(x_m)$$

We let $\lambda_m = \frac{1}{M}, m = 1, 2, ..., M$. Hence we will obtain:

$$ln(\frac{x_1 + x_2 + ... + x_m}{M}) \geq \frac{1}{M}[ln(x_1) + ln(x_2) + ... + ln(x_M)] = \frac{1}{M} ln(x_1 x_2 ... x_M)$$

We take advantage of the fact that $f(x) = lnx$ is strictly increasing and then obtain :

$$\frac{x_1 + x_2 + ... + x_m}{M} \geq \sqrt[M]{x_1 x_2 ... x_M}$$

Problem 1.41 Solution

Based on definition of $I[\mathbf{x},\mathbf{y}]$, i.e.(1.120), we obtain:

$$\begin{aligned} I[\mathbf{x},\mathbf{y}] &= -\int\int p(\mathbf{x},\mathbf{y}) ln \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})} d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x},\mathbf{y}) ln \frac{p(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})} d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x}) d\mathbf{x} d\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= -\int\int p(\mathbf{x}) ln p(\mathbf{x}) d\mathbf{x} + \int\int p(\mathbf{x},\mathbf{y}) ln p(\mathbf{x}|\mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] \end{aligned}$$

Where we have taken advantage of the fact: $p(\mathbf{x},\mathbf{y}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$, and $\int p(\mathbf{x},\mathbf{y}) d\mathbf{y} = p(\mathbf{x})$. The same process can be used for proving $I[\mathbf{x},\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]$, if we substitute $p(\mathbf{x},\mathbf{y})$ with $p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$ in the second step.

0.2 Probability Distribution

Problem 2.1 Solution

Based on definition, we can obtain :

$$\sum_{x_i=0,1} p(x_i) = \mu + (1-\mu) = 1$$

$$\mathbb{E}[x] = \sum_{x_i=0,1} x_i\, p(x_i) = 0\cdot(1-\mu) + 1\cdot\mu = \mu$$

$$\begin{aligned} var[x] &= \sum_{x_i=0,1} (x_i - \mathbb{E}[x])^2\, p(x_i) \\ &= (0-\mu)^2(1-\mu) + (1-\mu)^2\cdot\mu \\ &= \mu(1-\mu) \end{aligned}$$

$$H[x] = -\sum_{x_i=0,1} p(x_i) ln p(x_i) = -\mu ln\mu - (1-\mu) ln(1-\mu)$$

Problem 2.2 Solution

The proof in Prob.2.1. can also be used here.

$$\sum_{x_i=-1,1} p(x_i) = \frac{1-\mu}{2} + \frac{1+\mu}{2} = 1$$

$$\mathbb{E}[x] = \sum_{x_i=-1,1} x_i\, p(x_i) = -1\cdot\frac{1-\mu}{2} + 1\cdot\frac{1+\mu}{2} = \mu$$

$$\begin{aligned} var[E] &= \sum_{x_i=-1,1} (x_i - \mathbb{E}[x])^2 p(x_i) \\ &= (-1-\mu)^2\cdot\frac{1-\mu}{2} + (1-\mu)^2\cdot\frac{1+\mu}{2} \\ &= (1-\mu)^2 \end{aligned}$$

$$H[x] = -\sum_{x_i=-1,1} p(x_i) ln p(x_i) = -\frac{1-\mu}{2} ln\frac{1-\mu}{2} - \frac{1+\mu}{2} ln\frac{1+\mu}{2}$$

Problem 2.3 Solution

(2.262) is an important property of *Combinations*, which we have used before, such as in Prob.1.15. We will use the 'old fashioned' denotation C_N^m to represent choose m objects from a total of N. With the prior knowledge:

$$C_N^m = \frac{N!}{m!(N-m)!}$$

We evaluate the left side of (2.262) :

$$\begin{aligned} C_N^m + C_N^{m-1} &= \frac{N!}{m!(N-m)!} + \frac{N!}{(m-1)!(N-(m-1))!} \\ &= \frac{N!}{(m-1)!(N-m)!}(\frac{1}{m} + \frac{1}{N-m+1}) \\ &= \frac{(N+1)!}{m!(N+1-m)!} = C_{N+1}^m \end{aligned}$$

To proof (2.263), here we will proof a more general form:

$$(x+y)^N = \sum_{m=0}^{N} C_N^m x^m y^{N-m} \qquad (*)$$

If we let $y = 1$, $(*)$ will reduce to (2.263). We will proof it by induction. First, it is obvious when $N = 1$, $(*)$ holds. We assume that it holds for N, we will proof that it also holds for $N+1$.

$$\begin{aligned} (x+y)^{N+1} &= (x+y)\sum_{m=0}^{N} C_N^m x^m y^{N-m} \\ &= x\sum_{m=0}^{N} C_N^m x^m y^{N-m} + y\sum_{m=0}^{N} C_N^m x^m y^{N-m} \\ &= \sum_{m=0}^{N} C_N^m x^{m+1} y^{N-m} + \sum_{m=0}^{N} C_N^m x^m y^{N+1-m} \\ &= \sum_{m=1}^{N+1} C_N^{m-1} x^m y^{N+1-m} + \sum_{m=0}^{N} C_N^m x^m y^{N+1-m} \\ &= \sum_{m=1}^{N} (C_N^{m-1} + C_N^m) x^m y^{N+1-m} + x^{N+1} + y^{N+1} \\ &= \sum_{m=1}^{N} C_{N+1}^m x^m y^{N+1-m} + x^{N+1} + y^{N+1} \\ &= \sum_{m=0}^{N+1} C_{N+1}^m x^m y^{N+1-m} \end{aligned}$$

By far, we have proved $(*)$. Therefore, if we let $y = 1$ in $(*)$, (2.263) has been proved. If we let $x = \mu$ and $y = 1-\mu$, (2.264) has been proved.

Problem 2.4 Solution

Solution has already been given in the problem, but we will solve it in a

more intuitive way, beginning by definition:

$$\begin{aligned}
\mathbb{E}[m] &= \sum_{m=0}^{N} mC_N^m \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} mC_N^m \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} \frac{N!}{(m-1)!(N-m)!} \mu^m (1-\mu)^{N-m} \\
&= N \cdot \mu \sum_{m=1}^{N} \frac{(N-1)!}{(m-1)!(N-m)!} \mu^{m-1} (1-\mu)^{N-m} \\
&= N \cdot \mu \sum_{m=1}^{N} C_{N-1}^{m-1} \mu^{m-1} (1-\mu)^{N-m} \\
&= N \cdot \mu \sum_{k=0}^{N-1} C_{N-1}^{k} \mu^{k} (1-\mu)^{N-1-k} \\
&= N \cdot \mu [\mu + (1-\mu)]^{N-1} = N\mu
\end{aligned}$$

Some details should be explained here. We note that $m = 0$ actually doesn't affect the *Expectation*, so we let the summation begin from $m = 1$, i.e. (what we have done from the first step to the second step). Moreover, in the second last step, we rewrite the subindex of the summation, and what we actually do is let $k = m - 1$. And in the last step, we have taken advantage of (2.264). Variance is straightforward once *Expectation* has been calculated.

$$\begin{aligned}
var[m] &= \mathbb{E}[m^2] - \mathbb{E}[m]^2 \\
&= \sum_{m=0}^{N} m^2 C_N^m \mu^m (1-\mu)^{N-m} - \mathbb{E}[m] \cdot \mathbb{E}[m] \\
&= \sum_{m=0}^{N} m^2 C_N^m \mu^m (1-\mu)^{N-m} - (N\mu) \cdot \sum_{m=0}^{N} mC_N^m \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} m^2 C_N^m \mu^m (1-\mu)^{N-m} - N\mu \cdot \sum_{m=1}^{N} mC_N^m \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=1}^{N} m \frac{N!}{(m-1)!(N-m)!} \mu^m (1-\mu)^{N-m} - (N\mu) \cdot \sum_{m=1}^{N} mC_N^m \mu^m (1-\mu)^{N-m} \\
&= N\mu \sum_{m=1}^{N} m \frac{(N-1)!}{(m-1)!(N-m)!} \mu^{m-1} (1-\mu)^{N-m} - N\mu \cdot \sum_{m=1}^{N} mC_N^m \mu^m (1-\mu)^{N-m} \\
&= N\mu \sum_{m=1}^{N} m\mu^{m-1} (1-\mu)^{N-m} (C_{N-1}^{m-1} - \mu C_N^m)
\end{aligned}$$

Here we will use a little tick, $-\mu = -1 + (1-\mu)$ and then take advantage

of the property, $C_N^m = C_{N-1}^m + C_{N-1}^{m-1}$.

$$\begin{aligned}
var[m] &= N\mu \sum_{m=1}^{N} m\mu^{m-1}(1-\mu)^{N-m}\left[C_{N-1}^{m-1} - C_N^m + (1-\mu)C_N^m\right] \\
&= N\mu \sum_{m=1}^{N} m\mu^{m-1}(1-\mu)^{N-m}\left[(1-\mu)C_N^m + C_{N-1}^{m-1} - C_N^m\right] \\
&= N\mu \sum_{m=1}^{N} m\mu^{m-1}(1-\mu)^{N-m}\left[(1-\mu)C_N^m - C_{N-1}^m\right] \\
&= N\mu\left\{\sum_{m=1}^{N} m\mu^{m-1}(1-\mu)^{N-m+1}C_N^m - \sum_{m=1}^{N} m\mu^{m-1}(1-\mu)^{N-m}C_{N-1}^m\right\} \\
&= N\mu\left\{\cdot N(1-\mu)[\mu+(1-\mu)]^{N-1} - (N-1)(1-\mu)[\mu+(1-\mu)]^{N-2}\right\} \\
&= N\mu\left\{N(1-\mu) - (N-1)(1-\mu)\right\} = N\mu(1-\mu)
\end{aligned}$$

Problem 2.5 Solution

Hints have already been given in the description, and let's make a little improvement by introducing $t = y + x$ and $x = t\mu$ at the same time, i.e. we will do following changes:

$$\begin{cases} x = t\mu \\ y = t(1-\mu) \end{cases} \quad \text{and} \quad \begin{cases} t = x + y \\ \mu = \frac{x}{x+y} \end{cases}$$

Note $t \in [0, +\infty]$, $\mu \in (0,1)$, and that when we change variables in integral, we will introduce a redundant term called *Jacobian Determinant*.

$$\frac{\partial(x,y)}{\partial(\mu,t)} = \begin{vmatrix} \frac{\partial x}{\partial \mu} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial \mu} & \frac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} t & \mu \\ -t & 1-\mu \end{vmatrix} = t$$

Now we can calculate the integral.

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^{+\infty} exp(-x)x^{a-1}dx \int_0^{+\infty} exp(-y)y^{b-1}dy \\
&= \int_0^{+\infty}\int_0^{+\infty} exp(-x)x^{a-1}exp(-y)y^{b-1}dydx \\
&= \int_0^{+\infty}\int_0^{+\infty} exp(-x-y)x^{a-1}y^{b-1}dydx \\
&= \int_0^{1}\int_0^{+\infty} exp(-t)(t\mu)^{a-1}(t(1-\mu))^{b-1}t\,dt\,d\mu \\
&= \int_0^{+\infty} exp(-t)t^{a+b-1}dt \cdot \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu \\
&= \Gamma(a+b)\cdot\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu
\end{aligned}$$

Therefore, we have obtained :

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

Problem 2.6 Solution

We will solve this problem based on definition.

$$\begin{aligned}
\mathbb{E}[\mu] &= \int_0^1 \mu\, Beta(\mu|a,b)\,d\mu \\
&= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^a(1-\mu)^{b-1}\,d\mu \\
&= \frac{\Gamma(a+b)\Gamma(a+1)}{\Gamma(a+1+b)\Gamma(a)}\int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)}\mu^a(1-\mu)^{b-1}\,d\mu \\
&= \frac{\Gamma(a+b)\Gamma(a+1)}{\Gamma(a+1+b)\Gamma(a)}\int_0^1 Beta(\mu|a+1,b)\,d\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a+1+b)}\cdot\frac{\Gamma(a+1)}{\Gamma(a)} \\
&= \frac{a}{a+b}
\end{aligned}$$

Where we have taken advantage of the property: $\Gamma(z+1) = z\Gamma(z)$. For variance, it is quite similar. We first evaluate $E[\mu^2]$.

$$\begin{aligned}
\mathbb{E}[\mu^2] &= \int_0^1 \mu^2\, Beta(\mu|a,b)\,d\mu \\
&= \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a+1}(1-\mu)^{b-1}\,d\mu \\
&= \frac{\Gamma(a+b)\Gamma(a+2)}{\Gamma(a+2+b)\Gamma(a)}\int_0^1 \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)}\mu^{a+1}(1-\mu)^{b-1}\,d\mu \\
&= \frac{\Gamma(a+b)\Gamma(a+2)}{\Gamma(a+2+b)\Gamma(a)}\int_0^1 Beta(\mu|a+2,b)\,d\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a+2+b)}\cdot\frac{\Gamma(a+2)}{\Gamma(a)} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

Then we use the formula: $var[\mu] = E[\mu^2] - E[\mu]^2$.

$$\begin{aligned}
var[\mu] &= \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

Problem 2.7 Solution

The maximum likelihood estimation for μ, i.e. (2.8), can be written as :

$$\mu_{ML} = \frac{m}{m+l}$$

Where m represents how many times we observe 'head', l represents how many times we observe 'tail'. And the prior mean of μ is given by (2.15), the posterior mean value of x is given by (2.20). Therefore, we will prove that $(m+a)/(m+a+l+b)$ lies between $m/(m+l)$, $a/(a+b)$. Given the fact that :

$$\lambda\frac{a}{a+b} + (1-\lambda)\frac{m}{m+l} = \frac{m+a}{m+a+l+b} \text{ where } \lambda = \frac{a+b}{m+l+a+b}$$

We have solved problem. Note : you can also solve it in a more simple way by prove that :

$$(\frac{m+a}{m+a+l+b} - \frac{a}{a+b}) \cdot (\frac{m+a}{m+a+l+b} - \frac{m}{m+l}) \leq 0$$

The expression above can be proved by reduction of fractions to a common denominator.

Problem 2.8 Solution

We solve it base on definition.

$$\begin{aligned}
\mathbb{E}_y[\mathbb{E}_x[x|y]]] &= \int \mathbb{E}_x[x|y]p(y)dy \\
&= \int (\int x\,p(x|y)dx)p(y)dy \\
&= \int\int x\,p(x|y)p(y)dx\,dy \\
&= \int\int x\,p(x,y)dx\,dy \\
&= \int x\,p(x)dx = \mathbb{E}[x]
\end{aligned}$$

(2.271) is complicated and we will calculate every term separately.

$$\begin{aligned}
\mathbb{E}_y[var_x[x|y]] &= \int var_x[x|y]p(y)dy \\
&= \int (\int (x-\mathbb{E}_x[x|y])^2 p(x|y)dx)\,p(y)dy \\
&= \int\int (x-\mathbb{E}_x[x|y])^2 p(x,y)dx\,dy \\
&= \int\int (x^2 - 2x\mathbb{E}_x[x|y] + \mathbb{E}_x[x|y]^2)\,p(x,y)dx\,dy \\
&= \int\int x^2 p(x)dx - \int\int 2x\mathbb{E}_x[x|y]p(x,y)dx\,dy + \int\int (\mathbb{E}_x[x|y]^2)\,p(y)dy
\end{aligned}$$

About the second term in the equation above, we further simplify it :

$$\begin{aligned}\int\int 2x\mathbb{E}_x[x|y]p(x,y)dxdy &= 2\int \mathbb{E}_x[x|y](\int xp(x,y)dx)dy \\ &= 2\int \mathbb{E}_x[x|y]p(y)(\int xp(x|y)dx)dy \\ &= 2\int \mathbb{E}_x[x|y]^2p(y)dy\end{aligned}$$

Therefore, we obtain the simple expression for the first term on the right side of (2.271) :

$$\mathbb{E}_y[var_x[x|y]] = \int\int x^2p(x)dx - \int\int \mathbb{E}_x[x|y]^2p(y)dy \qquad (*)$$

Then we process for the second term.

$$\begin{aligned}var_y[\mathbb{E}_x[x|y]] &= \int(\mathbb{E}_x[x|y]-\mathbb{E}_y[\mathbb{E}_x[x|y]])^2p(y)dy \\ &= \int(\mathbb{E}_x[x|y]-\mathbb{E}[x])^2p(y)dy \\ &= \int \mathbb{E}_x[x|y]^2p(y)dy - 2\int \mathbb{E}[x]\mathbb{E}_x[x|y]p(y)dy + \int \mathbb{E}[x]^2p(y)dy \\ &= \int \mathbb{E}_x[x|y]^2p(y)dy - 2\mathbb{E}[x]\int \mathbb{E}_x[x|y]p(y)dy + \mathbb{E}[x]^2\end{aligned}$$

Then following the same procedure, we deal with the second term of the equation above.

$$2\mathbb{E}[x]\cdot\int \mathbb{E}_x[x|y]p(y)dy = 2\mathbb{E}[x]\cdot\mathbb{E}_y[\mathbb{E}_x[x|y]]] = 2\mathbb{E}[x]^2$$

Therefore, we obtain the simple expression for the second term on the right side of (2.271) :

$$var_y[\mathbb{E}_x[x|y]] = \int \mathbb{E}_x[x|y]^2p(y)dy - \mathbb{E}[x]^2 \qquad (**)$$

Finally, we add $(*)$ and $(**)$, and then we will obtain:

$$\mathbb{E}_y[var_x[x|y]] + var_y[\mathbb{E}_x[x|y]] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = var[x]$$